# ABSTRACT

Title of Document: TOP-DOWN ANALYSIS OF BACTERIAL PROTEINS BY HIGH-RESOLUTION MASS SPECTROMETRY

Colin Michael Wynne, Ph.D. 2010

Directed By: Dr. Catherine Fenselau, Department of Chemistry and Biochemistry

In the biodefense and medical diagnostic fields, MALDI mass spectrometry-based systems are used for rapid characterization of microorganisms generally by detecting and discriminating the highly abundant protein mass-to-charge peaks. It is important that these peaks eventually are identified, but few bacteria have publicly available, annotated genome or proteome from which this identification can be made. This dissertation proposes a method of top-down proteomics using a high-resolution, high mass accuracy analyzer coupled with bioinformatics tools to identify proteins from bacteria with unavailable genome sequences by comparison to protein sequences from closely-related microorganisms. Once these proteins are identified and a link between the unknown target bacteria and the annotated related bacteria is established, phylogenetic trees can be constructed to characterize where the target bacteria relates to other members of the same phylogenetic family.

First, the top-down proteomic approach using an Orbitrap mass analyzer is tested using a well known, well studied single protein.  After this is demonstrated to be successful, the approach is demonstrated on a bacterium without a sequenced genome, only matching proteins from other organisms which are thought to have 100% homology with the proteins studied by the top-down approach.  Finally, the proposed method is changed slightly to be more inclusive and the proteins from two other bacteria without publicly available genomes or proteomes are matched to known proteins that differ in mass and may not be 100% homologous to the proteins of the studied bacteria.  This more inclusive method is shown to also be successful in phylogenetically characterizing the bacteria lacking sequence information. Furthermore, some of the mass differences are localized to a small window of amino acids and proposed changes are made that increase confidence in identification while lowering the mass difference between the studied protein and the matched, homologous, known protein.

# TOP-DOWN ANALYSIS OF BACTERIAL PROTEINS BY HIGH-RESOLUTION MASS SPECTROMETRY

By

Colin Michael Wynne.

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2010

Advisory Committee:
Professor Catherine Fenselau, Chair
Professor Sang Bok Lee
Professor John Ondov
Professor Plamen Demirev
Professor Sergei Sukharev

# Dedication

To my parents, John and Ann, who have always provided sage advice, encouragement, and love.  And to my brother Andrew, his wife Carissa, and their newborn son Aiden, for keeping me grounded and bringing me some humor when I needed a pick-me-up.

# Acknowledgements

First and foremost, I have to acknowledge my advisor, Dr. Catherine Fenselau, who was always available to provide assistance, who answered every question, and knew what questions to ask to keep me on track.  The rest of the Fenselau lab have been my constant companions for the last four years, and have seen me at my best and worst.  I next have to acknowledge my bioinformatics collaborator, Dr. Nathan Edwards, who was very patient with my constant questions about how to work certain programs and who helped setup the bioinformatics infrastructure that was necessary to push my thesis project forward.  I also have to acknowledge the Detect to Protect team at Johns Hopkins Applied Physics Lab, especially Dr. Plamen Demirev, who allowed me to collaborate on their project, which led to the majority of my project.  Finally, I have to acknowledge the Achievement Rewards for College Scientists Foundation, who I am grateful for supporting me for two semesters, and the NIH for also supporting the research.

# Table of Contents

Dedication....................................................................................................ii
Acknowledgements.......................................................................................iii
Table of Contents..........................................................................................iv
List of Tables..................................................................................................v
List of Figures...............................................................................................vi
Chapter 1: Introduction..................................................................................1
Protein Mass Spectrometry.............................................................................1
Mass Analyzers................................................................................................5
Proteomic Workflows.....................................................................................10
Data Analysis.................................................................................................13
Objectives.......................................................................................................15
Chapter 2: Preliminary Top Down Study of Bovine Ubiquitin.................................17
Introduction....................................................................................................17
Experimental...................................................................................................18
Results and Discussion...................................................................................20
Summary.........................................................................................................25
Chapter 3: Top Down Analysis of *Yersinia rohdei* Lysate........................................27
Introduction....................................................................................................27
Experimental...................................................................................................29
Results and Discussion...................................................................................35
Phylogenetic Analysis....................................................................................51
Summary.........................................................................................................54
Chapter 4: Phlyoproteomic Characterization of Unsequenced Organisms by Top Down Identification of Proteins Using Capillary LC-MS/MS on a LTQ-Orbitrap XL..........55
Introduction....................................................................................................55
Experimental...................................................................................................57
Results and Discussion...................................................................................62
Phylogenetic Analysis....................................................................................95
Summary.......................................................................................................100
Chapter 5: Conclusions................................................................................102
Identification of Proteins...............................................................................102
Orbitrap Mass Spectrometry..........................................................................103
Sequence Homology......................................................................................104
Future Directions...........................................................................................105
Bibliography..................................................................................................106

# List of Tables

Table 2.1-Summary of Ubiquitin fragmentation showing precursor m/z ratio and the number of identified charge states.... ..........................................................................24

Table 3.1-Summary of *Bacillus anthracis* Sterne protein sequence matches......... ....36

Table 3.2-Table of *Yersinia rohdei* protein identifications when searched against a custom database from all *Yersinia* in the Swiss-Prot Database and other *Enterobacteriaceae*.........................................................................................38

Table 4.1-Database matches to *Erwinia herbicola* proteins.......................................63

Table 4.2-ProSight PC 2.0 matches to *Enterobacter cloacae*....................................81

# List of Figures

Figure 1.1-Cartoon image of Orbitrap mass analyzer.............................................................9

Figure 2.1-Precursor spectrum of ubiquitin showing charge states +7 through +14..............21

Figure 2.2-Fragmentation spectra of +8 charge state of ubiquitin with matched b- and y- ions highlighted in blue.........................................................................................................23

Figure 2.3-Fragmentation of the +9 charge state of ubiquitin with the matched b and y ion fragments highlighted in blue..........................................................................................24

Figure 2.4-the largest fragment highlighted in yellow in both the 3-D structure and the 1 letter representation of the amino acid structure of ubiquitin...............................................................................................................................25

Figure 3.1-Bruker Microflex MALDI-TOF signature of intact proteins from *Yersinia rohdei* with arrows indicating those proteins that were also confidently identified through the LC-MS/MS strategy.............................................................................................................39

Figure 3.2-Top:  MS/MS spectrum of the precursor ion at m/z 807.80 with a 9+ charge state for an intact mass of 7260.92 Daltons.  Middle:  The same MS/MS spectrum deconvoluted so that all ions are converted to +1 charge state.  Bottom:  Protein sequence assigned by ProSight PC 2.0, 50s ribosomal protein L29...........................................................................41

Figure 3.3-Top: MS/MS spectrum of the precursor ion at m/z 643.22 (14+ charge state, intact mass 8991.92 Da)). Middle: The same MS/MS spectrum with all fragment ions converted to zero charge state. Bottom: Protein sequence (ribosomal protein L27, Swiss-Prot A7FMT7) assigned by ProSightPC 2.0 showing observed fragmentation sites......................................42

Figure 3.4-Top: MS/MS spectrum of the precursor ion at m/z 682.68 (13+ charge state, intact mass 8862.89 Da)). Middle: The same MS/MS spectrum with all fragment ions converted to zero charge state. Bottom: Protein sequence (ribosomal protein L28, Swiss-Prot A1JHR2) assigned by ProSightPC 2.0 showing observed fragmentation sites......................................43

Figure 3.5-Top: MS/MS spectrum of the precursor ion at m/z 756.70 (8+ charge state, intact mass 6044.11 Da). Middle: The same MS/MS spectrum with all fragment ions converted to zero charge state. Bottom: Protein sequence (50s ribosomal protein L32, Swiss-prot A1JN60) assigned by ProSightPC 2.0 showing observed fragmentation sites......................................44

Figure 3.6-Top: MS/MS spectrum of the precursor ion at m/z 763.10 (11+ charge state, intact mass 8368.61 Da)). Middle: The same MS/MS spectrum with all fragment ions converted to zero charge state. Bottom: Protein sequence (ribosomal protein S21, Swiss-Prot P68684) assigned by ProSightPC 2.0 showing observed fragmentation sites......................................45

Figure 3.7-Top: MS/MS spectrum of the precursor ion at m/z 781.10 (8+ charge state, intact mass 6239.55 Da)). Middle: The same MS/MS spectrum with all fragment ions converted to zero charge state. Bottom: Protein sequence (ribosomal protein L33, Swiss-Prot A4W513) assigned by ProSightPC 2.0 showing observed fragmentation sites......................................46

Figure 3.8-Top: MS/MS spectrum of the precursor ion at m/z 802.90 (8+ charge state, intact mass 6413.56 Da). Middle: The same MS/MS spectrum with all fragment ions converted to zero charge state. Bottom: Protein sequence (50s ribosomal protein L30, Swissprot Q1C2W5) assigned by ProSightPC 2.0 showing observed fragmentation sites......................................47

Figure 3.9-Top: MS/MS spectrum of the precursor ion at m/z 857.72 (8+ charge state, intact mass 6852.67 Da)). Middle: The same MS/MS spectrum with all fragment ions converted to zero charge state. Bottom: Protein sequence (carbon storage regulator protein, Swiss-Prot Q1CL18) assigned by ProSightPC 2.0 showing observed fragmentation sites.......................48

Figure 3.10-Top: MS/MS spectrum of the precursor ion at m/z 1105.83 (11+ charge state, intact mass 12155.73 Da)). Middle: The same MS/MS spectrum with all fragment ions converted to zero charge state. Bottom: Protein sequence (ribosomal protein L22, Swiss-Prot A1JS31) assigned by ProSightPC 2.0 showing observed fragmentation sites........................49

Figure 3.11-Top: MS/MS spectrum of the precursor ion at m/z 1155.74 (13+ charge state, intact mass 15007.33 Da)). Middle: The same MS/MS spectrum with all fragment ions converted to zero charge state. Bottom: Protein sequence (ribosomal protein S6, Swiss-Prot A7FMW5) assigned by ProSightPC 2.0 showing observed fragmentation sites.....................50

Figure 3.12-Incidence matrix for observed *Y. rohdei* proteins in *Enterobacteriaceae* species.........................................................................................................................52

Figure 3.13-Phylogenetic trees based on top-down protein identifications (A) and 16S-rRNA sequence (B)..........................................................................................................53

Figure 4.1-Base peak chromatogram of *Erwinia herbicola* lysate separated on the Prominence Nanomate HPLC system (Shimadzu).......................................................................64

Figure 4.2-Top:  Deconvoluted tandem mass spectrum of precursor ion 1014.30 in charge state +8 identified as *Erwinia herbicola* protein Translation initiation factor 1A.  Bottom: Sequence of matched protein with 26 b- and y-ion fragment matches...................................65

Figure 4.3-Top: MS/MS of 1023.53 with charge +8 (molecular weight 9200.58).  Bottom: Sequence of Cell division protein zapB from Erwinia tasmaniensis with matched b- and y-ions.........................................................................................................................66

Figure 4.4-Top:  MS/MS of 761.98 with +11 charge (molecular weight of 8368.78).  Bottom: Amino acid sequence of ribosomal protein S21 with matched b- and y- ion matches............67

Figure 4.5-Top:  MS/MS of 686.39 with +15 charge (molecular weight 10285.10).  Bottom: Amino acid sequence of ribosomal protein S19 with b- and y- ion matches..........................68

Figure 4.6-Top:  MS/MS of 683.67 with +14 charge (molecular weight (9558.32).  Bottom: Sequence of ribosomal protein S17 with matched b- and y- ions...........................................69

Figure 4.7-Top:  MS/MS of 569.88 with +11 charge (molecular weight 6240.40).  Bottom: Sequence of ribosomal protein L33 with matched y-ions.......................................................70

Figure 4.8-Top: MS/MS of m/z 712.74 with +9 charge (molecular weight 6399.56). Bottom: Sequence of ribosomal protein L30 with matched b- and y- ions...........................................71

Figure 4.9-Top: MS/MS of m/z 1018.01 with +11 charge (molecular weight 11185.28). Bottom: Sequence of ribosomal protein L24 with matched b- and y- ions............................72

Figure 4.10-Top: MS/MS of 1138.13 with +6 charge (molecular weight 6820.01) Bottom: Sequence of carbon storage regulator protein with matched b-ions.........................73

Figure 4.11-Top: Deconvoluted tandem mass spectra of precursor 564.68 in charge state +11 identified as *Erwinia herbicola* protein 50s Ribosomal protein L32 with E-value 1.72e-7. Middle: Sequence of matched protein with 15 b- and y-ion fragments. Bottom: Highlighted substitution of serine to alanine at the 20th position, and rescored search with E-value now 5.85e-29 with 22 b- and y-ion fragment matches......................................76

Figure 4.12-Top: Tandem mass spectrum of precursor 1211.84 in charge state +6 identified as 50s Ribosomal protein L29 with E-value 2.71e-8. Middle: Original matching sequence with 13 b- and y- ion fragments. Bottom: Highlighted substitution of alanine to serine at C-terminus. Rescored E-value now 8.32e-20 with 14 b- and y- ion fragments.........................77

Figure 4.13-Top: tandem mass spectrum of precursor ion 732.71 in charge state +13 identified as DNA binding protein HU-alpha with E-value 7.5e-26. Middle: matched protein sequence with 31 b- and y-ion fragments. Bottom: Substitution of glutamate to aspartate at 38th position. Rescored E-value now 1.91e-58 with 41 b- and y- ions...................................78

Figure 4.14- Top: MS/MS of m/z 1221.98 with +6 charge (molecular weight 7318.23). Bottom: Sequence of cold shock protein with matched b- and y- ions..................................82

Figure 4.15- Top: MS/MS of m/z 1144.99 with +6 charge (molecular weight 6855.91). Bottom: Sequence of carbon storage regulator protein with matched b- and y- ions.............83

Figure 4.16- Top: MS/MS of m/z 798.45 with +6 charge (molecular weight 4781.59). Bottom: Sequence of DNA-binding protein HU-alpha with matched b- and y- ions.............84

Figure 4.17- Top: MS/MS of m/z 814.63 with +6 charge (molecular weight 4881.82). Bottom: Sequence of DNA-binding protein HU-alpha with matched b- and y- ions.............85

Figure 4.18- Top: MS/MS of m/z 1119.54 with +10 charge (molecular weight of 11185.20). Bottom: Sequence of ribosomal protein L24 with matched b- and y- ions............................86

Figure 4.19- Top: MS/MS of m/z 1306.71 with +8 charge (molecular weight 9137.52). Bottom: Sequence of DNA-binding protein HU-beta with matched b- and y-ions................87

Figure 4.20- Top: MS/MS of m/z 1189.65 with +8 charge (molecular weight 9505.00). Bottom: Sequence of DNA binding protein HU-alpha with matched b- and y- ions..............88

Figure 4.21- Top: MS/MS of m/z 1078.75 with +6 charge (molecular weight 6458.68). Bottom: Sequence of ribosomal protein L30 with matched b ions........................................89

Figure 4.22- Top: MS/MS of m/z of 809.36 with +11 charge (molecular weight 8891.36).

Bottom: Sequence of ribosomal protein L28 with matched b- and y-ions..............................90

Figure 4.23- Top: MS/MS of m/z 907.25 with +8 charge (molecular weight 7243.48). Bottom: Sequence of ribosomal protein L29 with matched b- and y- ions............................91

Figure 4.24- Top: MS/MS of m/z 937.16 with +11 charge (molecular weight 10299.10) Bottom: Sequence of ribosomal protein S19 with matched b- and y-ions..............................92

Figure 4.25- Top: MS/MS of m/z 1159.62 with +7 charge (molecular weight 8104.38) Bottom: Sequence of translation initiation factor 1A with matched b- and y- ions................93

Figure 4.26- Top: MS/MS of m/z 1464.22 of +9 charge (molecular weight 11705.50). Bottom: Sequence of thioredoxin protein with matched b- and y- ions.................................94

Figure 4.27- Top: MS/MS of 1221.64 with +6 charge (molecular weight 7318.23). Bottom: Sequence of cold shock protein Cold shock DNA protein-beta with matched b- and y- ions.........................................................................................................................95

Figure 4.28-Top: Deconvoluted MS/MS of precursor ion 732.91 with charge state +12, matching to Ribosomal protein S18 with an E-value of 6.28e-6. Bottom: ProSight PC 2.0 match with 11 b- and y- ion fragments. The highlighted portion indicates the N- and C-terminal amino acid sequences extracted for phylogenetic analysis......................................97

Figure 4.29-histogram denoting in blue where the N- and C-terminal sequences were extracted for phylogenetic analysis for the *Erwinia herbicola* matches.................................98

Figure 4.30-Phylogenetic trees constructed for *Erwinia herbicola* (A) and *Enterobacter cloacae* (B)..........................................................................................................99

# Chapter 1: Introduction

## *Protein Mass Spectrometry*

At its most basic concept, mass spectrometry is a technique that measures the mass and relative abundance of atoms and molecules[1]. In order to accomplish this, each mass spectrometer is composed of an ion source, an analyzer and a detector. The ionizer generates gas phase ions from the sample. The analyzer separates those ions by mass as well as allows for fragmentation of the precursor ion to create other, smaller ions. The detector finally provides the signal to be interpreted by the instrument software. The typical mass spectrum has two dimensions, the mass-to-charge ratio (m/z) ratio and the relative abundance or intensity. The most intense signal is generally set at 100% and the other signals having their height set in proportion to this "base" peak. In the past twenty-five years, protein mass spectrometry has grown due to advances in "soft" ionization[2]. "Soft" ionization provides charges to large molecules and biological materials without providing too much energy to cause fragmentation into smaller components. With this newer, less energetic ionization, the range of molecules able to be analyzed by mass spectrometry was expanded to the tens of thousands and hundreds of thousands of Daltons. This is the range necessary for whole protein and organism analysis and characterization.

One of the two most popular "soft" ionization techniques, especially for biological molecules, is matrix assisted laser desorption ionization (MALDI). MALDI was developed in the late 1980's by Koichi Tanaka[3]. The matrix is an organic acid, usually α-cyano-4-hydroxy-cinammic acid or sinapinic acid. The analyte is mixed with the matrix and allowed to dry on a metal plate. The plate is put under vacuum and the matrix-analyte spot is pulsed with a nitrogen laser at 337 nm. This excites the organic acid (and with it, the analyte) and the mixture desorbs into the gas phase as a plume of matrix-analyte clusters. The result of this plume produces mostly singly-charged ions, though the exact manner of ionization is still unclear. The leading thought on the dominant occurence of singly charged ions is the "lucky-survivor" theory proposed by Karas[4]. This theory states that the laser pulse initially creates a large charge imbalance towards positive ions in the desorbed clusters of matrix, analyte, and counter ions. However, as the plume is pulled towards the ion guides of the mass spectrometer, many of these charged clusters undergo charge reduction or neutralization through the capture of the electrons freed by the laser pulse. The exception to this is the singly charged ions, which have an unfavorable electron capture cross section. Therefore, these singly charged ions are the "lucky survivors" and move towards the analyzer within the mass spectrometer.

Because MALDI uses a rapidly firing laser and produces singly charged ions, it is widely used as a rapid screening tool for biological materials[5-7]. In a MALDI mass spectrum, most signals appear at their molecular weight, making for easy interpretation. Furthermore, separation of a mixture is not performed as much as in other ionization techniques since each analyte produces one set of isotopic peaks. This lack of separation reduces the sample preparation time, again benefiting rapid screening. However, because the charge is the denominator in the m/z ratio, MALDI mass spectrometers generally require an analyzer with a large range.

Electrospray ionization (ESI) is the other “soft” ionization technique widely used for protein mass spectrometry. Electrospray was developed in the mid-1980’s by John Fenn[8]. Electrospray utilizes an electric field and streaming nitrogen to ionize a mixture of dissolved target, and ion pair agent, and organic solvent. The ion pairing agent is generally acidic. The two most popular ion pairing agents are formic acid and acetic acid. The liquid mixture creates a Taylor cone of spray that spreads the solvated target ions. A nitrogen stream is used to keep this cone consistent and targeted towards the opening in the front of the mass spectrometer. Those solvated molecules become gaseous ions by a combination of two processes, field ionization and solvent evaporation. In solvent evaporation, the ion pairing agent imparts multiple charges to a droplet. As the solvent evaporates, the Coulombic repulsion becomes so great to break the surface tension of the droplet and place the target ion in the gaseous phase to enter the mass spectrometer.

Because electrospray ionization imparts multiple charges on the target, it is an ideal ionization method for a mass spectrometer with a limited mass-to-charge ratio. Generally, there are many isotopic envelopes for the same target due to different amounts of charge on the molecule. This causes a decrease in dynamic range, or the ability to detect distinct molecules at different concentrations. Because the molecules with higher concentrations will have many peaks, and these peaks will have similar abundance, the molecule with lower concentration will be more difficult to detect. ESI is also easily compatible with liquid chromatography to separate and analyze complex mixtures, including cell lysates and tryptic digests of proteins. This separation is necessary because of the lower dynamic range.

Recent advances in electrospray technology have given rise to lower flow rates through the electrospray needle into the nanoliter range. With the miniaturization of the pumps, fittings, and capillary tubes, less organic solvent and less sample is necessary to detect and identify proteins and peptides. Current technology allows for flow rates as little as 100 nL, bringing sensitivity of analyte to femtomole range [9]. This low flow rate also decreases the use of acetonitrile, which is increasingly expensive and dangerous to the environment, cutting the amount of waste production due to solvent injection.

Due to the multiple charges imparted by electrospray ionization, the mass analyzer used can have a limited mass-to-charge ratio range. The multiple charges also aid in any fragmentation that takes place. Multiple charges on a protein are generally distributed based on the position of the basic residues, lysine, arginine, and histidine. If the protein fragments at a different amino acid, then the basic residues will continue to carry one or more charges, leaving many fragments with a positive charge to be detected by the mass spectrometer in subsequent scans. Electrospray is generally the soft ionization technique used in experiments that require characterization of the fragment ions based on this reasoning.

## *Mass Analyzers*

The mass analyzer is the component of the mass spectrometer that separates the ions based on their mass-to-charge ratio. As with the ioniziation techniques, different analyzers have benefits and deficiencies in protein mass spectrometry and should be chosen based on the particular experiment that the user performed and the information desired by that user. Three types of analyzers widely used in protein mass spectrometry are the time-of-flight analyzer, the linear ion trap, and the Orbitrap.

Time-of-flight (TOF) mass analyzers separate the ions based on the time spent in a field free tube [3]. Once the ions enter the mass spectrometer, they are guided by RF and DC voltages to the beginning of this field free tube theoretically with the same kinetic potential. The other end of this field free tube is where the detector is usually placed. Because kinetic energy is proportional to mass multiplied by velocity squared, the ions will spend different velocities based on their mass. The ions with smaller mass will have a higher velocity and reach the detector first. The ions with a larger mass will have lower velocities and reach the detector later. Resolution of a time-of-flight mass analyzer is based on the length of this field free zone. TOF analyzers have a wide range of mass-to-charge ratios that can be detected, with upper bounds of more than 100,000 m/z [10]. This makes the TOF analyzer widely associated with MALDI ionization, since the single charge imparted by MALDI creates large m/z ratios, and TOF having such a wide range. Also, time-of-flight analyzers have a very fast scan rate, which accommodates pulsed ion production.

However, there are some applications that TOF analyzers are ill equipped for. Because of the need to allow ions to enter the field free zone at the same time, ions must be pulsed into the analyzer. With a liquid chromatography system in-line with the mass spectrometer, the ions are constantly flowing into the mass spectrometer, which makes it difficult to pulse the ions into the analyzer. Furthermore, because the TOF analyzer is a linear system dependant on one field free zone, fragmentation is a challenge within one analyzer. Generally, to characterize fragments, TOF analyzer are paired with another type of analyzer, or two TOF analyzers are placed with the mass spectrometer, one to analyze the precursor ions and the other to analyze the fragments. Fragmentation can also happen when the ions are initially injected into the mass spectrometer, either through high voltages near the inlet or increased laser power.

Another development in TOF technology has allowed for a longer field free zone, allowing for better resolution and increased detection of fragmentation. By placing a second detector on a different vertical path than the initial field free zone and curving the field using electric fields to guide the ions towards that 2nd detector, the field-free zone is lengthened. In this curved field reflectron [11], the resolution is increased. This CFR is also used for fragment detection since it allows for a longer activation time for fragmentation.

Linear ion traps are also used as analyzers in protein mass spectrometry. First described by Paul [12], linear ion traps use oscillating RF and DC voltages to contain ions of different m/z ratios in a small space. In a linear trap, higher DC potentials are applied to the ends to create an energy well to trap the ions in the middle, while RF potentials are oscillated to trap the ions inside in the x and y directions, with different m/z ratios oscillating at a different frequency. By ramping the main RF frequencies, the oscillations will eventually become unstable, again with each particular mass-to-charge ratio becoming unstable at a different RF frequency. This instability can cause ejection out of the trap through slits in the side, with the detectors on either side of the trap [13]. Resolution of the trap is based on the speed of change of the RF frequency along with the increment of change. This allows for higher resolution to be obtained by lowering the rate of scanning, but this method greatly increases the time needed to perform experiments. However, ion traps have a narrow mass-to-charge ratio range that can be analyzed. Large m/z ratios cannot be trapped by the high DC potentials at either end of the trap, and escape before being analyzed. Most ion traps have an m/z range of 0 to 2000. Because of this small range, ion traps are rarely coupled with MALDI ionization. Instead, ion traps are coupled with Electrospray ionization, with its multiply charged ions to increase the effective mass range [14]. With a resolution of about 2000, these linear ion traps can detect and determine the charge of a molecule with a +4 charge of a 3000 molecular weight ion using its normal scan rate.

One advantage that linear ion traps have is the ability to selectively trap or eject different m/z values. Because the RF frequency to destabilize trapped ions is different for every m/z, an ion trap can skip certain frequencies to isolate and trap an m/z value. This is beneficial to create fragment ions. By isolating a particular m/z, there should typically be one type of molecule in the trap. The trap then includes an excitation voltage and allows collisions between this type of molecule and an inert gas that the molecule fragments at its weakest point [15]. For peptides and proteins, this collisionally induced dissociation usually breaks the protein's weakest bond, the beta carbon to nitrogen amino bond linking one amino residue to another. Once the protein is fragmented, if the charge remains on the C-terminal side of the break, then that fragment is denoted a y fragment. If the charge remains on the N-terminal side of the break, then that fragment is a denoted b fragment. This is the nomenclature first proposed by Roepstorff and Fohlman[16] and later changed slightly by Biemann[17]. Once the fragments are formed, the trap then resumes full range scans of the RF frequencies to scan the complete m/z range. This ability to isolate and fragment a particular m/z is why linear ion traps are the analyzer most often used for experiments that necessitate the characterization of fragments.

The Orbitrap analyzer is one of the newest analyzers used in protein mass spectrometry. Invented by Makarov in 2000 [18, 19], the Orbitrap is an alternative to traditional superconducting magnet based FTMS systems for high resolution analysis. The Orbitrap is an ion trap with an oblong (football like) shape of its outer and central electrode. The ions are injected orthogonally to the central electrode and are attracted to an increasing voltage by this central electrode. The outer electrodes oscillate polarity, causing the ions to "orbit" around the central electrode, while endcap electrodes cause the ion to move back and forth across the central electrode. This is shown in Figure 1.1. The frequency of that movement is detected and subjected to an FT and is inversely proportional to the square root of the mass-to-charge ratio. All of the Orbitrap electrodes are controlled by electric potentials, not RF, so there is no need for a large superconducting magnet and the maintenance of that magnet. Because of the Fourier-transform, the resolving power of the Orbitrap mass analyzer is very large. While Makarov boasted 150,000 resolution in his initial experiments, the commercial version sold by Thermo Scientific has the capability to detect at 60,000 resolution with the LTQ-Orbitrap XL. Molecules with 20 or more positive charges can have their charge states determined, which would also determine their molecular mass. This higher resolution, and the high mass accuracy that goes with it, increases the effective range of an ion trap from about 8000 with a linear ion trap to over 100,000 without causing the scans to be so slow that peaks are missed when interfaced with liquid chromatography. However, the lack of RF means that a single m/z value cannot be isolated and fragmented, so any fragmentation must be done by another analyzer interfaced with the Orbitrap, such as a linear trap [20].

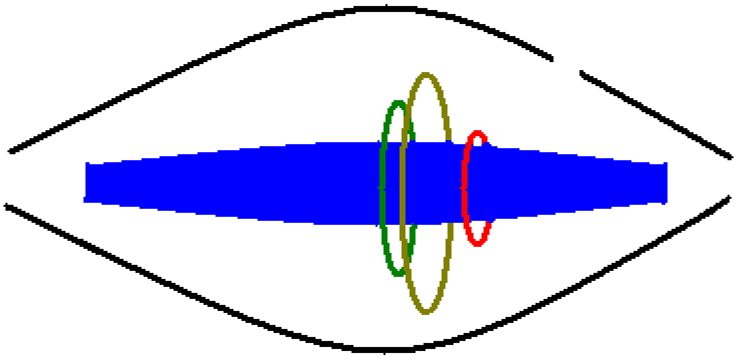


Figure 1.1-Cartoon image of three sets of ions in an Orbitrap mass analyzer, the red ions having the smallest m/z and the gold ions having the largest m/z.

### *Proteomic Workflows*

When trying to characterize proteins and organisms with mass spectrometry, there are some common practices and workflows that the community has developed and accepted. The mass spectrometrist has to choose which workflow to use based on the type of sample, the type of instrument available, and previous knowledge of the sample. Some of these workflows are better equipped to handle complex samples, some are better for rapid analysis, and some workflows allow for more information to be learned about a protein or organism. The proteomics [21] based approaches characterize fragments of a protein and match those fragments against a database of fragments generated *in-situ* from a collection of proteins. These proteins are entered into the database either through experimental discovery or translation from an annotated genome. These approaches are not dependant on using the same growth conditions or same sample preparation each time. The three proteomic approaches are bottom-up, top-down, and middle-down.analysis.

Bottom-up analysis uses enzymatic or chemical cleavage to hydrolyze the protein or group of proteins before injection into the mass spectrometer. Typically, trypsin enzyme is used to cleave the proteins at the C-terminus of the lysine or arginine residues. The products of that trypsin digest are separated by high pressure liquid chromatography, then they are ionized into the mass spectrometer. The trypsin digestion products are then fragmented in the mass spectrometer and (partially) sequenced based on the separation of the m/z ratios of their fragment ions. A search program is then used to match the digestion product to its original protein based on a theoretical digestion of a set of proteins. Because trypsin cuts at two amino acid residues, a protein digested by trypsin is usually cut into many pieces.  By digesting many proteins with trypsin at once, a bottom-up approach can cause a complex mixture as the sample.  The complex mixture places a strong emphasis on separation before ionization into the mass spectrometer and the ability for the mass spectrometer to isolate a particular m/z for fragmentation.  This emphasis, along with the fact that tryptic fragments generally have a mass of under 5000 Da, mean this type of analysis is generally performed with an ion-trap instrument with an inline high pressure liquid chromatography system [22].

The bottom-up approach is best used on systems that have been studied in the past and have the necessary sequence information in the publicly available proteome/genome databases .  Because of the complexity of the analyte and the fact that some digestion products ionize better than others, it is unlikely that the data from a bottom-up experiment will return 100% sequence coverage from every protein studied.  Database searching makes up for this lack of coverage by filling the gaps based on its theoretical digest.  However, if the protein hasn't been studied or previously sequenced, then there is nothing in the database to match the digestion products that can be identified.  There would be no indication of where the fragments fit into the amino acid sequence of the protein (except for the terminus not beginning or ending with cleavage site residue).  Furthermore, if the protein is modified, then one or more fragments will have a shifted mass from what is expected, unless that modification is already known to exist in that protein.  Therefore, the bottom-up approach is made much more difficult without prior knowledge about the target protein.

