STATISTICAL METHODS FOR INTERPRETATION OF HIGH RESOLUTION MASS SPECTRA

PARMINDER KAUR

Dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

BOSTON
UNIVERSITY

BOSTON UNIVERSITY

COLLEGE OF ENGINEERING

Dissertation

STATISTICAL METHODS FOR INTERPRETATION OF HIGH RESOLUTION MASS SPECTRA

by

PARMINDER KAUR

M.S., Boston University, 2005

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2007

Approved by

First Reader

Peter B. O'Connor, Ph.D.
Research Associate Professor
Department of Biochemistry and
Department of Electrical and Computer Engineering
Boston University, Boston

Second Reader

Mark E. McComb, Ph.D.
Research Assistant Professor
School of Medicine
Boston University School of Medicine, Boston

Third Reader

William C. Karl, Ph.D.
Professor
Department of Electrical and Computer Engineering
Boston University, Boston

Fourth Reader

William L. Oliver, Ph.D.
Associate Professor
Department of Electrical and Computer Engineering
Boston University, Boston

Acknowledgments

It is my pleasure to acknowledge a number of people whose contribution was critical to this work. I am highly grateful to Professor Peter B. O'Connor for his valuable advice, extraordinary support, and patience. I greatly appreciate the encouragement of scientific method, flexibility, and independence he allowed me at different stages of my work. I am grateful to him for exposing me to diverse and ambitious projects in the group, which helped me gain a broader perspective of the research. I would also like to thank Professor Catherine E. Costello for her generous support and promoting a very friendly and productive environment in the lab.

Many thanks to my committee members Professor W. Clem Karl, Professor Janusz Konrad, Professor Mark McComb, Professor Bill Oliver, and Professor Hanno Steen for their insightful comments and precious time to serve on my committee.

I am fortunate to have received great work environment created by very learned and supportive colleagues. I am particularly thankful to Konstantin Aizikov, Marina A. Belyayev, Dr. Bogdan Budnik, Jason J. Cournoyer, Vera B. Ivleva, Xiaojuan Li, Dr. Cheng Lin, Raman Mathur, Dr. Susanne C. Moyer, Renee Mullen, Dr. David H. Perlman, Dr. Jason L. Pittman, Nadezda Sargaeva, and Dr. Cheng Zhao for their suggestions and discussions. I am honored to have had the opportunity to collaborate with Dr. Nicolas Clavreul, Professor Richard A. Cohen, Professor Lawreen H. Connors, Claire Dauly, Dr. Hua Huang, Professor Marc Kirschner, Dr. Michael Rape, Professor Mahadevan Sethuraman, Professor Martha Skinner, Professor Judith J. Steen, Dr. Roger Theberge, and Dr. Bruce A. Thomson. I am highly thankful to the members of Cardiovascular Proteomics Center and Mass Spectrometry Resource group, particularly Professors John Cipollo and Joe Zaia for sharing their expertise. I would also like to acknowledge Dr. Alan Rockwood for sharing the Mercury algorithm used in my research, Professor Tom Brenna, Dr. Eugene Moskovets and Dr. Gavin Sacks for their helpful comments in the isotope ratio project. Special thanks

go to the administrators Kirsten Levy and Carly Marchioni for their reliable and timely support.

Very importantly, I am indebted to my maternal grandparents S. Jagroop Singh and Smt. Punjab Kaur for their endless support and encouragement for my education. Thanks to my husband Amit Juneja for his patience and constant support. I would like to acknowledge my parents, siblings, extended family members in Punjab, and friends for their cooperation, as and when was required.

Finally, I am very thankful to National Institutes of Health, National Council for Research Resources (Grant No. P41-RR10888) and the National Heart Lung and Blood Institute (NO1HV28178), for providing the funding support.

STATISTICAL METHODS FOR INTERPRETATION OF HIGH RESOLUTION MASS SPECTRA

(Order No.)

PARMINDER KAUR

Boston University, College of Engineering, 2007

Major Professor: Peter B. O'Connor, PhD, Department of Biochemistry,
Mass Spectrometry Resource,
Cardiovascular Proteomics Center,
Boston University School of Medicine
Department of Electrical and Computer Engineering,
Boston University

ABSTRACT

Mass spectrometry is a powerful analytical technique used to characterize various compounds. A mass spectrum is a graph of ion intensity as a function of mass-to-charge ratio.

Protein study experiments generate thousands of mass spectra, generating an overload of data that necessitates the development of sophisticated data analysis methods. Our work aims at developing the following methods that allow for extraction of biochemically relevant information from mass spectra.

The maximum likelihood estimator together with the non-random parameter estimation method has been used to derive the mathematical relationship between the number of ions generated in a mass spectrometry experiment and the variance in the experimental isotopic distribution in a spectrum. Performance analysis of the method has been carried out using simulated and experimental data. The method can show a factor of two improvement over a previously developed method, and is applicable for any isotopically resolved spectrum.

A theoretical framework has been developed and tested against experiments for estimating high-precision elemental isotopic abundances from the experimental isotopic distributions. Higher molecular weights are particularly useful for a better estimate because the higher number of carbon atoms and isotopic peaks observed lead to a greater amount of information. This method circumvents some of the limitations experienced by the traditional isotope ratio mass spectrometry.

Charge state determination requires methods to accurately estimate the m/z difference between adjacent isotopic peaks. A new method for charge state determination using the Matched Filter approach has been developed and compared with the established methods under various conditions. Matched Filter method performs significantly better than the existing methods and has a particular advantage in cases involving overlapping isotopic distributions and low signal-to-noise ratio cases.

Algorithms have been developed and integrated as MasSPIKE (Mass Spectrum Interpretation and Kernel Extraction) for isotopic cluster identification, charge state determination, resolving overlapping isotopic distributions, alignment of the experimental isotopic distribution with the theoretical isotopic distribution, and reducing the isotopically resolved mass spectrum to a monoisotopic mass list.

MasSPIKE has been used to characterize post-translational modifications for biologically interesting proteins Hemoglobin and H-Ras, allowing for differentiation of blood samples of diseased and healthy persons.

Contents

List of Tables

List of Figures

List of Abbreviations and Definitions

CAD	.....	Collisionally Activated Dissociation
1 Dalton (Da)	.....	$\frac{1}{12}$ the mass of Carbon 12, which is assigned a mass of 12 Daltons
ECD	.....	Electron Capture Dissociation
EID	.....	Experimental Isotopic Distribution
FT-ICR	.....	Fourier Transform Ion Cyclotron Resonance
FTMS	.....	Fourier Transform Mass Spectrometer
FWHM	.....	Full-Width at Half Maximum
HPLC	.....	High Pressure Liquid Chromatography
ICR	.....	Ion Cyclotron Resonance
ID	.....	Isotopic Distribution
IRMS	.....	Isotope Ratio Mass Spectrometry
LC	.....	Liquid Chromatography
Mass Accuracy (in parts-per-million)	.....	$\frac{(M_{Exp}-M_{Theo})\times 10^6}{M_{Theo}}$, M_{Exp}=Experimentally observed mass, M_{Theo}=Theoretical mass value
MasSPIKE	.....	Mass SPectrum Interpretation and Kernel Extraction
MS	.....	Mass Spectrometry
MS/MS	.....	Mass Spectrometry/Mass Spectrometry (Multistage mass spectrometry)
m/z	.....	mass-to-charge ratio

PTM	.....	Post-Translational Modification
Resolving Power	.....	$\frac{\frac{m}{z} \text{ location of peak}}{\text{peak width at FWHM}}$, FWHM=Full Width at Half Maximum
SNR	.....	Signal-to-Noise Ratio
SORI	.....	Sustained Off-Resonance Irradiation
TID	.....	Theoretical Isotopic Distribution
1 unit charge (e)	.....	magnitude of charge on an electron (1.6×10^{-19} Coulombs)

Chapter 1

Introduction

1.1 Significance

Mass Spectrometry (McLafferty and Turecek, 1993; McLafferty et al., 1999; Aebersold, 2003; Aebersold and Mann, 2003; Biemann, 1995) is a powerful analytical technique used for the analysis of large molecules. It is used to identify and quantify unknown compounds, determine molecular masses of large biological samples,(O'Connor and McLafferty, 1995) elucidate their structural and quantitative information, and investigate intermolecular reactions. These properties hold high significance for an analytical chemist or a life scientist in order to understand the behavior of biomolecules that control biological systems and, in turn, control our bodies. Mass spectrometry provides valuable information to a wide range of professionals: chemists, biologists, astronomers, and physicians, to name a few. For example, it is used to detect and identify the use of steroids in athletes, monitor the breath of patients by anesthesiologists during surgery, determine the composition of molecular species found in space, and determine how drugs are used by the body. It is a highly sensitive approach (one part in 10^{18} in a clean sample derived from chemically complex mixtures can be detected).(Moyer et al., 2003) One very important point is that mass spectrometers do not measure mass, they measure the mass-to-charge ratio of the ions formed from the molecules (called m/z ratio, where m=molecular weight in Daltons of the molecule under consideration; z=number of unit charges on the molecule; 1 unit charge $\approx 1.6 \times 10^{-19}$ Coulombs, 1 Dalton (Da)$\approx 1.6 \times 10^{-17}$ Kg). Thus, the molecules need to be ionized in order to be detected by an instrument.

Proteomics (Aebersold and Mann, 2003; Aebersold, 2003; Tyers and Mann, 2003) is the systematic and comprehensive study of diverse properties of proteins to unravel the biological processes responsible for health and disease. The rapid advancement of mass spectrometric technologies in the last two decades has revolutionized protein research. This revolution started with the invention of two new ionization techniques, Matrix Assisted Laser Desorption/Ionization (MALDI)(Karas et al., 1985; Karas et al., 1987; Karas and Hillenkamp, 1988; Tanaka et al., 1988), and Electrospray Ionization (ESI)(Fenn et al., 1990; Fenn et al., 1989), for which the Nobel prize in Chemistry was awarded in 2002. These methods allow a researcher to ionize proteins and peptides, transfer them into the gas phase and into the mass spectrometer for mass analysis, and to do so without analyte (fragile charged molecule) fragmentation.

Once ionized, a single mass from a protein/peptide mixture can be isolated and fragmented (a technique called tandem mass spectrometry or MS/MS) to generate structural information about the selection, such as sequence, post-translational modification (PTM) identity and isolation, crosslinks, etc. This can be done multiple times (called MS^n). Thus, methods to dissociate peptide and proteins are important, and further advancements in the field were made with the invention of new odd-electron fragmentation methods like Electron Capture Dissociation (ECD)(Zubarev et al., 1998; Zubarev, 2006), Electron Detachment Dissociation (EDD)(Budnik et al., 2001), and Electron Transfer Dissociation (ETD)(Coon et al., 2004; Syka et al., 2004). The odd-electron fragmentation methods complement the older collisionally activated fragmentation methods by allowing complete sequencing of peptides,(Nielsen et al., 2005) better localization of post-translational modifications (PTMs), and providing complementary information for the comparison of raw data against the protein sequence databases.

1.2 Fourier Transform-Ion Cyclotron Resonance Mass Spectrometer (FT-ICR MS or FTMS)

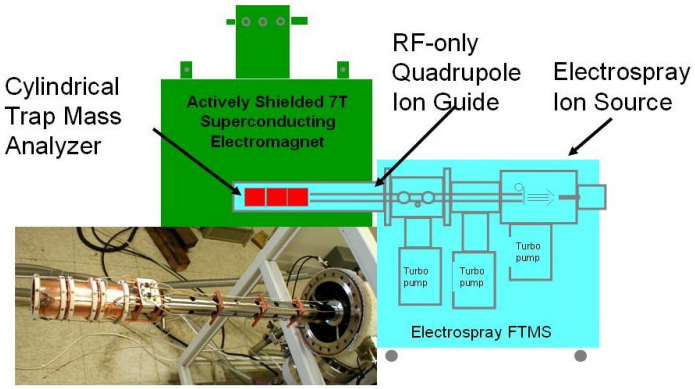
There are many different types of commercially available mass analyzers, each with different strengths and limitations. FTMS (Marshall and Verdun, 1990; Comisarow and Marshall, 1974; Marshall, 2000; Marshall et al., 1998; Amster, 1996; Gross and Rempel, 1984; Zhang et al., 2005) is a preferred type of instrument due to the superior resolving power (100,000 typically, and 1,000,000 with patience, discussed below) and mass accuracy (1 parts-per-million (ppm) when internally calibrated (calibration and measurement done in the same spectrum); 2-5 ppm when externally calibrated). High resolution is desirable in order to separate closely mass spaced mixture components, to observe fragments of the same component that are close in mass but have different elemental compositions, and for accurate assignment of masses.(Zhang et al., 2005; Spengler, 2004)

As shown in Fig 1·1a, an FTMS consists of an ion source, followed by ion optics to transfer the ions through the magnetic field gradient (in this case an RF-Only Quadrupole ion guide), into the ICR (Ion Cyclotron Resonance) cell or Penning trap. (Note: The two terms are often used interchangeably, but they rely on subtly different detection methods.) Alan G. Marshall and Mel Comisarow (Comisarow and Marshall, 1974; Marshall, 2000), were the first to recognize that inductive detection of the cyclotron motion (Lawrence and Cooksey, 1936) followed by use of the Fourier Transform (Oppenheim et al., 2002) (later the Fast Fourier Transform (Cooley and Tukey, 1965)) would allow the cyclotron to become a high performance mass spectrometer.

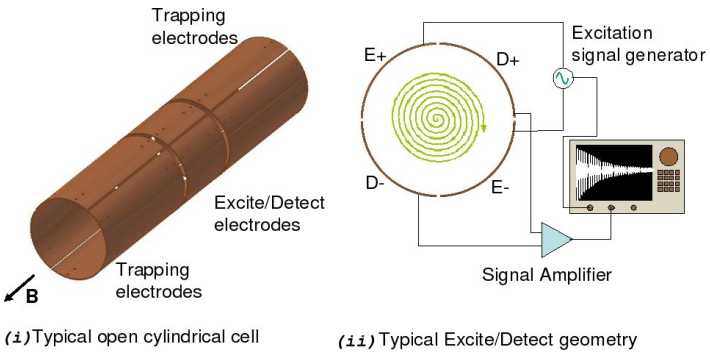
Ions have a fundamental oscillation frequency in a magnetic field given by:

$$\omega_c = \frac{zB}{m} \tag{1.1}$$

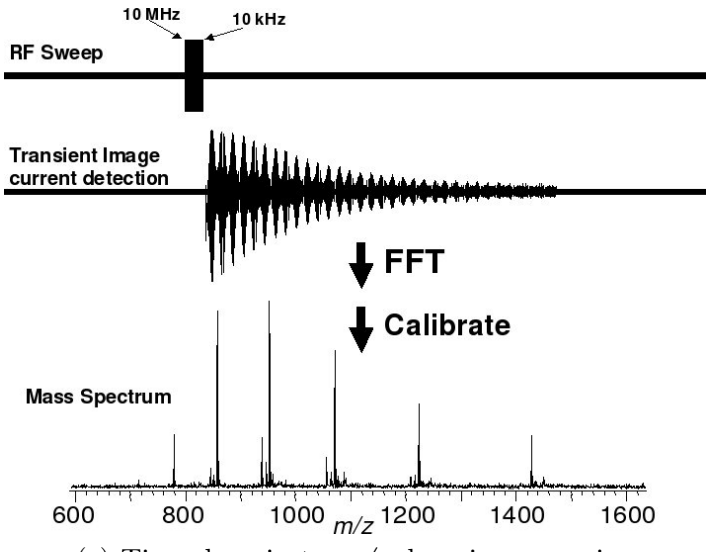
where ω_c=cyclotron frequency, z=elementary charge on the ion, B=strength of magnetic field, m=mass of the ion. In SI units, these can be measured in Hertz (Hz), Coulombs



(a) Block Diagram of an FTMS



(b) ICR cell operation



(c) Time domain to m/z domain conversion

Figure 1·1: A general diagram of FTMS operation

(C), Tesla (T), and kilograms (kg) respectively. Equation 1.1 states that the cyclotron frequency of an ion is independent of its velocity, and hence, it is also independent of its kinetic energy. This is not true for most other types of mass spectrometers, where the ions are separated by their m/z values due to the spread in their kinetic energies. (Note: The orbitrap is an exception.(Makarov et al., 2006)) The independence of an ion's cyclotron frequency on its kinetic energy is one of the key reasons why FTMS instruments are capable of extremely high resolving power.

After ions are trapped in the ICR cell (A modern ICR cell is drawn in Fig 1·1b(i), a typical diameter could be ∼7 cm), they are excited by a resonant excitation pulse into a coherent orbit as illustrated in Fig 1·1b(ii). The excitation amplifier is then turned off and the ions continue to orbit at their final radius. Ions moving near electrodes cause an image charge to form on these electrodes to balance the ions' electric field. Since the orbit is circular, the image charge induced on the detection plates will oscillate at the ions' resonant frequency, generating a sine wave between the detection plates which can be detected by a sensitive preamplifier circuit, digitized, and stored in computer memory. Since we wish to detect all ions (particularly those we don't expect) within some known mass range, we must sweep through all the possible frequencies (typically 10 KHz-10 MHz) to excite and detect all the corresponding ions. The detected transient image current is a time-domain signal which is then Fourier transformed to get the the signal intensity contributions as a function of frequency of various ions to the transient. Application of a dc voltage (V_{T}) to the trap plates of the ICR cell confines ions along the magnetic field direction, by introducing a potential that varies quadratically with axial z position (to a first order approximation). This leads to the harmonic oscillations of ions in the z direction at an angular frequency, ω_z (Brown and Gabrielse, 1986)

$$\omega_z = \sqrt{\frac{2\,q\,V_{\mathrm{T}}\,\alpha}{m\,a^2}} \tag{1.2}$$

in which α, the trapping scale factor, depends upon the trap geometry, and ranges typically from 2 to 4, and a is a characteristic dimension (i.e., distance between the trap plates) of the ICR cell. The quadratic variation in electrostatic potential as a function of z is also accompanied by a quadratic variation as a function of radial position r.(Marshall et al., 1998) This outward-directed force hence slightly reduces the effect of the radially inward-directed Lorentz force responsible for cyclotron rotation. In a plane perpendicular to the magnetic field, the force acting on an ion is:

$$F = m\,a = m\omega^2 r = qB_0\omega r - \frac{qV_{\mathrm{T}}\alpha}{a^2}r \tag{1.3}$$

where r is the radial distance between the ion and z axis. Rewriting the above equation:

$$\omega^2 - \frac{qB_0\omega}{m} + \frac{qV_{\mathrm{T}}\alpha}{ma^2} = 0 \tag{1.4}$$

Solving equation 1.4 yields the following two solutions for rotational frequency in a plane perpendicular to the magnetic field (with ω_z and ω_c defined above):

$$\omega_+ = \frac{\omega_c}{2} + \sqrt{(\frac{\omega_c}{2})^2 - \frac{\omega_z^2}{2}} \tag{1.5}$$

$$\omega_- = \frac{\omega_c}{2} - \sqrt{(\frac{\omega_c}{2})^2 - \frac{\omega_z^2}{2}} \tag{1.6}$$

ω_+ is close to the unperturbed cyclotron frequency (equation 1.1), and is called the "reduced ion cyclotron frequency", while ω_- is called the "magnetron frequency", representing a new "magnetron" motion.(Amster, 1996) Assuming that $\omega_+ = \omega_c$, rearranging equation 1.4, and substituting for ω_c from equation 1.1 gives:

$${\omega_+}^2 - \frac{qB_0\omega_+}{m} + \frac{qV_{\mathrm{T}}\alpha}{ma^2} = 0 \tag{1.7}$$

Using the fact that $q = ze$, in which z is the number of elementary charges per ion and e is the elementary charge, and multiplying equation 1.7 by $\frac{m}{{\omega_+}^2}$ leads to the following frequency to mass conversion relation (Ledford et al., 1984):

$$\frac{m}{ze} = \frac{A_{\text{Ledford}}}{\omega_+} + \frac{B_{\text{Ledford}}}{{\omega_+}^2} \tag{1.8}$$

where

$$A_{\text{Ledford}} = B_0 \tag{1.9}$$

$$B_{\text{Ledford}} = \frac{V_{\text{T}}\alpha}{a^2} \tag{1.10}$$

Another similar approach for calibration proposed by Francl *et al.* leads to the following relation between m/z and frequency:(Francl et al., 1983)

$$\frac{m}{z} = \frac{A_{\text{Francl}}}{B_{\text{Francl}} + \omega_+} \tag{1.11}$$

in which

$$A_{\text{Francl}} = eB_0 \tag{1.12}$$

$$B_{\text{Francl}} = \frac{V_{\text{T}}\alpha}{B_0 a^2} \tag{1.13}$$

During calibration, constants A_{Francl}, A_{Ledford}, B_{Francl}, and B_{Ledford} can be determined by calibrant samples of known m/z values, and Shi et al. have shown that these two calibration methods are essentially equivalent.(Shi, 2000)

The time-domain signal recorded from the detect plates is real, and its frequency-domain representation after fast Fourier transform is complex and symmetrical. The final mass spectrum plot is normally obtained by taking the magnitude of the Fourier transform of the detected signal. In a typical FT-ICR mass spectrum, the detection process has to be delayed until after the excitation in order to avoid the saturation of the detection preamplifier by capacitive crosstalk from the excite plates. That delay, along with contributions from other factors such as a temporally dispersed excitation event (e.g., frequency sweep), causes a continuous variation of phase with frequency in the time-domain data. Thus, the initial time-domain phase, $\phi(t_0)$, varies with frequency. However, this phase information is absent

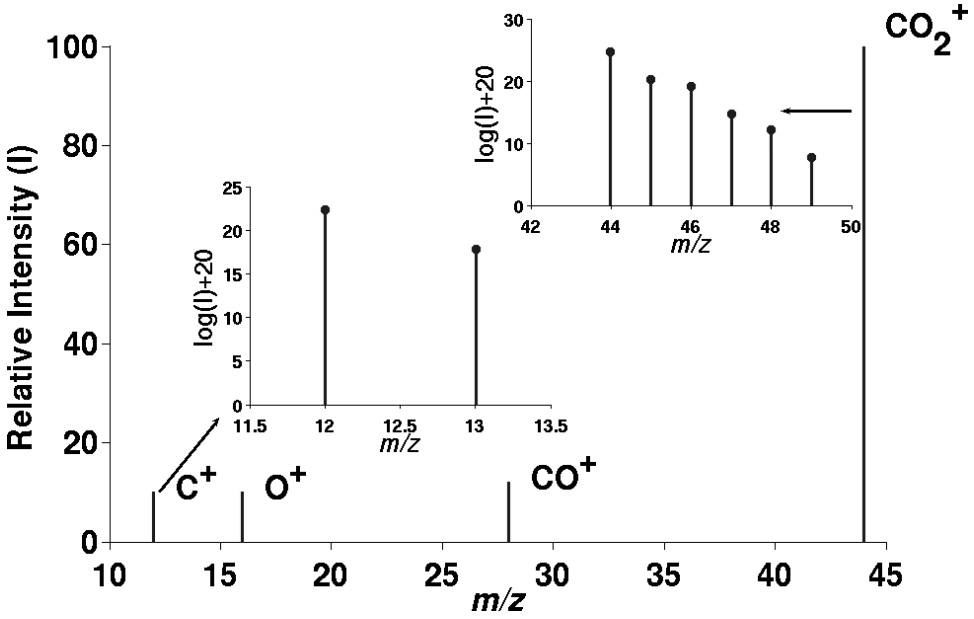


Figure 1·2: Theoretical mass spectrum of CO_2

from the detected time-domain signal due to the delayed detection process, and leads to loss in spectral resolution in the frequency-domain.(Beu et al., 2004) Special techniques have been proposed recently for simultaneous excitation and detection processes, leading to increased resolving power.(Beu et al., 2004)

1.3 Mass Spectrum

The output generated by a mass spectrometry experiment is a graph of ion intensity as a function of m/z ratio. Fig 1·2 shows a mass spectrum of the simple molecule carbon dioxide, CO_2. This record of ions and their intensities serves to establish the molecular weight and structure of the compound being mass analyzed. In this example, all the ions are positively charged (It is possible to generate and detect negative ions as well). The ionized CO_2 molecule (or molecular ion) appears at m/z 44. The ion is singly charged and the "nominal ion mass" (integer mass value) is 44 Da: carbon=12 and oxygen=16 (in calculating nominal ion mass, atomic masses are rounded to the nearest integer). Although only one peak is the most prominent at $m/z \approx$44 representing the monoisotopic mass of the

intact molecule CO_2(one ^{12}C and two ^{16}O atoms), there are six different isotopes possible, even for this simple molecule. For example, there is a possibility of observing the following combinations of isotopes at different nominal masses:

Nominal Mass	Isotope Combinations
44	$^{12}C_1\,^{16}O_2$
45	$^{12}C_1\,^{16}O_1\,^{17}O_1$, $^{13}C_1\,^{16}O_2$
46	$^{12}C_1\,^{17}O_2$, $^{12}C_1\,^{16}O_1\,^{18}O_1$, $^{13}C_1\,^{16}O_1\,^{17}O_1$
47	$^{12}C_1\,^{17}O_1\,^{18}O_1$, $^{13}C_1^{17}\,O_2$, $^{13}C_1\,^{16}O_1\,^{18}O_1$
48	$^{12}C_1\,^{18}O_2$, $^{13}C_1\,^{17}\,O_1\,^{18}O_1$
49	$^{13}C_1\,^{18}O_2$

Since the abundances are extremely small for most isotope combinations except for the first case representing the most abundant isotopes of carbon and oxygen, the insets in Fig 1·2 are drawn as log scale plots of the natural intensities as a function of *m/z* . Since the ionization process breaks up, or fragments, some of the CO_2 molecules, a fraction of the ions appear in the spectrum at *m/z* values less than 44. Cleavage of a carbon-oxygen bond in the molecular ion to produce ionized carbon monoxide or ionized atomic oxygen results in the fragment ions at *m/z* 28 and 16; loss of two neutral oxygen atoms results in an additional fragment at *m/z* 12 for carbon. The molecular ion is usually designated as M^+ or $CO_2{}^+$ and the fragment ions are designated as CO^+, O^+ and C^+. Since carbon is present in the form of two stable isotopes, ^{12}C and ^{13}C, there are two peaks at *m/z* values 12 and 13.003355 corresponding to these isotopes (bottom inset in Fig 1·2). In reality, modern mass spectrometers have much higher resolution than simply "nominal mass", so masses are usually calculated to an accuracy defined by the instrument. Since FTMS instruments yield ≈1 ppm mass accuracy, CO_2 is usually detected as (C=12.0000, O=15.9949) 43.9898 Da, and CO=27.9949 Da.

1.3.1 Resolving Power

One of the biggest advantages of an FTMS instrument is that it is able to provide much higher mass resolving power (100,000 typically) than most other types of mass spectrometers.(Marshall and Verdun, 1990; Shi et al., 1998) The resolving power of an instrument is defined as follows:

$$\text{Resolving Power} = \frac{\frac{m}{z} \text{ location of peak}}{\text{peak width at FWHM}} \tag{1.14}$$

where FWHM is Full Width at Half Maximum. The terms "resolving power" and "resolution" are often, and incorrectly, used interchangeably. "Resolution", measured in m/z units, usually refers to the mass spacing of FWHM, when two peaks of comparable height are just resolved at their half height. There are different ways to define "resolving power". The most commonly accepted definition is presented above. It is also defined in terms of FWHM when two neighboring peaks of comparable height are just resolvable at 10% of their height. Fig 1·3 illustrates the effect of varying resolving power on two consecutive Lorentzian peaks (Marshall and Verdun, 1990) of comparable height. The peaks are separated by 1.00 Dalton. As the resolving power drops, the peaks tend to exert greater influence on each other and get closer in proximity. If the resolving power further drops, the peaks will eventually merge together into a single peak with greater peak width. Fig 1·3 shows that, in order to separate isotopic peaks at their half height, a resolving power of at least 1.4×mass is required.

As equation 1.1 suggests, mass analysis in an FT-ICR mass spectrometer is based upon the measurement of ion cyclotron frequency. In more convenient units, the equation can be rewritten as:

$$\nu = 1.53561184 \times 10^7 \times \frac{zB}{m} \tag{1.15}$$

in which ν represents ion cyclotron frequency in Hz, z is ion charge in units of elementary charge, and m is the ion mass in Daltons. The electric-field-induced frequency shift has

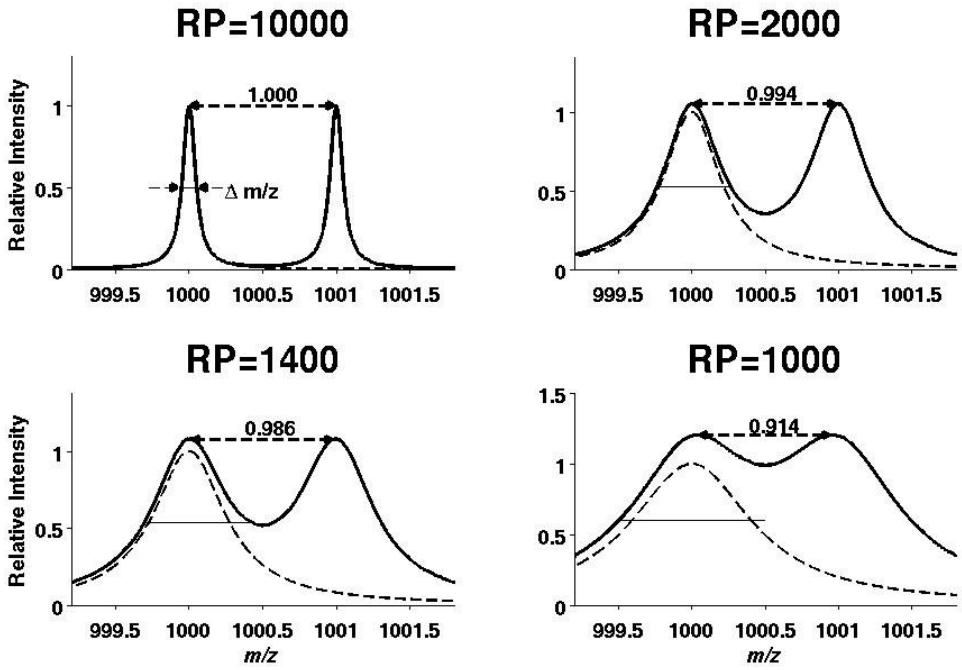


Figure 1·3: Effect of resolving power on the separation of two peaks of identical height, and space 1.000 m/z apart

been neglected for simplicity. Since equation 1.15 is linear, the following relations hold true:

$$\frac{m/z}{\Delta(m/z)} = \frac{m}{\Delta m} = \frac{\omega}{\Delta\omega} = \frac{\nu}{\Delta\nu} \tag{1.16}$$

where Δm, $\Delta\omega$, $\Delta\nu$ specify the FWHM in the respective domain. Under ideal vacuum conditions the frequency-domain magnitude mode spectral peak width at FWHM is given by (Comisarow and Marshall, 1976; Marshall et al., 1979; Marshall and Hendrickson, 2001):

$$\Delta\nu = \frac{1.2066}{T} \tag{1.17}$$

in which T is the total time for taken data acquisition. This equation assumes that (i) the time-domain signal, usually called the "transient", is not damped substantially during time T, and (ii) the transient is present for the entire time T. If either assumption is incorrect, $\Delta\nu$ increases, so equation 1.17 represents the lower bound for $\Delta\nu$. Equation 1.17 indicates that under zero pressure conditions FTICR width is independent of ion mass and charge.

Combining equation 1.15 and 1.17 leads to:

$$\frac{m}{\Delta m} = \frac{\nu}{\Delta \nu} = 0.8288\nu \times T \tag{1.18}$$

The above equation states that in the case of FTICR, mass resolving power is approximately equal to $\nu \times T$, which is the number of cyclotron orbits an ion makes during the data-acquisition process. This illustrates why FTICR is capable of such high resolving power. For example, in order to achieve a resolving power of 10^5 at m/z 1000 with a 7 Tesla magnet, data acquisition is required only for 1.12 seconds. Under these conditions, the ions make 112000 cyclotron orbits, which, with a typical orbit circumference of 10 centimeters represents a total flight distance of $\approx$ 12 kilometers. Substituting ν from equation 1.15 into equation 1.18, one obtained the following:

$$\frac{m}{\Delta m} = 1.2727 \times 10^7 \times \frac{zBT}{m} \tag{1.19}$$

Therefore, for a given acquisition time and magnetic field, mass resolving power is inversely proportional to m/z . So charge reduction is not desirable on a biomolecule in the case of electrospray ionization because a decrease in charge leads to an increase in m/z , and, hence, a decreased resolution which is proportional to z.(Marshall and Hendrickson, 2001) In general, mass accuracy is proportional to the resolving power, if all other factors remain the same. Hence, mass accuracy also drops with increase in m/z .

The mass resolution achieved by an instrument depends on both the type of analyzer and the experimental conditions. Higher resolving power is desirable in order to resolve ion species that are very close in their m/z values. This helps in the increased ability to assign the ion identities. For example, with an average resolving power of $\geq$80 000, across an m/z range of 200-1000 Daltons, Qian *et al.* were able to distinguish as many as 15 distinct chemical formulas within a 0.26 Da mass window, and more than 3000 chemically different elemental compositions were determined in a heavy crude oil sample.(Qian et al., 2001) Ultra-high resolving power has been used to determine the fine structure within an

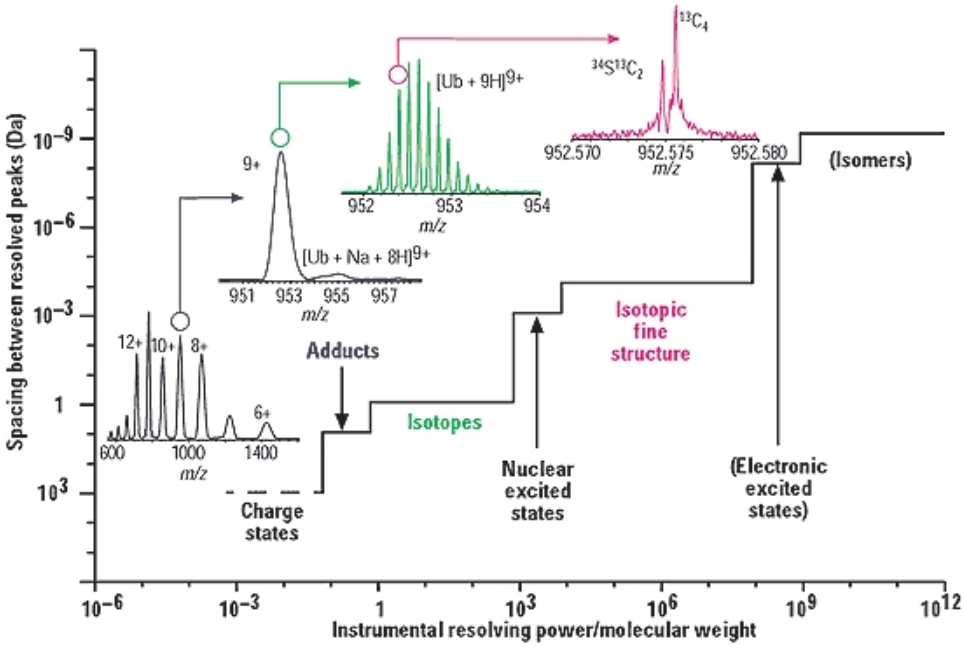


Figure 1·4: Mass spectral resolution versus the instrumental resolving power per unit of molecular weight, with experimental ESI FT-ICR mass spectra (insets) for the protein bovine ubiquitin with a monoisotopic mass of 8559.62 Da. As instrumental resolving power improves, ions of different charge (but the same mass) are resolved first, followed by ions differing in nominal closest-integer mass; ions of the same chemical formula but different isotopic composition; ions of the same nominal mass but different elemental composition; and ultimately, ions of different internal energy or isomers with different heats of formation. Parentheses indicate splittings that have not yet been observed experimentally. (Reprinted with permission from (Marshall et al., 2002). Copyright (2002) American Chemical Society)

isotopic distribution of high mass proteins, leading to results that confirm its molecular formula.(Shi et al., 1998; Spengler, 2004)

In a mass spectrum, m/z values are inherently quantized. This is because the charge, z, is quantized, and can take values only in integer multiples of the elementary charge e (charge on an electron). Mass values also progress in steps according to molecules, functional groups, elements, isotopes, and elemental compositions. As a result, with increasing resolving power, peaks suddenly start separating into finer and finer detail upon reaching certain thresholds or plateaus of resolving power. (Fig 1·4,(Marshall et al., 2002)) The

X-axis in the figure represents $\frac{1}{\Delta m_{50\%}}$ (where $\Delta m_{50\%}$ is the FWHM of mass spectral peak). It can be viewed as the ratio of mass resolving power to the ion mass m. Thus, it represents a mass-independent measure of resolving power. Fig 1·4 indicates that when the resolving power reaches the first plateau of about 0.8, different charge states of the same mass are resolved. The leftmost inset shows the charge states 6-15 for the protein ubiquitin (MW=∼8.5 kDa). The next spectral resolution plateau results in the separation of ions of the protein and protein with adducts like sodium (shown) with the same charge state. The subsequent step allows for the separation of ions with different nominal (nearest integer) isotopic masses. This permits the resolution of isotopes of the same molecule, having the same elemental but different isotopic composition, for example, substitution of ^{13}C for ^{12}C. If the resolving power rises farther to a level of 8×10^9, isotopic fine structure can be observed for ubiquitin, i.e., ions of the same nominal mass but different elemental composition and exact masses. The ultimate step is to resolve ions with different internal energies (because $E = mc^2$, where E=energy, m=mass, c=velocity of light; internal energy can be measured as mass), which has not been realized so far, except in some esoteric physics experiments.(Brown and Gabrielse, 1986; Gabrielse et al., 1999)

1.3.2 Mass Accuracy

Mass accuracy is a measure of how close the experimental mass value is to the theoretical value of a known analyte. It is defined in ppm units as follows:

$$\text{Mass Accuracy} = \frac{(M_{Exp} - M_{Theo}) \times 10^6}{M_{Theo}} \tag{1.20}$$

where M_{Exp} and M_{Theo} denote the experimentally observed and the theoretical values of the mass respectively. Mass measurement accuracy for a mass spectrometer depends primarily upon the type of mass analyzer being used and the calibration procedure/data processing methods used to calculate the mass values from the mass spectrum. FTMS instruments are capable of providing the best mass accuracy (low sub-ppm range with internal calibration) among the mass spectrometers currently available.

Accurate mass measurements are required for the characterization of both small molecules (like chemical synthesis products, metabolites, flavors, fragrances, etc.) (Zhang et al., 2005; Marshall et al., 1998; Marshall, 2000), and larger biomolecules.(Zubarev et al., 1995; Strittmatter et al., 2003; Kaiser et al., 2005; Henry et al., 1989; Zubarev et al., 1996; Spengler, 2004) Elemental composition of an unknown peptide can be determined on the basis of its accurate parent ion mass and a small number of fragment ion mass values.(Spengler, 2004; Zubarev et al., 1996) As the peptide mass increases, the mass accuracy requirements increase in order to uniquely assign the identity. For example, for a 200 Dalton peptide, an accuracy of 10 ppm is sufficient for unambiguous identification, while the identification of a peptide of molecular weight 1500 Daltons based upon the observed mass value alone requires an accuracy of greater than 0.01 ppm.(Spengler, 2004; He et al., 2004)

Increased resolving power is expected to lead to an increase in the mass accuracy in the spectrum, but this not always true (Marshall and Hendrickson, 2002) for an FTMS instrument. There are a number of factors such as space charge (Kaiser et al., 2005; Aizikov and O'Connor, 2006), peak coalescence (Huang et al., 1994; Mitchell and Smith, 1995) and mass calibration errors (Shi, 2000) that can limit the mass accuracy even under high-resolution conditions. Space charge effects arise because the electric fields associated with the ions influence each other, causing shifts in the cyclotron frequency in equation 1.1. Space charge essentially reduces the second term (the trapping field) by a time-varying quantity. This effect increases with both the increase in the number of ions present within the ICR cell and the amount of charge present on each ion. Peak coalescence is an extreme form of space charge effect. Under certain conditions, peaks that are closely spaced in frequency coalesce into one bigger peak. The mechanism proposed by Huang *et al.* (Huang et al., 1994; Mitchell and Smith, 1995) suggests that under high ion density conditions, ions with closely spaced cyclotron frequencies begin to move in synchronization with one another due to an interaction between the electric fields associated with each ion packet. This effect is particularly pronounced in cases where a low intensity peak is close (in frequency) to a

high intensity peak, and can be thought of as the fewer, low abundance ions being "swept up" into the electric field of the more numerous latter ions. This phenomenon depends upon the number of ions in the cell, the frequency spacing between the ions, the trapping voltage on the trap plates, the size and geometry of the ICR cell, and the radius of the ion's orbit.(Nikolaev et al., 1995) Mass calibration in an FT-ICR instrument refers to the process of converting the cyclotron frequency values obtained as the instrument output into mass-to-charge ratio values in the mass spectrum, generally by equations 1.8 and 1.11. The effect on calibration of the non-ideal behavior of the magnetic, electrostatic, and alternating electric fields created by space charge have been studied in great depth, resulting in the calibration equations yielding errors in the range of sub ppm.(Shi, 2000; Zhang et al., 2005; Amster, 1996; Kaiser et al., 2005)

1.3.3 Ion intensities

It is important to note that the ion intensities observed in a mass spectrum are a function of several parameters like the concentration of the given analyte, ionization efficiency, transmission efficiency across the ion optics, and detection efficiency of the instrument across the m/z range. For example, Fig 1·5 shows the mass spectrum of peptides from a protein called p21ras (Zhao et al., 2006), digested with trypsin, an enzyme known to cleave the given protein at specific positions of C-terminal to arginine and lysine residues.(Olsen et al., 2004) Since the whole protein was subjected to tryptic digestion, it is expected to result in an equal number of tryptic peptides from each region of the protein (assuming a full digestion). The spectrum (Fig 1·5) indicates varying ion intensities across different peptide positions within the protein. If the ion abundances depend only upon the analyte concentration, all the ion intensities of various peptides across the spectrum would be identical. Because of the various experimental factors mentioned above, the abundances of various peptides from the same protein can exhibit great variability. Clearly, peak intensities do not correlate well with the relative concentration of the components of a mixture, which has important implications in quantitative experiments.(Ong and Mann,

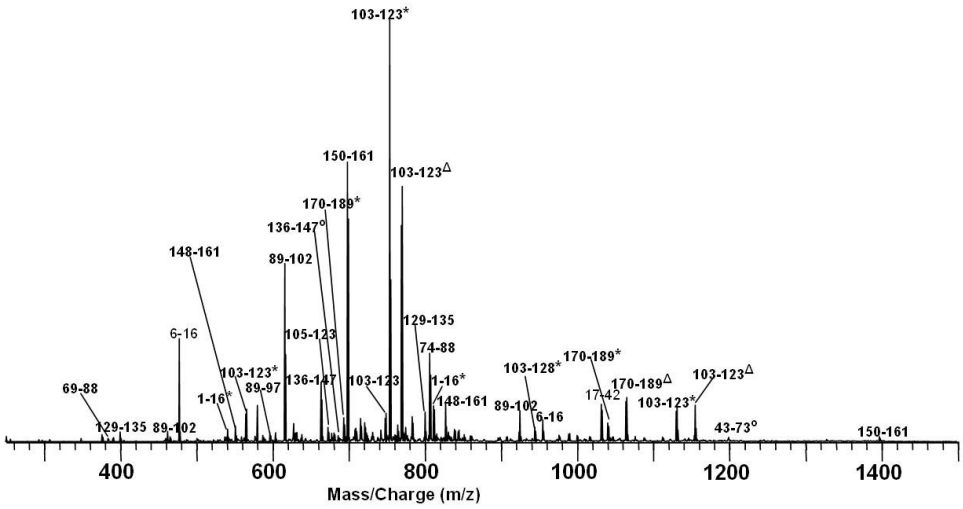


Figure 1·5: Mass Spectrum of p21ras protein digested with trypsin. The major peaks are labeled with peptide positions. (Reprinted with permission from (Zhao et al., 2006). Copyright (2006) American Chemical Society.)

2005)

1.4 Isotopic Distribution

Isotopes are atoms of the same element with the same atomic number (number of electrons or protons) but different atomic mass due to different number of neutrons. For example, carbon has two naturally occurring isotopes, ^{12}C and ^{13}C. Both isotopes are exactly the same except that ^{12}C has 6 neutrons, while ^{13}C has 7 neutrons, so they have an atomic mass of 12.00000 (by definition) and 13.00335 respectively. Some of the commonly occurring isotopes are shown in table 1.1. The successive isotopic elements are commonly referred to as A, A+1, and A+2 elements. For example, for oxygen, A denotes the ^{16}O isotope, A+1 refers to the ^{17}O isotope, and A+2 indicates the ^{18}O isotope.

