To the Graduate Council:

I am submitting herewith a dissertation written by Evgenia Razumovskaya entitled "Computational Analysis of Mass Spectrometric Data for Whole Organism Proteomic Studies." I have examined the final electronic copy of this dissertation for form and content and recommended that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Life Sciences.

Edward Uberbacher
Major Professor

We have read this dissertation
and recommended it acceptance:

Frank Larimer

Greg Hurst

Barry Bruce

Acceptance for the Council:

Anne Mayhew
Vice Chancellor and Dean
Of Graduate Studies

(Original signatures are on file with official student records.)

COMPUTATIONAL ANALYSIS OF MASS SPECTROMETRIC DATA FOR WHOLE ORGANISM PROTEOMIC STUDIES

A dissertation presented for the Doctor of Philosophy Degree
The University of Tennessee, Knoxville

Evgenia Razumovski

May 2006

ABSTRACT

In the last decades great breakthroughs have been achieved in the study of the genomes, supplying us with the vast knowledge of the genes and a large number of sequenced organisms. With the availability of genome information, the new systematic studies have arisen. One of the most prominent areas is proteomics. Proteomics is a discipline devoted to the study of the organism's expressed protein content. Proteomics studies are concerned with a wide range of problems. Some of the major proteomics focuses upon the studies of protein expression patterns, the detection of protein-protein interactions, protein quantitation, protein localization analysis, and characterization of post-translational modifications. The emergence of proteomics shows great promise to furthering our understanding of the cellular processes and mechanisms of life.

One of the main techniques used for high-throughput proteomic studies is mass spectrometry. Capable of detecting masses of biological compounds in complex mixtures, it is currently one of the most powerful methods for protein characterization. New horizons are opening with the new developments of mass spectrometry instrumentation, which can now be applied to a variety of proteomic problems. One of the most popular applications of proteomics involves whole organism high-throughput experiments. However, as new instrumentation is being developed, followed by the design of new experiments, we find ourselves needing new computational algorithms to interpret the results of the experiments. As the thresholds of the current technology are being probed, the new algorithmic designs are beginning to emerge to meet the challenges of the mass spectrometry data evaluation and interpretation.

This dissertation is devoted to computational analysis of mass spectrometric data, involving a combination of different topics and techniques to improve our understanding of biological processes using high-throughput whole organism proteomic studies. It consists of the development of new algorithms to improve the data interpretation of the current tools, introducing a new algorithmic approach for post-translational modification detection, and the characterization of a set of computational simulations for biological agent detection in a complex organism background. These studies are designed to further the capabilities of understanding the results of high-throughput mass spectrometric experiments and their impact in the field of proteomics.

TABLE OF CONTENTS

Chapter 1-Introduction to Proteomic Analysis with Mass Spectrometry and General Mass Spectrometry Data Analysis..... 1

Chapter 2-A Computational Method for Assessing Peptide Identification Reliability in Tandem Mass Spectrometry Analysis with SEQUEST……..... 19

Chapter 3-Charge Determination for Low Resolution Mass Spectrometry….. 43

Chapter 4-Computational Identification of Post-Translational Modifications from Shotgun Mass Spectrometry Data……..... 71

Chapter 5-Computational Simulations for Mass Spectrometry-Based Identification of Biological Agents……. 100

List of References……..... 142

Appendix1……... 149

Appendix2……... 157

Vita……... 160

LIST OF TABLES

Table 2.1: Peptide assignment results for the ISB protein mixture……. 25

Table 2.2: Peak assignment results for Sample I, Sample II and Sample III made by SEQUEST……... 30

Table 2.3: Performance of Neural Network 1 on the 8-protein dataset……... 37

Table 2.4: Performance of Neural Network 2 on the 8-protein dataset……... 38

Table 3.1: Charge state assignment results for neural net training with *Seattle dataset*……. 61

Table 3.2: Charge assignments for *ORNL dataset*……. 66

Table 3.3: Charge assignments for *E. coli dataset*……..... 67

LIST OF FIGURES

Figure 1.1: Top down vs. bottom up mass spectrometric techniques…….. 4
Figure 1.2: Tandem MS fragments…….. 8
Figure 1.3: Sequence dependent fragmentation pattern produced by CID…….. 9
Figure 2.1: Histogram of correct and incorrect peptide assignments made by neural network…….. 27
Figure 2.2: Specificity vs. sensitivity for ISB dataset…….. 32
Figure 2.3: Specificity vs. sensitivity of eight-protein dataset……. 33
Figure 2.4: Specificity vs. sensitivity of protein detection…….. 40
Figure 3.1: Charge dependent patterns…… 53
Figure 3.2: Significance of spectral parameters for charge state determination….. 56
Figure 3.3: The charge separation results for the *Seattle dataset*…… 60
Figure 3.4: Estimation of error rates for charge state assignment in *E. coli* and *ORNL datasets*……. 64
Figure 4.1: Post-translational modifications…… 75
Figure 4.2: *R. palustris* growth conditions……... 79
Figure 4.3: Effect of post-translational modifications on fragmentation patterns... 84
Figure 4.4: Example entries in the RESID database…… 87
Figure 4.5: Tandem MS spectrum of uridylated peptide…… 95
Figure 4.6: Tandem MS spectrum of lipoylated peptide…… 97
Figure 5.1: Component organisms used in the simulation……... 107
Figure 5.2: Protein mass distributions for the background and *E. coli* proteins….. 109
Figure 5.3: Frequency matrix based on family profile information…… 114
Figure 5.4: Unique "signature protein" masses as comparing to the background set as a function of measurement error…… 118
Figure 5.5: Unique "signature peptide" masses as comparing to the background set as a function of measurement error…… 120
Figure 5.6: Protein detection with fragmentation efficiency $N = 10$……... 123
Figure 5.7: Specificity of protein detection……. 125

Figure 5.8: Dependency of correct detection on the number of fingerprint fragments……. 127
Figure 5.9: Peptide detection……... 129
Figure 5.10: Specificity of peptide detection……... 131
Figure 5.11: Comparison between top down and bottom up detection specificity……. 133
Figure 5.12: Specificity of organism detection with *OrganismScore*…….. 136

LIST OF SYMBOLS AND ABBREVIATIONS

ORF	Open reading frame
CID	Collision induced dissociation
CAD	Collision activated dissociation
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
m/z	Mass to charge ratio
PCR	Polymerase chain reaction
PMF	Peptide mass fingerprint
PTM	Post-translational modification
X-correlation	Cross-correlation
FT	Fourier Transform
ES	Electrospray
MALDI	Matrix assisted laser desorption ionization
DTT	Dithiothreitol
HPLC	High performance liquid chromatography
LC	Liquid chromatography
TOF	Time of flight
2-D PAGE	Two-dimensional polyacrylamide gel electrophoresis
ESI	Electrospray ionization
DOE	Department of Energy
ISB	Institute for Systems Biology
ORNL	Oak Ridge National Laboratory
SNP	Single nucleotide polymorphism

Chapter 1

Introduction to Proteomic Analysis with Mass Spectrometry and General Mass Spectrometry Data Analysis

Introduction

The proteome can be defined as the set of all expressed proteins in a cell, tissue or organism. Proteomics is an emerging new discipline in the field of studying living organisms being a direct continuation to the area of genomics; where as genomics is concerned with the study of genetic codes of living organisms, proteomics is devoted to a study of the organism's expressed protein content. The study of genes is applied to uncovering the secrets of life focusing on the genetic makeup, including the sequencing and study of DNA patterns in nature. The Central Dogma of biology states: DNA is transcribed into mRNA which in turn is translated to produce proteins. As opposed to the relatively static study of a genome, proteomics is very dynamic, the protein content of a living organism ever changing, starting with the programmed changes in the proteome that deal with time points in life cycle and ending with the adapting the organism to varying environmental conditions. The study of a proteome can address a wide range of conditions such as discovering traces of disease, measuring body's response to a medication or detecting a particular protein compound in an organism. The area of proteomics involves a wide range of studies dealing with proteins, including the study of protein structure and function, post-translational modifications (PTMs), protein-protein interactions, protein regulation and the study of complex protein networks.

A major area of proteomics deals with characterization of an organism's proteomic content. It is focused upon studying what proteins are present in an organism

under a particular condition. While the genome sequences supply the full DNA information of an organism, proteomic studies measure the dynamics of an organism, supplying a variety of information about an organism at a particular time. With proteomics it became possible to measure the dynamics of an organism. In the last decades great breakthroughs were achieved in the study of the genome, supplying us with the vast knowledge of the genes and a massive amount of sequenced organisms. As the genome information becomes more and more accessible, the need for further systems level studies became clear; transcriptome studies and then proteomic studies became prominent. Some of the major proteomic studies involve the studies of protein expression patterns (protein cataloguing), detecting protein-protein interactions, protein localization analysis, and analysis of post-translational modifications. One of the main techniques used for high-throughput proteomic studies is mass spectrometry. It is capable of detecting masses of biological compounds in complex mixtures and currently one of the most powerful methods for protein detection and analysis.

Mass spectrometry

Mass spectrometry (MS) is one of the most indispensable tools for high-throughput proteomic analysis (Aebersold, 2003) due to its versatility and speed. It is a fast and reliable tool capable of measuring masses of biological molecules in complex mixtures (Pandley, 2000). Fundamentally, a mass spectrometer is an instrument consisting of three parts: ionization source, ion analyzer and ion detector. The ionization source is responsible for desorption and ionization of biological molecules into gas phase, the ion analyzer separates them according to their mass to charge ratios (m/z) and ion detector detects and multiplies the ion signal. The two ionization sources most frequently

used in proteomic analysis are electrospray (ESI) and matrix assisted laser desorption ionization (MALDI) sources. The MALDI ionization technique involves transferring biological molecules into gas phase from a solid matrix (Hillenkamp, 1991). This ionization method produces low charge state molecules -- the ions measured by a mass analyzer have in most cases single and double more rarely triple charges. ESI transfers biological molecules directly from liquid matrix into gas phase (Fenn, 1989), producing mainly multiply charged molecules. The masses of biological molecules reflect their composition and are used for their detection and characterization.

In addition to measuring the masses of biological molecules, mass spectrometers are also capable of producing sequencing information in a form of fragmentation pattern. The fragmentation pattern of a peptide is produced by a process called tandem MS (or MS/MS) (Hunt, 1981; Biemann, 1986). The tandem MS experiment involves the following three steps: isolation of ions in a particular mass to charge ratio, fragmentation of these ions, and the detection of the resulting fragment ions. The fragmentation is performed by the process called collision induced dissociation (CID), which involves colliding peptides of isolated m/z with inert gas, the collisions inducing breaks in the peptide bonds, and resulting in a spectrum of mass to charge ratios for peptide fragments, which are the function of the peptide sequence. Using the parent m/z of the peptide and its fragmentation pattern, the identity of a biological molecule can be established through a variety of database search algorithms.

The two major mass spectrometric techniques frequently used for protein analysis are referred to as top down and bottom up methods. These two techniques approach proteome analysis from the different angles as is illustrated in the figure 1.1.