The top-down proteomic approach ionizes the whole protein into the mass spectrometer, fragments the protein, then matches the fragments generated by isolating and fragmenting the molecular ion of an intact protein against a database of *in-situ* fragment masses from a database of intact protein sequences. By measuring the mass of the whole protein first, the top-down approach matches the protein mass first then confirms that identity through matching the masses of the fragments against the masses of the b and y fragments from that protein [23]. Furthermore, top down can better analyze previously unknown modifications or amino acid mutations through the molecular weight of the target. For example, if the observed mass is 80 Da more than the theoretical mass of a protein, then there is a good chance that there is a previously unknown phosphorylation on the protein.

The top-down approach has to be done with a mass spectrometer with a large molecular weight range. Most proteins weigh more than 6000 Da, meaning that the conventional linear ion trap would not be suitable for this type of analysis. Top-down mass spectrometry is generally done with either MALDI [5, 24], or with the high resolution FT mass spectrometers [25, 26]. Using the high resolution hybrid FT mass spectrometers, the data will have many peaks for the same protein because of different charges on the isotope envelope. Therefore this data will need to be decharged and deconvoluted before analysis by the mass spectrometrist. This deconvolution can be done by software comparing the highly resolved isotope envelope to a theoretical envelope based on the average mass of an amino acid. One such program, called THRASH [27], simplifies the spectra so that each protein or fragment only has one peak associated with it, as opposed to a cluster of isotope peaks.  Once deconvoluted, the mixture is not as complex and top-down is a faster approach than bottom-up.  Even with advances in trypsin technology, digestion still takes minutes to hours.  Using the top-down approach, fragmentation takes less than a second.  A barrier to the analysis of top-down proteomics is that, since top-down has been a relatively new technology, most of the current database search programs were created for the bottom-up approach and do not support analysis of top-down data.

There has been a third proteomic approach that has been revived in popularity[28, 29], middle-down analysis.  As the name implies, it takes the best aspects of the two previous approaches by using enzymatic or chemical cleavage to cleave the protein fewer times than bottom-up to create fewer, longer peptides for analysis. These longer peptides still carry many charges, so the high mass range is still necessary for complete analysis.  However, the data is better suited for the current search programs than the whole proteins in the top-down data.  By using a microwave-assisted chemical cleavage, the protein sample can be cleaved in 5 minutes.  Generally, the protein is only cleaved at 1 amino acid, as opposed to 2 with trypsin, so the mixture is only half as complex as it would be using the bottom-up approach.  This is also true for the enzyme Lys-C.  As the name implies, the enzyme digests the protein at the C-terminal end of the lysine residues and provides longer polypeptides.

### *Data Analysis*

All three of the proteomic approaches that were described in the last section can generate thousands of mass spectra in a single experiment.  The handling of that data takes on an increased importance.  Furthermore, all three proteomic approaches rely on matching the observed mass and observed fragments to theoretical *in-situ* cleavage or fragmentation of a protein database. Search programs have been created to provide statistics to how well the observed data matches the theoretical data and how these matches compare to a random match or a false match.

MASCOT (www.matrixscience.com) is a widely used search engine used to match bottom-up data to the correct protein and provide statistics about how probable is the match [30]. The user inputs which enzyme was used, which database to use, what possible modifications can be present, what mass tolerance should be used, and which instrument was used. The search program then performs the *in-situ* digest based on the inputs, and the program provides the matches in a results table. This result gives a score for each matched peptide and a score for the probability of a protein being present based on how many peptides are matched from that protein. Each peptide is also given an E-value, which measures how likely the match is opposed to a match to a random peptide. The higher the score and lower the E-value, then higher confidence can be given to the match of observed data and theoretical data. The observed data can also be searched against a "decoy" database (e.g. a database made of the reverse amino acid sequence of the constituent proteins), in order to calculate a false discovery rate.

ProSight PC is the search program that is widely used with the high mass accuracy FT top-down approach to match the protein and its fragment ions against a database of theoretical fragments [31]. First, the program uses the THRASH algorithm to deconvolute the multiply charged precursor and fragment data so that each protein and each fragment only have one mass to search. Then, the program uses an inputted mass tolerance (with lower tolerances for the fragment data due to the high mass accuracy of FT mass spectrometers) to allow analysis of top-down mass spectrometry done on a chromatographic time scale. Like MASCOT, ProSight PC provides the match to the theoretical protein based on matching the fragment masses within the assigned tolerance. Also like MASCOT, ProSight assigns a score and E-value to measure the confidence of the match versus a match to a random protein. However, ProSight PC is different in that there are no mass restrictions on the precursor to allow for better handling of top-down data. Furthermore, while MASCOT will only allow for defined modifications before a search, ProSight PC has a Sequence Gazer tool that allows the user to check for a previously unknown modification after the search has been completed. This is useful in top-down when dealing with an organism that is lacking some of the information needed in a bottom-up experiment.

### *Objectives*

Using the top-down proteomic approach with high accuracy precursor and fragment ions from a hybrid LTQ-Orbitrap mass spectrometer, we ask to what extent can an organism without a sequenced genome or proteome be characterized. As previously mentioned in this introduction, both the bottom-up and top-down approaches use theoretical masses based on a database of known amino acid sequence to identify the protein or organism observed in a protein mass spectrometry study. However, few microorganisms, only around 1200 archea and bacteria [32], have been sequenced and have had their sequences validated by the community. Therefore, there needs to be a method developed to be able to identify proteins from organisms without sequenced genomes in order to better study those bacteria and archea that have yet to be sequenced.  This thesis proposes using proteins from other species but believed to be homologous as the basis for database matching using a top-down approach with high mass accuracy.  In some cases, near homology will be shown to allow protein identification  The use of high mass accuracy will be used to localize and propose changes to those matches that have precursor mass differences within a reasonable tolerance.

# Chapter 2: Preliminary Top-Down Study of Bovine Ubiquitin

## *Introduction*

Ubiquitin is a highly studied, highly conserved, small protein that is expressed in most, if not all, eukaryotic cells. Most of the early structural and functional work was done by Irwin Rose and colleagues [33], and this work led to the Nobel Prize in Chemistry in 2004. Since then, this protein and its pathway have been one of the most studied systems in scientific research. The ubiquitin protein is relatively small, only having a mass of about 8,500 Daltons. Because its structure and amino acid sequence have been studied so thoroughly, it becomes an ideal analyte when testing new structural analysis instruments and approaches, in order to determine if the structure seen by the new approach matches the well-known ubiquitin structure. This has been the case with top-down mass spectrometry [34-36] using the MALDI-TOF analyzer and FT-ICR instruments.

As stated in the previous chapter, the Orbitrap mass analyzer was invented by Makarov in 2000 and later commercialized by Thermo in order to allow for high resolution, high mass accuracy analysis without having to buy and maintain a large electromagnet for use with an FT-ICR instrument. Its place as a relatively new mass analyzer meant there weren't many published procedures[37] on how to elucidate the structure of whole proteins at the time the LTQ-Orbitrap XL mass analyzer was installed at the University of Maryland. Therefore, I felt that using ubiquitin as a test of the capabilities of top-down mass spectrometry on whole proteins would be insightful.  Furthermore, the Fenselau lab's general knowledge of the software that was used to tune the features of the mass spectrometer and analyze the data was very little at that time.

The objectives of this experiment were to detect the molecular mass of ubiquitin at high charge states, fragment the ubiquitin into its b- and y- fragment ion, and use the Thermo Scientific software to be able to quickly analyze the data and be able to tell how much of the protein was covered by the fragment ions.

## *Experimental*

### Sample preparation

One milligram of lyophilized bovine ubiquitin powder (SigmaAldrich, St. Louis, MO), was dissolved in one milliliter of a 50% water, 40% acetonitrile, 10% acetic acid mixture.  This solution was then diluted by pipetting 300 microliters into 2700 microliters of the same 50% water, 40% acetonitrile, 10% acetic acid mixture, creating three milliliters of a 0.1 milligram per milliliter ubiquitin solution.

### Direct injection mass spectrometry

Before injection of the ubiquitin solution, the mass spectrometer was first mass calibrated by injecting a calibration mixture of caffeine, the quad-peptide MRFA, and Ultramark polymer solution, sprayed with 60% acetonitrile and 10% acid using electrospray ionization.  The LTQ-Orbitrap XL was automatically calibrated on the m/z ratios of 195.1 (for caffeine) 524.3 (for MRFA), 1222.1, 1322.1, 1522.1, and 1822.1 (for Ultramark polymer), for both the linear ion trap and the Orbitrap analyzers until the masses differed by less than 3 ppm in the Orbitrap analyzer. A 500 microliter syringe was then filled with the ubiquitin solution and interfaced to the electrospray. The syringe pump was set to spray 1 microliter per minute of the solution. The lenses and voltages on the mass spectrometer leading from the inlet to the linear ion trap were automatically tuned at m/z 857.37 (the +10 charge state of ubiquitin). The mass spectrometer was set to record a cycle of four spectra, one precursor ion acquired in the Orbitrap mass spectrometer and the fragmentation spectra of the three most abundant ions acquired in the linear ion trap. Collisionally induced dissociation was set at a normalized collision energy of 35% (Thermo's arbitrary units) with an isolation window of 3 m/z and the default activation time of 30 milliseconds. An exclusion list was used to make sure there was no carry over from the Ultramark polymer ions used in calibration. Dynamic exclusion was enabled with a repeat count of 2 and a duration of 15 seconds to prevent oversampling of any one particular charge state of ubiquitin. The LTQ-Orbitrap was set to collect spectra for 10 minutes.

### Data Analysis

The amino acid structure of ubiquitin was imported into the Bioworks software (ThermoFisher, San Jose, CA), and the theoretical b- and y- ion fragment m/z ratios were calculated with an upper limit the six most abundant charge states (+8 to +13) with an m/z range of 0 to 2000. For instance, for the +8 charge state, all fragments were calculated for +1 to +8. For the +13 charge, all fragments were calculated from +1 to +13 charge. The datafile was then imported and the fragmentation spectra with the most peaks were manually picked out. Those fragmentation spectra were then matched against the theoretical m/z ratios of all charge states up to the charge of the isolated precursor.  Because fragmentation spectra were acquired in the more sensitive but less accurate linear ion trap, the fragments were matched to the spectrum’s peaks with a tolerance of 0.6 Daltons.

## *Results and Discussion*

### Charge State Determination

The LTQ-Orbitrap XL collected 721 spectra over the ten minute acquisition time, which meant that one spectrum was acquired about every 800 milliseconds.  As noted in the experimental, the ubiquitin precursor consistently was acquired with a charge state range of +8 to +13.  In some cases over the ten minute acquisition, the +7 and +14 charge states were also visible, but in less than 10% of the most abundant charge state, which was usually +12.  With the high resolution and high mass accuracy of the Orbitrap, the acquisition software is able to calculate the charge on a particular set of peaks on the fly by measuring the distance between isotope clusters and matching known spacing.  This is shown in Figure 2.1, with the z below each mass-to-charge ratio being the charge.

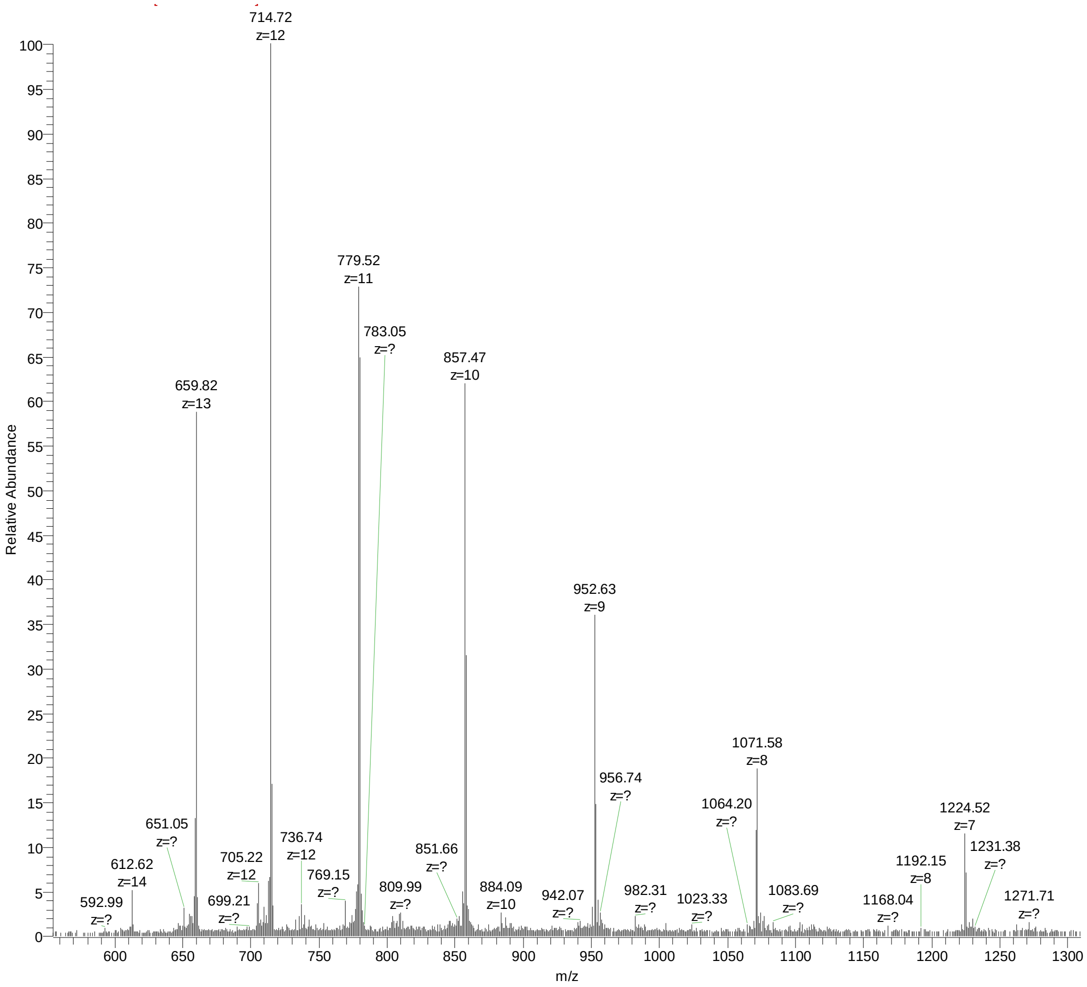


Figure 2.1-Precursor spectrum of ubiquitin showing charge states +7 through +14.

Fragmentation

Fragmentation spectra yielding many b- and y- ions were relegated to only the +8 and +9charge states, even though the +11, +12, and +13 were generally more abundant in the precursor spectra. Isolation and fragmentation of the +8 charge state at m/z 1071.58 yielded 17 identified fragments (Figure 2.2) and the isolation and fragmentation of the +9 charge state at m/z 952.63 yielded 11 identified fragments (Figure 2.3). The +10 through +13 charge states only yielded 14 combined, with the +12 charge state accounting for 5 of those identified fragments. This is summarized in Table 2.1.

These results mirror a study done in 2001 by Reid, McLucky, and colleagues where they used a quadropole ion trap interfaced with homebuilt ion/ion chemistry modifications to control charge states [34]. These were used to study the fragmentation of ubiquitin from the +1 charge state to the +12. In their study, Reid and McLucky also found that the +8 and +9 charge states yielded the best fragmentation. They postulated that it was due to the "mobile proton" theory [38, 39]. In short, the "mobile proton" theory states that at lower charge states, the protons tend to aggregate and stay at the most basic residues. When the charge on these basic residues is filled, then the remaining protons can move from amino acid to amino acid, depending on the particular molecule, which leads to higher sequence coverage by the fragments. Eventually, however, the molecule reaches a limit of charges at which time the Coulombic repulsion of the protons leads to a more regimented fragmentation pattern again.

The largest fragment ion confidently matched to the possible ubiquitin b- and y-ions in Bioworks was a y60 ion of charge state +7 (m/z 1072.4) which was fragmented off of the +8 precursor ion of m/z 1071.58. Figure 2.4 shows the ubiquitin amino acid sequence, the 3-D structure, and the highlighted fragment in yellow in both.

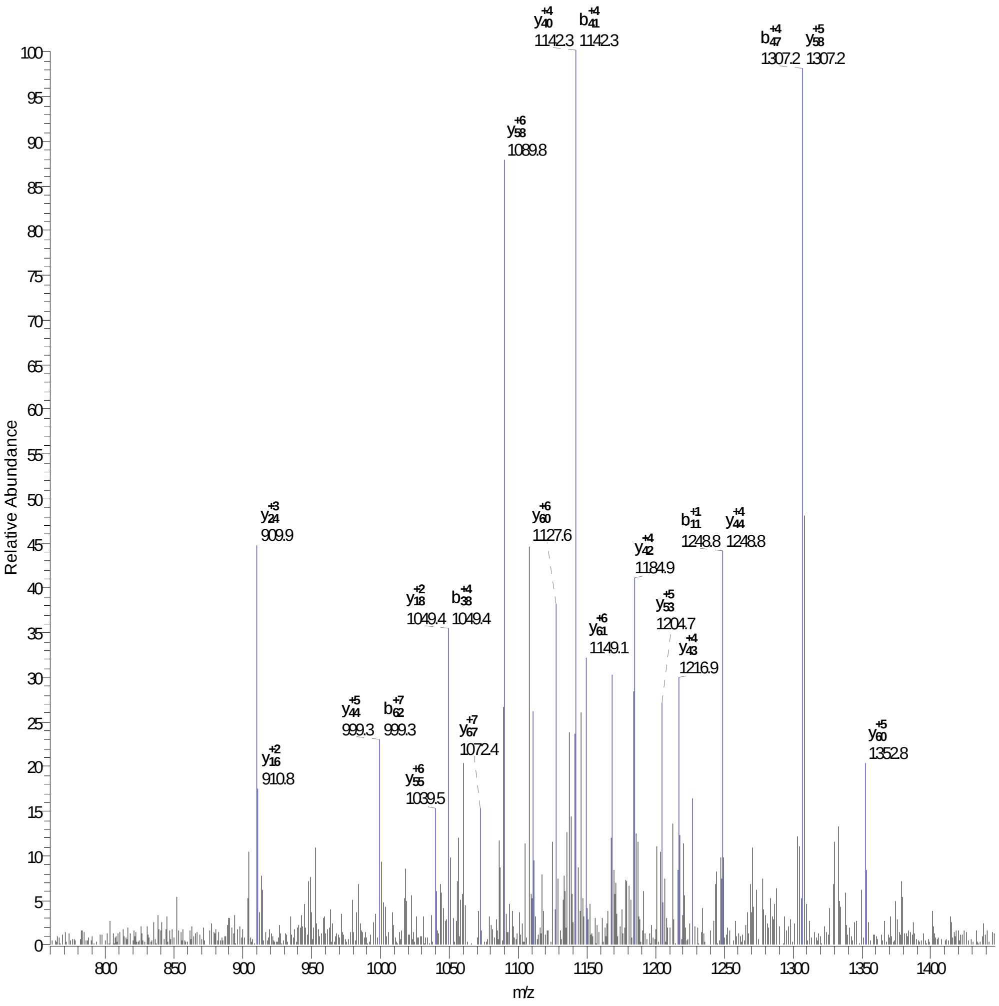


Figure 2.2-Fragmentation spectra of +8 charge state of ubiquitin with matched b- and y- ions highlighted in blue

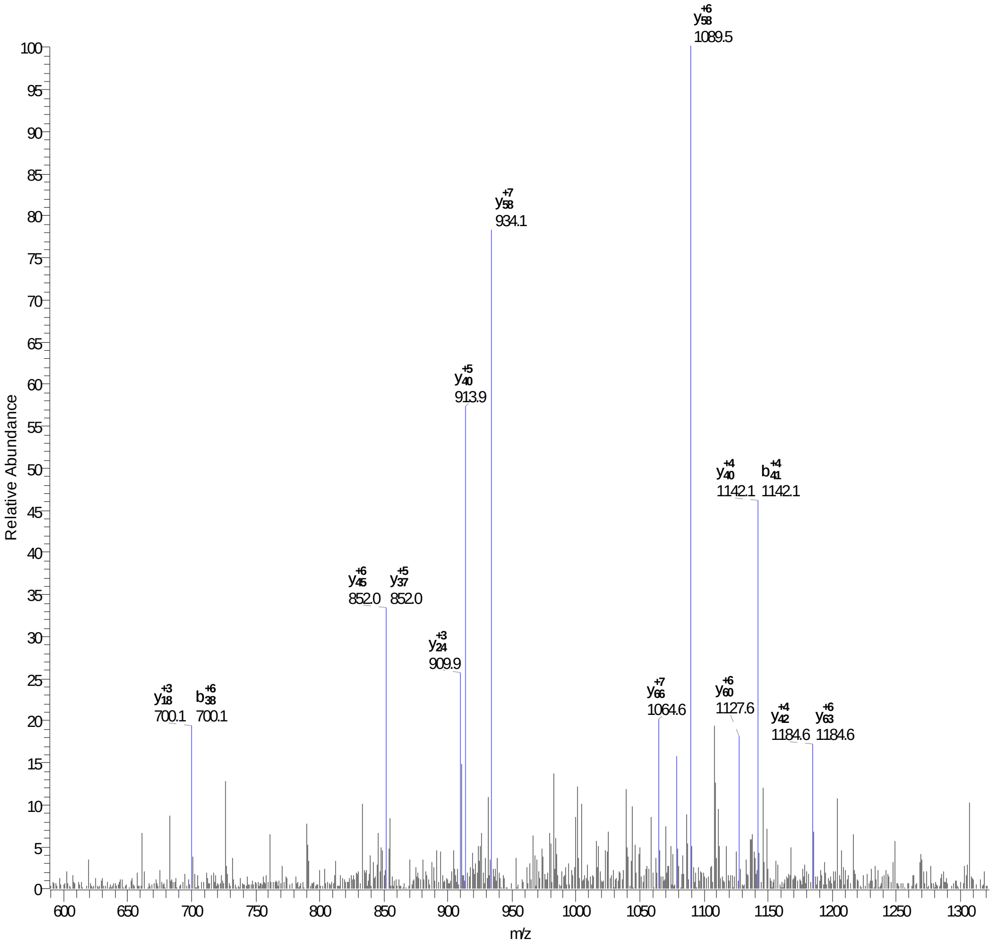
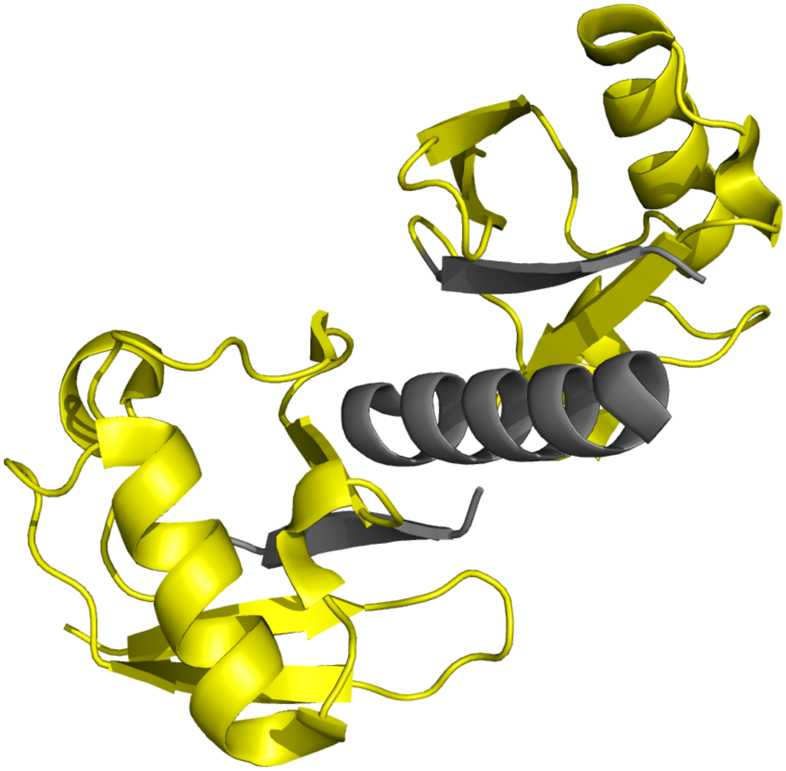


Figure 2.3-Fragmentation of the +9 charge state of ubiquitin with the matched b and y ion fragments highlighted in blue.

| Observed m/z | Calculated m/z | Mass Difference (Accuracy) | Charge | # of Identified Fragments |
|---|---|---|---|---|
| 1071.58 | 1071.60 | 0.02 | 8 | 17 |
| 952.63 | 952.64 | 0.01 | 9 | 11 |
| 857.47 | 857.48 | 0.01 | 10 | 2 |
| 779.52 | 779.62 | 0.10 | 11 | 4 |
| 714.72 | 714.73 | 0.01 | 12 | 5 |
| 659.92 | 659.83 | 0.10 | 13 | 3 |

Table 2.1-Summary of Ubiquitin fragmentation showing precursor m/z ratio and the number of identified charge states.



**MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPP**
**DQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRG**
**G**

Figure 2.4-the largest fragment highlighted in yellow in both the 3-D structure and the 1 letter representation of the amino acid structure of ubiquitin.

*Summary*

This chapter uses the LTQ-Orbitrap XL to study a well known analyte, bovine ubiquitin. By studying the +8 through +13 charge state, I showed that much, if not all, of the protein's amino acid sequence could be confirmed using CID and matching fragmentation spectra to already calculated b- and y- ion fragments. Furthermore, this chapter showed that if an analyte's amino acid sequence is already known and the charge of the precursor is already calculated, then the linear ion trap can be used to detect the fragments.  However, if the analyte is not previously known or has no known amino acid sequence, then the high mass accuracy must be used to identify the fragments.  These results, along with the results found in Reid's work, further confirm that there is a charge state range in every protein that yields the best fragmentation, and that more charges on the protein does not necessarily correlate to more fragmentation.  Therefore, it would be useful to sample many charge states of the same protein in order to determine which charge state yields the best fragmentation. This range of best fragmentation is most likely due to the "mobile proton" theory put forth by Harrison and Dongre.

# Chapter 3: Top Down Analysis of *Yersinia rohdei* Lysates

Taken from Wynne, C., Fenselau, C., Demirev, P.A., Edwards, N. Top Down Identification of Unsequenced Genomes. *Analytical Chemistry* 2009, 81, 9633-9642.

## *Introduction*

Rapid characterization of microorganisms has been considerably studied in the past 15 years. Many of these studies[5, 40] have used proteomic techniques with MALDI ionization to be used in fieldable instruments. These fieldable instruments are generally validated not by studying the possible pathogen, but studying another microorganism that is non-lethal yet shares many of the pathogen's characteristics. However, these test organisms are sometimes not studied as thoroughly as the actual pathogen. Only around 1200 bacteria and archea have their genomic or proteomic sequence in the publicly available database[32], so most of the non-lethal simulants are not in the databases used to match the observed masses with the theoretical sequence. If not all of the genome or proteome of the test organism is known, then some of the proteomic mass spectrometry techniques, such as bottom-up or molecular mass matching, cannot be used to identify the proteins or the organism. Therefore, new techniques must be used to provide the information needed to determine whether a test organism is close enough to validate these fieldable instruments.

The technique used in this study to find the primary sequence information of the target proteins is top-down analysis and identification of proteins from "near neighbor" organisms. Top-down analysis, which was described previously in the introduction, was used to obtain an accurate molecular mass of the protein along with the fragment masses used for extracted sequence tag[41] or for database matching. By constructing a database of proteins from organisms we assume are similar to the target organisms, it can also be assumed that many of the proteins will be homologous. These homologous proteins are what allow for the matching of the fragment masses to b and y ions of a particular protein. Finally, the phylogenic analysis is performed on the target organism based on the number of proteins that match from a particular "related" organism or group of organisms. This is one of the first studies to use top-down analysis on a chromatographic time scale using the LTQ-Orbitrap.

In this particular study, *Bacillus anthracis* Sterne was used in a feasibility test and *Yersinia rohdei* was used as the target organism. *Bacillus anthracis* Sterne is a BioSafety Level (BSL) 1 non-pathogenic bacterium. It is missing the plasmid that codes for the virulent proteins of anthrax. The proteins from this bacterium have been previously studied by the Fenselau lab[42]. *Yersinia rohdei* is another BSL-1 organism used as a simulant for *Yersinia pestis*, the bacterium that causes Bubonic plague[43]. This study was performed before the *Y. rohdei* genome project at the Naval Medical Research Center released a number of whole genome shotgun contigs to Genbank (June 2009).

For the *Yersinia rohdei* study, a MALDI mass spectrum was used to create a mass inclusion list for the high resolution high mass accuracy top-down analysis in order to better study and identify the same proteins that would be used to distinguish this particular bacteria from others using a MALDI-TOF based detection system that our collaborators constructed.

## *Experimental*

### Cell Culture

*Bacillus anthracis* Sterne cells were cultured on Nutrient Broth medium plates (ThermoFisher, Fair Lawn, NJ), setting four different colonies per plate using a tungsten loop sterilize over a flame. These four colonies were spread on the plate using a repeated S-turn motion. The bacteria were then left to grow on shelves in a room temperature controlled to 37°C. Once the bacteria were grown to cover most of the plate, those cells were scraped into 10 ml of broth in 15 ml tubes. These tubes were then incubated overnight in the 37°C room. Cell suspensions were centrifuged at 6000 rpm for 10 minutes, then washed with 3 ml of Milli-Q water, and centrifuged again for 5 minutes at 6000 rpm. This wash step was repeated two additional times, with the supernatant discarded each time. The pellet was then resuspended in 3 ml of 10% formic acid and centrifuged at 10000 rpm for 5 minutes. The supernatant was transferred to a vial for injection into the LC-MS/MS. Eight milliliters of a solution of $4.6x10^8$ cells per milliliters of *Yersinia rohdei* grown at the Johns Hopkins Applied Physics Lab under standard growth conditions[44] was further washed and lysed following the same procedures.

### MALDI-TOF MS Analysis

A Bruker Microflex MALDI-TOF (Bruker Daltonics, Billerica, MA) mass spectrometer was used to create a signature spectrum for intact *Yersinia rohdei*. This signature was created by creating an individual spectrum by shooting the laser 600 times at a sample well, then averaging multiple spectra. The Microflex had a resolving power of about 1000 at full width half maximum using the positive ion linear mode following standard sample preparation and data acquisition procedures[45].

### LC-MS/MS Analysis

An Accela HPLC unit (ThermoFisher) was used for the online separation of the intact proteins from the lysate prior to electrospray ionization into the mass spectrometer. The HPLC unit consists of two solvents. Solvent A was composed of 95% water, 4.9% acetonitrile, and 0.1% formic acid. Solvent B was composed of 95% acetonitrile, 4.9% water, and 0.1% formic acid. All solvents used were of HPLC grade (ThermoFisher). The proteins from both the *Bacillus anthracis* Sterne and *Yersinia rohdei* were separated on the same 1 millimeter inner diameter, 15 centimeter length BioBasic C-8 column (ThermoFisher). The gradient started out at 95% A for 5 minutes. This was followed by a linear climb from 5% to 65% solvent B over 45 minutes. The gradient was held at 65% B for 5 minutes, then the gradient quickly dropped back to the original 5% B for re-equilibration. This HPLC system was inline to the LTQ-Orbitrap XL mass spectrometer (ThermoFisher, San Jose, CA) for MS/MS analysis. Masses of both the precursor and fragment ions were collected at 30,000 resolving power at 400 m/z in the Orbitrap mass analyzer. Four product ion scans were acquired for every precursor scan, two based on the most abundant ions on a mass inclusion list and two based on the most abundant ions in the precursor spectrum as a whole. The inclusion mass list was based on the masses of the high abundance ions from the MALDI-TOF spectrum with charges of +5 to +10. CID was carried out in the LTQ analyzer using helium gas at the 35% activation setting. Each cycle of high resolution precursor and product ion scans took approximately 600 milliseconds. Dynamic exclusion was implemented with a 10 second exclusion period during which precursor ions were not resampled even if they were the most abundant in the preceding precursor spectrum. MS/MS was only performed on species with known charge states of +3 or higher.