1.4.1 Theoretical Isotopic Distribution (TID)

An Isotopic Distribution (ID) is an experimental measure of the probability distribution of the various isotopes in a molecule. The probability of any ion having a certain number of heavy isotopic atoms (e.g.,^{13}C, ^{2}H, ^{18}O etc.) can be calculated using the binomial

Element	Isotope	Accurate Mass	Natural Abundance
Carbon	12C	12.00000	98.9%
	13C	13.00335	1.1%
Hydrogen	1H	1.00782	99.985%
	2H	2.01410	0.015%
Nitrogen	14N	14.00307	99.63%
	15N	15.00010	0.37%
Oxygen	16O	15.99491	99.76%
	17O	16.99913	0.04%
	18O	17.99916	0.2%
Sulphur	32O	31.972070	95.02%
	33O	32.971456	0.75%
	34O	33.967866	4.21%

Table 1.1: Isotope table of natural elemental abundances (McLafferty and Turecek, 1993)

distribution (Yergey, 1983) for that particular elemental composition using the known natural abundance distribution for each element. For example, for carbon, the binomial distribution is

$$p_i = \binom{N_c}{i} {p_c}^i (1 - p_c)^{N_c - i} \tag{1.21}$$

where p_i represents the area of each individual peak in the ID (Fig 1·6a) (i.e., p_0 fraction of the total ions in the cell contain no ^{13}C, p_1 fraction of the total ions contain exactly one ^{13}C, etc.), N_c is the total number of C atoms in the molecule, p_c ($\cong 0.011$) is the natural abundance of ^{13}C isotope, and i is the total number of ^{13}C atoms in one molecule. Equation 1.21 accounts only for the stable isotopes of C, it may be extended by summation to account for other isotopic elements.(Yergey, 1983; Rockwood, 1995; Rockwood, 1996) Thus, the true theoretical isotopic distribution is a sum of these binomial distributions, one for each isotope. In general, if there is a compound with the composition $X_{N_x} Y_{N_y} Z_{N_z}$ (representing N_x, N_y, and N_z atoms of elements X, Y, and Z respectively), the theoretical isotopic distribution can be calculated by expanding the sum of binomial distributions using the polynomial method.(Yergey, 1983) Let P_{Xi}, P_{Yi}, and P_{Zi} represent the natural elemental abundances (Table 1.1) of the i^{th} isotopes of X, Y, and Z elements respectively. For an isotopic peak containing N_{x_j}, N_{y_j}, and N_{z_j} atoms of the j^{th} isotope of X, Y,

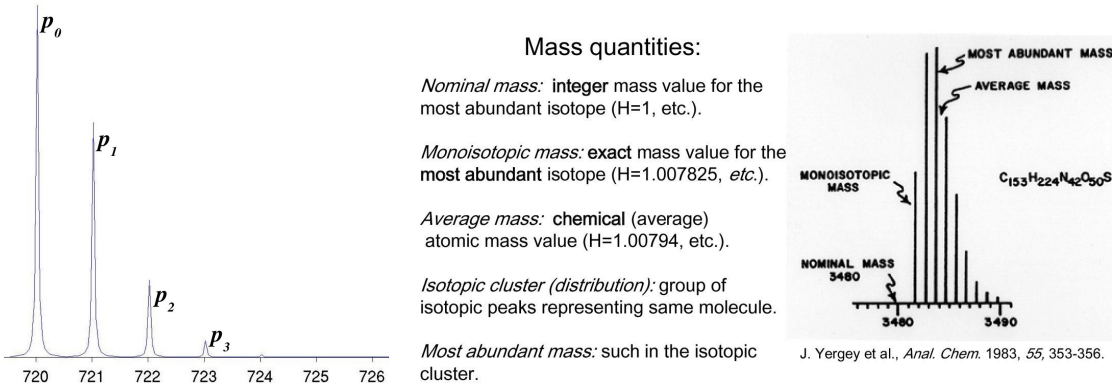
and Z elements respectively, the theoretical isotopic abundance can be represented by the following expression:

$$\frac{N_x!}{N_{x_1}!N_{x_2}!\ldots}P_{X1}^{N_{x_1}}P_{X2}^{N_{x_2}}\ldots\frac{N_y!}{N_{y_1}!N_{y_2}!\ldots}P_{Y1}^{N_{y_1}}P_{Y2}^{N_{y_2}}\ldots\frac{N_z!}{N_{z_1}!N_{z_2}!\ldots}P_{Z1}^{N_{z_1}}P_{Z2}^{N_{z_2}}\ldots \tag{1.22}$$

$$N_x=\sum_{i=1}^{I_X}N_{x_i},\ N_y=\sum_{i=1}^{I_Y}N_{y_i},\ N_z=\sum_{i=1}^{I_Z}N_{z_i} \tag{1.23}$$

where I_X, I_Y, and I_Z denote the total number of naturally occurring isotopes for each of the elements.

Direct implementation of the polynomial method has both computation and memory intensive requirements. This is because of the multiple factorial evaluations, multiplications, and divisions, which can lead to memory overflow problems, specially for larger numbers resulting from the calculations involving analysis of biomolecules like proteins. Moreover, there is a combinatorial explosion of terms with the increase in the number of atoms involved in the analysis, and a small protein like ubiquitin (average molecular weight=8565 Da) already has 1228 atoms ($C_{378}H_{627}O_{117}N_{105}S_1$). This problem was partly solved by optimizing certain factorial calculations, and rejecting certain terms below a certain threshold.(Yergey, 1983) These improvements helped solve some of the initial computational problems, but the scalability issues with the increasing complexity still remained. Also, pruning the low intensity terms led to computational errors, which is severe for elements which have a large number of natural isotopes, such as most metals. To alleviate these obstacles, Rockwood proposed a new approach using Fast Fourier Transform (Cooley and Tukey, 1965) methods to do the multiple convolutions required to generate the molecular isotopic distributions.(Rockwood, 1995; Rockwood, 1996) This approach reorganizes the polynomial multiplication problem as the convolution operation of individual isotopic abundances of each of the elements, and then maps the problem into the Fourier domain, converting the convolution operations in the mass domain into multiplications in the Fourier domain. This method produces fast and accurate results with minimal



(a) TID of C_{60} (b)Isotopic distribution represents the molecular mass

Figure 1·6: Isotopic Distribution

computational and memory overhead.

As the isotopic distribution illustrates, large molecules do not have a unique mass value due to the presence of multiple isotopes of the constituent elements. For example, Fig 1·6b shows different ways the molecular mass is defined. It is important, therefore, to specify what mass value out of this range of possibilities is being reported. One way is to report the average molecular mass, which is the average mass of all the isotopic species. However, there is a variability in the isotopic abundances of various elements across different organisms that limits the average mass accuracy to 10 ppm.(Beavis, 1993; Zubarev et al., 1996) The most significant and accurate value that can be reported is the monoisotopic mass, which is defined as the sum of the masses of the lowest-mass isotope for each of the constituent atoms of the molecule. This is because only the monoisotopic mass has a unique elemental composition, and remains unaltered by isotope ratio variability across various species. For example, in Fig 1·6b, the most abundant isotope is ≈2 Da heavier than the monoisotopic peak. This ≈2 Da can be either 2 ^{13}C substitutions for ^{12}C (2×(^{13}C-^{12}C)=2×(13.003355-12.0000)=2.0067) or one ^{18}O substitution for a ^{16}O (2.0042), etc.

Every isotopic peak (except for the monoisotopic peak) is located at approximately integer multiples of 1 Da higher in nominal mass than the monoisotopic mass; i.e., at unit

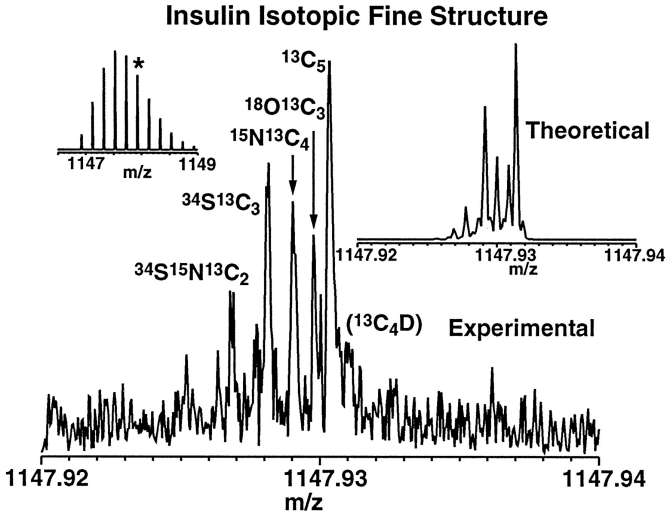


Figure 1·7: ESI FT-ICR mass spectrum (Upper Left), from a single time-domain data acquisition, of bovine insulin. Theoretical (Upper Right) and experimental (Lower) isotopic fine structure is shown for the isotopic peak (star $*$) $\sim$5 Da above the monoisotopic mass. Individual elemental compositions are clearly resolved at approximately correct relative abundances. Reproduced with permission from (Shi et al., 1998). Copyright 1998 National Academy of Sciences, U.S.A.

(nominal) mass resolution, the isotopic distribution consists of isotope peaks spaced $\sim$1 Da apart. Except for the monoisotopic peak, each other peak represents a sum of contributions from isotope combinations differing by a few mDa (e.g., two ^{13}C vs. two ^{15}N vs. one ^{13}C and one ^{15}N vs. ^{34}S, vs. ^{18}O, etc., at $\sim$2 Da higher in mass than the monoisotopic mass). At sufficiently high mass resolving power, each isotopic peak resolves into its isotopic fine structure, which means separation of the masses differing by a few milli-Daltons, and there are clearly several, low abundance contributions. Resolution of isotopic fine structures has been shown to confirm or determine the molecular formulas of certain molecules.(Shi et al., 1998; Stults, 1997) Fig 1·7 shows an example of the isotopic fine structure from the mass spectrum of bovine insulin.(Shi et al., 1998) The nominal mass of each of the components is $\sim$5 Daltons higher than the monoisotopic mass, but they are separated into five different components differing by a few milli-Daltons. Such fine structures are difficult to obtain for large molecules because of the tendency of closely spaced peaks to coalesce

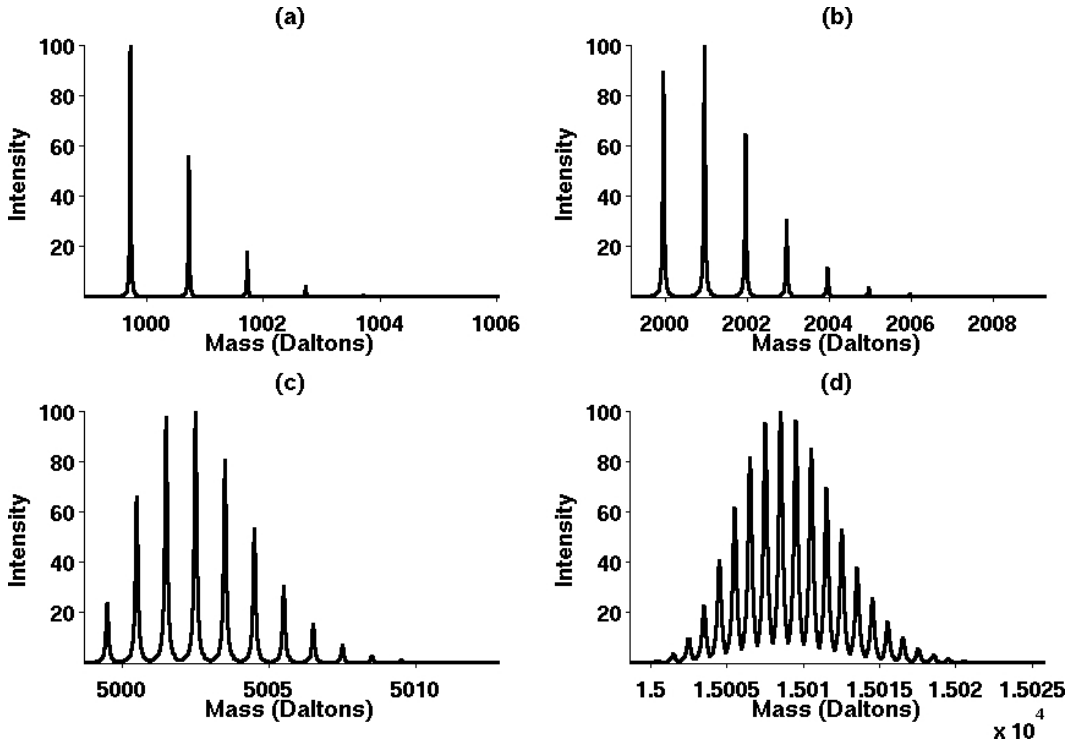


Figure 1·8: Variation in isotopic distributions with increasing molecular weight

into a single resonance.(Huang et al., 1994; Mitchell and Smith, 1995) The coalescence tendency decreases with the increasing magnetic field strength, so isotopic fine structures can be resolved at sufficiently high magnetic field strength (9.4 Tesla in this case).

The isotopic pattern of "average" proteins (Senko et al., 1995b) varies with the molecular mass. At low-mass values ($\leq$1100) the monoisotopic peak is the dominant peak in the spectrum as shown in Fig 1·8a. This is because the elemental abundance of lowest-mass isotopes is usually highest for most of the elements (Table 1.1). For low-mass molecules there is a high probability for all the atoms in the molecule to represent the smallest isotope, as can be calculated from equation 1.22. The relative intensity of the monoisotopic mass decreases with increasing molecular mass because of the increased probability of the presence of heavier isotopes. The monoisotopic peak is not visible experimentally for mass values greater than 5 kDa for most instruments because the tiny peak is buried in the back-

ground noise in the spectrum. When the molecular mass rises further, the isotopic pattern becomes broader, spanning a greater range of mass values, and the monoisotopic peak becomes vanishingly small, as seen in Fig 1·8d. This is due to the fact that combinations of higher isotopes become more probable with increasing mass (Equation 1.22). With bigger molecules the problem usually arises that the monoisotopic peak is not visible in the experimental isotopic distribution. In such cases the monoisotopic mass is usually estimated by comparing the experimental and theoretical isotopic distributions.(Horn et al., 2000; Kaur and O'Connor, 2006a) It is very important to correctly align the two distributions in order to obtain the correct value of monoisotopic mass. The value can be measured as accurately as the instrument allows, provided the monoisotopic peak has been correctly identified. If the distributions are misaligned, the mass value will be off by one or more Daltons, however many decimal places are present. Such situations illustrate the difference between accuracy and precision.

1.4.2 Experimental Isotopic Distribution (EID)

The discussion so far has been focused on the TID, where all the factors were deterministic, and the resulting distribution can be known exactly using the analysis discussed above. The TID determines the pattern of an isotopic distribution in theory. On the other hand, an EID is an experimental measure of the isotopic distribution, which involves certain random parameters, which vary from experiment to experiment. Other than the instrumentation-based parameters, one of the most important factors is the number of ions used to generate the EID. An EID can be interpreted as a result of a multinomial experiment (Papoulis, 1984) having multiple outcomes, with the number of trials being the number of charged molecules used to generate the distribution. Each of the outcomes is equivalent to an isotopic peak in the TID, and is associated with a certain probability, which can be determined using equations 1.21, 1.22, or the Mercury algorithm.(Rockwood, 1995; Rockwood, 1996) An EID can be reproduced *in silico* by means of generating a multinomial experiment, knowing the number of trials (equal to the number of ions gen-

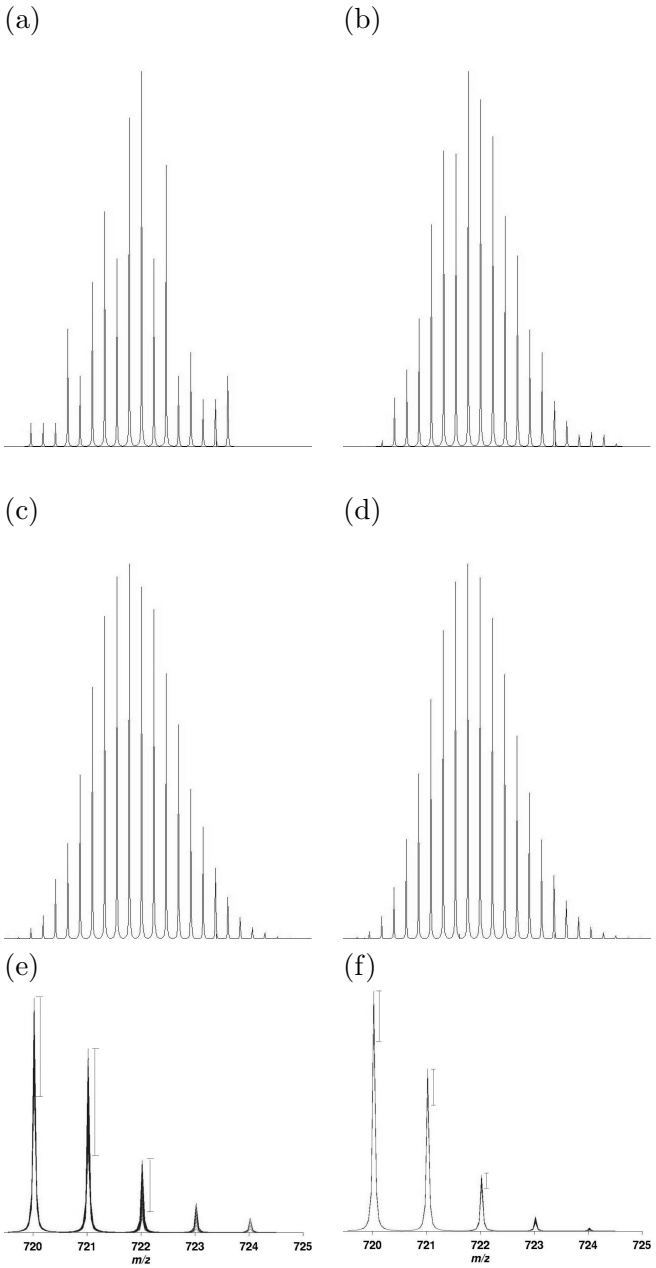


Figure 1·9: Experimental isotopic distributions approach theoretical isotopic distributions as the number of ions increase and variance decreases with increasing ions (a) 100 ions (b) 1000 ions (c) 10000 ions (d) infinite number of ions for myoglobin (e) over plotting 300 spectra of C_{60} with 100 ions (f) over plotting 300 spectra with 5000 ions (Reproduced with permission (Kaur and O'Connor, 2004). Copyright (2004) American Chemical Society)

erating the distribution) and likelihood of occurrence of each outcome. Fig 1·9a-d shows the simulated isotopic distributions of myoglobin, a single-chain protein with an average molecular weight of 16.7 kDa. It has been observed that the variance in the EID varies inversely with the number of ions used for its generation.(Senko et al., 1995b; Kaur and O'Connor, 2004) With only 100 ions in the cell (Fig 1·9a), the distribution can look quite "jagged", but as the number of ions increases (Fig 1·9b-c), the measured distribution approaches the theoretical (Fig 1·9d) distribution. Fig 1·9e-f are obtained by over plotting 300 Monte Carlo generated isotopic distributions of C_{60} with 100 ions (Fig 1·9e) and 5000 ions (Fig 1·9f). With only 100 ions in the cell, the scatter is higher than the case when there are 5000 ions. This topic shall be covered in greater depth in a later chapter.

For the analysis of unknown compounds, it is often useful to have a model based on their average molecular mass. This is particularly useful for analyzing the behavior of EIDs of a protein with a particular molecular weight. To this end, an approach was proposed that establishes a relationship between the average and monoisotopic mass of peptides and oligonucleotides.(Zubarev and Bondarenko, 1991) This method had a somewhat flawed assumption that all the amino acids (building blocks of proteins) have identical distributions across the proteins, which led to erroneous results. This limitation was later corrected, resulting in an improved model using the true distribution of amino acids from the Protein Identification Resource database.(Senko et al., 1995b) This approach led to a model amino acid, *averagine*, with molecular formula $C_{4.9384}H_{7.7583}N_{1.3577}O_{1.4773}S_{0.0417}$, and an average molecular mass of 111.1254 Da. This model helps to determine the "average" elemental composition of the molecule and isotopic distributions associated with that composition. To determine the model molecular formula, one calculates the total number of averagine units from its molecular weight, which is then multiplied by the number of atoms of each type in the averagine residue. Once the model molecular formula is established, this information can be used to determine a model TID and simulate the EID for a given number of ions, which can serve a variety of analytical purposes. The peaks of the EID are known to form

Lorentzian peak shapes (Marshall and Verdun, 1990) in an FT-ICR mass spectrometer, and an EID is the overall sum of individual Lorentzian peaks. It is very important to fit the EIDs to properly modeled distributions in order to do rigorous analysis. Fitting EIDs to improper peak shapes leads to mass assignment errors.

It has been observed that peaks in EIDs can be distorted by a variety of factors. For example, RF (radio frequency) interference peaks are frequently observed, which are attributed to the electronics used for controlling the instrument or inadequate shielding of detection electronics. Another cause of interference is chemical noise, which refers to the contaminants introduced during the sample preparation. These effects can be mitigated using rigorous sample preparation techniques. A common source of mass spectral signal perturbations is the interference from other isotopic distributions in the close vicinity of the peaks of interest. This happens frequently in the case of dense, complicated mass spectra resulting from the fragmentation of large biomolecules, so called "Top-down" protein analysis.(Reid and McLuckey, 2002; Kelleher et al., 1999) This is because multiple isotopic peaks are often produced at the same nominal m/z value. Automatic spectral interpretation becomes increasingly difficult under these circumstances, and special methods need to be developed to handle such complexity. The goal of this thesis is to develop such methods.

1.4.3 Non-natural Isotopic Distributions

The discussion so far has been based on the isotopic distributions resulting from the natural variation in the elemental isotopes. There are special situations when the natural distribution of isotopes is perturbed for the convenience of certain experimental studies. The methodologies for altering the isotope ratios include isotope labeling (Ong et al., 2002; de Godoy et al., 2006), radioactive labeling (Rice and Means, 1971), hydrogen-deuterium exchange (Engen and Smith, 2000; Wales and Engen, 2006; Jorgensen et al., 2005; Mandell et al., 1998), and isotopic depletion (Zubarev and Demirev, 1998; Marshall et al., 1997).

Isotope labeling is a method used for gauging the movement of a chemical through a system or a chemical reaction. The chemical is 'labeled' by including special isotopes in its composition i.e., by altering the natural isotopic abundances of certain elements. If these special isotopes are later discovered at some stage in the system, their source is attributed to the labeled element.

There are also ways other than mass spectrometry to detect the isotope labeling. Due to the difference in mass, molecules containing labeled isotopes have different vibrational modes. These can be detected by infrared spectroscopy. Another variation of this method is radioisotopic labeling, in which the specially introduced isotopes are radioactive and detected by their radioactivity.

Protein Nuclear Magnetic Resonance (NMR) spectroscopy (Wuthrich, 1990) uses NMR (Ramsey and Purcell, 1952; Bloch and Rabi, 1945; Jeener et al., 1979) spectroscopy to obtain information about the structure and dynamics of proteins. For protein NMR experiments, it is desirable to isotopically label the protein with ^{13}C or ^{15}N. This is because the predominant isotope ^{12}C has no net nuclear spin, which is the physical property nuclear magnetic resonance spectroscopy exploits, whereas the nuclear quadrupole moment of the predominant ^{14}N isotope prevents high resolution information from being obtained from this nitrogen isotope.

Hydrogen-deuterium exchange is a technique of studying proteins to gather information about their structure and dynamics. Some of the constituent hydrogen atoms in proteins exchange positions with the hydrogen atoms from the surrounding solvent molecules. If the solvent consists of the heavier isotope of hydrogen (deuterium), its heavier mass gets incorporated into the protein during the exchange. This increases the protein molecular weight, which can be detected in the mass spectrum. The exchange of hydrogens occurs at a specific rate at each position, which depends upon the protein structure and solvent accessibility. The measure of these exchange rates provides insights into the dynamics of

protein folding.

As discussed previously, it is difficult to observe the monoisotopic peak in large molecules. This is due to either low ion intensity of the peak or insufficient mass resolution. A potential solution to this problem can be achieved by enriching the proteins with the ^{12}C isotope, which is equivalent to depleting the relative content of the ^{13}C isotope .(Zubarev and Demirev, 1998; Charlebois et al., 2003; Marshall et al., 1997) This enrichment extends the isotopic distributions to lower mass values due to the greater percentage of the lighter isotope of carbon. This phenomenon causes the monoisotopic peak to become one of the most abundant species, resulting in more accurate characterization. This procedure is particularly beneficial for large molecules and low resolution mass analyzers.

Such perturbations in the isotopic abundances may cause the isotopic distributions to change drastically. These changes must be given due consideration for proper spectral analysis.

1.5 Charge State Determination

Spectrum interpretation represents one of the biggest bottlenecks in a mass spectrometry experiment. Manual analysis of such complex data is very tedious and time consuming. Hence, there is a great need for reliable sophisticated data analysis methods (Mann et al., 1989; Reinhold and Reinhold, 1992; Henry and McLafferty, 1990; Ferrige et al., 1991; Senko et al., 1995a; Senko et al., 1995b; Zhang and Marshall, 1998; Horn et al., 2000; Kaur and O'Connor, 2006a; Chen et al., 2006) in order to achieve high throughput results. One of the problems commonly encountered in automatic spectrum analysis is determining the charge state of ions representing the spectrum. For proteins in standard electrospray solutions, these charges usually arise by the adduction of available protons from the acidic solution to the protein. As both the solution and the protein itself partially shield the protons from each other, the number of charges can be quite large. Since all mass spectrometers measure mass-to-charge ratio (m/z), in order to measure the mass, the charge value must

be determined. A typical example is the charge state distribution of ubiquitin, a ∼8.5 kDa protein, whose charge states range from 6+ (m/z ∼1433) to 12+ (m/z ∼715).

1.5.1 Deconvolution

The first attempt for automatic charge state determination was (somewhat erroneously) called "deconvolution".(Mann et al., 1989; Reinhold and Reinhold, 1992; Zhang and Marshall, 1998; Henry and McLafferty, 1990) Since charge states can take only integer values, the idea behind "deconvolution" methods is to combine the isotopic peaks of the same mass with different charge states to determine the mass of the ion. For example, a molecule with mass=3000 Da will exhibit isotopic clusters roughly at m/z values of 1000, 1500 and 3000 corresponding to z=3, 2, and 1 respectively. By examining the locations of the isotopic clusters representing the same molecule with consecutive series of charge states, the mass value of the corresponding ion can be determined. These methods are particularly well suited for low resolution mass analyzers such as triple quadrupoles and ion traps, where charge states can be separated whereas isotopic peaks usually cannot.

The first "deconvolution" algorithm mathematically transforms a spectrum of several peaks for multiply charges ions into one peak corresponding to a singly charged ion.(Mann et al., 1989) It proposed the following function:

$$F(M^*) = \sum_{i=1}^{Z_{max}} f(\frac{M^*}{i} + m_a) \tag{1.24}$$

Here M^* denotes the mass value under consideration that takes on values from a certain range of values from, say, M_{min} to M_{max}, m_a indicates the mass of an adduct (such as a proton or sodium ion), i is an index that goes from 1 to the maximum possible charge state, Z_{max}. The function f represents the distribution function for peak heights in the measured spectrum. For example, if there is a peak of relative intensity 10 at m/z =800, then $f(800) = 10$. The function is evaluated for all possible values of masses M^*, with $M_{min} \leq M^* \leq M_{max}$. The result yields a transformed spectrum in which the peak with the

maximum height corresponds to the parent species with no charge. This method has been found to fail for complex mixtures of proteins, and also results in many false assignments of masses.

An alternative approach was later suggested, which is similar to the first one, except that it uses an entropy-based measure to detect the presence of a specific pattern, the envelope of charge state distributions corresponding to the trial parent ion mass, in the observed spectrum.(Reinhold and Reinhold, 1992) In this method the mass spectrum is interpreted as the output resulting from a large number of distinct random events of a Poisson distribution process, each event being the detection of an ion within the mass range of the experiment. The entropy method uses relative entropy as a measure of the difference between the two distributions as follows:

$$l_{\rho}^{2}(\nu) = -\sum_{i=1}^{N} \nu_i \log(\frac{\nu_i}{\rho_i}) \tag{1.25}$$

Here, vectors ν and ρ represent the normalized model charge state distribution for a given mass M and the observed spectrum respectively, N denotes the length of the shorter of the two vectors, and $l_{\rho}^{2}(\nu)$ denotes the "difference" between ν and ρ. For each parent mass M, a model distribution is constructed and the "difference" between this model distribution and the actual data is obtained. The plot of this difference as a function of parent mass is the deconvoluted spectrum. This method has been found to produce fewer false mass assignments and to be more discriminative than the previous method but results in substantial abundance distortion.

Another deconvolution method called Zscore is based on a charge scoring scheme that includes all the ion intensities above a user-defined threshold.(Zhang and Marshall, 1998) This technique uses various scoring schemes for charge assignments, which vary with the situation. For example, there are different scoring functions used when the spectrum is high resolution vs. low resolution cases, which are further classified into whether it's a low

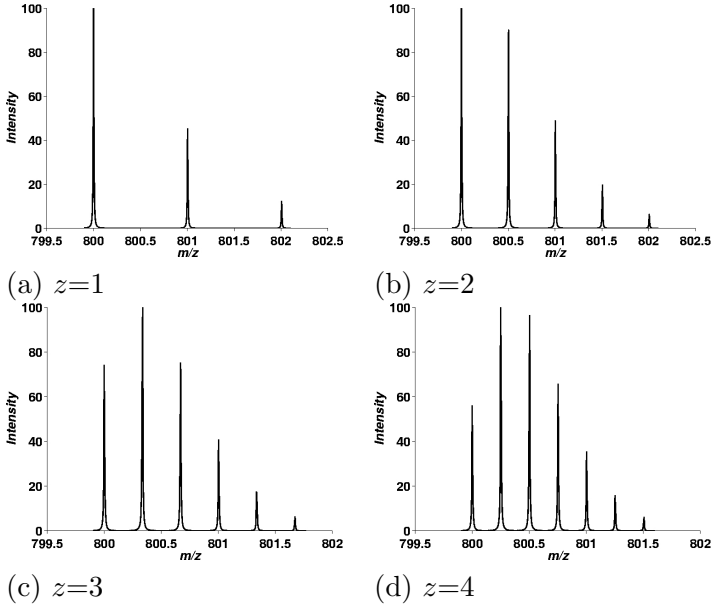


Figure 1·10: Spacing between isotopic peaks varies inversely with the charge state

charge state or a high charge state etc. This procedure was found to eliminate some of the artifacts associated the earlier methods.

The drawback of "deconvolution" methods is that they perform poorly if a given mass is represented by only one charge state, as is often the case in case of multistage mass spectrometry experiments. Due to these inherent problems in the "deconvolution" approach, techniques were developed for automated assignment of charge states from the isotopic spacings.(Senko et al., 1995a; Kaur and O'Connor, 2006a)

1.5.2 Charge state determination based on isotopic spacings

High resolution mass spectrometry such as that provided by the FTMS and Time-of-Flight instruments can generate resolving powers of greater than 10^4. When these high resolution instruments are used in conjunction with the modern ionization techniques that give rise to multiple charges, determination of a charge state, z, of an ion is simply a question of measuring the distance between neighboring isotopic peaks. This is because

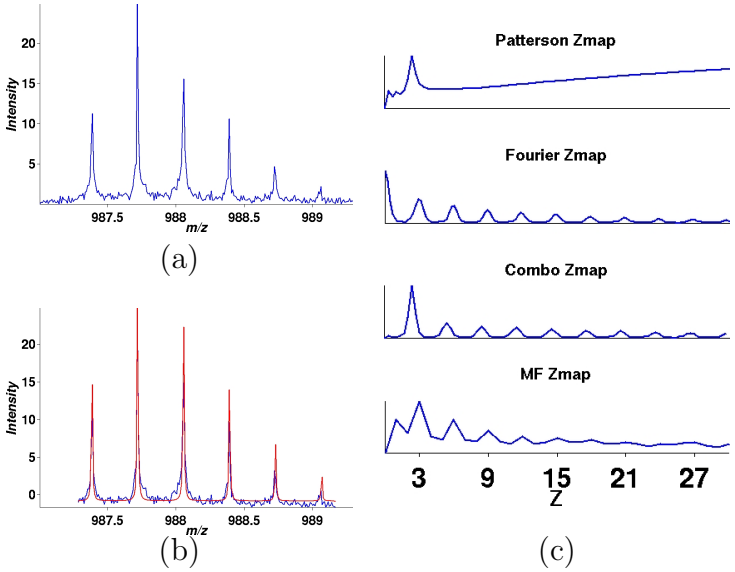


Figure 1·11: (a) EID from top-down spectrum of Bovine Carbonic Anhydrase (b) Shifted TID (red) (shift corresponding to maximum cross correlation coefficient (r=0.978 for Z=3)) plotted on the top of EID (blue) (c) Charge state maps using different methods. The Zmaps were imported from BUDA (Boston University Data Analysis)(O'Connor, 2004)

the m/z distance between adjacent isotopic ^{12}C and ^{13}C peaks is nearly $1.00235/z$.(Horn et al., 2000) Fig 1·10 shows isotopic distributions with the charge states (z) varying from 1 to 4, illustrating that the isotopic peak spacing decreases with the increasing charge state. Note that with the increasing charge states and m/z being the same, the molecular weight representing the IDs increases, and, hence, changing the isotopic pattern in Fig 1·10a-d. This "delta-mass" method to measure the distance between isotopic peaks works well when the signal intensities are high and there are no interfering peaks from other isotopic distributions. But difficulty arises with poorly resolved data or low signal/noise spectra. Under such conditions, it is difficult to pick the correct isotopic peaks. Senko *et al.* first explored alternative methods for automatic assignment of charge states(Senko et al., 1995a) based on the principle of isotopic spacing. These procedures include Patterson, Fourier transform, and a combination of the two, called Combo methods.

Patterson Method The Patterson routine (Senko et al., 1995a) uses a function similar to the Autocorrelation (Oppenheim et al., 2002) function, except that it uses a certain number of pre-determined lag values for calculating the autocorrelation values. Generally, there appears a maxima in the Patterson function plot corresponding to a lag of $1.00235/z$. Fig 1·11a shows an Experimental Isotopic Distribution (EID), taken from a tandem mass spectrum of carbonic Anhydrase, that corresponds to z=3. Fig 1·11c shows the "score" (called the Zmap) of each of the charge states as function of the charge state using various methods. The Patterson Zmap shows a strong peak at the correct value of z=3 (Zmap value for 0 lag was defined to be 0).

Fourier Transform Method In this case, the Fast Fourier Transform (FFT) of the EID is taken after zero-filling the input signal to the next power of 2 as shown in Fig 1·11c. Ignoring the large dc component at z=3, the largest peak is at z=3, with harmonics at z multiples.

Combo Method This method takes point-by-point multiplication of the above two methods to arrive at the Combo Zmap shown in Fig 1·11c. The peak in the Patterson and Fourier Zmaps corresponding to the true charge state is amplified by multiplication as shown in the Combo Zmap, with the highest peak being at z=3. This method suppresses the harmonic peaks observed in earlier methods.

Modified Fourier Transform Method Another variant to the above discussed Fourier transform approach first subjects the experimental isotopic distribution to Fourier transformation.(Tabb and Shah, 2006) The Fourier transform of each of the model isotopic distributions for all possible charge states is then computed. The EID's FFT is then compared to the FFT of the model isotopic distributions by normalized dot product. High scores indicate the most likely charge states.

Entropy-of-Fourier This unpublished method was developed by O'Connor at Boston University School of Medicine. It uses the Reinhold entropy distance function (Eq 1.25) to calculate the "entropy distance" between the FFT of the EID and the model distribution. The model distribution corresponding to the true charge state leads to the minimum distance between the two FFTs.

All of these methods were a big leap forward towards the automated interpretation of spectra. But they have certain limitations. These methods function very well when the the signal quality is high, but they all tend to break down when SNR is low or when the input signal represents multiple IDs overlapping with one another. So there was a need of a method that is capable of handling these limiting conditions. To this end, the Matched Filter approach for charge state determination was developed and is discussed in detail as a part of this dissertation. A brief overview is presented here.

Matched Filter (MF) This approach works by convolving the normalized EID with the TIDs from all possible charge states.(Kaur and O'Connor, 2006b) The TID representing the true charge state results in the highest cross-correlation value between the TID and the EID. Fig 1·11b illustrates detection of Z=3 using the MF approach. The coefficient values are plotted in the MF Zmap in Fig 1·11c as a function of z, with the highest value corresponding to z=3. Fig 1·11b shows TID (red) for z=3 plotted on the top of EID (blue), with the shift in TID corresponding to the maximum cross correlation coefficient value.

A detailed analysis of the comparison of the charge state determination methods has been carried out.(Kaur and O'Connor, 2006b) It concludes that the Patterson and Fourier Transform methods give poor performance under low charge states, both the Combo and Matched Filter performed much better under these conditions. The Patterson method was shown to break down (less discrimination) most rapidly as SNR decreased, followed by Fourier Transform, Combo, and Matched Filter method in that order. Furthermore, since the Matched Filter essentially matched a TID to the EID, the information about the

location of the EID is inherently present in the results, which is especially useful when overlapping IDs are present. This gives the Matched Filter an additional advantage over the previous methods. The Matched Filter (Chapter 4) gives improved performance at the expense of greater computational complexity due to the multiple convolution operations involved in the analysis.

1.6 Conclusions

Mass spectrometry is an indispensable tool for studying a wide variety of problems chemistry, biology, astronomy, and clinical applications, to name a few. FTICR instruments have been especially useful for solving mass spectrometry problems due to their high resolving power and mass accuracy. The output from a mass spectrometer, the mass spectrum, measures ion intensity as a function of mass-to-charge ratio. High spectral resolution is important to resolve species that are very close in their m/z values, a situation common in the analysis of larger molecules or complex mixtures. Mass resolving power progresses in a series of steps, with peaks separating into finer structures upon reaching certain thresholds, called MS plateaus. Higher resolving power typically leads to higher mass accuracy, provided other factors remain the same. If sufficient mass accuracy is available, elemental composition of an unknown molecule can be determined based upon the information of the accurate mass of parent ion and a certain number of its fragments.(Mann and Wilm, 1994; Mortz et al., 1996) Sophisticated mass calibration procedures, taking into account non-ideal behavior of electric and magnetic fields, play a key role in ensuring high mass accuracy.(Zubarev et al., 1995; Marshall et al., 2002; He et al., 2004; Zhang et al., 2005; Kaiser et al., 2005)

Ion intensities observed in a mass spectrum are a function of various parameters like concentration of an analyte, ionization and transmission efficiency across ion optics, detection efficiency in the mass analyzer, etc. Thus, ion intensities observed in a spectrum do not correspond only to its concentration in the original sample.

Isotopic distributions are generated as the result of observing heavier isotopes in larger molecules. Different methods have been proposed to generate the theoretical isotopic distributions for a given molecular formula.(Yergey, 1983; Rockwood, 1996) Experimental isotopic distributions can be modeled and generated *in silico* based on their theoretical counterpart for a given number of ions. As the number of ions generating an experimental isotopic distribution increases, the EID approaches the TID.

EID patterns experience distortions due to noise and other interfering sources. The noise sources could be chemical, arising from sample preparation artifacts, electronic, such as RF interference peaks from instrument electronics, etc. In a dense spectrum, overlapping isotopic distributions are commonly observed, making the data interpretation more challenging. There is a great need for sophisticated data analysis algorithms to interpret such complex data. Progress has been made in various facets of data analysis, including, but not limited to, spectral calibration, deconvolution, charge state determination, monoisotopic mass determination.

Non-natural isotopic distributions are of interest to handle specific analytical challenges. There are different methods for changing the isotope ratios include isotope labeling (Ong et al., 2002; de Godoy et al., 2006), radioactive labeling (Rice and Means, 1971), hydrogen-deuterium exchange (Engen and Smith, 2000; Wales and Engen, 2006; Jorgensen et al., 2005; Mandell et al., 1998), and isotopic depletion (Zubarev and Demirev, 1998). Isotope labeling is used for tracing the movement of a chemical through a system in vivo or a chemical reaction. Protein Nuclear Magnetic Resonance spectroscopy also uses isotopically labeled proteins for structural studies. Hydrogen-deuterium exchange is another methodology used for structural and dynamics characterization of proteins.

Outline of Dissertation The potential of biological mass spectrometry has been limited by the lack of fast and reliable methods for automated spectrum analysis methods. The employment of these methods in various applications can reveal very useful multi-

dimensional information. This dissertation aims at reducing the spectral interpretation bottlenecks experienced in mass spectrometry experiments. Towards this end, five specific problems are addressed:

1. **Estimation of the number of ions in the instrument cell by examining the isotopic distributions in the spectra**: Determination of the number of ions in a mass spectrometry experiment is required for the analysis of instrumental features such as ionization efficiency, collision-induced dissociation efficiency, ion transfer efficiency, ion trapping efficiency, preamplifier detection limit, calibration, the studying of space charge, etc. This project targets this problem by analyzing the variation in the intensities of the IDs, which depend upon the number of ions in the cell. The theory has been developed, based on the maximum likelihood estimation method, that estimates the number of ions in the ICR cell using non-random parameter estimation.(Kaur and O'Connor, 2004)

2. **Determination of high-precision isotope ratios from experimental isotopic distributions**: Isotope variability occurs in nature due to natural processes such as evaporation, photosynthesis, nitrogen fixation in forests, etc. This phenomenon provides insight into a diverse range of studies from authenticity control information for various foods (like fruit juices, butter, and cheese) to dietary patterns of ancient humans. These studies require the determination of elemental isotope ratios to a very high-precision. Isotope Ratio Mass Spectrometers (IRMS) are specialized mass spectrometers for such studies, which have some experimental limitations. This project develops the mathematical framework for estimating the elemental isotopic abundance from the experimental isotopic distributions.(Kaur and O'Connor, 2007)

3. **Developing and comparing charge state determination methods for high resolution mass spectra**: Charge state determination techniques are based upon accurately estimating the m/z difference between consecutive isotopic peaks. This becomes a particularly challenging problem under the conditions of low SNR and low

resolution data. A new method for charge state determination using a matched filter has been developed and compared with the established techniques under a variety of conditions.(Kaur and O'Connor, 2006b)

4. **Developing and integrating algorithms for isotopic cluster identification, resolving overlapping isotopic distributions, alignment of the experimental isotopic distribution with the theoretical isotopic distribution, and reducing the isotopically resolved mass spectrum to a monoisotopic mass list through MasSPIKE (Mass Spectrum Interpretation and Kernel Extraction)**: This work develops methods for reducing the dimensionality of a high resolution mass spectrum ($> 10^5$ data points) into a monoisotopic peak list (about 100 masses). The procedures are integrated into a suite of data reduction algorithms called MasSPIKE (Mass Spectrum Interpretation and Kernel Extraction). MasSPIKE models the noise across the spectrum, identifies isotopic clusters, determines the charge state of each identified isotopic cluster, resolves the overlapping isotopic distributions, aligns experimental and theoretical isotopic distributions for estimating the monoisotopic peak location, and finally, generates the monoisotopic mass list. The suite has been tested against a dense top-down spectrum of a protein, bovine carbonic anhydrase. Comparative performance analysis has been carried out with previously published work.(Kaur and O'Connor, 2004)

5. **Application of the above developed methods for characterization of Post Translational Modifications (PTMs) of biologically interesting proteins Hemoglobin and H-Ras**: Spectra obtained from real biological samples tend to have a higher degree of complexity. Thus, MasSPIKE has been used and tested against spectra from such samples. For example, Hemoglobin is the oxygen-carrying molecule of the red blood cells. This protein transports the oxygen molecules to the tissues of the body. It consists of four subunits; two are called alpha chains and the other two are called beta chains. Variants of Hemoglobin exist that result

in certain diseases. In this project, the goal is to confirm the results of a DNA sequence analysis of Hemoglobin from a normal person and persons with the variants of Hemoglobin.(Huang et al., 2005)

Ras proteins are regulatory guanosine triphosphate (GTP) binding proteins that serve as signal transducers, controlling cell growth and differentiation.(Berg et al., 2002) Impaired GTPase activity in a regulatory protein can lead to cancer.(Berg et al., 2002) Indeed, the gene for Ras is one of the genes most commonly mutated in human tumors.(Campbell et al., 1998; Shields et al., 2000) In a commonly studied mechanism of cancer, the Ras protein is trapped in the "on" position and continues to stimulate cell growth. Owing to its high biological significance, we carried out "top-down" and "bottom-up" (protein identification based on mass spectrometric analysis of peptides derived from breaking up of the protein into smaller fragments through an enzyme (like trypsin)) experiments to characterize any post translational modifications that affect the functionality of the protein.(Zhao et al., 2006)

Chapter 2

Use of Statistical Methods for Estimation of Total Number of Charges in a Mass Spectrometry Experiment

2.1 Introduction

The field of proteomics (Aebersold, 2003) attempts to catalog all proteins and their post translational modifications as a function of cell state. This field is possible largely because of the inventions of electrospray ionization (ESI),(Fenn et al., 1990; Fenn et al., 1989) and Matrix Assisted Laser Desorption/Ionization (MALDI)(Karas et al., 1985; Karas et al., 1987). However, cataloging the thousands of cell proteins, in even one cell, is an enormous undertaking. Two techniques that currently appear very promising in this regard are high throughput MALDI Fourier Transform Mass Spectrometry (FTMS)(Comisarow and Marshall, 1974; Marshall, 2000; Amster, 1996) and electrospray FTMS (Beu et al., 1993; Bakhtiar et al., 1993).

