THE EFFECT OF SEARCH ACCURACY ON SHOTGUN PROTEMOICS RESULTS USING A VALIDATED PROTEIN DATABASE AND HIGH RESOLUTION FRAGMENT SPECTRA

by

ARTHUR NUCCIO

(Under the Direction of Ron Orlando)

ABSTRACT

Mass spectrometry based proteomics has become one of the most powerful tools used to determine protein structure, function, and expression. The recent rapid expansion of the field is a result of significant improvements in the throughput of mass spectrometers and the availability of affordable and powerful computers. Due to the improvements to mass spectrometers, LC-MS/MS experiments are able to generate tens of thousands of MS/MS spectra making automated interpretation a necessity. The results of automated database searches must be scrutinized very carefully.

In this study, the incidence of false positives from search results was explicitly measured by using a database annotated with randomly generated proteins. True positives were measured as the identification of a subset of individually validated recombinant proteins that were spiked into

the experimental sample. Searches were performed under a variety of high and low mass accuracy settings, and the performance of a few methods of statistical validation was analyzed.

This investigation shows that post-search statistical validation should be a mandatory step in a shotgun proteomics experiment. Furthermore, high resolution mass accuracy and resolution ion fragment scanning should be used as soon as mass spectrometers can operate in those modes without significant losses to scan times and throughput.

INDEX WORDS: Shotgun Proteomics, False Discovery Rate, Peptide Probability, Mass Spectrometry, High Mass Accuracy, SEQUEST, Mascot, X!Tandem

THE EFFECT OF SEARCH ACCURACY ON SHOTGUN PROTEMOICS RESULTS USING A VALIDATED PROTEIN DATABASE AND HIGH RESOLUTION FRAGMENT SPECTRA

by

ARTHUR GAYOSA NUCCIO

B.S., University of Georgia, 2005

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

MASTER OF SCIENCES

ATHENS, GEORGIA

2011

THE EFFECT OF SEARCH ACCURACY ON SHOTGUN PROTEMOICS RESULTS USING A VALIDATED PROTEIN DATABASE AND HIGH RESOLUTION FRAGMENT SPECTRA

by

ARTHUR GAYOSA NUCCIO

Major Professor: Ron Orlando

Committee: Jonathan Amster
Lance Wells

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2011

ACKNOWLEDGEMENTS

This work would not have been possible without the assistance of many people. In order to achieve success, modern science requires a large and capable team of lab-mates and collaborates. My deepest personal gratitude goes to the many individuals who I been privileged to work with through my time at the University of Georgia. I am especially grateful to Ron Orlando. I have always been humbled by your creativity. I would also like to thank Carl Bergmann and Gerardo Gutierrez-Sanchez for all the time and advice they have given me as well as trusting me with routinely using their laboratory. I owe much to James Atwood III. You have always been extremely patient and a great motivator. I would also like to extend thanks to the numerous people I worked with in the Orlando lab at the CCRC and Coverdell Center; Kumar Kolli, Brent Weatherly, Gretchen Cooley, Rick Tarleton, Bryan Woosley, Peggi Angel, Punit Shah, Lei Chen, Xiang Zhu, Juli Botelho, Caroline Watson, DJ Johnson, Josh Sharp, ViLihn Tran, Jessie Saladino, Jesse Hines, Lance Wells, and Jae-Min Lim. By far the most important people to have helped me get here are my family members, especially my father who instilled in me a curiosity of the sciences from the beginning of my life.

TABLE OF CONTENTS

CHAPTER

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Mass Spectrometry-Based Proteomics

Proteomics is the "study of protein properties (expression level, post-translational modification, interactions, etc.) on a large scale to obtain a global, integrated view of disease processes, cellular processes and networks at the protein level."[1] "use of quantitative protein-level measurements of gene expression to characterize biological processes and decipher the mechanisms of gene expression control."[2]

Before groundbreaking progress in genomics, chemical and enzymatic methods utilizing UV detection, such as Edman degradation, were used to derive amino acid sequence information from highly purified proteins. Mass spectrometers have been increasingly integrated into the protein identification workflow, eventually surpassing Edman degradation as the most common method. The improvement of mass spectrometers over time as well as the availability of full genomes accelerated the ascension of mass spectrometry as the primary protein identification technology in proteomics. Unfortunately, the number of proteins in a species' proteome is far larger than the number of genes in the same species' genome. Furthermore, the concentration range of proteins in a proteome surpasses the dynamic range of any one analytical method. In order to account for the sometimes 10 order of magnitude difference in serum protein concentrations, several mass spectrometric and analytical techniques must be combined.[3]

There are two primary types of proteomics experiments performed today; bottom-up and top-down approaches. Bottom-up experiments identify protein sequences by first converting

proteins into peptides usually with an enzyme such as trypsin. Peptides are separated via gels or chromatography, ionized by MALDI or ESI, and injected into a mass spectrometer. The peptides are then fragmented in the mass spectrometer collision induced dissociation (CID) generating characteristic ion patterns that can be used to deduce amino acid sequence. These fragmentation spectra are searched by computer algorithms against protein databases that report matches based on varying criteria. Tryptic peptides are very advantageous because they are aqueous and spray with a +2 charge. This allows for a generally good ion ladder and therefore structural information to be derived from tryptic peptides. However, only a small percentage of the tryptic peptides are detected and even fewer yield useful fragmentation data. Additionally, characterizing PTM's with bottom-up proteomics is difficult.[4, 5]

Top-down proteomics is an approach where intact proteins are fragmented in the mass spectrometer. The intact protein mass and the peptide masses are observed, and if enough informative fragment peaks are detected, the entire protein sequence can be elucidated. Newer fragmentation techniques such as electron transfer dissociation (ECD) and electron transfer dissociation (ETD) have enhanced the capability of top-down approaches. High mass accuracy and resolution are critical in top-down proteomics leading to the popularity of FT-ICR and orbitrap instruments. Top-down proteomics is a useful tool for analyzing PTM's due to the preservation of their molecular features. Some disadvantages of top-down proteomics are only proteins samples of high purity can be used, and large proteins are difficult to analyze due to the numerous ways that they can fragment.[4, 5]

Ion Traps

Ion traps where first described by Steinwedel and Paul in 1960, and the principle of the trap was first applied by Stafford for the purpose of mass spectrometry. The trap itself consists of 2 cap electrodes positioned above and below a ring electrode. Ions travel through the trap in circular oscillations as if the trap was a quadrupole mass analyzer bent around itself. Describing the theory of the quadrupole mass analyzer can make this analogy clearer and facilitates understanding how the ion trap functions.

Quadrupole Mass Analyzers

A quadrupole mass spectrometer consists of an ion source, focus lenses, the quadrupole mass filters, and a detector. The quadrupole mass analyzer consists of four hyperbolic or cylindrical rods that are charged with radio frequency (RF) or direct current (DC) voltages. Ions traveling through the quadrupole (z-axis) have trajectories manipulated by the electric field within the quadrupole. This electric field is caused by potentials applied to the quadrupole rods

$$\Phi_0 = (U - V\cos\omega t) \text{ and } -\Phi_0 = -(U - V\cos\omega t) \quad (1)$$

Φ_0 is the potential applied to the rods, U is direct potential, V is the amplitude of the RF voltage, t is time, and ω is angular frequency (which is related to RF)[6, 7].

The motion of ions as they travel the length of the quadrupole can be modeled by using Mathieu equations

$$\frac{d^2u}{d\xi^2} + (a_u - 2q_u \cos 2\xi)u = 0 \quad (2)$$

$$\text{with } a_u = \frac{8eU}{mr_0^2\omega^2} \text{ and } q_u = \frac{4eV}{mr_0^2\omega^2} \quad (3)$$

In the equation above, *u* represents *x* or *y*, ξ substitutes for $\omega t/2$, *e* is an electron's charge, and *m* is the mass of the ion. The parameters *U*, *V*, and ω were defined above. Solutions to the Mathieu equation can be portrayed graphically as in Figure 1.1. For a given ion mass, the area within this stability region represents what RF and DC voltages applied to the quadrupole rods will permit the ion to successfully travel through the quadrupole region and ultimately reach the detector. This stability diagram is created by finding a region that overlaps between the diagram for stability in the *x* direction and the diagram for stability in the *y* direction as in Figure 1.2. A stability diagram of three ions with differing *m*/*z* is shown in Figure 1.3. As the RF and DC voltages are increased (maintaining the DC/RF ratio), ions of smaller *m/*z are forced into larger oscillations caused by the resulting changes in the quadrupole's electric field. The electric field manipulates the movement of the ions causing them to oscillate in the *xy* plane. If the oscillations are severe enough, the ion will impact the quadrupole rods where it is neutralized and will not reach the detector. Mathematically speaking, *x* or *y* from the equation above became equal to r_0.

Ion Traps

An ion trap is a three dimensional version of a quadrupole mass analyzer.[8] This means that ions within the trap do not have any direction where they can move free of manipulation from an electric field, such as the *z* direction in a quadrupole analyzer. Ion traps consist of a ring electrode and two hyperbolic end-cap electrodes as depicted in Figure 1.4. The equations of motion representing ion movement within the trap are very similar to those for quadrupoles

$$\frac{\mathrm{d}^2 z}{\mathrm{d}t^2} - \frac{4Ze}{m(r_0^2 + 2z_0^2)}(U - V\cos\omega t)z = 0 \quad (4)$$

$$\frac{\mathrm{d}^2 r}{\mathrm{d}t^2} - \frac{4Ze}{m(r_0^2 + 2z_0^2)}(U - V\cos\omega t)r = 0 \quad (5)$$

In the above equations, *r* can replace the *x* and *y* coordinates due to cylindrical symmetry. The *Z* term represents the number of charges on the ion, and the remaining terms are the same as defined for the quadrupole motion equations. The generalized form of the Mathieu equation for ion traps is

$$\frac{d^2 u}{d\xi^2} + (a_u - 2q_u\cos 2\xi)u = 0 \quad (6)$$

where *u* can be either *z* or *r*. The Mathieu equations of motion have the form

$$a_z = -2a_r = \frac{-16ZeU}{m(r_0^2 + 2z_0^2)\omega^2} \quad (7)$$

$$q_z = -2q_r = \frac{-8ZeU}{m(r_0^2 + 2z_0^2)\omega^2} \quad (8)$$

where again *u* is either *z* or *r*. Stable ions must not have their displacement coordinates equal r_0 or z_0 or they will hit the ring or endcap electrodes.

Most ion traps use electrospray ionization though some GCMS instruments still use internal ionization via electron beams. Analyte is ionized by electrospray and directed by a skimmer and gating lens into a series of hexa- or octapoles. The hexa/octapoles direct ions into the ion trap. Sample introduction from the spray into a capillary inlet occurs at atmospheric pressure while the hexa/octapoles are kept at a vacuum by differential pumping.

The kinetic energy of incoming ions is stymied by ~1 mTorr helium gas in the trap. This allows ions to more easily fall into an orbit in the center of the trap. Ion traps can hold between

10^5 and 10^6 ions before repulsions interfere with ion trajectories resulting in a loss of resolution.[8]

Just as with quadrupole analyzers, ion traps can control what *m*/*z* of ions are stable by altering the RF and DC voltage of the ring and endcap electrodes. As the RF/DC voltage is increased, ions with smaller *m*/*z* are pushed into orbits that increasingly approach the endcap electrodes. Once RF/DC voltage exceeds all the possible values in an ion's stability diagram, that ion's path becomes unstable and the on the ion's final oscillation within the trap, it falls into an exit slit and strikes the detector. Gradually increasing the RF/DC voltage of the trap electrodes will remove the ions from the trap increasing from smallest to larger *m*/*z*.

Ion trap instruments are also able to selectively eject specific *m*/*z* ions by resonance ejection. Each ion oscillates within the trap at a characteristic frequency called the secular frequency. If additional AC is applied to the cap electrodes at a particular ion's secular frequency, then that ion gains increasing amounts of kinetic energy until it oscillates out of the trap. This method creates what can conceptually be described as a hole in an ion's stability diagram. As the RF/DC voltage is increased and the resonance frequency is maintained, increasingly larger *m*/*z* ions "fall through the hole" and are ejected from the trap. The resonance ejection method effectively increases the mass detection range of the ion trap.

Orbitrap Mass Specrometers

An orbitrap mass spectrometer is a relatively new type of mass spectrometer created by Alexander Makarov in 2000.[9] Its design is similar to the Kingdon trap of 1923 which consisted of a central wire mounted axially in a surrounding cylinder with two rims containing the trapping volume. When a voltage is applied to the wire and cylinder, ions attracted to the

wire will survive only if they have enough tangential velocity to not collide with the wire. Because the rotational frequency of sample ions is dependent initial position and velocity, the resolution of mass to charge measurements derived from rotational frequencies is poor. By observing axial frequencies using image current detection and FT algorithms, Makarov was able to observe 100,000 to 200,000 resolution and 2-5 ppm mass accuracy.

The orbitrap mass analyzer of an inner spindle shaped electrode surrounded by a coaxial barrel like electrode. The electrostatic field created by the orbitrap has the distribution

$$U(r,z) = \frac{k}{2}\left(z^2 - \frac{r^2}{2}\right) + \frac{k}{2}(R_m)^2 \ln\left[\frac{r}{R_m}\right] + C \quad (9)$$

Where r and z are cylindrical coordinates, k is field curvature, R_m is the radius and C is a constant. The shape of the electrodes is given by

$$z_{i,o}(r) = \sqrt{\frac{r^2}{2} - \frac{(R_{i,o})^2}{2} + (R_m)^2 + \ln\left[\frac{R_{i,o}}{r}\right]} \quad (10)$$

where i represents the inner electrode and o represents the outer electrode. R_i denotes the maximum radius of the inner electrode, and R_o denotes the maximum radius of the outer electrode.

The equation of motion for ions moving in the axial direction is

$$z(t) = z_0 \cos(\omega t) + \sqrt{\left(\frac{2E_z}{k}\right)\sin(\omega t)} \quad (11)$$

where E_z is the energy characteristic for the z direction and ω is the frequency of axial oscillations in radians per second

$$qE_z = \left(\frac{m}{2}\right)(z_0)^2 \quad (12)$$

$$\omega = \sqrt{\left(\frac{q}{m}\right)k} \quad (13)$$

Orbitraps do not use RF or magnetic fields to trap ions. An electrostatic field causes an attraction that balances the path of moving ions against the centripetal force they feel when rotating around the central spindle of the orbitrap. Ions move around the central spindle in a complex spiral as shown in Figure 1.5. The rotational frequencies of these spirals depend upon the initial velocities and positions of injected ions. Ions of the same mass to charge can stay in phase in the z-axis (axial) for hundreds of thousands of oscillations. However, ion oscillations in the radial direction come out of phase in only 50 to 100 oscillations. This dephasing results in the ion cluster devolving into a thin oscillating ring causing measured radial image currents to cancel out. Mass to charge signal derived from these radial frequencies would have poor resolution. Fortunately, axial frequencies are independent of initial velocity and position and can be detected as an image current using electrodes surrounding the orbitrap. Fourier transforms can then be used to determine the axial oscillation frequencies of ions and their mass to charge ratio.[10, 11]

Though the orbitrap is an excellent mass analyzer, its absolute mass accuracy and resolving power fall behind those of FT-ICR instruments. However, the resolving power of an FT-ICR instrument varies inversely with measured mass to charge. At m/z 1000, an orbitrap's resolving power can outperform FT-ICR instruments of less than 4.7T strength for 1 second acquisitions. Additionally, due to the weaker mass dependence of the orbitrap, the orbitrap's resolving power would eventually surpass a 14.7T FT-ICR at higher m/z values. As a cheaper alternative to the FT-ICR, orbitraps are quite capable having ~150,000 resolution at 100 m/z, 2 to 5 ppm mass

accuracy using internal and external calibration respectively, 5000 dynamic range, and 6000 m/z charge limit.[12]

Thermo LTQ-Orbitrap

The Thermo LTQ-Orbitrap mass spectrometer consists of an LTQ mass spectrometer, C-trap, orbitrap mass spectrometer, and several connecting multipoles and ion guides. Combining the LTQ with the orbitrap allows for a system that has high trapping capacity, MS^n fragmentation capabilities, and automatic gain control in addition to very high mass accuracy, dynamic range, and resolving power.[13]

Figure 1.6 represents a diagram of the layout of Thermo's LTQ Orbitrap. The front end of the instrument has an LTQ mass spectrometer with an ESI ion source. Ions from the spray are guided into the linear trap via rf-only multipoles. Ions are analyzed and ejected radially into a pair of MCP detectors. A transfer octapole moves ions into a curved rf-only quadrupole called the C-trap. The C-trap is useful for performing radial rather than axial injections into the orbitrap. Radial injection allows for a fast, uniform, and large injection of ions into the orbitrap, thereby staying within the space charge limit of the orbitrap.[14] Ions in the C-trap lose energy by colliding with ~1mTorr nitrogen gas and come to rest after multiple passes in the C-trap. Collisions are low enough in energy such that no further fragmentation occurs.

After collisional cooling, ions are formed into a curved thin thread along the curve of the C-trap. The thread is compressed with 200V applied to both the gate and trap electrodes. Then to eject, the rf is ramped down and the following sequence occurs: 1200V to the push-out electrode, 1000V to the pullout electrode, and 1100V to the top and bottom electrodes. Ions are ejected orthogonally through a pull-out electrode. Ion optics under high vacuum accelerate the

ejected ions and they enter the orbitrap radially in a tight cloud. Ions within the orbitrap are at ~$2x10^{-10}$mbar. No excitation is necessary because the ions are injected offset from the equator of the orbitrap. Image current is detected with two outer electrodes and fast Fourier transformation yields spectral data.[13, 14]

Reverse-Phase High-Performance Liquid Chromatography

Revere-phase liquid chromatography (RP-HPLC) is an analysis technique that separates molecules based on their hydrophobicity. Hydrophobic effects allow the binding of solute molecules from the mobile phase to immobilized ligands, typically C18, attached to the stationary phase. Samples must be loaded in aqueous phase, and elutions are performed with increasing hydrophobicity. Analytes will elute in according to their hydrophobicity, with most hydrophobic analytes being retained the longest.