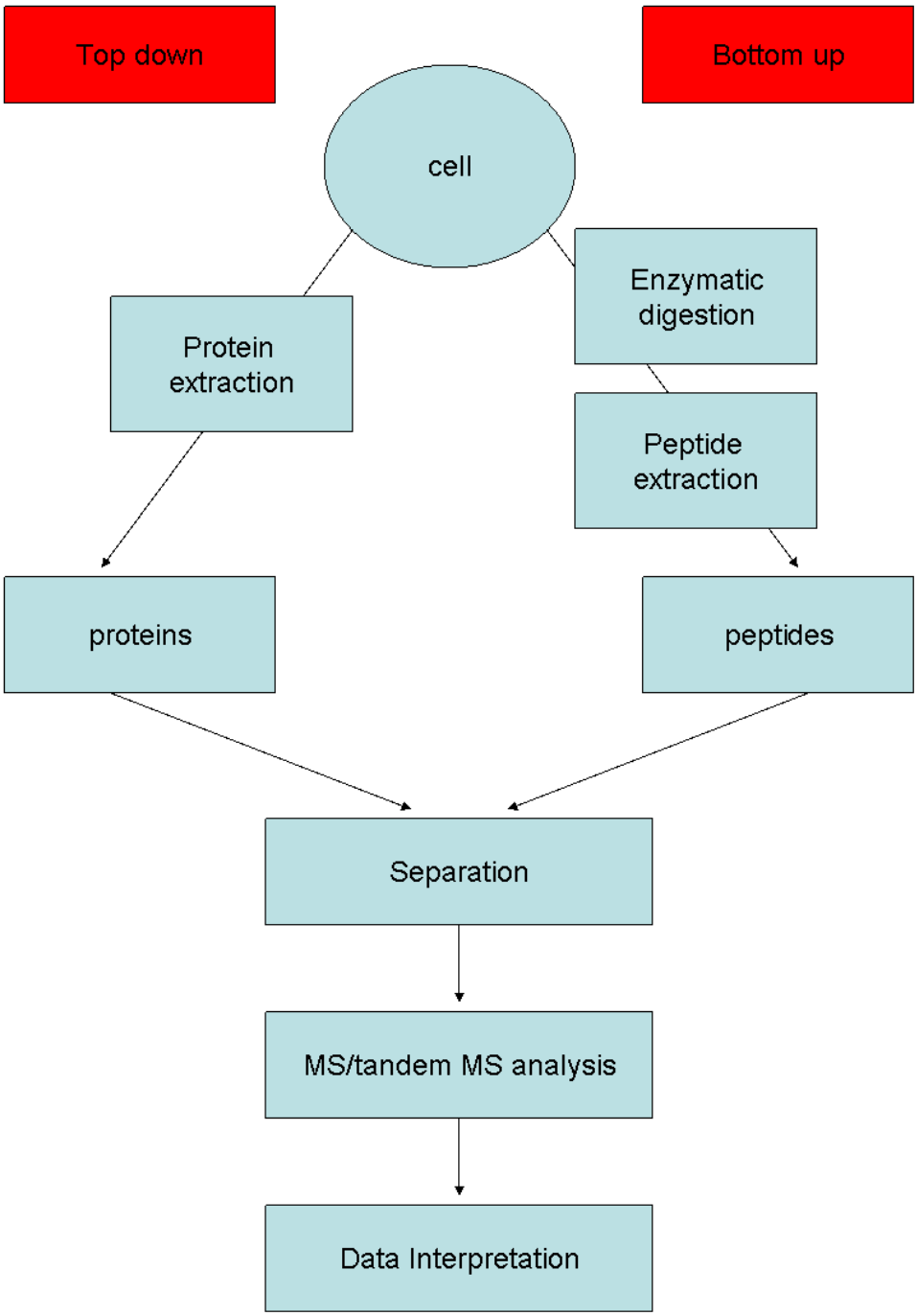

Figure 1.1 Top down vs. bottom up mass spectrometric techniques. Top down MS involves analyzing intact proteins, while the bottom up MS involves an extra step of digesting the proteins into smaller peptides, measuring and analyzing the peptides and then inferring the identities of the present proteins from the detected peptides.

The top down approach involves analyzing intact proteins which are extracted from the cell. The bottom up approach analyzes peptides, the fragments of proteins resulting from an enzymatic digestion, inferring the protein identities from the detected peptides. These techniques are often complementary to each other: while the bottom up approach produces more protein identifications, it does not supply the information about the state of an intact protein (PTMs, N-terminal protein processing, splice site variants, and amino-acid substitutions) and while top down experiment supplies the information about intact proteins, it generally produces less protein identifications due to the limitations of experimental technology and computational interpretations. Currently, the high-throughput proteomics tends towards bottom up techniques as the instrumental technology is more robust and better developed.

Top down proteomics is a technique involving protein characterization at the intact protein level by MS with possible following tandem MS analysis. Before proteins are analyzed by MS, they are separated into smaller fractions. This separation process of intact proteins is one of the most difficult challenges of current top down proteomics technologies. One of the most popular techniques for intact protein separation is two dimensional polyacrylamide gel electrophoresis (2-D PAGE gel), (O'Farell, 1975) separating proteins by two criteria: isoelectric point in the first dimension and molecular weight in the second dimension. While powerful, 2-D PAGE gel separation remains difficult and slow process impeding the speed of intact protein analysis. Furthermore, the extraction and subsequent MS analyses of intact proteins from 2-D PAGE gels is nearly impossible. The MS analysis of intact proteins is generally performed using electrospray ionization Fourier transform ion cyclotron resonance mass spectrometers (ES-FT-ICR-

MS) (Mortz, 1996; Kelleher, 1998). The instruments of this type have both high degree of resolution, and mass accuracy up to 10^{-4} Da, both of which parameters play a significant role in protein characterization. In addition, top down proteomic method can also employ the tandem MS (or MS/MS analysis), providing a limited sequence information for the measured protein. Top down analysis provides the information about the mature proteins expressed in the organism, including the post-translational modifications and amino acid substitutions. However, the interpretation of intact protein analysis is frequently complicated by the absence of the exact corresponding protein sequences in the database (PTMs, SNPs, etc.), while at experimental end separations with liquid chromatography and measuring large proteins by MS can often be very difficult. Top down mass spectrometry in general is not yet applied for high-throughput proteomic studies. While other methodologies are introduced in the dissertation, the work is focused upon the data interpretation for the bottom up "shotgun" proteomics introduced below and unless otherwise stated all of the references to MS and data analysis correspond to these experiments.

There are many avenues for bottom up proteomics which involve a multitude of techniques. The two most popular of the bottom up techniques are "shotgun" bottom up, (which will later be referred to simply as bottom up) and peptide mass fingerprint (PMF) analysis. The main technique for high-throughput protein identification using tandem mass spectrometry is called bottom up or "shotgun" approach. A typical complex mixture bottom up experiment involves a protein mixture digestion with an enzyme protease (such as trypsin, pepsin, glu-C, etc), a separation of the complex mixture into smaller fractions by such techniques by gel separations (Hess, 1993; Gatlin, 1998) or

liquid chromatography (McCormack, 1997; Martin, 2000; Shen, 2001) followed by MS analysis. The bulk of bottom up proteomic analysis is performed by ES ionization sources coupled with ion trap mass analyzers. This instrument is fast, reliable and affordable however, it has a limited resolution of approximately 0.5 Da. The MS analysis yields information about the mass to charge ratio of the examined peptide while MS/MS analysis provides information about the peptide amino acid sequence.

Another popular method for characterizations of complex mixtures is peptide mass fingerprinting (PMF), which is a different approach to bottom up experiment. Like the previously described "shotgun" bottom up technique, PMF involves digesting the proteins into peptides and measuring their masses with mass spectrometry. The key differences between PMF and shotgun bottom up technique is that PMF involves protein separation (generally 2-D PAGE) and that tandem MS experiment is not performed. Peptide masses are generally analyzed with MALDI TOF instruments. The protein identification is made based upon identifying it's peptides by their masses, the reliability of protein identification dependant on the detected sequence coverage and pattern of detected peptide masses.

Fragmentation pattern

Tandem MS or MS/MS peptide/protein fragmentation creates a sequence dependent fragmentation pattern. The peptides tend to fragment by breaking along peptide backbone bonds, each break creating a pair of fragment ions (Figure 1.2). The fragment ions retaining N-terminus of a peptide are referred to as 'a', 'b', 'c' ions, while the ions containing C-terminus are referred to as 'x', 'y', 'z' ions (Roepstorff, 1984; Biemann, 1988).

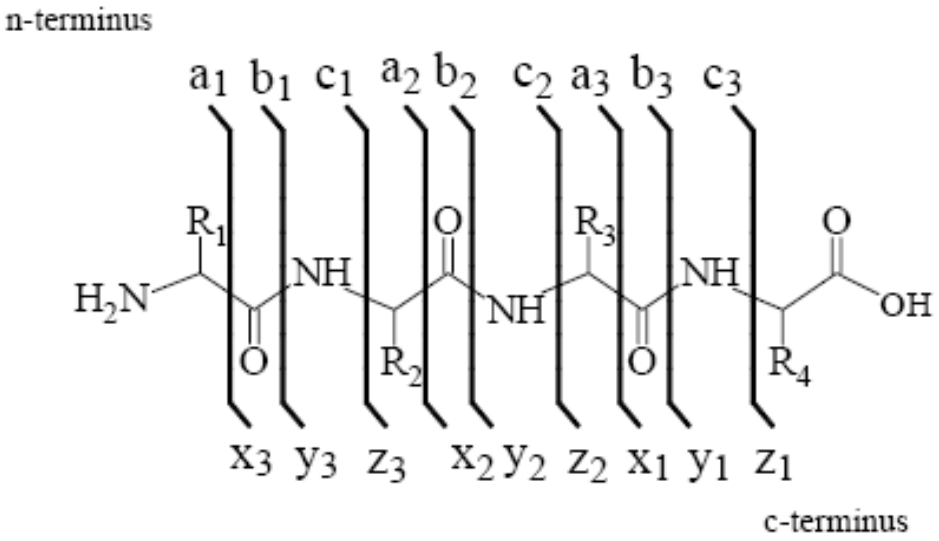

Figure 1.2 Tandem MS fragments. The representation of possible backbone fragmentations. The ions containing N-terminus are referred to as 'a', 'b', 'c' ions, while the ions containing C-terminus are referred to as 'x', 'y', 'z' ions.

In the case of low energy CID fragmentation, the 'b' and 'y' ions are the major ions in the spectrum, all of the other types of ions shown in the figure 1.2 can be present, but at a significantly lower abundance. In theory, the number of 'b' and 'y' ions for each given peptide is roughly equal to the number of the peptide bonds, if the number of amino acids in the peptide is equal to N, then the peptide will have a maximum of N-1 'b' and N-1 'y' ions (Figure 1.3). Each ion appears in the tandem MS spectrum in a form of a peak with an m/z and intensity. Intensity of a peak is related to the abundance of the ions of this m/z (Figure 1.3).

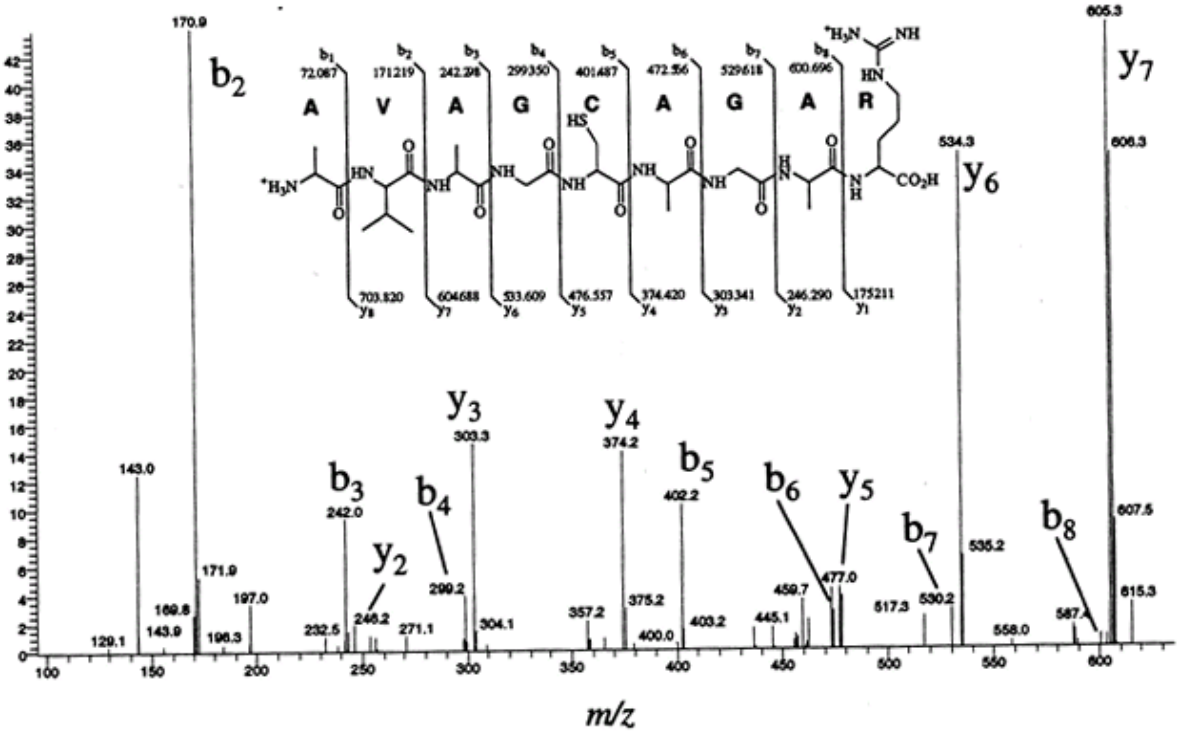

Figure 1.3 Sequence dependent fragmentation pattern produced by CID. The horizontal axis is representing the m/z ratios of the fragment ions and vertical axis is representing the intensities of the fragment ions. The major ions in the CID spectrum are the 'b' and 'y' ions. As shown, there is no easy differentiation between the 'b' and 'y' series in the ion trap tandem MS spectrum.