Proteomics experiments, particularly those using high throughput MALDI-FTMS instruments and ESI-FTMS, generate thousands of high resolution spectra a day. Such a high sample throughput generates an overload of data that necessitates the development of sophisticated data analysis methods. The work presented here proposes one such data analysis method to estimate the number of ions by examining the isotopic distributions in

the spectra.

Quantifying the number of trapped ions in a Ion Cyclotron Resonance (ICR) cell has been attempted previously in the literature.(Limbach et al., 1993) This task was first approached by comparing the experimentally observed signal voltage to that calculated for a single ion orbiting at the ICR orbital radius of the ion packet, assuming that all ions orbit in a tight, coherent packet. This method was able to estimate that, in a 3T cubic cell, under particular excitation and detection parameters, 177 ions generated a signal/noise level of 3:1. The disadvantage of this approach is that it needs instrument dependent parameters like capacitance of the detection circuit, gain of the amplifiers, peak-to-peak voltage corresponding to a measured FT/ICR mass spectral peak amplitude, ICR orbital radius, the distance between the centers of two detector electrodes, etc. Also, this method cannot be evaluated for its performance, i.e., how close is the estimate to the true number of ions.

An observation was made later that the error between an experimentally observed isotopic distribution (EID) and its theoretical isotopic distribution (TID) varies inversely with the number of contributing molecules.(Senko et al., 1995b) In this work, they used linear regression (LR) to determine the relationship between average error and number of ions. This method is preferred to the previous one as it requires no instrumental parameters for the estimate and can be tested by Monte-Carlo methods. It can be improved further by a more rigorous analysis using maximum likelihood parameter estimation.

In this project, we compare the TID of the molecule with the EID. The TID may be obtained knowing the elemental composition and isotopic abundances using the binomial distribution (Yergey, 1983). Statistical variance in the EID is used to estimate the number of ions in the cell.

The Maximum Likelihood (ML)(Poor, 1994) method is one of the most popular methods used for estimation of an unknown quantity from an observation. The ML estimator and the non-random parameter estimation (Poor, 1994) method are used to derive the mathematical relationship between the number of ions and the observed distribution. The performance of the method is shown to improve with increasing number of observations.

2.2 Theory

A mass spectrum is an experimental measure of the probability distribution of the various isotopes in a molecule. The probability of any ion having a certain number of heavy isotopic atoms (e.g.,^{13}C, ^{2}H, ^{18}O etc) can be calculated using the binomial distribution (Senko, 1998; Yergey, 1983) for that particular elemental composition using the known natural abundance distribution for each element. For example, for carbon, the binomial distribution is

$$p_i = \binom{N_c}{i} {p_c}^i (1 - p_c)^{N_c - i} \tag{2.1}$$

where p_i represents the area of each individual peak in the isotopic distribution, N_c is the total number of C atoms in the molecule, p_c ($\cong 0.011$) is the natural abundance of ^{13}C isotope, and i is the total number of ^{13}C atoms in one molecule. Equation 2.1 accounts only for the isotopes of C, it may be extended to account for other isotopic elements by convolution.(Yergey, 1983; Senko, 1998; Rockwood, 1996).

In order to estimate the total number of ions in a given EID, one can compare the EID to the TID and estimate the number of ions based upon the variance of the EID since variance depends on the number of ions. Ideally, the EID should look exactly like the TID, but this happens only in the limit of infinite or very large number (order of 10,000-15,000) of ions (or in the case of a statistical anomaly). Peak areas are calculated from the EID (denoted by y_i). The EID can be interpreted as the result of a multinomial experiment (with number of trials equal to the number of ions) having multiple outcomes, each with the probability p_i. The EID is represented by a vector Y, where y_i corresponds to p_i in

the TID. Since the components of Y are binomial random variables, the covariance matrix of Y is given by:

$$\Sigma_N = \frac{1}{N}\begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \dots & -p_1p_n \\ -p_2p_1 & p_2(1-p_2) & \dots & -p_2p_n \\ . & . & \dots & . \\ . & . & \dots & . \\ -p_np_1 & -p_np_2 & \dots & p_n(1-p_n) \end{pmatrix} \tag{2.2}$$

where N is the number of ions in the experiment, and n is the total number of isotopic peaks of interest. Defining

$$\Sigma = N\Sigma_N, \tag{2.3}$$

Σ is independent of N. If we assume that Y is a Gaussian random vector with mean P and covariance matrix Σ_N, where P is composed of p_i's as defined in equation 2.1 and has length n. The probability of observing Y, given that the number of ions in the cell is N, is given by (Poor, 1994):

$$P(Y|N) = \frac{e^{-0.5(Y-P)'{\Sigma_N}^{-1}(Y-P)}}{\sqrt{(2\pi)^n det(\Sigma_N)}} \tag{2.4}$$

where $(Y-P)'$ denotes the transpose of the vector $(Y-P)$, and ${\Sigma_N}^{-1}$ denotes the inverse of matrix Σ_N. (Note: The assumption that Y is a Gaussian random vector is a reasonable assumption because the abundance of Y will tend to fluctuate evenly around the mean value. Further, this assumption is supported by the consistent Monte-Carlo results obtained below. However, there is one problem with this assumption: In mass spectra, Y cannot have negative values, while a Gaussian random vector as defined by Y would have a (very small) probability of having negative values. In reality, this problem does not greatly affect the estimate because low abundance peaks go nearly unused in the estimate.)

Non-Random Parameter Estimation In this case, N is a constant but unknown variable, and the goal is to determine the best estimate of N based on the variance in the EID. One

can estimate N using the Maximum Likelihood (ML) Estimator (Poor, 1994):

$$N(Y) = \arg \max_{N} P(Y|N) \tag{2.5}$$

where $N(Y)$ denotes the number of ions as a function of an observed distribution Y and $\arg \max_N P(Y|N)$ is the value of N that is most likely to produce an isotopic distribution Y. Therefore, the task is to maximize the probability of observing Y as a function of the number of ions in the cell. Since $P(Y|N)$ is modeled as a Gaussian function, maximizing the probability implies (please see Appendix (section 2.6)):

$$\frac{\partial(ln(P(Y|N)))}{\partial N} = 0 \tag{2.6}$$

Solving eqn 2.6 using eqn 2.4, the estimate becomes:

$$N(Y) = \frac{n}{(Y-P)'\Sigma^{-1}(Y-P)} \tag{2.7}$$

This estimate uses one distribution, and the estimator can be improved by observing multiple isotopic distributions under a similar set of conditions. This improved estimation requires calculation of the following:

$$N(Y_1, Y_2, Y_3, \ldots, Y_M) = \arg \max_{N} P(Y_1, Y_2, Y_3, \ldots, Y_M|N) \tag{2.8}$$

where M is the total number of observations, and $Y_1, Y_2, Y_3, \ldots Y_M$ are the EID vectors for each of the observations. Since the observations are independent,

$$P(Y_1, Y_2, Y_3, \ldots, Y_M|N) = \prod_{i=1}^{M} P(Y_i|N) \tag{2.9}$$

Again, using,

$$\frac{\partial(ln(P(Y_1, Y_2, Y_3, \ldots, Y_M|N)))}{\partial N} = 0 \tag{2.10}$$

one obtains,

$$N(Y_1, Y_2, Y_3, \ldots, Y_M) = \frac{n \cdot M}{\sum_{i=1}^{M} (Y_i - P)'\Sigma^{-1}(Y_i - P)} \tag{2.11}$$

Equations 2.7 and 2.11 give the relationship between the number of ions and observed distribution(s), and, thus, can be used to directly calculate the number of ions from any given isotopically resolved spectrum.

2.3 Methods

62 Myoglobin spectra were collected using electrospray ionization on a previously described Fourier-Transform mass spectrometer (O'Connor et al., 2006) and 45 C_{60} spectra were generated using laser desorption on another, previously described MALDI Fourier-Transform mass spectrometer (O'Connor and Costello, 2001). Both of these mass spectrometers used a 7 T actively shielded magnet, the Ionspec "in-cell" preamplifier, and detection of 1 Megaword (12 bit) at 500 kilo samples-per-second. A Pentium 4 PC and BUDA (Boston University Data Analysis),(O'Connor, 2004) in the Windows XP environment were used to view and analyze the spectra. MATLAB 6.5 was used to carry out the comparison of the EID and TID and to calculate the observed number of ions. Theoretical distributions were generated using a previously published method (Yergey, 1983), as implemented by Isopro 3.0 (Senko, 1998). The ML method was tested extensively using Monte-Carlo methods in which a given number of ions were filled into an isotopic distribution based on the exact probabilities for each isotopic combination generated by the theoretical distribution (Yergey, 1983; Senko, 1998). This allows for the performance evaluation of the ML method.

In practice, not all the peaks are observed in the EID. Very low intensity peaks (peaks with intensity less than the noise baseline) cannot be observed. So while comparing the experimental distribution to the theoretical distribution, only the significant peaks (peaks with intensity greater than the noise baseline) of the experimental distribution are used. These peaks are aligned with the corresponding theoretical distribution peaks (Kaur and O'Connor, 2006a). Now the ML estimator is used to compare these "significant" experimental distribution peaks to the corresponding theoretical distribution peaks.

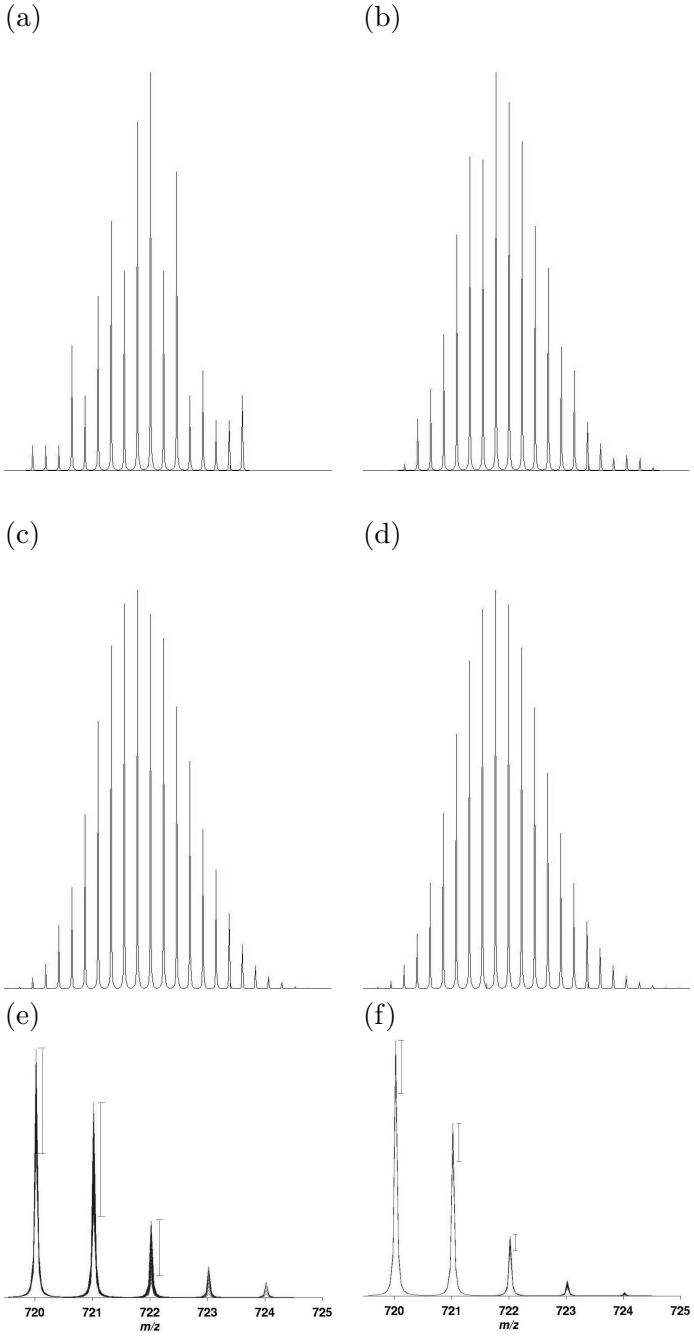


Figure 2·1: Experimental distribution approaches theoretical distribution as the number of ions increase and variance decreases with increasing ions (a) 100 ions (b) 1000 ions (c) 10,000 ions (d) infinite number of ions (e) overplotting 300 spectra with 100 ions (f) overplotting 300 spectra with 5000 ions

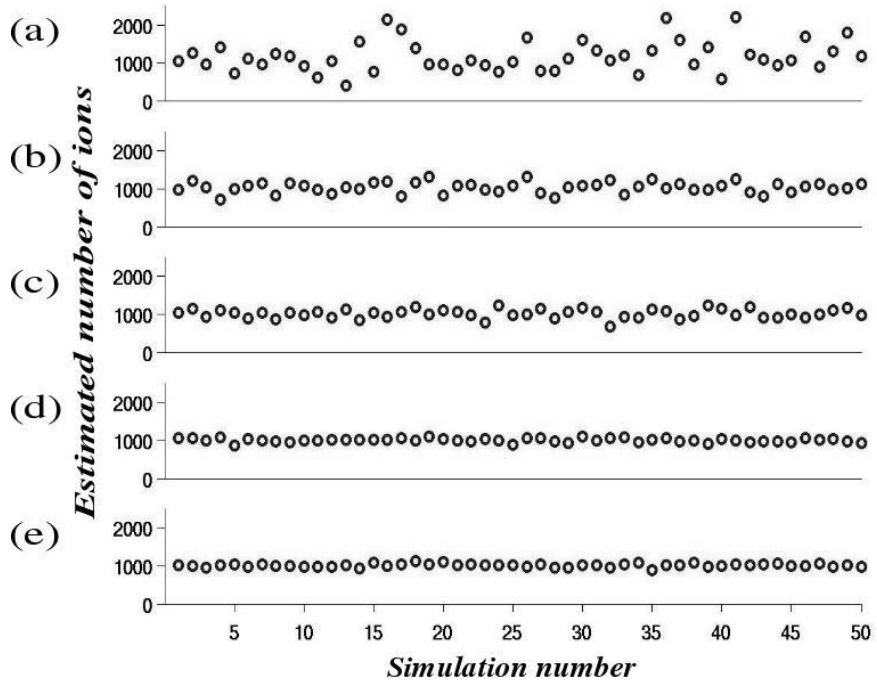


Figure 2·2: True number of myoglobin ions=1000 (a) 1 observation per simulation (b) 10 observations per simulation (c) 25 observations per simulation (d) 50 observations per simulation (e) 100 observations per simulation

2.4 Results and Discussion

Figure 2·1 shows the Monte-Carlo-generated isotopic distributions of myoglobin. With only 100 ions in the cell (Fig 2·1a), the distribution can look quite "jagged" but as the number of ions increases (Fig 2·1b-c), the measured distribution approaches the theoretical (Fig 2·1d) distribution. Figure 2·1(e-f) are obtained by overplotting 300 Monte-Carlo-generated isotopic distributions of C_{60} with 100 ions (Fig 2·1e) and 5000 ions (Fig 2·1f). With only 100 ions in the cell, the scatter is higher than the case when there are 5000 ions. This scatter is used to estimate the number of ions.

The estimate is expected to improve with increasing number of observations. Figure 2·2 shows this effect by comparing the results using 1, 10, 25, 50 and 100 observations (each observation is Monte-Carlo-generated isotopic distribution of myoglobin) per simulation in Figures 2·2a-e respectively. By using information from 100 observations, the method

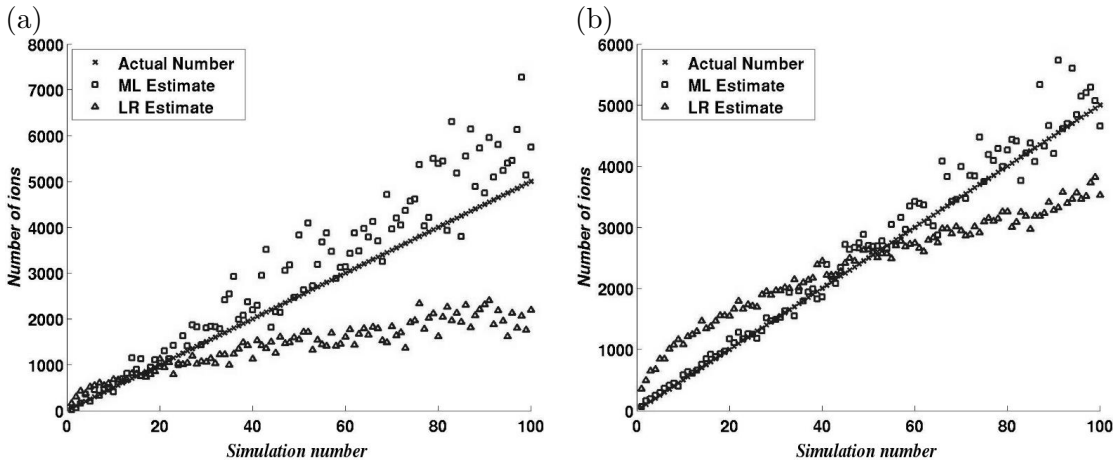


Figure 2·3: Performance evaluation using 15 observations per simulation (a) C_{60} ions estimate (b) myoglobin ions estimate

estimates within 5% of the actual value.

The ML estimator works best in the range of low number of ions. As Figs 2·3(a-b) show, the ML estimate is particularly accurate when the actual number of ions is small (50-1500), but error increases when the number of ions rises (though at 5000 ions, the results are still generally within 20%). This effect is intuitive because it uses variance in the distribution to calculate N (the actual number of ions) and variance changes as $\frac{1}{N}$ (Eq 2.2). Thus, the ML estimator works better in the range of low value of N, though better results can be obtained for higher values of N by using a greater number of observations (M). In general, determination of ionization efficiency (Keller and Li, 2001), preamplifier detection limit, etc require working with a low number of ions so that the improved accuracy of the ML estimator at low number of ions is advantageous. Fig 2·3 shows that the estimator also works better for higher mass ions like myoglobin (Fig 2·3b) than for low mass ions like C_{60}. This improvement is also intuitive since for low mass ions, the EID approaches the TID more rapidly than for the case of high mass ions. For example, for C_{60}, there are only about 5 significant peaks in the isotopic distribution while for myoglobin, there are about 26 significant peaks in the isotopic distribution. So it takes a greater number of ions for an EID of myoglobin to approach the TID than for C_{60}. Also, for myoglobin 16^+, for example,

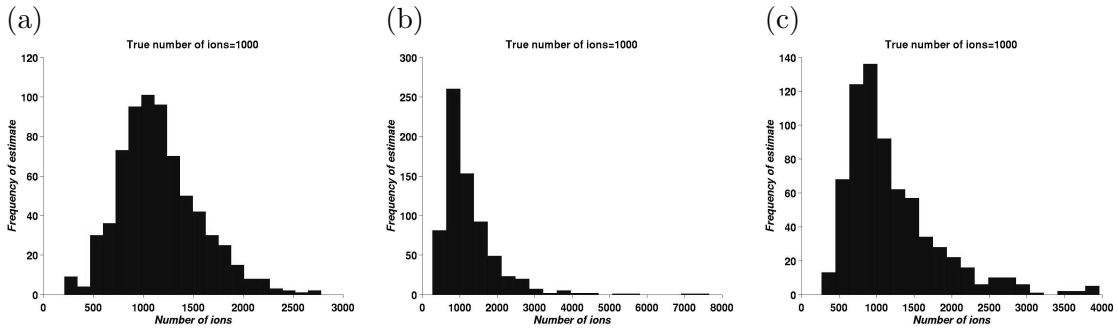


Figure 2·4: Effect of observing a limited number of peaks in the distribution (a) Histogram using all the peaks of myoglobin distribution (b) Histogram using peaks 7-16 of the myoglobin distribution (c) Histogram after eliminating outliers using peaks 7-16 of the myoglobin distribution

one ion accounts for 16 charges, but one elemental composition, thus amplifying the signal by 16 fold, allowing observation of smaller number of ions. Once the EID approaches the TID, it is hard for the estimator to tell whether it came from, say, 20,000 ions or 25,000 ions, since the variance does not change much, and the distributions are similar. This effect is also clear from Figure 2·1 in which the distribution containing 10,000 ions is virtually superimposable on the TID. Figure 2·3 also demonstrates performance evaluation (i.e., how close is the estimate to the true number of ions) of the ML and LR methods. The LR method works well for a limited range of number of ions (100-1000 in case of C_{60}, 2000-3000 in case of myoglobin) and digresses from the true value in other cases, whereas the ML method works best in the range of low number of ions. In some cases, the ML method shows an improvement by a factor of 2.

There are two known systematic biases to the ML estimate under these conditions. The first bias is the result of white noise adding to the randomness of the EID. Due to increased randomness, variance also increases, which causes the estimate to drop, so that for low signal/noise (SNR) data, this method will systematically underestimate the total number of ions. To partially correct for this, one subtracts the mean of the surrounding noise (noise in a neighboring m/z window having no isotopic distribution in it) from the EID before estimation. The second bias is caused by a statistical anomaly. There is

a certain probability that the EID will happen to be very similar to the TID even with limited number of ions (the "lucky guess" problem). This happens about 8-10% of the time when using a single observation per estimate. This will cause the estimator to overestimate highly. The "lucky guess" problem causes the "tail" of the estimate distribution to extend to the higher side (Fig 2·4a) because of overestimation to higher numbers. This problem can be reduced by using multiple obervations per estimate. For example, Fig 2·4a shows the estimate distribution when the true number of ions is 1000. This histogram of estimates was generated by applying equation 2.7 to 700 Monte-Carlo simulations. Clearly, the distribution has some outliers which occasionally overestimate the number by a factor of 2.5 or more. The rest of the distribution is centered at the histogram bin 1046±64, correctly predicting the expected 1000 ions.

These biases are affected by the fact that, in practice, only a limited number of peaks of the distribution are observed rather than the whole distribution. The observed peaks are aligned to their appropriate positions in the theoretical distribution using MasSPIKE (Kaur and O'Connor, 2006a). The observation Y now has a length 10-15 instead of 26 since we observe only about 10-15 peaks in the center of the distribution. Now Y is normalized to the sum of corresponding peaks in the theoretical distribution for comparison. Observation of less than the total number of peaks introduces a subtle effect, extending the tail of the estimate histogram (Fig 2·4b). Fig 2·4a shows the estimate histogram obtained when all the peaks (26 in case of myoglobin) in the distribution are observed. Each of the figures 2·4(a-c) is obtained using the Monte-Carlo method using 1000 ions with 1 observation per estimate, and a total of 700 simulations is used to obtain the histogram. Fig 2·4a has median estimate at 1046±64, the maximum value of the estimate extends to about 2750. So the estimate based on the maximum histogram frequency in this case will be 1046±64. Fig 2·4b shows the case when only peaks 7-16 are used to arrive at the estimate, as is the case with experimental data where noise can eliminate peaks on the edges of the distribution. The median estimate in this case is 822±184, but the maximum value of

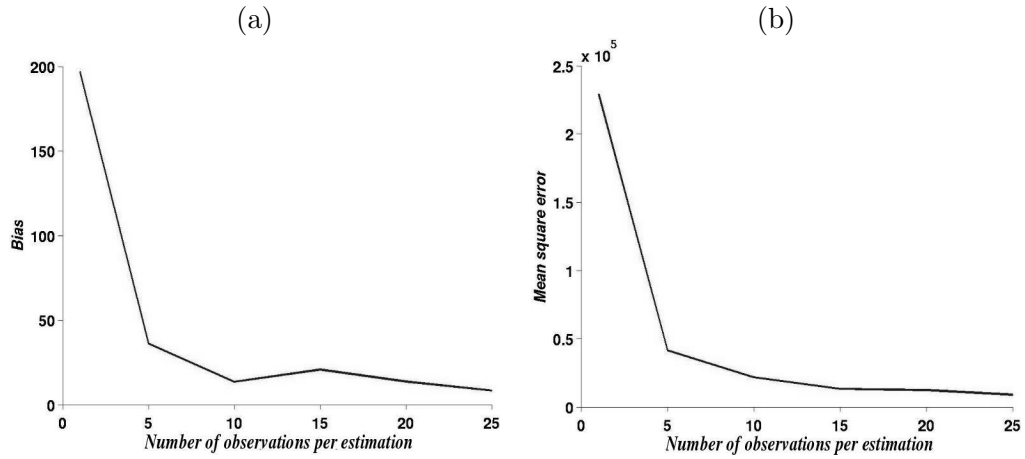


Figure 2·5: Effect of multiple observations per calculation on (a) Bias (b) Mean square error(MSE)

estimate extends to about 7600. This is again because of the "lucky guess" problem, but the effect is more pronounced in the case of a limited number of peaks, because now only 10 peaks (instead of 26) need to approach the theoretical distribution in order to generate the "lucky guess" problem. This causes the error in the estimate to increase ($\sim$ 22% (Fig 2·4b) for peaks 7-16 *vs.*$\sim$ 6% (Fig 2·4a) in case of all 26 peaks), which affects the accuracy of the estimate. This problem can be largely eliminated by rejecting the outliers (e.g. truncating the histogram once the frequency values reach less than 5% of the maximum frequency) and regenerating the histogram. Fig 2·4c is formed by truncating the histogram in Fig 2·4b when the frequency of the histogram falls below 5% of the maximum frequency and regenerating the new histogram by dividing the remaining data into 20 evenly spaced bins. Fig 2·4c calculates the median value that corresponds to 915$\pm$92, so the accuracy of the estimate now improves from $\sim$ 22% to $\sim$ 10%.

The net bias and mean square error (MSE) of an estimator can be calculated, which depend heavily upon the number of observations per estimate. In this case, bias and MSE are defined as follows: $bias = Mean[N_T - N_{Est}]$, $MSE = Mean[N_T - N_{Est}]^2$, where N_T=True number of ions, N_{Est}=ML estimate of N. Fig 2·5a shows that the bias falls rapidly as the number of observations per estimate increases. Similarly, MSE follows the

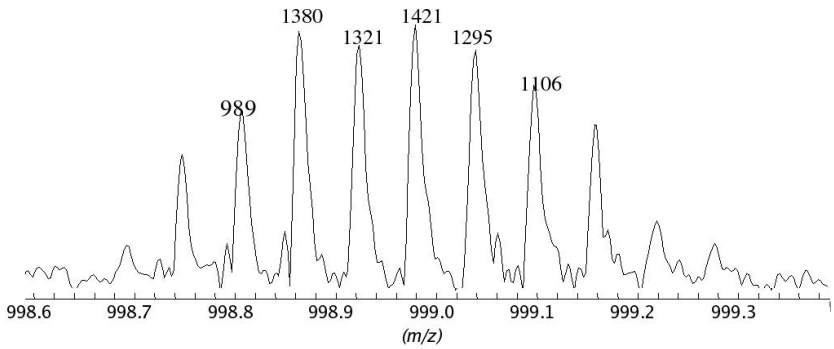


Figure 2·6: Estimation of total number of charges in each peak using equation 2.7, total number of ions in the whole 17^+ isotopic distribution=705

same pattern of abrupt drop with increase in the number of observations per estimate as shown in Fig 2·5. High values of the bias and MSE with a small number of observations per estimate are largely due to the presence of outliers due to the "lucky guess" problem. Thus, the bias and MSE can be improved either by increasing the number of observations per estimate or by increasing the total number of estimates and then using the median of the estimate histogram after ignoring the outliers as discussed above.

Figure 2·6 shows the estimated number of charges for each of the peaks in an electrospray mass spectrum of myoglobin for the 17^+ charge state. The central (labeled) 6 peaks comprise about 63% of the total charges in the distribution. This is an EID with typical broadband noise characteristics for this FTMS instrument. In order to determine the number of charges, the number of ions (N_{Est}) was estimated by observing the EID (Y) of myoglobin. Knowing the value of P (theoretical distribution for myoglobin), Σ (defined by equations 2.2 and 2.3) and n (number of expected abundant isotopes), N_{Est} is determined by equation 2.7. It was found that the whole distribution represents 705 (17^+) ions. The charge in each of the peaks was calculated by dividing the total charge ($705\times17\times0.63$) among the observed peaks proportional to their height.

One of the primary reasons for this study was to develop a test for the amplifier sensitivity on signal/charge basis. Fig 2·7 shows the preamplifier detection limit (number

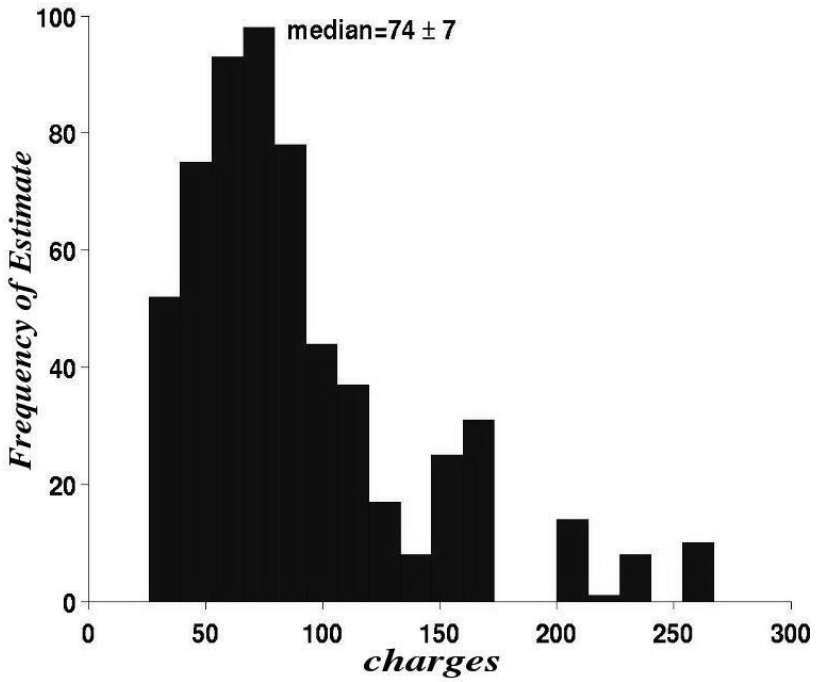


Figure 2·7: The number of charges needed for a SNR of 3 as estimated from a series of myoglobin spectra

of charges needed for SNR ratio of 3) values obtained from about 600 estimates, with 1 observation per estimate. The detection limit is defined as: $detection\ limit = \frac{N_{Est} \times Z \times 3}{SNR}$, where N_{Est}=Ion number estimate from the EID (calculated as described in above paragraph), Z=Charge state of the corresponding EID (calculated using BUDA (O'Connor, 2004; Senko et al., 1995a)) and SNR=Signal to noise ratio at the base peak of the EID. This distribution looks similar to that in Fig 2·4c, indicating that the detection limit value is 74±7. The distribution can be improved further by using more estimates, or more observations per estimate to give a better accuracy. For a comparison to the previous work (Limbach et al., 1993), the value can be adjusted to compensate for difference in magnetic field strength (3 T *vs.* 7 T), cell radius (0.5r *vs.* 0.8r). Thus, the corrected value of detection limit is 272 charges, which compares reasonably with the 177 charges determined previously.(Limbach et al., 1993)

2.5 Conclusions

A method has been developed for calculating the total number of trapped ions in an MS experiment from the statistical variation of EID in the mass spectra. The Maximum Likelihood estimator together with the Non-Random Parameter estimation method has been used to derive the mathematical relationship between the number of ions and the experimentally measured variance in the EID. This theory has been tested using Monte-Carlo simulations to compare the true ion number with the estimated number. The results improve rapidly with the increase in the number of observations since the estimator gets more information. This method is independent of the type of instrument used. In terms of performance evaluation, the method can show a factor of 2 improvement over a previously developed method,(Senko et al., 1995b) depending on the number of observations used for the calculation, and it can be used for any kind of mass spectra provided it can be resolved isotopically.

2.6 Appendix

Since N is a discrete variable, the behavior of the expression $ln(P(Y|N))$ is investigated to check the approximation of the derivative with respect to N in expression 2.6. Simplification of $ln(P(Y|N))$ yields the following form:

$$f(N) = ln(P(Y|N)) = -\frac{1}{2}[N\alpha - n\,ln(N) + \beta] \tag{2.12}$$

where α, β, and γ are constants with respect to N. Consider the following function:

$$g(x) = -\frac{1}{2}[x\alpha + -n\,ln(x) + \beta] \tag{2.13}$$

The function $g(x)$ is the same as $f(N)$, except that it is a function of a continuous variable x, as opposed to $f(N)$, which is a function of a discrete variable N. Hence, $f(N)$ can be viewed as a sampling of $g(x)$ at positive, discrete values of x. Maximizing $g(x)$ yields the

following:

$$g'(x) = -\frac{1}{2}(\alpha - \frac{n}{x}) = 0 \tag{2.14}$$

which yields the only one solution:

$$x = \frac{n}{\alpha} \tag{2.15}$$

This value of x can be either a maximum or a minimum. Second derivative $g''(x)$ yields the following:

$$g''(x) = -\frac{n}{2x^2} \tag{2.16}$$

Evaluating $g''(x)$ at the value of x in equation 2.15 results in the following:

$$g''(x) = -\frac{\alpha^2}{2\,n} \tag{2.17}$$

which is always negative, since α^2 and n (number of isotopes) are always positive. Hence, the extremum of $g(x)$ is indeed a maximum. Since $g(x)$ has exactly one maximum, its sampled function $f(N)$ also has only one maximum, which lies in the vicinity of the maximum for $g(x)$. Hence, for the purposes of analysis, the maximum for $f(N)$ can be determined by locating the maximum for $g(x)$, which has been used to derive the ion number estimate in section 2.2.

An estimator for a non-random parameter is called efficient if it meets the Cramer-Rao bound for the associated error covariance.(Poor, 1994) The existence of unbiased efficient estimator in this case was tested by evaluating the following expression:

$$exp = N + \frac{1}{I_Y(N)}\frac{\partial(ln(P(Y_1, Y_2, Y_3, \ldots, Y_M|N)))}{\partial N} = f(N) \tag{2.18}$$

The simplified form of the above expression is a function in terms on N, the number of ions, indicating that unbiased, efficient estimator does not exist for this problem.

Chapter 3

Quantitative Determination of Isotope Ratios from Experimental Isotopic Distributions

3.1 Introduction

Elemental isotopic composition variation in biological products due to natural processes is known and provides important information for a diverse variety of studies. The isotopic signatures of biomolecules depend upon the geographical parameters like latitude, distance from the sea, altitude, and seasonal effects.(Rozanski et al., 1992) For example, Oxygen isotope ratio ($^{18}O/^{16}O$) indicates the source along with the authenticity control information for products like fruit juices, wine, milk, butter, cheese etc.(Rossmann et al., 2000; Bricout and Koziet, 1973; Dunbar, 1982) Isotopic analysis of bone and dental remains of past species are used to determine the dietary patterns of ancient humans.(Stott and Evershed, 1996) Natural isotope ratio variability also provides information on studies like gender-specific physiology,(Dawson and Ehleringer, 1993) nitrification and nitrate turnover rates in forests,(Stark and Hart, 1997) history of earth's climate,(Schoell et al., 1994) determining origins of a given sample,(Rossmann et al., 2000) and diets of contemporary animals.(Hobson et al., 1997) In biomedical sciences, isotope ratios based tracer methods are employed for protein turnover studies, fat metabolism,(Brenna, 1997) and breath tests for clinical testing purposes,(Hoekstra et al., 1996) not to mention the drug testing of athletes.

Conventionally, high-precision carbon isotope ratio measurements are expressed in terms of the delta notation.(McKinney et al., 1950; Hayes, 1983) The Delta notation is defined as the relative difference in parts per thousand between the sample isotope ratio and an isotope ratio of an international standard. For carbon, the accepted international standard is PeeDee Belemnite (PDB), a belemnite from the Cretaceous Pee Dee formation, South Carolina USA. It is expressed as following:

$$\delta^{13}C_{PDB} = \frac{R_{SPL} - R_{PDB}}{R_{PDB}} \times 1000 = (\frac{R_{SPL}}{R_{PDB}} - 1) \times 1000 \tag{3.1}$$

where

$$R_x = \frac{[^{13}C_x]}{[^{12}C_x]} \tag{3.2}$$

where $[^{12}C_x]$ and $[^{13}C_x]$ are the abundances of the respective isotopes in the sample or PDB, R_{PDB}=0.01123720.0000090.(Craig, 1957) $\delta^{13}C$ values are expressed as "per mil" or ‰. Using this definition, C3 and C4 plants have a $\delta^{13}C$ value of -26.5‰ and -12.5‰,(Calvin and Benson, 1948; Smith and Epstein, 1971) with the corresponding abundances for ^{13}C being 1.082% and 1.097% of total carbon respectively, indicating the subtle differences in the isotopic signatures. Examples of C3 plants include trees, shrubs, flowering plants and temperate zone grasses. C4 plants include maize, sugar cane, and tropical grasses.

Since the isotopic variations in various categories are extremely subtle, measurements of δ values require very high-precision determination of the isotope ratios of a particular element involved. Isotope Ratio Mass Spectrometers (IRMS)(Brenna et al., 1997) are widely employed tools for such high-precision analysis. These instruments typically require complex compounds to be reduced to simpler molecules before measurement. For example, organic compounds are combusted to CO_2, H_2O and N_2 which are then separated and detected individually. Current IRMS instruments thus accept the sample analyte in the form of only a limited number of gases, which, in turn, must represent the isotopic characteristics

of the original sample. This represents a major limitation for the range of molecules that can take advantage of isotope ratio mass spectrometry. This work aims at overcoming the limitations inherent to IRMS by estimating the elemental isotopic abundance directly from the Experimental Isotopic Distribution (EID). This implies that any mass spectrometer capable of providing isotopic resolution can be used for isotope ratio analysis. Another substantial advantage of this method is that the samples do not need to be reduced to simpler molecules, extending the utility of isotope ratio mass spectrometry to molecules that are not amenable to the reduction process. This will also provide increased sensitivity since the sample preparation process will involve fewer steps, and hence, fewer losses.

Computational approaches have been proposed previously for stable isotope enrichment experiments.(Demirev and Fenselau, 2002; MacCoss et al., 2005) These allow for measuring the isotope enrichment ratios by comparing the predicted, enriched isotopic distributions with the measured isotopic distributions. Progress has also been made for the measurement of the distribution of isotopomers in a labeled compound by means of linear algebra.(Jennings and Matthews, 2005) However, the purpose of the current approach is to determine the natural isotopic abundances of elements from their experimental isotopic distributions. This work should not be confused with the approaches intended for monitoring the progress of isotope enrichment ratios used in isotope labeling experiments.

An isotopic distribution (ID)(Yergey, 1983; Rockwood, 1996) is a direct measure of the isotopic abundances of its constituent elements, convolved by the number of atoms of each element present in the molecule under consideration. If the isotopic abundance contributions from all but one element are known, an EID can be analyzed mathematically to determine the isotopic abundance contribution of the unknown element. Known isotopic abundance contributions can be deconvolved from the EID, allowing for solution for the unknown element.

3.2 Theory

Let the elemental composition of a given molecule be $C_{N_c}H_{N_h}N_{N_n}O_{N_o}S_{N_s}$. Let

$$\left.\begin{aligned} P_C &= [1-p_c \;\; p_c] \\ P_H &= [1-p_h \;\; p_h] \\ P_N &= [1-p_n \;\; p_n] \end{aligned}\right\} \quad \text{A+1 Elements} \tag{3.3}$$

$$\left.\begin{aligned} P_O &= [1-(p_{o_1}+p_{o_2}) \;\; p_{o_1} \;\; p_{o_2}] \\ P_S &= [1-(p_{s_1}+p_{s_2}) \;\; p_{s_1} \;\; p_{s_2}] \end{aligned}\right\} \quad \text{A+2 Elements} \tag{3.4}$$

be the isotopic abundances of C, H, N, O and S respectively, where $P_X(i)$ represents the i^{th} isotope of element X. Let $T[n]$ represent the Theoretical Isotopic Distribution (TID).(Yergey, 1983; Rockwood, 1996) Thus,

$$T[n] = P_C[n] \overset{N_c}{\circledast} P_C[n] * P_H[n] \overset{N_h}{\circledast} P_H[n] * P_O[n] \overset{N_o}{\circledast} P_O[n] * P_N[n] \overset{N_n}{\circledast} P_N[n] * P_S[n] \overset{N_s}{\circledast} P_S[n] \tag{3.5}$$

where $*$ denotes convolution operator (Proakis and Manolakis, 2003) and $\overset{N}{\circledast}$ denotes multiple convolutions defined as follows:

$z[n] = x[n] * y[n] = \sum_{j=1}^{+\infty} x[j]\; y[n+1-j]$

$x[n] \overset{N}{\circledast} x[n] = x[n] * x[n] * x[n] ... N$ times

Assuming that the isotopic abundances of all the elements except any one element, say carbon, are known, the goal is to find the unknown isotopic abundance of ^{13}C, denoted by p_c. In order to extract the p_c values from the convolved isotopic distributions, it is convenient to reorder T[n]. The theoretical isotopic distribution T[n] can be represented as the following:

$$T[n] = T_1[n] * T_2[n] \tag{3.6}$$

where $T_1[n]$ represents the isotopic abundance contribution from carbon as follows:

$$P_C[n] \overset{N_c}{\circledast} P_C[n] \tag{3.7}$$

and $T_2[n]$ contains known isotopic abundance information about the other elements defined as the following:

$$T_2[n] = P_H[n] \overset{N_h}{\circledast} P_H[n] * P_O[n] \overset{N_o}{\circledast} P_O[n] * P_N[n] \overset{N_n}{\circledast} P_N[n] * P_S[n] \overset{N_s}{\circledast} P_S[n] \tag{3.8}$$

$T_2[n]$ is assumed to be completely known since its constituents depend upon the known isotopic abundance values of all the elements other than carbon. Since $P_x[n]$ is non-zero only for positive values of n, where x denotes a particular element, the same holds true for $T_1[n]$ and $T_2[n]$. Thus,

$$T[n] = \sum_{j=1}^{N_c+1} T_1[j]\, T_2[n-j+1] \tag{3.9}$$

since the length of $T_1[n]$ is $N_c + 1$. Let the coefficients of expansion of $T_1[n]$ be α_1, α_2, $\cdots$, α_{N_c+1}. Thus, $T[n]$ can be expanded as follows:

$$T[1] = T_1[1]\, T_2[1] = \alpha_1\, T_2[1](1-p_c)^{N_c} \tag{3.10}$$

$$\begin{aligned} T[2] =& T_1[1]\, T_2[2] + T_1[2]\, T_2[1] \\ =& \alpha_1\, T_2[2](1-p_c)^{N_c} + \alpha_2\, T_2[1]\, (1-p_c)^{N_c-1}\, p_c \\ & \vdots \end{aligned} \tag{3.11}$$

$$T[k] = T_1[1]\, T_2[k] + T_1[2]\, T_2[k-1] + \cdots + T_1[k] T_2[1] \tag{3.12}$$

$$\begin{aligned} =& \alpha_1\, T_2[k](1-p_c)^{N_c} + \alpha_2\, T_2[k-1]\, (1-p_c)^{N_c-1}\, p_c + \cdots \\ & + \alpha_k\, T_2[1]\, (1-p_c)^{N_c-k+1}\, {p_c}^{k-1} \end{aligned} \tag{3.13}$$

When the number of ions representing an experimental isotopic distribution is extremely large, the experimental distribution approaches the theoretical isotopic distribution.(Kaur and O'Connor, 2004) Hence, the terms on the left of each of the above equations can be substituted by the corresponding values from the EID. Thereafter, each of the above polynomial equations can be solved for p_c, with each equation representing one of the isotopes observed in the experimental isotopic distribution.