RP-HPLC is extremely useful for the analysis of peptides and proteins for several reasons. This separation technique offers great resolution, ease in altering of selectivity by changing mobile phase, high recoveries, and excellent reproducibility. Complex mixtures of peptides and proteins are routinely separated at low pico- to femtomol range. Separations can also be manipulated by changing the gradient, temperature, and organic modifier. These boons overshadow the negative of RP-HPLC irreversibly denaturing sample proteins and peptides.

The stationary phase in RP-HPLC usually comprises alkyl groups bonded to glass beads. Alkyl chains are chemically bonded to silanol groups on the surface of the stationary phase. Typically, alkyl chains consisting of 18 carbon atoms (C18) are used, but C4 and C8 are also widely used. Immobilizing ligands onto the surface silanol groups only uses about one-half of the available sites due to steric hindrance. Smaller alkyl groups are used to end-cap the open

sites preventing non-specific binding. Acidic modifiers such as formic and acetic acid are also added to the mobile phase to reduce non-specific binding.

Detection of peptides and proteins using RP-HPLC offline from mass spectrometry is most often performed by UV-Vis spectroscopy. The amide stretch of peptide bonds has an absorption band at 210 to 220nm (214nm typically observed) and the aromatic amino acids, tryptophan and tyrosine, absorb at 280nm. The absorption at 280nm is much stronger than at 214nm, but peptides without aromatic amino acids will not be seen.[15]

Capillary Liquid Chromatography

Smaller scale chromatographic analysis is needed for rare peptides and proteins. Capillary HPLC concentrates dilute samples by retaining analyte on stationary phase and eluting it online with mass specs using organic gradients. By comparison, normal scale HPLC causes unacceptable losses of sample and sensitivity. Capillary HPLC also has the boon of being easily coupled to ESI and mass spectrometers.[16]

Capillary ESI using sub-µL/min flow rates has several advantages: Lower sample consumption, better limits of detection, and low femptomol sensitivity.[17] Conventional ESI requires post column or coaxial addition of organic to lower surface tension and sheath gas to increase the vaporization of mobile phase. With this setup, only a small percentage of flow reaches the vacuum inlet of the mass spectrometer. Scaling down to capillary ESI eliminates the need for a sheath gas and non-gradient additions of organic to the mobile phase. Lower scaling also reduces the potential needed for the spray needle allowing it to be physically moved much closer to the vacuum inlet, “within [the] vortex of gas” drawn into the mass spectrometer.[18]

With capillary ESI-LCMS, it is possible to detect 100amol of analyte using single ion monitoring, and 10fmol of tryptic digest using full scan MS/MS.[17]

High quality capillary columns are necessary for high sensitivity LC-MS/MS experiments. Fortunately, these can be made economically in laboratory settings. Capillary columns are made of fused silica capillary (FSC) tubing that is commercially available. Common specifications for the FSC columns are a 100µm inner diameter and 350µm outer diameter. Columns are 10 to 15cm in length and contain 6 to 8cm of stationary phase. The stationary phase is usually C18 beads with 5µm diameters and 300Å pores; however, strong cation exchange (SCX) material can also be used. Biphasic columns containing section packed with RP resin at the tip and SCX resin near the base are not uncommon in LC-MS/MS.[18] Applying an electric charge to the mobile phase requires a T-junction connecting a gold electrode to the chromatographic flow. The other two ends of the T-junction, the ones opposite each other, attach to the base of the capillary column and the transfer lines from the LC. Micro-spray voltages are in the 1.7 to 1.9kV range while nano-spray voltages are in the 700-900V range.[17]

Creating capillary columns requires heating FSC tubing and pulling it apart until the FSC separates and a tip is formed. This can be performed using a microtorch, but laser pullers are preferable due to greater reproducibility of tip shape. Using a laser puller, researchers can create columns with tips of 2.1µm internal diameter and 50 to 100nm FSC thickness. If a laser puller is not used, a silica scribe is used to chip the pulled capillary end into a fine tip. Capillary columns are capable of withstanding 8000psi and can be reused many times. Unfortunately, FSC at these dimensions is very fragile and prone to clogging without rigorous sample filtering.[17]

Loading stationary phase into a capillary column utilizes a method called bomb loading, depicted in Figure 1.7. A bomb is a steel high pressure chamber where the column is mounted with the tip facing up and away from the bomb. This orientation of the column allows a path for the high pressure within the bomb to escape to the atmosphere. The inner chamber of the bomb is large enough to house a microcentrifuge tube containing about 1mL of a slurry of C18 resin in an alcohol solvent. The base of the capillary column is within the slurry, and once high pressure is applied (400-800psi), the C18 solution is pushed through the capillary by pressure equilibration forces.[19] After washing and equilibrating a column, sample solutions can be loaded in the same manner by replacing the C18 slurry with a vial of sample.

The minute scale of capillary LC-MS/MS makes it favorable for proteomic applications. Typical run conditions for capillary LC-MS/MS are <5μL/min which allow for the analysis of dilute solutions. This sensitivity coupled with the excellent limit of detection of mass spectrometers makes capillary LC-MS/MS a popular choice for the analysis of biological samples where protein concentrations cover and incredible dynamic range.

Ionization Methods in Mass Spectrometry Based Proteomics

The two predominant ionization methods in mass spectrometry based proteomics are matrix assisted laser desorption ionization (MALDI) and electro-spray ionization (ESI). MALDI ionization involves using the rapid vaporization of a matrix to ionize sample molecules. A 1000 to 10,000 excess of matrix to analyte solution is spotted on a target plate. Matrices are acids that are capable of absorbing UV laser light, typically sinapinic or α-cyano-hydroxycinnamic acid when analyzing whole proteins and peptides respectively. The sample/matrix mixture spotted on the plate is allowed to dry causing sample and matrix to co-crystallize, and the target plate is then

inserted into a vacuum chamber in the mass spectrometer. Pulse UV laser light excites the matrix causing it to vaporized very quickly. This rapid vaporization also excites the sample molecules in addition to transferring protons to them, though the precise mechanism for this is not well understood. As little as one ~1 femtomole of sample can be detected from MALDI-ToF MS experiments.[4, 20]

Electrospray Ionization

Electrospray ionization was pioneered by Dole et al. in 1968.[21] Fenn et al. revitalized ESI in 1988 with an analysis of PEG molecules and then biomolecules 1989.[22, 23] Smith et al. were the first to analyze proteins and polypeptides using ESI in 1989.[24] Gas phase ion-transitions of large biological molecules are difficult processes to perform. Vaporization processes often cause catastrophic destruction of the very molecules desired for analysis. ESI is a soft ionization technique that solves this problem.

Gas phase ions are analyzed by spraying a protonated analyte ion solution from a high pressure, micro (to nano) flow capillary into an electric field under high vacuum, where they are then guided into the mass analyzer. Proteins and peptides are first ionized in acidic solution where basic amino acids and N-termini acquire an additional proton from solution. Liquid chromatography is typically used to interface liquid phase analytes into a mass spectrometer. Flow rates can range from tens of µL/min to less than one hundred nL/min. A spray voltage in the low kV range is applied to the spray needle. Once sample passes through the capillary needle tip, the applied electric field causes the spray at the tip to form into a Taylor cone. The applied electric field causes the spray tip to be enriched in positive ions and the inner capillary enriched in negative ions. Repulsion of positive ions at the surface combined with the pull of the applied

electric field on positive ions overcomes the surface tension of the solvent, causing the liquid to expand into the Taylor cone. The tip elongates into a filament and breaks into a plume. This plume consists of many small droplets containing solvent and analyte. The diameter of each droplet is a function of the applied electric potential, flow rate, and solvent properties. The electric field needed for electro spray can be stated as

$$E_0 = \sqrt{\frac{2\gamma \cos(49°)}{\varepsilon_0 r_c}} \quad (14)$$

where E_0 is the electric field strength, γ is the solvent surface tension, cos(49°) is the half-angle of a Taylor cone, ε_0 is the permittivity of vacuum, and r_c is the radius of the capillary. Liquids with high surface tensions sometimes require electric fields stronger than those required for an electric carona discharge.[25]

As the droplets travel further towards the inlet capillary of a mass spec, they undergo several rounds of fission. The division is caused by several factors. Evaporation occurs as droplets travel from the spray needle to the inlet capillary of the mass spectrometer. Each droplet can also undergo what is called a Coulombic explosion. As the droplet shrinks, more and more positive ions are forced nearer to each other. When the Rayleigh limit is reached, electric repulsion of the analyte ions becomes greater than the surface tension of the liquid droplet. At this point the droplet "explodes" into daughter droplets. This occurs continuously until single ion droplets are formed where further evaporation yields non-solvated ions as depicted in Figure 1.8.

Droplets originating from the spray tip are assisted to the mass spectrometer inlet capillary by a potential and pressure gradient. Ions initially sprayed into to atmospheric pressure are drawn

into the high vacuum within the mass spectrometer. Nebulization of analyte ions is facilitated by sheath gases and heating.[20, 26, 27]

Properties of Electroprayed Ions

The transfer of ions from liquid to gas phase via ESI is a low energy event. Since it is a "cooler" ionization method, large biological molecules tend to stay intact throughout the ionization process. Since biological molecules like peptides and proteins are so large, they tend to have multiple sites where it is possible to carry a charge. When large molecules carry multiple charges, the effective mass range of a mass spectrometer is increased. Basic sites like N-termini, lysine, arginine, and histidine are the sites for additional charges on proteins and peptides. This is part of the reason that trypsin is such a useful reagent for enzymatic digestions: The cleavage products of trypsin are typically doubly charged species.

Multiply charged analytes undergo ESI more efficiently and are more likely to fit within the mass range of an ion-trap or quadrupole mass spectrometer. Varying overall charge states of a particular peptide or protein lead to several isotopic envelopes existing within a mass spectrum consisting of a given peptide/protein species. If multiple charges are due to protons, assuming sufficient instrument resolution, each peak of a ^{13}C envelope is $\frac{1}{n}$ Thompsons apart, where *n* is the number of charges (See Figure 1.9). If another cation is responsible for the positive charge, then the peaks in the ^{13}C envelope will be $\frac{m}{n}$ Th apart, where *m* is the molecular mass of the cation. By observing the difference in *m/z* between each ^{13}C peak, the charge state of the analyte can be calculated. For adjacent ^{13}C peaks of an isotopic envelope x_1 and x_2, the charge *n* of the monoisotopic species can be determined by noting:

$$x_1 = \frac{M + n}{n} \quad (15)$$

$$x_2 = \frac{M + n + 1}{n + 1} \quad (16)$$

where *M* is the molecular mass of the ion of interest. Solving for *n* yields:

$$n = \frac{x_2 - 1}{x_1 - x_2} \quad (17)$$

Once the correct charge state of a peptide precursor ion is known, the molecular mass of the ion can be determined.[26, 27] It is essential to know the mass of the precursor peptide for protein identification to succeed.

Collision Induced Dissociation

MALDI and ESI ionization fall into the category of soft ionization techniques so it is necessary for further fragmentation of analyte ions. Energizing a stable ion after it has been mass selected is typically performed using Collision-Induced Dissociation (CID). The collision occurs with gas molecules where kinetic energy of the collision is converted to internal energy of the analyte molecule, typically a tryptic peptide. The conversion to internal energy causes the analyte molecule to become unstable, and fragmentation reactions occur prior to leaving the collision cell. After dissociation, fragment ions are mass analyzed.

The bulk of CID in proteomics is performed on tryptic peptides ionized by ESI. CID of tryptic peptides forms six predominant ion types: a-, b-, and c-ions that carry N-terminal protons, and x-, y-, and z-ions that. Of those six, b- and y- ions are by far the most prevalent in quadrupole and ion trap instruments.

Using ESI and a mass analyzer in positive ion mode, strongly basic amino acids (N-terminus, lysine, arginine, and histidine) are become ions in their protonated form. The proton on lysine, arginine, and histidine, much more basic than the N-terminus, is static even during collisional activation. The N-terminal proton can migrate via internal salvation pathways. It is best to view ESI ions as a heterogeneous population of peptide ions, but with varying subpopulations of ions having identical amino acid sequence and protons associated with each amide linkage. After mass selection, these populations are accelerated into multiple low energy (10eV to 50eV) collisions with a collision gas. The kinetic energy from collision gas is converted to vibrational energy in the peptide, which the peptide releases through fragmentation reactions determined by the site of the protonated amide bond. Proton migration dictating fragmentation patterns is dubbed the "mobile proton hypothesis".

The effect of the mobile proton changes dramatically with the charge state of the peptide and the presence of very basic amino acid residues in the peptide sequence. The N-terminal charge moves to create subpopulations of ions that produce varying fragmentation ion series. For peptides that have basic amino acids residues, such as lysine and arginine, and the N-terminus, doubly charged ions are formed who have essentially fixed charge sites. Fragments with lysine, arginine, and histidine that are singly charged do not have a mobile proton to direct fragmentation. Without a mobile proton, only limited fragmentation can occur at low energy CID. The most complete fragmentation, and therefore sequence information, is always seen by fragments of the highest charge state. For tryptic peptides +3 and greater, the charges are on the N-terminus, C-terminal lysine and arginine, and any internal lysine, arginines, or histidines resulting from missed cleavages. The positions of internal charges are critical for receiving more sequence information from a peptide MS/MS spectrum. Internal charges that are fixed alter

proton migration in negative ways. In this case, mobile protons would tend not to localize on amide bonds so little fragmentation would be observed at amide positions. Rearrangements lead to creation of daughter ions and can also lead to neutral losses. Daughter ions also continue to fragment leading to smaller b-ions becoming disproportionately abundant. Fragments ions survive long enough to by mass analyzed due to the stability of the oxazolone ring structure that forms.[28]

In CID, there are 3 types of amino acid structure: The C- and N-terminal amino acids and a series of internal amino acid residues. The formula mass of the amino acid residues is the cornerstone to determining peptide amino acid sequences via mass spectrometry. The residue masses are the differences between consecutive ions of a given ion series (b-, y- etc.). Not all amino acids can be easily distinguished. Isoleucine and leucine are structural isomers and have identical mass while glutamine and lysine are isobars. Compared to residue masses N-termini will be +1Da heavier from hydrogen, and C-termini will be +17 Da heavier from hydroxyl. If the proton charge resides on the carboxyl group (C-terminus) of a fragment ion, then it is a b-ion. If the proton charge resides on the terminal amine (N-terminus) of a fragment ion, then it is a y-ion. Figure 1.10 depicts the difference in the complimentary b- and y-ions. For +2 tryptic peptides, collisional cleavage at a protonated amide bond creates both a b- and y-ion daughter. These complimentary arrays of b- and y-ions are used to deduce peptide structure as given in Figure 1.11.

An experimentally obtained MS/MS spectrum of a peptide is far from ideal. Fragment ions occur in different abundances and there are more than b- and y-ions present. The relative abundances of MS/MS ions are a function of the frequency of the fragmentation reaction that creates a given fragment ion, including subsequent fragmentation reactions. Abundances range

widely and some products are not observed at all. Since only a finite amount of ions are in the mass spectrometer, fragmentation reactions at one amide bond compliment a reduction of fragmentation reactions at another amide bond. The wide variation in fragment ion intensities is dictated by the strength of amide bonds in the peptide backbone resulting from side chain chemistry.

Low energy CID also creates species other than b- and y-ions. Some of these ions are formed by the loss of small neutrals from b- and y-ions. For example, a decrease of 17Da corresponds to a loss of ammonia, a decrease of 18Da corresponds to a loss of water, and a decrease of 28Da corresponds to a loss of carbon monoxide. Another common and useful species are immonium ions ($H_2N\text{-}CHR^+$) which occur on the low mass end of the spectrum. Immonium ions are characteristic of tryptic peptides.

Tryptic peptides ionized by ESI and fragmented by CID dissociate with one major pathway and many minor ones. These fragments lead to product ions that are recorded as ion spectra and used to deduce peptide amino acid sequences. While many types of ions contribute information, b- and y- ions are the prime source of information in this technique. The mass difference between each consecutive member of an ion series corresponds to an amino acid residue mass, and those residue masses are used to elucidate peptide sequence.

Protein Identification

There have been several methods used since the 1950's to determine the amino acid sequence of proteins. Edman degradation, developed in 1950, was the first widely used form of protein sequencing. The process involves chemically labeling and then removing the N-terminal amino acid of a peptide.[29] This procedure is quite accurate, but only on peptides up to ~50 amino

acid residues. Eventually Edman sequencing was automated, increasing its usefulness greatly, but mass spectrometry based analyses eventually superseded it. The advent of 2-D electrophoresis gels (1975), ESI (1980's), and peptide mass fingerprinting (1993) pushed mass spectrometry based protein sequencing into the forefront.[30]

An early method of mass spectrometry based protein identification is peptide mass fingerprinting (PMF). In PMF, individual or very simple mixtures of proteins are enzymatically digested and the proteolytic fragments are analyzed by mass spectrometry. 2D gel electrophoresis can allow for individual proteins to be isolated from complex lysates. With a protein database, derived from a sequenced genome, observed fragment masses can be compared to the database derived theoretical fragment masses. This method works well but is very reliant on a few criteria[31]:

1) The peptide masses are representative of the sample protein
2) The sample is at most two proteins, ideally only one
3) The protein is known
4) The peptide masses are known with <20ppm mass accuracy

The 2nd point is frequently a limiting factor of PMF in that individually analyzing each individual gel spot of a 2D gel can be very time consuming. Another method of protein sequence determination involves the dissociation of peptides into smaller fragments. The method of sequence determination from continuously fragmenting a parent ion is called tandem mass spectrometry, or MS/MS.