Protein Identification

ProSight PC 2.0[31] (ThermoFisher) was used to decharge precursor and product ions via the THRASH[27] algorithm and to search the MS/MS spectra against a custom protein sequence database. Experimental measurements were compared to the average molecular weights of theoretical precursors and the monoisotopic molecular weights of theoretical fragments. The precursor mass tolerance was set to 150 Da to allow for N-terminal methionine cleavage. Fragment ion mass tolerance was set to 15 ppm. For the analysis of *Bacillus anthracis* Sterne spectra, a custom sequence database was constructed, containing all proteins from *B. anthracis* Sterne, *Bacillus thuringiensis* konkukian, *Bacillus cereus* AH167, and *Bacillus subtillus* 168 available in the Swiss-Prot database (Version 57.2, 5/5/09). For the analysis of *Y. rohdei* spectra, the custom database was composed of the protein sequences from *Yersinia* species, *Salmonella typhimurium*, *Escherichia coli*, *Shigella sonnei*, *Klebsiella pneumoniae*, *Enterobacter* sp. 638, and the partial proteome from *Enterobacter aggloramerans* (*Erwinia herbicola*). Identified proteins were checked for membership in highly homologous protein families by collecting and aligning cross-species orthologues, suing BlastP[46] and ClustalW[47].

Database matching using ProSight PC 2.0 is based on three values corresponding to the likelihood of providing a match of MS/MS data to random or generic amino acid sequences of the same quality as the match provided by the search program. The first value, x, is the probability of matching a particular m/z value of a tandem mass spectrum to a generic amino acid given the user defined criteria that was used in the search[48]. Equation 1 shows this calculation in three parts.

$$x = \frac{2^{m+1} \times (M_a \times 2)}{111.1} \qquad \text{(Equation 1)}$$

The first part, $2^{m+1}$, describes the maximum amount of fragments that can occur per each fragmentation of the protein with m equaling the number of modifications included in the search. This value includes both post-translational modifications, like oxidation of methionine, as well as sequence substitutions. For this experiment, m equals two, since the only modifications that were included in the search were protein N-terminal formylation and protein N-terminal acetylation. The first part is multiplied by the mass window available for the fragment matching, which is double the mass accuracy ($M_a$). In this experiment, the $M_a$ value would be 0.3 Daltons, or fifteen parts per million of a twenty thousand Dalton protein. The multiplication of these two parts are divided by the mass of an “averigine” residue[49]. The “averigine” residue mass is the weighted average mass of the twenty amino acids. In both the *Bacillus anthracis Sterne* and *Yersinia rohdei* experiments, the x-value is 0.0432.

The x value becomes a factor in the Poisson based probability of acquiring as good of a match between observed and matched fragments by chance. This p-score is calculated by Equation 2:

$$p(n) = 1 - \sum_{i=0}^{n-1} \frac{e^{-xf}(xf)^i}{i!} \qquad \text{(Equation 2)}$$

x is the value calculated from Equation 1, f is the number of observed fragments, and n is the number of fragments matched to the sequence from the protein database. Multiplying the x value by the number of observed fragments provides how many randomly matched fragments can be expected in the spectrum, and each iteration of the equation provides the distribution of successive random matches. The Poisson distribution allows n to be ever increasing, so one minus the distribution is used to determine the probability of a result as good coming by chance.

The Expectation value (or e-value) incorporates the size of the database into its calculation. This value determines how many sequences in the database used will provide matches to the fragment ions with p values of equal or better value. The e-value is calculated using a simple equation shown in Equation 3:

$$E = N \times p(n) \quad \text{(Equation 3)}$$

where p(n) is the calculated p value while N is the number of sequences used to construct the database used for the search. In the *Yersinia rohdei* experiment, this N value was 32901. This expectation value is the reported metric to compare how well a particular set of calculated b- and y-ions will match the observed masses in a MS/MS spectrum because this value is built from the both the x-value and the p-score. The x-value measures the randomness of a particular m/z value from the fragment spectrum. The p-score measures uses the x-value to determine how many random matches would occur from the same number of peaks, and the E-value determines how many sequences from the database would provide spectra generating p-values of equal or better value. All three of these models were tested by Meng and associates against data from 10 randomly selected *Methanococcus jannaschii* proteins from a database of 1,796. Meng demonstrated that the predicted number of random matches from the calculations discussed above matched within 5% the number of spurious matches from his empirical MS/MS data.

### Phylogenetic Analysis

The Rapid Microorganism Identification Database (RMIDb)[50], created by my bioinformatics collaborator Dr. Nathan Edwards, was used to construct a set of all Swiss-Prot, TrEMBL, RefSeq, Genbank, JCVI's CMR[51], and aggressive Glimmer3[52] predicted protein sequences from the *Enterobacteriaceae* family with molecular weights between 4000 and 16000 Daltons, grouped by PFam[53] protein family assignment. Sequences corresponding to the families of the 10 identified *Y. rohdei* proteins were extracted, and *Enterobacteriaceae* species with extracted protein sequences in all 10 families were identified. For each of these 27 species and each protein family, the protein sequence matching the identified *Y. rohdei* sequence best was selected using BlastP, and the selected sequences were concatenated in a predetermined order for phylogenetic analysis, using the web-server phylogeny.fr[54]. Similarly, identified *Y. rohdei* protein sequences were concatenated in the same order and added to the phylogeny analysis. The resulting 28 meta-sequences ranged from 759 to 770 amino-acids in length.

For the phylogenetic analysis using the traditional 16S-RNA sequences, the respective sequences were downloaded from the Ribosomal Database Project[55] for as many of these 28 species described above as possible. 21 out of the 28 species' sequences (including *Y. rohdei*) were assembled for phylogenetic analysis using the phylogeny.fr web-server, ranging in length from 1449 to 1540 nucleotides.

## *Results and Discussion*

### Feasibility Study

The study of a mixed culture (vegetative cells and spores) of the *Bacillus anthracis* Sterne was used to ensure that the strategy of top-down proteomics and database searching against protein sequences in related bacteria could work before moving onto a more complicated and less studied bacteria.  Even though the *B. anthracis* protein sequences were available, they were not included in the database. This ensured that only those proteins from related bacteria were matched to the target. Four protein sequences were identified, each with an E-value of at least 1e-10.  Two of the matched proteins were small acid-soluble proteins from spores, one was a cold-shock protein, and one of the proteins binds to the DNA.  All four were matched to the closely related species *B. cereus* and *B. thuringiensis* and further study showed these proteins are the same across the three *cereus* group species that were included in the database.  No MS/MS spectra matched to proteins from the species that was used as a negative control, *Bacillus subtilis.*  The absence of matches to this *B. subtilis* bacteria suggests that sequences from species that are very closely related to the target must be available for the discussed strategy to work.  A table summarizing the matching of proteins from the *cereus* group is shown in table 3.1.

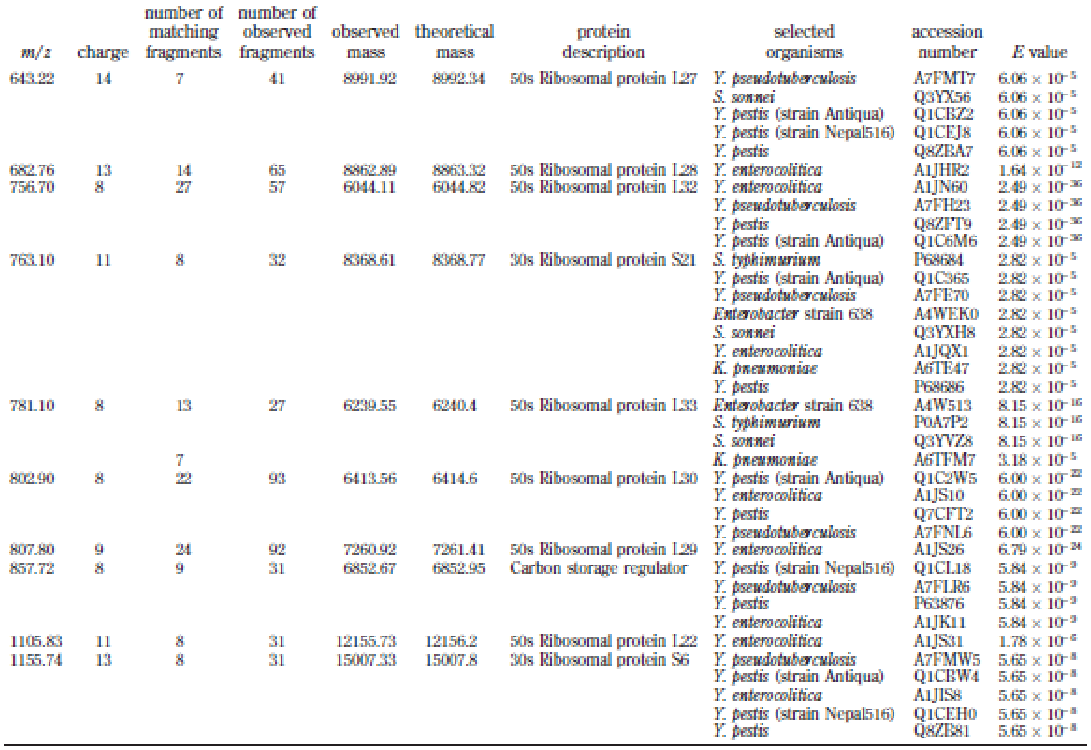
| m/z | Charge | Number of Matching Fragments | Number of Observed Fragments | Theoretical Mass | Observed Mass | Protein Description | Organism | Accession Number | E Value |
|---|---|---|---|---|---|---|---|---|---|
| 643.75 | 15 | 7 | 18 | 9642.06 | 9641.24 | DNA-binding protein HU | *Bacillus cereus* strain AH187<br>*Bacillus thuringiensis* konkukian | YP_002337635<br>YP_035726 | 5.2E-22<br>5.2E-22 |
| 954.95 | 7 | 25 | 57 | 6678.51 | 6678.43 | small, acid-soluble spore protein B | *Bacillus cereus* strain AH187<br>*Bacillus thuringiensis* konkukian | YP_002340704<br>YP_038695 | 3.3E-20<br>3.3E-20 |
| 977.37 | 7 | 17 | 40 | 6834.63 | 6834.45 | unknown, small acid-soluble spore protein | *Bacillus cereus* strain AH187<br>*Bacillus thuringiensis* konkukian | YP_002337009<br>YP_035107 | 2.5E-10<br>2.5E-10 |
| 1053.96 | 7 | 17 | 28 | 7366.13 | 7365.66 | cold shock protein CspB | *Bacillus cereus* strain AH187<br>*Bacillus thuringiensis* konkukian | YP_002339500<br>YP_037619 | 1.9E-25<br>1.9E-25 |

Table 3.1-Summary of *Bacillus anthracis* Sterne protein sequence matches

### *Y. rohdei* Analysis

Unlike the *Bacillus anthracis* Sterne study, there was no assembled, annotated genome available to access with *Yersinia rohdei*. As described by the experimental section, ProSightPC 2.0 was used to deconvolute the precursor and fragment mass spectra from 26 to 39 minutes of the HPLC separation run. This translates to an acetonitrile content of about 35% to 50% acetonitrile. The sequence matches that provide high confident identifications by the ProSight metrics for 10 *Y. rohdei* proteins are summarized in Table 3.2. Five of the ten high confident identifications are matched to proteins that are 100% homologous in more than one of the species that construct the custom database. Table 3.2 also shows the other factors that contribute to the high confident identifications, including the deconvoluted molecular masses, the number of amino acid backbone fragments identified, and the charge that resided on the identified ion. BlastP and ClustalW similarity searches confirmed the sequences across the species that composed the database that matched the target proteins, as were discussed earlier in this document.

Figure 3.1 indicates that 6 of the 10 proteins that were identified by the discussed method coincide with high abundance ions from the MALDI-TOF signature. Most of these intense ions are observed to be ribosomal proteins. Previous studies of vegetative bacteria[7, 42, 56] have shown these intense ions from a MALDI-TOF spectrum have been ribosomal proteins, which are highly abundant and highly basic in bacteria. With the sample preparation of 10% formic acid, those basic proteins will have many positive charges associated with them, making their detection in positive mode mass spectrometry easier.



| $m/z$ | charge | number of matching fragments | number of observed fragments | observed mass | theoretical mass | protein description | selected organisms | accession number | $E$ value |
|---|---|---|---|---|---|---|---|---|---|
| 643.22 | 14 | 7 | 41 | 8991.92 | 8992.34 | 50s Ribosomal protein L27 | *Y. pseudotuberculosis* | A7FMT7 | $6.06 \times 10^{-5}$ |
| | | | | | | | *S. sonnei* | Q3YX56 | $6.06 \times 10^{-5}$ |
| | | | | | | | *Y. pestis* (strain Antiqua) | Q1CBZ2 | $6.06 \times 10^{-5}$ |
| | | | | | | | *Y. pestis* (strain Nepal516) | Q1CEJ8 | $6.06 \times 10^{-5}$ |
| | | | | | | | *Y. pestis* | Q8ZBA7 | $6.06 \times 10^{-5}$ |
| 682.76 | 13 | 14 | 65 | 8862.89 | 8863.32 | 50s Ribosomal protein L28 | *Y. enterocolitica* | A1JHR2 | $1.64 \times 10^{-12}$ |
| 756.70 | 8 | 27 | 57 | 6044.11 | 6044.82 | 50s Ribosomal protein L32 | *Y. enterocolitica* | A1JN60 | $2.49 \times 10^{-36}$ |
| | | | | | | | *Y. pseudotuberculosis* | A7FH23 | $2.49 \times 10^{-36}$ |
| | | | | | | | *Y. pestis* | Q8ZFT9 | $2.49 \times 10^{-36}$ |
| | | | | | | | *Y. pestis* (strain Antiqua) | Q1C6M6 | $2.49 \times 10^{-36}$ |
| 763.10 | 11 | 8 | 32 | 8368.61 | 8368.77 | 30s Ribosomal protein S21 | *S. typhimurium* | P68684 | $2.82 \times 10^{-5}$ |
| | | | | | | | *Y. pestis* (strain Antiqua) | Q1C365 | $2.82 \times 10^{-5}$ |
| | | | | | | | *Y. pseudotuberculosis* | A7FE70 | $2.82 \times 10^{-5}$ |
| | | | | | | | *Enterobacter* strain 638 | A4WEK0 | $2.82 \times 10^{-5}$ |
| | | | | | | | *S. sonnei* | Q3YXH8 | $2.82 \times 10^{-5}$ |
| | | | | | | | *Y. enterocolitica* | A1JQX1 | $2.82 \times 10^{-5}$ |
| | | | | | | | *K. pneumoniae* | A6TE47 | $2.82 \times 10^{-5}$ |
| | | | | | | | *Y. pestis* | P68686 | $2.82 \times 10^{-5}$ |
| 781.10 | 8 | 13 | 27 | 6239.55 | 6240.4 | 50s Ribosomal protein L33 | *Enterobacter* strain 638 | A4W513 | $8.15 \times 10^{-16}$ |
| | | | | | | | *S. typhimurium* | P0A7P2 | $8.15 \times 10^{-16}$ |
| | | | | | | | *S. sonnei* | Q3YVZ8 | $8.15 \times 10^{-16}$ |
| | | 7 | | | | | *K. pneumoniae* | A6TFM7 | $3.18 \times 10^{-5}$ |
| 802.90 | 8 | 22 | 93 | 6413.56 | 6414.6 | 50s Ribosomal protein L30 | *Y. pestis* (strain Antiqua) | Q1C2W5 | $6.00 \times 10^{-22}$ |
| | | | | | | | *Y. enterocolitica* | A1JS10 | $6.00 \times 10^{-22}$ |
| | | | | | | | *Y. pestis* | Q7CFT2 | $6.00 \times 10^{-22}$ |
| | | | | | | | *Y. pseudotuberculosis* | A7FNL6 | $6.00 \times 10^{-22}$ |
| 807.80 | 9 | 24 | 92 | 7260.92 | 7261.41 | 50s Ribosomal protein L29 | *Y. enterocolitica* | A1JS26 | $6.79 \times 10^{-24}$ |
| 857.72 | 8 | 9 | 31 | 6852.67 | 6852.95 | Carbon storage regulator | *Y. pestis* (strain Nepal516) | Q1CL18 | $5.84 \times 10^{-9}$ |
| | | | | | | | *Y. pseudotuberculosis* | A7FLR6 | $5.84 \times 10^{-9}$ |
| | | | | | | | *Y. pestis* | P63876 | $5.84 \times 10^{-9}$ |
| | | | | | | | *Y. enterocolitica* | A1JK11 | $5.84 \times 10^{-9}$ |
| 1105.83 | 11 | 8 | 31 | 12155.73 | 12156.2 | 50s Ribosomal protein L22 | *Y. enterocolitica* | A1JS31 | $1.78 \times 10^{-6}$ |
| 1155.74 | 13 | 8 | 31 | 15007.33 | 15007.8 | 30s Ribosomal protein S6 | *Y. pseudotuberculosis* | A7FMW5 | $5.65 \times 10^{-8}$ |
| | | | | | | | *Y. pestis* (strain Antiqua) | Q1CBW4 | $5.65 \times 10^{-8}$ |
| | | | | | | | *Y. enterocolitica* | A1JIS8 | $5.65 \times 10^{-8}$ |
| | | | | | | | *Y. pestis* (strain Nepal516) | Q1CEH0 | $5.65 \times 10^{-8}$ |
| | | | | | | | *Y. pestis* | Q8ZB81 | $5.65 \times 10^{-8}$ |

Table 3.2-Table of *Yersinia rohdei* protein identifications when searched against a custom database from all *Yersinia* in the Swiss-Prot Database and other Enterobacteriaceae

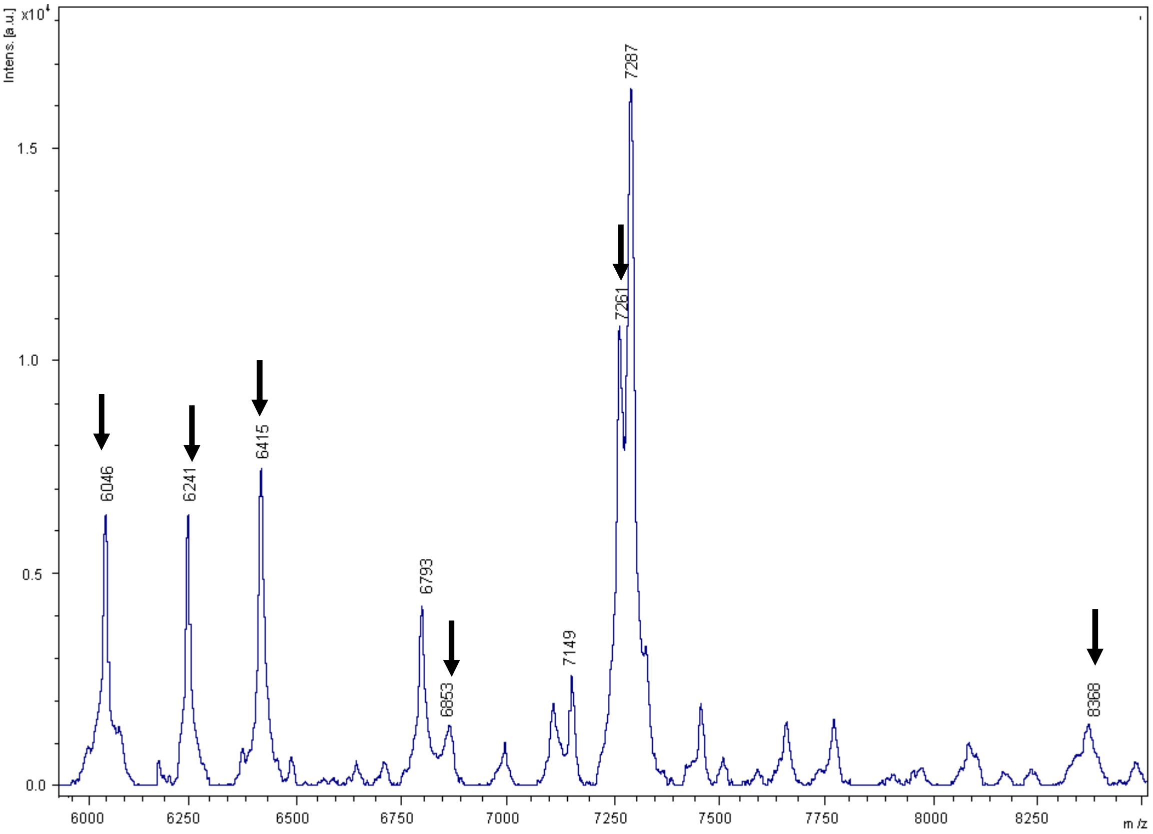


Figure 3.1-Bruker Microflex MALDI-TOF signature of intact proteins from *Yersinia rohdei* with arrows indicating those proteins that were also confidently identified through the LC-MS/MS strategy.

As Table 1 also shows, 3 out of the 10 matches have more than 20 matching fragments to the protein included in the custom database that the target protein was searched against. Many of these fragments are in the middle of the amino acid sequence, showing that the sample preparation allowed for access to the middle of the protein. Figure 3.3 shows one of the example workflow for the database searching. The workflow starts with the CID spectrum. Then, that spectrum is deconvoluted so it is easier to interpret, and then matched against the protein sequences in the database. In this particular case, the deconvoluted MS/MS spectrum was matched to the sequence of ribosomal protein L29 in *Yersinia enterocolitica*, a closely related species. As seen in the figure, there is a sequence tag of 8 consecutive amino acids that are characterized by a b or y ion, and that leads to an identification of high confidence with an E-value of 1e-28. The typical threshold that is used in the method such as this and the threshold that is the default for ProSight PC 2.0 for a confident identification is 1e-4, so 1e-28 is a very high confidence match.

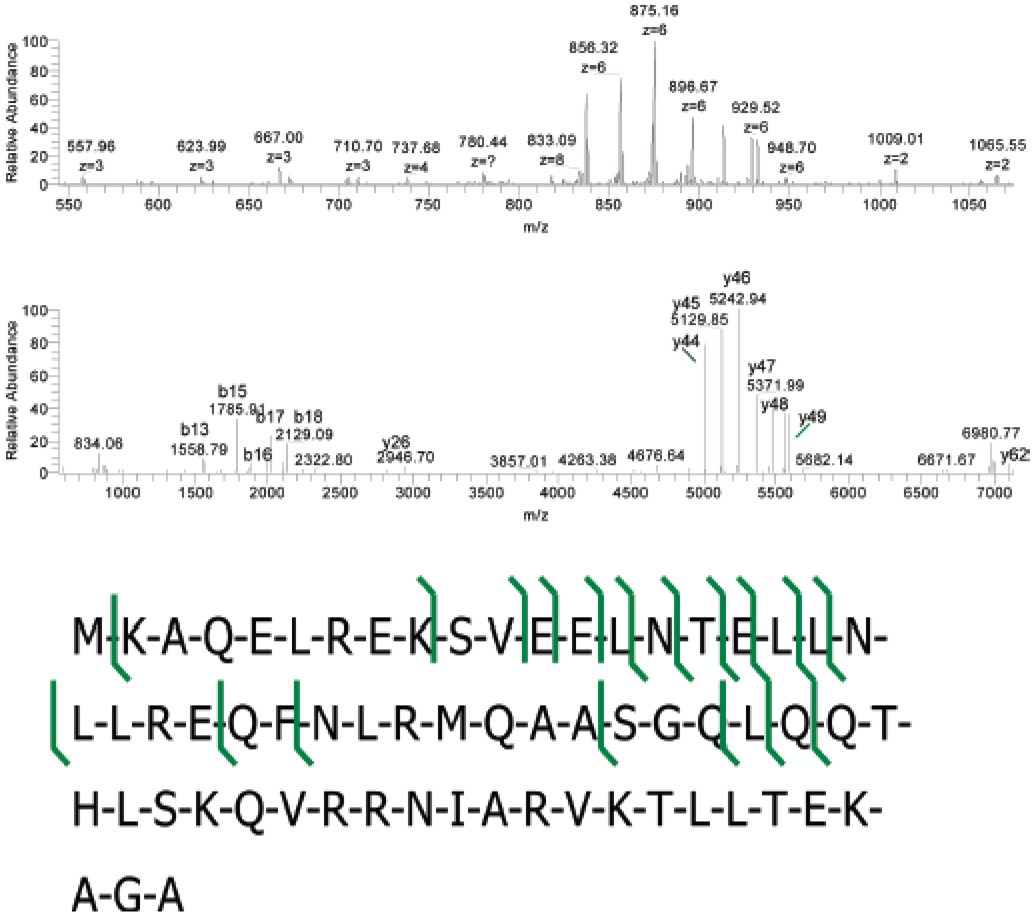


Figure 3.2-Top: MS/MS spectrum of the precursor ion at m/z 807.80 with a 9+ charge state for an intact mass of 7260.92 Daltons. Middle: The same MS/MS spectrum deconvoluted so that all ions are converted to +1 charge state. Bottom: Protein sequence assigned by ProSight PC 2.0, 50s ribosomal protein L29.

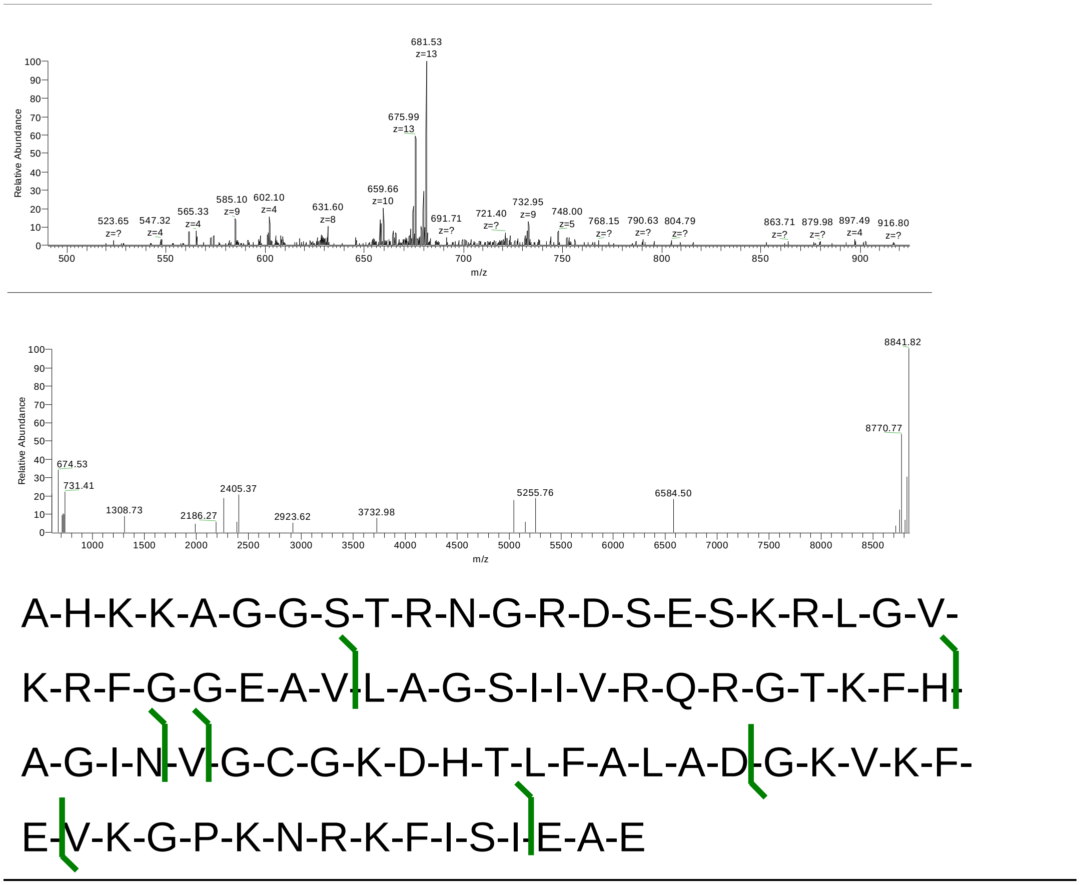


Figure 3.3-Top: MS/MS spectrum of the precursor ion at m/z 643.22 (14+ charge state, intact mass 8991.92 Da)). Middle: The same MS/MS spectrum with all fragment ions converted to zero charge state. Bottom: Protein sequence (ribosomal protein L27, Swiss-Prot A7FMT7) assigned by ProSightPC 2.0 showing observed fragmentation sites.

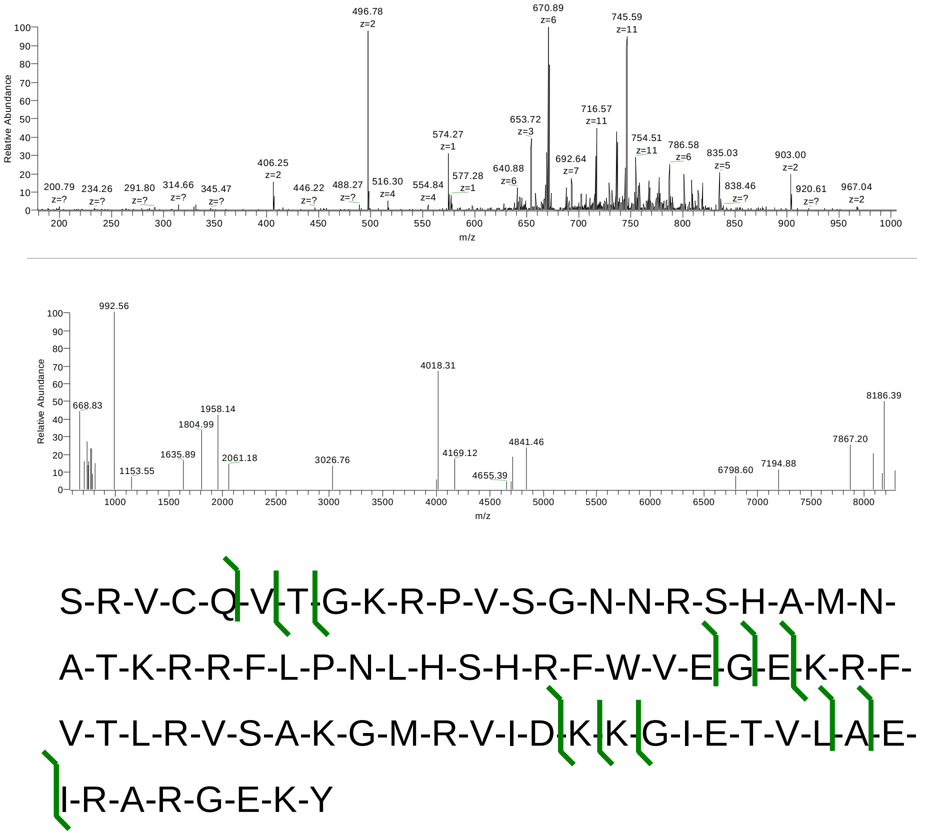


Figure 3.4-Top: MS/MS spectrum of the precursor ion at m/z 682.68 (13+ charge state, intact mass 8862.89 Da)). Middle: The same MS/MS spectrum with all fragment ions converted to zero charge state. Bottom: Protein sequence (ribosomal protein L28, Swiss-Prot A1JHR2) assigned by ProSightPC 2.0 showing observed fragmentation sites.