Using Chebyshev's inequality, it is possible to determine the number of ions required to generate an isotopic distribution that is sufficient to determine the $\delta^{13}C$ value within a given accuracy. This is done as follows. Assume that the true value of $\delta^{13}C$ is δ_{true}, and the estimate is desired within, say, $\pm\Delta$. Since the $\delta^{13}C$ value depends upon [^{13}C] (denoted by p_c) and [^{12}C] (1- p_c) as seen in equations 3.1 and 3.2, the corresponding values of p_c, say $p_{ctrue} \pm \Delta_{p_c}$, for $\delta_{true} \pm \Delta$ can be determined from equations 3.1 and 3.2. Now the isotopic abundances in an EID depend upon the value of p_c. Let the isotopic abundances for an ideal EID (which is the same as the TID) be represented by the vector E_{true} when δ_{true} is the true value of $\delta^{13}C$. Due to the limited number of ions, there is a variation in the EID.(Kaur and O'Connor, 2004) Let this variation be represented by $\mathrm{E}_{true}\pm$e, and be such that the variation is the same as that is caused by the variation in p_c in the range of $p_{ctrue} \pm \Delta_{p_c}$. The goal here is to determine the number of ions so that the vector of observed isotopic abundances, E, lie within $E_{true} \pm e$. Each of the isotopic peaks in the EID is treated individually for the analysis purposes. According to Chebyshev inequality,

$$P(|E(i) - E_{true}(i)| \geq e(i)) \leq \frac{\sigma_i^2}{e(i)^2} \tag{3.14}$$

where i denotes ith isotopic peak, and σ_i^2 indicates the abundance variance of the ith isotopic peak, and e(i) denotes the magnitude of deviation of the observed ith isotopic peak from its true value. It means that the probability that the observed isotopic abundances in EID, E, differs from the true values, E_{true} by more than a certain value e, is bounded by the expression of the right hand side of the above equation, which depends upon the variance in the given isotope and its deviation from the true value. The variance σ_i^2 is given by (Kaur and O'Connor, 2004):

$$\sigma_i^2 = \frac{E_{true}(i) \times (1 - E_{true}(i))}{N} \tag{3.15}$$

where N denotes the number of ions used to generate the EID. Now, assume that it is desired that the estimated delta value deviates from the true value by no more than Δ

with a probability $\hat{P}$. Then number of ions, N, can be solved for using the above equations and $\hat{P}$ as follows:

$$N(i) = \frac{E_{true}(i) \times (1 - E_{true}(i))}{\hat{P} \times e(i)^2} \tag{3.16}$$

Note that this number can be calculated for each isotopic peak in the EID. This number will ensure that if the EID is generated using a certain number of ions satisfying the above equation, $\delta^{13}C$ will be bounded within the limits as described above.

3.3 Methods

Experimental spectra of chlorophyll-b (Molecular formula = $C_{55}H_{70}MgN_4O_6$) purchased from Sigma (St. Louis, MO) were obtained with 1μM concentration using ethanol and 1% formic acid. The spectra were obtained on a previously described custom hybrid electrospray FTICR instrument.(O'Connor et al., 2006; Jebanathirajah et al., 2005) Bovine ubiquitin sample was prepared with 1μM concentration in a 50:50 mixture of water and methanol, and the spectra were obtained on the same instrument. Matlab 7.1 (Natick, MA) was used for the data analysis.

3.4 Results and Discussion

In order to test this theory (per equations 3.10, 3.11, and 3.12), performance analysis was first carried out on modeled EIDs generated using Monte-Carlo simulations in silico. EIDs were generated by varying the number of ions to study the effect of error in determining $\delta^{13}C$ as a function of the number of ions used to generate the EID. The number of ions was varied from 500 to 22000, in the increments of 500, and the true value of delta used to generate the isotopic distribution was -25.5‰ (typical value for C3 plants), corresponding to p_c = 1.0832%. A TID was generated for an "average" protein (Senko et al., 1995b) with a molecular weight of 9000 Da. Each estimate was generated using an average of 10 simulations, with number of ions being the same in each of the 10 simulations. Fig 3·1 illustrates that with the increase in the number of ions used to generate the EID, the error drops substantially. This is intuitive since the higher the number of ions, the closer the

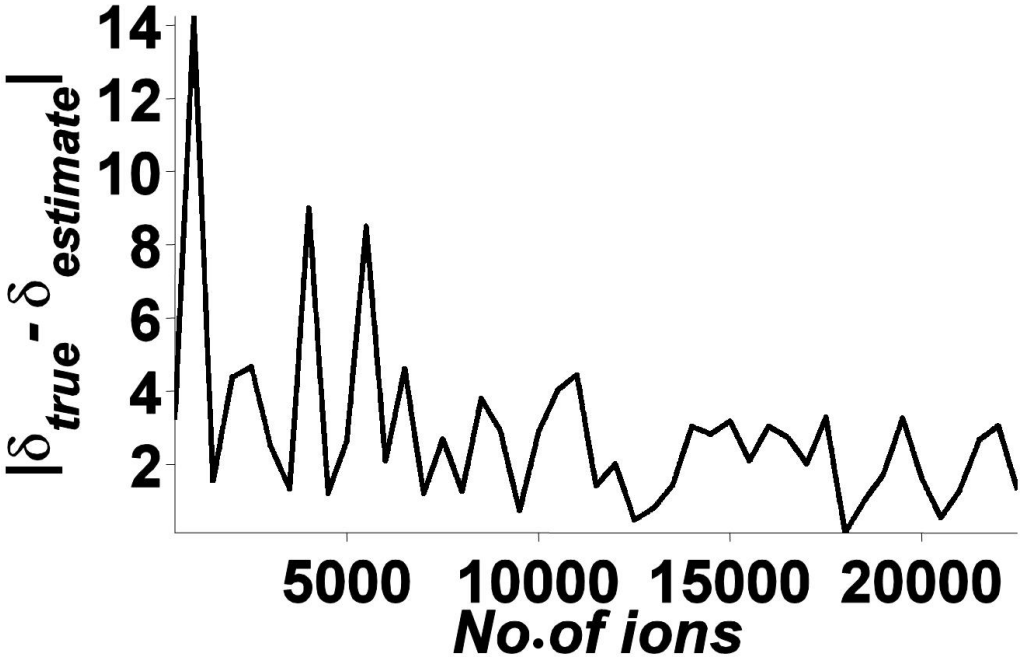


Figure 3·1: Estimate improves with the increase in the number of ions used to generate the simulated isotopic distribution, δ_{true}=-25.5, MW=9000, each estimate was generated using 10 simulations of isotopic distributions

EID approaches to the TID.

The higher the molecular weight of the biomolecule, the greater is the number of carbon atoms and the number of isotopic peaks present in its isotopic distribution, because a greater number of combinations of isotopes becomes possible with increasing molecular weight. So, an investigation of the effect of molecular weight on the delta estimate was carried out. Isotopic distributions were generated using Monte-Carlo simulations, with molecular weight varying from 1000 Da to 20000 Da, with increments of 500 Da. Each distribution was generated using 100000 ions, with the true value of delta being again -25.5‰. The results are plotted in Fig 3·2, which indicate that with the increasing molecular weight, the error in delta drops since more information is available from a higher number of isotopes present in the isotopic distribution. An important observation is that as the molecular weight increases, the number of ions required for the EID to approach the TID also increases (Kaur and O'Connor, 2004) due to the widening isotopic distribution. Thus,

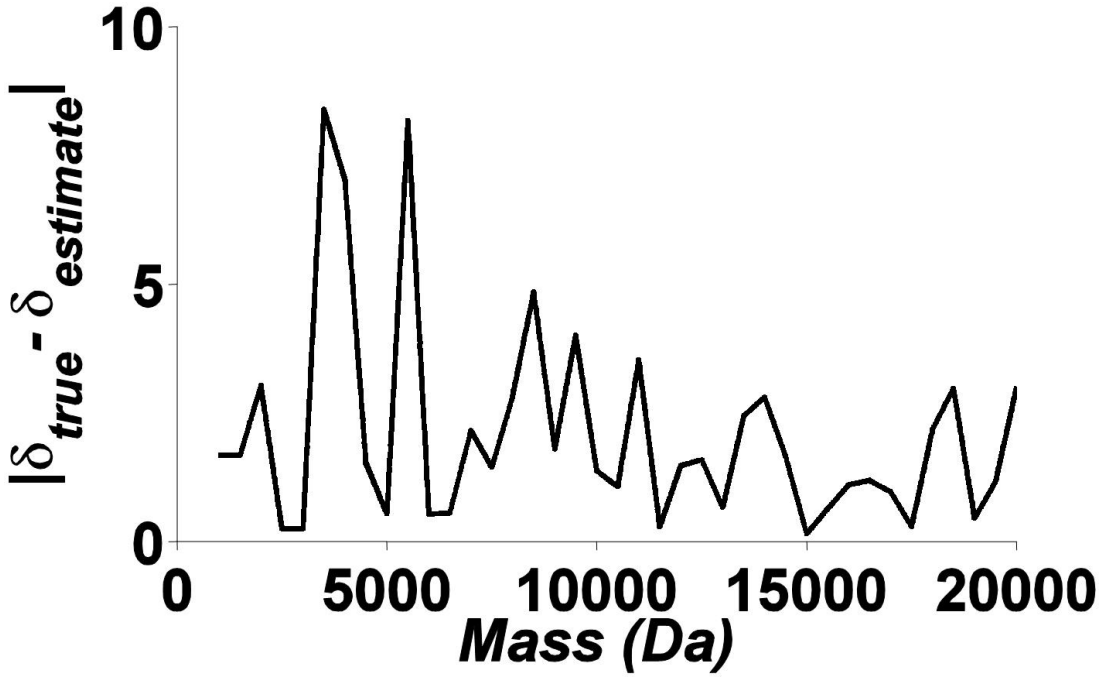


Figure 3·2: Estimate improves with the increase in molecular weight due to higher number of isotopes present in the isotopic distribution, δ_{true}=-25.5, δ_{Est}=-25.68

the improvement in the estimate with the increasing number of observed isotopic peaks may be somewhat offset by the higher variance in the EID for higher molecular weight for the same number of ions. Thus, Fig 3·2 shows that, for 100000 ions, error in the observation of $\delta^{13}C$ values is minimized in the mass range of ∼8-15 kDa, which includes the mass convenient for proteins such as ubiquitin, cytochrome c, and lysozyme, among others. All the isotopic peaks with peak heights greater than 5% of the highest peak were used to arrive at the final estimate, because in practice, not all the isotopic peaks are observed in the EID. The less abundant ones are difficult to observe since their abundance often falls below the noise baseline, and their abundance is usually highly distorted by noise making their value in the estimate highly suspect.

Thus, the theory from equations 3.10, 3.11, and 3.12 works well using simulated EIDs. But what about the performance for real life spectra? Thus, the theory established above

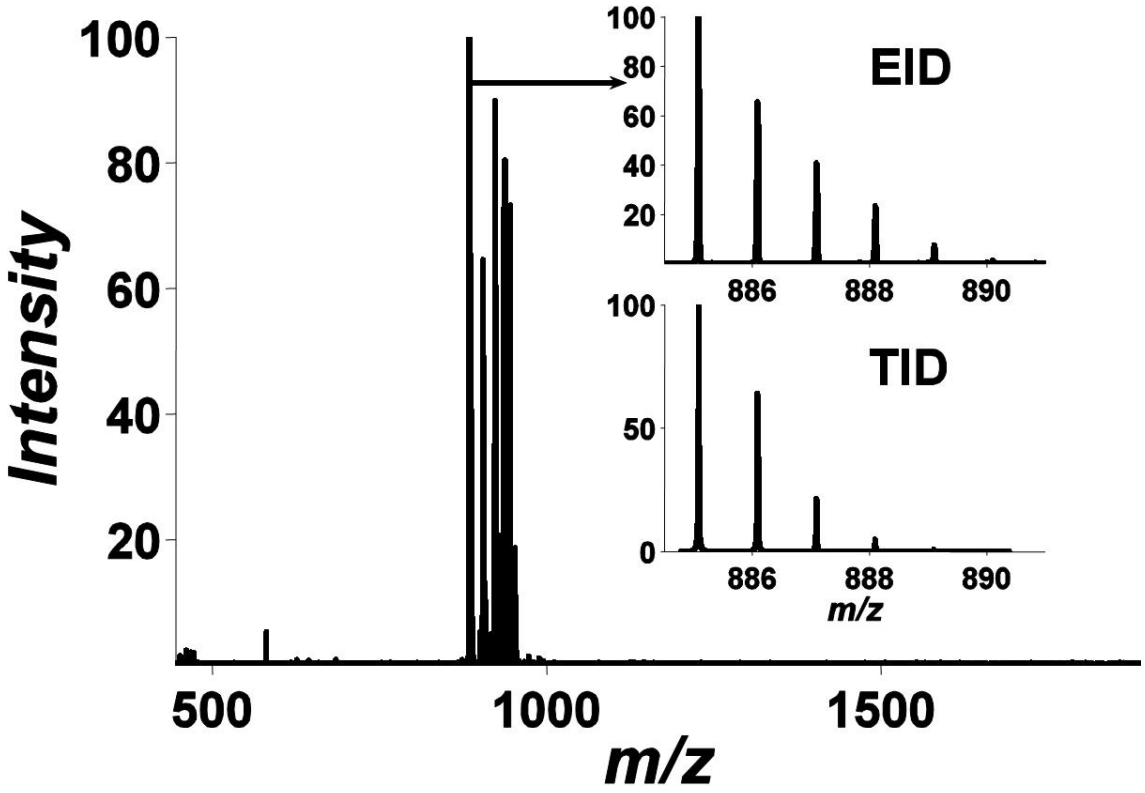


Figure 3·3: Mass Spectrum of chlorophyll-b ($C_{55}H_{70}MgN_4O_6$, MW=906.83, Mg is replaced by 2 H atoms in acidic medium, leading to MW=885.55) from spinach, the desired EID consists of multiple overlapping components demonstrating one of the difficulties of this approach.

was subsequently applied to experimental spectra obtained on a custom hybrid electrospray FTICR instrument.(O'Connor et al., 2006; Jebanathirajah et al., 2005) A chlorophyll b sample was prepared using ethanol and 1% formic acid. Chlorophyll b (C3 plant, $\delta^{13}C$ = -26.5‰) spectra were acquired and signal averaged over 100 scans. Mg present in the molecule is removed in the acidic medium and is replaced by 2 H atoms, leading to a monoisotopic molecular weight of 885.55 Da. Fig 3·3 shows one such spectrum of chlorophyll-b. The inset in the picture show the expanded *m/z* region of the EID, along with the TID. The EID, unfortunately, clearly consists of multiple overlapping components, and hence, has an isotopic pattern substantially different from the TID as the insets in the figure illustrate. Since the goal of this study is to differentiate between extremely subtle variations of the isotopic abundances, the interference in the EID of the desired molecule from any other components cannot be permitted, since it will distort the EID and lead to erroneous results. In this case, the overlapping isotopic distributions extend the EID to high mass, which would greatly overestimate C. This example illustrates one of the challenges faced by using this method. It is critical that any EID used for quantitative determination of isotope ratios consist of one and only one component, and that all of the isotopic variance be due to the variance in the isotope ratios.

Experimental spectra of bovine ubiquitin (8.5 KDa) were obtained in order to further evaluate the theory against the experimental data. Fig 3·4a shows the mass spectrum of bovine ubiquitin with isolation of 10+ charge state in the quadrupole, Q1, averaged over 550 scans. The insets show the expanded view of the EID and the TID. Fig 3·4b shows the plot of peak areas of the EID and the TID for the same isotopic distribution in circles and stars respectively. There appears to be a systematic bias in the EID when compared against the TID, the peak areas of the first few isotopes in the EID are smaller than their theoretical counterparts, and the difference becomes smaller for higher isotopes. This bias may be attributed to the ion selection in the front end quadrupole. Some of the isotopes may be preferentially accumulated over the others due to the *m/z* window on the

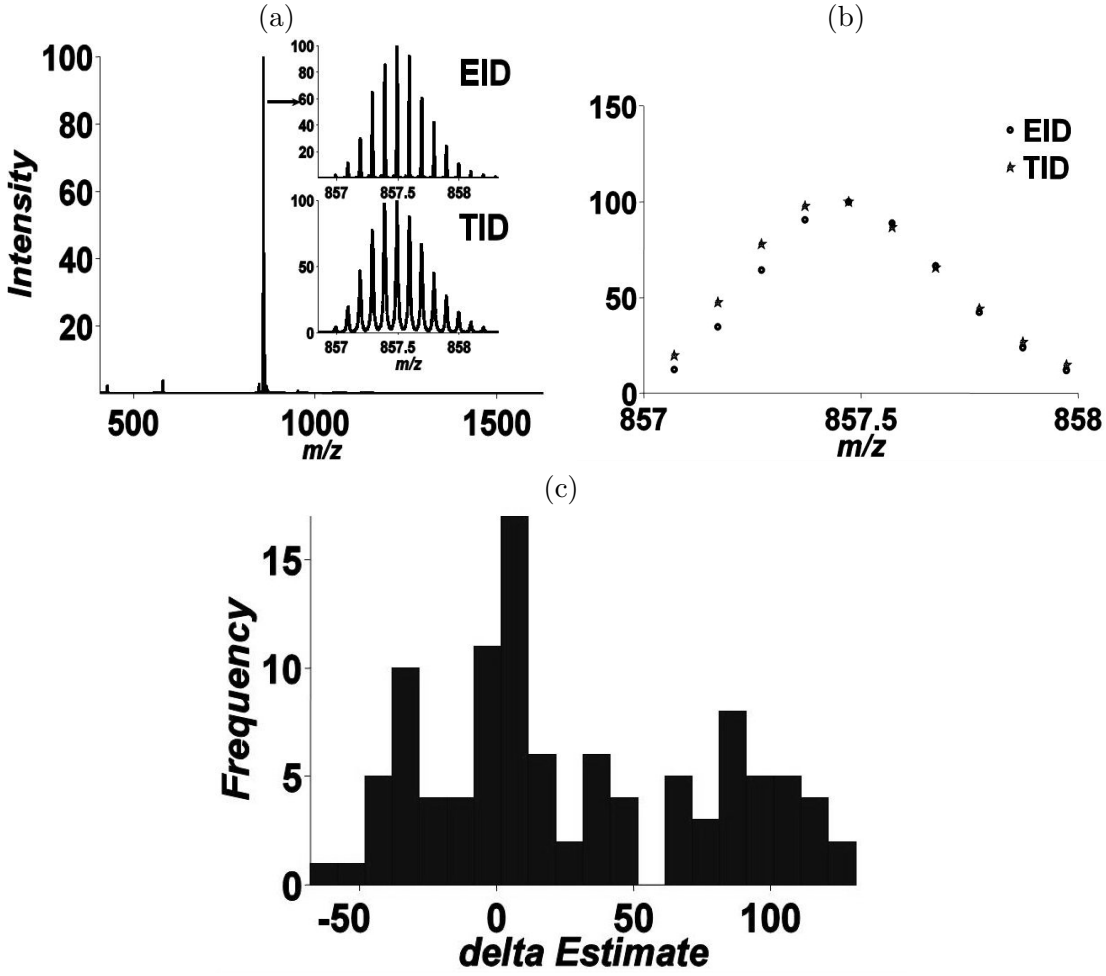


Figure 3·4: (a) Mass Spectrum of bovine ubiquitin with the front end isolation of 10+ charge state at m/z =857.5 (b) The EID (circles) differs from the TID (stars) due to the isolation artifacts, showing that care must be taken to prevent isotopic distribution distortion during the experiment (c) $\delta^{13}C$ estimate using 19 spectra of bovine ubiquitin, with isolation of 10+ charge state, delta median value=7.62 from the 103 estimated values from different isotopic peaks

quadrupole. This phenomenon was observed despite the fact that 10 Da wide window was used for ion selection. Fig 3·4c shows the histogram of the delta estimates obtained using 19 spectra of bovine ubiquitin. The results were generated from 103 isotopic peaks for the 10+ charge state. Due to the artifacts present in the EID because of isolation, the EID does not represent the true TID, and hence, the delta estimates obtained are also likely to be erroneous since the theory established above requires the EID to represent the TID very closely in order to give good estimates. Fig 3·4c shows that the delta values do not converge to a median value and represent a very wide range of values. Thus, experimental techniques which distort the isotopic distributions, such as quadrupole isolation shown here, prevent calculation of the true δ values. This is true even for extremely subtle distortions like the one shown.

Due to the problems associated with the EID distortion because of the isolation in the quadrupole, isolation was disabled, and all the charge states were allowed to pass through the ion optics and collected in the cell. The resulting spectrum with the averaging of 500 scans is shown in Fig 3·5a, with the inset showing the EID of 10+ charge state within the spectrum. The peak areas of the EID and the corresponding TID are shown in Fig 3·5b, with the EID and the TID being represented by circles and stars respectively. This figure shows that the EID matches very closely to the expected TID once the EID distortion caused by quadrupole isolation has been eliminated. For a complete evaluation, 24 bovine ubiquitin spectra were generated, and a total of 392 isotopic peaks from the charge states 9+ and 10+ were used for estimating the delta values. The resultant histogram of delta estimates is shown in Fig 3·5c. Note that the estimates tend to converge to a central median value of -27.55$\pm$2.89, which compares favorably with the expected $\delta^{13}C$ value (-25.5) for animals whose diet primarily consists of the C3 plants.(DeNiro and Epstein, 1978) It should be noted that this method may not suited for the analysis of heterogeneous populations of molecules, since it will increase the possibility of having multiple overlapping components in the isotopic distributions, and hence, distorting their shape, as was demonstrated with

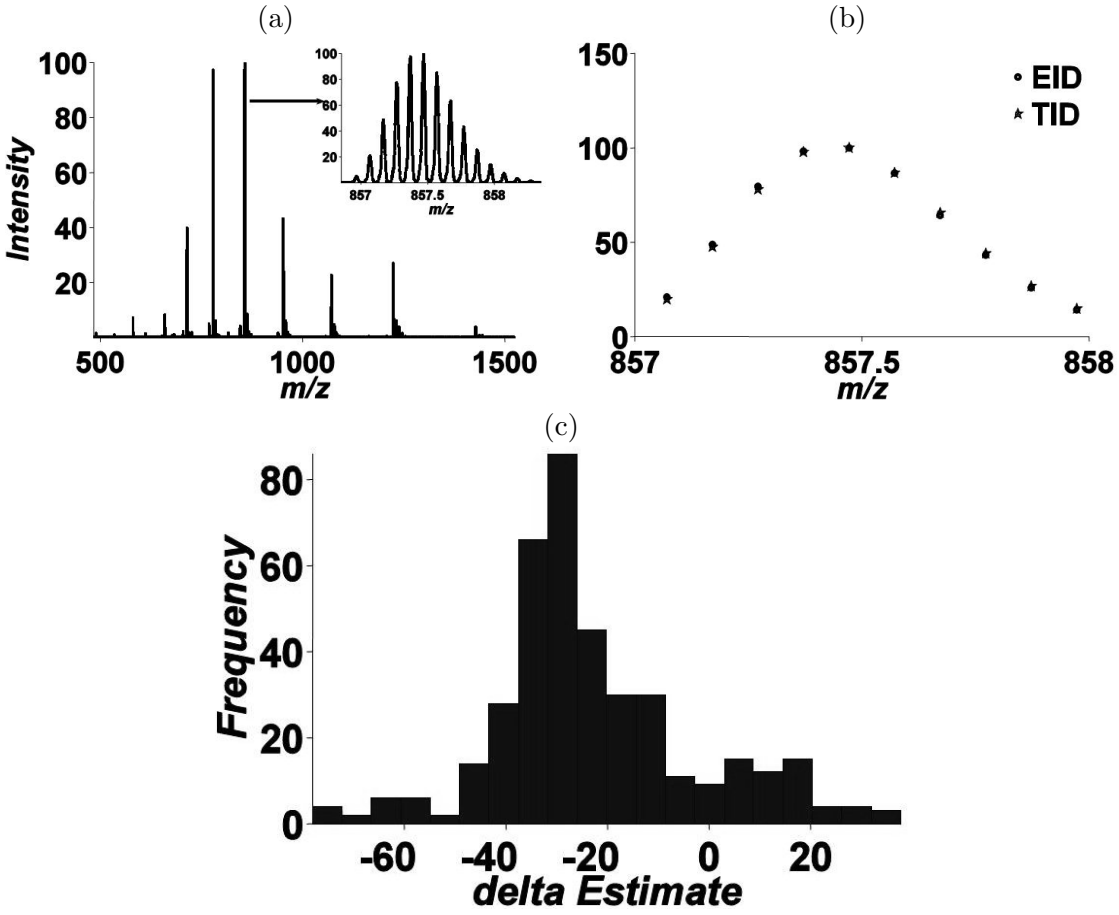


Figure 3·5: (a) Mass spectrum of bovine ubiquitin (b) The EID (circles) matches well with the TID (stars) (c) Delta estimate values using 26 spectra of bovine ubiquitin, median value=-27.55 from 392 estimated values from different isotopic peaks, indicating that the sample fed primarily on C3 plants

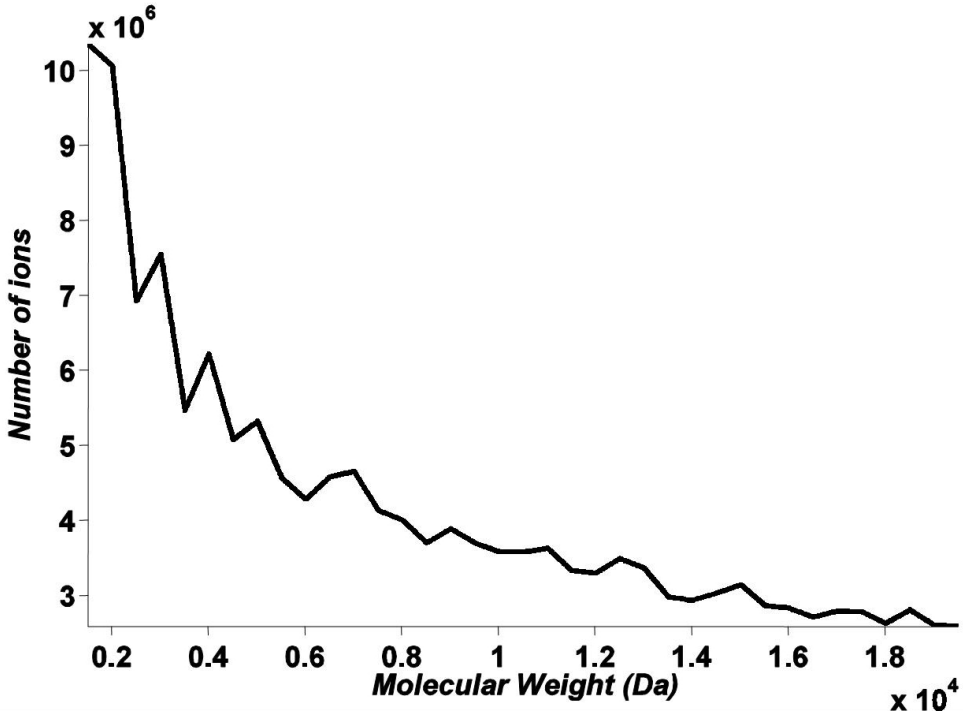


Figure 3·6: Number of ions required Vs Molecular Weight to measure the delta ^{13}C value within 1‰ with a probability of 0.95

the chlorophyll spectrum above.

Figure 3·6 shows the results of a Monte-Carlo simulation to determine the number of ions required to ensure with a 95% probability that the estimated delta value is within ±1‰ as a function of molecular weight. The true value of $\delta^{13}C$, δ_{true} was taken to be -25.5. The results were obtained as follows. Natural abundance values of ^{13}C (p_c) were calculated for the δ_{true}±1‰. TIDs were generated for the resulting p_c values. The ion number was estimated using Chebyshev's bound results in equation 3.16, e(i) was obtained by the difference between the ideal TID and the TID resulting from the p_c values corresponding to δ_{true}±1‰, and probability $\hat{P}$ was taken to be 0.95. The results demonstrate that with the increasing molecular weight, smaller number of ions are required in order to estimate the δ value with the same accuracy. This is intuitive since the higher molecular weight ions have greater number of carbon atoms and hence, subtle changes in δ, and thus p_c, will result

in greater changes in the isotopic distribution. This is because, as seen in equation 3.5, the isotopic distribution involves the term containing convolution of the natural abundances of carbon N_c times, where N_c is the number of carbon atoms present in the molecule. Hence, subtle changes in $\delta^{13}C$ values will produce a more pronounced effect for higher molecular weights due to greater number of multiplications of p_c involved than with the smaller molecules.

With FTMS, the space charge limit bounds the total number of charges that can be detected without peak distortion. Fig 3·6 indicates that, the larger the molecule, the smaller is the number of ions required to achieve the same level of accuracy in $\delta^{13}C$ values. Thus, for FTMS, increasing the mass of the molecule used for $\delta^{13}C$ measurements and decreasing the charge state increases the accuracy, provided baseline isotopic resolving power is achieved.

This approach can be used for the determination of isotope ratio for any element. The equations above can be modified to treat the isotopic abundance of the desired element to be unknown, and then solve for the unknown value. The discussion here was based primarily on the isotope ratio of carbon because of its utility in a wide variety of applications. This approach may be extended to determine the variations in natural abundances of multiple elements if mass spectra from multiple samples originating from the same source are available. If there are m unknown abundances, and $\geq$m different spectra from the same source are available, m different sets of equations (one from each spectra), similar to equations 3.9-3.12, can be used to solve for m variables using algebraic methods.

As shown, this method is very sensitive to the distortions introduced in the experimental isotopic distributions arising due to many factors such as quadrupole voltages, overlapping isotopic distributions, distortions due to modifications such as deamidation. Hence, this approach is not suitable for use with ion trap or quadrupole instruments. Time of Flight instruments may be better suited for this purpose due to minimal isotopic distribution

distortion, provided that the observed isotopic distribution provides baseline resolved isotopes.

It is possible to improve the performance of this method with an FTICR instrument using SWIFT (Stored Waveform Inverse Fourier Transform).(Wang et al., 1986) SWIFT excitation will allow equal excitation to all the isotopes within the isotopic distribution, and hence, minimize the signal artifacts. Also, using SWIFT, it is possible to eject ions that are not of interest, providing an opportunity to accumulate a greater number of ions of interest within the cell, which will lead to a smoother experimental isotopic distribution, and result in a more accurate final estimate, provided the space charge limit of the cell is not exceeded.

This approach can be useful for many applications that require detection of subtle distortions. For example, this may be a method of choice to check whether the instrument response is linear and stable in a given narrow m/z range. A known molecule, whose elemental isotopic abundances are known, can be used to generate the experimental spectrum, with signal averaging in large numbers. Experimental spectrum may then be compared against the theoretical spectrum in order to test for any isotopic distortions using this method. Any distortions observed may be attributed to the non-linear response of the instrument.

The number of ions sufficient for the experimental isotopic distribution to approach its theoretical counterpart depends upon the molecular weight of the molecule of interest. For higher the mass molecules, greater numbers of ions are thus required for the experimental and theoretical distributions to "match" against each other. There is a tradeoff between the number of ions and molecular weight for a "smooth" experimental isotopic distribution. Higher molecular weight molecules are advantageous because of greater number of carbon atoms and isotopic peaks in the distribution. Lower molecular weight molecules require fewer ions to provide ideal experimental isotopic distribution, but they contain fewer carbon

atoms and lead to smaller number isotopes, and hence, less information.

In this study, isotopic peaks were fitted to Lorentzian peak shapes (Marshall and Verdun, 1990), and the areas of the fitted peaks were calculated in order to calculate the peak intensities. Careful consideration must be given to avoid any peak distortions due to space charge effects in the ICR cell to avoid erroneous results. And care must be taken to use the best possible peak shape fitting method in order to calculate the most accurate peak areas which are true representatives of isotopic intensities. Peak height intensities are generally unacceptable due to approximation errors.

3.5 Conclusions

A theoretical framework has been developed and tested for estimating the elemental isotopic abundances from the experimental isotopic distributions. The estimate improves with increasing number of ions generating the isotopic distribution. Higher molecular weights are particularly useful for a better estimate because the higher number of carbon atoms and isotopic peaks observed lead to a greater amount of information. However, higher molecular weights also require a higher number of ions in order for the EID to converge to the theoretical isotopic distribution, which is required for reliable results. This method circumvents some of the limitations experienced by the traditional IRMS by providing a greater flexibility about the kind of samples that may be used for the analysis. This approach is applicable for isotopically resolved spectra from any kind of instrument. For optimal results, experimental isotopic distribution is required to have minimal artifacts due to the subtle nature of the measurements being made. It is very important to avoid any perturbations in the experimental isotopic distributions due to noise or other sources like influences from overlapping isotopic distributions, intensity artifacts due to bias in quadrupole voltages etc.

Chapter 4

Charge State Determination Methods for High Resolution Mass Spectra

4.1 Introduction

Electrospray ionization generates ions with multiple charge states. For proteins, in standard electrospray solutions, these charges arise by the adduction of available protons from the acidic solution to the protein. Since charge is quantized, it can take up only integer values. As both the solution and the protein itself partially shield the protons from each other, the number of charges can be quite large. A typical example is the charge state distribution of myoglobin, a ≈ 17 KDa protein, whose charge state envelope generally extends from 7^+ to 23^+. Since all mass spectrometers measure mass-to-charge ratio (m/z), this corresponds to m/z values of about 2423 at 7^+ down to 740 for 23^+. In order to measure the mass, the charge value must be determined.

Currently, spectrum interpretation represents one of the biggest bottlenecks in a mass spectrometry experiment. Manual analysis of such a complex data is very tedious and time consuming. Hence, there is a great need for reliable sophisticated data analysis methods (Mann et al., 1989; Reinhold and Reinhold, 1992; Henry and McLafferty, 1990; Senko et al., 1995b; Senko et al., 1995a; Zhang and Marshall, 1998; Horn et al., 2000) in order to achieve high-throughput results. An attempt has been made in this work to partly solve this problem by proposing a new approach for charge state determination using the Matched Filter (MF)(Haykin, 1994; Duda et al., 2001) technique. In this chapter, detailed comparison of the previous methods against the newly established method are done under

various conditions.

4.2 Previous Work

The first attempt for automatic charge state determination was (somewhat erroneously) called "deconvolution"(Mann et al., 1989; Reinhold and Reinhold, 1992; Zhang and Marshall, 1998; Henry and McLafferty, 1990). Since charge states can take only integer values, "deconvolution" methods combined isotopic peaks of the same mass with different charge states to determine the mass of the ion. For example, a molecule with mass=3000 Da will exhibit isotopic clusters roughly at m/z values of 1000, 1500 and 3000 corresponding to z=3, 2, and 1 respectively. The major drawback of "deconvolution" methods is that they perform poorly if a given mass is represented by only one charge state, as is often the case in case of multistage mass spectrometry experiments.

When high resolution instruments such as the Fourier Transform Mass Spectrometer (FTMS)(Marshall and Verdun, 1990; Amster, 1996) are coupled with the multiple charging effect of modern ionization techniques, determination of the charge state, z, of an ion is simply a question of measuring the distance between neighboring isotopic peaks, since, for example, the m/z distance between adjacent isotopic ^{12}C and ^{13}C peaks is $\sim 1.003355/z$.

Due to the above mentioned inherent problems in the "deconvolution" approach, techniques were developed for automated assignment of charge states from the isotopic spacings (Senko et al., 1995a). These methods are briefly discussed here and compared against the Matched Filter (MF) method.

Patterson Method The Patterson routine (Senko et al., 1995a) uses a function similar to Autocorrelation (Oppenheim et al., 2002) function, except that it uses a certain number of pre-determined lag values in order to calculate the autocorrelation values. An ID has a periodic structure depending upon the charge state. Generally, there appears a maxima in the Patterson function plot corresponding to a shift of $\sim 1.003355/z$. Fig 4·1a shows an

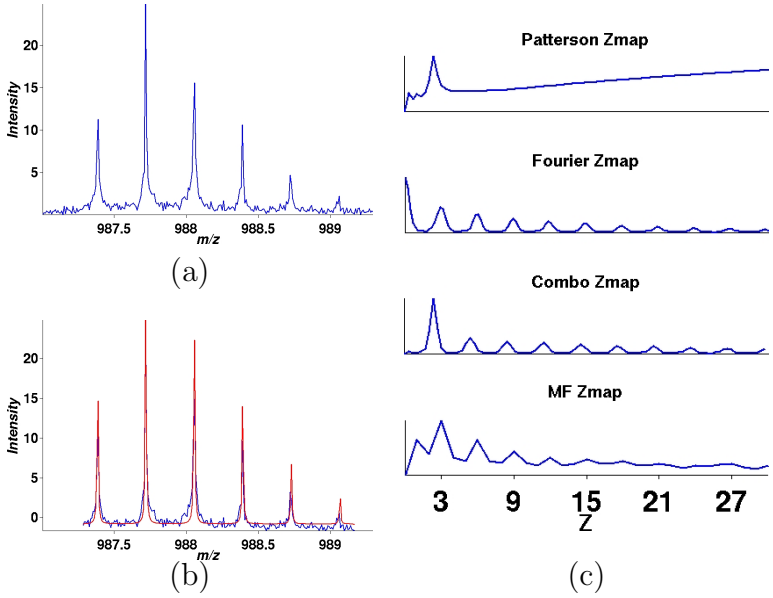


Figure 4·1: (a) EID from top-down spectrum of Bovine Carbonic Anhydrase (b) Shifted TID (red) (shift corresponding to maximum cross correlation coefficient (r=0.978 for z=3)) plotted on the top of EID (blue) (c) Charge state maps using different methods

Experimental Isotopic Distribution (EID), taken from a top-down spectrum of carbonic Anhydrase, that corresponds to z=3. Fig 4·1c shows the "score" (called the Zmap) of each of the charge states using various methods including Patterson method on the top. The Patterson Zmap shows the autocorrelation value as a function of the lag value, which is then mapped to the corresponding z value. For example, a lag of ∼1.003355/2 will correspond to z=2 on the X-axis of Zmap. Autocorrelation corresponding to 0 lag value give the maximum value, but that value was ignored and defined to be zero in the Zmap. The Patterson Zmap shows a strong peak at z=3 and a gradual rise at higher z.

Fourier Transform Method In this case, the Fast Fourier Transform (FFT) of the ID is taken after zero-filling the input signal to the next power of 2. The FFT of the ID in Fig 4·1a is shown in Fig 4·1c. The Fourier Zmap shows a strong signal at z=0 corresponding to the dc component of the signal, which is ignored. The next largest peak is at z=3, with harmonics at z=6, 9, 12, etc.

Combo Method This method is so called because its a combination of the Patterson and Fourier Transform method (Senko et al., 1995a). It takes point-by-point multiplication of the above two methods to arrive at the Combo Zmap shown in Fig 4·1c. It takes advantage of the fact that there will be a peak in both the Patterson and Fourier Zmaps corresponding to the true charge state and the peak will be amplified by multiplication as shown in the Combo Zmap in Fig 4·1c. The Combo routine gives a very good peak in the Zmap at z=3, but again shows harmonics at z=6, 9, 12, etc.

All of these methods work very well when the the signal quality is high, but they all tend to break down when signal-to-noise ratio (SNR) is low and when the input signal represents multiple isotopic distributions overlapping with one another. The Matched Filter method presented here is generally more rugged and reliable as shown below under a wide range of SNR and interference conditions. So there is a need to have a method that can handle the limitations discussed.

4.3 Current Approach

The problem of charge state determination can be posed as follows: Given an input signal called the Experimental Isotopic Distribution (EID), the goal is to design a classifier that will output the maximum value when the given EID most closely matches the Theoretical Isotopic Distribution (TID) corresponding to the true charge state. Let
E= normalized vector representing EID
T_z= normalized vector representing the TID corresponding to charge state z when m/z is given by the location of E in the mass spectrum. This is equivalent to T_z representing the isotopic distribution corresponding to an approximate Molecular Weight (MW)=$m/z \times z$.

Given the values of m/z and z, the Theoretical Isotopic Distribution (TID) is a well defined signal for proteins and can be constructed as follows. If m/z and z are known, an approximate molecular weight of the ion will be $m/z \times z$. Knowing the molecular weight, an average elemental composition of elements can be estimated using poly-averagine (Senko

et al., 1995b) as the "average" amino acid model. After determining the elemental composition of the molecule, its TID intensities can be calculated using binomial distribution calculations (Yergey, 1983) or using the Mercury (Rockwood, 1996) algorithm as is being done in the current work. Peak width at half height for generating the TID is determined by the EID, the value being the same as that of the highest peak in the EID. With the knowledge of the peak heights and width, each peak of the TID is generated assuming a Lorentzian (Marshall and Verdun, 1990) shape, the final TID being the sum of each of the individual Lorentzian peaks.

In the case of problems where the observed signal is very well defined for each of the classes, the Matched Filter (Haykin, 1994; Duda et al., 2001) method often proves to be quite useful for classification. In this case, the Matched Filter output is calculated as follows for each of the charge states z:

$$M_z(n) = E(n) * T_z(-n) \tag{4.1}$$

Equation 4.1 is equivalent to convolving the normalized (to mean zero and variance 1) observed signal E with the normalized, time-reversed theoretically expected signal, T_z. The "score" for each of the charge states is determined by the cross-correlation coefficient, $r(z)$ as follows:

$$r(z) = \max M_z(n) \tag{4.2}$$

Finally, the charge state is estimated by the following expression:

$$z_{MF} = \arg \max_z r(z) \tag{4.3}$$

Fig 4·1a shows an EID taken from top-down spectrum of Bovine Carbonic Anhydrase. Fig 4·1b illustrates detection of z=3 using the MF approach. TIDs corresponding to z=1 to 30 were generated and the cross-correlation coefficients for each of the charge states was calculated as shown in equation 4.2. The coefficient values are plotted in the MF Zmap

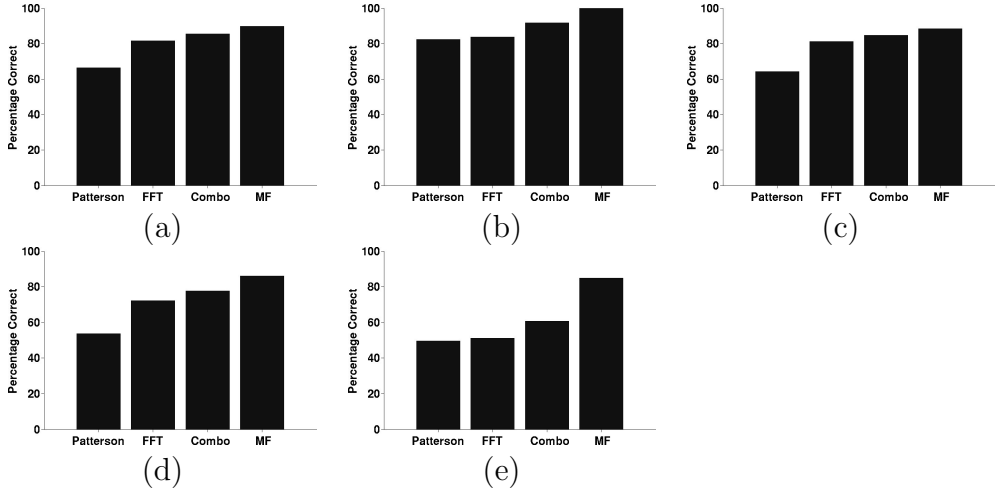


Figure 4·2: Comparison of various charge state determination methods using simulated isotopic distributions for (a) z=1 to 30 (b) Low charge state (z=1 to 3) cases (c) High charge state cases ($4 \leq z \leq 25$) (d) Low SNR (SNR$\leq$4) cases (e) Experimental isotopic distributions with charge states ranging from 1-28

in Fig 4·1c as a function of z. Since the highest coefficient value corresponds to z=3, the assigned charge state is 3. Fig 4·1b shows TID (red) for z=3 plotted on the top of EID (blue), with the shift in TID corresponding to the maximum cross correlation coefficient value.

4.4 Results and Discussion

In order to systematically compare the performance of various methods, 2800 isotopic distributions were generated using computer simulations. The generated data spanned the m/z range from 771-2100, with the SNR ranging from 2-15, with the number of ions (Kaur and O'Connor, 2004) per isotopic distribution varying from 100-900, each of the charge states 1-25 were equally represented by 112 simulations each. An automated analysis for charge state determination of the simulated data revealed that of the 2800 cases considered, Patterson classified the charge states correctly 66.5% times, Fourier Transform method - 81.5% of the time, Combo method - 85.6% time, MF approach - 89.9% time. The overall performance results are shown in figure 4·2a.