Protein Identification using Tandem Mass Spectrometry

Tandem mass spectrometry is a widely used tool for protein identification today. MS/MS protein identification has proven to be more robust than PMF, but MS/MS has its own limitation. Just as simultaneously analyzing more than one protein in PMF is detrimental, co-eluting peptides will badly obscure the sequence information in an MS/MS spectra. Some sequence ions are not observed at all and due to side reactions and rearrangements, there are frequently non b and y type ions present. The sequence information gleaned from b and y-ion series resulting from CID is commonly used in four different ways to identify proteins[32]:

1) Database searching
2) De novo sequencing
3) Peptide sequence tagging
4) Consensus database searching

Database Searching

MS/MS database search programs tend to work in a similar manner. Theoretical peptide masses are derived from protein databases, and sets of these peptides within a specified mass range of experimentally observed peptides are fragmented in-silico according to defined cleavage rules. In-silico enzymatic digests are compared to experimental spectra, and an alignment value (ion score) is given to represent the quality of the match. The most important characteristics of the comparison between experimental and virtual data are how close the peptide masses are in agreement and the number and intensity of shared fragment ion peaks. Database searching requires prior knowledge of all possible protein matches that could occur when analyzing a sample. This set of prior knowledge is derived from the genome of the

sample. Unfortunately, not all matches are correct. Probability based scoring algorithms attempt to reflect the probability of match occurring randomly. This is a major challenge in proteomics that will be addressed later.

There are many different search algorithms currently used in proteomics, some free and some commercial. Three very frequently used search engines are Sequest, Mascot, and X!Tandem.

Sequest uses as cross correlation function to evaluate matching between an experimental spectrum and a database peptide sequence. First a theoretical spectrum is calculated from a protein database. The alignment of the peaks shared between the theoretical spectrum and the experimental spectrum is displaced by τ over the range $-75 < \tau < 75$. The shift required for the best correlation is noted as $f(\tau)$. The final score for each peptide is equal to $f(0)$ minus the mean of $f(\tau)$. A large difference between the 1st and 2nd highest score suggests a true positive match.

Mascot is a probability based search algorithm that uses a modified version of MOWSE scoring. Each match between fragment ions of a database peptide entry and peaks of the experimental spectrum are regarded as random events. For each peptide match, the probability of the match being random is calculated. True peptide matches will therefore have a very small probability. As a convenience, Mascot converts the probability to an ion score where larger values are more desirable (Ion Score = -10log(Probability). The proteins in a target database are not random so the ion score is only a measure of how significant a match is, not whether or not is correct. As with Sequest, it is prudent to observe the difference between scores the 1st and 2nd best matches.

X!Tandem is a C++ based program that is free to download and use. Like Mascot and Sequest, X!Tandem searches for matched MS/MS ions, but there is a major difference in how the

algorithm works. X!Tandem breaks a search into two steps. The first step is a survey search that makes several confining assumptions about peptides and quickly identifies a set of protein sequences that are plausible candidates. These candidates are then refined with a more time consuming but more accurate scoring function that considers many characteristics of an MS/MS spectrum including incomplete enzyme cleavage, non-specific cleavages, and chemical modifications of amino acids. Since the survey step restricts the pool of peptides for later refinement, a faster result is obtained.[32]

Concatenated Databases

The use of concatenated databases offers an alternative to performing separate forward and reverse database searches. A concatenated database is a large (twice the size of normal) database containing both forward and reverse sequences. Two assumptions are presumed when using concatenated databases: (1) No true positive matches will come from both target and decoy, and (2) if false positives are equally likely to come from target and decoy sequences, then you can estimate the number of false positives that meet score cutoffs by doubling the number of selected decoy matches.

It is critical that no peptides exist in both the target and decoy portions of a concatenated database. The larger a peptide is the less likely for it to appear in both target and decoy sequences. As the length of a peptide increases, the chance for a tryptic peptide to appear in both target and decoy entries drops to 0.02% for peptides with eight amino acids. Tryptic peptides are typically greater than nine amino acids in length so it is highly unlikely for them to appear in both halves of a concatenated database. Additionally, small peptides are even less likely due to

frequently being singly charged, RPLC incompatible due to polarity, and weak in fragmentation potential.[33]

For concatenated databases, the total number of false positives that pass a given threshold is estimated by doubling the total number of decoy hits. This is only appropriate if there is an equal likelihood of selecting an incorrect match from the target and decoy portion of the concatenated database. The database search needs to be presented with equal numbers of target and decoy peptides because the number of possible incorrect peptides should be equally distributed between the target/decoy portions of the database. As a result, incorrect peptide identifications are equally selected from target and decoy sequences. Second ranked (typically incorrect) and lower ranked matches are distributed equally between target and decoy sequences showing that both assumptions are sound.[33]

One search on a concatenated database gives higher quality data compared to the two-search target and decoy method for a few reasons. In concatenated database searches, target and decoy sequences are forced to compete for the top score in a single search. Decoy sequences that partially match high quality MS/MS spectra may receive higher scores compared to other top-ranked hits from a forward search. Separate searches obstruct estimations of low-scoring correct identifications in the presence of high-scoring incorrect identifications. With one search, high scoring decoys won't be able to out-compete lower scoring correct identifications. Separate searches also force the assumption that app peptide assignments are incorrect below a score at which decoy hits outnumber targets leading to overestimated false positive rates.[33]

De Novo Sequencing

Database searching is only an option of a protein database is available and accurate. If no database exists, then protein identifications derived from MS/MS spectra must obtained using de novo sequencing. De novo searching algorithms calculate potential amino acid combinations that would best produce the b- and y-ion series observed in an experimental MS/MS spectrum. A scoring algorithm is needed to compute the best peptide due to the vast number of possible amino acid combinations. There are simply too many combinations to compute.

De novo sequencing has some difficulties that are minimized in database searching route of protein sequencing. Different amino acid combinations may have similar or identical masses, meaning differentiating between the correct amino acid sequence and one that is plausible according to the information in the MS/MS spectrum is quite a challenge. Also, the absence of peaks in the b- and y-ion series, and missed enzymatic cleavages present difficult challenge for de novo sequencing.

Peptide Sequence Tagging

Peptide sequence tagging (PST) determines protein sequences by searching databases with partial sequence information derived from MS/MS spectra. These partial sequences are determined from the MS/MS spectra by a de novo sequencing program. Amongst the partial sequences are mass values that represent the unknown amino acid combinations. The unknowns are due to missing information in the MS/MS spectra. The tags are then used to select peptides and proteins in a database and form a match despite the unknowns. A benefit of this approach of protein identification is that even if the correct protein match is not in the database, a homologue likely is and will be identified.

Consensus Database Searching

Manually removing all false positives from a database search is difficult and time consuming, and false negatives cause low coverage and identification confidence. With the availability of several search engines that are free, it is good practice to use many different search engines and compile the results. This will result in fewer false positives, better sequence coverage and higher confidence in search results. The idea of using different search engines is analogous to several independent interpretations of the same data. If two search engines make the same protein identification, then that identification is likely a true positive.

Sequest

Sequest was the first MS/MS algorithm capable of identifying peptides. The software was optimized for QToF and ion trap mass spectrometers. The software first uses a preliminary scoring algorithm to select the 500 best candidate peptide sequences for cross correlation. There are many scores and rankings for each candidate peptide. The preliminary score, S_p, is defined as:

$$S_p = (\Sigma i_m)(1+\beta)(1+p)\frac{n_i}{n_t} \quad (18)$$

where (Σi_m) is the sum of intensities of matched ions, ($1+\beta$) is the score of sequential ion series, ($1+p$) is the score of matched immonium ions, n_i is the number of matched ions, and n_t is the total number of ions. Data reduction is then performed and the 200 most intense peaks are selected and normalized to 100% intensity and compared to a theoretical spectrum.

The cross-correlation score (XCorr) is the primary score component of a Sequest search. It represents the alignment between a candidate peptide sequence and an experimental spectrum.

The XCorr is derived by reducing the MS/MS data into equal segments and normalizing each segment to 50. The reduced spectrum is then molded to look like an experimental spectrum. This means that y- and b-ions are normalized to 50 while a-ions and neutral losses are normalized to 10. Fast fourier transform are then used to compare a simulated spectrum with the modified experimental one. A large difference between the XCorr of the first and second match suggests that the first match is correct. This difference is noted by Sequest as the ΔCn score.

Sequest is a widely used search algorithm that is capable of finding many true matches. Unfortunately, Sequest also finds many false positives, but many validation algorithms have been developed assisting researchers in removing these false positive matches.[31, 34]

X!Tandem

X!Tandem is a free open source search program developed by Ron Beavis in 2004. The program is designed to be fast and run on modest PC's. X!Tandem's speed results from the program's approach to searching data. It is assumed that there exists at least one detectable tryptic peptide with zero or one missed cleavage for each identifiable protein in the sample mixture. A preliminary search is performed using this assumption and a list of protein matches smaller than the total database is taken to be a refined database. Experimental spectra are then stringently searched against this refined database where PTM's and non-specific hydrolysis is considered.

X!Tandem scoring is based on a dot product of the theoretical b- and y-ions and experimental MS/MS spectra. A score is then converted to an expectation value, the E-value. An E-value is the number of peptides in the target database expected to achieve that score by chance, meaning a lower E-value is more significant. The E-value is obtained by collecting statistics during a

search to estimate the distribution of scores for random and false identifications. The distribution of scores is hypergeometric. Figure 1.12 depicts a typical X!Tandem histogram. Outliers far from the main distribution of scores represent significant peptide matches. Dot product scores are then converted to a hyperscore by multiplying the score by the factorial number of matching b- and y-ions. The right half of the histogram is then log transformed as depicted in Figure 1.13. Scores that are higher than the intersection with zero are assumed significant and can have their E-values determined by extrapolation (E-value = $e^{-(\text{y-axis value})}$).

Mascot

Mascot is commercial proteomics software made by Matrix Science and is commonly considered the industry standard for database search software. Mascot supports data formats from all commercial vendors of mass spectrometers, has multithread and cluster versions, and has an excellent automation program called Daemon. The search mechanism that Mascot uses is to iteratively select subsets of the most intense peaks with the goal of finding subset which most clearly differentiates a top scoring peptide from the rest of the candidate matches. The different ion series and charge states in the matching process are tested independently and in combination in order to achieve the highest score. Mascot is able to search multiple charge states which has the benefits being good for low resolution data and not contributing to an increase in false positive matches. For each spectrum match, a probability base MOWSE score is calculated and converted to an ion score.

A MOWSE (molecular weight search) score is derived after several manipulations to a mass spectrum. First the proteins in the target database are grouped into 10kDa bins. Then for each protein, tryptic fragments are put into 100Da bins. The number of fragments in each bin is then

divided by the total number of fragments for each 10kDa protein interval, giving a frequency score. For each 10kDa interval, this number is normalized to the largest bin value. Spectrum masses are then compared to the mass list for each protein in the database. The frequency scores for each protein match are then retrieved and multiplied together giving P_N, the product of distribution frequency scores. This is used to determine the final MOWSE score:

$$Score = \frac{50{,}000}{P_N H} \quad (19)$$

where 50,000 is the average protein molecular weight, and H is the hit protein's molecular weight.[35]

The ion score reported by Mascot is a manipulation of the probability of the MOWSE-based scoring. The probability, P, that a significant match is a random event is small so the conversion:

$$Ion\ Score = -10\log(P) \quad (20)$$

conveniently changes a desirable small probability into a much larger number. Mascot is capable of assigning a score cutoff as a significance threshold. Calculating the threshold requires the size of the database and a percent chance that a match is random. The default percent used by Mascot is 5%. The relationship between the expectation value (5%) and the number of proteins in the database is:

$$P = E \times N^{-1} \quad (21)$$

E is the user specified expectation value (5% by default) and *N* is the number of entries in the target database. An ion score can then be calculated using *P*, and that is the significance threshold.[36]

Statistical Validation of Proteomic Data

Probability calculations are needed to account for the quickly improving throughput of mass spectrometry based proteomics. Mass spectrometers are capable of recording tens of thousands of spectra in one day of LC-MS experiment resulting in nearly as many database search matches. Manual validation of thousands of matches in a database search is not feasible. An automated way of distinguishing between correct and false identifications is necessary for datasets of significant size.

Mass spectral database search algorithms are potent tools in protein identification. Programs such as SEQUEST, Mascot, and X!Tandem are indispensible for analyzing the considerable amount of data modern mass spectrometers can generate. Unfortunately, when these search programs are tasked with identifying peptides from experimental spectra, there is much overlap between the ion scores values for correct and incorrect peptide identifications. One of the first methods of validating data was score cut-off filtering. There are several problems with this method: Each investigator will use a different threshold, each type of mass spectrometer will require a different threshold, each search algorithm returns different types of scores, and users can not assign an error rate.

In order to publish large amounts of shotgun proteomics data today, it has become common place to perform statistical validation of experimental datasets. Two of the most widely used methods are probability and false discovery rate validation. Assigning peptide/protein

probabilities and false discovery rates requires special kinds of database searches representing the null hypothesis. The rate of occurrence of null solutions is used to assign dataset validity.[37-40]

In order to apply statistical filtering to mass spectrometry-based proteomic data sets, target and decoy database searches must be performed. Decoy database searches serve as the null hypothesis for statistical analysis. This means a peptide identification in a decoy database represents the null hypothesis being correct, an undesirable event[41]. Decoy databases are typically formed by reversing, randomizing or shuffling target database entries. Typically, reverse databases are used because this will generate a decoy database that has the same amino acid composition and tryptic peptide lengths.[37] The null test can be performed by using two searches, one target and one decoy, or one search against a larger combined forward and reverse database dubbed a concatenated database.[33]

p-Value

A null distribution can be estimated for a particular matched spectrum by plotting a histogram of all scores for non-first ranked matches. Presumably, non-first ranked scores are incorrect and represent the null hypothesis. Matches that are furthest from the core distribution of scores from non-first matches are deemed more significant. This is essentially what the *p*-value is, where lower scores represent more believable peptide matches. Unfortunately with enough spectra searched, even random matches can have low *p*-values. Analysis of score histograms can be used to implement *p*-value filtering since histograms of peptide matches made against target databases are biased towards higher scores, reflecting scores corresponding to correct matches.

Hypothetical score histograms of proteomic data appear as in Figure 1.14. The curve shape between target and decoy distributions is very similar and only differs in that the target distribution has a heavier tail towards higher scores. Again, this represents the bias towards correct matches having higher scores. Using *p*-values to filter data requires using the ratio of target and decoy matches at a particular score filter. For a given peptide match, its *p*-value defined as

$$p = \frac{number\ of\ decoy\ matches\ at\ match's\ score\ or\ higher}{total\ number\ of\ decoy\ matches} \quad (22)$$

At a *p*-value of 0.01, there is a 1% chance that a decoy spectrum match was called a correct match.

Expectation values (*E*-values) are used more than *p*-values in data set validation. An expectation value represents the number of peptides with scores equal to or better than an observed score assuming that peptides are matching the experimental spectrum by chance. X!Tandem is an example of a search algorithm who's primary score parameter is a log of *E*-values.[39]

False Discovery Rates

The use of *p*-values is inadequate with large datasets because the test is performed so many times. For example, at a *p*-value of 0.01, if 4,000 target peptides were found amongst 35,000 spectra, then 350 of the peptides were found by chance. Multiple testing corrections are needed with so many statistical tests are performed. FDR's are a widely accepted method for validating data under these circumstances.

FDR's are frequently and erroneously referred to as false positive rates (FPR). An FDR associated with a score threshold is the expected percentage of accepted peptide matches that are incorrect. Accepted peptide matches are the subset that score above this threshold. The false positive rate is the rate of "true null tests" that are called significant.[39, 41] For example, at a 5% FPR, 5% of null results will be called significant. FDR's are used to gauge the quality of entire datasets; they can't be used to validate spectrum matches, unlike probabilities. For example, if a 5% FDR search yields 1000 matches, then 50 of them are expected to be incorrect.[42] However, the FDR validation does designate which matches are incorrect.

The calculation for FDR's is similar that for *p*-values. For a given threshold, sum the number of decoy and target peptide matches above the threshold and compute the ratio of the values. For example, with an XCorr cutoff of 3.0 in a hypothetical SEQUEST search, if 4000 target matches and 250 decoy matches meet the threshold, then the dataset has a 6.25% FDR.

Searching against decoy databases is essential for establishing statistical filtering criteria. Decoy peptide matches, by design, are incorrect, but not all target peptide matches are necessarily correct. There may be some portion of target matches that were random but counted as correct. These incorrect target matches need to be factored into FDR calculations. The distributions of target and decoy matches are similar except target matches have a heavier tail to the right due to a bias of target matches having better scores. Target matches are a mixture of correct and incorrect peptides as show in Figure 1.15.

The mean for incorrect matches is 1.0, and 3.0 for correct matches. At XCorr 1.0, there are 650 target matches compared to 800 decoy matches. This means that the percentage of incorrect targets (PIT) for the example data is 81.2%.[41] The PIT allows for a reduction of estimated

FDR associated with a given group of target matches. In this example, if a certain number, *x*, of decoy matches are chosen at a specified threshold, then 0.812*x* incorrect target matches are expected. Most of the incorrect target matches should occur at the low scoring region of the histogram, and this is a conservative estimate because there will very few correct matches a low scores. This method should not be applied to high scoring matches as nearly all of them will be correct matches.