As a result of CID fragmentation the expected outcome is a pair of 'b' and corresponding 'y' ions at each amino acid position, producing so called ion series. The m/z difference between two consecutive 'b' (and 'y') ions of a single charge (equivalent to a single amino acid mass) can be used to infer the identity of an amino acid between them as shown in figure 1.3, from "Introduction to Proteomics" by Dan Liebler (Page 95, Figure 4). This property is extensively used by many of the peptide identification algorithms to infer the sequence information from the tandem MS spectrum.

Mass spectrometry is used to detect and identify protein content of complex biological mixtures. Each of the outlined approaches, top down, bottom up and PMF is applied to the proteomic studies. The majority of high-throughput proteomic analysis, as previously stated, is performed by the bottom up mass spectrometry. A significant portion of bottom up analysis is performed with the use of ES ion trap instruments. The instrumentation and procedures are well characterized and developed, making proteome analysis a comparably simple and routine task. Bottom up mass spectrometry can be used for protein detection in complex mixtures; however, the identity of a protein in bottom up analysis must be inferred from its detected peptides. Since not all of the peptides are generally analyzed and interpreted, only partial sequence coverage is possible and the protein in the complete form is generally not detected. In some cases bottom up MS can be used to detect post-translational modifications if they are present in the detected peptides. However, the analysis of post-translational modifications using the current data interpretation methods is not straightforward. Some of the major problems with PTM detection are: a) difficulty with ionizing modified peptides, as modifications frequently change peptide properties, b) modifications can cause complications in tandem MS spectra such as reducing peptide fragmentation, c) there are no rigorous computer algorithms for identification of modified peptides. Chapter 4 of this dissertation addresses this very important and interesting task of high-throughput PTM detection by bottom up mass spectrometry.

The top down approach is used to detect intact proteins as they appear in the sample. As opposed to the bottom up technique, the entire protein is characterized by top down mass spectrometry, including the possible post-translational modifications and

sequence mutations. As is explored in Chapter 5 of the dissertation, it can be applied to detecting a particular protein or even an organism in a biological sample. In addition, top down approach can be expected to be less computationally intensive as there is significantly smaller number of intact proteins that can be measured with mass spectrometry than the peptides. However, the top down approach is more difficult experimentally, the difficulties laying in the protein separation, gas phase conversion and performing MS/MS on intact proteins.

Mass spectrometric data analysis

While mass spectrometers produce meaningful information about masses of biological molecules, computational data interpretation is necessary to discover the identity of the measured biological molecules. In order to interpret mass spectrometric data, a multitude of computation methods have been developed. There are three basic approaches to bottom up MS data interpretation: database search algorithms, *de novo* sequencing algorithms and hybrid algorithms. The database search algorithms are currently the main tools for bottom up MS data interpretation, the *de novo* sequencing and hybrid algorithms are relied upon in cases when the database is not available or the database searches do not produce adequate results.

Database search algorithms

The most commonly used methods for bottom up mass spectrometry data interpretation are the database searches. Database searches are robust, reliable, sensitive and fast approaches for peptide identification. All of the database search algorithms have an inherent similarity: they operate based on the sequence database generally specific to the organism of interest. Database search algorithms rely on the comparison between the

theoretical fragmentation patterns of peptides derived from the database and the experimental peptide fragmentation pattern. Such tools produce a list of possible peptide assignments and theoretical fragmentation patterns for each experimentally measured tandem MS fragmentation pattern based on m/z of the parent peptide. The theoretical fragmentation patterns are scored against the experimental tandem spectrum and the theoretical peptide that displays the highest similarity to the experimental measurements is accepted as the best candidate. There are basic two assumptions that database search algorithms make that must be met in order for the identification to be successful. The first assumption is that the peptide represented by the tandem MS spectrum is present in the database in exactly the same form as it is in the sample. The second assumption is that if the peptide which gave rise to the tandem MS spectrum has been found in the database, its theoretical spectrum is more similar to the tandem MS spectrum than that of any other peptide in the database. There is a multitude of database search algorithms for MS data interpretation such as MASCOT (Perkins, 1999), SEQUEST (Eng, 1994), DBDigger (Tabb, 2005), Sonar (Field, 2002), ProteinProspector (Clauser, 1999), and OMSSA (Geer, 2004). While the software design of the algorithms varies to improve speed and flexibility, the main differences in performance are dependent on two factors: the selection of candidate peptides and the scoring scheme used for spectral comparison. The selection of candidate peptides for database search algorithms is generally based on the mass window of the measured precursor peptide. The scoring schemes can vary between spectral comparisons, which take into account the similarity of two spectra, and complex probabilistic approaches, which attempt to assess the probability of match by incorporating the whole database into the comparison. One of the oldest and most

established database search program is SEQUEST (Eng, 1994), developed in 1994. It involves an adaptation of a cross correlation scoring scheme for experimental and theoretical spectra comparison. While sensitive, the X-correlation scorer does not provide information on the identification probability. Chapter 2 of this dissertation addresses the need for a reliability scheme for SEQUEST scoring for statistically sound peptide and protein identification.

***De Novo* algorithms**

De Novo algorithms have always been the "silver bullet" of mass spectrometry data interpretation, in theory abolishing the need for databases by deriving sequence information directly from the tandem MS spectra. When sequence databases are not available or there is an inconsistency between the database and the protein of interest people often resort to "sequencing by hand". The term "sequencing" in application to tandem MS data means reading sequence off the spectrum and in the majority of cases produces either a partial (several amino acids long) or a full sequence of the peptide in question. *De Novo* is the name that people apply to the algorithms that use tandem MS spectrum to derive the full length sequence (or it's majority), using the same concepts as do humans while performing "sequencing" by hand. Unfortunately at this time, this promising and inspiring approach does not demonstrate any degree of robustness and sensitivity that has been demonstrated by generally much more simplistic database search algorithms for the low resolution ion trap data and while it is promising technique for higher resolution instrumentation such instruments are not currently used for shotgun proteomic experiments. In addition, *de novo* approaches are generally very slow, taking significantly more time than the database search algorithms, while producing a

significantly greater number of false positives. The difference between *de novo* and database search approaches is such that their performance is considered not comparable. *De novo* approaches are more likely to be used when a tandem spectrum of a good quality cannot be identified by a database search, and then the likelihood of success is often small. Similarly to the database search, *de novo* methods generally provide an answer, however, even though a partial sequence might be found correctly; to find the fully correct sequence or even predict which part of the sequence is correct is very hard. Some of examples of well known *de novo* programs are PEAKs (Ma, 2003), Lutefisk (Taylor, 2001), and Sherenga (Dancik, 1999).

Hybrid approach

Hybrid data interpretation is a more flexible method than the database search. It is based on performing the database search based on short amino acid tag (peptide tag), rather than on the mass of the peptide which allows for a better PTM search as PTMs affect the mass of measured peptide, preventing the mass filtering from finding correct candidate peptide from the database. There are several approaches to finding a peptide tag, for example, Mann and Wilm in 1994 introduced a reasonably successful *Peptide Sequence Tag* approach (Mann, 1994). The other representatives of the hybrid approach are GutenTag, developed by David Tabb (Tabb, 2003), and one of the most recent hybrid approach MultiTag (Sunyaev, 2003). The basic flexible hybrid approach is based on finding a continuous short amino acid sequence tag from a tandem MS, searching the database based on the tag and masses flanking the tag, scoring the database peptide sequences against the original tandem MS, and reporting the best fitting candidate sequence. This type of approach might show an improvement over simple peptide mass

database search, being less sensitive to one or two amino acid substitutions in the peptide sequence, since it is possible to search the database by degenerative tags and either or none of the flanking masses. However, it is difficult to rigorously assess the reliability of an amino acid tag. In addition, many of the short amino acid tags match the database, the possibility of degenerate tag and flanking masses creates a large number of false positives as well as false negatives.

Overview of the dissertation

The goal of this dissertation is to address some of the problems in mass spectrometric data interpretation. As the field of mass spectrometry expands, the experiments are becoming more complex and demand better data interpretation. The majority of algorithmic tools currently widely used for the data interpretation are a few years old. While being well tested and robust, they often lack the flexibility and accuracy required to deal with the increasing demands of the field. It can therefore be expected that with this rising demand new data interpretation tools will be continuously developed in the future years. This dissertation addresses a few of the concerns of mass spectrometric data interpretation for bottom up shotgun high-throughput proteomics. Some of the common problems in bottom up MS data involve assigning reliability to the peptides and proteins identified by SEQUEST, one of the accepted database search algorithms up to date; determination of peptide mass based on mass to charge ratio (m/z deconvolution); and the detection of post-translational modifications. These three problems of MS data interpretation comprise three chapters of this dissertation. Mass spectrometry is becoming a promising tool for organism detection based on its proteomic content which can be an invaluable resource for detection of biological agents in the

environment. The fifth chapter of the dissertation presents a series of computer simulations made to probe the capabilities of mass spectrometry as biological agent detector.

The second chapter of the dissertation is focused on designing a new scheme to assess the reliability of peptide and protein detection made by SEQUEST. SEQUEST is one of the most popular database search tools used for bottom up MS data interpretation, which currently employs simple filtering procedures for peptide and protein detection without assessing the likelihood of the correct identification. The results of SEQUEST data interpretation are sorted and ranked by its scoring scheme and the peptide identification receiving the highest score is considered to be the correct answer. The people using the software are then free to accept or reject any of the identifications made based on the score cutoffs accepted in their laboratory. However, the score is frequently influenced by such factors as peptide length, peptide charge, spectral quality and even digest procedures. The lack of accepted reliability scheme causes difficulties in accurate peptide detection as well as making it difficult to assess the likelihood associated with protein identification. The presented work suggests a new neural network based approach to assess the reliability of peptide identifications by SEQUEST and, using the new strategy, proposes a scheme for reliability of protein identifications.

A new algorithmic approach for charge state deconvolution for low resolution mass spectrometry is presented in the third chapter of the dissertation. While other methodologies are emerging, the tools employed as driving force of proteomic efforts today are electrospray ion-trap instruments. While they are robust, sensitive, fast and easily coupled to liquid chromatography separations, they are also low-resolution

techniques. Electrospray ionization technique produces multiple charged peptides and MS analyzers measure m/z ratio. The low resolution of ES ion-trap instruments does not allow to definitively recognize the charge state of the peptide ion measured by MS which leads to an ambiguity in the peptide mass. The presented work explores the fragmentation pattern for charge specific characteristics and employs a trained artificial neural network to differentiate between the multiply charged spectra. This methodology both reduces the number of spectra by eradicating precursor mass ambiguities and improves the performance of the analysis by potentially decreasing the number of incorrect identifications. Additionally, in the chapter, a new performance evaluation method is introduced and used to evaluate the presented charge state determination method.

In the fourth chapter of the dissertation, an organism-specific detection of post-translational modifications by bottom up proteomic analysis is addressed. A new PTM driven database search algorithm is introduced and tested on several growth conditions of the metabolically versatile prokaryotic organism *Rhodopseudomonas palustris*. The detection of post-translational modifications is an extremely difficult and important problem. Mass spectrometry is uniquely qualified for high throughput protein modification detection due to the changes in the protein mass. However, interpretation of mass spectrometric data for post-translationally modified proteins is especially difficult due to the amount of false identifications and explosive database sizes. A detection of biologically sound post-translational modifications in an organism of interest was approached by building an organism specific post-translational modifications annotated homology based database and developing a set of criteria for a reliable identification.

The database was built to include only the experimentally observed post-translational modifications, greatly reducing the number of false positive identifications, database size and time of analysis. The method for reliable detection of post-translational modifications included a combination of a database search approach combined with a set of rules increasing the reliability of detection and reducing the number of incorrect identifications.

The fifth chapter of the dissertation is devoted to a set of computational simulations for biological agent detection via mass spectrometry. With the use of mass spectrometers, proteins present in complex mixtures can be analyzed and identified. This concept can be extended to proteome based organism detection in a complex multi-organism mixture. Organism detection in a complex background can be useful in detecting harmful organisms in the environment, water supplies and food sources and is already considered for use in the detection of biological weapons. However, there are many limitations to the current technology which must be overcome before the instruments and data analysis algorithms are capable of undertaking high-throughput organism detection. The simulations presented in Chapter 5 of this dissertation are designed to examine the pros and cons of top down and bottom up MS detection techniques and explore the instrumental parameters needed for detection of an organism in a complex environmental sample.