4.4.1 Low Charge States

There are mainly two types of popular ionization mechanisms in mass spectrometry - MALDI and electrospray ionization. MALDI mechanism usually produces low charge state (usually 1-3) ions, while electrospray ionization produces highly charged ions and the charge states can vary between 1 to 30 in most cases depending on the size of the biomolecule under investigation. So it is important to analyze the performance for low charge states in order to determine which methods are suitable for MALDI data. Fig 4·2b depicts a performance plot for various methods when the charge state varies from 1 to 3. Out of the 336 cases analyzed, the accuracy results were as follows: Patterson 82.44%, Fourier Transform method 83.63%, Combo 91.67%, Matched Filter 100%.

4.4.2 High Charge States

For electrospray experiments, it is important that the method under consideration gives good results when the input signal represents a high charge state. To this end, performance of various methods for charge states 4 to 25 was analyzed. Fig 4·2c shows that out of the 2464 cases for higher charge states, Patterson performed correctly 64.28%, Fourier gave 81.21% results, while Combo and MF approaches gave 84.74% and 88.47% results respectively.

4.4.3 Low SNR Cases

It has been observed that mostly all the methods give great performance when SNR is high. However, the real test of a method requires it to perform well under low SNR cases. So the robustness of each of the methods was tested under low SNR conditions. For this purpose, an analysis of the 1750 cases, with isotopic distributions having SNR range between 2 to 4 and charge states varying from 1-25, was carried out and final outcome is as shown in Fig 4·2d: Patterson - 53.71 %, Fourier Transform method - 72.23%, Combo-77.66%, MF-86.06%. On the other hand, when test data comprised of isotopic distributions having SNR greater than or equal to 4, the following accuracy was observed: Patterson

- 81.93%, Fourier Transform method - 93.57%, Combo - 95.79%, MF - 94.79%. Thus, while going from high to low SNR cases, the accuracy showed the following absolute drop - Patterson - 81.93-53.71 =28.22%, Fourier Transform method - 21.34%, Combo - 18.13%, MF - 8.73%. These tests demonstrate that Patterson method is most sensitive to the SNR of isotopic distribution, followed by Fourier Transform and Combo methods, while MF approach is least sensitive, and hence, most robust with respect to SNR value.

4.4.4 Experimental Data

After doing performance analysis of the simulated data, tests were carried out on the experimentally generated mass spectra from MALDI and electrospray experiments. A total of 24 different mass spectra of various molecules (ECD spectra of Ubiquitin (charge states 7 through 11), top down spectra of bovine carbonic anhydrase, myoglobin, C_{60}) representing 253 EIDs, with charge states ranging from 1 to 28, and SNR varying from 2 to 15, were analyzed. The complete analysis of the experimental data revealed the performance results as shown in Fig 4·2e, which suggests that the Patterson method gave 49.60% results, Fourier Transform resulted in 51.2% correct answers, while Combo and MF approach gave 60.72% and 84.92% results respectively. The overall performance for each of the methods was dropped because in the simulated data, it is difficult to model completely some of the instrumental artifacts like shift in the spectrum due space charge (Amster, 1996), effects due to neighboring isotopic distributions, low resolution under high pressure in the instrument etc. So the performance usually is lower with experimental data as compared to the simulated data.

Fig 4·3 illustrates the performance characteristics of various methods in a particularly low SNR case taken from an experimental spectrum of bovine carbonic anhydrase. Fig 4·3a shows an EID that is difficult to be detected by an eye when present as a part of the spectrum. The true charge state in this case is 4. According to the Zmaps shown in Fig 4·3c, the assignments were as follows: $z_{Patterson}$=13, $z_{Fourier}$=1, z_{Combo}=2, z_{MF}=4 with

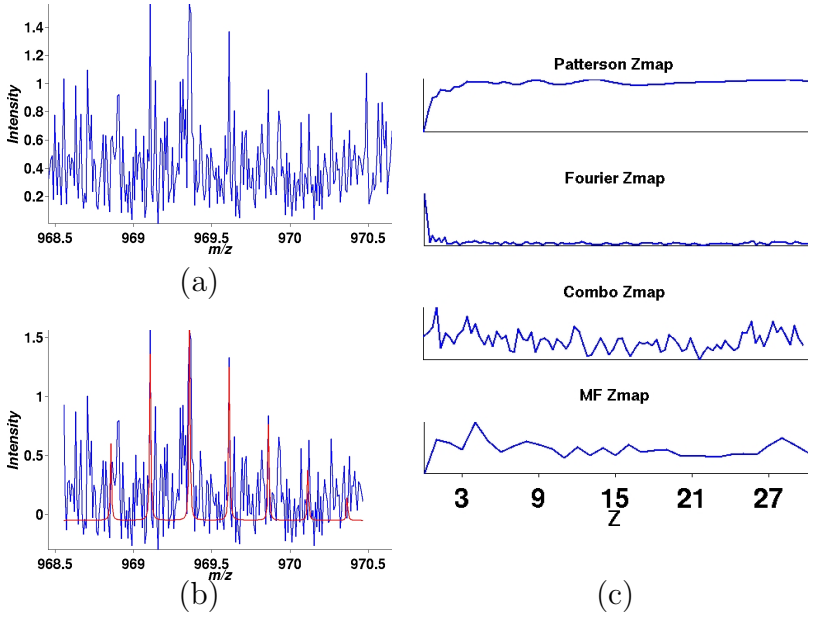


Figure 4·3: (a) An EID from the top-down spectrum of Bovine Carbonic Anhydrase (b) TID (red) (shifted corresponding to the maximum cross correlation coefficient (r=0.5 for z=4)) plotted on the top of EID (blue) (c) Zmaps using different methods

cross-correlation coefficient, r=0.5. Despite the fact that the cross-correlation coefficient is low, the Matched Filter is able to make the correct assignment since the value is higher than the corresponding value for any other charge state. Fig 4·3b shows the plot of EID (blue) and TID (red) corresponding to z=4 that gives the best cross-correlation value.

4.4.5 Overlapping Distributions

Due to the abundance of information present in complex mass spectra, multiple isotopic distributions are commonly observed so that they are present at the same m/z location within the mass spectrum. A good charge state determination method is expected to handle these complicated cases. All the methods discussed so far result in a single output for the estimated charge state in the input signal even if it represents multiple charge states. Thus, an important feature in the automated analysis of spectra is to locate the position of the isotopic distribution so that it can be removed from the input signal, and then followed by further analysis of the residual signal in order to look for any more isotopic

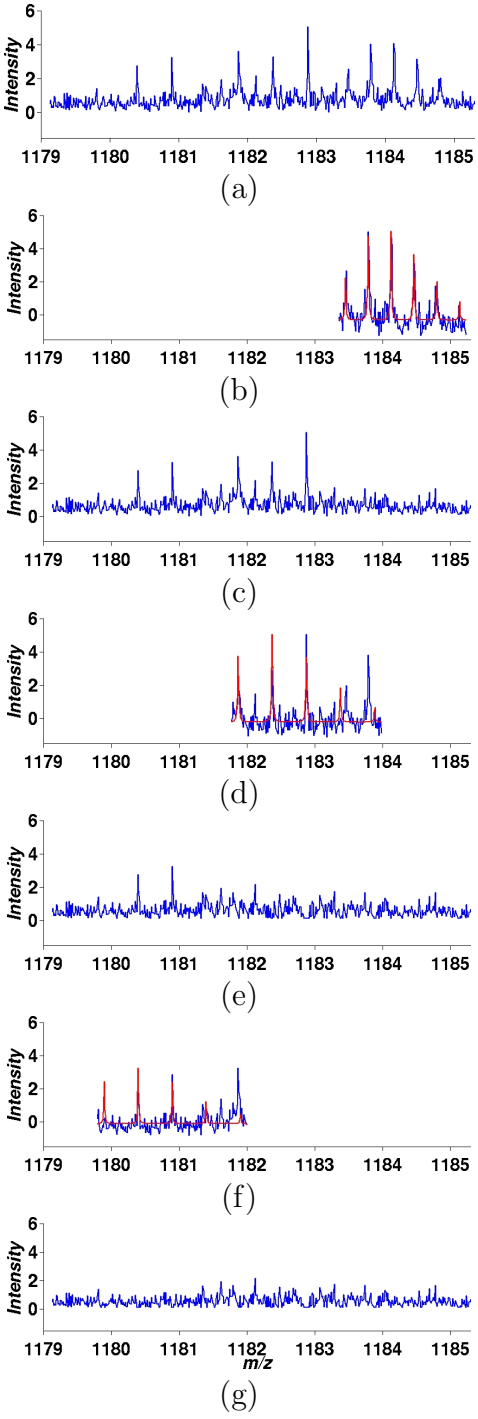


Figure 4·4: (a) Input signal containing multiple EIDs (b) z=3 detected, cross correlation coefficient, r=0.82 (c) residual after subtracting TID for z=3 from (a) (d) z=2 detected, r=0.591 (e) residual signal (f) z=2 detected, r=0.46 (g) Final residual signal

distributions. None of the methods discussed above except the Matched Filter gives such an information about the location of isotopic distribution. The Matched Filter method works by shifting the the TID through the EID and the best "alignment" between the two distributions results in the maximum value of the cross correlation coefficient. This best "alignment" position indicates the position of the EID so that it can be subsequently subtracted from the input signal and the residual signal can undergo further analysis to detect further charge states in the spectrum.

Fig 4·4 gives an example of the case when the input signal EID (Fig 4·4a) represents 3 isotopic distributions in close proximity to each other so that it is difficult to separate them individually even by the eye of the observer. When the Matched Filter was used to determine that charge state, it first detected the presence of z=3 at m/z 1183.3-1185.2 (Fig 4·4b showing TID (red) plotted on EID (blue)). The detected signal was subtracted from the input signal and the residual is shown in Fig 4·4c. When the residual (Fig 4·4c) was subjected to the MF method again, z=2 was detected at m/z 1181.8-1184 (Fig 4·4d), leaving behind the residual as shown in Fig 4·4e. The residual thus obtained represented one more isotopic distribution corresponding to z=2 (Fig 4·4f), which was subtracted to give the final residual as shown in Fig 4·4g. When Fig 4·4g was further subjected to determination of any more charge states, none of the charge states gave the value of cross-correlation coefficient greater than 0.4, which was assigned to be the threshold for a positive identification for a charge state. Thus, a total of three isotopic distributions were detected in the input signal.

4.5 Conclusions

A detailed comparison of the various methods for charge state determination has been carried out. An automatic comparison of the different methods was done using 2800 simulated isotopic distributions, with each of the charge states 1 to 30 being represented equally by 112 isotopic distributions, and SNR was varied from 2 to 15. The results indicated the following performance for the simulated data: Patterson - 66.5 %, Fourier Transform method

- 81.5%, Combo method - 85.6%, Matched Filter - 89.9%. Performance comparison was made under low and high charge state conditions separately, and the results indicated that Patterson and Fourier Transform methods give comparable performance under low charge states (82.4% and 83.6% respectively), while Combo (91.7%) and Matched Filter performed much better (100%). All the methods gave relatively lower performance under high charge state conditions since its harder to distinguish between consecutive charge states when their values become high because the isotopic peaks become really close to each other. Patterson method was shown to be the most sensitive to SNR value, followed by Fourier Transform, Combo and Matched Filter method in that order. Analysis of the experimentally generated isotopic distributions revealed the following performance: Patterson - 49.6%, Fourier Transform -51.2%, Combo - 60.7%, Matched Filter - 86.0%. Results for experimental isotopic distributions were also relatively lower because of instrumental artifacts which cannot be appropriately modeled in the simulated data, making the handling of the experimental data analysis more difficult. The information about the location of the EID is inherently present in the intermediate results produced using Matched Filter approach. This gives Matched Filter method an additional advantage over the previous methods which fail to give such an information. This is especially useful when overlapping isotopic distributions are present. Matched Filter allows for subtraction of the detected EID signals which can be subsequently removed and the residual signal can undergo further analysis.

Chapter 5

Algorithms for Automatic Interpretation of High Resolution Mass Spectra

5.1 Introduction

The wide employment of Fourier Transform Mass Spectrometry (FTMS)(Marshall and Verdun, 1990; Amster, 1996) instruments for Matrix Assisted Laser Desorption/Ionization (MALDI)(Karas et al., 1987) and Electrospray Ionization (ESI)(Fenn et al., 1989) experiments results in thousands of high resolution mass spectra every day, creating an information overload. Due to the high mass accuracy and resolving power of an FTMS, MALDI-FTMS (O'Connor and Costello, 2001) and ESI-FTMS (Henry et al., 1991; Shen et al., 2001) are becoming the instruments of choice for proteomics (Aebersold, 2003) experiments on proteins and large fragments of proteins, so called "top-down"(Kelleher et al., 1999; Reid and McLuckey, 2002; Zubarev et al., 1998) mass spectrometry. These experiments tend to slow down due to the lack of sophisticated methods for automatic spectrum analysis. Currently, spectrum interpretation is one of the biggest bottlenecks in a proteomics experiment. Manual interpretation of such complex data is very tedious and time consuming. While some instrument manufacturers have developed reasonably effective programs for this problem, they rarely publish that algorithm, and thus the strengths and limitations of these methods are difficult or impossible to assess. Hence, there is need for the development of advanced data analysis algorithms (Mann et al., 1989; Reinhold

and Reinhold, 1992; Henry and McLafferty, 1990; Senko et al., 1995b; Senko et al., 1995a; Zhang and Marshall, 1998; Horn et al., 2000). In this work, several new algorithms are discussed that allow for the improved automated reduction of a high resolution mass spectra into a monoisotopic peak list. The proposed name for the unified suite of methods is Mass Spectrum Interpretation and Kernel Extraction (MasSPIKE).

The m/z ratio of most ESI product ions lies in the range of 500-5000 Daltons. Since the same mass can have multiple charge states and there can be multiple isotopic peaks at each nominal m/z value, a very dense, complicated spectrum can be generated. The first attempt for automated spectrum interpretation was (somewhat erroneously) called "deconvolution"(Mann et al., 1989; Reinhold and Reinhold, 1992; Zhang and Marshall, 1998; Henry and McLafferty, 1990). "Deconvolution" was based upon the principle that charge states can take only integer values. It combined peaks of the same mass but different charge states to determine the mass of the ion. Currently, most of the published "deconvolution" algorithms result in spurious peaks due to mis-assignment of charge states. Furthermore, these methods generally perform poorly with low signal-to-noise ratio and complex spectra resulting in missed peaks (false negatives). Also, most "deconvolution" methods bias against peaks represented by only one charge state. Such cases will be poorly represented in these deconvolution approaches, though the Z score (Zhang and Marshall, 1998) algorithm does not suffer from this drawback.

To overcome these limitations, a computer algorithm called THRASH (Thorough High Resolution Analysis of Spectra by Horn) was developed.(Horn et al., 2000) THRASH was the first comprehensive "non-deconvolution" algorithm that addressed the problem of reducing a complex mass spectrum into a mass list with minimal human intervention. It combines various modules of SNR (Signal to Noise) calculation, charge state determination using the Fourier/Patterson (Senko et al., 1995a) method, and least squares fitting for determination of monoisotopic mass. It was a remarkable step towards automated spectrum interpretation and represents the current benchmark in the field. However, THRASH

is based upon certain modules that can be approached differently in order to achieve significantly better results. The work presented here aims to develop better individual modules, and then combine them together for improved data reduction. The comparative results are presented in each section.

5.2 Experimental

The methods are presented here, but their performance characteristics are discussed later. All the methods are being integrated as part of the open source software package BUDA (O'Connor, 2004) and will be available at www.bumc.bu.edu/ftms. MasSPIKE starts with modeling the mean of the noise across the selected m/z range of the spectrum. It then identifies isotopic distributions, marks their location, and determines the charge state for each of the identified isotopic distributions in order to map m/z values to corresponding mass values. Overlapping isotopic distributions are then separated, and the charge state is assigned to each of the resolved overlapping distributions. Then each Experimental Isotopic Distribution (EID) is aligned with its Theoretical Isotopic Distribution (TID) to arrive at the best alignment index for the two distributions. Finally, the monoisotopic mass for each of the resolved isotopic clusters is calculated using results from the previous steps, and the final, minimal, monoisotopic peak list is generated. The mathematical basis of each of these methods is discussed, and the critical equations are"boxed" for the convenience of the reader.

5.2.1 Modeling noise

Baseline noise in a Fourier transform mass spectrum typically has white noise characteristics. Other sources of noise include random electronic RF (Radio Frequency) interference peaks and chemical noise due to unevaporated solvent clusters, which may make the noise non-white. In order to detect peaks, it is critical to know noise levels in a particular region of spectra in the m/z domain. This module aims at modeling the mean of the noise. In order to calculate the SNR across the spectrum, noise mean is characterized as follows.

1. Find the mean of the signal every 1 Da, with 0.5 Da overlap between consecutive m/z windows to assure completeness

2. Every 10 Da (value can be changed by user), the window with the minimum mean is assumed to be the noise window

3. A histogram of the intensity values in the noise window is plotted. The histogram is then truncated to eliminate high intensity values caused by signal or RF interference noise peaks, i.e., the histogram is truncated once the intensity occurrence values reach less than 5% of the maximum occurrence value. Then the mean of the intensity values corresponding to the truncated histogram is calculated and this value is defined as the local noise value. This process is done iteratively until the mean converges. A typical example of this procedure is shown in Fig 5·1(a) and (b). Note that this procedure does not take into account RF interference peaks, but is designed to find the baseline noise level. RF interference peaks will be filtered out in the subsequent modules in order to eliminate false positives.

5.2.2 Isotopic Distribution Identification

The goal here is to identify the locations of isotopic distributions (IDs) based upon the SNR in the spectrum. Here, an ID is identified based upon the fact that SNR for an ID is higher than a particular user defined threshold. This principle is similar to that used in THRASH (Horn et al., 2000), but this module defines the isotopic distribution boundaries before assigning the charge state, instead of taking a ± 0.5 Da window around the highest intensity peak. This step is very important for good performance of charge state determination. It is carried out by the following steps:

1. Scan the spectrum every 1 m/z unit from low m/z to high m/z

2. Check $\frac{S}{N}$ in every 1 m/z window

3. $\frac{S}{N} = \frac{\max(signal)}{\text{mean}(noise)}$, with noise value defined as above

4. If $\frac{S}{N}$ is greater than the user-defined threshold (default value=3), mark the window as potentially containing an ID

5. Combine together consecutive potential ID windows and output: C1 (Start ID m/z value), C2 (End ID m/z value)

6. If there is only one peak within {C1,C2} with $\frac{S}{N}$ greater than the threshold, it typically indicates an RF interference noise peak and is discarded because real mass spectral peaks almost always have an isotopic signature.

Note that C1 and C2 are those values of m/z where the $\frac{S}{N}$ hits the threshold value in the window first and last respectively. A table of {C1,C2} values is constructed and used as input to the charge state determination routine. A bovine carbonic anhydrase "top-down" spectrum (Fig 5·1a) was used to test MasSPIKE and is discussed below. {C1,C2} values for the 1103-1133 m/z region of this spectrum are plotted in Fig 5·1c as arrows below the spectrum with the "up" arrows indicating the start, C1, and "down" arrows indicating the end, C2, of individual IDs. If the threshold value is kept too low, some random noise spikes may be picked, while a value too high value will miss the low SNR isotopic distributions. So a value of threshold=3 was found empirically to be an optimal balance between the two cases, but it can be adjusted by the user.

5.2.3 Charge State Determination

Each entry in the {C1,C2} table constructed above is subjected to the process of charge state determination. Previously this problem has been approached by taking the Fourier Transform, Patterson, and combination (Senko et al., 1995a) charge state maps (Z-maps) of the isotopic distribution. These methods generally work well with good signals, but all charge state determination methods fail under conditions of low SNR or overlapping IDs so that these methods should be compared against each other under those conditions. In addition, a new method using the Matched Filter (Haykin, 1994; Duda et al., 2001)

approach has been developed and compared to the previous methods in the discussion section.

The Matched Filter (MF) Method for Charge State Determination In charge state determination, the goal is to design a detector for a specific known pattern (in this case, the theoretical isotopic distribution for a particular molecular weight while varying charge state). In pattern recognition literature, a standard method for approaching this problem is the use of a matched filter (Duda et al., 2001). Let

E= vector representing Experimental Isotopic Distribution (EID),

T= matrix with N_z (number of possible charge states) rows, such that: Zth row of T, $T(Z)$ is a vector representing the Theoretical Isotopic Distribution (TID) for a given charge state Z. $T(Z)$ is constructed as follows. An approximate average molecular weight ($\mathrm{MW_{approx.}}$) can be calculated from the location of EID and the charge state, Z, under consideration ($\mathrm{MW_{approx.}}$=$m/z \times Z$, where m/z is the location of the center of the EID under investigation). For a given $\mathrm{MW_{approx.}}$, elemental composition is determined using the average composition of a model amino acid, averagine (Senko et al., 1995b). Based on the elemental composition, the Mercury (Rockwood, 1996) algorithm is used to generate the peak intensities of the TID. Peak width at half height for generating the TID is determined from EID, the value being the same as that of the highest peak of the EID. Knowing the peak heights and the width, each peak is generated assuming a Lorentzian (Marshall and Verdun, 1990) peak shape. The TID is finally generated as the sum of individual Lorentzian peaks.

Given an observation E, $T(Z)$ vectors for all the different possible Z values are generated as discussed above. For each charge state Z, the matched filter output is then calculated as follows:

$$M(Z,n) = \sum_{k=-L}^{L} E(k)T(Z,k-n) \tag{5.1}$$

where L is the maximum of the lengths of E and T. This is equivalent to

$$M(Z,n) = E(n) * T(Z,-n) \tag{5.2}$$

where $*$ denotes the convolution operator. Note that this is also equivalent to taking a cross-correlation of the EID and TID. Define

$$M_{\max}(Z) = \max_n M(Z,n) \tag{5.3}$$

$$N(Z) = \arg\ \max_n M(Z,n) \tag{5.4}$$

where $\arg\max_n(M(Z,n))$ indicates the value of n that corresponds to the maximum value of $M(Z,n)$. Since the signal intensities and length of E may vary highly in a given experiment, it is important to normalize both E and T while calculating the "score" of closeness of E and $T(Z)$. This "score" is given by the value of cross-correlation coefficient $r(Z)$, which is given by the following expression:

$$\boxed{r(Z) = \frac{\sum_i (E(i) - M_E)(T(Z, i - N(Z)) - M_{T(Z)})}{\sqrt{\sum_i (E(i) - M_E)^2}\sqrt{\sum_i (T(Z, i - N(Z)) - M_{T(Z)})^2}}} \tag{5.5}$$

where M_E and $M_{T(Z)}$ are the means of E and $T(Z)$ respectively.

The theory of matched filters (Duda et al., 2001) tells us that the value of $r(Z)$ will be maximum when E and $T(Z)$ belong to the same class, which in this case means that they represent the same Z. So the charge state is estimated as follows:

$$\boxed{Z_{\text{est}} = \arg\ \max_Z r(Z)} \tag{5.6}$$

This means that the charge state that corresponds to the maximum value of $r(Z)$ can be assigned as the estimated true charge state. This works satisfactorily provided the given input signal E is composed of only one charge state. In practice, a given input signal

may represent multiple isotopic distributions of different charge states. Thus, the Z values corresponding to $r(Z)$ greater than a certain user-defined threshold are assigned to be the true charge states. Isotopic cluster(s) corresponding to above determined charge state(s) are then subtracted from the observed distribution, and the residual signal undergoes the same procedure to look for any more charge states represented by the residual data similar to the procedure used by THRASH. The process continues till the final residual cannot be assigned any charge state since $r(Z)$ value is below the threshold for all values of Z. The procedure for determining useful threshold values is discussed below.

5.2.4 Alignment between Theoretical and Experimental Isotopic Distribution

A mass spectrum does not generate a unique mass value for large molecules due to the presence of multiple isotopes of the constituent elements. So the question arises as to what mass value should be reported. One way is to report the chemical average mass using average of isotopic peaks, but this suffers from the problem that carbon isotope variability across different organisms limits the mass accuracy (Beavis, 1993; Zubarev et al., 1996) to about 10 ppm. The most significant and accurate mass that can be reported is the monoisotopic mass because its value is unaltered by isotopic variability. The monoisotopic mass (M) of a molecule is the sum of the masses of the lowest mass isotope for each of the elements present in the molecule. The relative abundance of the monoisotopic peak decreases with increase in the molecular weight because of the increased probability for the presence of heavier isotopes with increasing molecular mass. The monoisotopic peak is typically not visible experimentally when molecular weight is higher than 5 kDa because the tiny peak is buried in the noise. Thus, there is need for the development of a method that can estimate the monoisotopic mass based upon the experimentally observed isotopic profile. Previously, this problem was approached by Senko *et al.* (Senko et al., 1995b) and Horn *et al.* (Horn et al., 2000) using a least squares fit between the theoretical and experimental isotopic distribution. This method generally works well, but breaks down in the limit of low number of ions or low SNR. This module targets at solving this problem

rigorously by analyzing the isotopic distributions.

As discussed previously (Kaur and O'Connor, 2004), the EID can be interpreted as a result of a multinomial experiment (with number of trials equal to the number of ions) having multiple outcomes, each with the probability t_i, where t_i represents the area of each individual peak in the TID (i.e., t_0 fraction of the total ions in the cell contains no higher isotopes, t_1 fraction of the total ions contain exactly one +1 Dalton higher isotope (e.g., ^{13}C), etc.). Let vector E represent EID peaks areas, where e_i corresponds to t_i in the TID. E is a Gaussian random vector with mean T and covariance matrix Σ_N (Eq 5.8), where T is composed of t_is ($0 \leq i \leq n$), and is obtained using the poly-averagine (Senko et al., 1995b) model and the Mercury (Rockwood, 1996) algorithm.

The probability of observing E, given that the number of ions in the cell is N, is the given by (Poor, 1994):

$$P(E|N) = \frac{e^{-0.5(E-T)'{\Sigma_N}^{-1}(E-T)}}{\sqrt{(2\pi)^n det(\Sigma_N)}} \tag{5.7}$$

where Σ_N is given by the following expression:

$$\Sigma_N = \frac{1}{N}\begin{pmatrix} t_1(1-t_1) & -t_1t_2 & \dots & -t_1t_n \\ -t_2t_1 & t_2(1-t_2) & \dots & -t_2t_n \\ . & . & \dots & . \\ . & . & \dots & . \\ -t_nt_1 & -t_nt_2 & \dots & t_n(1-t_n) \end{pmatrix} \tag{5.8}$$

where t_is are the components of T, defined by the theoretical isotopic abundances. For big molecules, only a part of E is observed. The goal, therefore, is to align it with the

appropriate indices of T to determine the monoisotopic mass. Thus:

$$P(E|N,i) = \frac{e^{-0.5(E-T_i)'{\Sigma_{N_i}}^{-1}(E-T_i)}}{\sqrt{(2\pi)^n det(\Sigma_{N_i})}} \tag{5.9}$$

which means that E is a Normal (Gaussian) random vector with mean T_i and covariance matrix Σ_i (both mean and covariance matrix vary with the index). The index of T corresponding to first "visible" value of E is estimated using a Maximum Likelihood estimator as follows:

$$\boxed{\text{index} = \arg \max_i P(E|T_i, \Sigma_{N_i})} \tag{5.10}$$

where

$$T_i = T(i, ..., i + L_E - 1) \tag{5.11}$$

where L_E=Length of E.

This means that the first L_E values of T are aligned with E, calculate the probability of its occurrence from expression 5.9, and then T is shifted by 1, until all the possibilities have been considered. The shift of T that corresponds to the highest value of the probability is assigned to be the true index of the first observed value of E. This procedure is illustrated in detail in the Results and Discussion section.

Finally, the monoisotopic mass (M) is calculated as following:

$$M = \frac{m_1}{z} \times Z_{\text{est}} - (\text{index} - 1) \times 1.00235 - Z_{\text{est}} \times M_+ \tag{5.12}$$

where $\frac{m_1}{z}$ is the location of the first "visible" isotopic peak in the EID (i.e., peak location corresponding to $Y_T(1)$), Z_{est} is the estimated charge state (equation 5.6), 1.00235 is the average mass difference between the centroid of each adjacent isotopic peak for polyaveragine (Horn et al., 2000), and M_+ is the mass of the charge carrier (e.g., 1.0073 for a proton). Assuming the experiment was run in a positive ion mode, a charge state of Z

usually means the ion carries Z protons, so the corresponding mass of Z protons (default) is subtracted to get M.

5.3 Results and Discussion

Fig 5·1a shows a top down spectrum of carbonic anhydrase against which MasSPIKE has been tested. Fig 5·1b shows zoomed-in view of the baseline of Fig 5·1a and noise mean variation as a function of m/z (white line passing through the baseline). The plot is consistent with the variation of baseline noise in the spectrum. Noise modeling serves to provide a noise mean value to be used in SNR calculation for the identification of ID locations. Note that noise is not truly white (flat across m/z range), which is due to the "chemical noise" effect caused by unevaporated solvent clusters formed by the electrospray source.

Fig 5·1c shows the result of the ID identification module applied to one low SNR region of the spectrum. Fig 5·1c shows the ID boundaries for the m/z range of 1103-1132 with up and down arrows indicating the start and end of an ID respectively. Very closely spaced IDs (e.g., between m/z 1114 and 1117) are not separated as seen in the figure. Such cases and overlapping distributions are separated later in the charge state determination routine. ID determination allows MasSPIKE to identify the IDs representing both low and high charge states without bias. This method was found to correct a limitation of THRASH, which uses ± 0.5 m/z window around the maximum intensity peak for the charge state determination, restricting the analysis to charge states greater than 2. MasSPIKE, therefore, can be used for both MALDI (typically representing 1^+ or perhaps 2^+ charge states) and electrospray (typically representing high charge states) spectra. Also, THRASH assumes that the isotopic distribution has a symmetrical Gaussian shape around the highest peak, which holds true for molecular weights greater than $\approx$5 kDa, while MasSPIKE makes no such assumption, so is suited for any kind of ID shape.

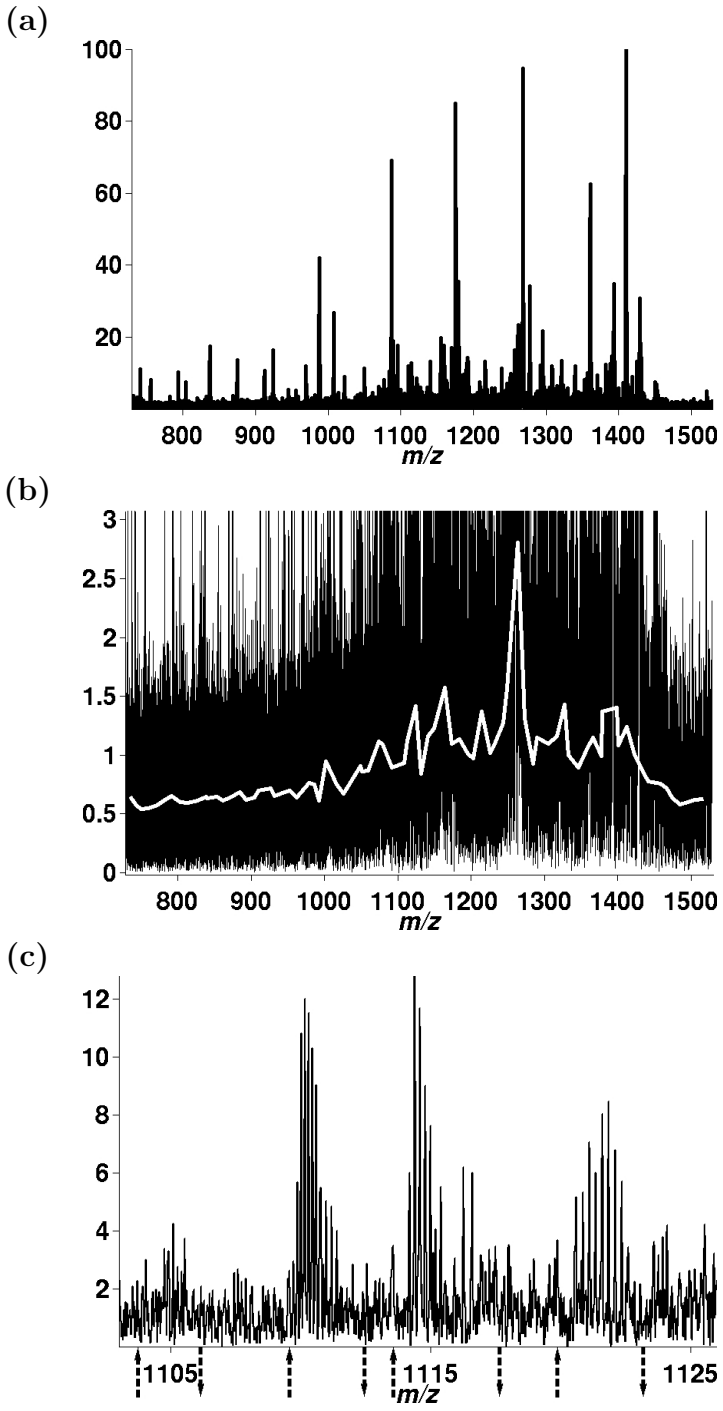
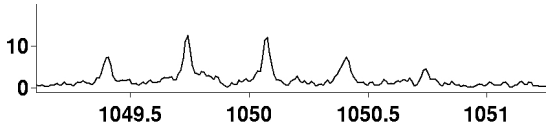
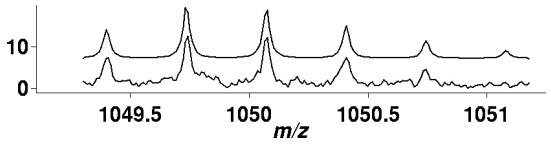


Figure 5·1: (a) Top down spectrum of bovine carbonic anhydrase (b) Zoomed in view of the baseline (black), modeled noise baseline (white) (c) Zoomed-in view of the spectrum,"up" and "down" arrows denote the start and end of an ID respectively

(a) Raw Spectrum



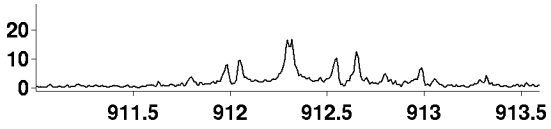
(b) Z=3 detection



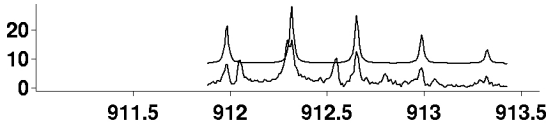
(c)

smz	emz	Z	Res	Ion	MIO	MIT	MIO-MIT
1049.136	1051.012	3	233G27	y	3144.696	3144.713	-0.018

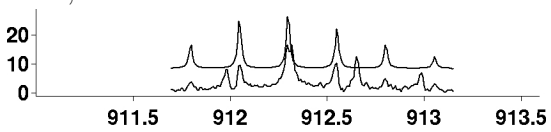
(d) Raw Spectrum



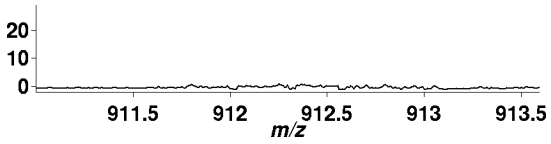
(e) Z=3, r=0.74



(f) Z=4, r=0.64



(g) Residual



(h) Output Listing

smz	emz	Z	Res	Ion	MIO	MIT	MIO-MIT
911.57	913.02	4	V209:L240	IFy	3642.65	3642.93	-0.27
911.76	913.29	3	237L23	y	2732.56	2732.55	0.002

Figure 5·2: (a) Experimental data from Fig 5·1 showing an ID of a bovine carbonic anhydrase fragment (b) TID with Z=3 (top) and EID (bottom). (c) Output listing corresponding to above fragment, y27 (d) EID from Fig 5·1 showing two overlapping distributions (e) Z=3, r= 0.74 (f) Z=4, r=0.64 (g) Residual after subtracting TIDs of (e) and (f) (h) Final output listing.

One of the major challenges encountered in the interpretation of dense, complex spectra is that there is a high chance that the peak of interest is affected by interfering noise peaks or peaks from other signal components (e.g., other isotopic distributions). Fig 5·2 shows a simple case when input signal EID (Fig 5·2a) represents only Z=3. Fig 5·2b shows the plot of $T(3)$ (shifted up) and EID, with shift in $T(3)$ corresponding to maximum value of cross-correlation coefficient (0.954) between the two. Note that $r(Z)$ varies from 0 to 1, so $r(3)$=0.954 indicates a very good match between EID and $T(3)$. The output list from MasSPIKE, Fig 5·2c, shows the starting and ending values of the distribution (smz,emz), the charge state (Z), the assigned amino acid residue region (Res, given the sequence) and ion type (Ion), as well as the observed monoisotopic ion mass (MIO), theoretical ion mass (MIT), and the mass error in Daltons. However, this is an easy case with good SNR, and no overlapping distributions. The real test of automated analysis methods comes at low SNR with distorted peak shapes. Figures 5·2, 5·3, and 5·4 show a couple of such cases extracted from Fig 5·1a. It is important to note that Figs 5·2-5·3 are drawn on the same vertical scale as Fig 5·1a (which is normalized to 100). Thus, Figs 5·2-5·3, with highest intensity values in the 3-15 range, represent parts of the spectrum where the SNR is the lowest, and in particular, Fig 5·3 depicts a case where input signal came from one of the noisiest portions of the spectrum.

Fig 5·2d shows the case when input signal represents two charge states (Z=3 and Z=4), which share a central peak at m/z=912.3. The two charge states are successfully identified and subtracted from the input signal as shown, with TID shifted and plotted on the top of the EID. Note that here MasSPIKE is simultaneously detecting Z=3 and 4 (Figs 5·2e and 5·2f) and the residual after subtraction is free of peaks (Fig 5·2g). By comparison, THRASH proceeds by identifying the charge state represented by the combo (Senko et al., 1995a) routine, and then subtracts the TID from the experimental data. With such an approach, if any of the Z=3 or Z=4 is detected by the combo routine (which is likely), the peak at m/z=912.3 (common peak to both Z=3 and Z=4) will be removed and the

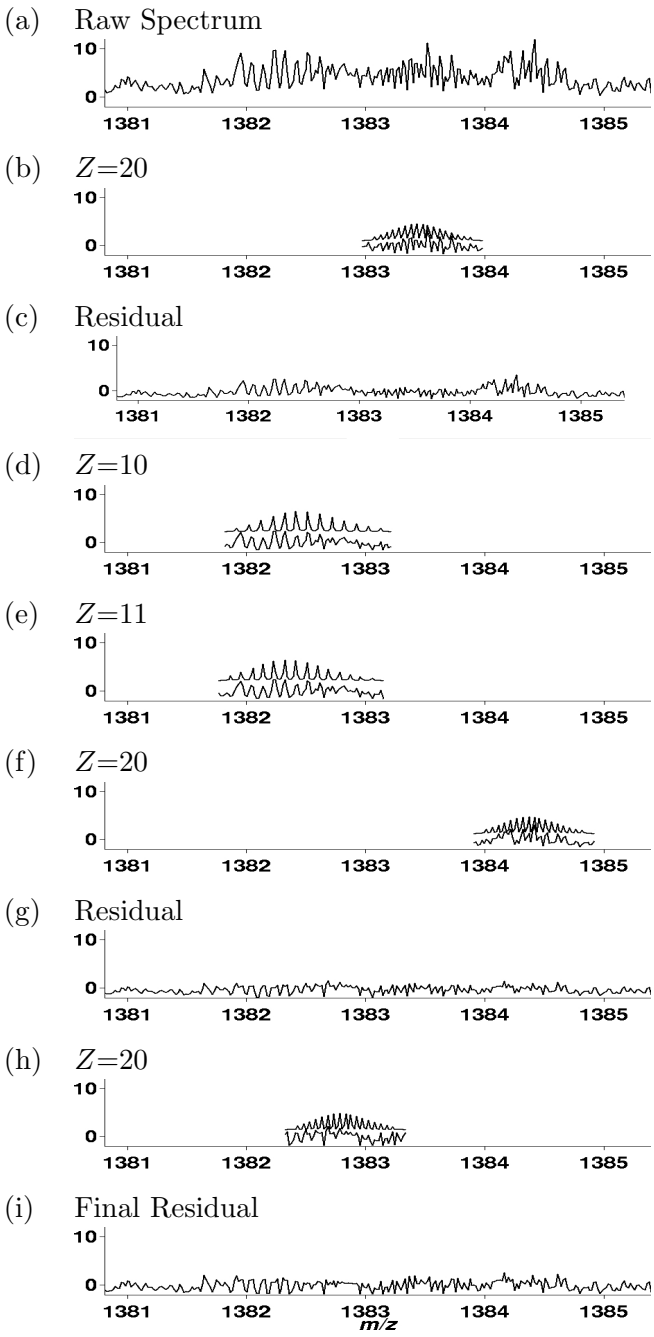
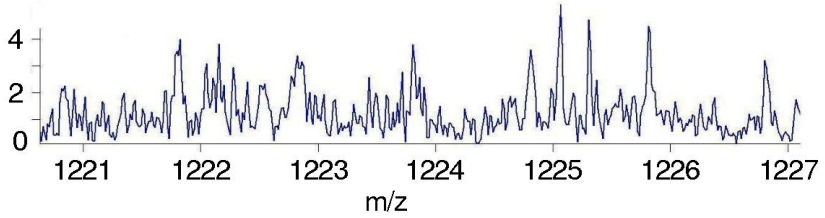


Figure 5·3: Experimental data from Fig 5·1 showing an very low SNR region of spectrum containing 4 overlapping IDs (a) Raw data (b) Z=20, r=0.68 (c) Residual after subtraction of (b) (d) Z=10,r=0.576 (e) Z=11, r=0.56 and (f) Z=20, r=0.65 detected simultaneously (g) Residual after subtraction of (d), (e) and (f); (h) Z=20 (r=0.496); (i) Residual of experimental signal after subtraction of (h)

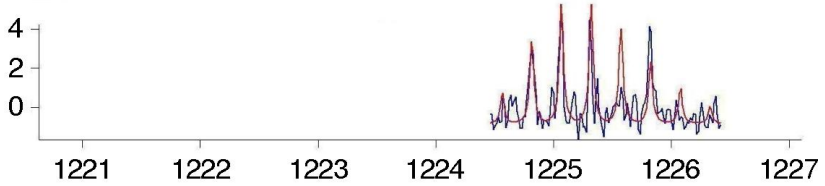
next charge state will not be assigned because the isotopic pattern is perturbed due to subtraction. MasSPIKE attempts to find all the charge states that give cross-correlation coefficient,$r(Z)$, value greater than a certain threshold (default=0.45) before carrying out the subtraction. This allows for assignment of a greater number of charge states. Note that assignment of this threshold represents a balance between missing peaks and generation of false positives. The default threshold value of 0.45 was empirically determined to be a moderate value, but this value can also be altered by the user.