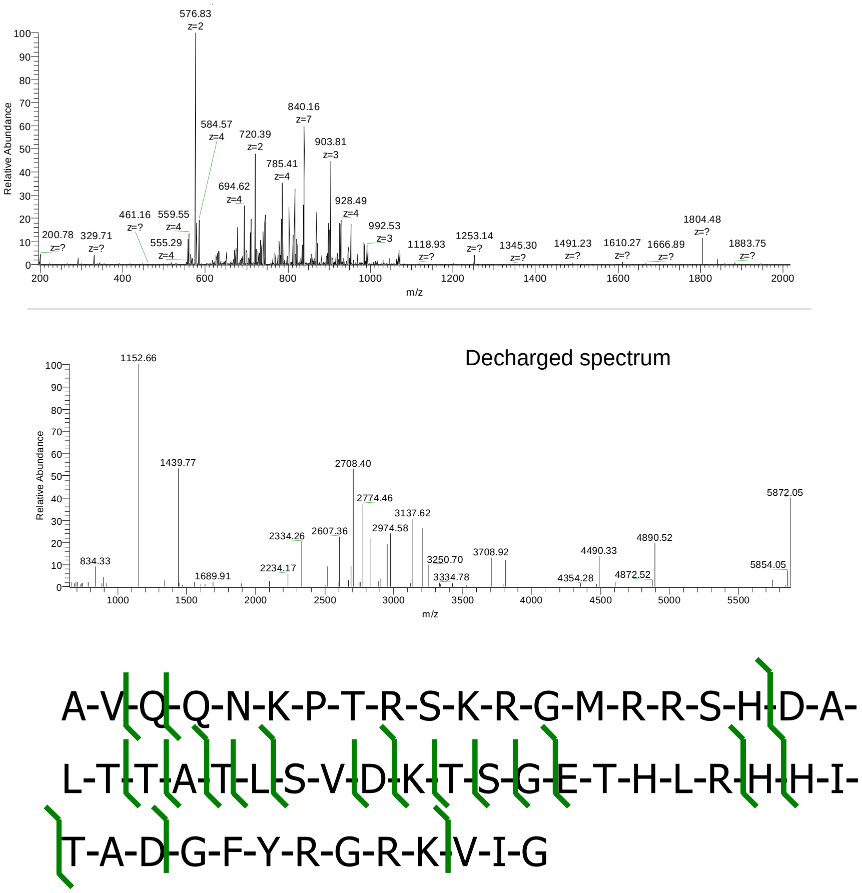


Figure 3.5-Top: MS/MS spectrum of the precursor ion at m/z 756.70 (8+ charge state, intact mass 6044.11 Da). Middle: The same MS/MS spectrum with all fragment ions converted to zero charge state. Bottom: Protein sequence (50s ribosomal protein L32, Swiss-prot A1JN60) assigned by ProSightPC 2.0 showing observed fragmentation sites.

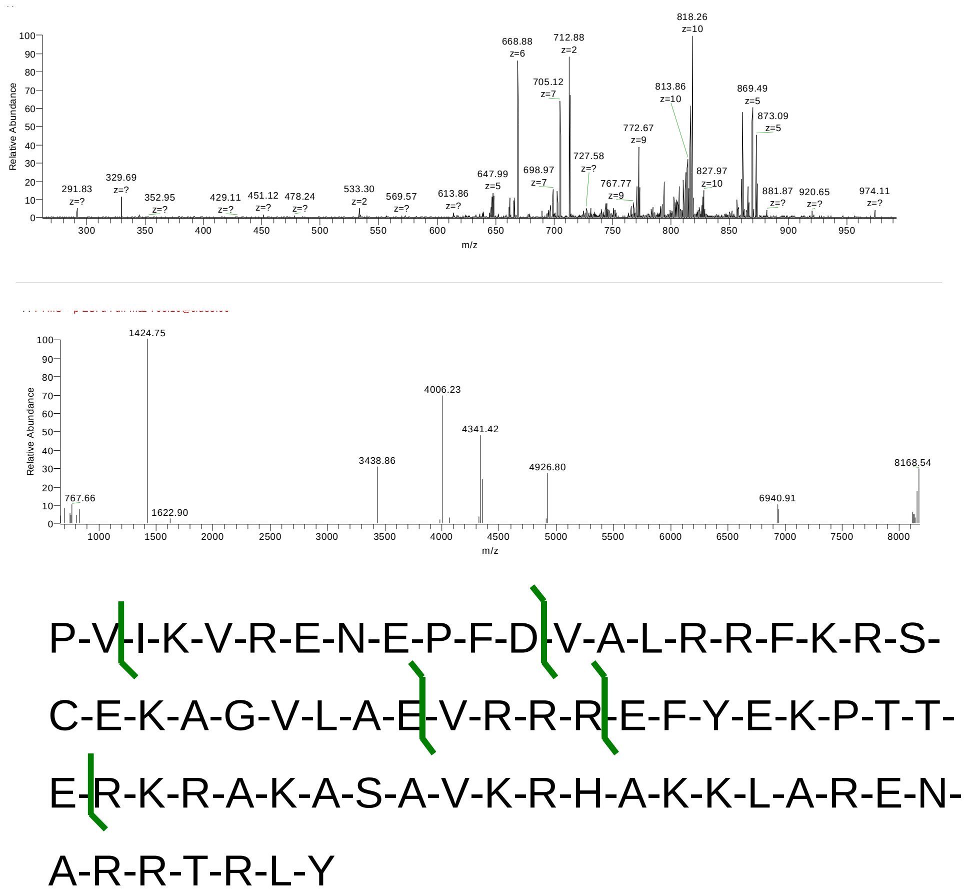


Figure 3.6-Top: MS/MS spectrum of the precursor ion at m/z 763.10 (11+ charge state, intact mass 8368.61 Da)). Middle: The same MS/MS spectrum with all fragment ions converted to zero charge state. Bottom: Protein sequence (ribosomal protein S21, Swiss-Prot P68684) assigned by ProSightPC 2.0 showing observed fragmentation sites.

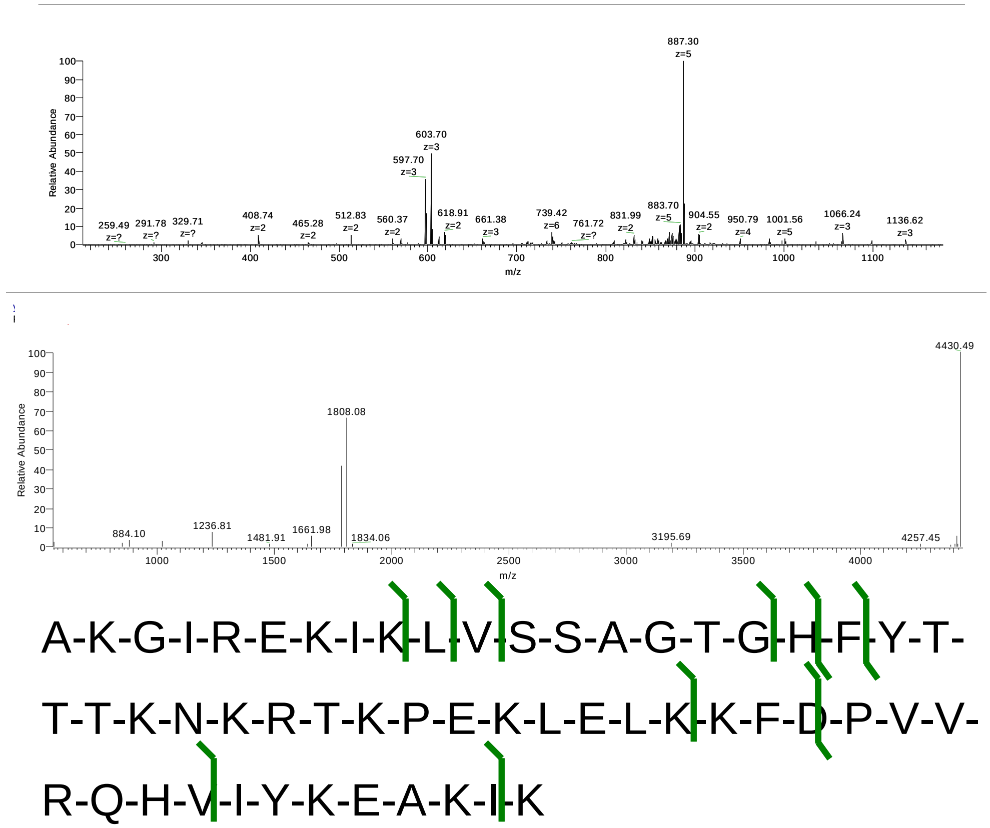


Figure 3.7-Top: MS/MS spectrum of the precursor ion at m/z 781.10 (8+ charge state, intact mass 6239.55 Da)). Middle: The same MS/MS spectrum with all fragment ions converted to zero charge state. Bottom: Protein sequence (ribosomal protein L33, Swiss-Prot A4W513) assigned by ProSightPC 2.0 showing observed fragmentation sites.

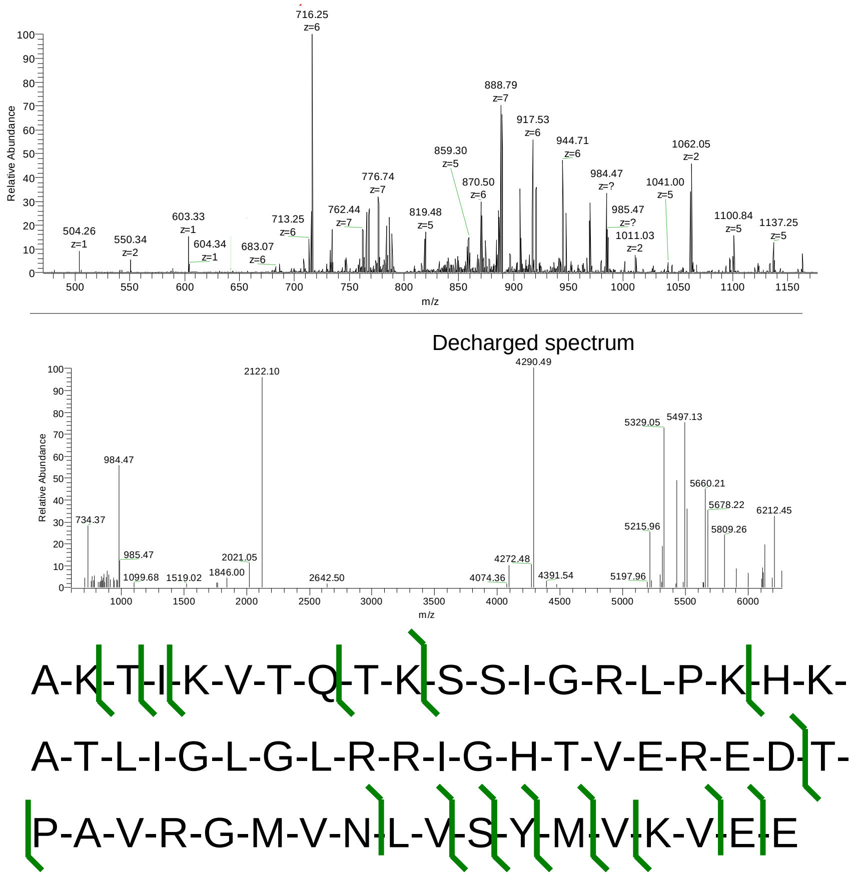


Figure 3.8-Top: MS/MS spectrum of the precursor ion at m/z 802.90 (8+ charge state, intact mass 6413.56 Da). Middle: The same MS/MS spectrum with all fragment ions converted to zero charge state. Bottom: Protein sequence (50s ribosomal protein L30, Swissprot Q1C2W5) assigned by ProSightPC 2.0 showing observed fragmentation sites.

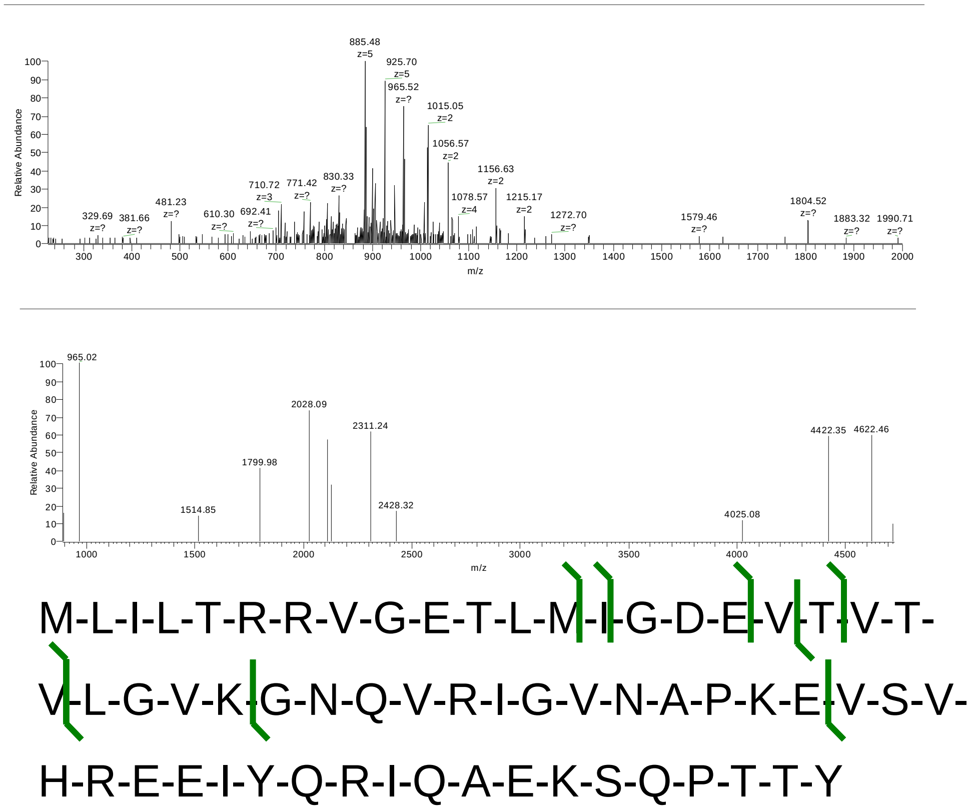


Figure 3.9-Top: MS/MS spectrum of the precursor ion at m/z 857.72 (8+ charge state, intact mass 6852.67 Da)). Middle: The same MS/MS spectrum with all fragment ions converted to zero charge state. Bottom: Protein sequence (carbon storage regulator protein, Swiss-Prot Q1CL18) assigned by ProSightPC 2.0 showing observed fragmentation sites.

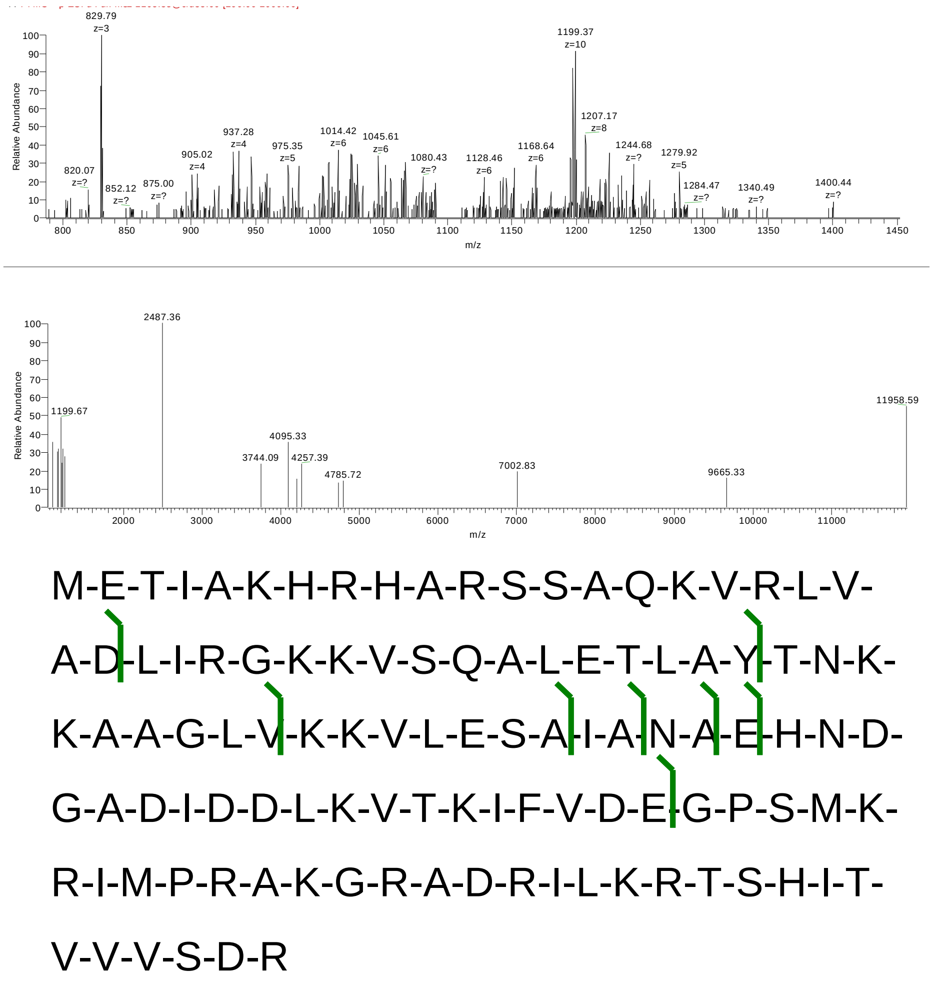


Figure 3.10-Top: MS/MS spectrum of the precursor ion at m/z 1105.83 (11+ charge state, intact mass 12155.73 Da)). Middle: The same MS/MS spectrum with all fragment ions converted to zero charge state. Bottom: Protein sequence (ribosomal protein L22, Swiss-Prot A1JS31) assigned by ProSightPC 2.0 showing observed fragmentation sites.

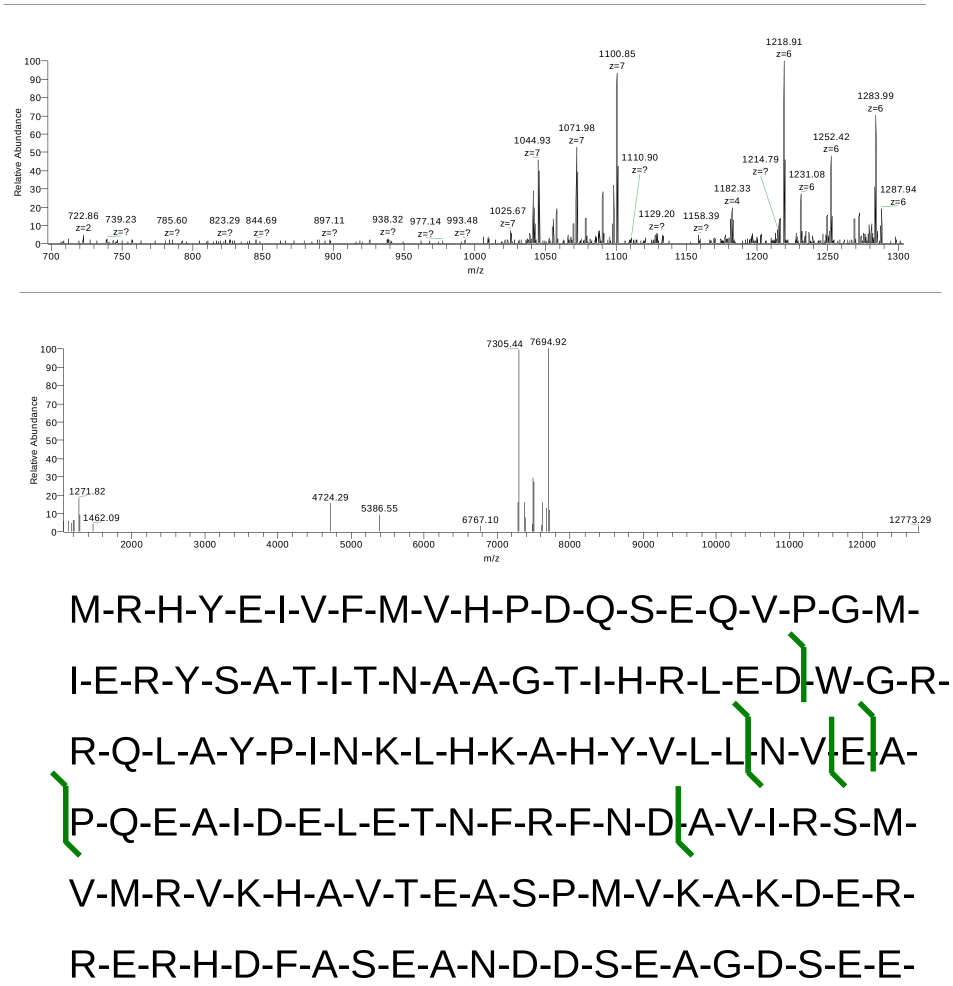
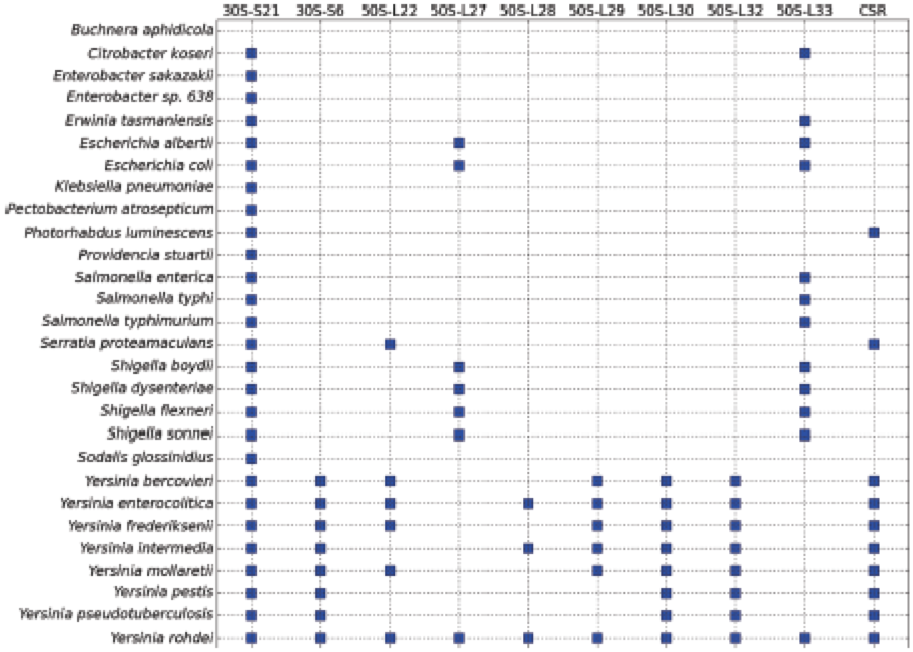


Figure 3.11-Top: MS/MS spectrum of the precursor ion at m/z 1155.74 (13+ charge state, intact mass 15007.33 Da)). Middle: The same MS/MS spectrum with all fragment ions converted to zero charge state. Bottom: Protein sequence (ribosomal protein S6, Swiss-Prot A7FMW5) assigned by ProSightPC 2.0 showing observed fragmentation sites.

## *Phylogenetic Analysis*

### Incidence matrix

Figure 3.3 shows an incidence matrix[57] that is based on the results from Table 3.2 with the proteins whose sequence was matched to the target proteins from *Yersinia rohdei* in the columns. The rows of the incidence matrix show which of the organisms whose sequences were searched against have the same homologous sequence for a protein. The last row of the incidence matrix denotes the target organism, which will match all ten of the proteins. Blue squares across a row show how many protein sequences a particular organism has in common with the target organism. Blue squares down a column show how conserved a protein's sequence is conserved across the organisms included in the custom database. For example, ribosomal protein 30s-S21 is has the most highly conserved sequence since all but one organism in the incidence matrix has the blue square. The least conserved according to this matrix would be the 50s-L29 ribosomal protein, since only two organisms other than *Yersinia rohdei* share the same protein sequence. As expected, the *Yersinia* species that were used in the database had the most protein sequences in common with the target organism, *Yersinia rohdei*. The organism sharing the most protein sequences with the ten matched protein sequences to *Yersinia rohdei* would be *Yersinia enterocolitica* with 8 sequences in common, followed by *Yersinia frederiksenii* and *Yersinia intermedia* with 7 sequences in common. All of the other genera only had 2 or 3 proteins in common with the *Yersinia rohdei*, usually only matching the higher conserved sequences.



| | 30S-S21 | 30S-S6 | 50S-L22 | 50S-L27 | 50S-L28 | 50S-L29 | 50S-L30 | 50S-L32 | 50S-L33 | CSR |
|---|---|---|---|---|---|---|---|---|---|---|
| *Buchnera aphidicola* | | | | | | | | | | |
| *Citrobacter koseri* | ■ | | | | | | | | ■ | |
| *Enterobacter sakazakii* | ■ | | | | | | | | | |
| *Enterobacter sp. 638* | ■ | | | | | | | | | |
| *Erwinia tasmaniensis* | ■ | | | | | | | | ■ | |
| *Escherichia albertii* | ■ | | | ■ | | | | | ■ | |
| *Escherichia coli* | ■ | | | ■ | | | | | ■ | |
| *Klebsiella pneumoniae* | ■ | | | | | | | | | |
| *Pectobacterium atrosepticum* | ■ | | | | | | | | | |
| *Photorhabdus luminescens* | ■ | | | | | | | | | ■ |
| *Providencia stuartii* | ■ | | | | | | | | | |
| *Salmonella enterica* | ■ | | | | | | | | ■ | |
| *Salmonella typhi* | ■ | | | | | | | | ■ | |
| *Salmonella typhimurium* | ■ | | | | | | | | ■ | |
| *Serratia proteamaculans* | ■ | | ■ | | | | | | | ■ |
| *Shigella boydii* | ■ | | | ■ | | | | | ■ | |
| *Shigella dysenteriae* | ■ | | | ■ | | | | | ■ | |
| *Shigella flexneri* | ■ | | | ■ | | | | | ■ | |
| *Shigella sonnei* | ■ | | | ■ | | | | | ■ | |
| *Sodalis glossinidius* | ■ | | | | | | | | | |
| *Yersinia bercovieri* | ■ | ■ | ■ | | | ■ | ■ | ■ | | ■ |
| *Yersinia enterocolitica* | ■ | ■ | ■ | | ■ | ■ | ■ | ■ | | ■ |
| *Yersinia frederiksenii* | ■ | ■ | ■ | | | ■ | ■ | ■ | | ■ |
| *Yersinia intermedia* | ■ | ■ | | | ■ | ■ | ■ | ■ | | ■ |
| *Yersinia mollaretii* | ■ | ■ | ■ | | | ■ | ■ | ■ | | ■ |
| *Yersinia pestis* | ■ | ■ | | | | | ■ | ■ | | ■ |
| *Yersinia pseudotuberculosis* | ■ | ■ | | | | | ■ | ■ | | ■ |
| *Yersinia rohdei* | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |

Figure 3.12-Incidence matrix for observed *Y. rohdei* proteins in *Enterobacteriaceae* species.

### Phylogenetic Trees

Based on the results of the incidence matrix, a phylogenetic tree was created. The FASTA sequences from the identified proteins in *Yersinia rohdei* were added to the protein sequences from the other *Enterobacteriaceae* family members used in the custom database. These protein amino acid sequences were inputted into the webserver phylogeny.fr and then the one-touch application was used to create the phylogenetic trees, as described in the Experimental section. Another phylogenetic tree was created using the traditional 16S-rRNA sequences. These two phylogenetic trees are shown in Figure 3.4. Both trees have *Yersinia rohdei* as a branch next to *Yersinia enterocolitica*. The tree created based on the top-down results separates all the *Yersinia* species as well as the 16S-rRNA.

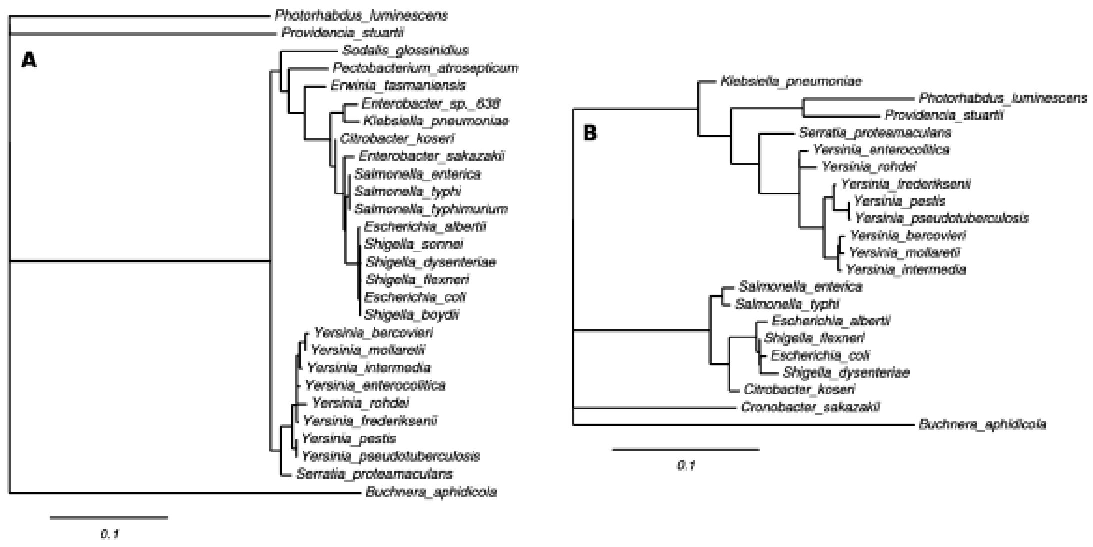


Figure 3.13-Phylogenetic trees based on top-down protein identifications (A) and 16S-rRNA sequence (B)

### Comparison to WGS contigs submitted to Genbank

As noted in the introduction, whole genome shotgun contigs of *Yersinia rohdei* were submitted to Genbank in June 2009. These genome contigs covered genes that coded for 4 out of the 10 proteins that the top-down results were able to identify. Those DNA sequences supported the amino acid sequences of all four proteins after RNA replication and translation. However, the translation start site was only consistent with one protein identified by the top-down results, and that sequence was labeled as a different protein by the Genbank additions. The WGS contig labeled the matching sequence lactoylglutathione lyase while the top-down results credited this sequence to 30S S6 ribosomal protein. The sequence for the top-down database match was taken from *Yersinia pestis* and *Yersinia enterocolitica*. According to the Genbank additions, the other 3 proteins started at a different place in the translation process.  For example, the protein that both the WGS contigs and the top-down results agree is the sequence 50S ribosomal protein L32 has a 1583 Dalton difference due to this discrepancy in translational start site.  Therefore, the submitted WGS contigs are not enough information for the method proposed, and mass-spectrometry based protein identifications can correct the bacterial genome annotation.

### *Summary*

This chapter identifies a strategy to identify proteins from an organism without a sequenced genome.  By using high mass accuracy and high resolution, the fragmentation produced by CID in a linear ion trap can provide enough fragmentation to confidently match the measured protein's amino acid sequence to a database of proteins from closely related organisms with known sequences.  This strategy was demonstrated to identify proteins between 5000 Daltons to 15000 Daltons from two bacteria, *Bacillus anthracis* Sterne and *Yersinia rohdei*.  The proteins that were identified would be the same proteins that would have a high ion intensity in a MALDI mass spectrometry based, rapid bacterial characterization system.  The high abundance ribosomal proteins that were identified by the top-down strategy did provide some degree of differentiation and allowed for the determination of the closest neighboring organism and the position of the target organism in a phylogenetic tree.  This tree was similar to the traditional 16S-rRNA method of determining phylogeny, and the top-down mass spectrometric information proved favorable to the later whole genome shotgun contigs that were introduced into the Genbank database from the Naval Medical Research Center.

# Chapter 4: Phyloproteomic characterization of unsequenced organisms by top-down identification of proteins using capillary LC-MS/MS on a LTQ-Orbitrap XL

Adapted from Wynne C., Edwards, N.J., and Fenselau, C. Phyloproteomic classification of unsequenced organisms by top-down identification of bacterial proteins using capLC-MS/MS on an Orbitrap. *Proteomics*, in press.

## *Introduction*

As with the last chapter, this chapter uses a top-down proteomic approach to tackle the problem of identification of microorganism biomarker proteins that would be found in a rapid MALDI-TOF detection system. In the last chapter, identification of these biomarker proteins was demonstrated using 100% sequence homology and less than a 1 Dalton difference in precursor mass. This chapter uses a similar top-down proteomic approach, but the mass tolerance of the precursor mass was extended to 250 Daltons on another simulant of *Yersinia pestis, Erwinia herbicola,* and a third microorganism without a fully sequenced genome, *Enterobacter cloacae*. By widening the precursor tolerance and allowing for changes through either post-translational modifications or through changes in the amino acid structure, this extension of the previous method should allow for more protein identifications.