Statistical Filtering Using ProValT

ProValT is a software algorithm that calculates FDR's for proteins using Mascot, Sequest, and/or X!Tandem peptides. ProValT extracts peptide matches from search results, eliminates redundancy, and then clusters peptides to corresponding proteins. Homologous proteins are put into groups since finding one tryptic peptide will not allow the differentiation of two similar proteins if they both contain the observed tryptic peptide. A protein FDR uses random databases and peptide probabilities to calculate expected protein FDR's for a minimal number of expressed proteins identified by matched peptides. Random databases represent the null hypothesis: Matches from the random database are considered random, and scores from matches follow a quasi-normal distribution with a false positive rate related to each score. ProValT compares score distributions from searching normal and random databases to calculate FDR's for each specified score threshold. The overall goal is for the threshold to maximize true positive matches while minimizing false positive matches.

The ProValT algorithm has an extensive workflow that organizes peptides matched by search programs. ProValT first extracts all matched peptides and corresponding scores from forward and reverse database searches. The results are combined and filtered generating a non-redundant

list of peptides. Peptides are grouped as a function of score into bins (B_i) containing all peptides equal to or exceeding each search algorithm score (i). The score, i, embodies the ion score, S, assigned by the search program used. Each ion score ranges from M to N where

$$M = \text{Min}(S|\text{n}PEP(\text{S}) > \text{r}PEP(S)) \quad (23)$$

$$N = \text{Min}(S|\text{r}PEP(S) = 0) \quad (24)$$

$nPEP(S)$ = number of peptides in normal database $\geq S$

$rPEP(S)$ = number of peptides in random database $\geq S$

The peptides in each score bin are clustered to their corresponding protein. Proteins are then selected based on their degree of peptide coverage, c, where

$$c = (C, C - 1,\ldots,1) \quad (25)$$

and C is the user defined maximum peptide coverage. Starting with C, a histogram is created based on the frequency of protein identifications within each bin for matches in the normal and random databases. FDR's are calculated for each protein identification in a score bin.

$$Protein\ FDR_C(S) = \left[\frac{rPRO_C(S)}{nPRO_C(S)}\right] \times 100\% \quad (26)$$

where $nPRO_C(S)$ is the number of proteins identified in the normal database with sequence coverage c in the bin containing S, and $rPRO_C(S)$ represents the same for proteins identified in the random database. ProValT then determines the threshold S_c for when $c = C$, where S_c is given by

$$S_C = Min\big(S|Protein\ FDR_C(S)\big) \leq Max\ Protein\ FDR \quad (27)$$

for all *S* between *M* and *N*. *Max Protein FDR* is the user defined maximum protein false discovery rate. The minimum ion score threshold necessary to achieve specified Protein FDR's given by peptide coverage, c, is thus found. Peptides that meet this criteria are stored while remaining peptides not matched to proteins are grouped as a function of ion score. New bins, B_i, are formed who contain all peptides equal to or exceeding ion scores, *i*. For the next degree of coverage ($C = C - 1$), distribution of ion scores will have a minimum, *M*, dependent on score threshold, S_c, which is determined for the previous degree of peptide coverage in the following manner

$$M = S_{c+1} \text{ for } c < C \tag{28}$$

$$N = \text{Min}(S|rPEP(S) = 0) \tag{29}$$

$$rPEP(S) = \text{number of peptides in random database} \geq S$$

Previously unmatched peptides in each bin are clustered to corresponding proteins along with peptides matched at previous iterations. For the next degree of coverage ($C - 1$), another histogram is formed, and protein FDR is calculated for each score bin. The score threshold for $c = C$ -1 is calculated, and peptides meeting the criteria are stored. This process is repeated until $c = 1$.

The overall goal for ProValT is to separate valid identifications from incorrect ones. An FDR finds the proportion of random matches among all peptide matches deemed significant. A peptide FDR provides no information about the error rate of a specific protein identification. For this statistical modeling to work well, large data sets on the order of hundreds of proteins are required. FDR models are based on the assumption that null hypothesis follow normal distributions with minimal ion scores. Some distributions may not appear normal, but they

approach this as the data set size increases. Manual identification is possible for data sets falling short of this range.[43]

Probability Calculations Using ProteoIQ

The software program ProteoIQ uses the ProValT algorithm to calculate protein and peptide FDR's. ProteoIQ can also calculate probabilities by using PeptideProphet, a statistical model made by Andrew Keller in 2002.[44] Probability calculations were needed to account of the quickly improving throughput of mass spectrometry based proteomics. Mass spectrometers became capable of recording tens of thousands of spectra in one day of LC-MS experiment resulting in nearly as many database search matches. Manual validation of thousands of matches in a database search is not feasible. An automated way of distinguishing between correct and false identifications is necessary for datasets of significant size.

One of the first methods of validating data was score cut-off filtering. There are several problems with this method: Each investigator will use a different threshold, each type of mass spectrometer will require a different threshold, each search algorithm returns different types of scores, and users can not assign an error rate. Peptide/protein probability and FDR filtering are currently the most commonly used methods. PeptideProphet is a robust and accurate model for determining the quality of peptide identifications from MS/MS spectra. Each peptide match, including false positives, is compared to every other match made in the database search. Using the ion scores assigned by a search algorithm, PeptideProphet is able to differentiate correct from incorrect matches and calculate peptide probability.

PeptideProphet uses discriminant function analysis to combine database search scores. This method is a strong statistical model for validating peptide matches and distinguishing between

true and false peptide matches using spectral information. Each different search algorithm outputs different scores grading experimental and theoretical spectral alignment. Bayes' Theorem is useful for calculating peptide correctness. The probability of scores x_1, x_2, and x_3 being correct can be modeled as in Equation 30

$$p(+|x_1, x_2, \dots, x_s) = \frac{p(x_1, x_2, \dots, x_s|+)p(+)}{p(x_1, x_2, \dots, x_s|+)p(+) + p(x_1, x_2, \dots, x_s|-)p(-)} \quad (30)$$

where $p(x_1, x_2, \dots, x_s|+)$ is the probability of scores $x_1, x_2, \dots, x_s$ among true positive peptide matches and $p(x_1, x_2, \dots, x_s|-)$ is the probability of scores $x_1, x_2, \dots, x_s$ among false positive peptide matches. The quantities $p(+)$ and $p(-)$ are the prior probabilities of previous true positive and false positive matches respectively. Prior probabilities are the overall proportion of true positive and false positive peptide matches in the dataset. Keller et al chose to compute probabilities by finding joint probability distributions for false positive and true positive matches using a standardized dataset of known quality. A discriminant function analysis combines database search scores $x_1, x_2, \dots, x_s$ into a single discriminant score that best differentiates the training data into true and false positives. The discriminant score, F, is a weighted combination of database search scores:

$$F(x_1, x_2, \dots x_s) = c_0 + \sum_{i=1}^{s} c_i x_i \quad (31)$$

where c_0 is a constant and c_i is a weighing factor. The term c_i is derived such that the "ratio of between-class variation to within-class variation is maximized under the assumption of multivariate normality."[44] Deriving a discriminant function from these scores requires data points from a standardized dataset. Discriminant scores from the standard dataset can be substituted into Equation 31 instead of the original database scores to allow a manageable

determination of probabilities that have just as much discriminating capability while using a single weighted combination of ion scores. Using discriminant scores, *F*, Equation 30 becomes

$$p(+|F) = \frac{p(F|+)p(+)}{p(F|+)p(+) + p(F|-)p(-)} \quad (32)$$

where $p(+|F)$ is the probability that a peptide assigned with discriminant score *F* is correct. The quantities $p(F|+)$ and $p(F|-)$ are the probabilities of *F* according to the discriminant score distributions for true and false positive matches.

The calculations of peptide probabilities require an equation to model the distribution of discriminant scores. If discriminant score is plotted versus the number of spectra, a Gaussian model can be derived for correct peptide matches from the plot.

$$p(F|+) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(F-\mu)^2/2\sigma^2} \quad (33)$$

In Equation 33, μ is the calculated mean and σ is standard deviation. For false matches, a gamma distribution yields

$$p(F|-) = \frac{(F-\gamma)^{\alpha-1} \mathrm{e}^{-(F-\gamma)/\beta}}{\beta^{\alpha}\Gamma(\alpha)} \quad (34)$$

where γ is set as less than the minimum value for F, and the contants α and β are calculated using the method of moments. Substituting Equations 33 and 34 into Equation 32 with calculated prior probabilities allows the calculation of accurate peptide probabilities.

High-Resolution Mass Spectrometry

High resolution scanning during mass spectrometry experiments is highly desirable though it does have its downfalls. It has been employed more and more frequently as high resolution

instrumentation modernizes. The newest generations of mass spectrometers used in proteomics have the sensitivity nearly that of ion traps, and the mass accuracy and resolution of QToF mass spectrometers. High mass accuracy and resolution have the boons of isotopic resolution, monoisotopic peak determination, charge state determination, and with enough mass accuracy, elemental composition determination. However, no matter how advanced the instrumentation, higher mass accuracy and resolution scans require more time and this leads to losses in sensitivity.

Very complex peptide mixtures are routinely analyzed in proteomics. These samples can contain tens of thousands of unique peptides covering a large dynamic range. For lower resolution instruments, well below 100,000 FWHH, coeluting peptides with similar *m*/*z* will overlap. As a result, accurate mass measurements and charge state determination are not possible. The old choice of instruments for mass spectrometry based proteomics was between the ion trap and the QToF. Traps offered ~300 resolution but high scan speeds and sensitivity. ToF's offered 10,000 but less sensitivity and robustness. The advent of the FT-ICR mass spectrometer made high resolution capabilities widely available to the field. Higher mass accuracy acts as a filter that reduces the number of "false peptides" chosen for fragmentation. At 1ppm precursor mass accuracy, 99% of amino acid compositions for a given nominal mass are ruled out.[45, 46] This in turn leads to higher database search scores. LTQ-FTICR instruments use low ppm precursor mass accuracy compared to the several Da (1000's of ppm) mass accuracy of ion traps; however, MS/MS is still typically done at unit resolution in the LTQ. The combination of MS1 in the FT and MS2 in the trap has the most favorable duty cycle leading to the most CID spectra collected.

Orbitrap mass spectrometers enable researches to obtain high resolution MS/MS spectra while limiting the loss of MS/MS spectra due to longer scan times. Precursor ions can be fragmented during ion transfer through the C-trap to the orbitrap. Ion *m*/*z* scans are performed in the orbitrap to take advantage of the high mass accuracy and resolution of the orbitrap. Despite the popularity of QToF's, FT-ICR's, and orbitraps, most database search programs don't take advantage of high accuracy and resolution MS/MS spectra. This is another factor leading to precursor masses typically being scanned in FT-ICR's or orbitraps while fragment spectra are taken in ion traps.

Fragment scanning performed in an ion trap offers comparable performance to fragment scans performed in the orbitrap. Ion injection times and duty cycle are longer using C-trap fragmentation and orbitrap scanning compared to ion trap fragmentation and *m*/*z* scanning. This results in a lower number of peptide identifications during database searching. However, the orbitrap can perform fragment scans at ~7500 resolution which is sufficient for accurate charge state determination and accurate mass measurement. Higher quality MS/MS spectra lead to better specificity in protein identifications from database searching, somewhat offsetting the fewer overall protein identifications found from ion trap MS/MS data.[47] LTQ-FTICR instruments, with a duty cycle larger than one second, scan too slowly to outperform orbitrap high resolution MS/MS scanning.

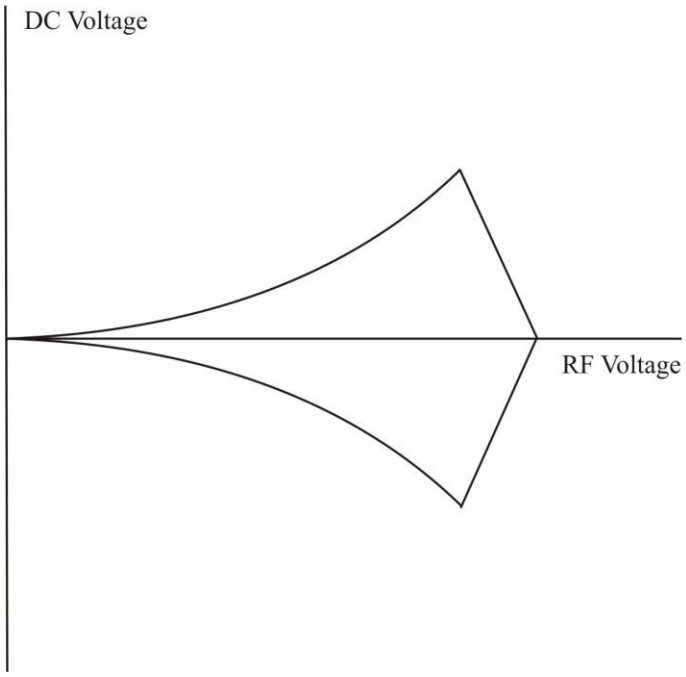

Figure 1.1 Mathieu Solutions for a Quadrupole

An ion will only be stable in the trap if the correct combination of RF and DC voltages are applied. Values within the enclosed area will provide stable ion trajectories.

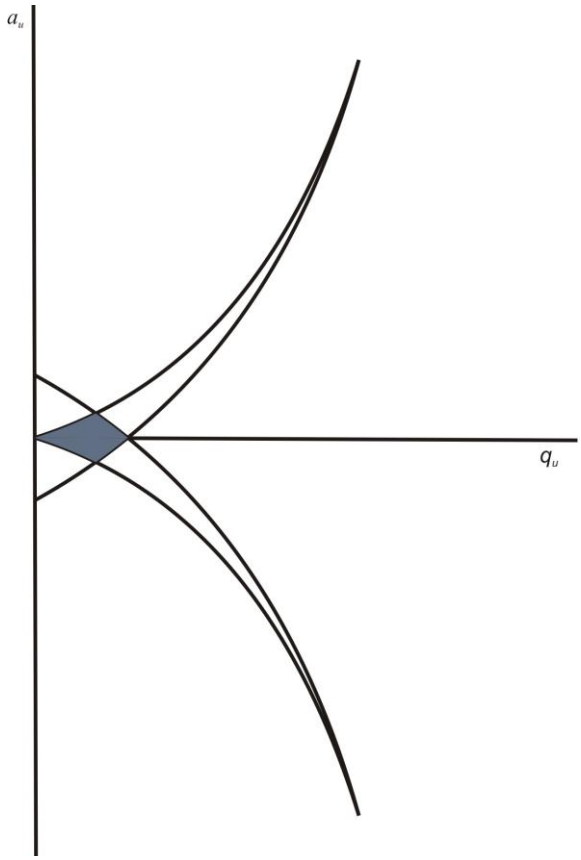

Figure 1.2 Ion Stability Diagram

The ideal charges to apply to the ion quadrupole electrodes are found by determining the overlap in the *x* and *y* stability curves. The shaded area represents the area plotted in Figure 1.

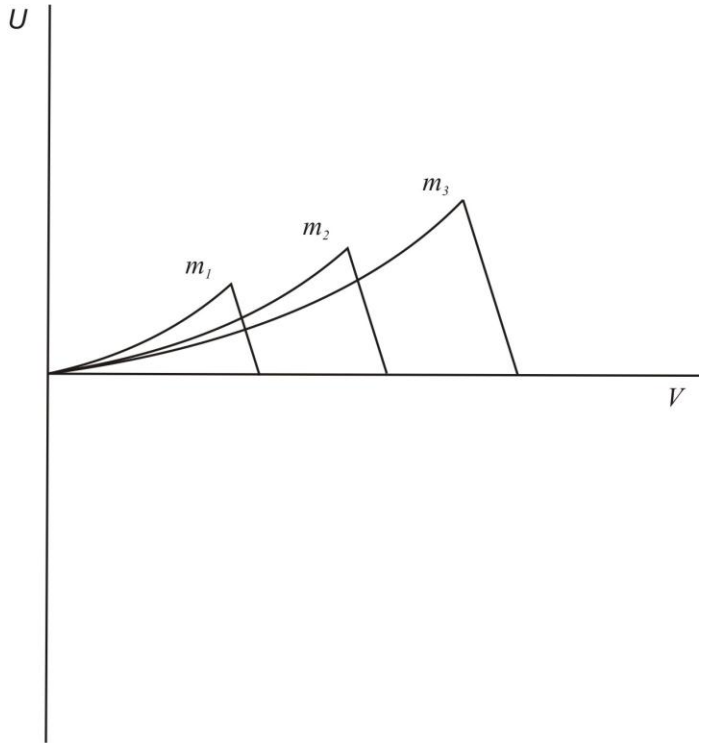

Figure 1.3 Matheiu Solutions for Three Different Ions

Ions of different *m*/*z* have differing stability diagrams. Ejecting ions requires increasing the DC and RF until their increasing oscillations force them into one of the quadrupoles.