Fig 5·3a shows an input signal from region m/z=1380.7-1385.5 of the Bovine Carbonic Anhydrase spectrum. MasSPIKE was used for determination of various charge states present in the signal. In this case, four isotopic distributions are identified with multiple distributions sharing isotopic peaks and the Z=20 distribution at m/z is identified (Fig 5·3b) and removed (Fig 5·3c). For higher charge states, especially when the sampling rate of the spectrum is low (which is the case at higher m/z since sampling rate in m/z domain drops as m/z increases in FTMS instruments), it is sometimes difficult to distinguish between the consecutive charge state values. For example, for the m/z region between 1381.6 and 1383.2 (Figs 5·3d and 5·3e), the method identifies the charge state values to be either 10 or 11 (though 10 is slightly more likely to be true, r=0.576, than the case of Z=11 where r=0.566). In ambiguous cases like this, a flag is marked and it is left for the user to decide about the true charge state based on the knowledge from the protein sequence, or supplementary information from other portions of the spectrum. Furthermore, MasSPIKE identified two more EIDs with Z=20 in this region of mass spectrum. These masses could not be assigned to a particular fragment ion from the given sequence. However, the approximate difference between the two higher Z=20 ion masses corresponds to the loss of a water molecule, which commonly appears at high molecular weight. For example, the approximate molecular weight for the EID represented by the m/z region 1383.9-1384.9 is 1384.4×20=27688, while that for the m/z region 1383-1384 is 1383.5×20=27670. The difference of the two species (27688-27670=18) corresponds to the loss of a water molecule,

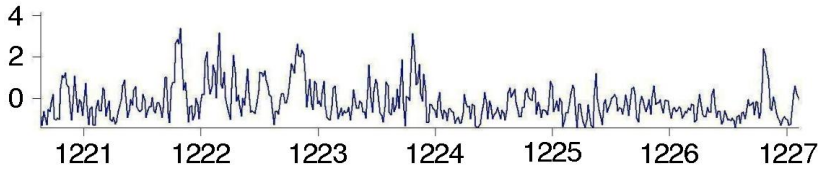
(a) Raw Spectrum



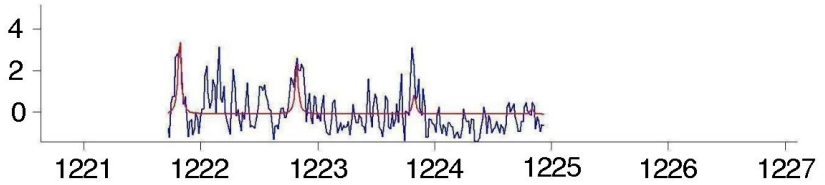
(b) Z=4, r=0.731



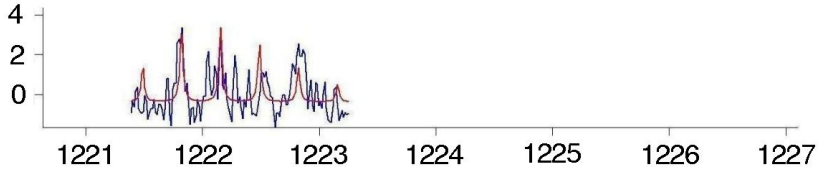
(c) Residual



(d) Z=1, r=0.54



(e) Z=3, r=0.495



(f) Final Residual

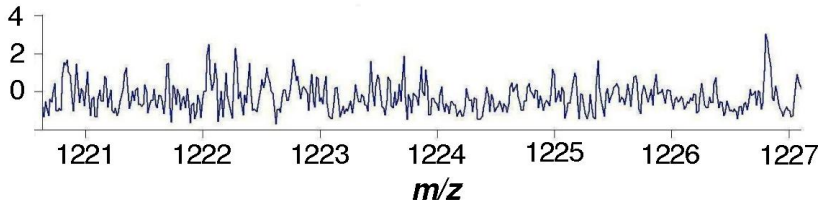


Figure 5·4: (a) EID of a Bovine Carbonic Anhydrase fragment (b) Z=4, r=0.731 (c) Residual after subtraction (d) Z=1, r=0.54 (e) Z=3 sharing 2 peaks with (d), r=0.495 (f) Residual after subtraction of assigned charge states

which suggests the assignment of Z=20 is correct. Also, the final residual from this region, Fig 5·3i seems to contain one or two remaining isotopic distributions. This is an artifact that arises due to the imperfect subtraction of TID from EID, and often happens because of the non-ideal peak shapes of the EID in low SNR conditions as seen in Fig 5·3d and e. Since the residual in Fig 5·3i contains an artifact and not real signal, no further charge state assignments are generated because MasSPIKE does not yield high enough quality assignment (cross-correlation coefficient, r>0.45) for any further charge states. Note that EID and TID take on negative values in some cases (Fig 5·3b-i) because both the EID and TID are normalized, which involves subtraction of the mean, while computing their cross-correlation coefficient as shown in equation 5.5. Fig 5·4 shows the case when the input signal represents 3 isotopic distributions (Z=1, 3 and 4), sharing multiple peaks in the region of m/z 1221-1227.

It is important to test the matched filter method of charge state determination against established methods in an unbiased manner. To this end, 26 electrospray spectra of myoglobin, representing 775 isotopic distributions (resulting from charge states for the whole molecule, water losses and phosphate adducts for the whole protein, and one contaminant species with Z=1) with SNR of 1-100, were acquired and each m/z region corresponding to Z=1-22 in each spectrum (regardless of the presence/absence of signal) was analyzed by five different methods. The percentage correct answers for each method are plotted in Fig 5·5. BUDA (Boston University Data Analysis) (O'Connor, 2004) was used to determine the charge states using the Fourier, Patterson, and combo charge state determination methods (Senko et al., 1995a). In this analysis, the MF method gave correct answers 91% of the time. Of the missed 77 assignments, manual post analysis showed no apparent signal in 50 of them, and the remaining 27 misassigned the charge state by $\pm$1.

There are certain points that need to be addressed while generating the TID. Generating good model distributions is the key to good results. Although the sampling rate is constant in the frequency domain, due to the inverse proportionality relation between frequency

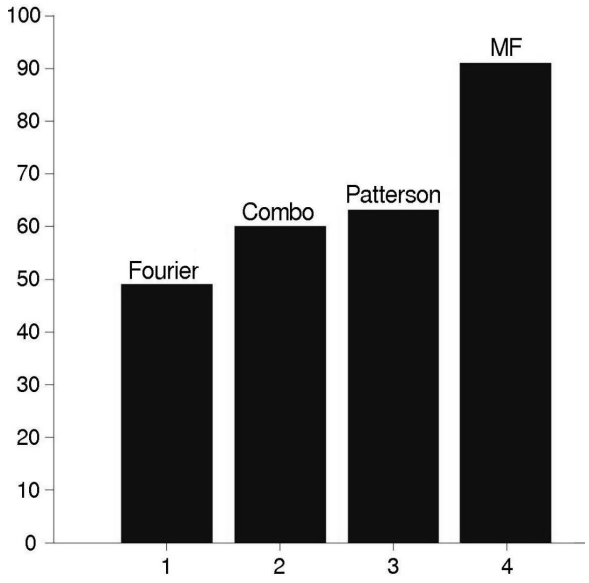


Figure 5·5: Comparison of different charge state determination methods on 775 isotopic distributions from 26 electrospray spectra of myoglobin

and m/z, the sampling rate of an FT mass spectrum is not the same over the whole m/z range, it decreases with an increase in m/z. Thus, parameters for generating good model distributions (peak width, sampling rate, maximum and minimum possible Z ($MAXZ$ and $MINZ$)), must vary with the mass spectral region of interest, and in MasSPIKE they are based upon the observed data and shift through the m/z range. Furthermore, it is important to use unapodized spectra, zero filled once for the experimental data and true line shapes for the TIDs to generate the best matches. For the TID model, the peak width for generating the Lorentzian peaks in the TID is defined by the width of the highest peak in the EID. $MINZ$ is defined by the observed isotopic distribution width. e.g., If the EID spans 1.1 Dalton, $MINZ$=1 but if the EID is only 0.9 Dalton wide, $MINZ$=2, so that it contains at least 2 peaks. This helps eliminate most of the RF interference noise peaks which usually consist of a single high spike. Also, special consideration is given to the number of TID peaks involved in the resulting cross-correlation coefficient. For example, if there is only one peak of the TID matching with the EID that results in the maximum cross-correlation value, it is discarded as a false positive, since it is highly

unlikely for biomolecules to have an isotopic distribution with only one peak. $MAXZ$ is defined by the peak width of highest peak, e.g., Peak Width at Half Height$<$ ($\frac{1}{MAXZ}$). Sometimes, when there are many noise peaks around the main peak, a false identification for a high charge state is generated. Filters have been added to remove these false positives by comparing the peaks of Fourier charge state maps of the theoretical and experimental data. Also, it is required that for a particular molecular weight, the observed isotopic distribution should be wide enough to represent the peaks that are of intensity at least 60% or greater than the maximum intensity. For example, for an observed distribution of molecular weight 10000, the distribution should be wide enough to encompass at least 5 peaks ($>$ 60% of max intensity). Otherwise, it most likely arises as a false positive. These considerations lead to reduced number of false positives and overall better performance.

Fig 5·6 demonstrates the alignment of a typical experimental isotopic distribution (Fig 5·6a) with the theoretical isotopic distribution (Fig 5·6b). The EID and TID are represented by grey and black stick plots respectively in Fig 5·6c-e. Fig 5·6c-e shows the alignment of the EID with the TID, with the TID being shifted by 5, 6, and 7 in Fig 5·6c, d, and e respectively. Fig 5·6f shows the probability of alignment of the EID against the TID with varying shift of TID. A shift of 6 in TID gives the best alignment as depicted in Fig 5·6d and 5h. The normalized probability plot (Fig 5·6h) shows that the probability EID and the TID are aligned properly when the shift is 6 is much higher than its nearest neighbor (index=5). These results are typical with such high SNR ($\approx$20) clean isotopic distributions. However, all alignment methods will work well under these conditions. It is important to test these methods under low SNR and low ion count conditions where large statistical variance occurs in isotopic abundance (Kaur and O'Connor, 2004).

When only 100 ions are present in an isotopic distribution, large statistical variation in isotopic abundance occurs; A typical 100 ion isotopic distribution for myoglobin (16.7 kDa, 16^+) is shown in Fig 5·6g. In order to test the ML method versus the least squares method, 3150 Monte Carlo simulated distributions were generated with only 100 ions per

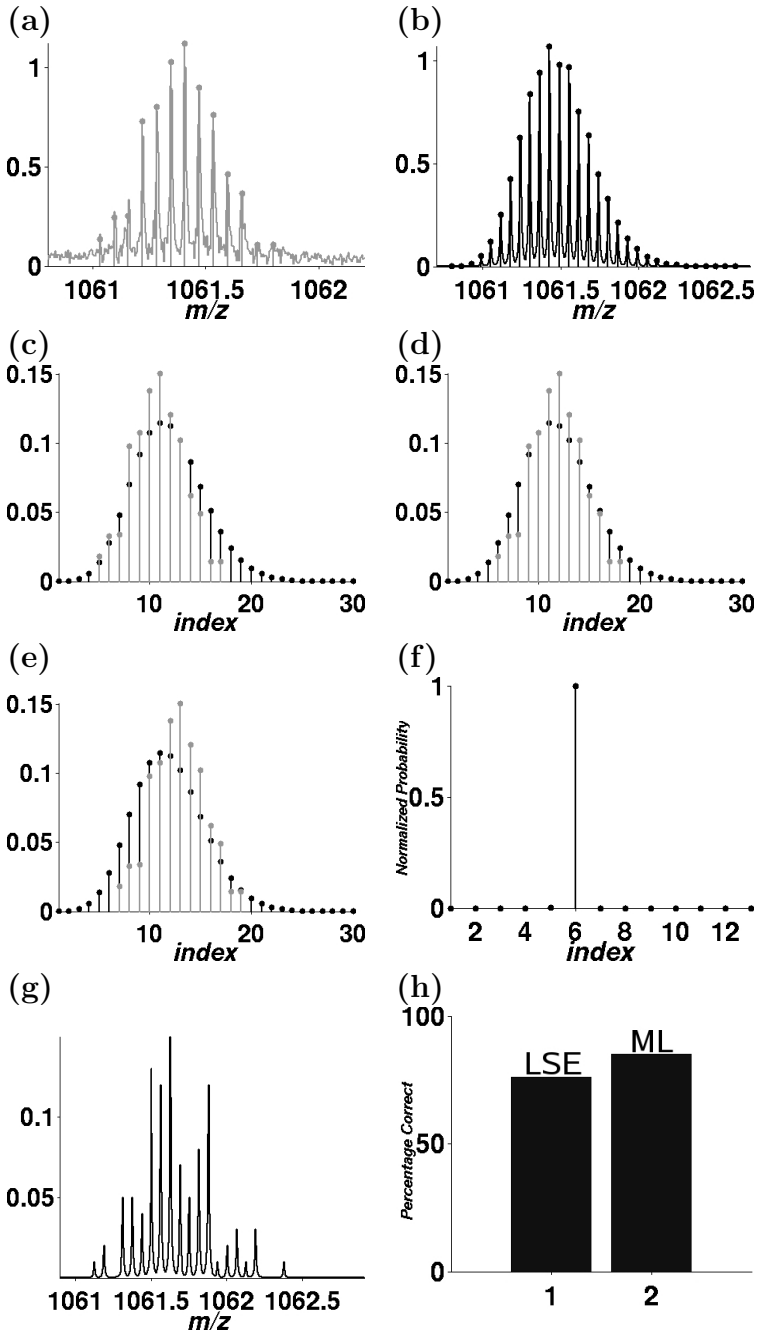


Figure 5·6: (a) EID of myoglobin when Z=16 (b) TID of myoglobin, Alignment of the EID with (c) TID shifted by 5 (d) TID shifted by 6 (e) TID shifted by 7 (f) Normalized probability of alignment as a function of varying TID indices (g) Alignment of myoglobin IDs using 3150 simulations (100 ions in each simulation) (h) A typical Monte-Carlo generated myoglobin isotopic distribution with only 100 ions

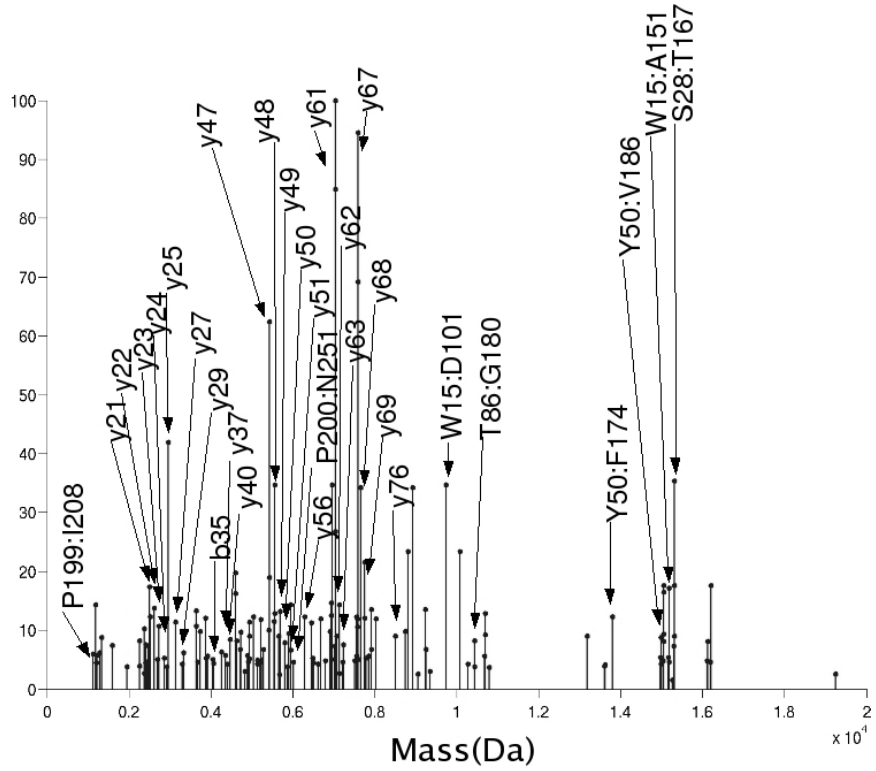


Figure 5·7: Final monoisotopic mass plot of Bovine Carbonic Anhydrase (The full table of masses is included in Table 5.1)

simulation, and the two alignment methods were tested against these distributions. The tests revealed that ML method works correctly 85% of the time, as compared to the least squares error method which gave 76% correct results (Fig 5·6h). Note that it is more difficult to estimate the true index when the distribution is generated by a fewer number of ions since the EID deviates from the TID due to high variance among the isotopic peaks as discussed in our previous work (Kaur and O'Connor, 2004).

After the determination of monoisotopic masses (as discussed above), it is desirable to automatically assign the protein fragments that generated those masses. This requires the knowledge of how a protein or peptide fragments in an experiment (Roepstorff and Fohlman, 1984). MasSPIKE was used to generate theoretical masses of the b and y ions. Internal fragment masses and masses with common losses (e.g. water loss from a molecule)

were calculated knowing the sequence of the protein. The observed masses that match with the theoretical masses of the whole protein and its fragments are then evaluated. A complete analysis of the Bovine Carbonic Anhydrase spectrum revealed the presence of 165 isotopic clusters after eliminating all false positives, which were matched to the closest masses of b or y ions, the corresponding internal fragment ions, and some common losses like water loss, or ammonia loss from a y-ion. The complete de-convolved spectrum representing monoisotopic masses is shown in Fig 5·7. Only abundant peaks are labeled, but the complete monoisotopic mass list is included in Table 5.1. Due to the high energy used for fragmentation, the precursor ion is not observed.

One important limitation of MasSPIKE at this time is the assumption implicit in the poly-averagine model, specifically that the molecule of interest is an "average" protein. Clearly, this assumption fails routinely. A future modification to MasSPIKE will include a DNA and Glycan model as well as the ability to adjust the model manually.

Conclusions MasSPIKE (Mass Spectrum Interpretation and Kernel Extraction), a suite of data analysis algorithms, has been developed. The goal is to reduce a high resolution mass spectrum into a monoisotopic peak list. MasSPIKE identifies isotopic peak cluster locations, determines the charge state for each of the isotopic clusters, resolves overlapping isotopic distributions, aligns the experimental and theoretical distributions, and generates a monoisotopic mass list. If the protein sequence is available, the calculated masses are matched for possible assignments. The method has been applied and tested against complex top-down spectra of Bovine Carbonic Anhydrase. The isotopic distribution identification method is able to identify and mark locations corresponding to both low and high charge states. The Matched Filter charge state determination routine worked correctly 91% of the time for unbiased test data as compared to the standard routines (Senko et al., 1995a), which vary from 48%-64% accuracy. MasSPIKE is capable of identifying multiple charge states in the input signal sharing multiple peaks. Alignment of the theoretical and experimental isotopic distributions with only 100 ions (and hence, high statistical variance)

in the distribution gave 85% correct results as compared to 76% given by the least-squares fitting method.

start m/z	end m/z	Z	Residues	Ion type	Obs. M	Theo. M	Error (ppm)
1136.46	1137.66	1	P199:I208	IFy	1135.6250	1135.6380	11.44
803.36	805.07	2	P200:P213	IFy	1604.9200	1604.9280	5.60
755.00	756.53	3	241A19	y	2262.2720	2262.2610	4.86
792.68	794.22	3	240L20	y	2375.3560	2375.3450	4.63
1239.06	1240.44	2	Q220:L240	IFy	2475.3140	2475.2760	15.35
836.37	837.91	3	239M21	y	2506.3840	2506.3860	0.39
874.06	875.60	3	238L22	y	2619.4700	2619.4700	0.38
911.75	913.29	3	237L23	y	2732.5560	2732.5540	0.73
954.76	956.04	3	236E24	y	2861.5810	2861.5960	5.59
981.22	982.50	3	235P25	y-H2O	2940.6390	2940.6490	3.40
987.12	988.99	3	235P25	y	2958.6480	2958.6490	0.33
1049.13	1051.01	3	233G27	y	3144.6960	3144.7130	5.72
1115.82	1117.70	3	231A29	y	3344.7430	3344.7930	14.94
1299.94	1300.96	3	L210:W243	IFy-H2O	3896.9580	3897.0170	15.13
1311.61	1312.56	3	F129:T167	IFIF2-H2O	3930.1050	3930.1470	10.68
1354.52	1355.70	3	35K225	b	4057.9540	4057.9210	8.13
1092.72	1093.45	4	223K37	y	4365.3330	4365.3430	2.51
1121.24	1122.70	4	T191:N230	IFy-2H2O	4480.3850	4480.4190	7.36
1185.50	1186.77	4	220Q40	y	4737.5450	4737.5260	3.79
1209.34	1209.99	4	P213:V254	IFy	4832.5070	4832.5480	8.48
1238.81	1239.76	4	H121:T167	IFIF2	4949.6150	4949.5820	6.66
1307.33	1309.28	4	215S45	y	5225.7100	5225.7500	7.65
1322.77	1324.00	4	T86:F129	IFIF1	5286.5570	5286.5340	4.35
1355.79	1357.36	4	E212:P258	IFy	5418.8170	5418.8340	3.13
1359.86	1361.82	4	213P47	y	5435.8380	5435.8860	9.01
1088.07	1089.67	5	213P47	y	5435.8480	5435.8860	6.99
1114.01	1115.30	5	212E48	y	5564.9130	5564.9290	2.87
1392.37	1394.52	4	P200:P248	IFy-2H2O	5565.9080	5565.9630	9.88
1136.70	1137.11	5	211K49	y-NH3	5676.0320	5676.0240	1.40
1424.15	1426.35	4	G127:L182	IFIF2-H2O	5692.9450	5692.9630	3.16
1139.53	1141.30	5	211K49	y	5693.0400	5693.0240	2.81
1162.35	1163.76	5	210L50	y	5806.0910	5806.1080	2.92
1182.14	1183.38	5	209V51	y	5905.1450	5905.1760	5.24
1192.59	1194.40	5	P200:N251	IFy	5957.2000	5957.1850	2.51
1299.64	1300.46	5	204S56	y	6491.4930	6491.4880	0.77
1104.42	1105.49	6	F92:A151	IFIF2	6620.2470	6620.2220	3.77
1389.85	1391.38	5	200P60	y	6943.6910	6943.7510	8.64
1158.59	1159.32	6	200P60	y	6943.7980	6943.7510	6.76
1405.85	1407.44	5	199P61	y-H2O	7022.6830	7022.8040	17.22
1171.90	1172.91	6	199P61	y-NH3	7023.7700	7023.8040	4.98
1174.52	1176.05	6	199P61	y	7040.7520	7040.8040	7.38
1006.88	1008.22	7	199P61	y	7040.7850	7040.8040	2.69

1409.25	1411.06	5	199P61	y	7040.7850	7040.8040	2.69
1191.40	1192.23	6	T191:Q253	IFy-H2O	7140.8530	7140.7940	8.40
1191.37	1192.72	6	198T62	y	7141.8570	7141.8520	0.70
1449.85	1451.20	5	W15:K79	IFIF1-H2O	7242.7300	7242.6590	9.80
1208.40	1209.39	6	197T63	y	7242.8390	7242.9000	8.42
1081.50	1082.24	7	193P67	y-2H2O	7560.9790	7561.0900	14.68
1261.57	1262.27	6	193P67	y-2H2O	7561.0210	7561.0900	8.99
1264.43	1265.12	6	Y192:P258	IFy-2H2O	7578.0460	7578.0580	1.58
1083.77	1085.07	7	193P67	y-H2O	7579.0580	7579.0900	4.22
1267.24	1268.94	6	193P67	y	7597.0220	7597.0900	8.95
1086.34	1087.82	7	193P67	y	7597.0700	7597.0900	2.63
1109.63	1111.12	7	192Y68	y	7760.0840	7760.1530	8.89
1294.43	1296.13	6	192Y68	y	7760.1700	7760.1530	2.19
1311.41	1312.94	6	191T69	y	7861.0540	7861.2010	18.69
1417.64	1419.34	6	S104:L182	IFIF2-H2O	8499.3290	8499.3930	7.52
1251.15	1252.00	7	184P76	y	8748.5160	8748.6190	11.77
1133.37	1133.93	8	Y50:D128	IFIF1-2H2O	9055.3460	9055.3110	3.97
1320.97	1321.88	7	T86:G169	IFIF1-2H2O	9236.6730	9236.7180	4.98
1392.70	1394.48	7	W15:D101	IFIF1-2H2O	9740.8360	9740.8090	2.66
1306.72	1307.22	8	T86:G180	IFIF1-2H2O	10442.2970	10442.2540	4.11
1075.46	1075.74	14	Y50:V186	IFIF1-H2O	15035.4400	15035.4630	1.52
1255.88	1256.65	12	Y50:V186	IFIF1	15053.2460	15053.4630	14.34
1159.35	1160.11	13	Y50:V186	IFIF1	15053.2960	15053.4630	11.02
1076.60	1077.25	14	Y50:V186	IFIF1	15053.2970	15053.4630	11.02
1369.88	1370.68	11	Y50:V186	IFIF1	15053.4140	15053.4630	3.18
1169.14	1169.84	13	W15:A151	IFIF1	15181.2470	15181.5480	19.82
1013.40	1014.04	15	W15:A151	IFIF1	15181.4750	15181.5480	4.80
804.80	805.26	19	135Q125	b	15267.6210	15267.4050	14.14

Table 5.1: Final output mass table: the output list generated by MasSPIKE resulting from the interpretation of Bovine Carbonic Anhydrase spectrum of Fig 5·1. All the assignments that match the given sequence fragments within an error of 20 ppm are listed. The columns indicate the start m/z and end m/z locations of isotopic distributions within the spectrum, charge state (Z), amino acid residues corresponding to the cleavage site, ion type, observed/estimated monoisotopic mass from the spectrum, theoretical monoisotopic mass for the fragment, and the error in ppm.

Chapter 6

Application of MasSPIKE in the Real World

The analysis of large, biologically derived molecules, such as proteins, oligosachcharides, DNA (Deoxyribonucleic Acid), and RNA (Ribonucleic Acid) poses significant analytical challenges for mass spectral interpretation. This chapter is aimed at demonstrating the capabilities of MasSPIKE for assisting in interpretation of complex, information rich spectra obtained from biologically interesting proteins such as Hemoglobin, Ras, and Transthyretin (TTR). MasSPIKE has also been utilized to characterize a home built qQq-FTMS instrument (O'Connor et al., 2006) by analyzing the experimental data originating from a variety of proteins.

6.1 Characterization of Hemoglobin variants by Mass Spectrometry

The term Hemoglobin (Hb) is formed by the combination of heme and globin, meaning that each subunit of hemoglobin is a globular protein (globelike proteins that are soluble in aqueous solutions) with an embedded heme group.(Campbell, 1999; Reece, 2005) Each heme group contains an iron atom, responsible for the binding of oxygen. Hemoglobin is present in the red blood cells in humans and other animals, primarily for the transportation of oxygen from the lungs to the rest of the body. The most common types of hemoglobin contains four such subunits, each with one heme group as shown in Fig 6·1.

A commonly prevalent genetic disorder in humans is the mutation of the genes responsible for coding for hemoglobin, resulting in a group of hereditary diseases called hemoglobinopathies. The dysfunction mechanism for hemaglobinopathies was one of the first human diseases to be understood down to the molecular level. However, not all

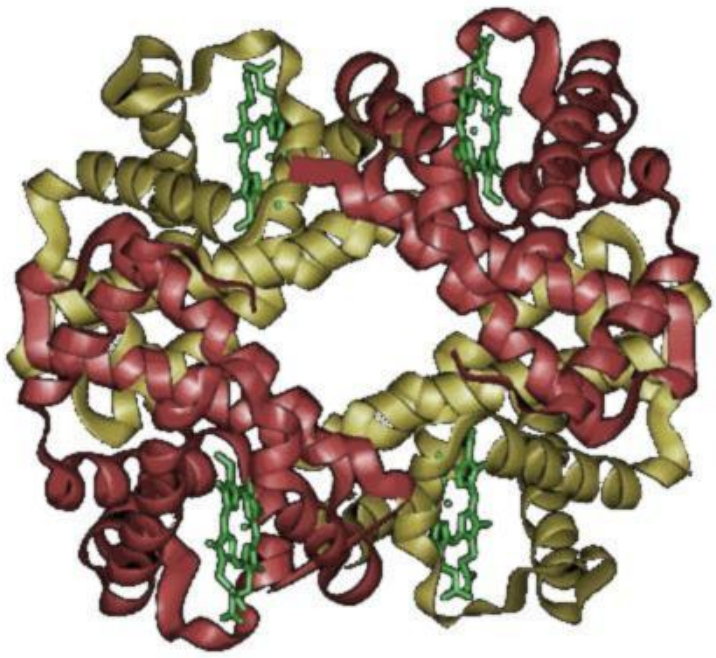


Figure 6·1: 3-D structure of human hemoglobin. The four subunits are shown in red and yellow, and the heme groups in green

such mutations manifest themselves as a disease, and are formally termed as hemoglobin variants.(Campbell, 1999; Reece, 2005) DNA analysis is commonly employed for clinical disease diagnosis. However, such an analysis alone may not be able to detect important Post-Translational Modifications (PTMs) at the molecular level, necessitating the use of protein sequence analysis using mass spectrometry.(McComb et al., 1998; Caruso et al., 2004)

As part of a Cardiovascular Proteomics Center collaboration, several hemoglobin variants were tested by mass spectrometry. Genetic analysis of a blood sample from a patient had previously found the alpha chain to be normal, without any mutations. Analysis of the same blood sample revealed heterozygotic mutation of hemoglobin beta chain codon 6 into GTG, indicating a mutation, commonly associated with sickle cell anemia, which means glutamic acid had been replaced by valine (Glu6→Val). Isoelectric focusing (IEF) gel electrophoresis analysis resulted in a band pattern abnormally shifted to a lower pI range

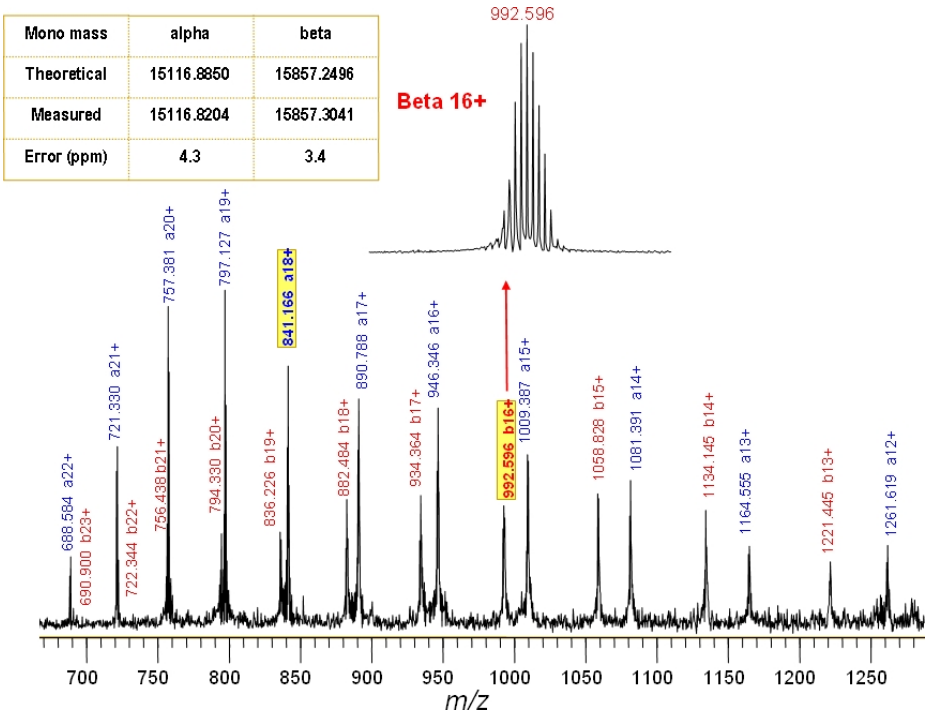


Figure 6·2: ESI spectrum of intact hemoglobin chains from a non diseased sample. Different charge states of alpha chains are marked as blue, while those for beta chains are shown in red.

compared to that of Hb S ($2\beta^s$+2α) and Hb A (2β+2α). The pI shifts were not explainable by the sickle mutation alone, indicating the possibility of additional modifications which went undetected by DNA analysis. This led to the current study of systematic investigation to explore the variations using mass spectrometry.(Huang et al., 2007; Huang et al., 2005)

The custom ESI qQq-FTICR-MS instrument with a nanoelectrospray ion source was used for all the experiments presented here.(O'Connor et al., 2006) First, the blood sample containing normal hemoglobin chains was analyzed using MasSPIKE, and the intact masses of alpha and beta chains were determined within 4 ppm as shown in Fig 6·2. An 18+ charge state of alpha chain at an m/z value of 841 was subjected to top-down analysis using collisionaly-activated-dissociation (CAD), and the resulting spectrum is shown in Fig 6·3. Automated analysis of the spectrum using MasSPIKE yielded an assignment of the monoisotopic mass list to C-terminal ions (called b ions), N-terminal ions

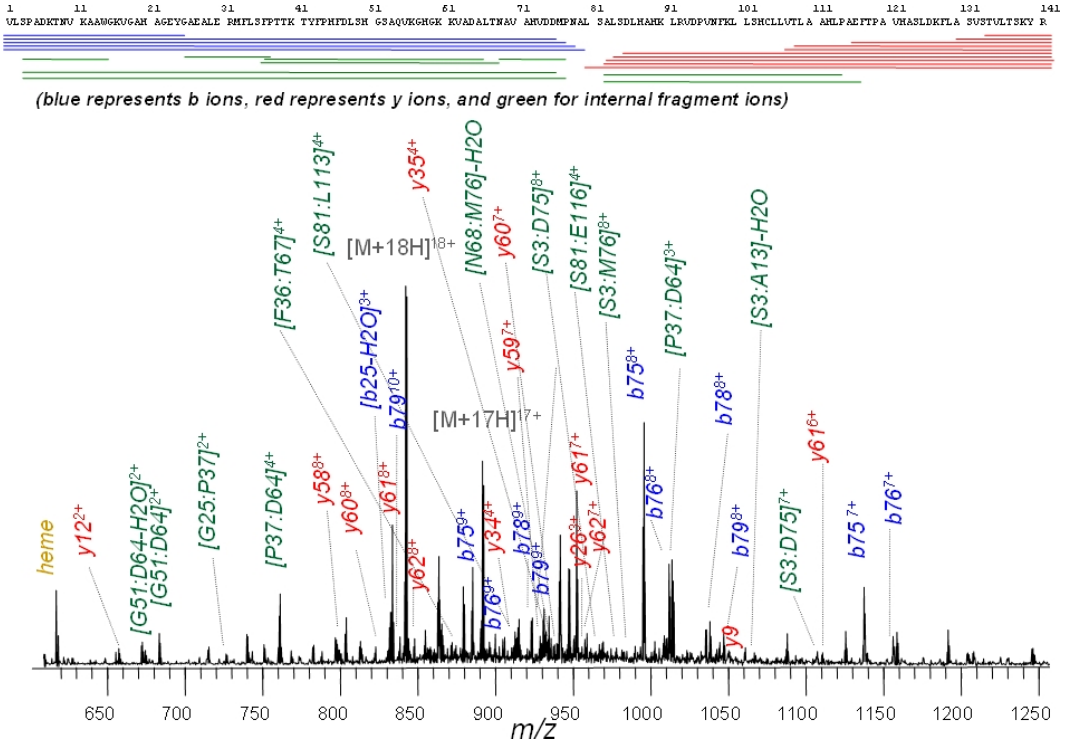


Figure 6·3: Q2 CAD spectrum of the Q1 isolated alpha chain 18+ at m/z 841

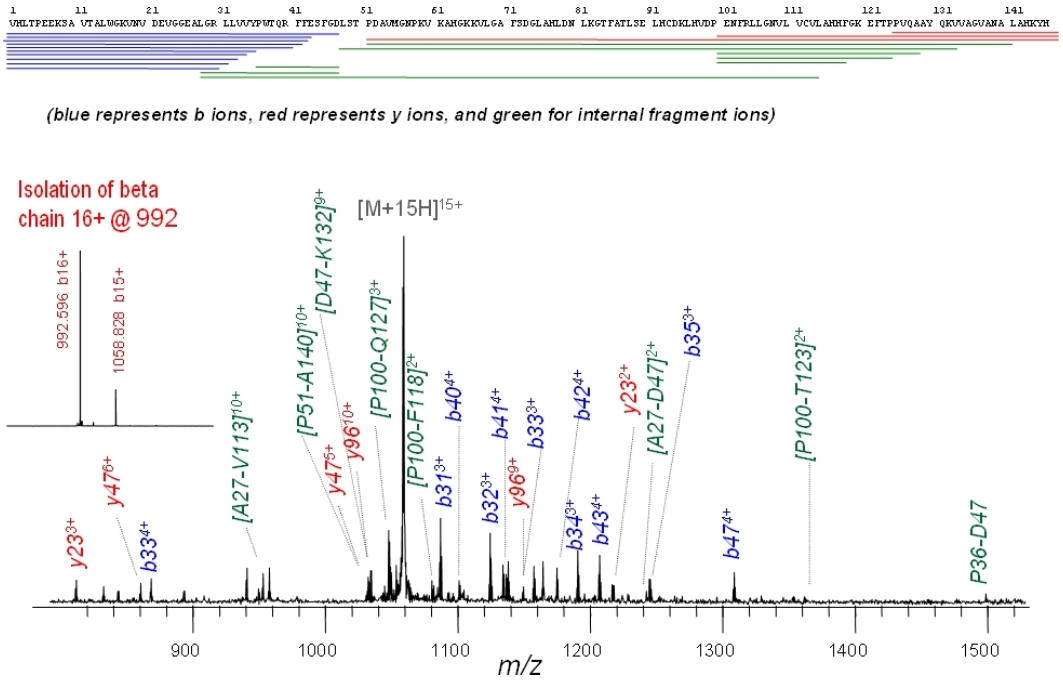


Figure 6·4: SORI-CAD spectrum of the Q1 isolated beta chain 16+ at m/z 992

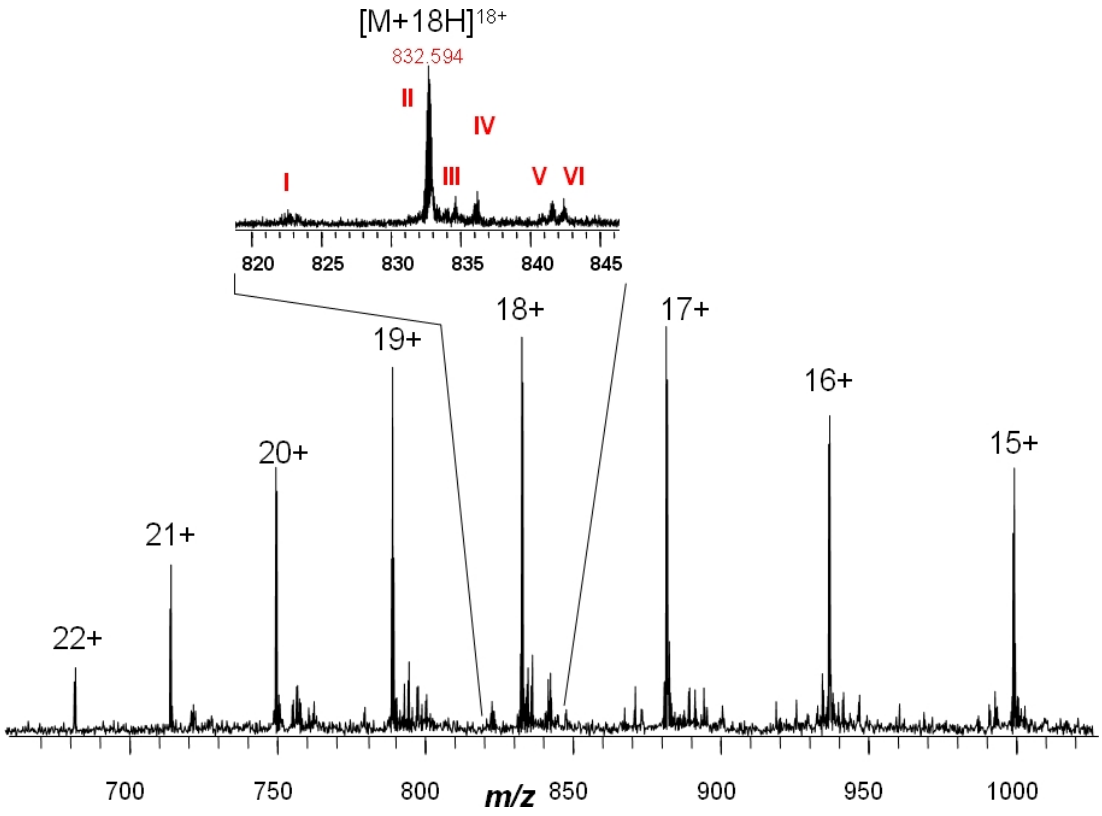


Figure 6·5: NanoESI-FT-MS of the patient Hb sample with labeled charge states. The inset, an expansion of the range m/z 820 to 845, shows the six species present in this sample all at a charge state of 18+. The monoisotopic mass value of the 18+ charge state of the major species (labeled II, at m/z 832.594) was evaluated to be 14960.7143, which matches that of the alpha chain minus the mass of an arginine residue (14960.7839).

(y ions),(Roepstorff and Fohlman, 1984) and the internal fragments, showing an extensive sequence coverage. C-terminal ions are marked in red, N-terminal ions are marked in blue, and the internal fragments are represented in green in Fig 6·3. A similar automated analysis of a top-down spectrum of normal beta chain of charge state 16+ at m/z location of 992 (15857.2496 Da) yielded results shown in Fig 6·4, with the monoisotopic masses assigned to the appropriate fragment ions.

MasSPIKE was used to determine the position of isotopic clusters, to assign the charge states to each isotopic distribution, and align the experimental isotopic distribution against

number	Protein	Theo. (mono)	Obs. (mono)	Error (ppm)
I	alpha chain b139	14779.7100	14779.5300	12.1
II	Alpha - R141	14960.7839	14960.7143	4.6
III	beta sickle	15827.2755	15827.2565	1.2
IV	beta	15857.2496	15857.1762	4.6
V	alpha	15116.8850	15116.8204	4.3
VI	alpha + H_2O	15134.8950	15134.9299	-2.3

Table 6.1: Mass determination of hemoglobin and variants/modifications

the theoretical isotopic distribution of an "average protein"(Senko et al., 1995b). The intact masses identified six different species present in the sample, as shown in Fig 6·5 and Table 6.1, which includes beta sickle mutation with an error of 1.2 ppm, confirming earlier results from genetic analysis. An experimental mass of 14960.7143 was detected for the most abundant species in the spectrum, as compared to the theoretical mass of 15116.8204 for the normal alpha chain, indicating a mass shift of -156.106. One of the possible explanations for this shift in observed mass is the loss of an ariginine residue from the alpha chain, with an error of 4.6 ppm. The top-down MS/MS analysis of the isolated ion with the 18+ charge state at m/z 832 (14960.7839 Da) indeed confirmed it to be an alpha chain derivative.

The hemoglobin sample was subsequently digested with Endoproteinase AspN, and the resulting digests were subjected to nano-ESI-FTMS (Fig 6·6) to confirm and localize the arginine deletion on the alpha chain. The AspN hemoglobin peptide mass mapping yielded full sequence coverage of the alpha chain, including the detection of the sickle mutated peptide β^s-D1, as well as two peptides (αˆD9 and αˆD7-9) that were consistent with an Arg truncation at the extreme C-terminus of the alpha chain (Fig 6·6). Because αˆD7-9 is a large peptide with two missed cleavages, multiple charge states (6+, 7+, and 8+) were observed for this peptide. The doubly charged ion at m/z 830 was isolated for MS/MS

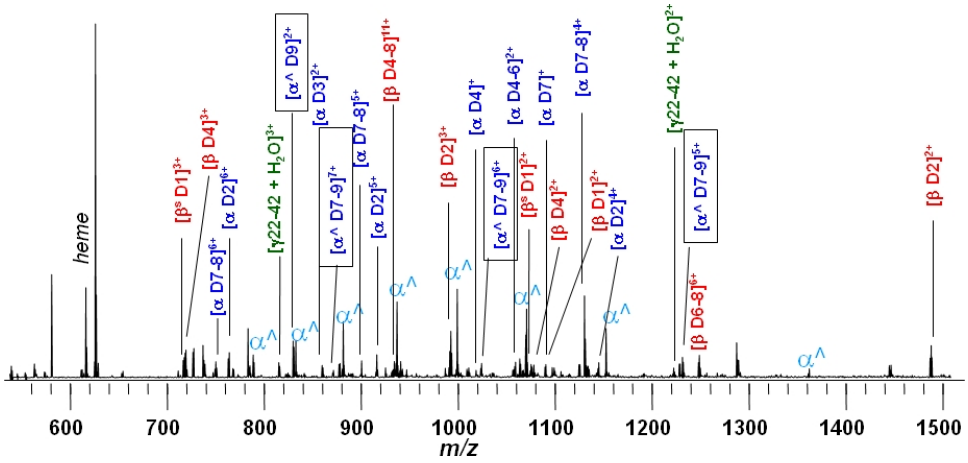


Figure 6·6: Accurate AspN peptide mass mapping by ESI-FT-MS. Peptide ions are labeled with their globin chain of origin (blue, alpha; red, beta; green, gamma), amino acid interval, charge state, and water adducts. Ions of the peptide β^s-D1 containing the beta sickle mutation were detected (labeled $[\beta^s\ \mathrm{D1}]^{2+}$ and $[\beta^s\ \mathrm{D1}]^{3+}$). High coverage of truncated alpha chain (96%), beta chain (97%), and beta sickle chain (97%) was obtained. Ions of the truncated alpha peptides D9 and D7-9 were detected in multiple charge states (boxed labels $[\alpha\hat{}\ \mathrm{D9}]^{2+}$, $[\alpha\hat{}\mathrm{D7\text{-}9}]^{5+}$, $[\alpha\hat{}\ \mathrm{D7\text{-}9}]^{6+}$, $[\alpha\hat{}\ \mathrm{D7\text{-}9}]^{7+}$). $\alpha\hat{}$designates Arg-141 truncated alpha chain.

sequencing, and SORI-CAD (Sustained Off Resonance Irradiation Collisionally Activated Dissociation)(Gauthier et al., 1991) was applied to fragment it generating tandem mass spectrum shown in Fig 6·7. A series of b and y ions were detected that matched those of the truncated alpha peptide D9 as shown in Fig 6·7. Moreover, LC-MS and MS/MS of this sample also detected and sequenced the same truncated alpha D9 peptide, which further confirmed the loss of the arginine from the alpha chain C-terminus.(Huang et al., 2005; Huang et al., 2007)

C-terminal Arginine has been found to play an important role for salt bridge formation in hemoglobin, and its removal can have important medical implications. Higher oxygen affinities have been reported for hemoglobin molecules missing the C-terminal arginine residue in the alpha chain, or the beta chain with C-terminal histidine deletions.(Kavanaugh et al., 1995; Bettati et al., 1997) Thus, in the sample under investigation, the removal of

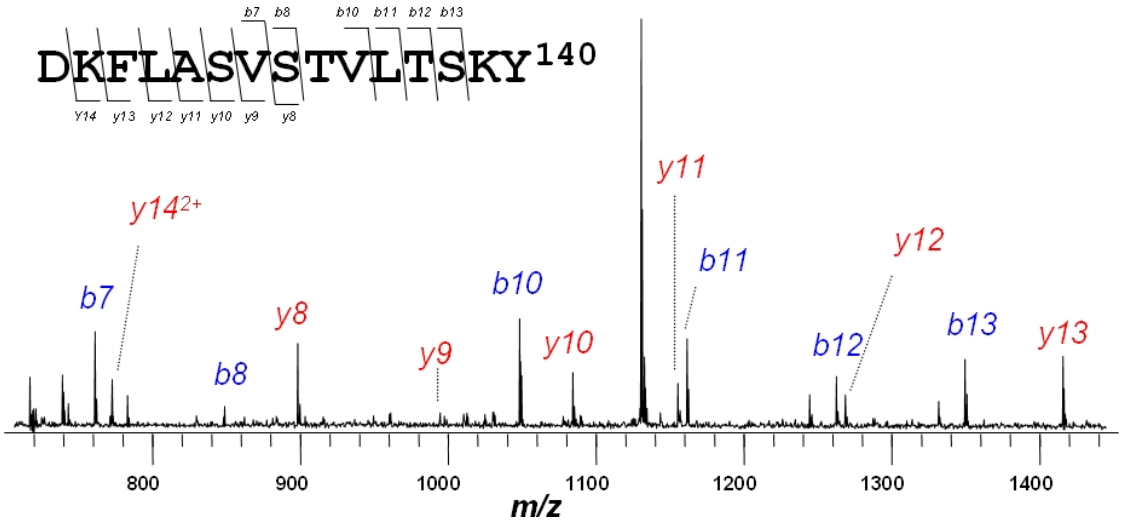


Figure 6·7: Tandem mass spectrum over the range m/z 700 to 1450 of the fragment ions after the α^-D9 peptide ion was subjected to SORI-CAD. Fragment ions are labeled with their b and y ion assignment . The reconstructed sequence of the truncated alpha chain D9 peptide containing the truncation of Arg-141 is shown inset above the spectrum and includes flags that designate the detected b and y ions.