The top-down proteomic approach has been demonstrated multiple times to be a sensitive and robust technique for the identification of biomarker proteins [25, 41], but the target sequence's genome must be available. The previous chapter demonstrated that this is no longer a prerequisite as long as 100% homologous proteins are included in any database used to search the spectra collected from a top-down experiment. The whole sequence from these homologous proteins, and the organisms they came from, were then used to create phylogenetic trees to classify the microorganism without a sequenced genome in context with other sequenced organisms. This chapter takes this approach one step further by using only partial N- and C-terminal amino acid sequences that can be confidently identified by three or more fragment ions to be used in the construction of these phylogenetic trees. Therefore, while the target proteins have to be somewhat homologous to the already annotated proteins, the target proteins can have some differences. The identification and characterization of proteins that are not 100% homologous could lead to identifications that could possibly be unique to the target organism and set it apart in other detection systems used to screen for possible pathogen agents in defense and homeland security applications.

The high mass accuracy of the precursor ions and the amount of identified b- and y-ions in a particular target protein help localize any mass difference and narrow the list of available modifications that could cause the mass difference between the target protein and the protein that it was matched to through database searching. By knowing the precursor mass and mass difference between target and matched proteins up to tenths of a Dalton, the number of modifications that could cause said difference greatly narrows. By determining gaps in the b- and y-ions, the location of that mass difference can also be narrowed to that gap. By limiting the possible amino acids changed, as well as knowing the exact mass of the possible difference, the number of possible modifications can be small enough that a fast trial and error method of proposing changes could be undertaken.

Like *Yersinia rohdei*, *Erwinia herbicola* is a biosafety level 1 (BSL 1) bacterium commonly used as a non-lethal simulant for the lethal *Yersinia pestis* [58]. The bacterium was isolated from plant leaves, and is also known as *Pantoea agglomerans* or *Enterobacter agglomerans*. There is little annotated information about the proteome or genome, which gives rise to a lack of a commonly accepted name for the organism. In agricultural studies, *Erwinia herbicola* was used to control other *Erwinia* species[59]. *Enterobacter cloacae* is another BSL-1 bacterium. It is also used in agricultural studies and has been known to cause infection in hospitals after a patient has had open surgery[60].

## <u>*Experimental*</u>

### Preparation of Bacterial Lysates

*Erwinia herbicola* bacteria were grown at Johns Hopkins Applied Physics Lab under standard conditions and transferred in 50 ml tubes. Once at the University of Maryland, the tubes were split into 15 ml tubes and frozen at -20$^{o}$C. A 15 ml tube was centrifuged at 8000 RPM for 5 minutes and the cell pellet was transferred to 1.5 ml microcentrifuge tubes. The cell pellet was washed with Milli-Q water and centrifuged at 10,000 RPM for 5 minutes and the supernatant was discarded. This wash was repeated twice more. The pellet was suspended in 100 microliters of a 10% formic acid solution. The suspension was then centrifuged a final time at 10,000 RPM for 5 minutes. This time, the supernatant was transferred to a YM 3,000 Dalton molecular weight cutoff filter (Millipore, Billerica, MA) and centrifuged at 14,000

RPM for 60 minutes, as directed by the cutoff filter manual. The liquid in the top portion of the filter (theoretically the portion over 3000 Daltons) was pipetted out of the filter and into a vial to be injected for LC-MS/MS analysis.

Enterobacter cloacae was obtained from the American Type Culture Collection (Manassas, VA) and grown using typical bacteria growth practices [10] on Nutrient broth plates (ThermoFisher, Fairlawn, NJ). Cells were grown in 3 ml cultures overnight to saturation, then treated the same as the *Erwinia herbicola* preparation prior to LC-MS/MS analysis, with one exception. Instead of the YM-3000 molecular weight cutoff filter, an Amicon Ultra 3,000 Dalton molecular weight cutoff filter was used and centrifuged at 14,000 RPM for 30 minutes in accordance with the instructions from its manual.

### LC-MS/MS Analysis

Ten microliters of bacteria lysate solution was injected onto a 0.1millimeter by 15 centimeter Magic C-8 column with 5 micron particles (Michrom, Auburn, CA) using a 2-D Prominence Nanomate pump (Shimadzu, Columbia, MD) inline to a LTQ-Orbitrap XL mass spectrometer (ThermoFisher, San Jose, CA). Solvent A was a 95% water, 4.9% acetonitrile, 0.1% formic acid mixture. Solvent B was a 4.9% water, 95% acetonitrile, 0.1% formic acid mixture. In both mixtures, the water, acetonitrile, and formic acid were all HPLC grade solvents (ThermoFisher). The gradient program used a 15 minute sample load with a 1 milliliter flow rate onto a trapping cartridge at 10% B, followed by a 50 minute gradient from 10%B to 70%B with a 500 nanoliter flow rate using the nanopumps with an internal split. Next, a column cleaning step at 80%B for 10 minutes and a reequilibration step back to

10%B.  The LTQ-Orbitrap XL was set to record the MS/MS spectra of the 5 most abundant signals for each precursor scan.  Each precursor scan was acquired at 30,000 resolving power at 400 m/z, while the MS/MS scans were acquired at 15,000 resolution at 400 m/z.  This decrease in resolution from the precursor scans to the MS/MS scans was to ensure the size of the data file remained less than 2 Gigabytes to avoid problems with the Bioworks software (ThermoFisher, San Jose, CA). Fragmentation occurred using collisionally induced dissociation in the linear ion trap with high purity helium gas.  The CID settings were set at 35% level with activation for 60 milliseconds.  Dynamic exclusion was set at 10 seconds, meaning that once a particular m/z was sampled, it could not be sampled again for the next 10 seconds. This reduced the oversampling of abundant precursor ions.  MS/MS analysis was restricted to precursor ions with known charge states of +3 or more.  Prior to analysis, the LTQ-Orbitrap XL was mass calibrated using the Thermo mix of caffeine, the quad-peptide MRFA, and an Ultramark polymer.

### Protein Sequence Database

A custom FASTA format sequence database of *Enterobacteriaceae* protein sequences was constructed from all protein sequences from Swiss-Prot, TrEMBL, RefSeq, Genbank, and the Venter Institute's CMR annotated as from the *Enterobacteriaceae* family, which contains *Erwinia herbicola*.  In addition, Glimmer3[52] was used to predict primary and alternative translation start-site protein sequences on RefSeq *Enterobacteriaceae* genomes.  The set of sequences was further filtered for molecular weights between 1000 Daltons and 20000 Daltons.  In total, over 1 million sequences were merged  to 253,626 distinct protein sequences representing 256 Enterobacteriaceae species. The FASTA sequence database was built using infrastructure developed for the Rapid Microorganism Identification Database (RMIDb)[50], and can be downloaded from the ProteomeCommons.org Tranche network.

### Protein Identification

ProSight PC 2.0[31] was used to deconvolute precursor and fragment spectra, and to search the MS/MS spectra, in absolute mass mode, against the custom FASTA database described above. The THRASH[27] algorithm was used to deconvolute the spectra as part of ProSight PC 2.0. After deconvolution, the MS/MS were filtered so that only those spectra with 3 or more fragment ions were searched against the custom databse. A loosely constrained precursor mass tolerance was used, matching protein sequences within 250 Daltons of experimental precursor masses (both smaller and larger) with corresponding MS/MS spectra. The loose precursor constraint was applied to allow for a small number of post-translational modifications and/or amino acid substitutions. Experimental fragment mass measurements were matched against theoretical monoisotopic fragment masses of the protein sequences using a mass tolerance of 15 parts per million. The "ΔM" feature of ProSight PC was used to check for unexpected mass shifts at the protein N- or C-terminii. Post-translational modifications and amino acid substitutions were manually investigated using the Sequence Gazer tool in ProSight PC.

### Phylogenetic Analysis

ProSight PC protein identifications with expect values less than 1e-4 were analyzed, and N- and C- terminal amino-acid sequence supported by at least three high accuracy b- or y-ions, respectively, were established. For identifications in which the confident established N- and C-terminus meet or overlap, the entire protein sequence is established. For identifications in which less than 3 b-ions or y-ions were observed, only the C-terminus or N-terminus was considered established. Confidently established sequences of at least ten amino-acids from each protein identification were searched against the custom *Enterobacteriaceae* FASTA sequence database using BlastP[46]. Extracted sequences that did not align to any *Enterobacteriaceace* sequence with E-values less than 1e-4 were discarded, as were those with exact or near exact alignments with another extracted sequence. Species with significant alignments (E-value less than 1e-4) to all remaining query sequences were noted, and the corresponding amino-acid sequence of these species' best alignment were retained. A random ordering of the identified sequences was fixed, and the corresponding amino-acid sequences from each of the retained species were concatenated in the same order, to form a FASTA file suitable for multiple sequence alignment and phylogeny analysis. In all, the confidently established amino-acid sequences from the N- and C-terminus of identified proteins from *Erwinia herbicola* could be matched in 27 other species in which the bacteria was represented by a meta-sequence of length 795 (median) amino acids. Phylogenetic analysis was carried out using the "one-click mode" at www.phylogeny.fr[54].

### Genome Annotation Analysis

The source of *Enterobacteriaceae* protein sequences matched to *Erwinia herbicola* and *Enterobacter cloacae* spectra using ProSight PC, and homologous *Enterobacteriaceae* sequences matched to confidently established *Erwinia herbicola* and *Enterobacter cloacae* N- and C- terminal amino-acid sequences using BlastP were tabulated to find species and proteins observed only in the Glimmer3 protein sequence set.

## Results and Discussion

### *Erwinia herbicola* proteins

Table 4.1 shows the proteins identified from *Erwinia herbicola*, which have molecular weights in the range 4 – 12 kDa. 72 intact molecular masses yielded 14 identified proteins. Ten of the 14 identified proteins are ribosomal, whose high abundance has been previously observed and discussed in the last chapter. Some of the identified protein sequences, such as 50S ribosomal protein L29, are genus specific and are only matched in the closely related *Erwinia tasmaniensis*. Other ribosomal proteins, such as 30S ribosomal protein S18, are matched in a number of different organisms, including all of the available *Yersinia* species. Six proteins matched with a protein molecular weight delta of 15 ppm or less, indicating that *Erwinia herbicola* has the same protein sequence as the related organisms supplying the identifying protein sequence. The base peak chromatogram is shown in Figure 4.1 and an example of a tandem mass spectrum are shown in Figures 4.2, with the annotated with its protein sequence and b- and y-ion fragments matched by ProSightPC 2.0.

| m/z | # Matching Fragments | Observed Mass | Theoretical Mass | Mass Diff | Best Hit Protein Family | Best Hit Organism | Best Hit E-value |
|---|---|---|---|---|---|---|---|
| 566.92 | 13 | 4530.128 | 4632.230 | -102.102 | Ribosomal protein L32 | *Erwinia tasmaniensis* | 8.44E-06 |
| 564.68 | 15 | 6199.083 | 6215.080 | -15.997 | Ribosomal protein L32 | *Erwinia tasmaniensis* | 1.72E-07 |
| 569.88 | 12 | 6240.392 | 6240.400 | -0.008 | Ribosomal protein L33 | *Escherichia coli, Salmonella enterica, Salmonella typhimurium, Erwinia tasmaniensis, Shigella flexneri, Citrobacter koseri, Shigella sonnei, Shigella dysenteriae, Shigella boydii, Salmonella choleraesuis* | 7.85E-05 |
| 712.74 | 9 | 6399.555 | 6442.570 | -42.015 | Ribosomal protein L30 | *Sodalis glossinidius* | 5.43E-05 |
| 1138.13 | 15 | 6820.008 | 6867.960 | -47.952 | Carbon storage regulator | *Serratia plymuthica* | 4.84E-11 |
| 1211.84 | 13 | 7262.980 | 7247.420 | 15.960 | Ribosomal protein L29 | *Erwinia tasmaniensis* | 2.71E-08 |
| 1014.30 | 26 | 8105.355 | 8104.380 | 0.025 | Translation initiation factor 1A | *Sodalis glossinidius,Yersinia pseudotuberculosis, Yersinia pestis, Yersinia intermedia, Yersinia enterocolitica, Enterobacter sp. 638, Photorhabdus luminescens, Serratia proteamaculans, Yersinia bercovieri, Erwinia tasmaniensis, Pectobacterium atrosepticum, Yersinia frederiksenii* | 2.34E-28 |
| 761.98 | 14 | 8368.777 | 8368.770 | 0.007 | Ribosomal protein S21 | *Sodalis glossinidius,Yersinia pseudotuberculosis, Yersinia pestis, Yersinia intermedia, Yersinia enterocolitica, Enterobacter sp. 638, Photorhabdus luminescens, Serratia proteamaculans, Yersinia bercovieri, Erwinia tasmaniensis, Pectobacterium atrosepticum, Yersinia frederiksenii, Escherichia coli, Salmonella enterica, Salmonella typhimurium, Erwinia tasmaniensis, Shigella flexneri, Citrobacter koseri, Shigella sonnei, Shigella dysenteriae, Shigella boydii, Salmonella choleraesuis, Klebsiella pneumoniae, Providencia stuartii, Enterobacter sakazakii* | 1.07E-08 |
| 742.91 | 11 | 8900.285 | 8900.350 | -0.065 | Ribosomal protein S18 | *Yersinia pseudotuberculosis, Yersinia pestis, Yersinia intermedia, Yersinia enterocolitica, Enterobacter sp. 638, Photorhabdus luminescens, Serratia proteamaculans, Yersinia bercovieri, Erwinia tasmaniensis, Pectobacterium atrosepticum, Yersinia frederiksenii* | 6.28E-06 |
| 1023.53 | 12 | 9200.258 | 9076.200 | 124.058 | Cell division protein zapB | *Erwinia tasmaniensis* | 2.55E-05 |
| 732.71 | 31 | 9507.190 | 9520.970 | -14.128 | DNA-binding protein HU-alpha | *Salmonella enterica, Salmonella typhimurium, Citrobacter koseri, Salmonella choleraesuis* | 7.50E-26 |
| 683.67 | 10 | 9558.321 | 9559.220 | -0.899 | Ribosomal protein S17 | *Serratia proteamaculans* | 1.67E-07 |
| 686.39 | 21 | 10285.003 | 10285.100 | 0.007 | Ribosomal protein S19 | *Salmonella enterica, Klebsiella pneumoniae, Salmonella typhimurium, Enterobacter sakazakii, Enterobacter sp. 638, Salmonella choleraesuis* | 1.96E-16 |
| 1018.01 | 20 | 11185.280 | 11185.294 | -0.014 | Ribosomal protein L24 | *Escherichia coli, Salmonella enterica, Salmonella typhimurium, Erwinia tasmaniensis, Shigella flexneri, Citrobacter koseri, Shigella sonnei, Shigella dysenteriae, Shigella boydii, Salmonella choleraesuis, Klebsiella pneumoniae, Enterobacter sakazakii* | 7.79E-16 |

Table 4.1-Database matches to *Erwinia herbicola* proteins

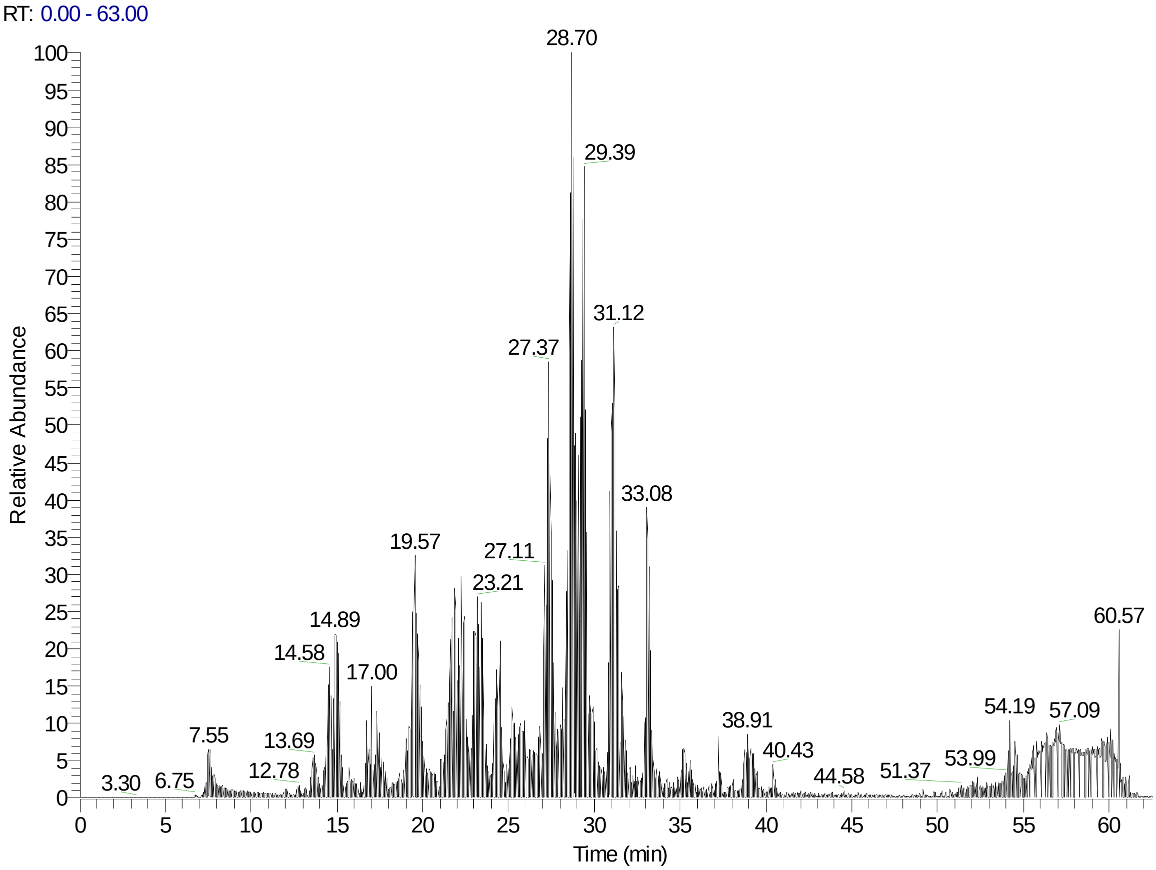


Figure 4.1-Base peak chromatogram of *Erwinia herbicola* lysate separated on the Prominence Nanomate HPLC system (Shimadzu)

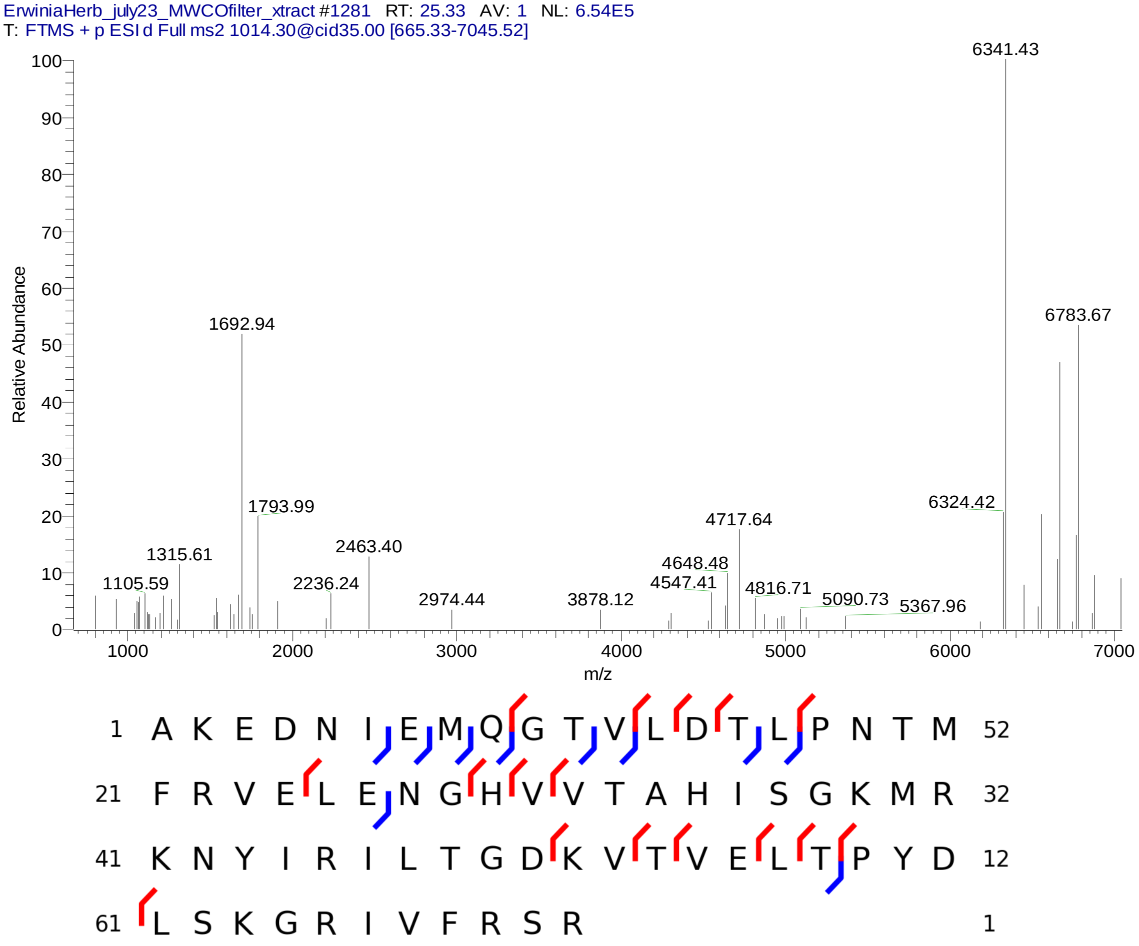


Figure 4.2-Top: Deconvoluted tandem mass spectrum of precursor ion 1014.30 in charge state +8 identified as *Erwinia herbicola* protein Translation initiation factor 1A. Bottom: Sequence of matched protein with 26 b- and y-ion fragment matches.

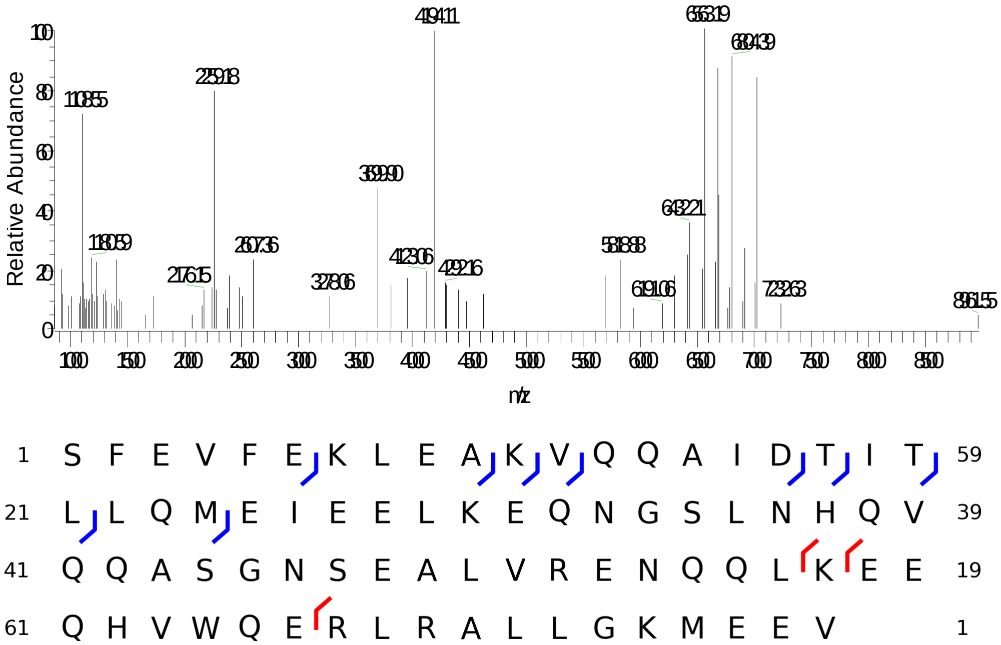


Figure 4.3-Top: MS/MS of 1023.53 with charge +8 (molecular weight 9200.58).

Bottom: Sequence of Cell division protein zapB from Erwinia tasmaniensis with matched b- and y-ions.

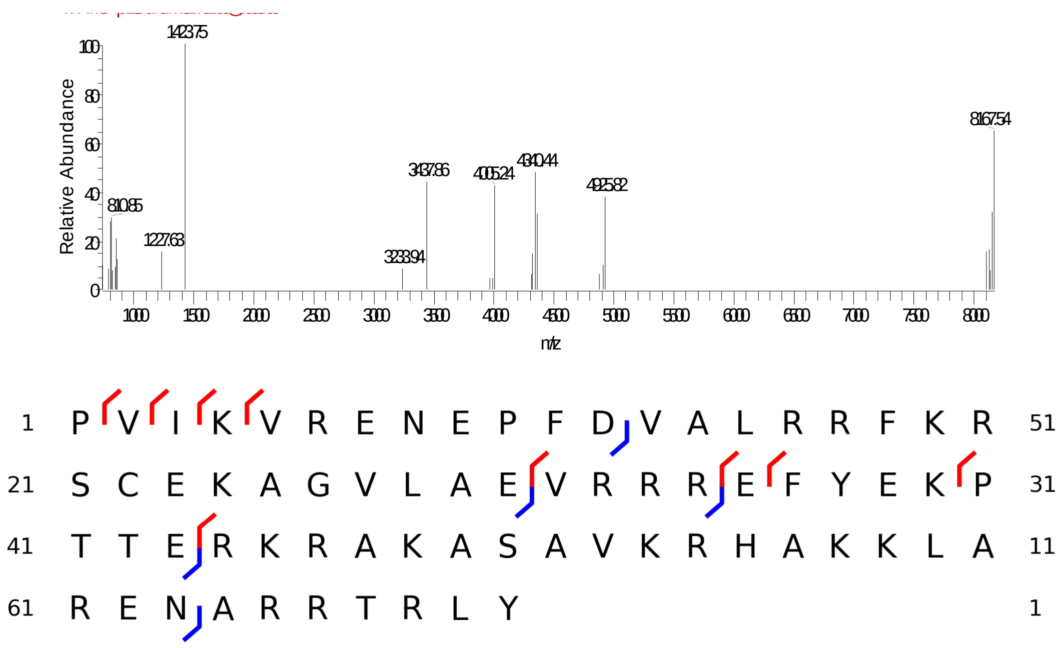


Figure 4.4-Top: MS/MS of 761.98 with +11 charge (molecular weight of 8368.78). Bottom: Amino acid sequence of ribosomal protein S21 with matched b- and y- ion matches.

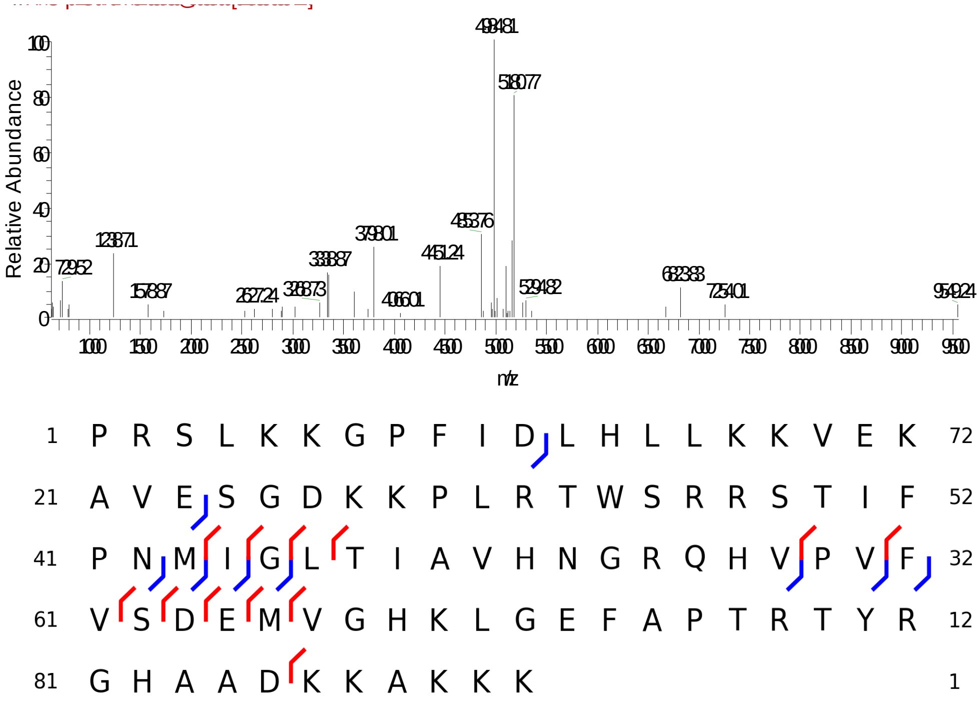


Figure 4.5-Top: MS/MS of 686.39 with +15 charge (molecular weight 10285.10).

Bottom: Amino acid sequence of ribosomal protein S19 with b- and y- ion matches.

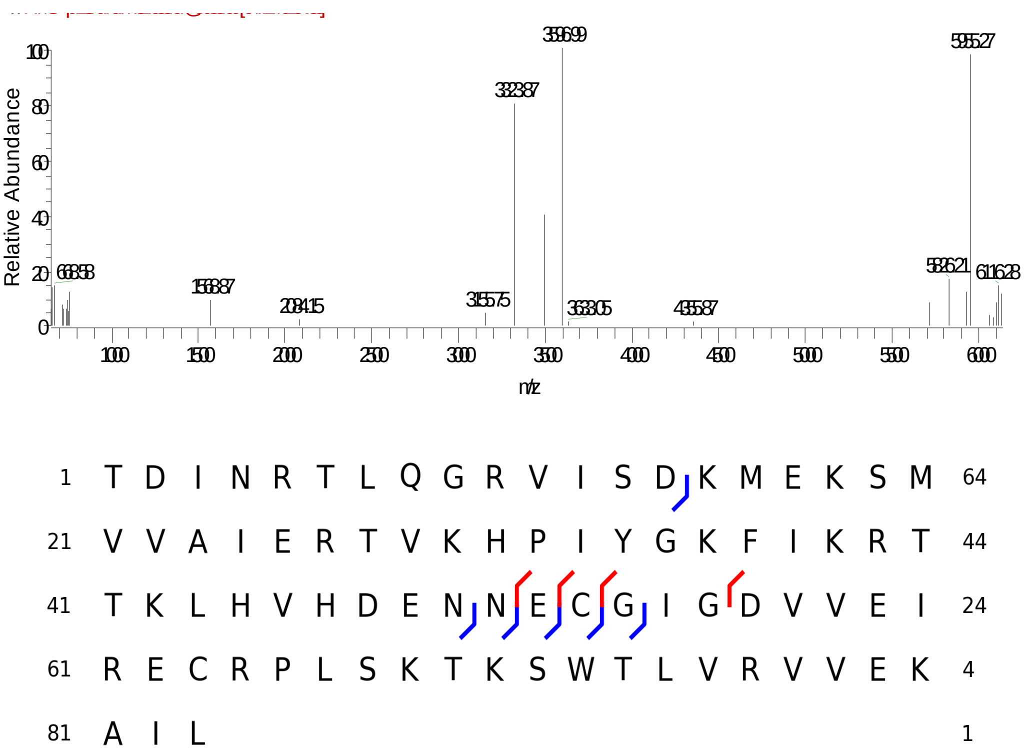


Figure 4.6-Top: MS/MS of 683.67 with +14 charge (molecular weight (9558.32).

Bottom: Sequence of ribosomal protein S17 with matched b- and y- ions.