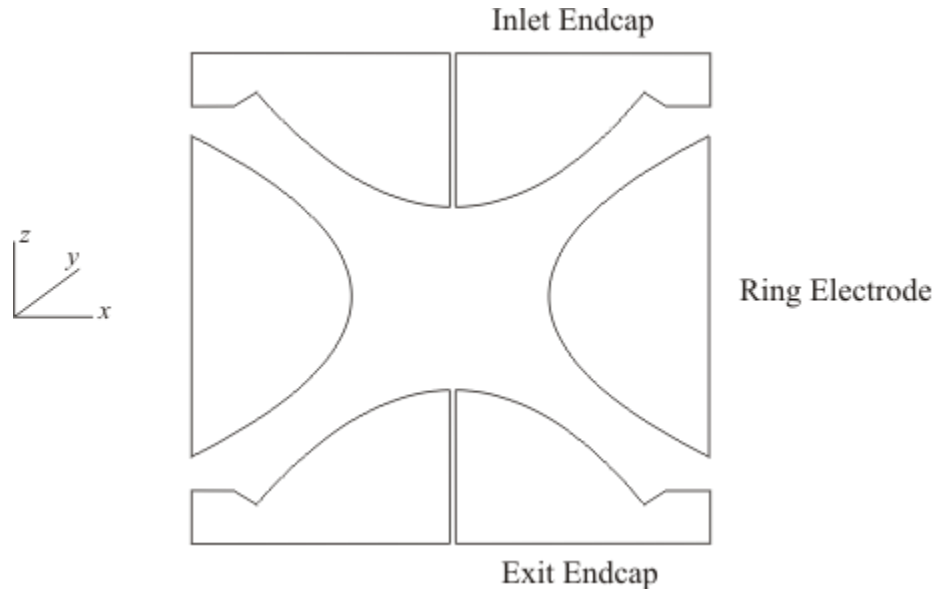

Figure 1.4 Cross Section of an Ion Trap

The ion trap consists of a ring electrode two endcap electrodes. Ions enter and are ejected along the *z*-axis and revolve in the *xy* plane.

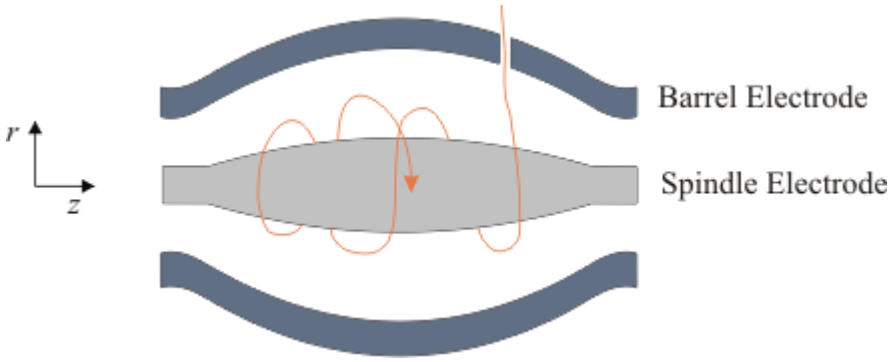

Figure 1.5 Cross section of an Orbitrap

An orbitrap contains a spindle shaped electrode within a barrel shaped electrode. Ions rotate around the spindle electrode and *z*-axis oscillation frequencies are used to determine *m*/*z*.

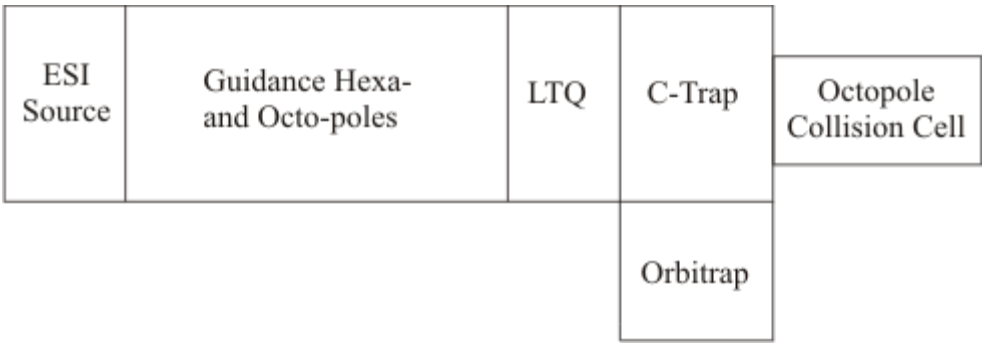

Figure 1.6 Conceptual Diagram of the Thermo LTQ Orbitrap XL

Ions from the ESI source are guided by mutli-poles to the LTQ. Ions can then be sent to the Orbitrap via the C-Trap for high mass accuracy and resolution scanning. Several fragmentation methods are available with this instrument.

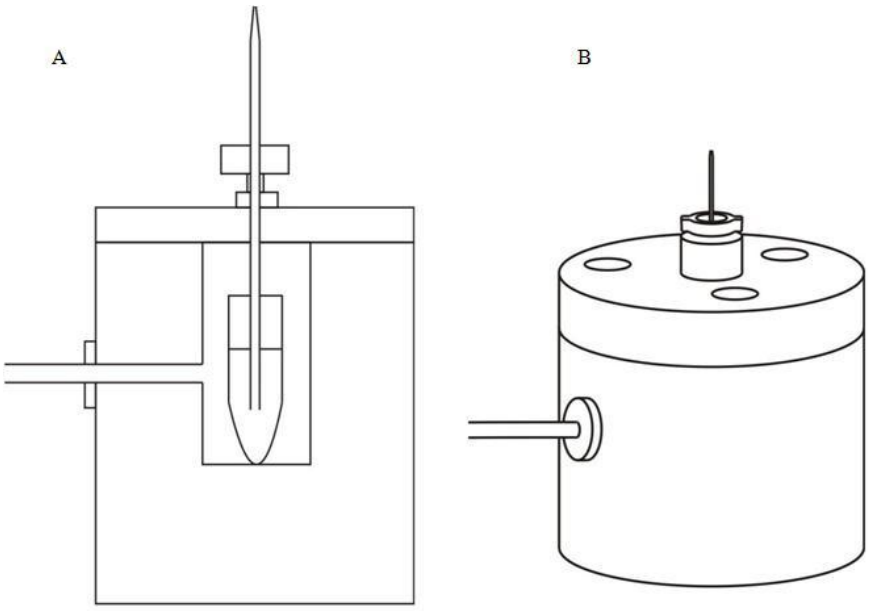

Figure 1.7 High Pressure Bomb Diagrams

Figure 7A shows a cross section of a bomb during sample loading. High pressure forces the solution within a vial up through the capillary column. Figure 7B shows an orthogonal exterior view of a bomb.

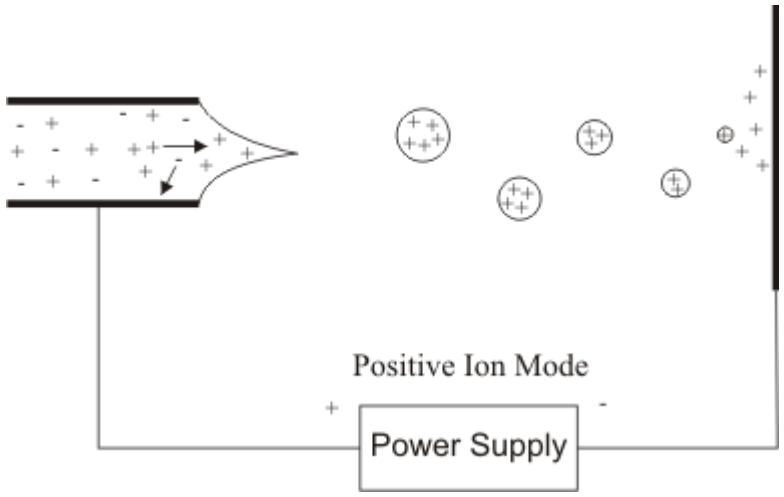

Figure 1.8 Electrospray Ionzation in Positive Ion Mode

Charged solvent and analyte spray out the tip of a capillary column into an electric field. Coulombic repulsion reduces the size of droplets until unsolvated ions enter the mass spectrometer.

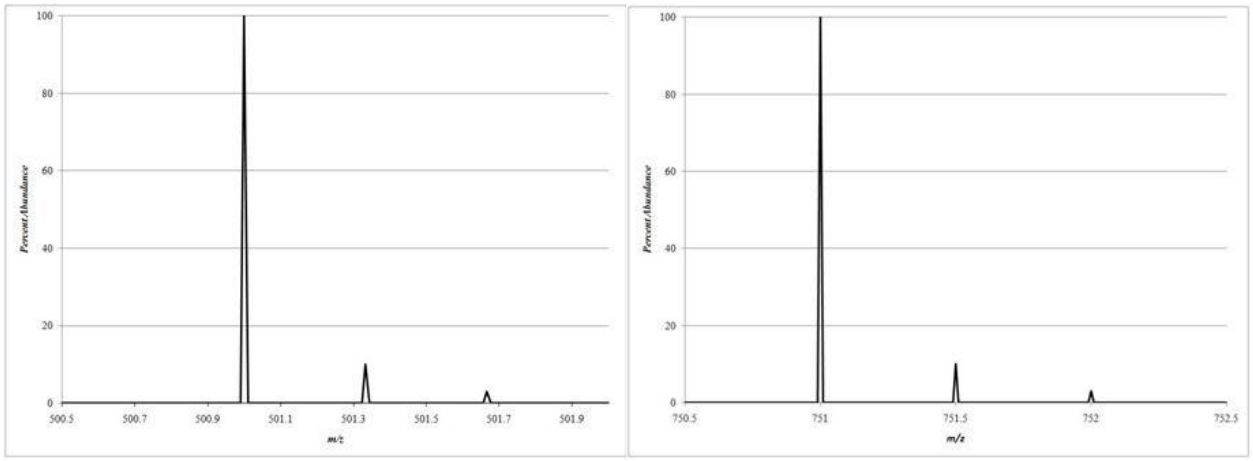

Figure 1.9 Isotopic Envelopes of Multiply Charged Ions

Multiply charged peptides will have spacing between the ^{12}C, $^{13}C1$, and $^{13}C2$ that can be used to determine the charge state of the ion.

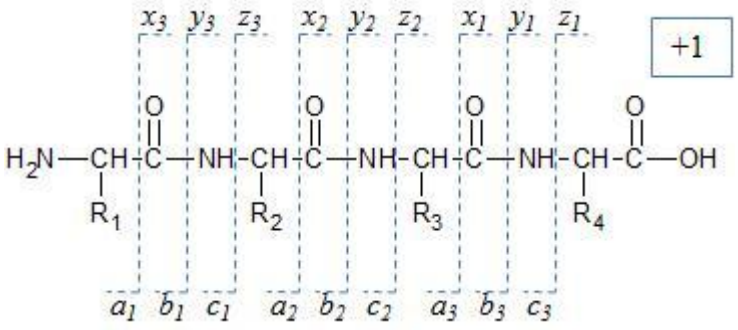

Figure 1.10 Fragment Ion Types Generated from CID

Several types of fragment ions can form with *a*, *b*, and *c* ions complementary to *x*, *y*, and *z* ions. The difference between the two groups is determined by where the proton charge lies.

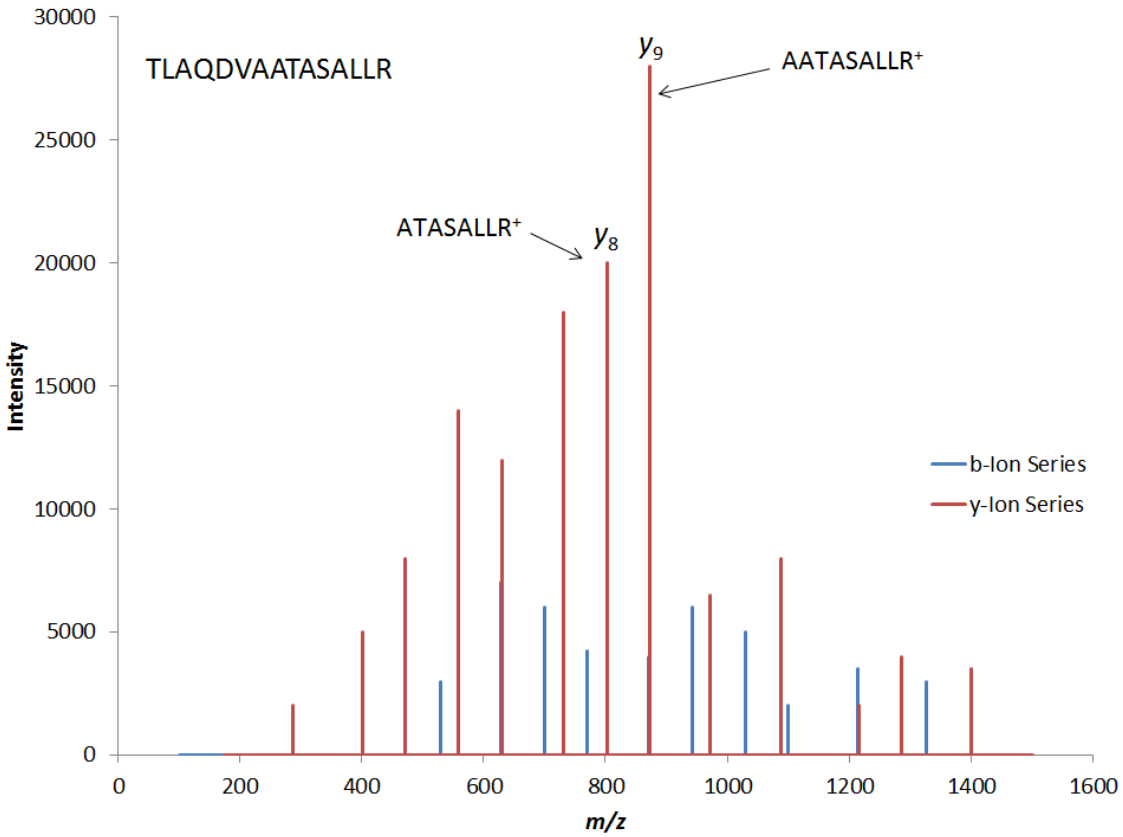

Figure 1.11 Amino Acid Determination From *b*- and *y*-Ion Ladders

The difference in mass between two consecutive ions in a *b*- or *y*-ion series determines the identity of one of the amino acids of the analyte peptide. Isomers such as Leu and Ile cannot be identified explicitly using this method.

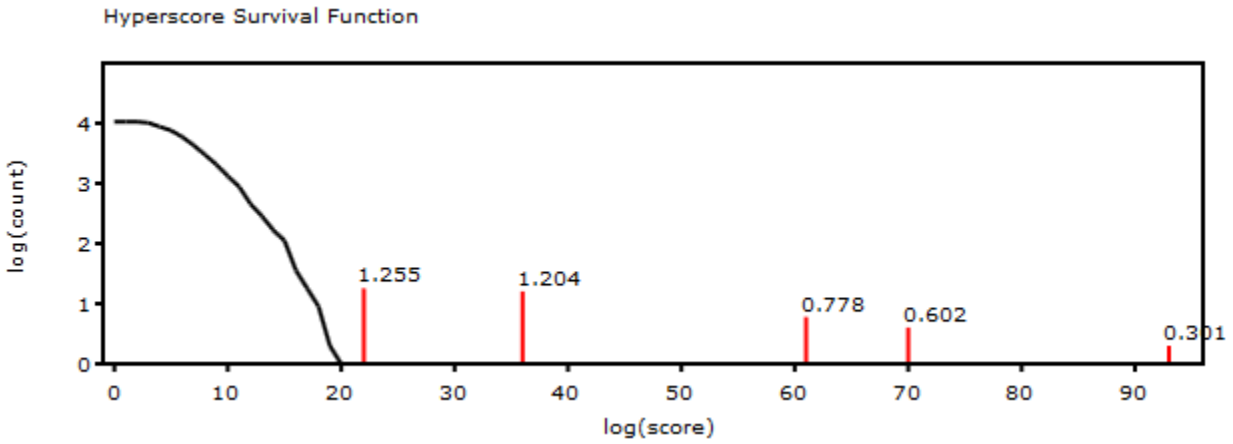

Figure 1.12 X!Tandem Peptide Match Histogram

Significant matches are tagged along the *x*-axis. Matches furthest from the curve are considered the most significant because they are least likely to occur by chance.

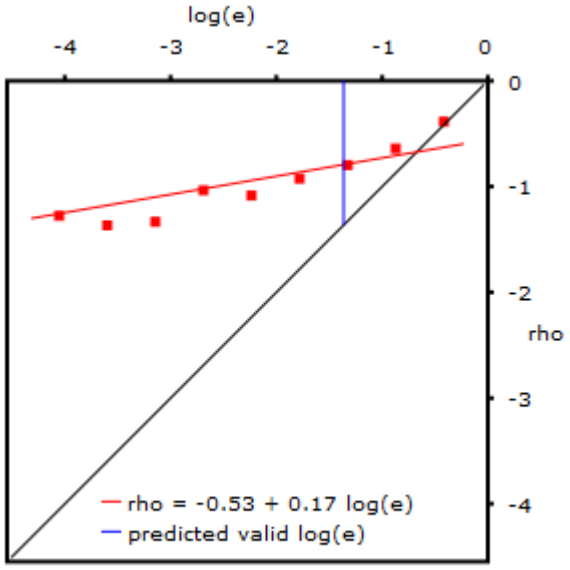

Figure 1.13 Log(e) Transform of a Hyperscore Histogram

X!Tandem will predict a threshold where log(e)'s become valid. The scores can have their E-value determined by extrapolation.