C-terminal arginine would be expected to increase the molecule's oxygen affinity, and may have significant clinical implications for the patient. However, the possibility of this truncation having occurred *ex-vivo* due to proteolysis cannot be excluded.

Overall, MasSPIKE has been utilized for the identification of a sickle mutation of hemoglobin beta chain (Glu6 replaced by Val), successfully confirming the genetic analysis results. Automated interpretation results revealed six different variants of alpha and beta chain, and indicated an important molecular structural change resulting from the truncation of C-terminal arginine residue from alpha chain. The results have been verified using different approaches. The methodology can serve as a useful diagnostic tool for clinical applications.

6.2 Mapping Oxidative Post-Translational Modifications (PTMs) of human P21Ras using FTMS

The Ras proteins are small guanine nucleotide exchange proteins that play an important function of signal transduction for the regulation of a number of cellular processes such as cell growth, differentiation, and apoptosis.(Shields et al., 2000; Campbell et al., 1998; van der Schroeff et al., 1990) They cycle between the inactive GDP-bound state and the active GTP-bound state regulated by guanine nucleotide exchange factors and GTPase-activating proteins.(Boriack-Sjodin et al., 1998; Vetter and Wittinghofer, 2001; Scheffzek et al., 1997) These regulatory proteins bind to ras and regulate the exchange of GDP with GTP. The gene coding for Ras is one of the genes most commonly mutated in human tumors.(Shields et al., 2000; Campbell et al., 1998) In cancerous cells, the Ras protein has been found to be trapped in the "on" position and continues to stimulate cell growth. Thus, it is an important potential target for pharmaceutical therapeutic intervention in cancer.

Ras can be modified and activated by reactive nitrogen species (RNS), including nitric oxide ($\cdot NO$), nitrogen dioxide ($\cdot NO_2$), dinitrogen trioxide (N_2O_3), and peroxynitrite ($ONOO^{-\cdot}$), as well as reactive oxygen species (ROS), such as superoxide anion radical ($O_2^{-\cdot}$) and hydrogen peroxide (H_2O_2).(Stamler et al., 2001; Jaffrey et al., 2001; Eu et al., 2000; Sun et al., 2001; Klatt and Lamas, 2000; Adachi et al., 2004; Heo and Campbell, 2004; Mallis et al., 2001)

Peroxynitrite anion ($ONOO^{-\cdot}$), the product of the reaction between superoxide anion ($O_2^{-\cdot}$) and nitric oxide ($^{\cdot}NO$),(Radi et al., 1991; Koppenol et al., 1992) is a potent oxidant formed in endothelial cells which oxidizes a wide range of biological targets.(Radi et al., 1991; Beckman et al., 1990; Moreno and Pryor, 1992) Ras activity is potentially modulated by reversible as well as irreversible oxidative post-translational modifications. In particular, four surface cysteines (C118, C181, C184 and C186) with reactive thiol groups can regulate

H-ras activity in response to oxidative modifications.

Glutathione (GSH), a tripeptide γ-glutamyl cysteinyl glycine, is oxidized by ROS/RNS to glutathione disulfide, GSSG. In mammalian cells, GSH is the most abundant low molecular weight thiol, and thus it frequently binds to protein thiols to form mixed disulfides, a process termed S-glutathiolation. S-glutathiolated Ras (GSS-Ras) has been suggested to take part in cellular regulation.(Klatt and Lamas, 2000; Adachi et al., 2004) It has been generally accepted that S-oxidation, S-nitrosation, and S-glutathiolation of Ras is involved in cell signaling,(Kuster et al., 2005; Teng et al., 1999) but the relative importance by which these modifications directly regulate activity is not well understood. Moreover, there have been limited studies of structural characterization of post-translational modifications on full-length Ras protein. In general, PTM mapping for large proteins is challenging biologically, chemically, and technically due to the sub-stoichiometric level of modifications and their labile nature. This study was focused on structural characterization of oxidant-induced PTMs on p21ras in order to generate a complete PTM map of p21ras. The number of glutathiolated cysteines in p21ras treated with oxidized glutathione disulfide (GSSG) was determined by the intact mass difference between glutathiolated p21ras and the control sample, and the major site of S-glutathiolation was identified by top-down analysis.(Zhao et al., 2006) This structural information will aid in understanding the function of p21ras.

Unmodified Ras Fig 6·8A (left) shows the ESI spectrum of a sample of intact purified p21ras prepared in the presence of DTT (dithiothreitol). The resolved experimental isotopic distributions of several high abundance peaks were aligned against the model isotopic distribution (dots) using MasSPIKE for the evaluation of the monoisotopic molecular weight (MI), which was determined to be 21284.4032 Da, consistent with the theoretical mass of intact p21ras, 21284.5511 Da. This 7 ppm difference is within the expected error range of $\approx$5-10 ppm for externally calibrated FTMS data. When comparing the mass of intact p21ras protein prepared in the absence (Figure 6·8B), or in the presence of DTT (Figure 6·8A (left)), a 2 Da mass difference was detected, suggesting the presence of a

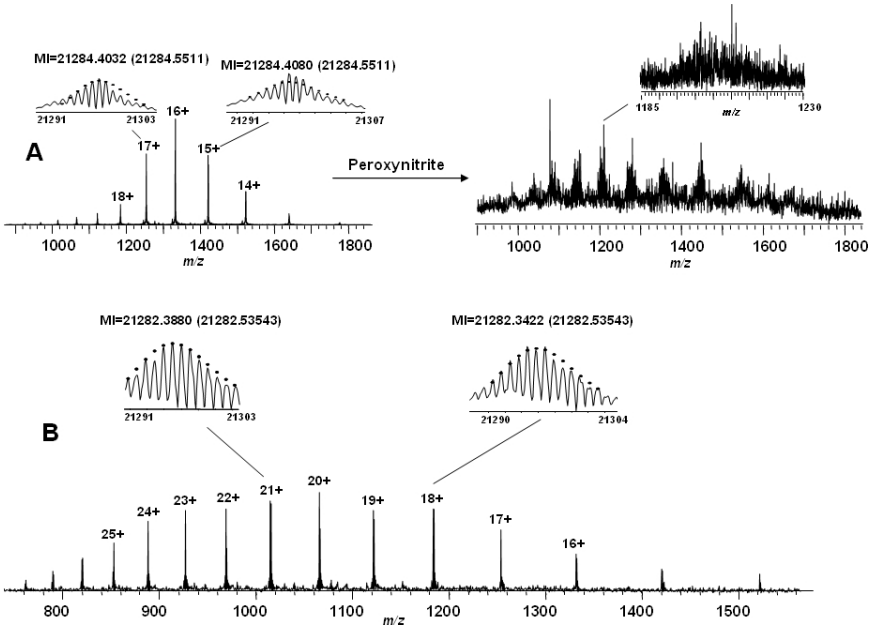


Figure 6·8: A. (left) ESI spectrum of purified unmodified p21ras treated with DTT. (right) ESI spectrum of purified p21ras treated with peroxynitrite. Inset shows m/z 1185 - 1230. B. ESI spectrum of purified unmodified p21ras that is not treated with DTT.

disulfide bond. Because the p21ras sample was purchased in buffer containing DTT, this disulfide bond probably re-formed after removal of the DTT by dialysis.

In order to localize this disulfide bond, these two samples were digested using trypsin. A peptide spanning amino acids 170-185 including Cys 181 and Cys 184 was identified in both samples which was 2 Da lighter in the sample without DTT than in the sample treated with DTT. Q2 CAD MS allowed localization of a disulfide bond to the two cysteines of the peptide.(Zhao et al., 2006)

Peroxynitrite (PN) treated p21ras The Ras sequence suggests that oxidative modifications could happen on many different sites such as methionines, cysteines, and tyrosines. The ESI spectrum of the intact purified p21ras treated with a 10-fold excess peroxynitrite is shown in Figure 6·8A (right). Due to the complex and heterogeneous oxidative modifications on p21ras, the spectrum yielded a signal to noise ratio <5 despite several steps of

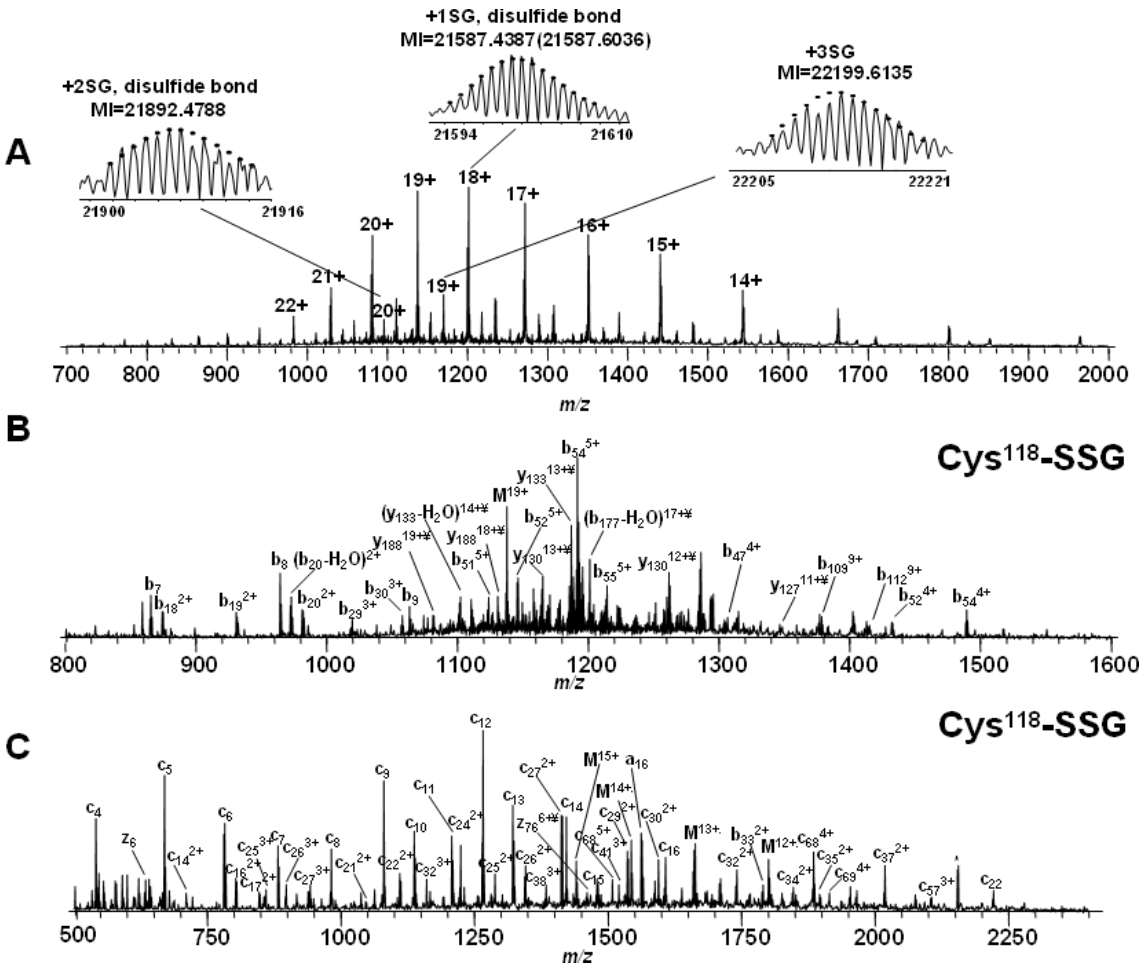


Figure 6·9: Top down spectra of p21ras. A. ESI spectrum of purified glutathiolated p21ras. B. CAD MS/MS spectrum and C. ECD MS/MS spectrum on isolated +19 singly glutathiolated p21ras. The peak labeled with $*$ indicates electronic noise.

sample cleaning. The isotopic distribution of individual charge states cannot be resolved. Top-down analysis is very difficult under such conditions. Thus, the bottom-up approach was employed for further analysis of this sample.(Zhao et al., 2006) Figure 6·10A plots the modifications detected in peroxynitrite-treated p21ras including irreversible methionine, tyrosine, and cysteine oxidations. Five methionines (M1, M67, M72, M111, M182) were oxidized and five tyrosines (Y4, Y40, Y96, Y137, Y157) were nitrated. Cys118, which resides in the "NKCD" loop, the GTP binding site, was oxidized. Cys118 sulfenic acid (SOH), sulfinic acid (SO_2H), sulfonic acid (SO_3H), and S-nitrosothiol (SNO) were all observed.

S-Glutathiolated p21ras Figure 6·9A shows the ESI spectrum of intact glutathiolated p21ras which shows at least three groups of charge state distributions. Compared to the ESI spectrum of unmodified p21ras in Figure 6·8B, the most intense charge state distribution, whose monoisotopic molecular weight (MI) was 21257.4387 Da (+18 charge state), indicated one glutathione addition to the p21ras and the other two indicated two and three glutathione additions, respectively. For the triply glutathiolated p21ras, the disulfide bond at Cys 181- Cys-184 was cleaved. For the bottom-up experiments on tryptic peptides, all the cysteines, including the terminal three cysteines were glutathiolated (Figure 6·10B), although it is possible that some of them were formed by exchange of disulfide bonds between GSSG and p21ras protein/tryptic peptides during sample digestion. Although with bottom-up data, it is impossible to determine which cysteine is the major site of glutathiolation on p21ras, this difficulty is solved with top-down analysis, in which the specific intact modified protein ions can be selectively isolated and fragmented. The major glutathiolated p21ras ion (Figure 6·9A) was isolated by Q1 and then directly fragmented in Q2 by CAD and in the ICR cell by Electron Capture Dissociation (ECD) and the resulting spectra are shown in Figure 6·9B and Figure 6·9C. Although only 64 of 188 inter-residue bonds were cleaved, the fragment ions provided enough information to localize the modification. The CAD spectrum in Figure 6·9B did not show any glutathiolation on the b112 fragment ion but b131 did include one glutathiolation indicating that a cysteine between amino acid 112 and 131 from the N-terminus was involved. Similarly, the ECD spectrum in Figure 6·9C, showed no glutathiolation at z38 but showed one at z79, indicating that the glutathiolation was on a cysteine between amino acid 38 and 76 from the C-terminus, thus identifying the site as C118. Therefore, these data identify C118 as the site of most abundant glutathiolation in GSSG-treated p21ras. Thirty-two cleavages (b6-9, b18-20, b29-31, b47, b51-55, b109, b112, b131, b137, b177, y31, y123, y126, y129, y131-133, y135-136, y179, y187) were obtained in CAD top-down analysis, 46 cleavages (c4-9, c11-27, c29-35, c37-38, c41, c46, c57, c68-69, c83-84, c87, z5-6, z11-12, z38, z76) were obtained in ECD top-down analysis, and 4 complementary pairs (b53 and y136, b54 and

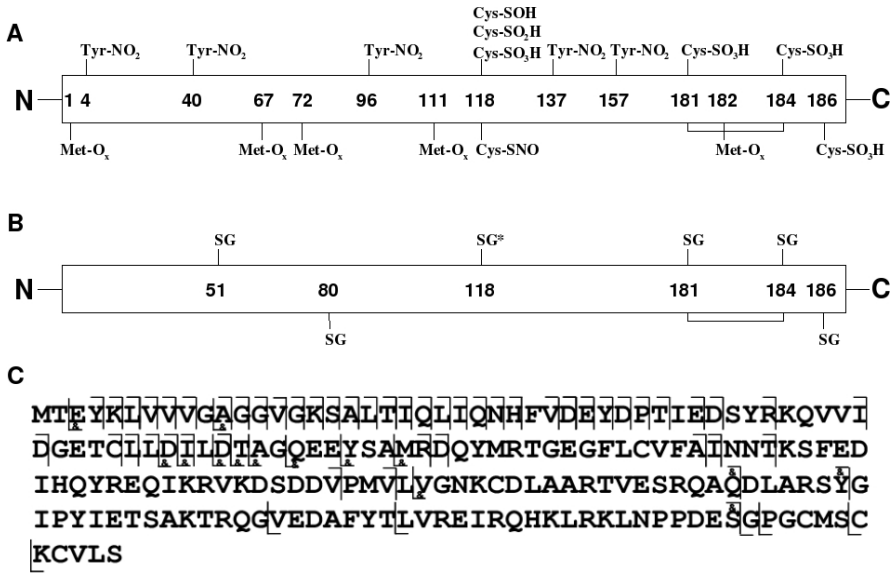


Figure 6·10: A. Map of oxidative post-translational modifications detected on peroxynitrite-treated p21ras. B: Detected modifications on glutathiolated- p21ras. The most abundant glutathiolation on C118 confirmed by top-down analysis is labeled with $*$. C: Top down map of the major singly glutathiolated p21ras. The cleavages labeled with & include C118 glutathiolation.

y135, c57 and y132, b177 and z12) were obtained, resulting in 100% sequence coverage on the singly glutathiolated p21ras. Also, the consecutive sequence tags such as c11-27 and c29-35 with $\approx$2 ppm mass accuracy is more than sufficient to unambiguously assign the protein identity.

In order to further analyze Cys118-SSG, the 103-123 peptide with glutathiolated C118 was isolated by Q1 from the digestion mixture and fragmented using low energy (15-25 eV) CAD and a ECD MS/MS experiments. The fragment map confirmed the Cys118-SSG modification.(Zhao et al., 2006) The detailed results of these studies are plotted in Figure 6·10.

In conclusion, MasSPIKE has been employed for the automated interpretation of spectra indicating complex modifications in peroxynitrite-treated p21ras. Interpretation results revealed many oxidative modifications and some low abundance modifications such

as Cys118-SNO, Cys118-SOH and Cys118-SO_2H. In addition, five oxidized methionines, five nitrated tyrosines, and at least two oxidized cysteines, including Cys-118 and one of the terminal cysteines, were identified. MasSPIKE confirmed the structure of the most abundant oxidative modification on Cys-118, Cys118-SO_3H, using experimental data from low energy CAD and ECD MS/MS experiments. From top-down analysis, Cys-118 is identified as the major glutathiolated cysteine on p21ras. Mass accuracy of the final monoisotopic mass list for peptides or fragment ions was ≈2 ppm with internal calibration, except for some very weak peaks (signal/noise < 5) for which accuracy suffers due to noise distortion of the peak shapes. This high mass accuracy helps for both protein and peptide identification.

6.3 Top-down analysis of Transthyretin using ESI FTMS

Transthyretin (TTR) is a 55 kDa homotetramer with a dimer of dimers configuration that is synthesized in the liver. Each monomer is a polypeptide chain consisting of 127 amino acid residues. Amino acid substitutions resulting from gene mutations in monomeric TTR are hypothesized to destabilize the tetramer and cause the TTR to form intermediates that self-associate into amyloid fibrils. A substitution of valine by methionine at position 30 (TTR Val30Met) is the mutation most commonly found in Familial Amyloid Polyneuropathy (FAP). A position 122 substitution of valine by isoleucine (TTR Val122Ile) is carried by 3.9% of the African-American population, and is the most common cause of Familial Amyloid Cardiomyopathy (FAC). Familial transthyretin amyloidosis (ATTR) is associated with the deposition of TTR variants as amyloid fibrils in various organs and tissues,(Lim et al., 2002) causing neurodegeneration and organ failure. Since TTR is primarily produced in the liver, current treatment of ATTR involves the replacement of a liver containing a mutant TTR gene with a normal gene in order to replace the mutant TTR in the body. Correct diagnosis is critical in early stages of the disease to avoid the complicated liver transplantation procedure.

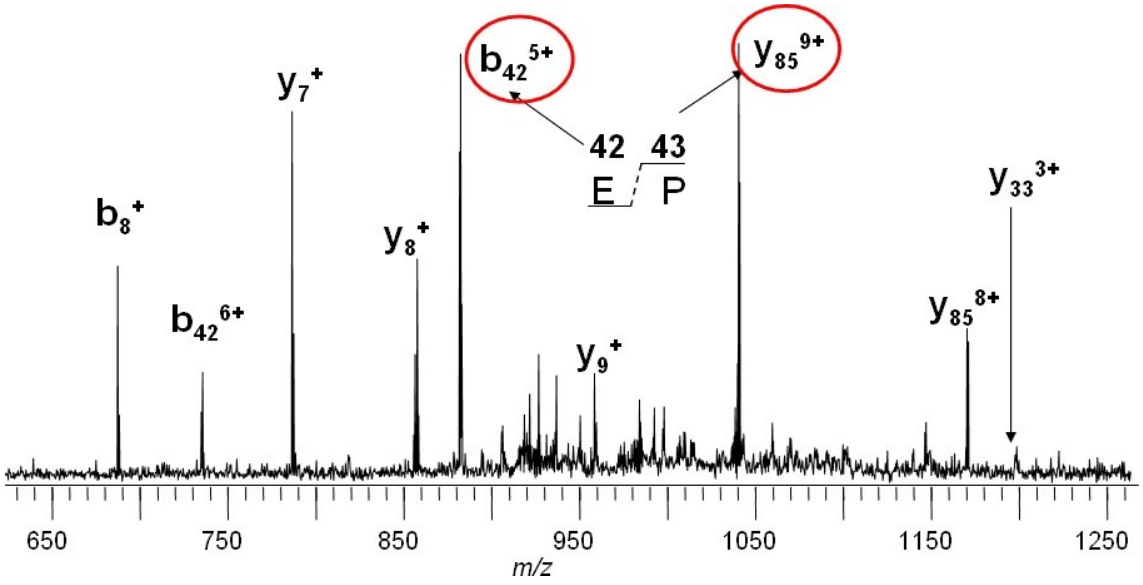


Figure 6·11: Fragment ion mass spectrum obtained from the Q2 CAD of the 15+ charge state of wild type TTR

The clinical significance of TTR inspired the current investigation of the protein sequencing using mass spectrometry.(Theberge et al., 2005) A custom hybrid ESI qQq-FTMS (O'Connor et al., 2006) was used to sequence the wild type TTR by pre-selecting the 15+ charge state of the intact protein and subjecting the accumulated ions to fragmentation using CAD in the quadrupole generating the spectrum (Fig 6·11). The spectrum was dominated by two complementary fragment ions, b42 and y85, providing the complete sequence coverage. The high abundance of these species was expected since it results from the cleavage of the relatively weak glutamic acid-proline bond between positions 42 and 43. The complete automated interpretation of the spectrum yielded the mass assignments to the fragments as shown in Fig 6·11.

Tandem mass spectrometry approach was also used to characterize a Val30Met mutation responsible for Familial Amyloid Polyneuropathy, a form of Paramyloidosis. The variant and wild type protein were pre-selected for Q2 CAD. The CAD spectrum thus obtained (Fig 6·12) exhibited the peaks corresponding to b42 fragment and b42+32 Da peaks not present in the wild type TTR CAD spectrum, which is consistent with Val30Met muta-

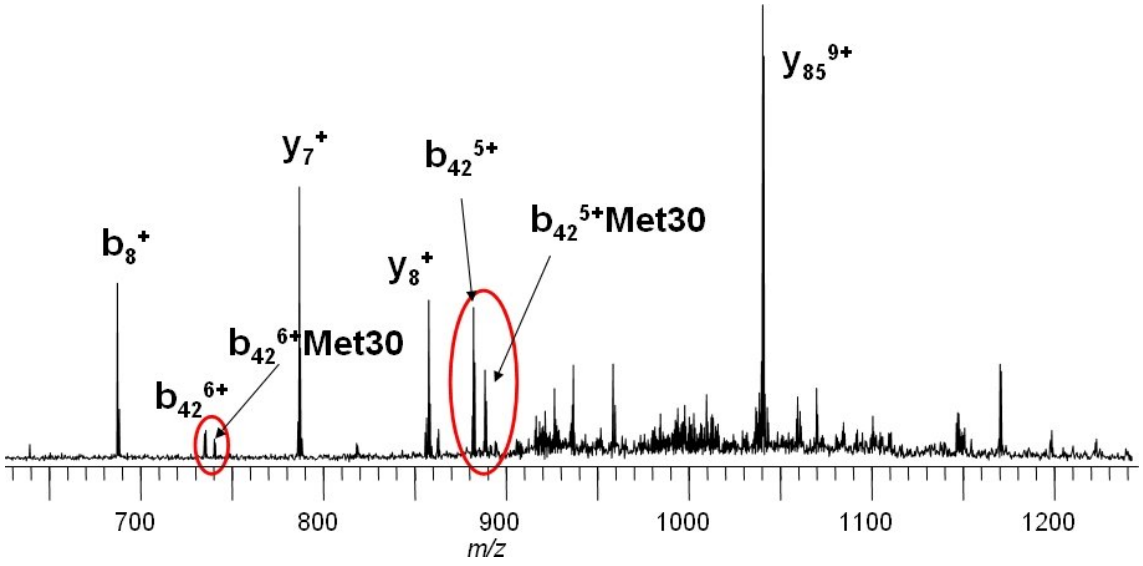


Figure 6·12: Q2 CAD spectrum of m/z 924 (charge state 15+) from both Val30Met variant and wild type TTR

tion. The isolation and SORI CAD fragmentation of the b42 fragment bearing the variant localized the variant position to 19-32 but did not yield data that specified the mutation site. Efforts are ongoing to explore further the mutation site of the b42 fragment using other complementary fragmentation mechanisms such as ECD.

A similar approach was successfully used to analyze a TTR sample containing Val122Ile mutation by isolating y85 produced from Q2 CAD of m/z 924 to undergo SORI CAD to yield sequence information localizing the mutation site to 122 (Fig 6·13). An immunoglobulin light chain involved in primary amyloidosis (AL) was also investigated. The protein isolated from the urine of a patient was analyzed using the same method that was applied to TTR variants. Q2 CAD followed by SORI CAD was necessary to generate fragmentation of this 23 kDa protein, probably due to the presence of intermolecular disulfide bonds (Cys23 to Cys88 and Cys134 to Cys194). Fragmentation was observed to occur almost exclusively the middle of the molecule.

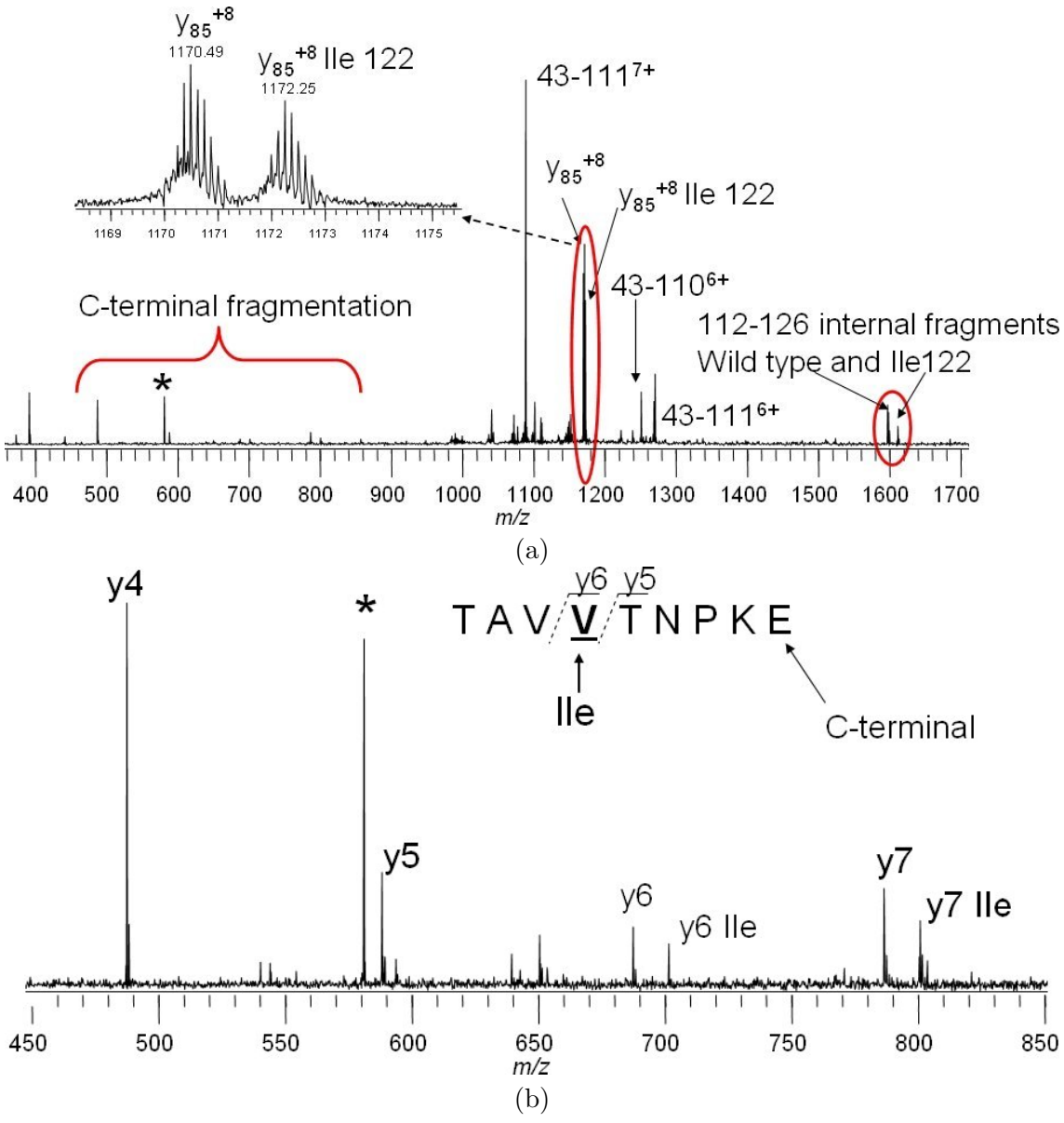


Figure 6·13: (a) Fragment ion mass spectrum obtained from the SORI CAD of the Y85 fragment generated by Q2 CAD of the 15+ charge state of Val122Ile TTR immunoprecipitated from patient serum. (b) Expanded mass scale of (a) to show C-terminal fragmentation. $*$ indicates electronic noise.

Wild type TTR, Val30Met, and Val122Ile variants have been investigated using top-down spectrometry and the resulting spectra were subjected to MasSPIKE for interpretation.(Theberge et al., 2005) MasSPIKE was able to localize the Val30Met mutation to positions 19-32, and identified the correct mutation site for Val122Ile variant. Automated analysis of immunoglobulin light chain again revealed the fragmentation pattern to dominate at the middle of the molecule. This is an ongoing project, with the efforts directed towards obtaining more sequence information for immunoglobulin light chain, and localizing the mutation site for Val30Met mutation. This constitutes a useful clinical application of top down mass spectrometric analysis.

6.4 Testing new mass spectrometry instrumentation

MasSPIKE has also been used for the analysis of a number of proteins that were used to characterize custom qQq-FTICR mass spectrometer (O'Connor et al., 2006), designed for the study of post-translationally modified proteins and for top-down analysis of biologically interesting protein samples.(Jebanathirajah et al., 2005) The performance of the instrument was evaluated for the analysis of a commonly occurring, but challenging post-translational modification, phosphorylation.

Top-down sequencing was performed on various proteins, including commercially available ones and biologically derived samples such as the human E2 ubiquitin conjugating enzyme, Ubch10.(Jebanathirajah et al., 2005) Sometimes, a cloning site can shift the open reading frame (ORF), resulting in changes in protein size/sequence. In the case of Ubch10, the 5' cloning site was not well characterized by nucleotide sequencing methods and restriction enzyme mapping. However, with the availability of the translated product, the protein Ubch10, it was possible to obtain information about the 5' cloning site. The accurate assignment of masses for MS and MS/MS spectra enabled verification of the sites used for the cloning and identified the 5' linker region N-terminal to the C-terminal His-tag that had been introduced to facilitate purification. A good sequence tag was obtained

for the human recombinant protein Ubch10, allowing the unambiguous identification of the protein. A total of 40 fragment ions were identified from the spectrum, covering the complete sequence of the protein. Twenty one unique ions (17 y-ions and 4 b-ions) were identified after taking into account the multiple charge states for the same fragment ion. Unfortunately, a number of ions could not be identified in this spectrum. Some of these ions could be a result of incomplete desolvation in the source or secondary fragmentation in Q2.

6.5 Conclusions

The use of MasSPIKE has been demonstrated under a variety of practical applications. Results from hemoglobin spectra included two major modifications on the protein. The beta chain was found to be modified as beta sickle, confirming the earlier genetic analysis results. Investigation of alpha chain led to the conclusion that the last residue arginine was missing from the sequence, resulting in truncated chain. This can have important clinical implications, participating in the pathophysiology of an individual.

Ras proteins have been found to be commonly responsible for human tumors. The functionality of Ras proteins is strongly influenced by their oxidation state. With the assistance of MasSPIKE, a map of oxidative post translational modifications of Ras proteins has been drawn, using a combination of bottom-up and top-down protein sequencing approaches. Many oxidative modifications including some low abundance modifications were identified. Five oxidized methionines, five nitrated tyrosines, and at least two oxidized cysteines, including Cys-118 and one of the terminal cysteines, were identified. Most abundant oxidative modification was found to be on Cys-118, as Cys118-SO_3H, which was confirmed by low energy CAD and ECD MS/MS experiments. Top-down analysis confirmed that Cys-118 is the major glutathiolated cysteine on p21ras.

Mutations in transthyretin (TTR) protein have been hypothesized to destabilize the structure of the protein and causing the TTR to form intermediates that self-associate into

amyloid fibrils, resulting in a disease called familial transthyretin amyloidosis. Interpretation of the top down mass spectrum of wild type TTR revealed major fragmentation species of b42 and y85 resulting from the cleavage of the bond between glutamic acid and proline at positions 42 and 43 respectively. TTR samples of patients with Val30Met and Val122Ile variants were analyzed to characterize the mutations. MS/MS analysis of y85 ion of Val122Ile mutant was able to identify and localize the site of mutation.

The characteristics of a home built qQq-FTICR instrument have been studied by means of analyzing a number of proteins, including commercially available and biologically relevant proteins such as human E2 ubiquitin conjugating enzyme, Ubch10. Interpretation of Ubch10 spectra was able to identify the cloning site, and a good sequence tag was obtained. A total of 40 fragment ions were identified, enabling the complete sequence coverage.

All the results were obtained using MasSPIKE in an automated manner, illustrating its utility for real world proteomics applications.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

A mass spectrum presents an information rich volume of data useful to researchers in multiple disciplines. In order to be meaningful and usable, this data needs to be "mined" or transformed into an appropriate form, often by reducing the dimensionality of data by several orders of magnitude. The type of transformation is highly dependent upon the application of interest, and the complexity of data in a mass spectrum makes the transformation process highly challenging. This dissertation has been directed towards designing reliable methods to automate these transformations, with the goal of eliminating the time consuming, inefficient, somewhat unreliable, tedious, and sometimes impossible task of manual interpretation.

A number of problems have been addressed that require sophisticated spectral analysis techniques. These include estimation of the number of ions generating an isotopic distribution in a mass spectrometry experiment, determination of high-precision isotope ratios from experimental isotopic distributions, development and comparison of charge state determination methods for high resolution mass spectra, and development and integration of algorithms for mass spectral interpretation into a suite of algorithms called MasSPIKE (Mass Spectrum Interpretation and Kernel Extraction). MasSPIKE has been applied to a variety of biologically interesting and challenging proteins.

Estimation of the number of ions generated in a mass spectrometry experiment is required to determine instrumentation parameters such as ionization efficiency, ion transfer

efficiency, collision induced dissociation efficiency, ion trapping efficiency, preamplifier detection limit, etc. An approach has been developed for estimating the total number of ions in a mass spectrometry experiment by observing the statistical variation of the resulting experimental isotopic distributions in the mass spectra. The maximum likelihood estimator in conjunction with the non-random parameter estimation method has been used to establish the mathematical relationship between the number of ions and the experimentally measured variation in the resulting EID. The theory has been first tested *in silico* for performance evaluation. Increasing the number of observations has been shown to greatly improve the estimate since the estimator gets more information. The estimator shows a positive bias and mean square error which depend upon the quantity being estimated, i.e., number of ions. The bias and mean square error drop substantially with increase in the number of observations. The estimator gives best performance in the limit of low number of ions. In general, determination of ionization efficiency, preamplifier detection limit, etc. requires working with a low number of ions so that the improved accuracy of the ML estimator under those conditions is advantageous. Experimental spectra were subjected to ion estimation analysis to characterize the sensitivity of the preamplifier used in the mass spectrometer. This approach is independent of the type of instrument used, and can be used for any kind of isotopically resolved mass spectrum. It is capable of showing a factor of 2 improvement over a previously developed method, depending on the number of observations used for the calculation.

Isotope variability due to natural processes provides important insights into a variety of complex natural phenomena ranging from the origins of a particular sample to the traces of biochemical reaction mechanisms. These measurements require very high-precision determination of isotope ratios of the particular element involved. A computational method has been developed and tested for estimating the elemental isotopic abundances from the observed isotopic distributions. Increasing the number of ions generating the isotopic distribution results in a much improved estimate. Higher molecular weights are particularly

advantageous for an accurate estimate since the higher number of carbon atoms and isotopic peaks observed provide a greater amount of information. However, higher molecular weights also need a higher number of ions in order for the experimental isotopic distribution to converge to its theoretical counterpart, which is a must for reliable results. This method is applicable for isotopically resolved spectra from any kind of instrument. It eliminates some of the limitations experienced by the conventional isotope ratio mass spectrometry by providing a greater flexibility about the kind of samples that may be utilized for the analysis, and makes the high resolution instruments such as Fourier transform mass spectrometer available for isotope ratio analysis. Any perturbations in the experimental isotopic distributions must be avoided. Such perturbations may arise due to various sources such as electronic noise, chemical noise, influences from overlapping isotopic distributions, intensity artifacts due to bias in quadrupole voltages, etc. For optimal results, experimental isotopic distribution must have minimal artifacts due to the subtle nature of the measurements.

Since mass spectrometers measure mass/charge ratio rather than mass, determination of charge state is crucial for the determination of mass. Charge state determination requires accurate estimation of the m/z difference between adjacent isotopic peaks. This poses a difficult problem under conditions of low signal-to-noise and when isotopic clusters are poorly resolved. A new method for charge-state determination using the Matched Filter approach has been developed and compared in detail with the established methods under various conditions. An automated comparison of the different methods was done under various conditions using 2800 simulated IDs. Overall, the following performance results were obtained: Patterson - 66.46 %, Fourier Transform method - 81.5%, Combo method - 85.57%, MF - 89.96%. Comparison was made under low and high charge state conditions separately, and the results indicated that Patterson and Fourier Transform methods give comparable performance under low charge states while Combo and MF performed much better. The Patterson method was observed to degrade in performance most rapidly as signal-to-noise decreased, followed by Fourier Transform, Combo and MF method in that

order. Analysis of the experimentally generated IDs revealed the following performance: Patterson - 49.6%, Fourier Transform - 51.2%, Combo - 60.72%, MF - 86.06%. Matched filter is capable of providing the information about the locations of experimental isotopic distributions, which is particularly useful for resolving overlapping isotopic distributions. This gives MF an additional advantage over the previous methods.

Automatic spectral interpretation has been one of the biggest bottlenecks in a mass spectrometry experiment, and, hence, limits the potential of mass spectrometry. This is because reliable and efficient automated analysis of spectra is a highly challenging computational problem. Algorithms have been developed to approach this problem, and integrated into a suite called MasSPIKE (Mass Spectrum Interpretation and Kernel Extraction). The algorithms are aimed towards reducing a high resolution mass spectrum into a monoisotopic peak list. MasSPIKE proceeds by modeling the noise across the mass spectrum, identifies the locations of isotopic distributions, determines the charge state for each of the isotopic clusters, resolves overlapping isotopic distributions, and aligns the experimental and theoretical distributions, and the final result is a monoisotopic mass list. Modeling of noise is done by analyzing the baseline of the spectrum across the whole m/z range. Isotopic distributions are located based upon the signal-to-noise ratio within the spectrum. The located isotopic distributions are subjected to charge state determination using the matched filter approach, and overlapping isotopic distributions are resolved from each other. The Matched Filter charge state determination routine worked correctly 91% of the time for test data as compared to 64% for standard combo routine. The observed and theoretical isotopic distributions are subsequently aligned against each other for the determination of monoisotopic mass, generating the final mass list. Alignment of the theoretical and experimental isotopic distributions with only 100 ions (and hence, high statistical variance) in the distribution gave 85% correct results as compared to 76% for the least-squares fitting method. If the protein/peptide sequence representing the spectrum is known, the calculated masses are matched for possible assignments. The suite has been applied and tested

against complex top-down spectra of Bovine Carbonic Anhydrase.

The utilization of MasSPIKE for the analysis of large, biologically derived molecules has been demonstrated under a variety of real world applications. Investigation of hemoglobin variants suggested two major modifications on the protein. The beta chain was found to be mutated as beta sickle, which was consistent with the earlier DNA analysis results. Analysis of the alpha chain revealed that the last residue arginine was missing from the sequence, resulting in a truncated chain. These modifications are likely to participate in the pathophysiology of an individual, suggesting important clinical implications.

Another important family of proteins, called Ras proteins, has been suggested to be responsible for human tumors. Oxidant-induced post-translational modifications of p21ras have been demonstrated to modulate its activity. MasSPIKE has been used to assist in drawing the map of oxidative post-translational modifications of Ras proteins, using a combination of bottom-up and top-down protein sequencing. Several oxidative modifications which included some low abundance modifications were identified. Analysis using MasSPIKE revealed major oxidative modification of C118, Cys118-SO_3H, which was confirmed by several tandem mass spectrometry experiments. Top-down analysis confirmed that Cys-118 is the major glutathiolated cysteine on p21ras.

Amino acid substitutions in Transthyretin (TTR), a 55 kDa homotetramer protein, are known to cause the deposition of the protein as amyloid fibrils, causing neurodegeneration and organ failure. TTR is also known to be associated with the amyloid diseases. Investigation of the top-down mass spectrum of wild type TTR using MasSPIKE revealed high abundance fragmentation species of b42 and y85 resulting from the cleavage of the bond between glutamic acid and proline at positions 42 and 43 respectively. MasSPIKE was also used to characterize the Val30Met and Val122Ile substitutions contained in patient TTR samples. MS/MS analysis of the y85 ion of Val122Ile mutant was able to identify and localize the site of mutation.