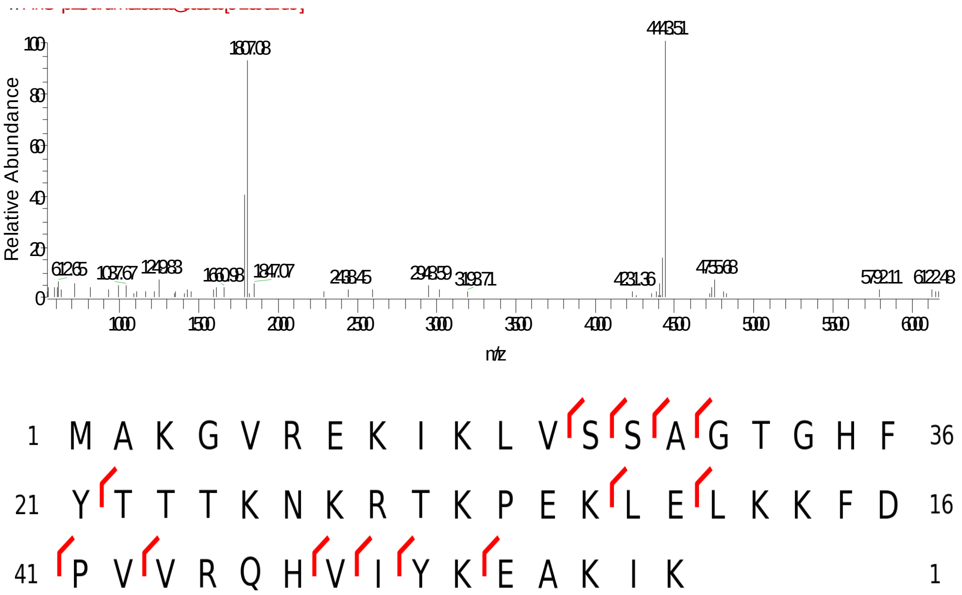


Figure 4.7-Top:  MS/MS of 569.88 with +11 charge (molecular weight 6240.40).

Bottom:  Sequence of ribosomal protein L33 with matched y-ions.

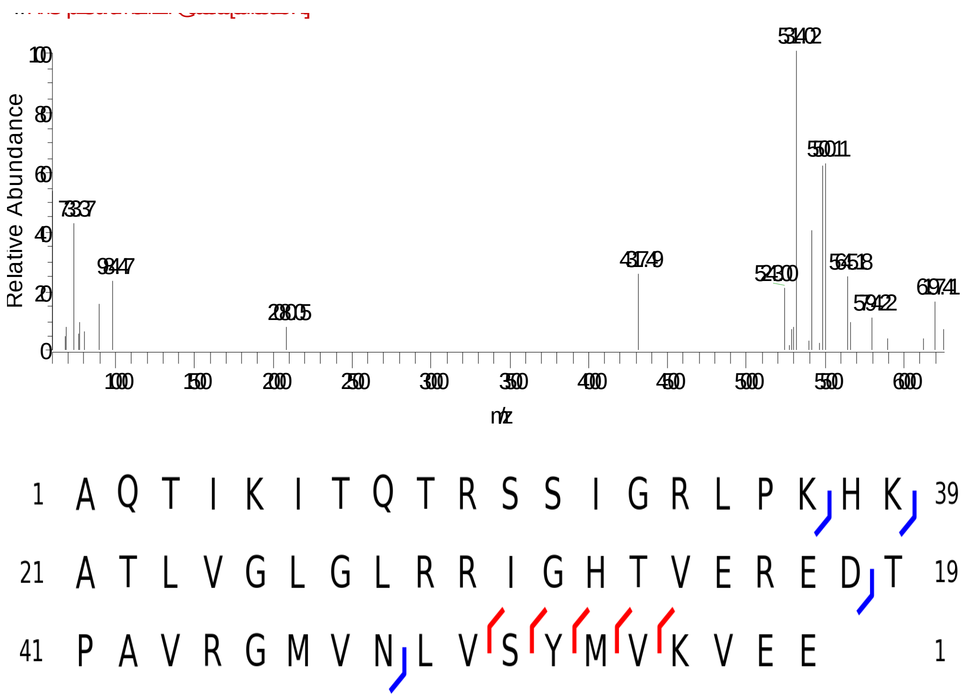


Figure 4.8-Top: MS/MS of m/z 712.74 with +9 charge (molecular weight 6399.56).

Bottom: Sequence of ribosomal protein L30 with matched b- and y- ions.

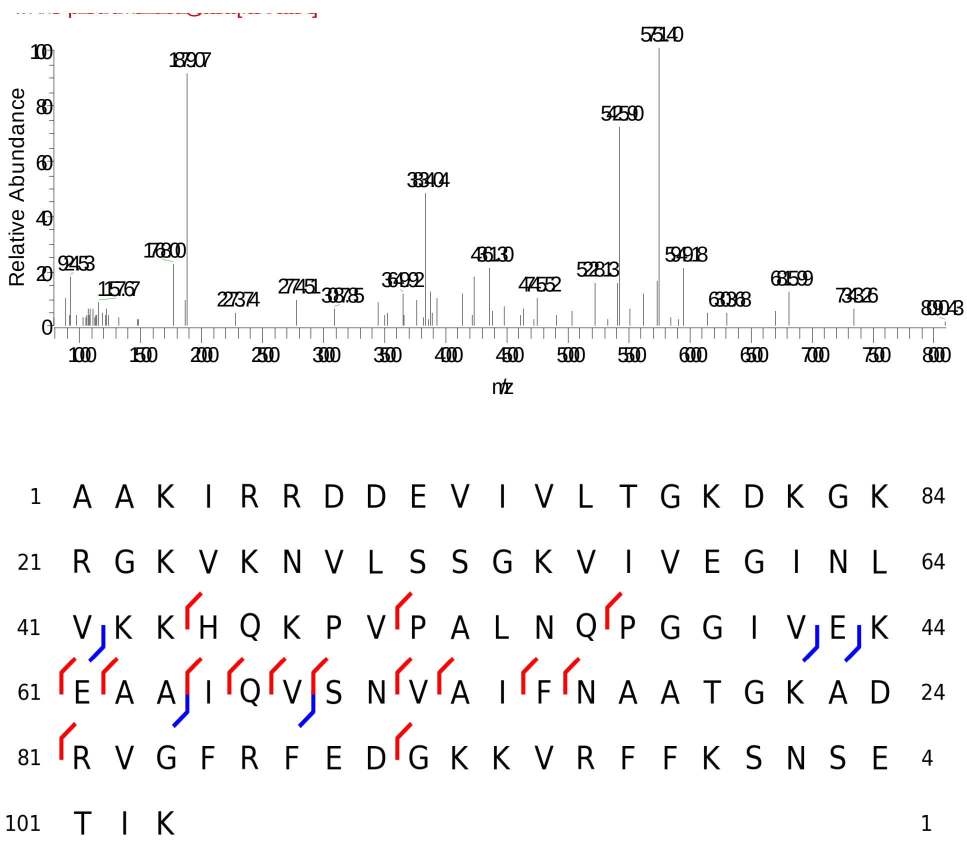


Figure 4.10-Top: MS/MS of m/z 1018.01 with +11 charge (molecular weight 11185.28). Bottom: Sequence of ribosomal protein L24 with matched b- and y-ions.

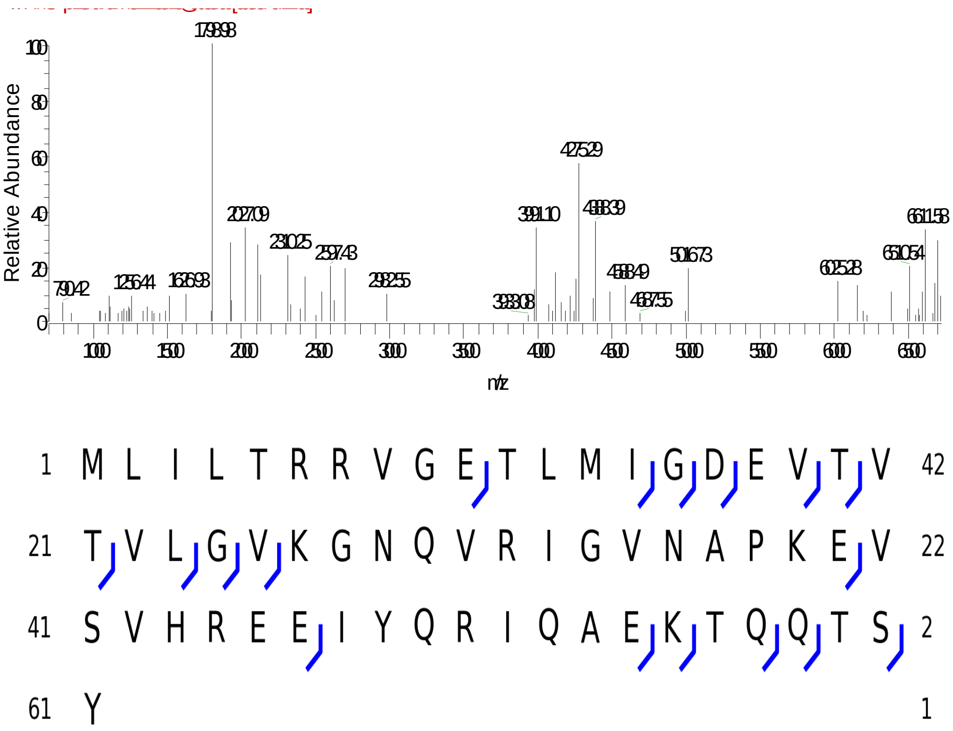


Figure 4.10-Top: MS/MS of 1138.13 with +6 charge (molecular weight 6820.01)

Bottom: Sequence of carbon storage regulator protein with matched b-ions.

While *E. herbicola* protein identifications with E-values less than 1.0e-4 are shown in Table 4.1, many of these are much more statistically significant, with five identifications having E-values less than 1.0e-10. These highly significant protein identifications are made primarily due to the number, and position, of the matched high-accuracy fragment ion measurements.

The additional eight proteins which match due to the loose precursor mass search tolerance more than double the number of identified *E. herbicola* proteins. Four of these have mass deltas of 16 Da or less, with the remaining four having much larger mass deltas. For three of the small mass-deltas, the likely position of the required mass-shift can be constrained by the positions of matching b- and y-ions, in each case suggesting a putative amino-acid substitution in the *E. herbicola* protein.

An *E. herbicola* spectrum with precursor 564.68 in charge-state +11 matched 50S ribosomal protein L32 from *Erwinia tasmaniensis* with E-value 1.72e-7 and mass delta of -15.997 Da. While this identification matches two b-ions and 13 y-ions, the N-terminus fragment ions stop at $b_{19}$, while C-terminus fragment ions end at $y_{34}$, a gap of just two amino-acids, suggesting a -15.997 mass shift on serine at position 20 or leucine at position 21. Changing the twentieth amino acid from serine to alanine and checking the result using the Sequence Gazer tool in ProSightPC, the mass delta becomes -0.01 Da, and six additional b-ions and one additional y-ion is matched, improving the E-value to 5.85e-29. Figure 4.3 shows this tandem mass spectrum, plus the protein sequence and the matched b- and y- fragment ions.

The top-down spectrum from the charge state +6 precursor ion at m/z 1211.84 matched to *Erwinia tasmaniensis* 50S ribosomal protein L29 with E-value 2.71e-8 and mass delta of +15.960. The 12 matching b-ions end at $b_{61}$, establishing the amino-acid sequence of all but the last two amino-acids. By changing the C-terminal residue from alanine to serine, the mass difference is becomes -0.024, two new y-ions are matched, and the resulting E-value improves to 8.32e-20. Oxidation of methionine or histidine residues cannot readily be placed at the C-terminus, and placement elsewhere in the sequence results in the loss of a significant number of b-ion fragment matches. (Figure 4.4)

The spectrum of precursor m/z 732.71, in charge state +13, matches DNA binding protein HU-alpha with E-value 7.5e-26 and mass delta -14.128, matching 12 b-ions and 19 y-ions. The b-ion fragment matches run to residue 37 while the y-ion fragments start at residue 42, leaving a 5 residue gap. An aspartate for glutamate substitution results in an additional 7 b-ion and 3 y-ion matches, improving the E-value to 1.91e-58, and changing the mass-delta to 0.110 Da. (Figure 4.5)

The mass-shifts responsible for the remaining five protein identifications remain unexplained at this time, but we stress that the intact protein mass and the protein *family* identity is not in doubt. Due to the number (at least ten in each case) and position of the accurate mass fragment matches, these identifications are highly statistically significant, even if the entire amino acid sequence of the protein cannot be asserted.

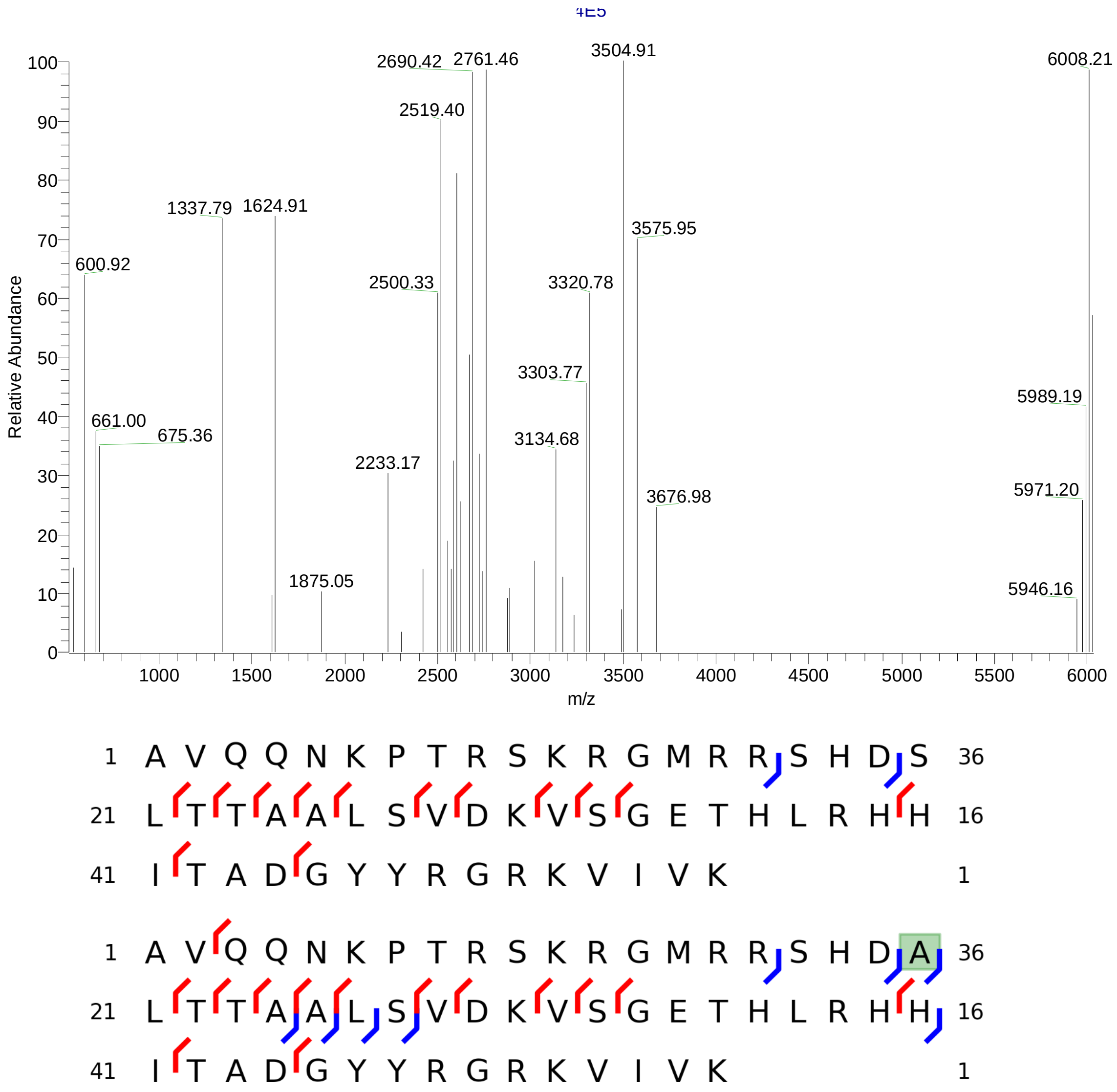


Figure 4.11-Top: Deconvoluted tandem mass spectra of precursor 564.68 in charge state +11 identified as *Erwinia herbicola* protein 50s Ribosomal protein L32 with E-value 1.72e-7. Middle: Sequence of matched protein with 15 b- and y-ion fragments. Bottom: Highlighted substitution of serine to alanine at the 20th position, and rescored search with E-value now 5.85e-29 with 22 b- and y-ion fragment matches.

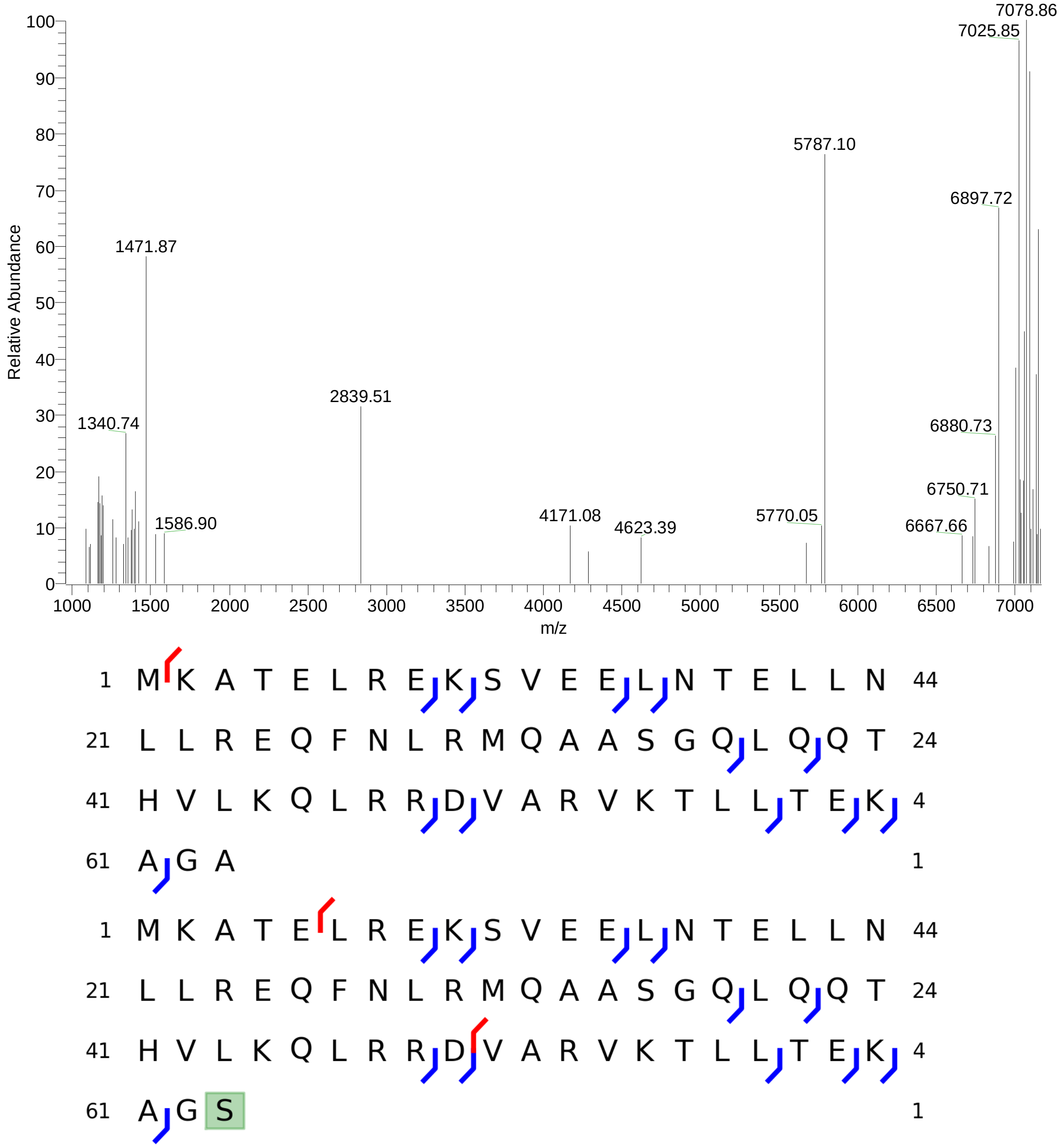


Figure 4.12-Top: Tandem mass spectrum of precursor 1211.84 in charge state +6 identified as 50s Ribosomal protein L29 with E-value 2.71e-8. Middle: Original matching sequence with 13 b- and y- ion fragments. Bottom: Highlighted substitution of alanine to serine at C-terminus. Rescored E-value now 8.32e-20 with 14 b- and y- ion fragments.

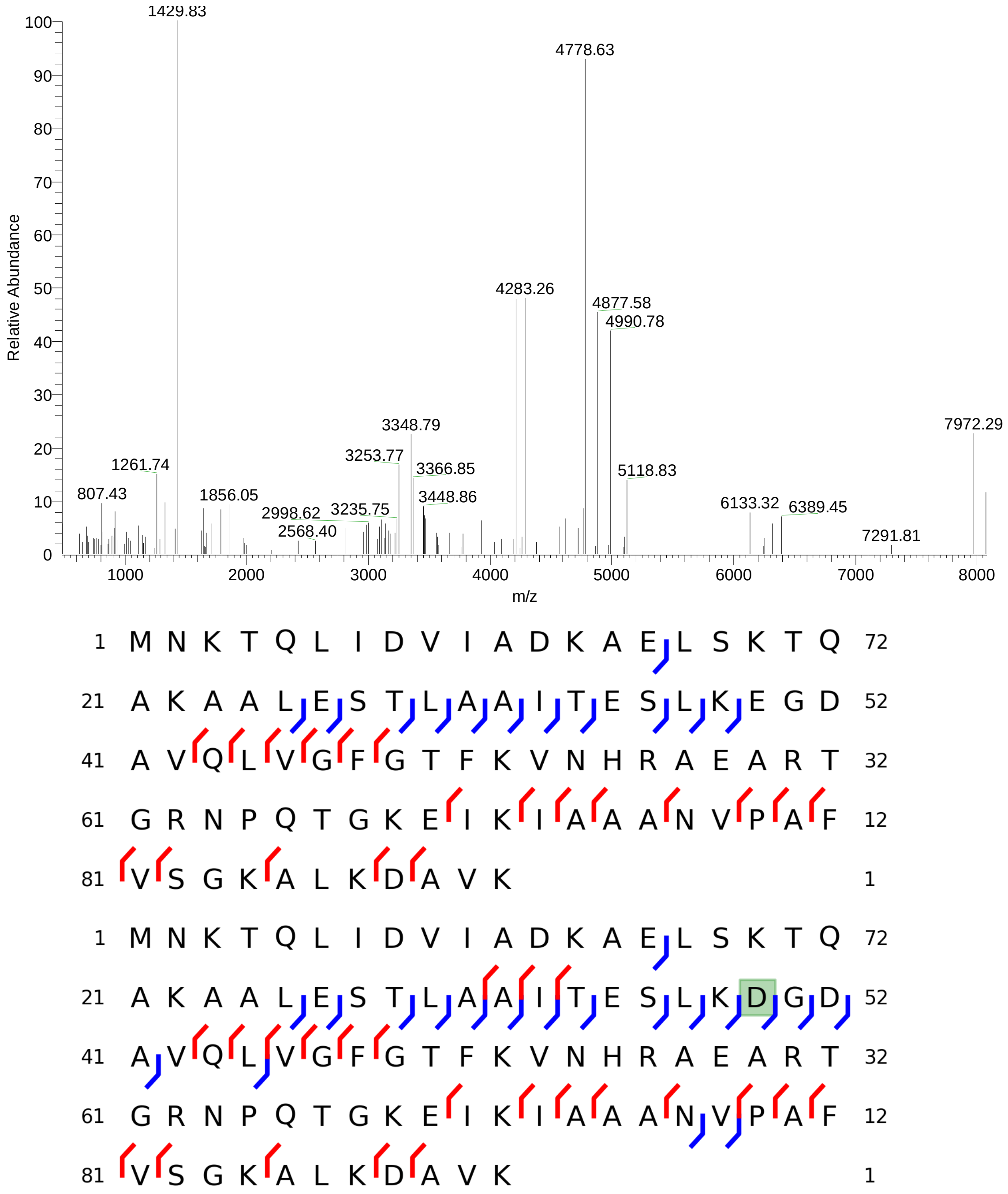


Figure 4.13-Top: tandem mass spectrum of precursor ion 732.71 in charge state +13 identified as DNA binding protein HU-alpha with E-value 7.5e-26. Middle: matched protein sequence with 31 b- and y-ion fragments. Bottom: Substitution of glutamate to aspartate at 38th position. Rescored E-value now 1.91e-58 with 41 b- and y- ions.

*Enterobacter cloacae* proteins

Table 2 presents a summary of proteins identified with ProSightPC E-values less than 1.0e-4 from *Enterobacter cloacae*.  Fifteen proteins were identified from 129 intact molecular masses. Five of the 15 identified proteins are ribosomal, with three cold shock protein spectra also observed. This may be due to the fact that the *Enterobacter* cells were lyophilized and stored for 4 years while the *Erwinia* cells were frozen in water and stored for 6 months. The best matches to three of the *E. cloacae* precursors, 50S ribosomal protein L30, 50S ribosomal protein L28, and DNA-binding protein HU-alpha (m/z 1189.65) were found only in *Enterobacter sp. 638*. Other organisms with matching protein sequences include *Klebsiella pneumoniae*, *Enterobacter sakazakii*, and *Escherichia coli*. The relationship between *Enterobacter cloacae* and *Escherichia coli* has been shown in the similarity of their 16S-rRNA previously by Clementino[61], and this is confirmed with the top-down proteomics analysis here.

Table 4.2 includes six identifications with E-values smaller than 1.0e-10, reinforcing the observation that accurate mass fragment ion matches can be sufficient for highly significant protein identifications. Eleven precursor masses are within 15ppm of the matched proteins' molecular weight. These matches indicate that the experimental proteins from *Enterobacter cloacae* have the same amino acid sequence as the matched proteins from near neighbor organisms. An additional identification to *Escherichia coli* 50S ribosomal protein L24 has mass delta 18.5ppm. Finally, three proteins have a mass difference of 100 Daltons or more. Again, while we cannot claim to have established the full amino-acid sequence of these proteins, the protein *family* identity is not in doubt, and in each case, a significant proportion of the amino-acid sequence of the protein can established by the number and position of the fragment ion matches.

| m/z | # Matching Fragments | Theoretical Mass | Observed Mass | Mass Diff | Best Hit Protein Family | Best Hit Organism | Best Hit E-value |
|---|---|---|---|---|---|---|---|
| 798.45 | 8 | 4923.660 | 4781.589 | -142.071 | DNA-binding protein HU-alpha | *Serratia proteamaculans* | 9.35E-07 |
| 814.63 | 10 | 5054.860 | 4881.825 | -173.035 | DNA-binding protein HU-alpha | *Serratia proteamaculans* | 2.41E-10 |
| 1078.75 | 11 | 6458.660 | 6458.683 | 0.023 | Ribosomal protein L30 | *Enterobacter sp. 638* | 3.00E-07 |
| 1144.99 | 13 | 6855.910 | 6855.920 | 0.010 | Carbon storage regulator | *Escherichia coli, Shigella flexnari, Salmonella enterica, Shigella sonnei, Klebsiella pneumoniae, Citrobacter koseri, Salmonella typhermurium, Enterobacter sakazakii, Shigella boydii, Shigella dysenteriae* | 1.42E-12 |
| 907.25 | 15 | 7243.480 | 7243.480 | 0.000 | Ribosomal protein L29 | *Klebsiella pneumoniae* | 1.14E-14 |
| 1455.96 | 13 | 7271.130 | 7271.096 | -0.034 | Cold shock DNA binding protein | *Escherichia coli, Shigella flexnari, Salmonella enterica, Shigella sonnei, Klebsiella pneumoniae, Citrobacter koseri, Salmonella typhermurium, Enterobacter sakazakii, Shigella boydii, Shigella dysenteriae* | 1.87E-11 |
| 1221.64 | 13 | 7318.230 | 7318.213 | -0.017 | Cold shock DNA binding protein | *Enterobacter sakazakii, Klebsiella pneumoniae* | 2.10E-11 |
| 1062.35 | 11 | 7287.180 | 7441.750 | 154.570 | Cold shock DNA binding protein | *Yersinia pestis, Yersinia pseudotuberculosis, Serratia proteamaculans, Yersinia bercovieri, Yersinia enterocolitica, Yersinia mollaretii* | 1.99E-11 |
| 1159.62 | 10 | 8104.380 | 8104.360 | -0.020 | Translation initiation factor 1A | *Sodalis glossinidius,Yersinia pestis, Yersinia pseudotuberculosis, Serratia proteamaculans, Yersinia bercovieri, Yersinia enterocolitica, Yersinia mollaretii, Enterobacter sp. 638, Photorhabdus luminescens, Pectobacterium atrosepticum, Erwinia tasmaniensis* | 1.63E-08 |
| 809.36 | 9 | 8891.310 | 8891.355 | 0.045 | Ribosomal protein L28 | *Enterobacter sp. 638* | 9.33E-06 |
| 1306.71 | 17 | 9137.480 | 9137.522 | 0.042 | DNA-binding protein HU-beta | *Klebsiella pneumoniae* | 5.42E-13 |
| 1189.65 | 14 | 9504.970 | 9505.004 | 0.034 | DNA-binding protein HU-alpha | *Enterobacter sp. 638* | 3.08E-10 |
| 937.16 | 8 | 10299.100 | 10299.157 | 0.057 | Ribosomal protein S19 | *Escherichia coli, Shigella flexnari, Salmonella enterica, Shigella sonnei, Klebsiella pneumoniae, Citrobacter koseri, Salmonella typhermurium, Enterobacter sakazakii, Shigella boydii, Shigella dysenteriae* | 4.41E-05 |
| 1119.54 | 7 | 11185.000 | 11185.207 | 0.207 | Ribosomal protein L24 | *Escherichia coli, Shigella flexnari, Salmonella enterica, Shigella sonnei, Klebsiella pneumoniae, Citrobacter koseri, Salmonella typhermurium, Enterobacter sakazakii, Shigella boydii, Shigella dysenteriae* | 1.60E-06 |
| 1464.22 | 8 | 11705.500 | 11705.447 | -0.053 | Thioredoxin protein | *Enterobacter sakazakii, Klebsiella pneumoniae* | 4.66E-05 |

Table 4.2-ProSight PC 2.0 matches to *Enterobacter cloacae*

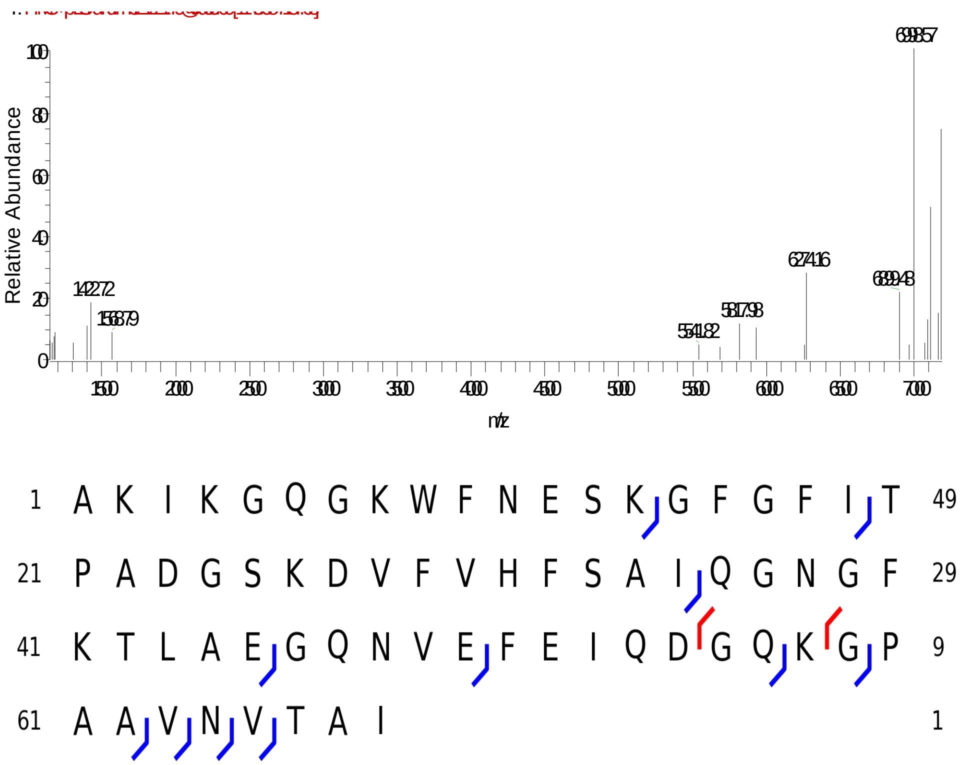


Figure 4.14-Top: MS/MS of m/z 1221.98 with +6 charge (molecular weight 7318.23). Bottom: Sequence of cold shock protein with matched b- and y- ions.