Chapter 2

A Computational Method for Assessing Peptide-Identification Reliability in Tandem Mass Spectrometry Analysis with SEQUEST

Some of the text presented below has been published as Razumovskaya J., Olman V., Xu D., Uberbacher E., Nathan Verberkmoes, Hettich R.L., Xu Y., *Proteomics*, 2004., Apr; 4(4):961-9.

Introduction

One of the most important goals in systems biology is to identify and characterize the protein composition of cells as a function of conditions (Pandley, 2000). Mass spectrometry has become a fast and reliable tool for determining the protein composition of complex mixtures by measuring mass to charge ratios of proteins and peptides in mixtures. A very popular technique for whole proteome characterization is bottom up shotgun mass spectrometry.

Many tools have been developed for high-throughput peptide identifications for bottom up shotgun mass spectrometry such as SEQUEST (Eng, 1994; Yates, 1995; Yates, 1995*), MASCOT (Perkins, 1999) SONAR (Field, 2002) and others. Most of the current applications are database search based: they rely on the comparison between theoretical peptides derived from the database and experimental mass spectrometric tandem spectra. The database theoretical peptides are scored against the experimental tandem spectrum and the theoretical peptide that displays the highest similarity to a corresponding experimental spectrum, according to accepted scoring scheme is considered to be the best hit. SEQUEST is one of earliest developed and still a very popular tool for peptide identification. From a tandem mass spectrometry experiment, SEQUEST produces a list of possible peptide assignments in a protein mixture. For each

candidate peptide, it assigns scores, including the X-correlation score, the final score produced by SEQUEST, the charge state of the peptide and several others. The peptide identification process takes place after SEQUEST produces peptide identifications and involves a number of filtering steps based on the aforementioned scores. The two main SEQUEST scores used for filtering are the X-correlation score and the number of charges, though other scores can also be used. SEQUEST peptide hits are generally ranked based on their X-correlation scores. One problem with the current SEQUEST scoring scheme is that a SEQUEST score, say the X-correlation score = 2.5, may have different meanings for different peptides with different lengths and charges, making it difficult to interpret the SEQUEST identification results automatically. For some annotated data sets (Keller, 2002), the distributions of X-correlation score along with the charge states for the correct and incorrect hits do not have a clear separation.

A possible solution to the problem is to develop a scheme to estimate the identification reliability for each SEQUEST hit, based on the SEQUEST scores. The preliminary analyses of SEQUEST search results have suggested that it is possible to achieve this by combining different SEQUEST scores.

There have been several attempts to separate correct SEQUEST assignments from the incorrect ones (Yates, 1995). Recently, a statistical approach was reported to assign reliabilities to peptide hits using a database consisting of 18 protein sequences -- these are *Drosophila* proteins with possible human contaminants (Keller, 2002). The score distributions for the correct and incorrect peptide assignments were used to create a statistical model from which the probabilities of correct and incorrect assignments were derived. This method was designed to filter out a large number of database search results

with predictable false identification error rates. The probability distributions were represented as normal and gamma distributions. This approach relies on fitting the experimental data to these distributions without theoretical justification for the phenomena which might not be reflected under different conditions. The method is focused on a specific experimental design involving tryptic digest coupled with non-specific digest SEQUEST data analysis, as one of the given parameters is NTT (number of tryptic termini), which although might insure an improvement under this specific experiment, is not applicable to other type of experiments.

In another recent study, a support vector machine (SVM) technique was applied to separate correct SEQUEST identifications from the incorrect ones (Keller, 2002). SVM is a binary classifier that learns to distinguish between correctly and incorrectly identified peptides by using a vector of parameters describing each peptide identification. This method improved upon the simple cutoff approach, currently used in SEQUEST, for separating the correct and the incorrect peptide identifications, but it does not provide an estimate of the reliability of each identified peptide.

In this chapter, a new scheme is described for assessing reliabilities of peptide identifications made by SEQUEST. In the scheme, peptide scores are normalized and their probabilities to be correct are statistically estimated. These peptides and assigned probabilities are then used to provide a statistical assessment for protein identification. This method is based on a combined application of a statistical decision-making procedure and a neural network. The training of the neural network was accomplished using a set of tryptic peptides from known proteins measured by mass spectrometry and analyzed by SEQUEST. The SEQUEST results were separated into correct and incorrect

identifications by careful manual analysis. Once trained, the neural network provides a score between 0.0-1.0 for each peptide, reflecting the probability of a peptide to be the correct identification.

One advantage of this approach is that it provides improved resolution of assignments for peptides that SEQUEST scores in the "gray area". In this current approach, each peptide hit has a particular level of confidence associated with its SEQUEST score. This confidence value can then be used in conjunction with other parameters to assign a reliability estimate for protein identification, which typically corresponds to a number of peptides identified in a protein.

The trained neural network was evaluated on two sets of data, one representing a relatively simple mixture of proteins, and one complex mixture. The evaluation was based on two accepted parameters often used in method comparisons: sensitivity and specificity. Sensitivity is the ratio between the number of correctly predicted hits and the number of all correct peptide assignments, while the specificity is defined as the ratio between the number of correctly predicted hits and the total number of hits. The formula for sensitivity is described as

$$\frac{TP}{TP + FN} \qquad [1]$$

and specificity is

$$\frac{TP}{TP + FP} \qquad [2]$$

where TP denotes the number of true positives, FN denotes the number of false negatives, and FP, the number of false positives. The test results showed a significant improvement

in both the identification sensitivity and specificity for peptides by this method, compared to the standard SEQUEST filtering procedure. Based on this peptide-identification neural network, we have developed a statistical model for protein identification, through combining peptide-identification reliability estimates. To evaluate this approach, comparisons were performed on our statistical model and one current filtering procedure of SEQUEST, called DTASelect (Tabb, 2002). This method yielded a significantly larger set of protein identifications than the filtered DTASelect, showing 20% improvement in sensitivity over the filtered DTASelect in the first 70 ranked protein identifications (with the same specificity). These results demonstrate that the combined neural network method and statistical model is more sensitive than filtered DTASelect while maintaining the same specificity, and more specific than the results of DTASelect without application of filters (later to be referred to as unfiltered DTASelect), which was applied for higher sensitivity of protein identifications, while maintaining the same sensitivity.

Materials and Methods

Data set for neural network training

An 18 protein sequence dataset was used as the training data set for the neural network. This set was obtained from a mixture analyzed by the Institute for Systems Biology (ISB) (Keller, 2002*). The 18 purified proteins, placed in the mixture and digested with trypsin, were: bovine β-casein, bovine carbonic anhydrase, bovine cytochrome c, bovine β-lactoglobulin, bovine α-lactalbumin, bovine serum albumin, chick ovalbumin, bovine transferrin, rabbit GAPDH, rabbit phosphorylase b, *E. coli* β-

galactosidase, bovine γ-actin, bovine catalase, rabbit myosin, *E. coli* alkaline phosphatase, horse myoglobin, *B. lichenformis* α-amylase, and *S. cerevisiae* phosphomannose isomerase. Keller et al. performed an analysis, using SEQUEST, to identify the peptides. The search was performed against the human protein database plus these eighteen protein sequences. The assignments of spectra to peptides were confirmed through thorough manual examination. Keller et al. (Keller, 2002*) did peptide assignment in the following way: if a peptide did not belong to the set of expected proteins, its assignment was considered incorrect; otherwise they were manually examined before being considered as correct. Detailed manual analysis also revealed that this set contains additional proteins from human contamination. Hence the set actually consists of 29 proteins.

The final list of correct assignments consists of 2,784 peptides, confidently identified in the mixture; and the list of incorrect assignments contains 34,287 peptide hits by SEQUEST (Table 2.1). The incorrect assignments could be due to limitations of SEQUEST interpretation, or to the presence of bad spectra. These data were used to train our neural network. Using this dataset, the sensitivity using a "normal" X-correlation cutoff (1.8 for charge state 1, 2.5 for charge state 2, and 3.5 for charge state 3) was found to be 66%, with a specificity of 89%, while the sensitivity using a "minimal" cutoff (1.5 for charge state 1, 2.3 for charge state 2, and 3 for charge state 3) was found to be 75%, with a specificity of 84%, in comparison to the neural network approach, which at specificity of 89% has sensitivity of 89%.

Table 2.1. Peptide assignment results for the ISB protein mixture.

Charge state	Correct	Incorrect
+1	125	379
+2	1649	16856
+3	1010	17052

Assignments used in the neural network training. The first column indicates the charge state. The second column shows the number of the correct peptide assignments found by SEQUEST through database search and confirmed by manual interpretation. The third column represents the incorrect peptide assignments.

Neural network selection and training

The first goal with neural network application is to identify which peptide-assignments by SEQUEST are correct and which ones are incorrect, through applications of other parameters in addition to the X-correlation scores and the charge states.

The preliminary studies have suggested the following six parameters should be useful in helping to achieve this goal: the SEQUEST X-correlation score (measure of likelihood of an experimental spectrum to be a representation of a theoretical peptide), peptide charge state (1, 2 or 3), ΔCn (ΔCn - the difference between X-correlations of the top and the second top hits in the SEQUEST output for a particular experimental spectrum), SpRank (rank of the peptide in the preliminary scoring), ion coverage (percent of matched peaks), and the length of the peptide. Several other scores were also evaluated in the preliminary studies, including dM (mass difference between theoretical and experimental parent ions) and Sp (SEQUEST preliminary score). However, it was found that these do not improve the performance of our neural networks. Each of the training data (2784 correct ones and 34287 incorrect ones) has six parameters associated with it and a 0/1 (for “incorrect” and “correct”, respectively) label as the desired output value.

70% of the data points from this data set were randomly selected as the training data and the remaining 30% of the data as the testing set. The architecture of our neural network has six input nodes corresponding to the selected input parameters, one hidden layer, and one output node corresponding to the result of the neural net. Nodes of adjacent layers are fully connected. Throughout the training process various neural network architectures with different numbers of hidden nodes were evaluated.

A SNNS 4.2 (Stuttgart Neural Network Simulator) was used to train a neural network to distinguish correct from incorrect peptide identifications. We used the back-propagation learning algorithm to train the connection weights. Performance results are saved every 60 cycles throughout the training process. The training stops when no improvement in the error rate could be achieved. Each of the resulting nets was then tested for performance and the best was selected based on resulting sensitivity and specificity. Neural network testing was performed using a jackknife approach where 30% of the ISB data was held out for testing, while the rest was used for training. The resulting neural network has a hidden layer with four hidden nodes. The neural network output ranges from 0.0 to 1.0, where the output represents the network's estimate of assignment correctness. Each result of SEQUEST, its associated neural network score, and its a priori classification can be used to generate histograms of the network's performance. An example histogram is shown in the figure 2.1.