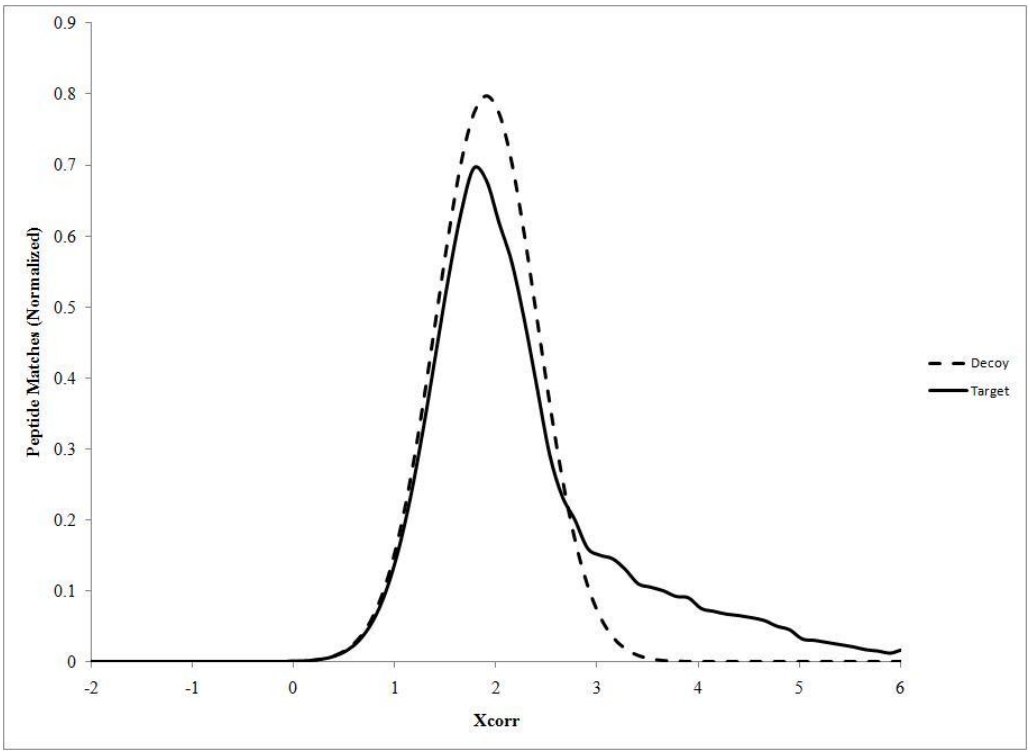

Figure 1.14 Histogram of Target and Decoy Scores

Decoy scores form a Gaussian centered over poorer Xcorr values. This represents the random nature of Decoy matches. Target matches also yield many random matches, however there is a bias towards higher scores due to high quality spectra originating from real peptides.

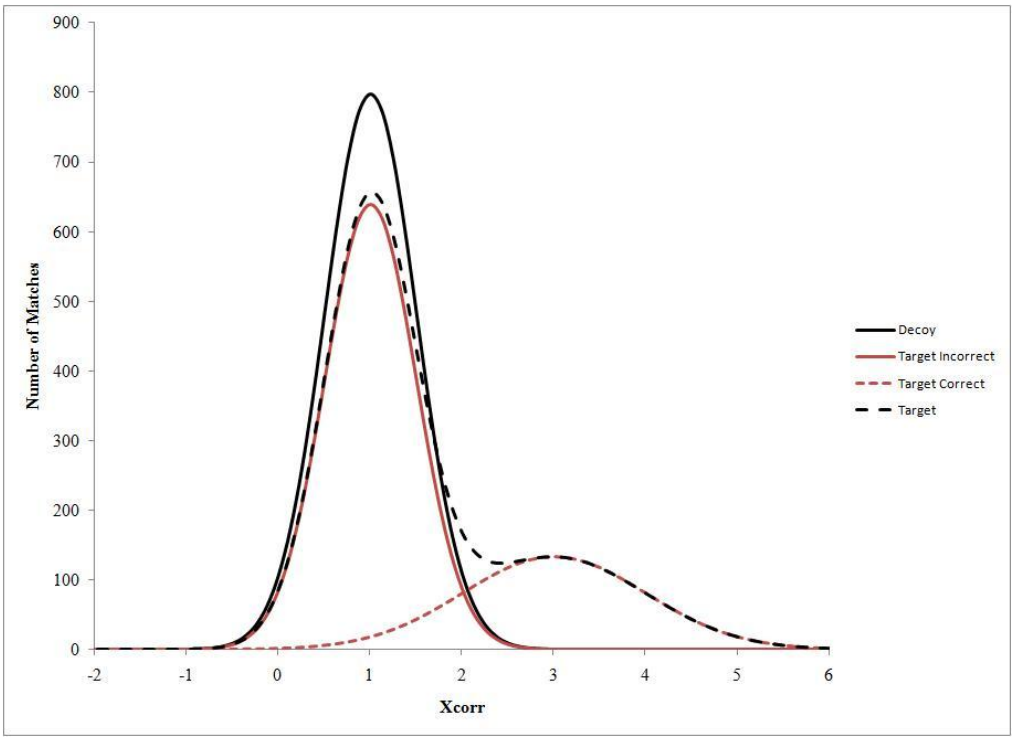

Figure 1.15 Target/Decoy Makeup of a Search Histogram

Target searches are assumed to contain both correct and incorrect matches. Searching against a decoy database will allow for the creation of a gamma distribution that can help determine which target matches are incorrect.

CHAPTER 2

MASS ACCURACY EFFECTS ON SHOTGUN PROTEOMICS DATABASE SEARCHES

Nuccio, A.G. and Orlando, R. To be submitted to *Molecular and Cellular Proteomics*

ABSTRACT

High mass resolution and accuracy are extremely important for mass spectrometry based proteomics. Using highly accurate and resolved *m*/*z* measurements, precursor and fragment ion masses and their charge states can be determined. Whether or not precursor or fragment accuracy is more important for ion scores depends on the scripting of the search algorithm used. Accurate *m*/*z* measurements generally decrease FPR's and increase the speed of protein database searches. Despite these improvements, with large enough datasets false positive matches are inevitable. FDR and probability filtering are effective and commonly used methods for removing false positive matches while preserving true positive matches. FDR filtering should be can be used when only a modest amount of MS/MS spectra are available while probability filtering is most powerful when several thousand spectra are available. Popular search algorithms have results scores which can be used as a gauge for correctness, but it has become the industry standard to use statistical validation programs to generate publication worthy data. Though precursor ion accuracy is extremely influential on the quality of database search results, fragment ion masses are central to the identification of proteins. As a result, high mass accuracy fragment ion measurements have the greater effect on decreasing the FPR of database search results. However, combining high mass accuracy and resolution precursor and fragment is most desirable if there is no significant loss in sensitivity and throughput.

INTRODUCTION

Mass spectrometers with high sensitivity, mass accuracy, and mass resolution and fast scan times are relatively new to the proteomics field. These new instruments are capable of generating 10,000's of spectra during lengthy LC-MS/MS experiments. The most popular search engines used today were created before today's high performance mass spectrometers were available. As a result, many of today's database search programs are not ideally suited for processing large amounts of high mass accuracy peptide fragment data. Sequest and Mascot were originally introduced into the field when ion traps and QToF's were the workhorses of proteomics. Neither of these instruments excels at generating high resolution fragmentation spectra, and as a result, Sequest and Mascot were not designed to weight matches based on the fragment ion tolerance window used during search. X!Tandem is an example of a newer, and publically available, search program that heavily weights the fragment ion tolerance window rather than the precursor ion tolerance window. It is difficult to gauge which search algorithm is the best, and while literature comparisons tend to favor Mascot and Sequest, other search programs are still viable and can show better sensitivity for certain types of proteins and PTM's.

Regardless of the strengths and weaknesses of each search algorithm, the increasing performance and availability of mass spectrometers suited for shotgun proteomics has made searching extremely large datasets a normal occurrence. High resolution and accuracy mass spectrometry greatly reduces the false positive rates obtained during database searching. Unfortunately, in a large dataset a low false positive rate will still lead to an unacceptably high number of false positive matches. With biomarker discovery as one of the most desirable goals of shotgun proteomics, great efforts must be taken to minimize false positives matches. Combining high

mass accuracy and resolution mass spectra with FDR and probability filters is a potent way of achieving this.

METHOD

Protein Sample Preparation

A standard protein set was made containing 62 Trypanosoma cruzi recombinant proteins, verified by LC-MS/MS, spiked into an Escherichia coli whole cell lysate. The recombinant proteins were obtained from Rick Tarleton's lab at the Center for Tropical and Emerging Global Diseases, UGA. The E. coli was obtained from Sigma Aldrich (EC11303) as lyophilized cells and ruptured using a tip sonicator. E. coli cells were dissolved in lysis buffer (50mM Tris-HCl pH 7.5, 200mM NaCl, and 5mM DTT) and freeze-thawed at -20°C five times. Cells were then ruptured using a tip sonicator using 5 ten second bursts with 30 second cooling intervals. The solution was always kept on ice. Cell debris was removed by centrifugation at 4°C for 30 minutes under 45,000RPM in a Beckman ultracentrifuge. Concentrations of the *T. cruzi* proteins were determined by BCA assay (Pierce). The molar concentrations of the *T. cruzi* proteins span four orders of magnitude. Lyophilized *T. cruzi* and E. coli protein were combined in a 1:1 dry mass to mass ratio.

The standard protein set was reduced (DTT), carboxyamidomethylated (IDA), and then digested using sequencing grade modified trypsin (Promega). The lyophilized protein mixture was first rehydrated in 50mM ammonium bicarbonate (Sigma), reduced using 10mM DTT at 55°C for 1 hour, and carboxyamidomethylated in the dark for 30 minutes at room temperature. Trypsin was then added in a 50:1 ratio and the solution was incubated at 37°C for 8 hours. Prior to LC-MS/MS analysis, samples were filtered using 0.2µm Millipore Amicon centrifuge filters.

LC-MS/MS

An Agilent 1100 capillary LC (Palo Alto, CA) was attached to a T splitter to deliver nL flow rates into the mass spectrometer. Capillary columns were made using five µm diameter C18 beads (Rainin, Woburn, MA) packed into a pulled fused silica capillary (10.5cm x 75µm ID) under 1,000 psi pressure using nitrogen gas. Peptides were then eluted with a gradient using 0.1% formic acid (buffer A) and 99.9% acetonitrile/0.1% formic acid (buffer B). Following the initial wash with 95% buffer A for 10 minutes, peptides were eluted from the column during a 75 minute gradient of 5-50% buffer B at ~200nL/min directly into an LTQXL-Orbitrap mass spectrometer (Thermo Fisher, San Jose, CA). The orbitrap was set to acquire MS/MS spectra on the six most abundant precursor ions from each MS scan with a repeat count set to 1 and repeat duration of 6. Dynamic exclusion was enabled for 30 seconds. Both precursor and fragment ions were detected in the orbitrap at a resolution of 30,000 and 7,500 M/ΔM respectively, providing m/z accuracies of about 20ppm.

Database Analysis

Mascot, Sequest, and X!Tandem were used to search this data against both target/decoy and concatenated databases. The target database consisted of the 62 *T. cruzi* proteins, 4243 E. coli K12 proteins (NCBI) and 5,000 randomly generated protein sequences, while the decoy database consisted of a reversed copy of the target database. The *T. cruzi* protein and Random protein sequences were obtained from Brent Weatherly in the Tarleton laboratory. The randomly generated proteins were designed with two major constraints: No in-silico tryptic fragments would match any reverse, *T. cruzi*, or E. coli sequences, and protein/peptide sizes would mimic those of E. coli proteins. The concatenated database contains both the forward and reverse

protein sequences. In all of these searches carbamidomethylation (+57.03404Da) of cysteine (C) was set as a fixed modification while deamidation (+0.9839Da) of asparagine (N) and glutamine (Q) with oxidation (+15.99492Da) of methionine (M) were set as variable modifications. Up to two missed cleavages were allowed. For the study of precursor m/z accuracy's effect on ion scores, the fragment m/z tolerance was held at 0.6 or 0.02Da while precursor m/z tolerances were varied. For the study of fragment mass accuracy's effect on ion scores, precursor mass tolerance was held at 1000 or 20ppm while fragment m/z tolerances were varied. Mascot does not allow ppm searches for fragment tolerance, so 0.02Da was chosen because it represents 20ppm of a 1000Da polypeptide. Table 1 summarizes the precursor and fragment combinations chosen.

The database has 3 protein components: 62 *T. cruzi* proteins, 5000 Random proteins, and 4243 E. coli proteins. The *T. cruzi* entries serve as confirmed true positive matches, the Random represent false positive matches, and the E. coli proteins help to simulate the complexity of a real biological sample. While tabulating the number of peptides discovered, E. coli matches were disregarded because it would be impossible to tell if they were true or false positive matches. They were included in the mixture because without a sufficiently complex mixture, it's possible that the search algorithms would successfully identify all members of the small *T. cruzi* protein set without mistakenly assigning real spectra to the wrong proteins. By increasing the number of peptides and MS/MS events, it becomes much more likely that search algorithms will match MS/MS data to some of the Random entries of the database. Both concatenated and forward/reverse databases were used for comparison.

ProteoIQ was used to apply statistical filtering to the database search results. 5% Peptide FDR and 95% peptide probability filters were predominantly used while monitoring the effects of search accuracy on minimizing FP rates. Peptide identifications were tabulated rather than

proteins due to the relatively small size of the dataset (being only one LC-MS/MS bomb loaded run).

RESULTS AND DISCUSSION

Effect of MS1 Tolerance on Sensitivity and False Positive Rate

Each search algorithm uses precursor tolerance to determine a list of candidate tryptic peptides potentially responsible for a given observed peak that triggered an MS/MS event. When fragment tolerance is held at 0.6Da, as precursor tolerance narrows from 1000 to 20ppm the number of *T. cruzi* peptides increases by 27% while the number of Random proteins decreases to nearly zero. When fragment tolerance is held at 0.02Da and precursor tolerance narrows from 1000 to 20ppm, the number *T. cruzi* and Random matches was unchanged with the number of Random matches being held to a minimum.

With a wide fragment tolerance window, increasing the search accuracy shrank the pool of possible peptide matches. Since real peptides are more likely to be very close in *m*/*z* to predicted peptides, it is logical that the number of Random matches decreases. The simultaneous increase in the number of *T. cruzi* matches for a narrower window occurs due to how FDR's are calculated by ProteoIQ. With more false positives ruled out by the narrow search tolerance window, the assigned score threshold is lowered allowing more *T. cruzi* peptide matches to pass the 5% peptide FDR filter. When the FDR algorithm delves further into the low scoring matches, more Random peptides are considered; however, Random peptides are infrequently matched by Mascot when narrow fragment ion tolerances are used. Sometimes a peptide match being scored could be mistakenly attributed to an MS/MS spectrum belonging to a different peptide because by chance it has a fragment ion series with more peak matches than the actual

peptide fragmented to generate the experimental spectrum. These cases are exceedingly rare and typically only occur because a wide window is chosen for precursor tolerance, allowing for an exceptionally large amount of MS/MS spectra to be considered.

When fragment tolerance is held to 0.02Da, precursor ion tolerance has no effect on the number of Random or *T. cruzi* peptides identified. Furthermore, the number of Random peptides was reduced to 1 and the same number of *T. cruzi* peptides was identified at 38. This indicates that fragment tolerance has the strongest effect in what peptides are identified. Mascot, Sequest and X!Tandem all had similar results in this case.

Effect of MS2 Tolerance on Sensitivity and False Positive Rate

Fragment ion tolerance directly effects how a search algorithm assigns b- and y-ions and therefore peptide sequence identification. With wide fragment tolerances, more peaks can be considered in b- and y-ion assignments. Mascot and Sequest determine scores based on how many b- and y-ions fit in the fragment tolerance bin. No bonus is given based on how narrow the window is, unlike X!Tandem where log(e) values are increased for more narrow fragment tolerance windows. Therefore, fragment ion tolerance does not affect Mascot and Sequest scores, unless the search window is so narrow that it causes some b- and y-ions to fall outside the bin. Figure 2.1 plots search score as a function of ion tolerance for four different *T. cruzi* peptides.

When precursor tolerance is held at 1000ppm and fragment tolerance is narrowed from 0.6Da to 0.02Da, the number of *T. cruzi* and Random matches are constant until 0.06Da where the number of *T. cruzi* matches then increases by 7% and the number of Random matches decreases from three to one. This effect is similar to what was observed when precursor tolerance was

narrowed with fixed 0.6Da fragment tolerance. The cause for this effect is also the same. A narrow ion tolerance will remove false positives far more frequently than true positives resulting in the FDR script looking further into the list of peptides until the specified 5% FDR is reached.

Holding precursor tolerance at 20ppm while fragment tolerance narrows from 0.6Da to 0.02Da illustrates the strength of high accuracy MS1 scans in removing false positives. With the list of candidate matches reduced to precursor ions within a 20ppm window, only precursors that nearly match a peptide's theoretical *m*/*z* will be considered by the search engine, greatly reducing Random matches. Figure 2.2 shows the lack of an effect of fragment tolerance in reducing FPR's until 0.06Da. After that, at 0.02Da a loss of identifications is found showing "in or out of the bin" treatment of b- and y-ions by Mascot. Figure 2.3 shows the importance of fragment tolerance in X!Tandem scoring. Lower FPR's are achieved by entering the narrow MS2 window to match the performance the orbitrap is capable of. Real MS/MS spectra are far more likely to be compatible with narrow search windows and as a result, the FDR script can include more *T. cruzi* matches before it obtains enough Random matches to achieve a 5% FDR. Furthermore, different Random proteins/peptides appear and disappear as fragment tolerance is narrowed from 0.1Da to 0.02Da. No such change is observed for *T. cruzi* protein identifications suggesting that they are true positive identifications.