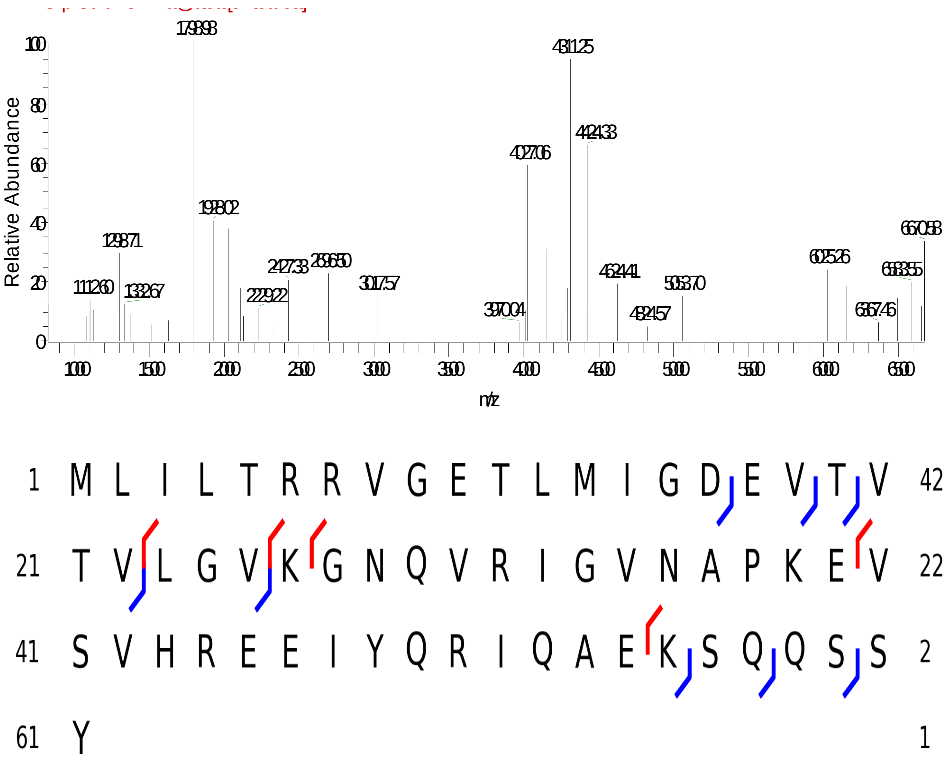


Figure 4.15-Top: MS/MS of m/z 1144.99 with +6 charge (molecular weight 6855.91). Bottom: Sequence of carbon storage regulator protein with matched b- and y- ions.

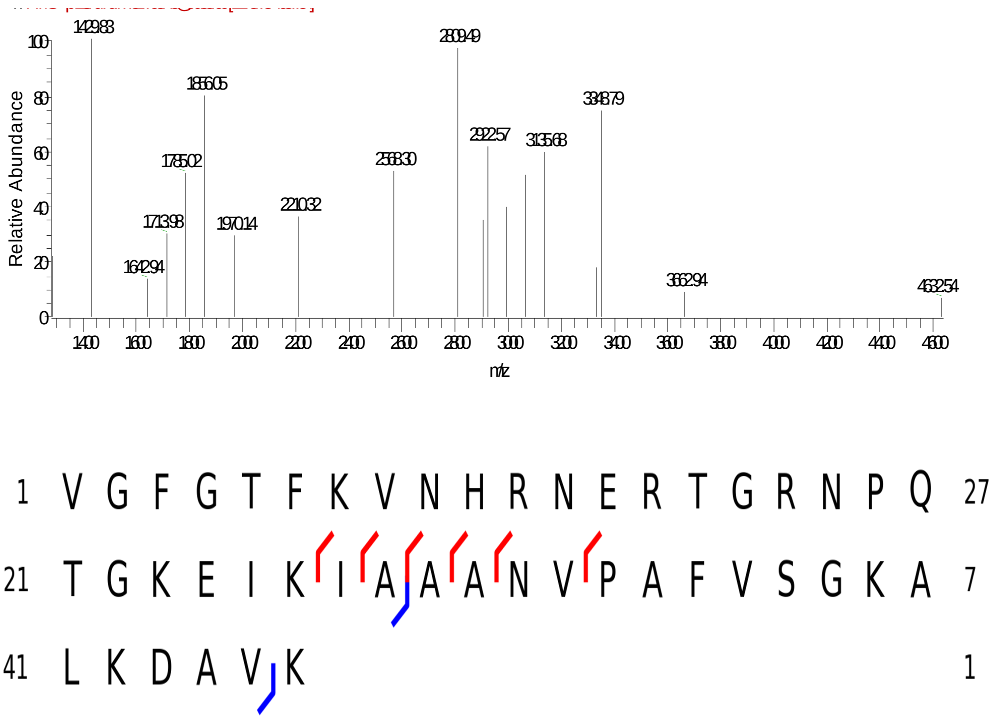


Figure 4.16-Top: MS/MS of m/z 798.45 with +6 charge (molecular weight 4781.59).

Bottom: Sequence of DNA-binding protein HU-alpha with matched b- and y- ions.

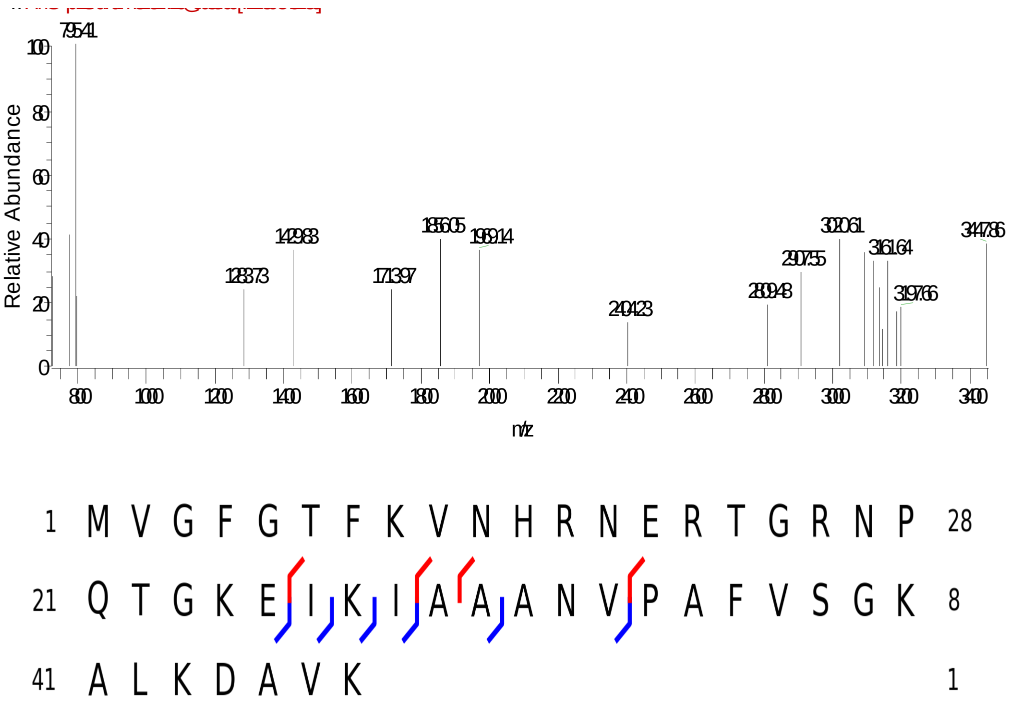


Figure 4.17-Top: MS/MS of m/z 814.63 with +6 charge (molecular weight 4881.82).

Bottom: Sequence of DNA-binding protein HU-alpha with matched b- and y- ions.

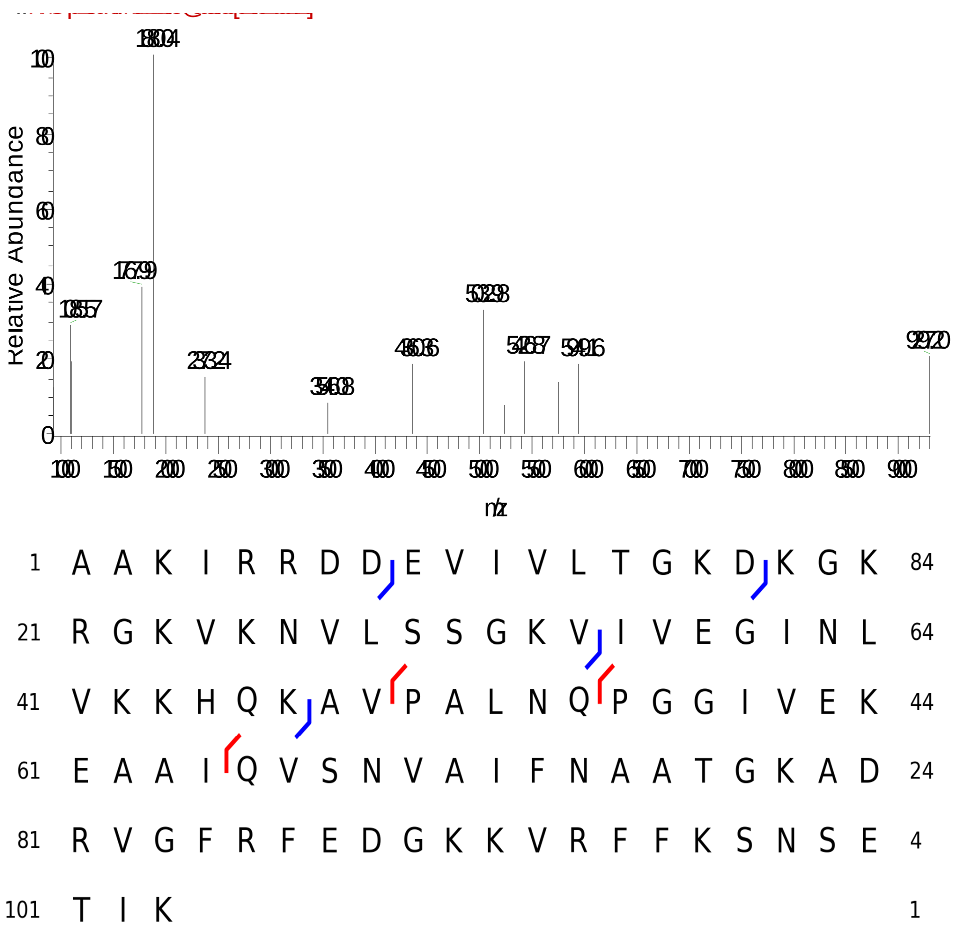


Figure 4.18-Top: MS/MS of m/z 1119.54 with +10 charge (molecular weight of 11185.20). Bottom: Sequence of ribosomal protein L24 with matched b- and y- ions.

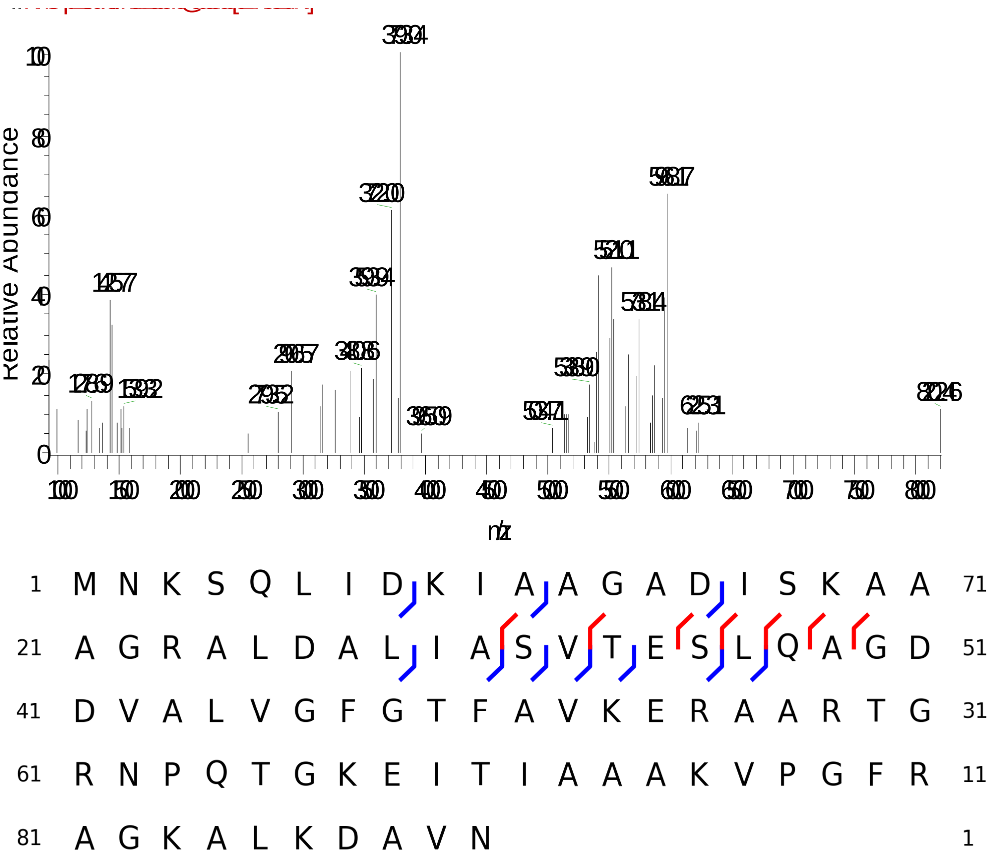


Figure 4.19-Top: MS/MS of m/z 1306.71 with +8 charge (molecular weight 9137.52). Bottom: Sequence of DNA-binding protein HU-beta with matched b- and y-ions.

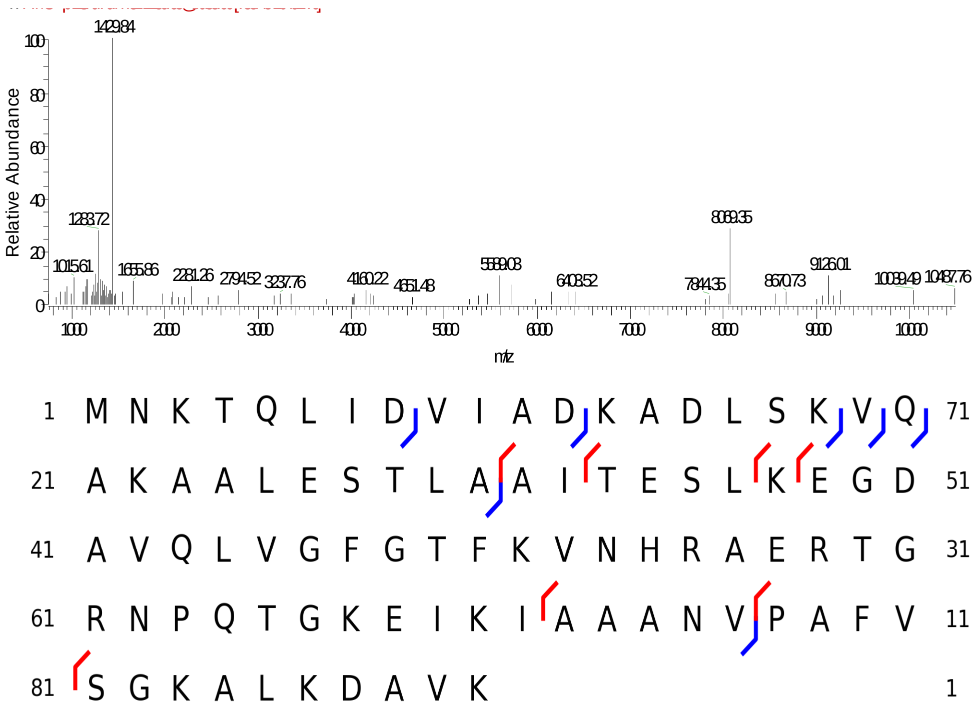


Figure 4.20-Top: MS/MS of m/z 1189.65 with +8 charge (molecular weight 9505.00). Bottom: Sequence of DNA binding protein HU-alpha with matched b- and y- ions.

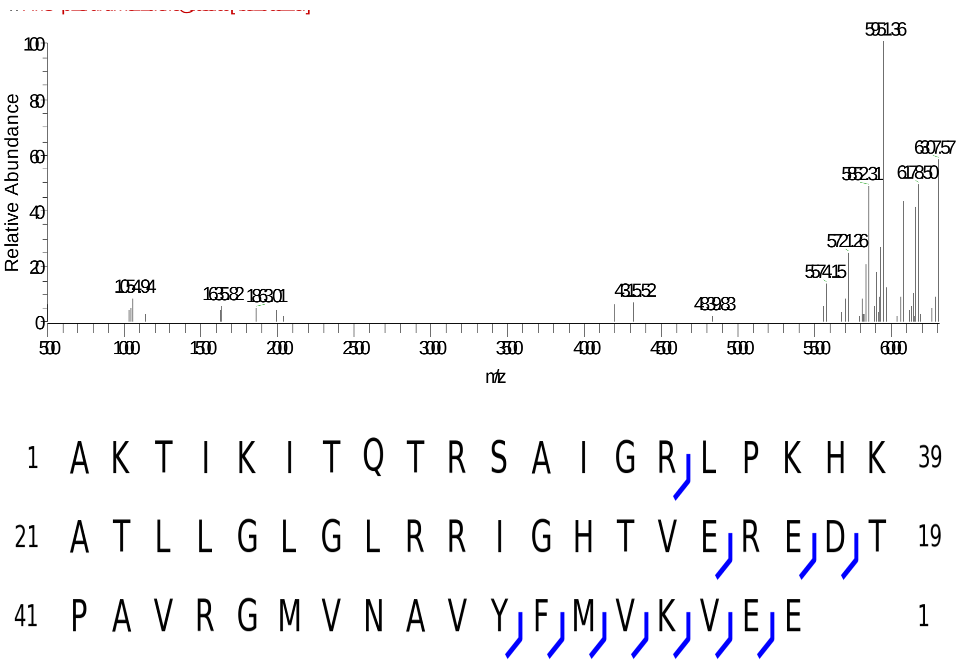


Figure 4.21-Top: MS/MS of m/z 1078.75 with +6 charge (molecular weight 6458.68). Bottom: Sequence of ribosomal protein L30 with matched b ions.

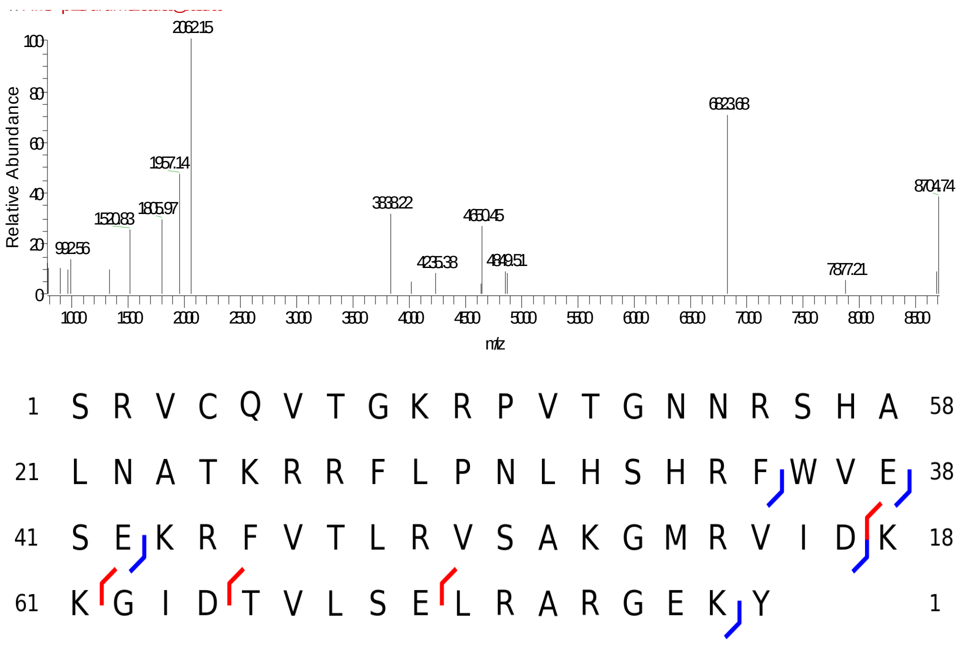


Figure 4.22-Top: MS/MS of m/z of 809.36 with +11 charge (molecular weight 8891.36). Bottom: Sequence of ribosomal protein L28 with matched b- and y-ions.

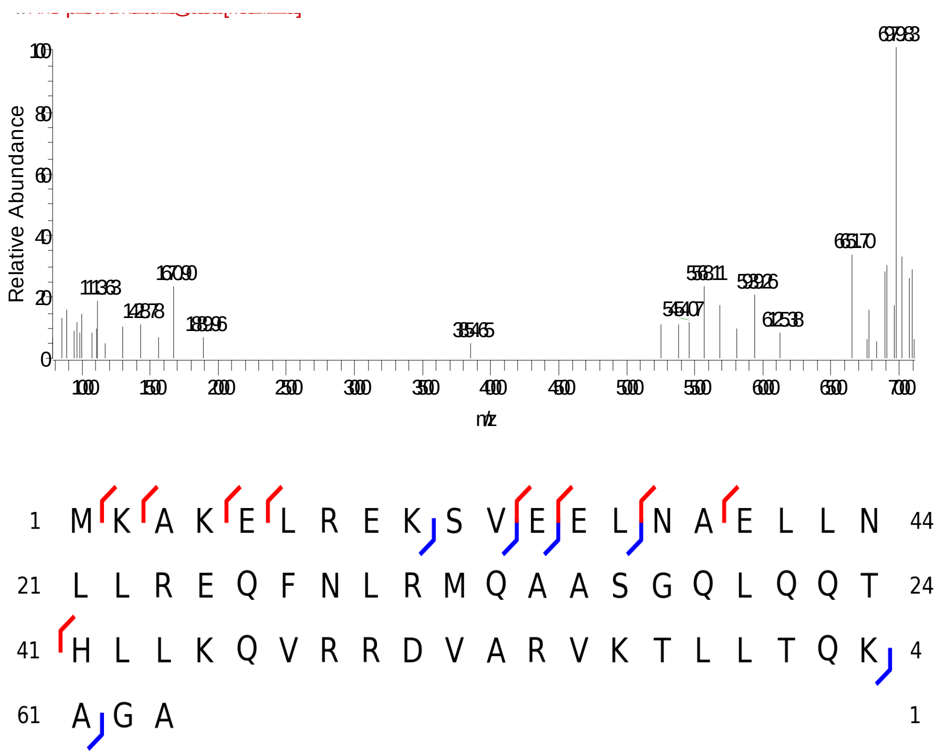


Figure 4.23-Top: MS/MS of m/z 907.25 with +8 charge (molecular weight 7243.48).

Bottom: Sequence of ribosomal protein L29 with matched b- and y- ions.

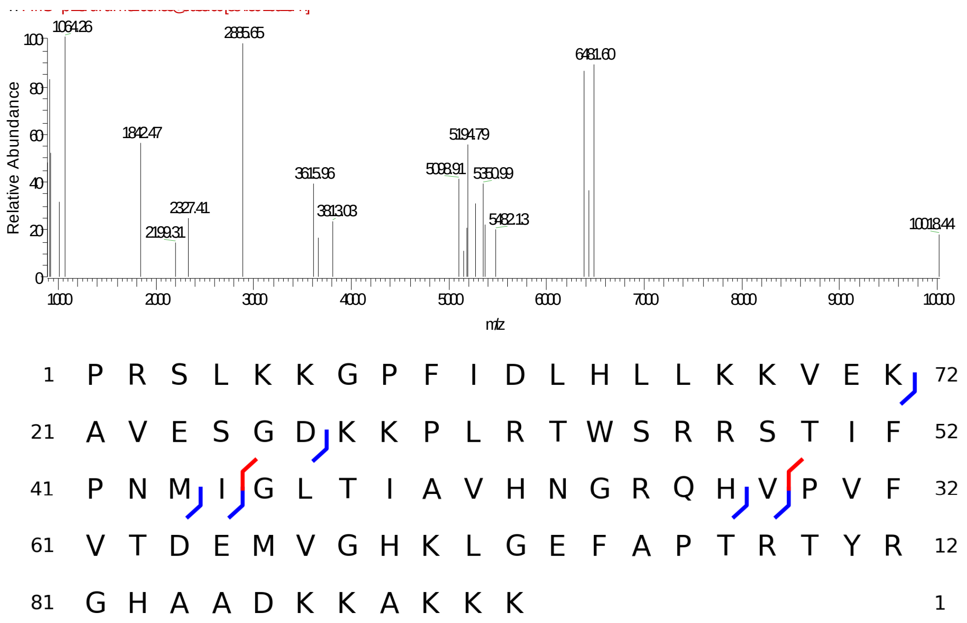


Figure 4.24-Top: MS/MS of m/z 937.16 with +11 charge (molecular weight 10299.10). Bottom: Sequence of ribosomal protein S19 with matched b- and y-ions.

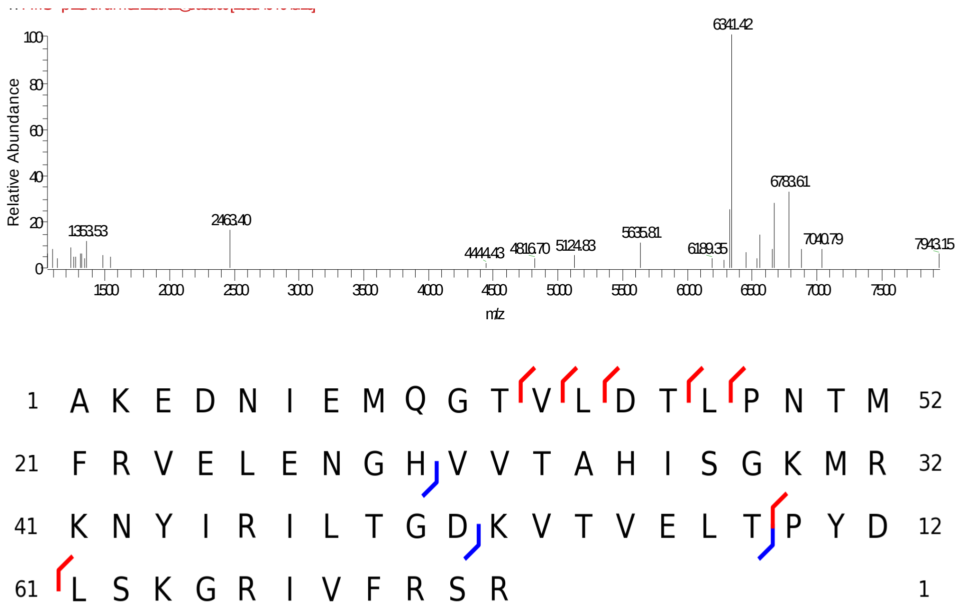


Figure 4.25-Top: MS/MS of m/z 1159.62 with +7 charge (molecular weight 8104.38)

Bottom: Sequence of translation initiation factor 1A with matched b- and y- ions.

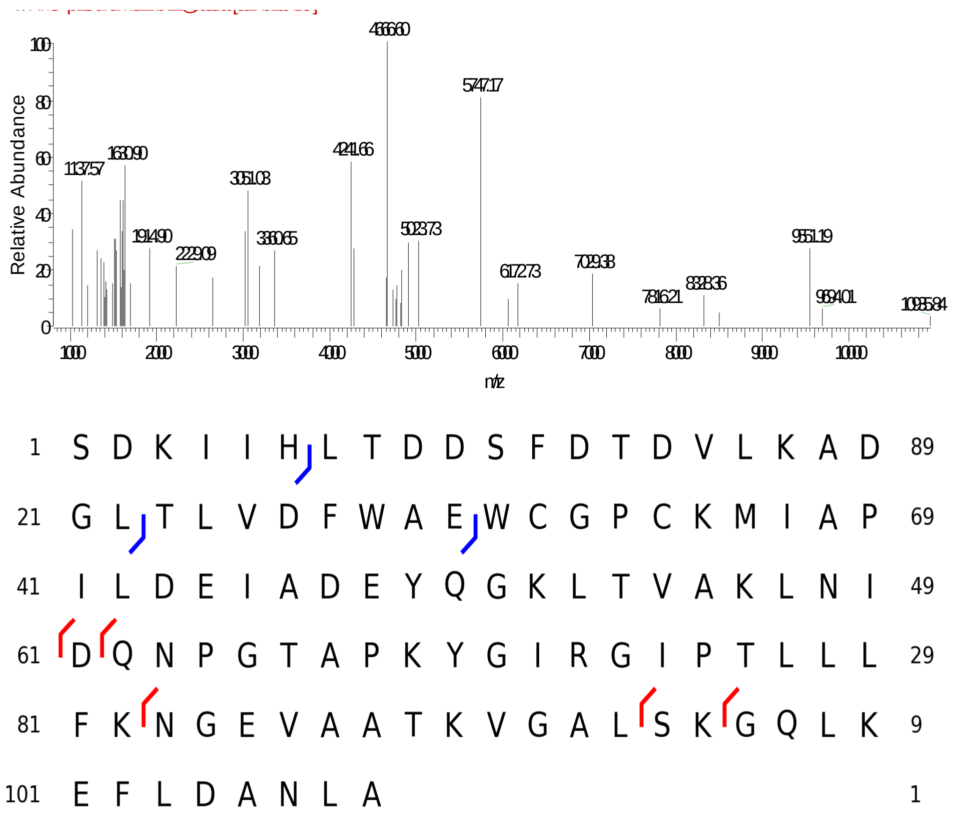


Figure 4.26-Top: MS/MS of m/z 1464.22 of +9 charge (molecular weight 11705.50).

Bottom: Sequence of thioredoxin protein with matched b- and y- ions.

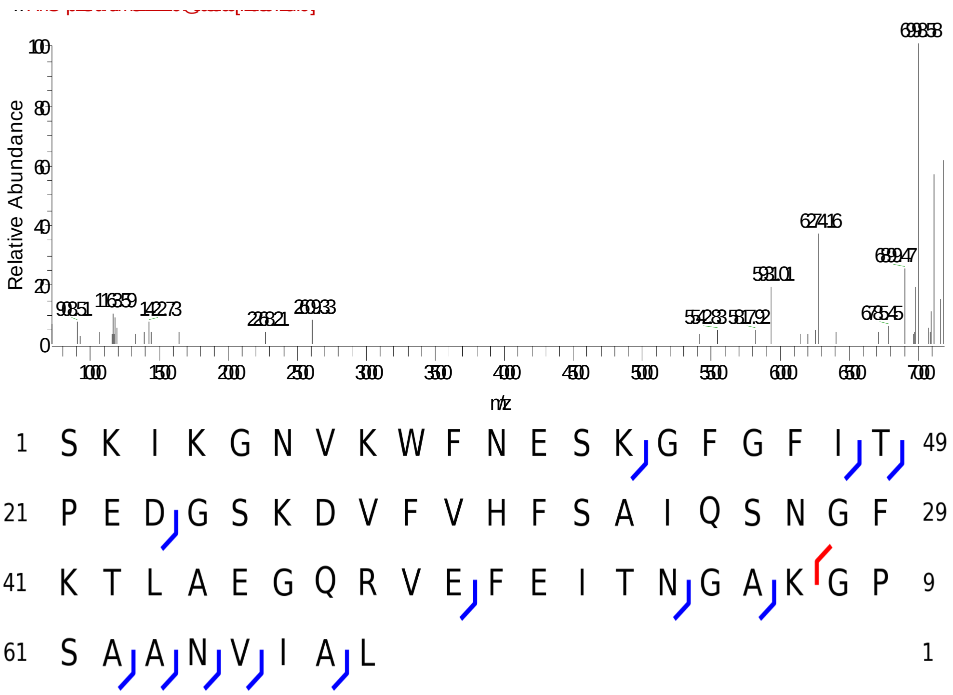


Figure 4.27-Top: MS/MS of 1221.64 with +6 charge (molecular weight 7318.23). Bottom: Sequence of cold shock protein Cold shock DNA protein-beta with matched b- and y- ions.