Additional dataset for testing

A mixture of eight proteins was prepared as an additional test set: bovine hemoglobin alpha chain (2mg), bovine hemoglobin beta chain (2mg), bovine carbonic

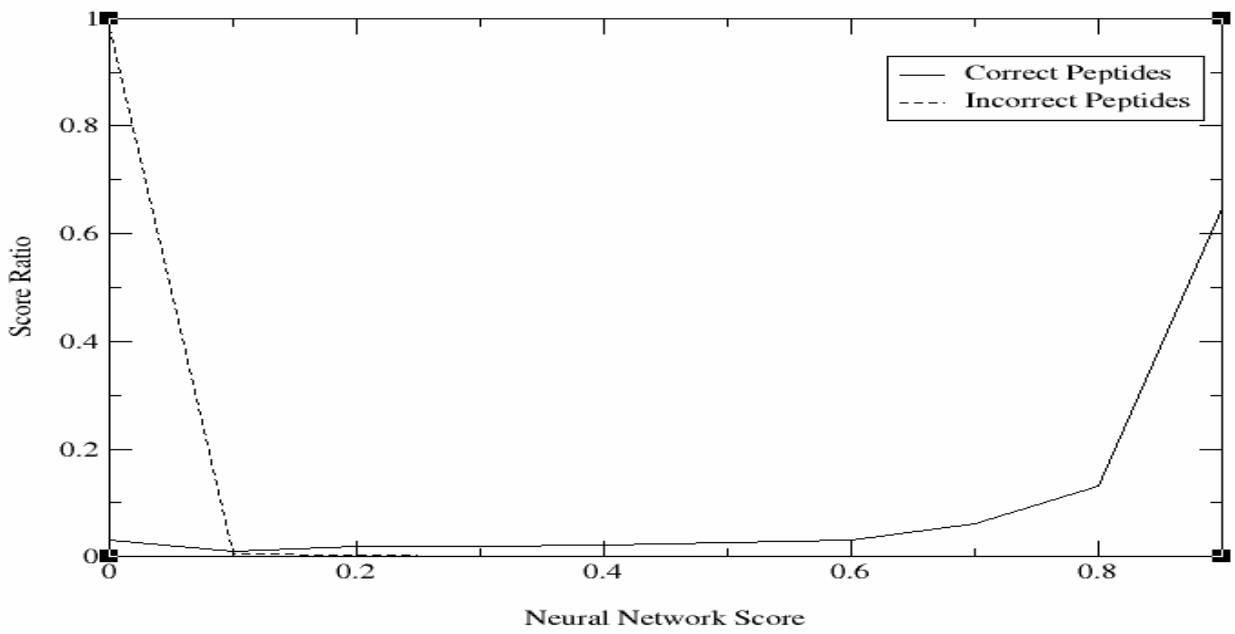

Figure 2.1 Histogram of correct and incorrect peptide assignments made by neural network. The majority of incorrect peptide assignments are at a range of 0-0.3 while the majority of correct peptide assignments have scores that range between 0.75-1.0. This histogram can be used to estimate the probability that a peptide with a given neural network score is correctly assigned. The neural network output ranges from 0.0 to 1.0, where the output represents the network's estimate that each assignment is correct. Each result of SEQUEST, its associated neural network score, and its *a priori* classification can be used to generate histograms of the network's performance.

anhydrase (2mg), horse myoglobin (1mg), bovine albumin (5mg), yeast alcohol dehydrogenase chain I (3mg), yeast alcohol dehydrogenase chain II (3mg), and chicken lysozyme (1mg). In order to denature the protein mixture, it was dissolved in 1ml of 6M guanidine and 5mM DTT and heated at 60C for 1hour. Then 2ml of solution was diluted to 10ml in Tris buffer and digested overnight with 1:50 aliquot of trypsin (60ug of trypsin). The sample was treated with 10mM DTT for 1 hr at 60C, desalted with a SepPak, completely dried, re-dissolved in 1ml of HPLC Buffer A and filtered resulting in sample concentration of ~2mg/ml. This sample was then analyzed with ion trap instruments (Thermo Finnigan LCQ-Deca). The sample was separated into three equal parts: Sample I was analyzed using LC-MS/MS with a 30ul injection and the short LCMS run. Sample II was concentrated to 10mg/ml and run with a 50ul injection with long LC-MS run. Sample III was concentrated to 10mg/ml and run with 50ul injection with short LC-MS run. The three samples were created in order to check the performances with different concentration related detection levels.

The trained neural network was tested on the mass spectrometry data collected under different conditions as defined in Samples I, II and III, respectively. In order to assign peptides a *Shewanella* database was used in addition with the eight proteins present in the mixture. The *Shewanella* database was selected for this task as there were little sequence similarity between the organism's proteins and the proteins present in the sample. Thus, it was considered to be a good background for SEQUEST search. SEQUEST (TurboSEQUEST v. 27) was used to search the database, using the default parameters. 2980 assignments were made for Sample I; 2163 assignments for Sample II; and 1186 assignments for Sample III. All the peptides identified by SEQUEST as

belonging to the eight proteins present in the protein mixture were considered correct. Peptides that correspond to both *Shewanella* proteins and the proteins from the test mixture were taken out of the analysis. Remaining peptides attributed to proteins coming from *Shewanella* were, therefore, considered to be incorrectly assigned. Results for SEQUEST on this dataset are summarized in Table 2.2.

Probability model for peptide identification

Using the result of neural network training, P, the conditional probability of true peptide identification given neural network score can be estimated using Bayesian formula as follows. A probability of a peptide being correctly assigned, given a particular neural network *Score*, is the ratio of the previously observed number of correct peptide identifications with that given *Score* to the total number of peptides with that *Score*. Formally, let P(C | Score) and P(I | Score) as the probabilities for a assignment to be correct (C) and incorrect (I), respectively. P is estimated as P(C | Score) as frequency(C , Score)/frequency(Score), where frequency(E) is a frequency of event E from histogram of neural network scores. Similarly, we estimate P(I | Score) as frequency(I , Score)/frequency(Score).

Protein identification

The above discussion is about identification of a peptide. Identification of a protein, from the mass spectrometry data, is based on the identification of peptides that come from the protein. In order to quantify the accuracy of protein assignment we have assessed the likelihood of a false protein to be identified by chance. For estimating the reliability of proteins in the mixture, the identified peptides are treated as *independent* observations (of a potential protein) within the mixture. Due to the assumption of peptide

Table 2.2 Peak assignment results for Sample I, Sample II and Sample III made by SEQUEST.

Sample set	Correct	Incorrect
Sample I	**487**	**2493**
Sample II	**345**	**1818**
Sample III	**245**	**941**

The first column indicates the analyzed sample. The second column shows the correct peptide assignments made by SEQUEST. The third column represents the incorrect peptide assignments made by SEQUEST.

assignment independence, the probability of a false protein assignment can be calculated by combining the probabilities of incorrect identification of its peptides as follows: Let peptides $a_1, a_2 ... a_n$ be a complete set of peptides that belong to protein A (resulted from the mixture analysis), the probability for a_i to be a true hit is defined as p_i, and m_i is a number of proteins that were found to contain peptide a_i. Then the probability that protein A is not in the mixture is estimated by value:

$$\text{PScore} = \prod (1 - \frac{1}{m_i} p_i \mid a_i \in A) \qquad [3]$$

The result shows the likelihood of a protein being identified by chance, given the collection of peptides that belong to the protein. The smaller this PScore, the less is the chance that the protein identification occurred by chance and therefore, greater the certainty that the protein is actually present in the mixture. Another method that utilizes a similar protein reliability model is described by MacCoss et al (MacCoss, 2002), however, with a very different peptide probability estimation (further discussed in Results section).

In this method, if multiple peptide assignments in the mixture correspond to the same protein, the likelihood that that protein is present in the mixture increases. The difference between this method and simple addition of number of peptides per protein, as it has been done by DTASelect (where every accepted peptide contributes the same amount) which also uses SEQUEST's peptide identifications is that the contribution of peptides to the final protein likelihood is based on their probability of correct identification. In the current scheme, unreliable peptide hits contribute less than reliable ones, but as the number of hits per protein increases, so does the likelihood of its presence in the mixture. In addition to the assumption of independence of peptide observations, it is assumed that non-unique peptides have equal chance to be produced by any parent protein.

Results and Discussion

Peptide identification

We present the peptide identification results on the two test sets: the ISB set and our own eight protein set, and compare these results with the simple SEQUEST identification results, using both the normal and minimal cutoffs. We also provide a performance comparison between our program and PepProphet, a software program written by Keller et al. To facilitate the comparison, the identification sensitivity and specificity were calculated for each range of the neural network score, between 0.0 and 1.0 with a 0.01 increment. The sensitivity and specificity of the SEQUEST normal and minimal cutoffs were calculated. The comparisons between the neural network results, the cutoffs and PepProphet are presented in figures 2.2 and 2.3. Figure 2.2 shows the plot

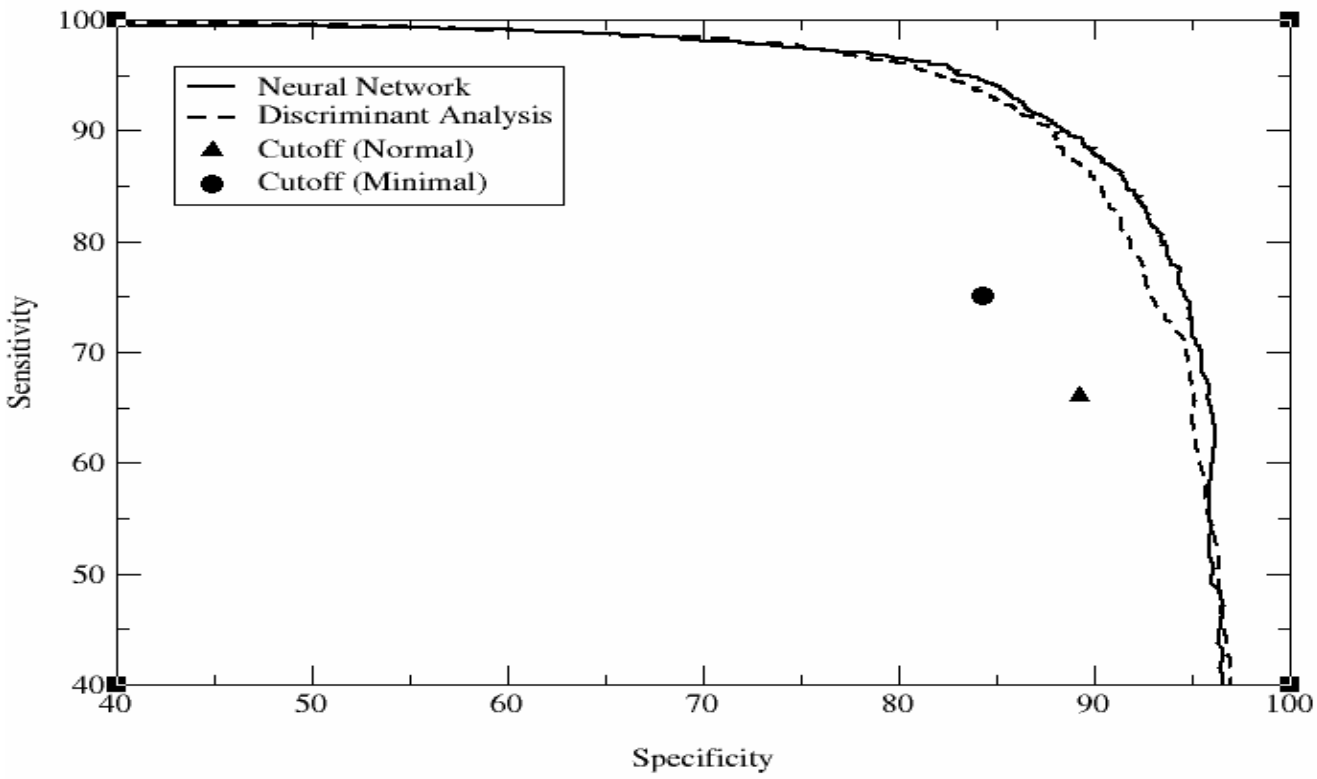

Figure 2.2 Specificity vs. specificity for ISB dataset. Prediction specificity vs. sensitivity on the ISB test set for different neural network scores, as compared to PepProphet results and to the SEQUEST predictions using both normal and minimal cutoffs through DTASelect.