Figure 2.4 displays a histogram of Mascot rounded to the next higher integer. For the histogram plot, the number Random peptide matches have been normalized by dividing by 12. Each extreme end of search parameters was used to compare the effects of high precursor accuracy, high fragment accuracy, and both high precursor and fragment accuracy. The Random matches in the 20ppm/0.6Da search are greater in number than in the 1000ppm/0.02Da search showing that fragment tolerance is more important than precursor tolerance in reducing FPR's for

Mascot searches. Furthermore, there does not appear to be a significant loss in *T. cruzi* matches when comparing 1000ppm/0.02Da to 20ppm/0.6Da. More *T. cruzi* and Random peptides are removed, but Randoms are removed at a much faster rate. This is a result of a 0.02Da fragment tolerance window largely succeeding at selecting the correct peptide candidates pooled in the large 1000ppm precursor window. When only high scoring peptides matches are considered (Score > 30) the same trends are seen. The 1000ppm/0.02Da and 20ppm/0.02Da Score>30 searches have the same number of *T. cruzi* and Random matches, more evidence that MS2 tolerance is the more important search parameter. Table 2 summarizes the percent change in total raw peptide identifications from the widest search window, 1000ppm/0.6Da. High accuracy precursor and fragment ion tolerance, 20ppm/0.02Da, has the best sensitivity and lowest FPR; however, a significant number of *T. cruzi* matches are lost.

Exacerbating this problem is the relative lack of MS/MS spectra in the data set under investigation. Figure 2.4 displays histograms for Mascot searches of 1000ppm/0.6Da and 20ppm/0.02Da tolerance. With more data from replicates or MudPIT fractions, the observed values would form a much smoother curve with a higher peak. These plots are strong evidence that probability filtering is not the optimal choice in statistically validating this particular dataset. The throughput and sensitivity of orbitrap is not maximal due to the fragmentation method used. Multiple replicates and/or MudPIT fractions would have made statistical filtering models fit better. High mass accuracy precursor and fragment scans with corresponding narrow database search tolerances and several replicates would make yield results with the most *T. cruzi* identifications confidently identified.

The Necessity of Statistical Filtering

Each database search program comes with a native scoring mechanism that, while capable, has limitations. The search algorithms are far too lenient in reporting false positives that pass scoring thresholds given by the algorithm. Figures 2.5 and 2.6 show the effect of precursor mass on the number of peptides found by Mascot to be above the default $p<0.05$ threshold which suggest identification or homologue identification. The threshold assigned by Mascot is dependent on the precursor mass tolerance entered into the search fields, and as the tolerance window is narrowed, the score threshold is lowered. This results in a large increase of Random identifications passing the $p<0.05$ threshold even when fragment tolerance is 0.02Da.

X!Tandem exhibits more consistent FPR's when using accepted log(e) values. However, its native validation thresholds are fairly effective, producing a FPR of 9.7% for 20ppm/0.02Da. This is too high for publishable data, but it is far better than the 33.8% FPR given by the Mascot native validation. 5% and 1% FDR's significantly reduce the number of both Random and *T. cruzi* matches, but the reduction of false positives is of paramount importance. 1% FDR's reduce the FPR to 0% and 2.5% for Mascot and X!Tandem respectively. For this dataset, 95% peptide probability filters also reduce the number of Random peptides to zero; however, *T. cruzi* matches were also reduced significantly. It is likely that some of the *T. cruzi* matches are matches that occurred by chance, but they are assumed true positives in order to assess the performance of FDR and probability based statistical filtering in this investigation. The effect of statistical validation is far greater on Random peptide matches than on *T. cruzi* matches, suggesting that this is a safe assumption for this investigation. For datasets this small (only 1800 spectra), probability based statistical models may not have enough data points to fit properly. Figure 2.7 compares different statistical filtering methods for each search algorithm used. The version of

Sequest used does not state an XCorr or ΔCn value at which matches are taken to be true positive identifications so that bar is omitted in Figure 2.8

When no statistical filters are used for the 20ppm/0.02Da search, there is still an unacceptable number of Random matches. Search results at 20ppm/0.02Da accuracy in Mascot yield 589 Random, 92 *T. cruzi*, and 509 *E. coli* peptides. If one were to assume that all *T. cruzi* and *E. coli* matches are true positives, this still results in an unacceptable 49% false positive rate. It is more likely that all of the Random and a significant portion of the *T. cruzi* and *E. coli* matches are false positives, exacerbating the poor false positive rate. Table 2 summarizes the results of different validation methods using a 20ppm/0.02Da search in Mascot. A 5% FDR is the most liberal but still serviceable statistical filter for this dataset. High mass accuracy spectra and searches are not a sufficient replacement for statistical filtering.

The difficult issue here is how gauge the importance in the loss of false positives compared to the loss of true positives. The lesser of the two evils will depend on the goal of an investigator's proteomics experiment. For biomarker research, it is critical that false positives be eliminated, though any protein of interest should be manually validated anyway. For a quantitative experiment, one might still be interested in poor scoring matches so long as the precursor peptide has an expected chromatographic retention time. For extremely large datasets encompassing many LC-MS/MS experiments, there will be enough spectra to apply good fits for Bayesian and Gaussian models. In the dataset used in this investigation, either probability filtering or FDR filtering were employed with FDR filtering proving more favorable. With larger datasets both types of statistical validation can be employed for the same dataset. Additional parameters frequently used include minimum spectra/peptides found, score cutoffs, and consensus database identifications.

CHAPTER 3

CONCLUSIONS

This investigation analyzed a validated subset of 62 *T. cruzi* proteins spiked into an *E. coli* lysate and searched against a custom protein database containing the 62 *T. cruzi* proteins, 5000 Random proteins, and the *E. coli* proteome. By narrowing the input search parameters from 1000ppm/0.6Da to 20ppm/0.02Da, Mascot identified 82% fewer Random peptides and 23% fewer *T. cruzi* peptides (28 fewer *T. cruzi* peptides discovered). Individually narrowing fragment tolerance to 0.02Da had 3 times the effect of reducing Random matches compared to individually reducing precursor tolerance to 20ppm. This shows the ion fragment tolerance is more important for reducing false positive rates. Obtaining a false positive match for a given MS/MS spectrum is unlikely if the tolerance window is very narrow due to how many peaks would have to coincidentally be aligned. Mascot and Sequest do not weight fragment tolerance input when assigning scores. X!Tandem behaves in the opposite manner and will assign higher scores for any matches that occur in high accuracy fragment ion searches.

Statistical validation is still necessary for high accuracy searches. Even with a 20ppm/0.02Da search, Mascot identifies 589 Random peptides. The native statistical filtering function of Mascot claims that 21 Random peptides are protein or homologue identifications. X!Tandem's log(e) parameter fairs slightly better allowing 7 Random peptide matches to be deemed significant.

Database searches for shotgun proteomics experiments generate so many peptide/protein identifications that automated validation steps are necessary for speedily removing false positive

matches. There are roughly a dozen commonly used search programs used in the proteomics field today. Mascot and Sequest have been the industry standard for many years, but debate about the best search program is ongoing. The combination of statistical validation with high accuracy and resolution precursor and fragment scanning is a very powerful tool in minimizing false positive rates. With sufficient replicate analysis and newer fast scanning high accuracy instruments, losses in sensitivity can also be minimized and high accuracy MS/MS should become the preferred method of scanning. Even with search engines created in the era of ion traps, these search engines generate results with lower false positive rates and modest losses in sensitivity.

Table 1 Mass Tolerance Combinations for Database Searching

Precursor Tolerance (ppm)	**Fragment Tolerance (Da)**
1000	0.6
500	0.6
250	0.6
100	0.6
50	0.6
20	0.6
1000	0.02
500	0.02
250	0.02
100	0.02
50	0.02
20	0.02
20	0.3
20	0.1
20	0.06
20	0.02
1000	0.3
1000	0.1
1000	0.06
1000	0.02

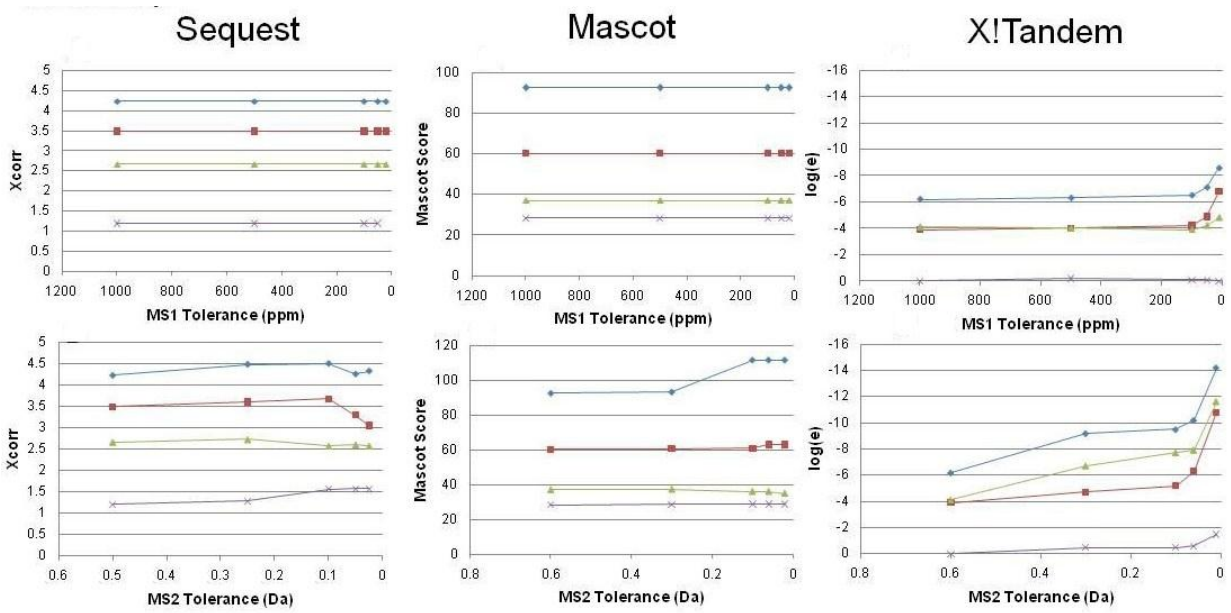

Figure 2.1 Search Tolerance vs Ion Score for Sequest, Mascot, and X!Tandem

The peptides KQFTELFATGER, RLPIGPITTVEK, VEVLQTQLPAYNR, and STGIDLSNER had their ion scores plotted as a function of ion search tolerance. Sequest and Mascot are consistent with their ion scores regardless of search tolerance. X!Tandem will adjust scores significantly for matches made in a narrow MS2 window.

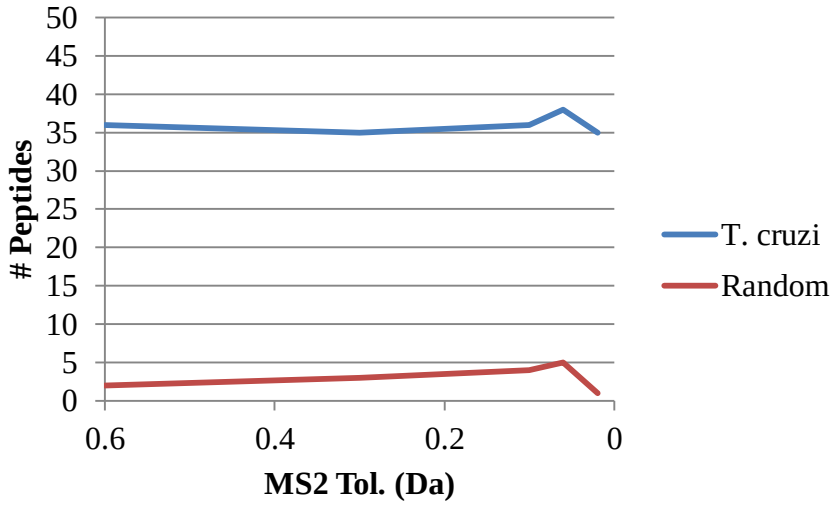

Figure 2.2 Fragment Ion Tolerance vs Number of Peptides, 20ppm Precursor Tolerance

Above is a Mascot search against the 9000 entry concatenated database with 20 ppm precursor mass held constant. 5% Peptide FDR filtering was used. When MS2 tolerance is lowered to 0.02Da, both correct and incorrect matches are removed from consideration by Mascot.

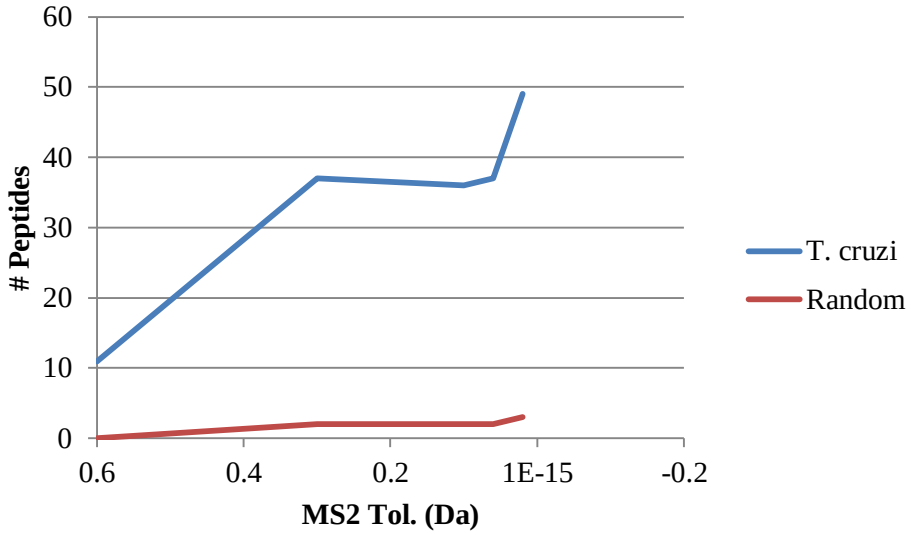

Figure 2.3 Fragment Ion Tolerance vs Number of Peptides, 20ppm Precursor Tolerance

X!Tandem search against the 9000 entry concatenated database with 5% peptide FDR filtering. X!Tandem weights matches made in high accuracy MS2 searches very favorably.

Table 2 *T. cruzi* and Random Matches at Specified Input Parameters

		T. Cruzi	**Random**
A) Gross Results	1000ppm/0.6Da	120	3167
	1000ppm/0.02Da	104 (-13%)	1176 (-63%)
	20ppm/0.6Da	113 (-5.8%)	2597 (-18%)
	20ppm/0.02Da	92 (-23%)	584 (-82%)
B) 20ppm/0.02Da	$p < 0.05$	41	21
	Score > 30	40	6
	5% FDR	35	1
	1% FDR	27	0
	95% Pep Prob	27	0
C) Score >35	1000ppm/0.6Da	45	42
	1000ppm/0.02Da	40 (-11%)	10 (-76%)
	20ppm/0.6Da	43 (-4%)	16 (-62%)
	20ppm/0.02Da	40 (-11%)	10 (76%)

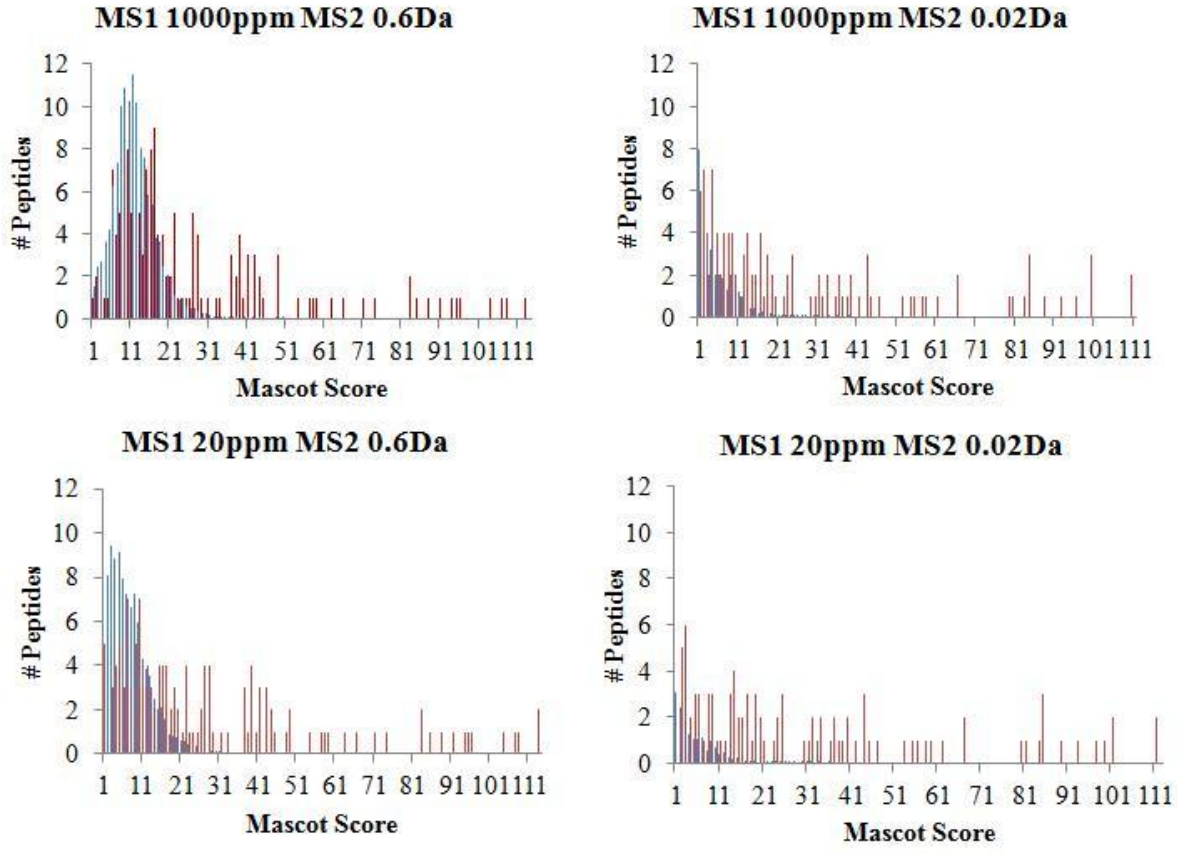

Figure 2.4 Histogram of Mascot Scores Under Minimum and Maximum Search Tolerance

Mascot searches provided enough datapoints to generate score histograms. Every combination of minimum and maximum search tolerance is plotted. MS2 tolerance has the greatest effect on reducing Random matches.