## *Phylogenetic Analysis*

This phylogenetic analysis had to be done in a slightly different manner than the previous chapter due to the extended mass tolerance of the precursor. Because the mass of the entire protein cannot be confidently matched to the proteins in the FASTA database, it can no longer be asserted that the whole protein sequence can be confidently identified for the construction of a phylogenetic tree. To overcome this problem, partial N- and C- terminal amino-acid sequences that could be confidently matched were extracted. The threshold used to verify that the partial sequence was correct was that an N-terminal sequence must have been supported by at least 3 b-ions, and the C-terminal sequences must have had at least 3 y-ions to be considered confidently identified. The whole protein sequence was considered to be confidently established only when these extracted N- or C-terminal sequences met or overlapped, and only those N- or C-terminal sequences consisting of ten amino acids or longer were retained for the phylogenetic analysis.

For *Erwinia herbicola*, the entire sequence was confidently established from only five of the fourteen ProSight PC matches.  N- and C- terminal sequences with 32 or more amino acids were established in three others.  Three more only had N-terminal sequences extracted, while three other had C-terminal sequences extracted. These extracted sequences for *Erwinia herbicola* averaged 60 amino acids per ProSight PC top-down match.  Figure 4.6 provides an example of both N- and C-terminal extracted fragments, while Figure 4.7 shows where the extracted sequences resided on each of the *Erwinia herbicola* proteins.

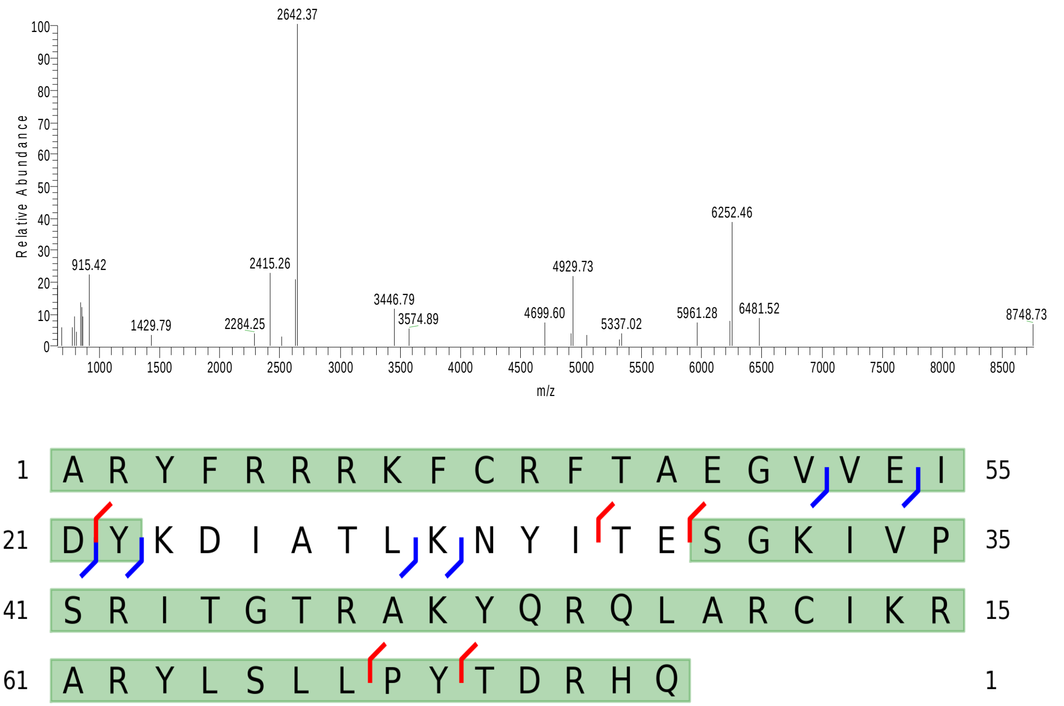


Figure 4.28-Top: Deconvoluted MS/MS of precursor ion 732.91 with charge state +12, matching to Ribosomal protein S18 with an E-value of 6.28e-6. Bottom: ProSight PC 2.0 match with 11 b- and y- ion fragments. The highlighted portion indicates the N- and C-terminal amino acid sequences extracted for phylogenetic analysis.

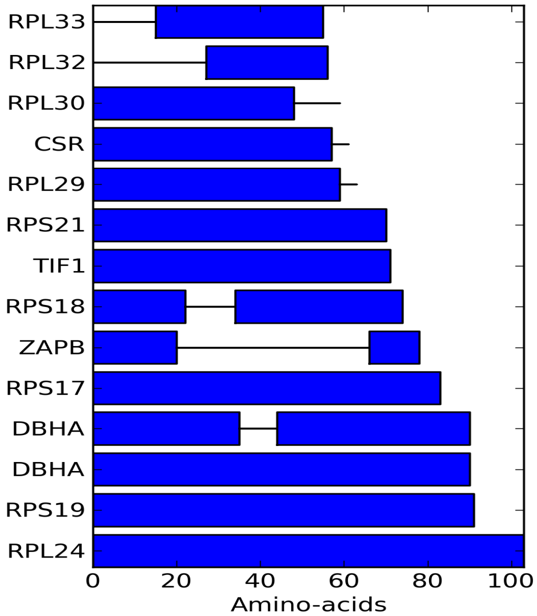


Figure 4.29-histogram denoting in blue where the N- and C-terminal sequences were extracted for phylogenetic analysis for the *Erwinia herbicola* matches.

Similarly, there were five full length sequences extracted from the protein matches of *Enterobacter cloacae*. Three other matches had N- and C-terminal sequence pairs extracted, while five matches only had their N-terminal sequences extracted. Two other matches had only C-terminal sequences extracted for the phylogenetic analysis. On average, fifty amino acids were extracted for each protein matched to *Enterobacter cloacae* to construct its phylogenetic tree.

The confidently extracted amino-acid sequences were aligned with proteins from the *Enterobacteriaceae* proteins to identify homologous regions in the related species. Those N- and C- terminal sequences could not be aligned to any homologous sequences with E-values of 1e-4 or less and were discarded. Furthermore, any extracted sequence with an exact alignment with another extracted sequence was also discarded. Any species with a significant alignment to all remaining sequences was used for multiple-sequence alignment and phylogenetic tree construction. The best alignment to each extracted sequence was concatenated in the same order for each species. These alignments were combined with the extracted sequences for *Erwinia herbicola* and *Enterobacter cloacae*, just as in the previous chapter. *Erwinia herbicola* could be compared with twenty-seven other species using about eight hundred amino acid positions, while *Enterobacter cloacae* could be compared to twenty-six other species at about half the amino acid positions as *E. herbicola*. Figure 4.8 shows the phylogenetic trees for each species, as created in the same manner as in the previous chapter using the phylogeny.fr web-server.

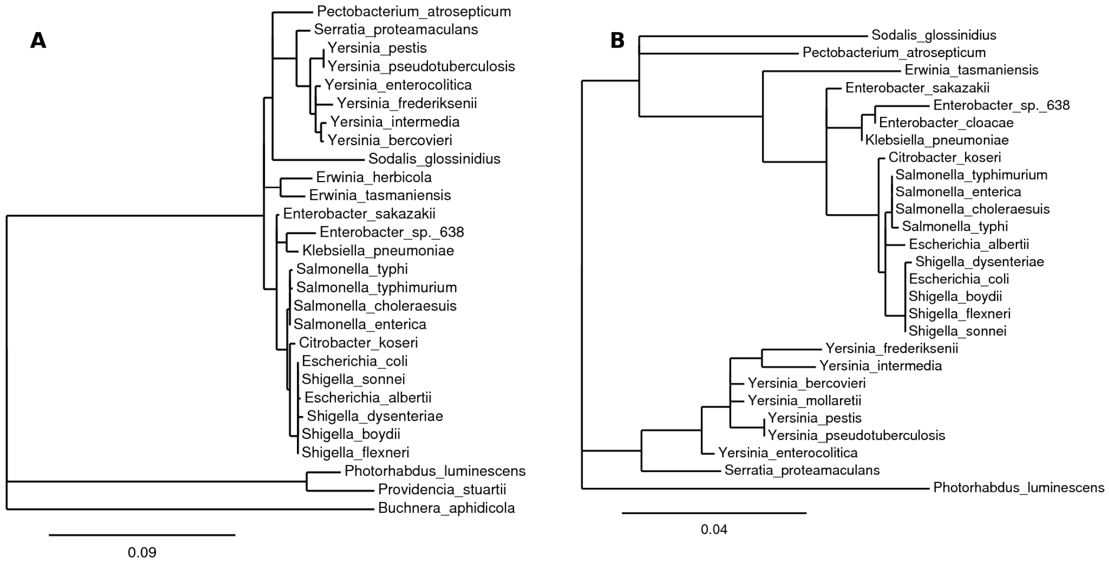


Figure 4.30-Phylogenetic trees constructed for *Erwinia herbicola* (A) and *Enterobacter cloacae* (B)

### *Summary*

The previous chapter of this dissertation laid out the top-down strategy that could be used to identify microorganisms without sequenced genomes by using proteins from related organisms that have the same mass as the target proteins, with the presumption that those proteins will be 100% homologous. This chapter expands this strategy to allow for a 250 Dalton mass difference between the target proteins characterized in the top-down experiment and the proteins from the same phylogenetic family using two different bacteria. Fourteen proteins were identified from *Erwinia herbicola* and fifteen proteins were identified from *Enterobacter cloacae* using the proteins from the entire *Enterobacteriaceae* family. While the whole protein sequence could not be characterized in most cases, precursor and fragment mass spectra acquired with high mass accuracy and high resolution allowed for confident characterization of long N- and C-terminal pieces of those proteins identified by their b- and y- ions. These partial sequences were then used in phylogenetic analysis and discovered enough differentiation to construct complex phylogenetic trees for each of the target bacteria.

High mass accuracy of the precursor and fragment spectra also provided enough information in some cases to be able to localize the mass changes between the target and matched amino acid sequences, and, in some cases, allowed for the presumption of amino acid mutations. These presumed mutations identified more b- and y- ions than the initial match and increased the confidence in those identifications by a significant margin. Three examples of this were shown in *Erwinia herbicola* proteins.

This expansion of the previous method should allow for the classification of more bacteria without sequenced genomes since it is not reliant on 100% homology between the target proteins and the proteins from the sequenced members of the same family. However, it does still rely on some prior knowledge of the bacteria to create a database of adequate size. By limiting the database to only those proteins derived from sequenced organisms within the same family, the author feels that the database will be inclusive enough to provide enough homology matches without the database file becoming so large that a regular PC does not take days to attempt to match the target proteins.

Furthermore, this expansion of the method still characterized many of the highly abundant proteins that would be found in a typical MALDI-TOF mass spectrum that resulted from some of the detection systems used in defense and homeland security. By characterizing these proteins, the information could be incorporated into these systems and allow for better discrimination between a lethal pathogen and non-lethal simulant organism.

# Chapter 5: Conclusions

## *Identification of Proteins*

Few bacteria and archea have publicly available genome or proteome sequences, so any mass spectrometry-based method studying those without publicly available sequences needs to accommodate mass differences stemming from amino acid differences and unexpected post-translational modifications. The strategy proposed in this dissertation uses top-down mass spectrometry and the existing proteome/genome database to characterize proteins from these unsequenced bacteria and identify the organism's place on a phylogenic tree. The method was tested on three bacteria that are used as simulants for biohazards. First, the method was used on *Yersinia rohdei* to measure only those proteins that had the same exact masses as proteins from a custom database. Extensive fragmentation of proteins with the same masses as known proteins in the database often allow the assignment of 100% homology. Next, the method was expanded to accommodate a 250 Dalton difference between the masses known and unknown proteins from two other proteins without sequenced genomes, *Erwinia herbicola* and *Enterobacter cloacae*, and the database of proteins was extended to all sequenced bacteria from the Enterobacteria family. Proteins from all three of our tests were identified and characterized very confidently using the metrics that were laid out in the dissertation. From those protein identifications, phylogenic trees were constructed to determine which of the sequenced bacteria were closest to the unsequenced target.

Top-down mass spectrometry was the proteomic approach taken because of this approach provides complete coverage of the amino acid sequence of each protein detected. The top-down approach first measures the mass of the whole protein, which would include any amino acid changes or post translational modifications to the protein. In the alternative approaches, bottom-up or middle-down, the protein is broken into peptides before mass spectrometric analysis. Due to incomplete ionization, differences in relative abundance in co-eluting peptides, and incomplete separation, not every peptide is detected. Chances are that the peptide will not be detected. To make a similar bottom-up approach as successful to the method proposed in this dissertation, multiple enzymatic or chemical cleavage methods would have to be employed. Use of multiple methods to digest the proteins will make the sample more complex, and analysis would take much longer than the times used in the top-down studies already discussed. In the present study, seven out of the fourteen proteins identified from *Erwinia herbicola* and three of the fifteen proteins identified from *Enterobacter cloacae* had masses different from the masses of the protein sequences in the database. Therefore, a top-down proteomic approach is favorable when matching proteins from an unsequenced bacteria to bacterial proteins in a database.

### *Orbitrap Mass Spectrometry*

High resolution and high mass accuracy provided by the Orbitrap analyzer was used in conjunction with collisionally induced dissociation to identify proteins ranging from 5000 Daltons to 15000 Daltons. Molecular weights could be automatically determined in ions with fifteen charges in the timeframe allowed by online HPLC fractionation. The Fourier transform used by the Orbitrap provides resolution that can detect a m/z difference in the isotopes of 0.067, which results from that high of a charge. Many fragments generated by CID remain charged and also require high resolution for analysis. Charge deconvolution software, like THRASH, allows for automated analysis of these highly charged molecular and fragment ions by bringing all precursor and fragment ions to only one m/z peak per ion.  Simplifying the mass spectra in this manner simplifies the peak picking needed in the database searching, which makes those searches faster eases interpretation of those results.

### Sequence Homology

The method proposed in this dissertation allowed for some phylogenic differentiation even though it identified only those proteins that are highly abundant in the bacteria.  Most of the proteins identified in the three target bacteria were characterized as ribosomal proteins.  In the case of *Yersinia rohdei,* all but one of the proteins identified were ribosomal proteins.  These ribosomal proteins are generally have highly conserved amino acid sequences among members of the same family, and these experiments were able to differentiate the bacteria based on those identifications.  As shown by the two incidence matrices, ribosomal proteins L32 and L29 allowed for separation due to different sequences.  In the case of the *Erwinia herbicola,* this identification led to the determination that *Erwinia tasmaniensis* was the closest known relative to the target species.  Other ribosomal proteins, such as S21, were so conserved that no differentiation between the members of the family could be determined by itself.

In this top-down strategy, the whole sequence does not have to be characterized by fragment b and y ions for the protein to be confidently identified. The *Erwinia herbicola* and *Enterobacter cloacae* studies both showed that gaps in the fragment ions and mass differences between the target proteins and the proteins contained in the database do not prevent confident identifications.  Gaps between the identified b and y ions in a protein occur in most of the proteins that have mass differences. These gaps help localize where these mass differences can occur, either through post translational modifications or differences in the amino acid sequence.

### *Future Directions*

The next iteration of the method demonstrated in this dissertation would be to improve sensitivity of detection and separation of the proteins while also increasing the precursor mass tolerance. These changes would probably allow for more protein identifications and identifications of less abundant proteins. This would lead to more reliable phylogenic identification. Implementing an alternative sample preparation is expected to allow for the identification of more acidic proteins, which could also lead to more unique identifications. Rapid and effective methods that provide additional protein identifications and phylogenetic characterization of bacteria that lack sequence information will be of value in homeland security, epidemiologic and medical diagnostics, and food safety, as well as enhance basic research.

# Bibliography

1. Dass, C., *Principles and Practice of Biological Mass Spectrometry*. Wiley-Interscience: New York, 2001.

2. Griffiths, W. J.; Jonsson, A. P.; Liu, S. Y.; Rai, D. K.; Wang, Y. Q., Electrospray and tandem mass spectrometry in biochemistry. *Biochemical Journal* **2001,** 355, 545-561.

3. Tanaka, K.; Hiroaki, W.; Yutaka, I.; Satoshi, A.; Yoshikazu, Y.; Tamio, Y.; Matsuo, T., Protein and polymer analyses up to <I>m/z</I> 100 000 by laser ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry* **1988,** 2, (8), 151-153.

4. Karas, M.; Gluckmann, M.; Schafer, J., Ionization in matrix-assisted laser desorption/ionization: singly charged molecular ions are the lucky survivors. *Journal of Mass Spectrometry* **2000,** 35, (1), 1-12.

5. Fenselau, C.; Demirev, P. A., Characterization of intact microorganisms by MALDI mass spectrometry. *Mass Spectrometry Reviews* **2001,** 20, (4), 157-171.

6. Holland, R. D.; Wilkes, J. G.; Rafii, F.; Sutherland, J. B.; Persons, C. C.; Voorhees, K. J.; Lay, J. O., Rapid identification of intact whole bacteria based on spectral patterns using matrix-assisted laser desorption/ionization with time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry* **1996,** 10, (10), 1227-1232.

7. Demirev, P. A.; Ho, Y. P.; Ryzhov, V.; Fenselau, C., Microorganism identification by mass spectrometry and protein database searches. *Analytical Chemistry* **1999,** 71, (14), 2732-2738.

8. Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M., ELECTROSPRAY IONIZATION FOR MASS-SPECTROMETRY OF LARGE BIOMOLECULES. *Science* **1989,** 246, (4926), 64-71.

9. Gatlin, C. L.; Kleemann, G. R.; Hays, L. G.; Link, A. J.; Yates, J. R., 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, (1), 93-101.

10. Pribil, P.; Fenselau, C., Characterization of Enterobacteria using MALDI-TOF mass spectrometry. *Analytical Chemistry* **2005,** 77, (18), 6092-6095.

11. Cornish, T. J.; Cotter, R. J., A CURVED-FIELD REFLECTRON FOR IMPROVED ENERGY FOCUSING OF PRODUCT IONS IN TIME-OF-FLIGHT MASS-SPECTROMETRY. *Rapid Communications in Mass Spectrometry* **1993,** 7, (11), 1037-1040.

12. Paul, W.; Reinhard, H. P.; Vonzahn, U., DAS ELEKTRISCHE MASSENFILTER ALS MASSENSPEKTROMETER UND ISOTOPENTRENNER. *Zeitschrift Fur Physik* **1958,** 152, (2), 143-182.

13. Schwartz, J. C.; Senko, M. W.; Syka, J. E. P., A two-dimensional quadrupole ion trap mass spectrometer. *Journal of the American Society for Mass Spectrometry* **2002,** 13, (6), 659-669.

14. McLuckey, S. A.; Vanberkel, G. J.; Glish, G. L.; Huang, E. C.; Henion, J. D., ION SPRAY LIQUID-CHROMATOGRAPHY ION TRAP MASS-

SPECTROMETRY DETERMINATION OF BIOMOLECULES. *Analytical Chemistry* **1991,** 63, (4), 375-383.

15. Biemann, K.; Scoble, H. A., CHARACTERIZATION BY TANDEM MASS-SPECTROMETRY OF STRUCTURAL MODIFICATIONS IN PROTEINS. *Science* **1987,** 237, (4818), 992-998.

16. Roepstorff, P.; Fohlman, J., PROPOSAL FOR A COMMON NOMENCLATURE FOR SEQUENCE IONS IN MASS-SPECTRA OF PEPTIDES. *Biomedical Mass Spectrometry* **1984,** 11, (11), 601-601.

17. Biemann, K., NOMENCLATURE FOR PEPTIDE FRAGMENT IONS (POSITIVE-IONS). *Methods in Enzymology* **1990,** 193, 886-887.

18. Makarov, A., Electrostatic axially harmonic orbital trapping: A high-performance technique of mass analysis. *Analytical Chemistry* **2000,** 72, (6), 1156-1162.

19. Hu, Q. Z.; Noll, R. J.; Li, H. Y.; Makarov, A.; Hardman, M.; Cooks, R. G., The Orbitrap: a new mass spectrometer. *Journal of Mass Spectrometry* **2005,** 40, (4), 430-443.

20. Makarov, A.; Denisov, E.; Kholomeev, A.; Baischun, W.; Lange, O.; Strupat, K.; Horning, S., Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. *Analytical Chemistry* **2006,** 78, (7), 2113-2120.

21. Wasinger, V. C.; Cordwell, S. J.; Cerpapoljak, A.; Yan, J. X.; Gooley, A. A.; Wilkins, M. R.; Duncan, M. W.; Harris, R.; Williams, K. L.; Humpherysmith, I., PROGRESS WITH GENE-PRODUCT MAPPING OF THE MOLLICUTES - MYCOPLASMA-GENITALIUM. *Electrophoresis* **1995,** 16, (7), 1090-1094.

22. Qin, J.; Fenyo, D.; Zhao, Y. M.; Hall, W. W.; Chao, D. M.; Wilson, C. J.; Young, R. A.; Chait, B. T., A strategy for rapid, high confidence protein identification. *Analytical Chemistry* **1997,** 69, (19), 3995-4001.

23. Reid, G. E.; McLuckey, S. A., 'Top down' protein characterization via tandem mass spectrometry. *Journal of Mass Spectrometry* **2002,** 37, (7), 663-675.

24. Demirev, P. A.; Feldman, A. B.; Kowalski, P.; Lin, J. S., Top-down proteomics for rapid identification of intact microorganisms. *Analytical Chemistry* **2005,** 77, (22), 7455-7461.

25. Parks, B. A.; Jiang, L.; Thomas, P. M.; Wenger, C. D.; Roth, M. J.; Boyne, M. T.; Burke, P. V.; Kwast, K. E.; Kelleher, N. L., Top-down proteomics on a chromatographic time scale using linear ion trap Fourier transform hybrid mass spectrometers. *Analytical Chemistry* **2007,** 79, (21), 7984-7991.

26. Kelleher, N. L.; Lin, H. Y.; Valaskovic, G. A.; Aaserud, D. J.; Fridriksson, E. K.; McLafferty, F. W., Top down versus bottom up protein characterization by tandem high-resolution mass spectrometry. *Journal of the American Chemical Society* **1999,** 121, (4), 806-812.

27. Horn, D. M.; Zubarev, R. A.; McLafferty, F. W., Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *Journal of the American Society for Mass Spectrometry* **2000,** 11, (4), 320-332.

28. Yergey, J. A.; Cotter, R. J.; Heller, D.; Fenselau, C., RESOLUTION REQUIREMENTS FOR MIDDLE-MOLECULE MASS-SPECTROMETRY. *Analytical Chemistry* **1984,** 56, (12), 2262-2263.

29. Rubino, F. M.; Danieli, B.; Chillemi, F.; Cremonesi, A., FAST-ATOM-BOMBARDMENT AND TANDEM MASS-SPECTROMETRY AT HIGH AND LOW COLLISION ENERGY FOR THE SEQUENCE-ANALYSIS OF LOW TO MIDDLE-MASS PEPTIDES. *Biological Mass Spectrometry* **1992,** 21, (9), 451-462.

30. David, N. P.; Darryl, J. C. P.; David, M. C.; John, S. C., Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999,** 20, (18), 3551-3567.

31. Boyne, M. T.; Garcia, B. A.; Li, M. X.; Zamdborg, L.; Wenger, C. D.; Babai, S.; Kelleher, N. L., Tandem Mass Spectrometry with Ultrahigh Mass Accuracy Clarifies Peptide Identification by Database Retrieval. *Journal of Proteome Research* **2009,** 8, (1), 374-379.

32. Bernal, A.; Ear, U.; Kyrpides, N., Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Research* **2001,** 29, (1), 126-127.

33. Haas, A. L.; Rose, I. A., THE MECHANISM OF UBIQUITIN ACTIVATING ENZYME - A KINETIC AND EQUILIBRIUM-ANALYSIS. *Journal of Biological Chemistry* **1982,** 257, (17), 329-337.

34. Reid, G. E.; Wu, J.; Chrisman, P. A.; Wells, J. M.; McLuckey, S. A., Charge-state-dependent sequence analysis of protonated ubiquitin ions via ion trap tandem mass spectrometry. *Analytical Chemistry* **2001,** 73, (14), 3274-3281.

35. Novak, P.; Kruppa, G. H.; Young, M. M.; Schoeniger, J., A top-down method for the determination of residue-specific solvent accessibility in proteins. *Journal of Mass Spectrometry* **2004,** 39, (3), 322-328.

36. Stephenson, J. L.; McLuckey, S. A.; Reid, G. E.; Wells, J. M.; Bundy, J. L., Ion/ion chemistry as a top-down approach for protein analysis. *Current Opinion in Biotechnology* **2002,** 13, (1), 57-64.

37. Waanders, L. F.; Hanke, S.; Mann, M., Top-down quantitation and characterization of SILAC-labeled proteins. *Journal of the American Society for Mass Spectrometry* **2007,** 18, (11), 2058-2064.

38. Dongre, A. R.; Jones, J. L.; Somogyi, A.; Wysocki, V. H., Influence of peptide composition, gas-phase basicity, and chemical modification on fragmentation efficiency: Evidence for the mobile proton model. *Journal of the American Chemical Society* **1996,** 118, (35), 8365-8374.

39. Harrison, A. G.; Yalcin, T., Proton mobility in protonated amino acids and peptides. *International Journal of Mass Spectrometry* **1997,** 165, 339-347.

40. Fagerquist, C. K.; Garbus, B. R.; Miller, W. G.; Williams, K. E.; Yee, E.; Bates, A. H.; Boyle, S.; Harden, L. A.; Cooley, M. B.; Mandrell, R. E., Rapid Identification of Protein Biomarkers of Escherichia coil O157:H7 by Matrix-Assisted Laser Desorption Ionization-Time-of-Flight-Time-of-Flight Mass Spectrometry and Top-Down Proteomics. *Analytical Chemistry* 82, (7), 2717-2725.

41. Demirev, P. A.; Ramirez, J.; Fenselau, C., Tandem mass spectrometry of intact proteins for characterization of biomarkers from Bacillus cereus T spores. *Analytical Chemistry* **2001,** 73, (23), 5725-5731.

42. Ryzhov, V.; Fenselau, C., Characterization of the protein subset desorbed by MALDI from whole bacterial cells. *Analytical Chemistry* **2001,** 73, (4), 746-750.

43. Aleksic, S.; Steigerwalt, A. G.; Bockemuhl, J.; Huntleycarter, G. P.; Brenner, D. J., YERSINIA-ROHDEI SP-NOV ISOLATED FROM HUMAN AND DOG FECES AND SURFACE-WATER. *International Journal of Systematic Bacteriology* **1987,** 37, (4), 327-332.

44. Sambrook, J.; Fritsch, E. F.; Maniatis, T., *Molecular cloning: a laboratory manual*. 2nd ed.; Cold Spring Harbor Press: Cold Spring Harbor, N.Y., 1989.

45. Pineda, F. J.; Antoine, M. D.; Demirev, P. A.; Feldman, A. B.; Jackman, J.; Longenecker, M.; Lin, J. S., Microorganism identification by matrix-assisted laser/desorption ionization mass spectrometry and model-derived ribosomal protein biomarkers. *Analytical Chemistry* **2003,** 75, (15), 3817-3822.

46. Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J., BASIC LOCAL ALIGNMENT SEARCH TOOL. *Journal of Molecular Biology* **1990,** 215, (3), 403-410.

47. Thompson, J. D.; Higgins, D. G.; Gibson, T. J., CLUSTAL-W - IMPROVING THE SENSITIVITY OF PROGRESSIVE MULTIPLE SEQUENCE ALIGNMENT THROUGH SEQUENCE WEIGHTING, POSITION-SPECIFIC GAP PENALTIES AND WEIGHT MATRIX CHOICE. *Nucleic Acids Research* **1994,** 22, (22), 4673-4680.

48. Meng, F. Y.; Cargile, B. J.; Miller, L. M.; Forbes, A. J.; Johnson, J. R.; Kelleher, N. L., Informatics and multiplexing of intact protein identification in bacteria and the archaea. *Nature Biotechnology* **2001,** 19, (10), 952-957.

49. Senko, M. W.; Beu, S. C.; McLafferty, F. W., DETERMINATION OF MONOISOTOPIC MASSES AND ION POPULATIONS FOR LARGE BIOMOLECULES FROM RESOLVED ISOTOPIC DISTRIBUTIONS. *Journal of the American Society for Mass Spectrometry* **1995,** 6, (4), 229-233.

50. Edwards, N. J., Pineda, R. In *Rapid Microorganism Identification Database (RMIDb)*, 54th Annual Conference of the American Society for Mass Spectrometry, Seattle, WA, May 28-June 1, 2006; Seattle, WA, 2006.

51. Peterson, J. D.; Umayam, L. A.; Dickinson, T.; Hickey, E. K.; White, O., The Comprehensive Microbial Resource. *Nucleic Acids Research* **2001,** 29, (1), 123-125.

52. Delcher, A. L., Bratke, K.A., Powers, E.C., Salzberg, S.L, Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **2007,** 23, (6), 673-679.

53. Finn, R. D.; Tate, J.; Mistry, J.; Coggill, P. C.; Sammut, S. J.; Hotz, H. R.; Ceric, G.; Forslund, K.; Eddy, S. R.; Sonnhammer, E. L. L.; Bateman, A., The Pfam protein families database. *Nucleic Acids Research* **2008,** 36, D281-D288.

54. Dereeper, A.; Guignon, V.; Blanc, G.; Audic, S.; Buffet, S.; Chevenet, F.; Dufayard, J. F.; Guindon, S.; Lefort, V.; Lescot, M.; Claverie, J. M.; Gascuel, O., Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Research* **2008,** 36, W465-W469.

55. Cole, J. R.; Wang, Q.; Cardenas, E.; Fish, J.; Chai, B.; Farris, R. J.; Kulam-Syed-Mohideen, A. S.; McGarrell, D. M.; Marsh, T.; Garrity, G. M.; Tiedje, J. M., The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Research* **2009,** 37, D141-D145.

56. Fenselau, C.; Russell, S.; Swatkoski, S.; Edwards, N., Proteomic strategies for rapid characterization of micro-organisms. *European Journal of Mass Spectrometry* **2007,** 13, (1), 35-39.

57. Teramoto, K.; Sato, H.; Sun, L.; Torimura, M.; Tao, H.; Yoshikawa, H.; Hotta, Y.; Hosoda, A.; Tamura, H., Phylogenetic classification of pseudomonas putida strains by MALDI-MS using ribosomal subunit proteins as biomarkers. *Analytical Chemistry* **2007,** 79, (22), 8712-8719.

58. Buttner, M. P.; Cruz, P.; Stetzenbach, L. D.; Cronin, T., Evaluation of two surface sampling methods for detection of Erwinia herbicola on a variety of materials by culture and quantitative PCR. *Applied and Environmental Microbiology* **2007,** 73, (11), 3505-3510.

59. El-Masry, M. H.; Brown, T. A.; Epton, H. A. S.; Sigee, D. C., Transfer from Erwinia herbicola to Escherichia coli of a plasmid associated with biocontrol of fire blight. *Plant Pathology* **1997,** 46, (6), 865-870.

60. Breathnach, A. S.; Riley, P. A.; Shad, S.; Jownally, S. M.; Law, R.; Chin, P. C.; Kaufmann, M. E.; Smith, E. J., An outbreak of wound infection in cardiac surgery patients caused by Enterobacter cloacae arising from cardioplegia ice. *Journal of Hospital Infection* **2006,** 64, (2), 124-128.

61. Clementino, M. M.; De Filippis, I.; Nascimento, C. R.; Branquinho, R.; Rocha, C. L.; Martins, O. B., PCR analyses of tRNA intergenic spacer, 16S-23S internal transcribed spacer, and randomly amplified polymorphic DNA reveal inter- and intraspecific relationships of Enterobacter cloacae strains. *Journal of Clinical Microbiology* **2001,** 39, (11), 3865-3870.