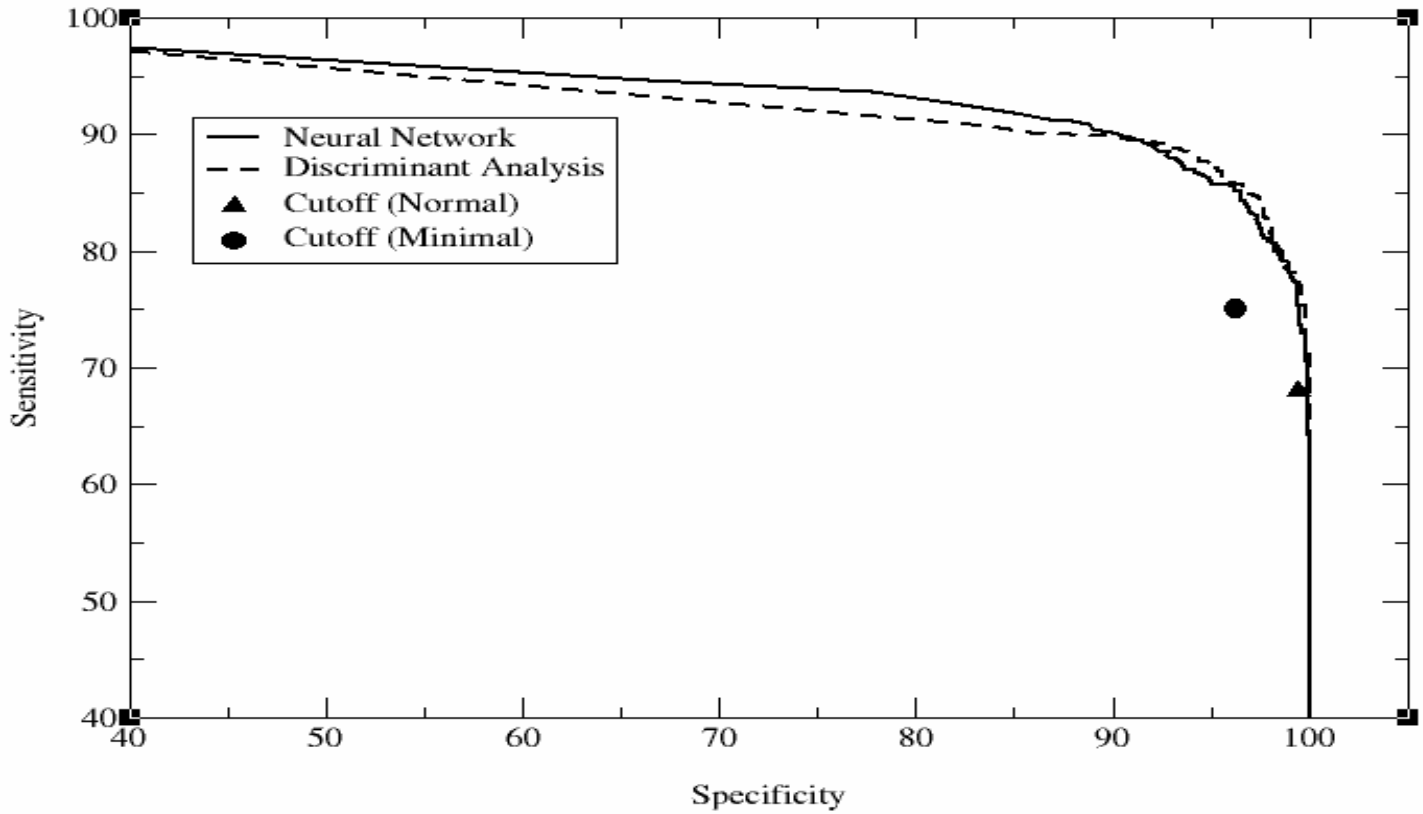

Figure 2.3 Specificity vs. sensitivity of eight-protein dataset. Prediction specificity vs. sensitivity on the eight-protein set for different neural network scores, as compared to PepProphet results and to the SEQUEST predictions using both normal and minimal cutoffs through DTASelect.

of sensitivity and specificity for different values of the neural network score, as well as for the minimal and normal cutoffs for the ISB test set and the results of PepProphet. Here, the neural network significantly outperforms the SEQUEST cutoff method for both the normal and minimal cutoffs. For example, at the 89% specificity, the SEQUEST normal cutoff achieves a sensitivity level at 66% compared to 89% by our method. The performances of the neural network and PepProphet are shown to be relatively similar though neural network consistently outperforms PepProphet throughout the sensitivity-specificity plot.

Figure 2.3 shows the sensitivity versus specificity plot for different values of our neural network score on the eight protein test set. Note that SEQUEST's cutoff performance is much better on the second set than on the first one because the first dataset is more complex than the second, in terms of peptide compositions and spectral quality. In addition, the first database for peptide searching is significantly larger than the second one. The performance for our neural network and PepProphet, however, appears to be somewhat lower in sensitivity on the eight protein set. The explanation for this lies in the definition of correct hits for the eight protein set. All the peptides that were assigned to the expected proteins were considered to be correct, even though the assignment could happen by chance (with X correlation score sometimes less than 0.8). These random hits were considered incorrect by all the methods and thus, their sensitivity was penalized. The manually validated ISB set, however, does not contain any random hits and therefore, the sensitivity of all the methods is higher. The performance of PepProphet on the eight protein data set is very similar to the performance of the neural network (neural network performs better than PepProphet in sensitivity except in the

specificity range greater than 92%). The similar level of performance on both sets, by our neural network, indicates that our method is robust. While our neural network approach improves on the prediction performance compare to SEQUEST's cutoffs, a key advantage of using this approach lies in its ability to provide probability assignments for peptide identification and the potential to combine these for protein identification reliability estimation.

Protein identification

SEQUEST generally provides a list of top ten hits for each experimental spectrum that it assigns. It is a general practice to take the top hit as the correct assignment, discarding the other nine. However, many people use the difference between the top ranked hit (will be referred to as X-correlation(1)), and the second ranked hit (X-correlation(2), the number in parenthesis ranges from 1 to 10, indicating the number of rank) to assess the goodness of the identification in the first hit (ΔCn). If there is no significant difference between X-correlation(1) and X-correlation(2), the first hit's reliability is undermined. However, many researchers are still concerned with the possibility of the second or even the third hit being correct rather than the first one. It is difficult to assess statistics when SEQUEST identification with the first hit was incorrect, but the second (or lower) hit was correct, since it involves ten times the manual effort than that required for just the first hit. The ISB test set used in this paper for neural network training does not contain any correct hits that are not ranked one. Thus, it was not possible to train a neural net with the consideration of rank in the identification as a parameter. We attempted to use the information from all ten SEQUEST identification hits for protein reliability assignment. It was done with the use of the neural network

(previously described) and without any consideration as to the rank of the peptide hit in the SEQUEST identification (each peptide hit, regardless of rank, was treated equally). For example, it is possible to have a peptide for which X-correlation(1) is lower than the X correlation(2-10) of another peptide. This treatment of ranks is, of course, naive. However, we decided to attempt this approach, which we call Neural Network 2, which relies on hits with X-correlation(1-10) rather than Neural Network 1, which only relies on hits with X-correlation(1). Thus, in Neural Network 2, the overall number of peptides assigned to proteins is roughly ten times higher than in Neural Network 1. It was hypothesized that even though the number of random hits in Neural Network 2 will be increased, the number of correct hits will contribute more, achieving a greater separation between signal and noise in protein identification.

To evaluate our method for estimating the reliability of protein assignments, we compared the following four approaches: (1) protein assignment based on the neural network scores for the top hit only (Neural Network 1), (2) protein assignment based on the neural network output for the 10 top hits (Neural Network 2), (3) DTASelect with the normal filter in SEQUEST, and (4) DTASelect without filters. On the eight protein test set (the simpler one), both Neural Network 1 and Neural Network 2, as well as DTASelect using the normal cutoff, identified all eight proteins as their top 8 predictions. Results of Neural Network 1and Neural Network 2 are shown in Table 2.3 and Table 2.4 respectively. A Pscore (Table 2.3, Table 2.4) is computed by combining the probabilities of a protein's peptides being incorrectly identified. The value, therefore, represents the likelihood that the protein in the sample has been identified by chance. Thus, the lower is Pscore, the higher the reliability of protein identification.

Table 2.3 Performance of Neural Network 1 on the 8-protein dataset.

Rank	PScore	#peptide Hits	Protein Name
1	1.45417902267382e-145	150	gi|2190337|gnl|PID|e321614
2	3.16526496278431e-119	105	gi|2506462|sp|P02188|MYG_HORSE
3	1.15508933747051e-109	120	gi|1168350|sp|P00330|ADH1_YEAST
4	8.23525223578428e-63	53	gi|122361|sp|P01966|HBA_BOVIN
5	1.86784851604797e-61	58	gi|122572|sp|P02070|HBB_BOVIN
6	3.24834848406027e-38	40	sp|P00698|LYC_CHICK
7	5.91934606492622e-33	40	gi|115453|sp|P00921|CAH2_BOVIN
8	0.000420000000000001	6	gi|113380|sp|P00331|ADH2_YEAST
9	0.2288	5	Contig7971.revised.gene2331.protein
10	0.33	3	Contig7971.revised.gene1007.protein

The first 8 proteins are correctly identified with given probabilities in both neural net results. Proteins with ranks 9 and 10 are not present in the sample and identified incorrectly. In Neural Network 1 the difference in probabilities between protein with rank 8 (correct) and protein with rank 9 (incorrect) is on the order of 1 order of magnitude.

Table 2.4 Performance of Neural Network 2 on the 8-protein dataset.

Rank	PScore	#peptide hits	Protein Name
1	3.04681564494944e-146	199	gi|2190337|gnl|PID|e321614
2	1.43470065809146e-120	130	gi|2506462|sp|P02188|MYG_HORSE
3	3.75183070708015e-112	138	gi|1168350|sp|P00330|ADH1_YEAST
4	7.74113710163722e-63	60	gi|122361|sp|P01966|HBA_BOVIN
5	1.50924027945192e-61	68	gi|122572|sp|P02070|HBB_BOVIN
6	9.09537575536875e-39	48	sp|P00698|LYC_CHICK
7	5.68434802614865e-33	52	gi|115453|sp|P00921|CAH2_BOVIN
8	4.17398305533061e-29	75	gi|113380|sp|P00331|ADH2_YEAST
9	1.09152586669637e-05	20	Contig7971.revised.gene2048.protein
10	0.0004899779900928	15	Contig7971.revised.gene2240.protein

The first 8 proteins are correctly identified with given probabilities in both neural net results. Proteins with ranks 9 and 10 are not present in the sample and identified incorrectly. In Neural Network 2 the difference in probabilities between protein with rank 8 (correct) and protein with rank 9 (incorrect) is 28 orders of magnitude.

For example (Table 2.3), according to the results of Neural Network 1, the likelihood of seeing myoglobin (second hit) in the sample by chance is 3.16e-119 based on its peptide hits. The likelihood of seeing yeast alcohol dehydrogenase II by chance, according to Neural Network 1, is much higher, with a probability of 4.2e-3 while the first incorrect protein has a Pscore of 3e-1.

On the more complex ISB dataset, protein identification results were found to be significantly different by these different methods. The comparison of the four methods in terms of sensitivity and specificity of protein identifications is shown in Figure 2.4. The first 500 protein identifications for each method were used in the analysis, showing the change in sensitivity and specificity based on the rank of protein identifications. The plot was made by creating rank cutoffs for each method in incremental step of 10 from 10 to 100 consequently computing sensitivity and specificity of each method. After the hundredth rank, the step size was increased to 100. Of the 29 proteins manually identifiable in the ISB set, DTASelect with the normal filter found 17 proteins ranked between 1 and 100 while Neural Network 1 found 23 proteins in the first 100 hits. A more detailed comparison is given in figure 2.4. From this figure, we can see that Neural Network 1 consistently outperforms DTASelect (SEQUEST) by a large margin. However, Neural Network 2 performs significantly worse than Neural Network 1, having identified only 62% of the proteins within the first 100 hits, while Neural Network 1 identified 79% and DTASelect with normal cutoff identified 58% (Figure 2.4). Protein assignments based on DTASelect unfiltered, regardless of protein rank or dataset, showed no usable specificity (Figure 2.4). The comparison of performance between the four methods shows that overall, the neural network approach is superior to simple filtering.