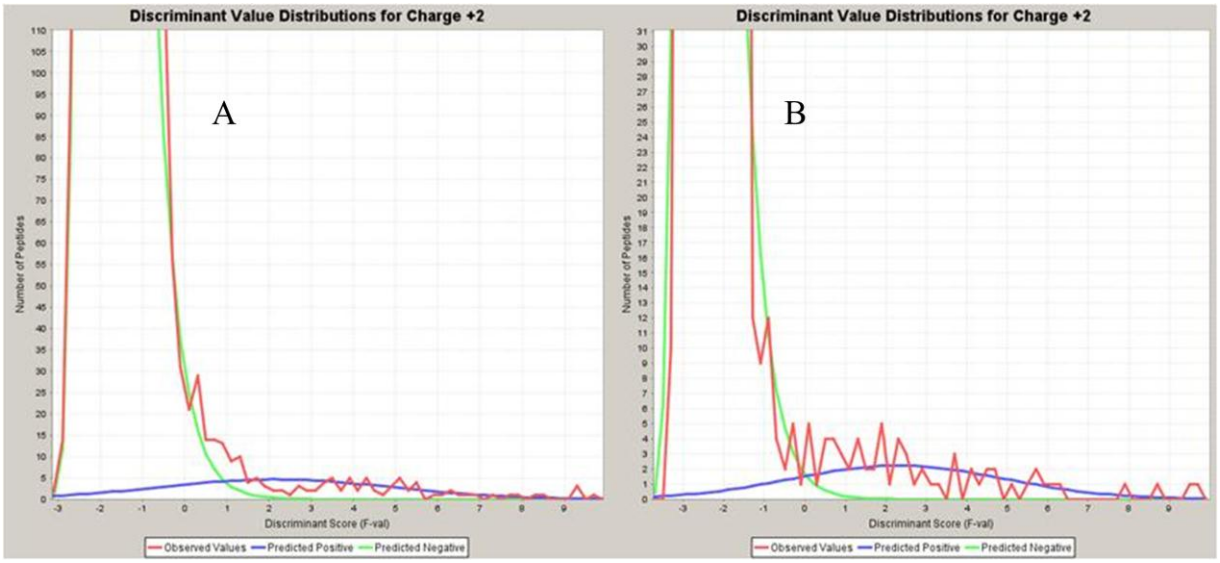

Figure 2.5 Histogram of Mascot Scores (F-Value)

Both searches were performed against the 9000 entry concatenated database (18,000 with reverse sequences). Figures A and B show 1000ppm/0.6Da and 20ppm/0.02Da tolerance respectively. The modest amount of spectra prevents high scoring experimental data from forming a smooth curve as depicted by the idealized "Predicted Positive".

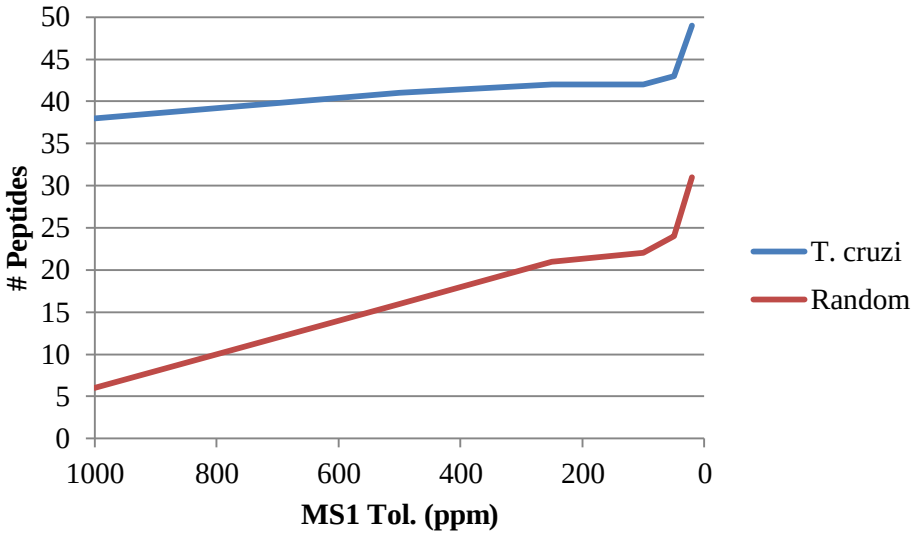

Figure 2.6 Precursor Tolerance vs. Number of Peptides at $p < 0.05$ Validation, 0.6Da MS2 Fixed

Fragment tolerance is held at 0.6Da while Precursor tolerance is narrowed from 1000ppm to 20ppm.

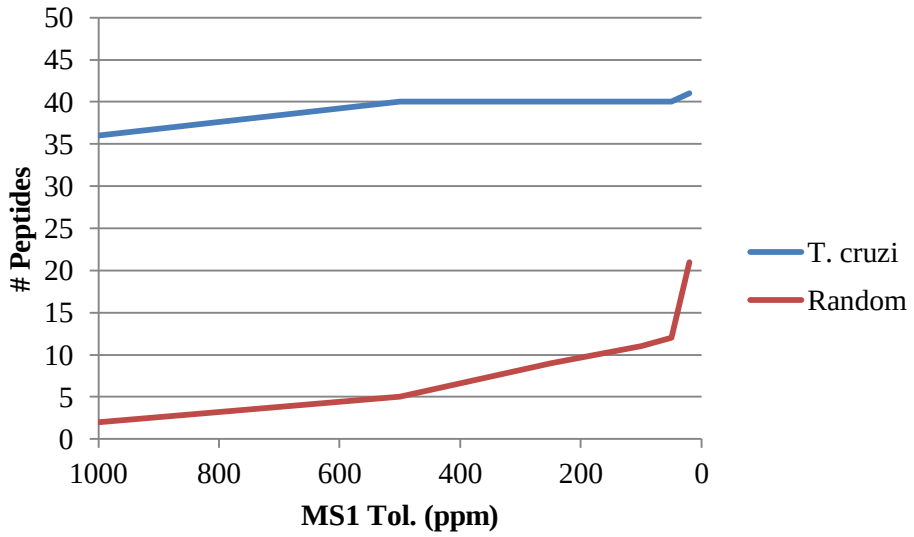

Figure 2.7 Precursor Tolerance vs. Number of Peptides at $p < 0.05$ Validation, 20ppm MS2 Fixed

Fragment tolerance is held at 0.02Da while Precursor tolerance is narrowed from 1000ppm to 20ppm.

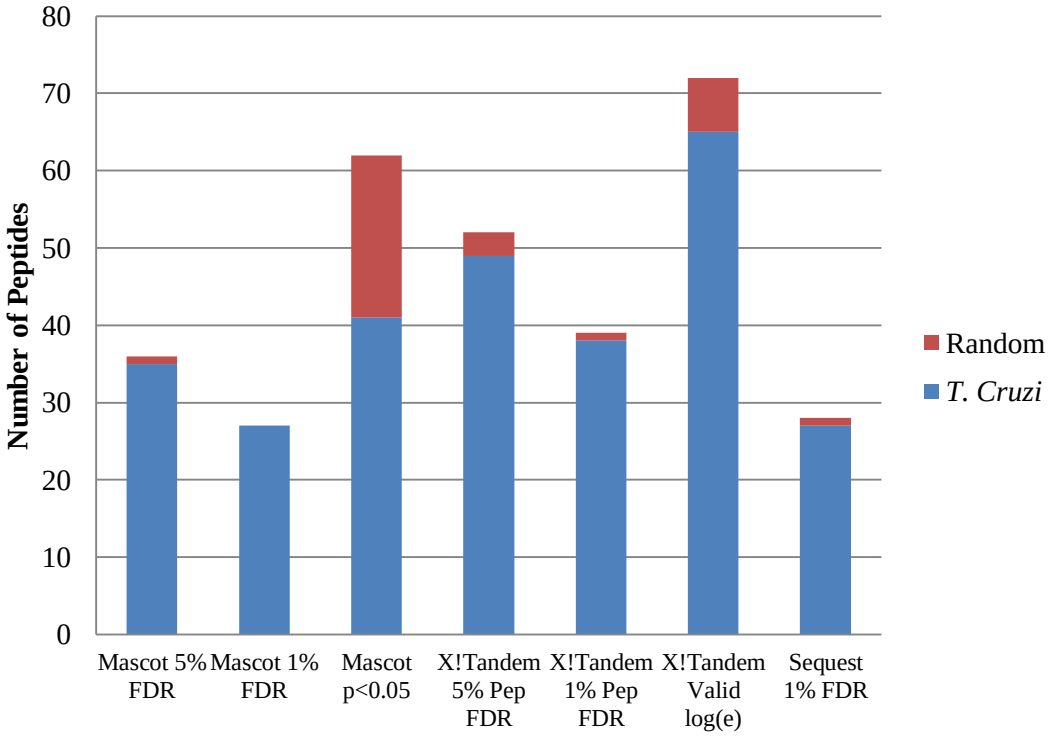

Figure 2.8 Comparison of Validation Techniques for a 20ppm/0.02Da Search

Various statistical validation techniques were employed and compared. Probability filtering was not effective due to insufficient spectra to form accurate probability models.

REFERENCES

1. Blackstock, W.P. and M.P. Weir, *Proteomics: quantitative and physical mapping of cellular proteins.* Trends in Biotechnology, 1999. **17**(3): p. 121-7.
2. Anderson, N.L. and N.G. Anderson, *Proteome and proteomics: New technologies, new concepts, and new words.* Electrophoresis, 1998. **19**(11): p. 1853-61.
3. Domon, B. and R. Aebersold, *Mass Spectrometry and Protein Analysis.* Science, 2006. **312**: p. 212-7.
4. Chen, C.-H., *Review of a current role of mass spectrometry for proteome research.* Analytica Chimica Acta, 2008. **624**: p. 16-36.
5. Chait, B.T., *Mass Spectrometry: Bottom-Up or Top-Down?* Science, 2006. **314**: p. 65.
6. Kero, F.A., R.A. Yost, and R.E. Pedder, *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*. 2005.
7. de Hoffmann, E. and V. Stroobant, *Mass Spectrometry: Principles and Applications*. 2001: Wiley.
8. Cooks, R.G., *Ion Trap Mass Spectrometry.*
9. Makarov, *Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis.* Analytical Chemistry, 2000. **72**(6): p. 1156-1162.
10. Hu, Q., et al., *The Orbitrap: a new mass spectrometer.* Journal of Mass Spectrometry, 2005. **40**: p. 430-443.

11. Scigelova, M. and A. Makarov, *Orbitrap Mass Analyzer - Overview and Applications in Proteomics.* Proteomics, 2006. **6**(S2): p. 16-21.

12. Perry, R.H., R.G. Cooks, and R.J. Noll, *Orbitrap Mass Spectrometry: Instrumentation, Ion Motion and Applications.* Mass Spectrometry Reviews, 2007. **27**: p. 661-699.

13. Makarov, A., et al., *Performance Evaluation of a Hybrid Linear Ion Trap/Orbitrap Mass Spectrometer.* Analytical Chemistry, 2006. **78**(7): p. 2113-2120.

14. Perry, R.H., R.G. Cooks, and R.J. Noll, *Orbitrap Mass Spectrometry: Instrumentation, Ion Motion and Applications.* Mass Spectrometry Reviews, 2008. **27**: p. 661-699.

15. Aguilar, M.-I., *HPLC of Peptides and Proteins: Methods and Protocols*. 2003: Humana Press.

16. Davis, M.T. and T.D. Lee, *Analysis of peptide mixtures by capillary high performance liquid chromatography: A practical guide to small-scale separations.* Protein Science, 1992. **1**: p. 935-944.

17. Gatlin, C.L., et al., *Protein Identification at the Low Femtomole Level from Silver-Stained Gels Using a New Fritless Electrospray Interface for Liquid Chromatography-Microspray and Nanospray Mass Spectrometry.* Analytical Biochemistry, 1998. **263**: p. 93-101.

18. Davis, M.T., et al., *A Microscale Electrospray Interface for On-Line Capillary Liquid Chromatography/Tandem Mass Spectrometry of Complex Peptide Mixtures.* Analytical Chemistry, 1995. **67**: p. 4549.

19. Delahunty, C. and J.R.Y. III, *Protein identification using 2D-LC-MS/MS.* Nature Methods, 2005. **35**: p. 248-255.

20. Cañas, B., et al., *Mass spectrometry technologies for proteomics.* Briefings in Functional Genomics and Proteomics, 2006. **4**(4): p. 295-320.

21. Dole, M., et al., J. Chem. Phys., 1968. **1968**(49).

22. Fenn, J.B., S.F. Wong, and C.K. Meng, *Multiple Chargin in Electrospray Ionization of Poly(ethylene glycols).* J. Phys. Chem., 1988. **92**: p. 546-550.

23. Fenn, J.B., et al., *Electrospray ionization for mass spectrometry of large biomolecules.* Science, 1989. **246**(4926): p. 64-71.

24. Smith, R.D., J.A. Loo, and H.R. Udseth, *Peptide and Protein Analysis by Electrospray Ionization-Mass Spectrometry and Capillary Electrophroesis-Mass Spectrometry.* Analytical Biochemistry, 1989. **179**: p. 404-412.

25. Kebarle, P., *A brief overview of the present status of the mechanisms invovled in electrospray mass spectrometry.* Journal of Mass Spectrometry, 2000. **35**: p. 804-817.

26. Gaskell, S.J., *Electrospray: Principles and Practice.* Journal of Mass Spectrometry, 1997. **32**: p. 677-688.

27. Griffiths, W.J., et al., *Electrospray and tandem mass spectrometry in biochemistry.* Biochem. J., 2001. **355**: p. 545-561.

28. Vaisar, T. and J. Urban, *Low-energy collision induced dissocitation of protonated peptides. Importance of an oxazolone formation for a peptide bond cleavage.* Eur. Mass. Spectrom., 1998. **4**: p. 359-364.

29. Edman, P., *A method for the determination of amino acid sequence in peptides.* Arch Biochem., 1949. **22**(3).

30. Matthiesen, R., *Methods, algorithms and tools in computational proteomics: A practical point of view.* Proteomics, 2007. **7**: p. 2815-2832.

31. Kapp, E. and F. Schutz, *Overview of tandem mass spectrometry (MS/MS) database search algorithms.* Current Protocols in Protein Science, 2007.

32. Xu, C. and B. Ma, *Software for computational peptide identification from MS-MS data.* Drug Discovery Today, 2006. **11**(13/14): p. 595-600.

33. Elias, J.E. and S.P. Gygi, *Target-decoy search strategy for increaased confidence in large-scale protein identification by mass spectrometry.* Nature Methods, 2007. **4**(3): p. 207-214.

34. John R. Yates, I., J.K. Eng, and A.L. McCormick, *An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database.* Journal of the American Society for Mass Spectrometry, 1994. **5**: p. 976-989.

35. Pappin, D.J.C., P. Hojrup, and A.J. Bleasby, *Rapid identification of proteins by peptide-mass fingerprinting.* Current Biology, 1993. **3**(5): p. 327-332.

36. Perkins, D.N., et al., *Probability-based protein identification by searching sequence databases using mass spectrometry data.* Electrophoresis, 1999. **20**: p. 3551-3567.

37. Higdon, R., et al., *Randomized Sequence Databases for Tandem Mass Spectrometry Peptide and Protein Identification.* OMICS, 2005. **9**(4): p. 364-379.

38. Nesvizhskii, A.I., O. Vitek, and R. Abersold, *Analysis and validation of proteomic data generated by tandem mass spectrometry.* Nature Methods, 2007. **4**(10): p. 787-797.

39. Choi, H. and A.I. Nesvizhskii, *False Discovery Rates and Related Statisitical Concepts in Mass Spectrometry-Based Proteomics.* Journal of Proteome Reasearch, 2008(7): p. 47-50.

40. Tabb, D.L., *What's Driving False Discovery Rates.* Journal of Proteome Reasearch, 2008(7): p. 45-46.

41. Käll, L., et al., *Assignig Significance to Peptides Identified by Tandem Mass Spectrometry Using Decoy Databases.* Journal of Proteome Reasearch, 2008(7): p. 29-34.

42. Storey, J.D. and R. Tibshirani, *Statistical significance in genomewide studies.* Proceedings of the National Academy of Sciences, 2003. **100**(16): p. 9440-9445.

43. Weatherly, D.B., et al., *A Heuristic Method for Assigning a False-discovery Rate for Protein Identifications from Mascot Database Search Results.* Molecular & Cellular Proteomics, 2005. **4**(6): p. 762-72.

44. Keller, A., et al., *Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search.* Anal. Chem., 2002. **74**: p. 5383-5392.

45. Zubarev, R. and M. Mann, *On the Proper Use of Mass Accuracy in Proteomics.* Molecular & Cellular Proteomics, 2007. **6**: p. 377-381.

46. Mann, M. and N. Kelleher, *Precision proteomics: The case for high resolution and high mass accuracy.* Proceedings of the National Academy of Sciences, 2008. **105**(47): p. 18132-18138.

47. Scherl, A., et al., *On the Benefits of Acquiring Peptide Fragment Ions at High Measured Mass Accuracy.* Journal for the American Society of Mass Spectrometry, 2008. **19**: p. 891-901.