The characteristics of a custom qQq-FTICR instrument have been studied by the analysis of a number of proteins. Such proteins included both commercially available and biologically derived proteins such as human E2 ubiquitin conjugating enzyme, Ubch10. Interpretation of Ubch10 spectra using MasSPIKE was able to identify the cloning site, and a good sequence tag was obtained. A total of 40 fragment ions were identified, with complete sequence coverage.

7.2 Future Work

Although significant advances have been made recently in the spectral analysis aspects for mass spectrometry data, there are many more open problems waiting to be solved, providing ample opportunity for future research in this direction. Some ideas for further progress are listed below.

- **Integrating MasSPIKE with database searching**

  The monoisotopic mass list generated as output from MasSPIKE can be used as an input for database searching in order to reveal the identity of unknown proteins. This is especially useful for spectra resulting from peptide analysis, when the database search engines use the peptide mass information to assign the protein identity by means of a process known as peptide mass fingerprinting. If the spectrum has been internally calibrated, the error tolerance input given to the search engine should be very low. This will help eliminate false positive results. Robust denovo sequencing methods can be employed to construct the protein sequence from the masses generated from top-down sequencing analysis.

- **Optimizing MasSPIKE for use with online Liquid Chromatography/Mass Spectrometry (LC/MS) experiments**

  MasSPIKE has been used extensively for offline data analysis for top-down protein sequencing experiments. Analysis of peptide data is much simpler due to their low molecular weight. Hence, certain modules of MasSPIKE can be simplified for suit-

ability towards peptide data, and can be optimized for speed so that it can be used online in conjunction with LC/MS experiments.

- **Determination of elemental compositions using experimental isotopic distributions**

  An experimental isotopic distribution is a function of the elemental isotopic abundances and number of atoms of each type present in the molecule. Assuming that elemental isotopic abundances for each element are known, an observed isotopic distribution can be used to estimate the elemental composition of the molecule under consideration. This can serve as particularly useful information to filter the possible candidates in order to assign the identity of the protein. Experimental isotopic distributions should be free of any sort of distortions for the optimal performance of such experiments.

- **Analysis of Cramer-Rao bound for biased estimator**

  The estimator developed for ion number estimation as a part of this dissertation has been found to be biased, and the bias depends upon the quantity being estimated, i.e., the number of ions. However, the bias and mean square error in the estimator drop substantially with the increase in the number of observations. It has also been established that an unbiased, efficient estimator does not exist for this problem. It would be of analytical interest to evaluate the Cramer-Rao bound for the biased estimator. If an estimator exists that meets the Cramer-Rao bound, a comparison of that estimator against the one established in this work would be useful and may provide insights into achieving improved estimates.

- **Establishing the elemental composition of a model glycan**

  The field of glycomics has been mostly unexplored in terms of data processing methods. One such useful method would be to develop an average glycan model in order to do *in silico* studies for sugars, i.e., given the molecular weight for a glycan, researchers are interested to know the "average" composition of a glycan. This will

provide groundwork for a number of sugar-related data analysis studies. For example, it can be used to construct the theoretical isotopic distribution for a glycan, which can be compared against the experimentally observed distribution for analytical purposes.

References

Adachi, T., Pimentel, D. R., Heibeck, T., Hou, X. Y., Lee, Y. J., Jiang, B. B., Ido, Y., and Cohen, R. A. (2004). S-glutathiolation of Ras mediates redox-sensitive signaling by angiotensin II in vascular smooth muscle cells. *The Journal of Biological Chemistry*, 279:29857–29862.

Aebersold, R. (2003). A mass spectrometric journey into protein and proteome research. *Journal of the American Society for Mass Spectrometry*, 14:685–695.

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

Aizikov, K. and O'Connor, P. B. (2006). Use of the filter diagonalization method in the study of space charge related frequency modulation in FTMS. *Journal of the American Society for Mass Spectrometry*, 17:836–843.

Amster, I. J. (1996). Fourier transform mass spectrometry. *Journal of Mass Spectrometry*, 31:1325–1337.

Bakhtiar, R., Hofstadler, S. A., and Smith, R. D. (1993). Fourier-transform electrospray instrumentation for tandem high-resolution mass-spectrometry of large molecules. *Journal of the American Society for Mass Spectrometry*, 4:557–565.

Beavis, R. C. (1993). Chemical mass of carbon in proteins. *Analytical Chemistry*, 65:496 – 497.

Beckman, J. S., Beckman, T. W., Chen, J., Marshall, P. A., and Freeman, B. A. (1990). Apparent hydroxyl radical production by peroxynitrite - implications for endothelial injury from nitric-oxide and superoxide. *Proceedings of the National Academy of Sciences of the United States of America*, 87:1620–1624.

Berg, J. M., Tymoczko, J. L., Stryer, L., and Clarke, N. D. (2002). *Biochemistry - Fifth Edition*. W. H. Freeman and Company, New York.

Bettati, S., Kwiatkowski, L. D., Kavanaugh, J. S., Mozzarelli, A., Arnone, A., Rossi, G. L., and Noble, R. W. (1997). Structure and oxygen affinity of crystalline des-His-146beta human hemoglobin in the T state. *The Journal of Biological Chemistry*, 272(52):33077–33084.

Beu, S., Blakney, G., Quinn, J., Hendrickson, C., and Marshall, A. (2004). Broadband phase correction of ft-icr mass spectra via simultaneous excitation and detection. *Analytical Chemistry*, 76(19):5756–5761.

Beu, S. C., Senko, M. W., Quinn, J. P., Wampler, F. M., and McLafferty, F. W. (1993). Fourier-transform electrospray instrumentation for tandem high resolution mass spectrometry of large molecules. *Journal of the American Society for Mass Spectrometry*, 4:557–565.

Biemann, K. (1995). The coming of age of mass-spectrometry in peptide and protein chemistry. *Protein Science*, 4:1920–27.

Bloch, F. and Rabi, I. I. (1945). Atoms in variable magnetic fields. *Reviews of Modern Physics*, 17(2-3):237–244.

Boriack-Sjodin, P. A., Margarit, S. M., Bar-Sagi, D., and Kuriyan, J. (1998). The structural basis of the activation of Ras by Sos. *Nature*, 394:337–343.

Brenna, J. T. (1997). Use of stable isotopes to study fatty acid and lipoprotein metabolism in man. *Prostaglandins, Leukotrienes, and Essential Fatty Acids*, 57:467–472.

Brenna, J. T., Corso, T. N., Tobias, H. J., and Caimi, R. J. (1997). High-precision continuous-flow isotope ratio mass spectrometry. *Mass Spectrometry Reviews*, 16:227–258.

Bricout, J. and Koziet, J. (1973). Control of the authenticity of orange juices by isotopic analysis. *Journal of Agricultural and Food Chemistry*, 35:758–760.

Brown, L. S. and Gabrielse, G. (1986). Geonium theory: Physics of a single electron or ion in a penning trap. *Reviews of Modern Physics*, 58(1):233–311.

Budnik, B. A., Haselmann, K. F., and Zubarev, R. A. (2001). Electron detachment dissociation of peptide di-anions: An electron-hole recombination phenomenon. *Chemical Physics Letters*, 342:299–302.

Calvin, M. and Benson, A. A. (1948). The path of carbon in photosynthesis. *Science*, 107(7):476–480.

Campbell, M. K. (1999). *Biochemistry (Third Edition)*. Harcourt College Publishers.

Campbell, S. L., Khosravi-Far, R., Rossman, K. L., Clark, G. J., and Der, C. J. (1998). Increasing complexity of Ras signaling. *Oncogene*, 17:1395–1413.

Caruso, D., Crestani, M., DaRiva, L., Mitro, N., Giavarini, F., Mozzi, R., and C, C. F. (2004). Mass spectrometry and DNA sequencing are complementary techniques for characterizing hemoglobin variants: the example of hemoglobin j-oxford. *Haematologica*, 89(5):608–609.

Charlebois, J. P., Patrie, S. M., and Kelleher, N. L. (2003). Electron capture dissociation and 13c,15n depletion for deuterium localization in intact proteins after solution-phase exchange. *Analytical Chemistry*, 72:3263–3266.

Chen, L., Sze, S., and Yang, H. (2006). Automated intensity descent algorithm for interpretation of complex high-resolution mass spectra. *Analytical Chemistry*, 78(14):5006–5018.

Comisarow, M. B. and Marshall, A. G. (1974). Fourier transform ion cyclotron resonance spectroscopy. *Chemical Physics Letters*, 25:282–283.

Comisarow, M. B. and Marshall, A. G. (1976). Theory of fourier transform ion cyclotron resonance mass spectroscopy. i. fundamental equations and low-pressure line shape. *The Journal of Chemical Physics*, 64:110–119.

Cooley, J. W. and Tukey, J. W. (1965). An algorithm for the machine calculation of complex fourier series. *Mathematics of Computation*, 19:297–301.

Coon, J. J., Syka, J. E. P., Schwartz, J. C., Shabanowitz, J., and Hunt, D. F. (2004). Anion dependence in the partitioning between proton and electron transfer in ion/ion reactions. *International Journal of Mass Spectrometry*, 236:33–42.

Craig, H. (1957). Isotopic standards for carbon and oxygen and correction factors for mass-spectrometric analysis of carbon dioxide. *Geochimica et Cosmochimica Acta*, 12(1-2):133–149.

Dawson, T. E. and Ehleringer, J. R. (1993). Gender-specific physiology, carbon isotope discrimination, and habitat distribution in boxelder, acer negundo. *Ecology*, 74:798–815.

de Godoy, L. M., Olsen, J. V., de Souza, G. A., Li, G., Mortensen, P., and Mann, M. (2006). Status of complete proteome analysis by mass spectrometry: Silac labeled yeast as a model system. *Genome Biology*, 7:R50.

Demirev, P. A. and Fenselau, C. (2002). Determination of isotope-enrichment ratios in proteins by high-resolution fourier transform ion cyclotron resonance mass spectrometry. *European Journal of Mass Spectrometry*, 2(8):163–167.

DeNiro, M. J. and Epstein, S. (1978). Influence of diet on the distribution of carbon isotopes in animals. *Geochimica et Cosmochimica Acta*, 42(5):495–506.

Duda, R. O., Hart, P. E., and Stork, D. H. (2001). *Pattern Classification*. Wiley Interscience, New York.

Dunbar, J. (1982). A study of the factors affecting the 180/160 ratio of the water of wine. *European Food Research and Technology*, 174:355–359.

Engen, J. R. and Smith, D. L. (2000). Investigating the higher order structure of proteins: Hydrogen exchange, proteolytic fragmentation & mass spectrometry. *Methods in Molecular Biology*, 146:95–112.

Eu, J. P., Sun, J. H., Xu, L., Stamler, J. S., and Meissner, G. (2000). The skeletal muscle calcium release channel: Coupled O-2 sensor and NO signaling functions. *Cell*, 102:499–509.

Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., and Whitehouse, C. M. (1989). Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246:64–71.

Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., and Whitehouse, C. M. (1990). Electrospray ionization-principles and practice. *Mass Spectrometry Reviews*, 9:37–70.

Ferrige, A. G., Seddon, M. J., Jarvis, S., Skilling, J., and Aplin, R. (1991). Maximum entropy deconvolution in electrospray mass spectrometry. *Rapid Communications in Mass Spectrometry*, 5:374–377.

Francl, T. J., Sherman, M. G., Hunter, R. L., Locke, M. J., Bowers, W. D., and McIver, R. T. (1983). Experimental determination of the effects of space charge on ion cyclotron resonance frequencies. *International Journal of Mass Spectrometry and Ion Processes*, 54:189–199.

Gabrielse, G., Khabbaz, A., Hall, D. S., Heimann, C., Kalinowsky, H., and Jhe, W. (1999). Precision mass spectroscopy of the antiproton and proton using simultaneously trapped particles. *Physical Review Letters*, 82(16):3198–3201.

Gauthier, J. W., Trautman, T. R., and Jacobson, D. B. (1991). Sustained off-resonance irradiation for collision-activated dissociation involving Fourier transform mass spectrometry. collision-activated dissociation technique that emulates infrared multiphoton dissociation. *Analytica Chimica Acta*, 246:211–225.

Gross, M. L. and Rempel, D. L. (1984). Fourier transform mass spectrometry. *Science*, 226(4672):261–268.

Hayes, J. M. (1983). Practice and principles of isotopic measurements in organic geochemistry. *Organic Geochemistry of Contemporaneous and Ancient Sediments*.

Haykin, S. (1994). *Communications Systems*. John Wiley and Sons, Inc., Singapore.

He, F., Emmett, M. R., Hakansson, K., Hendrickson, C. L., and Marshall, A. G. (2004). Theoretical and experimental prospects for protein identification based solely on accurate mass measurement. *Journal of Proteome Research*, 3(1):61–67.

Henry, K. D. and McLafferty, F. W. (1990). Electrospray ionization with fourier-transform mass spectrometry. charge state assignment from resolved isotopic peaks. *Organic Mass Spectrometry*, 25:490–492.

Henry, K. D., Quinn, J. P., and McLafferty, F. W. (1991). High-resolution electrospray mass spectra of large molecules. *Journal of the American Chemical Society*, 113:5447–5449.

Henry, K. D., Williams, E. R., Wang, B. H., McLafferty, F. W., Shabanowitz, J., and Hunt, D. F. (1989). Fourier-Transform Mass Spectrometry of Large Molecules by Electrospray Ionization. *Proceedings of the National Academy of Sciences of the United States of America*, 86(23):9075–9078.

Heo, J. Y. and Campbell, S. L. (2004). Mechanism of p21(Ras) S-nitrosylation and kinetics of nitric oxide-mediated guanine nucleotide exchange. *Biochemistry*, 43:2314–2322.

Hobson, K. A., Sease, J., Merrick, R. L., and Paitt, J. F. (1997). Investigating trophic relationships of pinnipeds in alaska and washington using stable isotope ratios of nitrogen and carbon. *Marine Mammal Science*, 13:114–132.

Hoekstra, J., Aker, J. H. v. d., Kneepkens, C. M., Stellaard, F., Geypens, B., and Ghoos, Y. F. (1996). Evaluation of 13co2 breath tests for the detection of fructose malabsorption. *The Journal of Laboratory and Clinical Medicine*, 127:303–309.

Horn, D. M., Zubarev, R. A., and McLafferty, F. W. (2000). Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *Journal of the American Society for Mass Spectrometry*, 11:320–332.

Huang, H., Budnik, B. A., Perlman, D. H., Kaur, P., OConnor, P., McComb, M. E., Costello, C. E., Eung, S. H., Luo, H.-Y., Skelton, T. P., Steinberg, M. H., and Chui, D. H. K. (2007). Identification of human hemoglobin alpha chain truncation by mass spectrometry. manuscript in preparation.

Huang, H., McComb, M. E., Perlman, D. H., Budnik, B. A., Kaur, P., Skelton, T. P., Chui, D. H. K., O'Connor, P. B., and Costello, C. E. (2005). Proteomics approach for identification of hemoglobin variants and post-translational modifications. In *Proceedings of 53rd American Society of Mass Spectrometry conference on Mass Spectrometry*.

Huang, J., Tiedemann, P. W., Land, D. P., McIver, R. T., and Hemminger, J. C. (1994). Dynamics of ion coupling in an FTMS ion trap and resulting effects on mass spectra, including isotope ratios. *International Journal of Mass Spectrometry*, 134(1):11–21.

Jaffrey, S. R., Erdjument-Bromage, H., Ferris, C. D., Tempst, P., and Snyder, S. H. (2001). Protein S-nitrosylation: a physiological signal for neuronal nitric oxide. *Nature Cell Biology*, 3:193–197.

Jebanathirajah, J. A., Pittman, J. L., Thomson, B. A., Budnik, B. A., Kaur, P., Rape, M., Kirschner, M., Costello, C. E., and O'Connor, P. B. (2005). Biological applications of a novel qQq-FTICR mass spectrometer: Characterization

of post-translational modifications and top-down sequencing. *Journal of the American Society for Mass Spectrometry*, 16(12):1985–1999.

Jeener, J., Meier, B. H., Bachmann, P., and Ernst, R. R. (1979). Investigation of exchange processes by two-dimensional nmr spectroscopy. *The Journal of Chemical Physics*, 71(11):4546–4553.

Jennings, M. and Matthews, D. (2005). Determination of complex isotopomer patterns in isotopically labeled compounds by mass spectrometry. *Analytical Chemistry*, 77(19):6435–6444.

Jorgensen, T., Gardsvoll, H., Ploug, M., and Roepstorff, P. (2005). Intramolecular migration of amide hydrogens in protonated peptides upon collisional activation. *Journal of the American Chemical Society*, 127(8):2785–2793.

Kaiser, N. K., Anderson, G. A., and Bruce, J. E. (2005). Improved mass accuracy for tandem mass spectrometry. *Journal of the American Society for Mass Spectrometry*, 16(4):463–70.

Karas, M., Bachmann, D., Bahr, U., and Hillenkamp, F. (1987). Matrix-assisted ultraviolet laser desorption of non-volatile compounds. *International Journal of Mass Spectrometry and Ion Processes*, 78:53–68.

Karas, M., Bachmann, D., and Hillenkamp, F. (1985). Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules. *Analytical Chemistry*, 57:2935–2939.

Karas, M. and Hillenkamp, F. (1988). Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical Chemistry*, 60:2299–2301.

Kaur, P. and O'Connor, P. B. (2004). Use of statistical methods for estimation of total number of charges in a mass spectrometry experiment. *Analytical Chemistry*, 76:2756–2762.

Kaur, P. and O'Connor, P. B. (2006a). Algorithms for automatic interpretation of high resolution mass spectra. *Journal of the American Society for Mass Spectrometry*, 17:459–468.

Kaur, P. and O'Connor, P. B. (2006b). Comparison of charge state determination methods for high resolution mass spectra. In *IEEE International Conference on Granular Computing*, pages 550–553.

Kaur, P. and O'Connor, P. B. (2007). Determination of high-precision isotope ratios from experimental isotopic distributions. Analytical Chemistry. In Press.

Kavanaugh, J. S., Chafin, D. R., Arnone, A., Mozzarelli, A., Rivetti, C., Rossi, G. L., Kwiatkowski, L. D., and Noble, R. W. (1995). Structure and oxygen affinity of crystalline of DesArg141a human hemoglobin A in the T state. *Journal of Molecular Biology*, 248:136–150.

Kelleher, N., Lin, H., Valaskovic, G., Aaserud, D., Fridriksson, E., and McLafferty, F. (1999). Top down versus bottom up protein characterization by tandem high-resolution mass spectrometry. *Journal of the American Chemical Society*, 121(4):806–812.

Keller, B. O. and Li, L. (2001). Detection of 25,000 molecules of substance p by maldi-tof mass spectrometry and investigations into the fundamental limits of detection in maldi. *Journal of the American Society for Mass Spectrometry*, 12:1055–1063.

Klatt, P. and Lamas, S. (2000). Regulation of protein function by S-glutathiolation in response to oxidative and nitrosative stress. *European Journal of Biochemistry*, 267:4928–4944.

Koppenol, W. H., Moreno, J. J., Pryor, W. A., Ischiropoulos, H., and Beckman, J. S. (1992). Peroxynitrite, a cloaked oxidant formed by nitric-oxide and superoxide. *Chemical Research in Toxicology*, 5:834–842.

Kuster, G. M., Pimentel, D. R., Adachi, T., Ido, Y., Brenner, D. A., Cohen, R. A., Liao, R., Siwik, D. A., and Colucci, W. S. (2005). Alpha-adrenergic receptor-stimulated hypertrophy in adult rat ventricular myocytes is mediated via thioredoxin-1-sensitive oxidative modification of thiols on Ras. *Circulation*, 111:1192–1198.

Lawrence, E. O. and Cooksey, D. (1936). On the apparatus for the multiple acceleration of light ions to high speeds. *Physical Review*, 50:1131–1140.

Ledford, E. B., Rempel, D. L., and Gross, M. L. (1984). Space charge effects in fourier transform mass spectrometry. mass calibration. *Analytical Chemistry*, 56:2744–2748.

Lim, A., Prokaeva, T., McComb, M., O'Connor, P., Theberge, R., Connors, L., Skinner, M., and Costello, C. (2002). Characterization of transthyretin variants in familial transthyretin amyloidosis by mass spectrometric peptide mapping and DNA sequence analysis. *Analytical Chemistry*, 74(4):741–751.

Limbach, P. A., Grosshans, P. B., and Marshall, A. G. (1993). Experimental determination of the number of trapped ions, detection limit, and dynamic range in fourier transform ion cyclotron resonance mass spectrometry. *Analytical Chemistry*, 65:135–140.

MacCoss, M. J., Wu, C. C., Matthews, D. E., and Yates, J. R. (2005). Measurement of the isotope enrichment of stable isotope-labeled proteins using high-resolution mass spectra of peptides. *Analytical Chemistry*, 77(23):7646–7653.

Makarov, A., Denisov, E., Lange, O., and Horning, S. (2006). Dynamic range of mass accuracy in ltq orbitrap hybrid mass spectrometer. *Journal of the American Society for Mass Spectrometry*, 17:977–982.

Mallis, R. J., Buss, J. E., and Thomas, J. A. (2001). Oxidative modification of H-ras: S-thiolation and S-nitrosylation of reactive cysteines. *The Biochemical Journal*, 355:145–153.

Mandell, J. G., Falick, A. M., and Komives, E. A. (1998). Identification of protein-protein interfaces by decreased amide proton solvent accessibility. *Proceedings of the National Academy of Sciences of the United States of America*, 95(25):14705–14710.

Mann, M., Meng, C. K., and Fenn, J. B. (1989). Interpreting mass spectra of multiply charged ions. *Analytical Chemistry*, 61:1702–1708.

Mann, M. and Wilm, M. (1994). Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Analytical Chemistry*, 66:4390–4399.

Marshall, A. G. (2000). Milestones in fourier transform ion cyclotron resonance mass spectrometry technique development. *International Journal of Mass Spectrometry*, 200:331–356.

Marshall, A. G., Comisarow, M. B., and Parisod, G. (1979). Relaxation and spectral line shape in fourier transform ion resonance spectroscopy. *The Journal of Chemical Physics*, 71:4434–4444.

Marshall, A. G. and Hendrickson, C. L. (2001). Charge reduction lowers mass resolving power for isotopically resolved electrospray ionization fourier transform ion cyclotron resonance mass spectra. *Rapid Communications in Mass Spectrometry*, 15:232–235.

Marshall, A. G. and Hendrickson, C. L. (2002). Fourier transform ion cyclotron resonance detection: principles and experimental configurations. *International Journal of Mass Spectrometry*, 215(1):59–75.

Marshall, A. G., Hendrickson, C. L., and Jackson, G. S. (1998). Fourier transform ion cyclotron resonance mass spectrometry: A primer. *Mass Spectrometry Reviews*, 17(1):1–35.

Marshall, A. G., Hendrickson, C. L., and Shi, S. D. (2002). Scaling ms plateaus with high-resolution ft-icrms. *Analytical Chemistry*, 74:252A–259A.

Marshall, A. G., Senko, M. W., Li, M., Dillon, S., Guan, S., and Logan, T. M. (1997). Protein molecular mass to 1 da by 13c, 15n double-depletion and ft-icr mass spectrometry. *Journal of the American Chemical Society*, 119:433–434.

Marshall, A. G. and Verdun, F. R. (1990). *Fourier Transforms in NMR, Optical, and Mass Spectrometry*. Elsevier, Amsterdam.

McComb, M. E., Oleschuk, R. D., Chow, A., Ens, W., Standing, K. G., Perreault, H., and Smith, M. (1998). Characterization of hemoglobin variants by MALDI-TOF MS using a polyurethane membrane as the sample support. *Analytical Chemistry*, 70(24):5142–5149.

McKinney, C. R., McCrea, J. M., Epstein, S., Allen, H. A., and Urey, H. C. (1950). Improvements in mass spectrometers for the measurement of small differences in isotope ratios. *The Review of Scientific Instruments*, 21:724–730.

McLafferty, F. W., Fridriksson, E. K., Horn, D. M., Lewis, M. A., and Zubarev, R. A. (1999). Biomolecule mass spectrometry. *Science*, 284:1289–1290.

McLafferty, F. W. and Turecek, F. (1993). *Interpretation of Mass Spectra*. University Science Books.

Mitchell, D. W. and Smith, R. D. (1995). Cyclotron motion of 2 coulombically interacting ion clouds with implications to fourier-transform ion-cyclotron resonance mass-spectrometry. *Physical Review E*, 52:4366–4386.

Moreno, J. J. and Pryor, W. A. (1992). Inactivation of alpha-1-proteinase inhibitor by peroxynitrite. *Chemical Research in Toxicology*, 5:425–431.

Mortz, E., O'Connor, P. B., Roepstorff, P., Kelleher, N. L., Wood, T. D., McLafferty, F. W., and Mann, M. (1996). Sequence tag identification of intact proteins by matching tandem mass spectral data against sequence databases. *Proceedings of the National Academy of Sciences of the United States of America*, 93(16):8264–8267.

Moyer, S. C., Budnik, B. A., Pittman, J. L., Costello, C. E., and O'Connor, P. B. (2003). Attomole peptide analysis by high pressure matrix-assisted laser desorption/ionization fourier transform mass spectrometry. *Analytical Chemistry*, 75:6449–6454.

Nielsen, M. L., Savitski, M. M., and Zubarev, R. A. (2005). Improving Protein Identification Using Complementary Fragmentation Techniques in Fourier Transform Mass Spectrometry. *Molecular and Cellular Proteomics*, 4(6):835–845.

Nikolaev, E. N., Miluchihin, N. V., and Inoue, M. (1995). Evolution of an ion cloud in a fourier transform ion cyclotron resonance mass spectrometer during signal detection: its influence on spectral line shape and position. *International Journal of Mass Spectrometry*, 148:145–157.

O'Connor, P. B. (2004). Boston university data analysis. www.bumc.bu.edu/ftms.

O'Connor, P. B. and Costello, C. E. (2001). A high pressure matrix-assisted laser desorption/ionization fourier transform mass spectrometery ion source for thermal stabilization of labile molecules. *Rapid Communications in Mass Spectrometry*, 15:1862–1868.

O'Connor, P. B. and McLafferty, F. W. (1995). Oligomer characterization of 4-23 kda polymers by electrospray fourier transform mass spectrometry. *Journal of the American Chemical Society*, 117:12826–12831.

O'Connor, P. B., Pittman, J. L., Thomson, B. A., Budnik, B. A., Cournoyer, J. C., Jebanathirajah, J., Lin, C., Moyer, S., and Zhao, C. (2006). A new hybrid electrospray fourier transform mass spectrometer: Design and performance characteristics. *Rapid Communications in Mass Spectrometry*, 20:259–266.

Olsen, J. V., Ong, S. E., and Mann, M. (2004). Trypsin cleaves exclusively c-terminal to arginine and lysine residues. *Molecular and Cellular Proteomics*, 3(6):608–14.

Ong, S.-E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., and Mann, M. (2002). Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics. *Molecular and Cellular Proteomics*, 1(5):376–386.

Ong, S.-E. and Mann, M. (2005). Mass spectrometry-based proteomics turns quantitative. *Nature Chemical Biology*, 1(5):252–262.

Oppenheim, A. V., Schafer, R. N., and Buck, J. R. (2002). *Discrete-Time Signal Processing*. Pearson Education Pvt. Ltd., Delhi.

Papoulis, A. (1984). *Probability, Random Variables, and Stochastic Processes, 2nd edition*. McGraw-Hill, New York.

Poor, H. V. (1994). *An Introduction to Signal Detection and Estimation (Springer Texts in Electrical Engineering), 2nd edition*. Springer Verlag, New York.

Proakis, J. G. and Manolakis, D. G. (2003). *Digital Signal Processing*. Prentice-Hall India, New Delhi, 3 edition.

Qian, K., Robbins, W., Hughey, C., Cooper, H., Rodgers, R., and Marshall, A. (2001). Resolution and identification of elemental compositions for more than 3000 crude acids in heavy petroleum by negative-ion microelectrospray high-field fourier transform ion cyclotron resonance mass spectrometry. *Energy & Fuels*, 15(6):1505–1511.

Radi, R., Beckman, J. S., Bush, K. M., and Freeman, B. A. (1991). Peroxynitrite oxidation of sulfhydryls - the cytotoxic potential of superoxide and nitric-oxide. *The Journal of Biological Chemistry*, 266:4244–4250.

Ramsey, N. F. and Purcell, E. M. (1952). Interactions between nuclear spins in molecules. *Physical Review*, 85(1):143–144.

Reece, J. B. (2005). *Biology (Seventh Edition)*. Benjamin Cummings.

Reid, G. E. and McLuckey, S. A. (2002). 'top down' protein characterization via tandem mass spectrometry. *Journal of Mass Spectrometry*, 37:663–675.

Reinhold, B. B. and Reinhold, V. N. (1992). Electrospray ionization mass spectrometry: Deconvolution by an entropy based algorithm. *Journal of the American Society for Mass Spectrometry*, 3:207–215.

Rice, R. H. and Means, G. E. (1971). Radioactive Labeling of Proteins in Vitro. *The Journal of Biological Chemistry*, 246(3):831–832.

Rockwood, A. L. (1995). Relationship of fourier transforms to isotope distribution calculations. *Rapid Communications in Mass Spectrometry*, 9:103–105.

Rockwood, A. L. (1996). Ultrahigh-speed calculation of isotope distributions. *Analytical Chemistry*, 68:2027–2030.

Roepstorff, P. and Fohlman, J. (1984). Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomedical Mass Spectrometry*, 601:601.

Rossmann, A., Haberhauer, G., Holzl, S., Horn, P., Pichlmayer, F., and Voerkelius, S. (2000). The potential of multielement stable isotope analysis for regional origin assignment of butter. *European Food Research and Technology*, 211:32–40.

Rozanski, K., Araguas-Araguas, L., and Gonfiantini, R. (1992). Relation between long-term trends of oxygen-18 isotope composition of precipitation and climate. *Science*, 258:981–985.

Scheffzek, K., Ahmadian, M. R., Kabsch, W., Wiesmuller, L., Lautwein, A., Schmitz, F., and Wittinghofer, A. (1997). The Ras-RasGAP complex: Structural basis for gtpase activation and its loss in oncogenic Ras mutants. *Science*, 277:333–338.

Schoell, M., Schouten, S., Sinninghe Damste`, J. S., and de Leeuw, J. W. (1994). A molecular organic carbon isotope record of miocene climate changes. *Science*, 263:1122–1125.

Senko, M. W. (1998). Isopro 3.0. http://members.aol.com/msmssoft/.

Senko, M. W., Beu, S. C., and McLafferty, F. W. (1995a). Automated assignment of charge states from resolved isotopic peaks for multiply charged ions. *Journal of the American Society for Mass Spectrometry*, 6:52–56.

Senko, M. W., Beu, S. C., and McLafferty, F. W. (1995b). Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *Journal of the American Society for Mass Spectrometry*, 6:229–233.

Shen, Y. F., Tolic, N., Zhao, R., Pasa-Tolic, L., Li, L. J., Berger, S. J., Harkewicz, R., Anderson, G. A., Belov, M. E., and Smith, R. D. (2001). High-throughput proteomics using high efficiency multiple-capillary liquid chromatography with on-line high-performance esi fticr mass spectrometry. *Analytical Chemistry*, 73:3011–3021.

Shi, S. D. H. (2000). Comparison and interconversion of the two most common frequency-to-mass calibration functions for fourier transform ion cyclotron resonance mass spectrometry. *International Journal of Mass Spectrometry*, 195:591–598.

Shi, S. D.-H., Hendrickson, C. L., and Marshall, A. G. (1998). Counting individual sulfur atoms in a protein by ultrahighresolution Fourier transform ion cyclotron resonance mass spectrometry: Experimental resolution of isotopic fine structure in proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 95(20):11532–11537.

Shields, J. M., Pruitt, K., McFall, A., A, S., and Der, C. J. (2000). Understanding Ras: 'it ain't over 'til it's over'. *Trends in Cell Biology*, 10:147–154.

Smith, B. N. and Epstein, S. (1971). Two categories of $^{13}C/^{12}C$ ratios for higher plants. *Plant Physiology*, 47(3):380–384.

Spengler, B. (2004). De novo sequencing, peptide composition analysis, and composition-based sequencing: A new strategy employing accurate mass determination by fourier transform ion cyclotron resonance mass spectrometry. *Journal of the American Society for Mass Spectrometry*, 15(5):703–714.

Stamler, J. S., Lamas, S., and Fang, F. C. (2001). Nitrosylation: The prototypic redox-based signaling mechanism. *Cell*, 106:675–683.

Stark, J. M. and Hart, S. C. (1997). High rates of nitrification and nitrate turnover in undisturbed coniferous forests. *Nature*, 385:61–64.

Stott, A. W. and Evershed, R. P. (1996). 13c analysis of cholesterol preserved in archaeological bones and teeth. *Analytical Chemistry*, 24:4402 –4408.

Strittmatter, E. F., Rodriguez, N., and Smith, R. D. (2003). High mass measurement accuracy determination for proteomics using multivariate regression fitting: Application to electrospray ionization time-of-flight mass spectrometry. *Analytical Chemistry*, 75(3):460–68.

Stults, J. T. (1997). Minimizing peak coalescence: High resolution separation of isotope peaks in partially deamidated peptides by matrix assisted laser desorption/ionisation fourier transform ion cyclotron resonance mass spectrometry. *Analytical Chemistry*, 69(10):1815–1819.

Sun, J., Xin, C., Eu, J. P., Stamler, J. S., and Meissner, G. (2001). Cysteine-3635 is responsible for skeletal muscle ryanodine receptor modulation by NO. *Proceedings of the National Academy of Sciences of the United States of America*, 98:11158–11162.

Syka, J. E. P., Coon, J. J., Schroeder, M. J., Shabanowitz, J., and Hunt, D. F. (2004). Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, 101:9528–9533.

Tabb, D. L. and Shah, M. B. (2006). Determination of peptide and protein ion charge states by Fourier transformation of isotope-resolved spectra. *Journal of the American Society for Mass Spectrometry*, 17:903–915.

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

Teng, K. K., Esposito, D. K., Schwartz, G. D., Lander, H. M., and Hempstead, B. L. (1999). Activation of c-Ha-Ras by nitric oxide modulates survival responsiveness in neuronal PC12 cells. *The Journal of Biological Chemistry*, 274:37315–37320.

Theberge, R., Budnik, B. A., Connors, L. H., Skinner, M., Kaur, P., Connor, P. B. O., and Costello, C. E. (2005). Top down analysis of transthyretin using ESI FTMS. In *Proceedings of 53rd American Society of Mass Spectrometry conference on Mass Spectrometry*.

Tyers, M. and Mann, M. (2003). From genomics to proteomics. *Nature*, 422(6928): 193–197.

van der Schroeff, J. G., Evers, L. M., Boot, A. J. M., and Bos, J. L. (1990). Ras oncogene mutations in basal cell carcinomas and squamous cell carcinomas of human skin. *Journal of Investigative Dermatology*, 94(4):423–425.

Vetter, I. R. and Wittinghofer, A. (2001). Signal transduction - the guanine nucleotide-binding switch in three dimensions. *Science*, 294:1299–1304.

Wales, T. E. and Engen, J. R. (2006). Hydrogen exchange mass spectrometry for the analysis of protein dynamics. *Mass Spectrometry Reviews*, 25:158–70.

Wang, T. C. L., Ricca, T. L., and Marshall, A. G. (1986). Extension of dynamic range in fourier transform ion cyclotron resonance mass spectrometry via stored

waveform inverse Fourier transform excitation. *Analytical Chemistry*, 58:2935–2938.

Wuthrich, K. (1990). Protein structure determination in solution by NMR spectroscopy. *The Journal of Biological Chemistry*, 265(36):22059–22062.

Yergey, J. A. (1983). A general approach to calculating isotopic distributions for mass spectrometry. *International Journal of Mass Spectrometry and Ion Processes*, 52:337–349.

Zhang, L.-K., Rempel, D., Pramanik, B. N., and Gross, M. L. (2005). Accurate mass measurements by fourier transform mass spectrometry. *Mass Spectrometry Reviews*, 24:286–309.

Zhang, Z. and Marshall, A. G. (1998). A universal algorithm for fast and automated charge state deconvolution of electrospray mass-to-charge ratio spectra. *Journal of the American Society for Mass Spectrometry*, 9:225–233.

Zhao, C., Sethuraman, M., Clavreul, N., Kaur, P., Cohen, R., and O'Connor, P. B. (2006). Detailed map of oxidative post-translational modifications of human p21ras using fourier transform mass spectrometry. *Analytical Chemistry*, 78(14):5134–5142.

Zubarev, R. A. (2006). Protein primary structure using orthogonal fragmentation techniques in fourier transform mass spectrometry. *Expert Review of Proteomics*, 3(2):251–261.

Zubarev, R. A. and Bondarenko, P. V. (1991). An a priori relationship between the average and monoisotopic masses of peptides and oligonucleotides. *Rapid Communications in Mass Spectrometry*, 5(6):276–277.

Zubarev, R. A. and Demirev, P. A. (1998). Isotope depletion of large biomolecules: Implications for molecular mass measurements - instrumentation and applications in biological research. *Journal of the American Society for Mass Spectrometry*, 9:149–56.

Zubarev, R. A., Demirev, P. A., Haakansson, P., and Sundqvist, B. U. R. (1995). Approaches and limits for accurate mass characterization of large biomolecules. *Analytical Chemistry*, 67(20):3793–3798.

Zubarev, R. A., Hakansson, P., and Sundqvist, B. (1996). Accuracy requirements for peptide characterization by monoisotopic molecular mass measurements. *Analytical Chemistry*, 68(22):4060–4063.

Zubarev, R. A., Kelleher, N. L., and McLafferty, F. (1998). Electron capture dissociation of multiply charged protein cations. a nonergodic process. *Journal of the American Chemical Society*, 13:3265 – 3266.

Curriculum Vitae

Parminder Kaur
Born in 1977 at Patiala, Punjab, India
Email: pkhatra@bu.edu

Education

June 2002 - Jan 2007	Ph.D., Computer Engineering, Boston University, Boston Thesis topic : Statistical Methods for Interpretation of High Resolution Mass Spectra
Jan. 2001 - May 2002	M.S., Computer Engineering, Boston University, Boston. Major : Computer Networks,
Aug. 1994 - May 1998	B. Tech., Computer Science and Engineering, Punjab Technical University, Jalandhar, Punjab, India.

Experience

June 2002 - present	Graduate Research Assistant at Mass Spectrometry Resource, Boston University School of Medicine, Boston, MA. Job responsibilities: Conducting research in the field of statistical data analysis for high resolution mass spectra.
Jan 2001 - May 2002	Graduate Teaching Fellow, Department of Electrical and Computer Engineering, Boston University, Boston, MA. Courses taught: Electric Circuits Theory, Computer Communication and Networks. Job responsibilities: Assisting students with coursework and laboratory assignments, grading homeworks and exams.
Aug 1998 - Dec 2000	Lecturer, Department of Electrical and Computer Engineering, Guru Nanak Dev Engineering College, Ludhiana, Punjab, India. Courses taught: Computer Architecture, Fundamentals of Computer Programming and Information Technology. Job responsibities: Planning and teaching the coursework, scheduling and grading exams and homeworks, designing and conducting laboratory assignments.

Publications

1. Kaur, P.; O'Connor, P. B. *Determination of High-Precision Isotope Ratios from Experimental Isotopic Distributions*, Anal. Chem., In Press

2. Kaur, P.; O'Connor, P. B. *Algorithms for automatic interpretation of high resolution mass spectra*, J. Am. Soc. Mass Spectrom. 2006, 17, 459-468

3. Kaur, P.; O'Connor, P. B. *Comparison of Charge State Determination Methods for High Resolution Mass Spectra*, IEEE Int. Conf. on Granular Computing, 10-12 May

2006, 550 - 553

4. Zhao, C.; Sethuraman, M.; Clavreul, N.; Kaur, P.; Cohen, R. A.; O'Connor, P. B. *Detailed Map of Oxidative Post-Translational Modifications of Human P21Ras Using Fourier Transform Mass Spectrometry*, Anal. Chem. 2006, 78, 5134-5142

5. Jebanathirajah, J. A.; Pittman, J. L.; Thomson, B. A.; Budnik, B. A.; Kaur, P.; Rape, M.; Kirschner, M.; Costello, C. E.; O'Connor, P. B. *Characterization of a new qQq-FTICR mass spectrometer for post-translational modification analysis and top-down tandem mass spectrometry of whole proteins*, J. Am. Soc. Mass Spectrom. 2005, 16, 1985-1999

6. Kaur, P.; O'Connor, P. B. *Use of statistical methods for quantitative determination of the number of trapped ions*, Anal. Chem. 2003, 76, 2756-2762

7. O'Connor, P. B.; Budnik, B. A.; Ivleva, V. B.; Kaur, P.; Moyer, S. C.; Pittman, J. L.; Costello, C. E. *A High Pressure Matrix-Assisted Laser Desorption Fourier Transform Mass Spectrometry Ion Source Designed to Accommodate Large Targets with Diverse Surfaces*, J Am Soc Mass Spectrom 2003, 15, 128-132

Presentations

1. Kaur, P; Zhao, C; O'Connor, P. B. *Determination of High-Precision Isotope Ratios from Experimental Isotopic Distributions* ASMS 2006 Conference Proceedings

2. O'Connor, P. B.; Lin C.; Mathur, R.;Aizikov, K.;Cournoyer, J.; Kaur, P.; Ivleva, V. B.; Zhao, C. *A fourier transform ion cyclotron resonance mass spectrometer designed to operate at cryogenic temperatures* ASMS 2006 Conference Proceedings

3. Zhao, C.; Sethuraman, M.; Clavreul, N.; Kaur, P.; Cohen, R.; O'Connor, P. B. *A Detailed Map of Oxidative Post-translational Modifications of Human p21ras using Fourier Transform Mass Spectrometry* ASMS 2006 Conference Proceedings

4. Kaur, P.; Aizikov, K.; Budnik, B. A.; O'Connor, P. B. *MasSPIKE (Mass Spectrum Interpretation and Kernel Extraction) for Biological Samples* ASMS 2005 Conference Proceedings

5. Theberge, R.; Budnik, B. A.; Kaur, P.; Connors, L. H.; Skinner, M.; O'Connor, P. B.; Costello, C. E.*Top Down Analysis of Transthyretin Using ESI FTMS* ASMS 2005 Conference Proceedings

6. Huang, H; McComb, M. E.; Perlman, D. H.; Budnik, B. A.; Kaur, P.; Skelton, T. P.; Chui, D. H. K.; O'Connor, P. B.; Costello, C. E. *Proteomics Approach for Identification of Hemoglobin Variants and Post-Translational Modifications* ASMS 2005 Conference Proceedings

7. McComb, M. E.; Perlman, D. H.; Huang, H.; Budnik, B. A.; Kaur, P.; O'Connor P. B.; Costello, C. E. *Direct Protein 2D-LC MALDI With On-Target Digestion for High-Throughput Proteomic Analyses* ASMS 2005 Conference Proceedings

8. Dauly, C.; Odhiambo, A.; Perlman, D. H.; Huang, H; Budnik, B. A.; Kaur, P.; O'Connor, P. B.; Steinberg, M. H.; Farber, H. W.; Klings, E. S.; McComb, M. E.; Costello, C. E. *Comparative Proteomics of Abundant Protein-depleted Plasma from Patients with Sickle Cell Disease-related Pulmonary Hypertension* ASMS 2005 Conference Proceedings

9. Kaur, P.; Aizikov, K; O'Connor, P. B. *Improved Algorithms for Interpretation of High Resolution Mass Spectra* ASMS 2004 Conference Proceedings

10. Kaur, P.; O'Connor, P. B. *Data Mining Methods for MALDI-FTMS Data* ASMS 2003 Conference Proceedings

11. Budnik, B. A.; Moyer, S. C.; Kaur, P.; Costello, C. E.; OConnor, P. B. *Automated high pressure MALDI FTMS - Advantages of a new design* ASMS 2003 Conference Proceedings

12. Moyer, S. C.; Budnik, B. A.; Kaur, P.; Costello, C. E.; O'Connor, P. B. *Sensitivity Increase Resulting from Design Improvements for a High Pressure MALDI Source on an FTMS* ASMS 2003 Conference Proceedings

Graduate Courses

Detection and Estimation Theory, Statistical Pattern Recognition, Stochastic Processes, Mass Spectrometry based Proteomics, Biomolecular Architecture, Digital Signal Processing, Discrete Mathematics.

Awards

University Gold Medal for BS in Computer Engineering, Punjab Technical University, India
American Society of Mass Spectrometry travel grant award, 2003
Boston University Electrical & Computer Engineering travel grant award, 2006