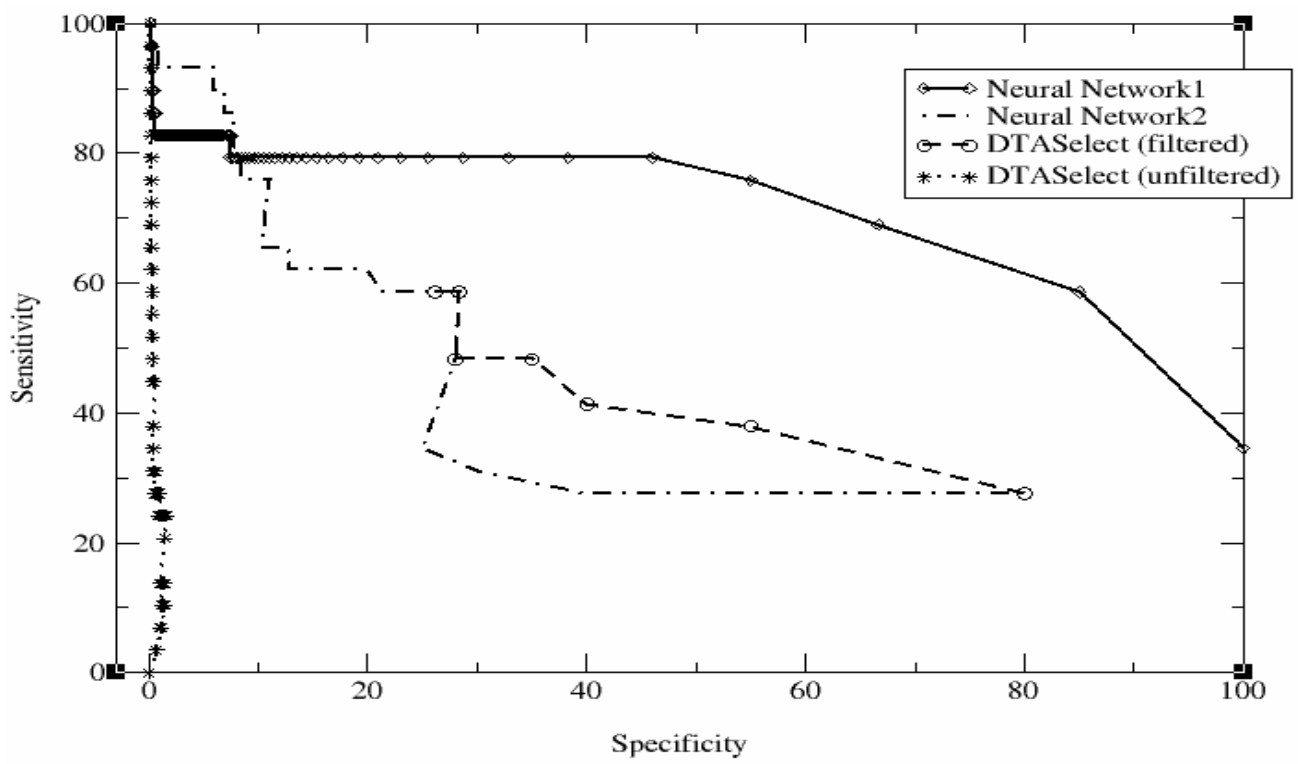

Figure 2.4 Specificity vs. sensitivity of protein detection. Comparison of sensitivity and specificity for identified proteins between DTASelect filtered, DTASelect unfiltered and the two neural network approaches as the rank cutoffs increase for the expected 29 proteins. Sensitivity and specificity are computed for rank windows (1-10, 1-20, 1-30 … 1-200…), as rank increases, the specificity of identification decreases while sensitivity increased.

DTASelect filtered lacks the sensitivity needed for protein identification in complex mixtures. Because there is no cutoff applied at the peptide level, it is understandable that the other three methods have sufficient sensitivity to identify most of the proteins with different false positive rates. In particular, DTASelect unfiltered and Neural Network 2 lack practical specificity. Neural Network 1, however, provides superior sensitivity and specificity compared to conventional methods.

Another approach for protein reliability estimation based on SEQUEST scores, presented by MacCoss et al (MacCoss, 2002) appears very similar to the method proposed in this chapter. The differences between the two approaches lay in the treatment of peptide probabilities and the implications in the model of protein reliability. MacCoss's reliability of peptide matching appears to solely depend on the use of normalized X-correlation score. X-correlation score is normalized through dividing it by the autocorrelation score of the experimental spectrum; that is expected to minimize the effect of peptide charge and length on the X-correlation. This approach to normalization of X-correlation assumes that the users have SEQUEST source code to modify the resulting X-correlation (as autocorrelation has to be calculated exactly as X-correlation is in order to provide correct results). It does not take into account the effects of ΔCn, SpRank, and ion coverage on the probability of correct peptide identification, excluding them from MacCoss's $Pept_{prob}$ model. We expect, however, that the peptide probability model would be improved by using these parameters, which are also often used in simple filtering procedures. The $Prot_{prob}$ proposed by MacCoss is depicted as a real probability model of identifying protein from its peptides. However, it ignores an important fact of non-unique peptides that are not an evidence of any one protein. Additionally, due to the

greatly varied number of peptide matches per protein, the resulting probability is often misleading. For example, from the Neural Network 1 part of Table 2.3, it is shown that according to the $Prot_{prob}$ the proteins that rank 1 and 7 have probability of 1 to be present in the sample (as their probability of being identified by chance is negligibly small comparing to 1), and the incorrectly identified proteins with the rank 9 and 10 will have probability of detection respectively 0.99 and 0.98. The approach proposed in this chapter is preferable, as the difference between more likely proteins and less likely proteins is more clearly seen by the difference in exponents. It will eventually be possible to automatically detect where the most likely cutoffs lie between correct and incorrect protein identifications.

Significant improvement to the conventional methods of peptide filtering and protein assignment has been demonstrated, compared to standard cutoffs. This improvement was achieved by training a neural network, which utilizes a number of additional parameters not usually considered in filtering. The neural network score provides a more accurate basis for estimating peptide identification likelihood, as well as a foundation for statistical scoring of protein assignments. The results also show, however, that even with these improvements the methodology is far from perfect.

At the current stage of development, it remains difficult to distinguish between the scores of small proteins present in the sample. There is also a problem in that often times a particular peptide may be observed and counted multiple times, artificially inflating the protein identification score. It may be possible to remedy this situation by constructing a more complex statistical model that would involve coverage of proteins by peptides.

Chapter 3

Charge Determination for Low Resolution Mass Spectrometry

Some of the data presented below has been presented as Razumovskaya J., Fridman T., Day R., Borziak A., VerBerkmoes N., Hettich R., Uberbacher E., Gorin A., poster presentation at ASMS 2004

Introduction

Mass spectrometry

Analysis of protein mixtures is one of the most important and difficult technological challenges for high-throughput proteomics. Mass spectrometry is an analytical technique that is capable of accurately measuring the mass-to-charge ratio (m/z, where 'm' stands for the mass of the measured molecule, while 'z' represents its charge state) of gas-phase ions originating from biological molecules**.** One of the most popular instruments used today is robust, sensitive and inexpensive quadrupole ion trap mass spectrometer used in bottom up shotgun proteomics approaches (McCormack 1997, Martin, 2000, Shen, 2001). With all the advantages of using quadrupole ion trap instruments for high throughput protein identification, they lack the necessary resolution (at least under high-throughput operating modes) to measure directly the charge state of peptides, making the determination of *mass* from the m/z measurement ambiguous. While more sophisticated mass spectrometry instrumentation can record spectra with sufficient resolution for unambiguous charge state determination, the abundance of proteomics data generated by quadrupole ion trap instruments dictates the need for substantial advances in computational algorithms for interpretation of spectra from these instruments, including robust charge determination algorithms. Here a new approach for

accurate charge state assignment is presented involving using a trained neural network to detect charge-specific spectral characteristics.

Charge determination problem for database search algorithms

One of the main approaches to tandem MS data interpretation is database searching. As described in the Introduction "Database search algorithms" section, a typical database search involves scanning a protein database for potential candidate sequences that match the experimentally observed peptide's spectrum by such parameters as the peptide's m/z or a short sequence tag obtained from the peptide's fragmentation pattern (MS/MS spectrum) through sequencing. In this analysis every piece of information can play a critical role in the reliable identification of the peptide. Some of the most frequently used methods, such as mentioned SEQUEST (Eng, 1994), MASCOT (Perkins, 1999), SONAR (Field, 2002) rely on using the peptide m/z ratio as a step for obtaining the candidate sequences from the database. The database typically consists of the predicted peptide masses derived from *in-silico* digestion of the proteins present in the database. In order to obtain database candidate sequences that match the experimentally measured peptide's m/z, the masses of the theoretical peptides must be compared to the m (mass) of the peptide, for which its z (charge) must be determined or at least assumed. In case of incorrect mass assumption, the true peptide's sequence does not appear in the list of the candidate peptides leading to either no identification or a false positive identification.

The resolution of an ion trap instrument does not allow for direct determination of charge state of peptides by looking at the isotopic packet distances as it is done with some mass spectrometers that are capable of higher resolution (Loo, 1992). In case of a

peptide carrying a single charge, the CID produces MS/MS spectrum where all the real fragment ions fall below the precursor mass. Thus the singly charged state can generally be easily determined by comparing the percent of spectral intensity before and after the precursor mass (95% of intensity below precursor ion generally constitutes a singly charged peptide as mentioned in the work describing 2to3 (Sadygov, 2002)). However, in case of higher charge states, the fragment ions can appear in different charge states, as well as either above or below the precursor m/z. From CID spectrum of a doubly-charged peptide parent ion, we can observe fragment ions of charge one or two below the precursor m/z, and of charge one above the precursor m/z. In case of multiply charged peptides, the spectral density alone does not allow for a good discrimination between the charge states.

Current charge determination approaches

The widely applied approach is to differentiate between singly and multiply charged ion spectra. Since obtaining tandem mass spectra from parent ions with charge states higher than three is considered rare (Sadygov, 2002), all the multiply charged spectra are considered to be either doubly or triply charged. Therefore, the charge states of two or three are assumed and then the corresponding molecular weight for each charge state is computed. The database search then proceeds to search for database peptides with both calculated molecular weights. The actual charge state of a peptide is then inferred from the results of the database search – the charge state with the higher score is assumed to be the correct charge state with consideration for the charge dependant increase in X-correlation. While this approach prevents loss of peptide assignments due

to the incorrect charge determination, the search time is nearly doubled by the approach, and the possibility of false positives is increased.

A recently published approach, 2to3, (Sadygov, 2002) assigns multiply charged spectra to charge states +2 or +3, while the un-assignable spectra remain in the duplicate versions. The approach is based on counting all the pairs of fragment peaks that match the expected parent masses (for either charge 2 or charge 3) and then choosing the charge state that accounts for the most of the fragment ions found in the spectrum.

Another relevant paper by Colinge et al (Colinge, 2003) reviews three algorithms focused at peptide charge assignment. The algorithms are denoted as Algorithm (N) and Algorithm (K), "posteriori charge assignment" algorithms and Algorithm (B), "integrating observations" algorithm. Algorithm (N) involves dividing a tandem mass spectrum into a set of intervals according to parent mass with different possible charges. They propose designing a stochastic model to evaluate the distributions of fragment m/z values in the specific intervals defined by the parent mass to charge ratio and compare the modeled distributions to each of the considered spectra. Algorithm (B) is based on complementary ions, much as previously discussed 2to3 algorithm and parent mass correction algorithm (Dancik, 1999). The third Algorithm (B) is a combination of Algorithm (N) and Algorithm (K).

In this chapter, a new method for charge state assignment is presented. It involves using an artificial neural network to determine a charge state of a precursor ion based on a set of charge-specific features found in its tandem mass spectrum. The features that are used to differentiate between the multiple charged peptides involve: long distance information (amino acid differences), short distance information (small neutral losses),

the number of fragment ions consistent with possible parent masses and relative densities of peaks in regions of the spectrum. The last two features were also used in the previous approaches. The addition of new characteristics such as long distance information and the short distance information coupled with a neural network allows us to significantly improve the quality of charge determination. The results of our charge determination show an accuracy of roughly 99% for complex samples in charge state assignment, with only 10% of spectra with unassigned charge state.

Materials and Methods

Training sets

A protein standard mixture dataset, the "*Seattle dataset"*, provided by the Institute for Systems Biology (ISB, Seattle, Washington), published by Keller et. al (Keller, 2002*), was used as the training data for the systems. It includes 18 proteins, and additionally considers a few human contaminants, bringing the number of proteins in the sample to 29. The peptide data used in our analysis includes 1,565 identified peptides of charge 2 and 914 identified peptides of charge 3.

An additional dataset from a standard mixture of proteins was used for testing the system. A protein standard mixture of eight proteins including 5442 tandem mass spectra was prepared and analyzed by 1D-LC-ES-MS/MS. This "*ORNL dataset*", was fully described in a previous study on reliability assessment (Razumovskaya, 2004).

An *Escherichia coli* proteome data set, the "*E. coli dataset*" was used in both testing and training. This dataset was generated from an *E. coli* K-12 strain grown deep into stationary phase and analyzed by 1D-LC-MS/MS with multiple mass range scanning and contained 35,486 tandem MS spectra. Briefly, the cells from a 2-L culture grown

deep into stationary phase were harvested, washed twice with Tris buffer (pH 7.5) with 10mM EDTA and lysed with sonication. Four crude protein fractions were created by ultracentrifugation (100,000g for 1 hour creates membrane and crude fraction and then for 24 hours creates pellet and cleared fraction). Protein fractions were denatured, reduced and digested with sequencing grade trypsin. The resultant peptide mixtures were de-salted with solid phase extraction (C-18), concentrated and filtered to give a final concentration of ~10μg/μL based on starting material. All tryptic digestions of all fractions were analyzed via one-dimensional LC-MS/MS experiments performed with an Ultimate HPLC (LC Packings, a division of Dionex, San Francisco, CA) coupled to an LCQ DECA XP ion trap mass spectrometer (Thermo Finnigan, San Jose, CA) equipped with an electrospray source operated at 4.5kV. Injections were made with a Famos (LC Packings) autosampler onto a 50μl loop. Flow rate was ~4μL/min with a 240-min gradient for each LC-MS/MS run. A VYDAC (Grace-Vydac, Hesperia, CA) C18 column (300μm id x 15cm, 300Å with 5μm particles) was directly connected to the Finnigan electrospray source with 100μm id fused silica. For all LC/MS/MS data acquisition, the LCQ was operated in the data dependent mode with dynamic exclusion enabled, where the top four peaks in every full MS scan were subjected to MS/MS analysis. To increase dynamic range in the 1D-LC-MS/MS analysis separate injections were made with a total of 8 overlapping segmented m/z ranges scanned (referred to as gas phase fractionation or multiple mass range scanning). The resultant MS/MS spectra files from all fractions were searched with SEQUEST against all predicted ORFs from *E. coli*. The raw SEQUEST output files were filtered and sorted with DTASelect (Tabb, 2003) with the following parameters: fully tryptic peptides only, with delCN of at least

0.08 and cross-correlation scores (X-correlations) of at least 1.8 (+1), 2.5 (+2) and 3.5 (+3). All peptides passing these criteria were kept for further analysis.

Neural network training and testing

Stuttgart Neural Network Simulator 4.2 (SNNS) (http:/www-ra.informatik.uni-tuebingen.de/SNNS/) was used to train an artificial neural network to assign charge states to multiply charged peptides measured by quadrupole ion trap. The neural network connection weights were trained using the back-propagation learning algorithms, with all the standard parameters suggested in the SNNS package. Performance results were saved every 60 cycles throughout the training process, the training procedure stopping when no improvement in the error rate could be achieved. Each of the resulting nets was then tested for performance and the best was selected based on performance on the training and testing sets. The neural network output ranges from 0 to 10, where the output represents the charge state of the precursor ion – 0, stands for charge 2+, while 10, stands for charge 3+.

The three non-overlapping datasets described in the "Training Sets" section, the *ORNL dataset*, the *E. coli dataset* and the *Seattle dataset* were used to train and test the neural network for charge determination. The well-characterized, manually curated *Seattle dataset* (Keller, 2002) was used for training of the neural net; it includes 1565 examples of charge 2+ tandem MS spectra and 914 examples of charge 3+ tandem MS spectra. The method was tested on two non-overlapping data-sets acquired at Oak Ridge National Laboratory 1) the *ORNL dataset*, acquired from a standard mixture of 8 proteins, and 2) the *E. coli dataset*, derived from a complex mixture of proteins, the proteome of a whole organism. In order to increase the size of the training set, the high

confidence identifications from *E. coli dataset* were included for *ORNL dataset* testing and the high confidence identification from *ORNL dataset* were included in the *E. coli dataset* testing, creating the unbiased non-overlapping training datasets. The high confidence identifications included into the training sets from *ORNL* and *E. coli* datasets were chosen as follows: all spectra assigned to charge 2+ peptides with SEQUEST's X-correlation score of 2.8 and higher, (X-correlation scores of 2.5 or above are generally required for confident identifications of charge 2+ peptides) and all spectra assigned to charge 3+ peptides with SEQUEST's X-correlation of 3.8 and higher (confindent identification requires X-correlation score of 3.4 or above). The training set for *ORNL dataset* included the full *Seattle dataset* and a filtered collection of spectra from *E. coli dataset*, including 1963 examples of charge 2+ tandem MS spectra and 1001 examples of charge 3+ tandem MS spectra. The training set designed to test the performance of the charge determination approach on the *E. coli dataset* included *Seattle dataset* and 395 examples of charge 2+ spectra and 131 examples of charge 3+ spectra from *ORNL dataset*. Thus the neural network trained with the first training set was only tested on a non-overlapping *ORNL dataset* and neural network trained on the second training set was tested on non-overlapping *E. coli dataset*.

Algorithmic approach

The development of the charge determination approach involved the following three steps: 1) identifying the fragmentation pattern features that are related to the charge of precursor ion, 2) training an artificial neural network to recognize the charge state of the precursor ion based on the set of features found in its fragmentation pattern, 3) selecting the neural network cutoffs for the assignment of the charge states. The rationale

behind choosing the fragmentation pattern features is discussed in the following three sections: “CID spectrum” and “Underlying principles” and “Charge Determining Spectral Features”. The spectral features used for the charge identification are: long distance information (amino acid differences) for each of the charge states considered (3 parameters), short distance information (neutral losses) (3 parameters), parent masses for the charge 2+ and charge 3+ states (2 parameters) and relative densities below and above parent mass (2 parameters) each of which is discussed separately in the “Charge determining spectral features” section. Using the set of ten parameters derived from the mentioned features we trained an artificial neural network, as described in “Neural network training and testing” section and based on the results selected reasonable neural network score cutoffs for the final charge state determination.

CID spectrum

Tandem mass spectrometry produces a sequence dependent fragmentation pattern and the mass analyzer records the m/z of the resulting ions. The number of positive charges on the parent ion (H+) determines the total charge of the peptide and therefore real parent ion mass (PM) can be computed as follows: PM(real) = PM(observed)*charge - charge or PM = (m/z)*z - z. As a result of CID fragmentation, a peptide tends to fragment by breaking along the backbone bonds and creating a pair of fragments, which if they retain a charge are referred to as ‘b’ and ‘y’ ions (Roepstorff, 1984; Biemann, 1988). The number of singly charged ‘b’ and ‘y’ ions for each given peptide is equal to the number of the peptide bonds. A peptide containing N amino acids, has N-1 peptide bonds, it’s CID spectrum is expected to have N-1 possible ‘b’ and N-1 possible ‘y’ ions of charge 1+. The sum of masses between the singly charged ‘b’ and the corresponding

‘y’ ions in tandem MS spectrum produce the mass of the parent ion PM= b + y -1, and the sum of charges between the ‘b’ and corresponding ‘y’ ions, should produce the total charge on the parent ion: $z_P=z_b + z_y$ where z_p is the charge on the peptide p. Many of the ‘b’ and ‘y’ ions have been shown to have additional fragmentation, resulting in a series of trailing peaks formed by the losses of various chemical groups, such as loss of water and loss of ammonia, etc (Dancik, 1999, Fridman, 2003) (the considered chemical losses off the ‘b’ and ‘y’ ions were: H, H_2O, NH_3, CO, CO-H_2O, CO-NH_3, NH_3-H_2O, where ‘-‘ refers to the loss of the corresponding chemical group).

Underlying principles

The main idea of the charge determination described here is based on the concept that different charge states of a parent peptide lead to variations in the fragmentation patterns. For example, a peptide with a total charge of 1+ will only produce fragment ions of 1+ charge, while the same peptide with charge 2+ might produce 1+ and 2+ fragment ions, and a peptide with charge 3+ could have all the charge states present in the fragmentation pattern, as illustrated in Figure 3.1. The peptide with total charge of 2+ is likely to fall apart in the following patterns: ‘b’ ion of charge 1+ and corresponding ‘y’ ion of charge 1+, or ‘y’ ion of charge 2+, with unobservable neutral fragment in place of a ‘b’ ion and conversely ‘b’ ion of charge 2+, with unobservable ‘y’ fragment ({1+,1+} pattern or {0,2+} pattern). The peptide with total charge of 3+ can fall apart into {1+,2+} pattern or {0,3+} pattern. It was also noted that, due to charge repulsion, protons favor to be at a distance from each other making {1+,1+} and {1+,2+} patterns more frequent (Figure 3.1). Since the patterns are different, it should be possible to differentiate between the precursors of charge 2+ and charge 3+ (as well as higher charge states).

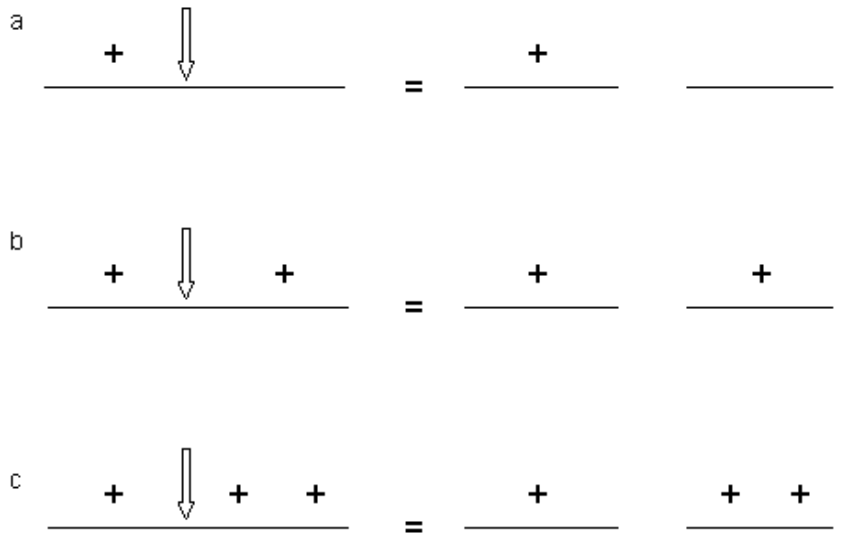

Figure 3.1 Charge dependent patterns. The most frequently observed charge fragmentation patterns in the SEATTLE dataset. The vertical arrows indicate the cleavage site for each parent ion.
a. Peptides of charge 1+ fragment into an ion with 1+ charge and a neutral fragment,
b. Peptides of charge 2+ are likely to fragment into b ion with charge 1+ and a y ion with charge 1+,
c. Peptides of charge 3+ are likely to fragment into b and y ions with one ion carrying 1+ charge and the other with 2+ charge.

It was noted that more fragments from charge 2+ precursors carry 1+ charge with little of 2+ charged ions, while fragments from charge 3+ precursor have roughly equal amount of 1+ and 2+ charged ions.

Charge determining spectral features

Long range information (Amino acid differences)

Most of the software developed for interpretation of tandem mass spectra relies on the fact that in an ideal case a peptide is expected to fragment along the peptide backbone, producing a series of consecutive ions separated by m/z of one amino acid residue as described in "CID spectrum" section. In the case of a continuous succession of 'b' (or 'y') fragment ions with charge 1+, the mass differences between consecutive ions would correspond to single amino acid masses. In the case of a series of fragment ions with charge 2+, the mass differences should match amino acid masses divided by 2, etc. This insight can be used to assess the number of differently charged fragment ions present in the spectrum. It would be expected that different charge spectra have a different number of 1+, 2+ and 3+ amino acid differences. To use this information for charge state determination, the instances of amino acid differences of charge 1+, charge 2+ and charge 3+ are counted and weighted. Computing these weights involves the summation of the normalized intensity products of peak pairs, whose m/z are separated by mass of one amino acid or one amino acid divided by the respective charge. For example, as the differences between the m/z values of all pairs of consecutive charge 1+ 'b' (or 'y') ions yield to the masses of amino acids, *AA1* is the sum of the product intensities of the two consecutive ions, normalized to the square of the highest peak in the spectrum. Since the ion identity is unknown, all the peak pairs located at the amino acid

distances with an error of 1 Da of each other are accepted in the analysis, which potentially includes the neutral loss ions and the noise peaks. The three charge states of amino acids provide the three parameters (*AA1*, *AA2* and *AA3*) for the long range information in the spectrum. The contribution of long range information to the differentiation between charge 2 and charge 3 spectra is shown on Figure 3.2a. The separation based only on the long range information is not in itself conclusive, as most of charge 2 and charge 3 spectra have similar scores derived from the neural net analysis.

Short range information (Event differences)

In addition to the amino acid differences, it is expected that the differences between main ion types and their corresponding satellite peaks, described above in the ”CID spectrum” section, will be similarly affected by the charge states. Since this information is contained within approximately 50 m/z region surrounding each b or y ion peak, we refer to this as the “short range” information. For example, a ‘b’ fragment ion with a charge state 1+ has a water loss ion peak 18 m/z units lighter, while a ‘b’ fragment ion with a charge state of 2+ having a water loss ion peak only 9 m/z units lighter. The full set of such small neutral losses that we considered in our model is listed in the “CID spectrum” section. The procedure for computing the short range parameters is very similar to the one described above in the section describing the “Long Range Information”. Computing *ED1* involves summation of the normalized intensity products between the expected charge 1+ ions and their corresponding neutral loss satellite peaks, while *ED2* and *ED3* are computed by summing the normalized intensities of ion pairs located at the event masses divided by 2 and by 3 respectively. The short range

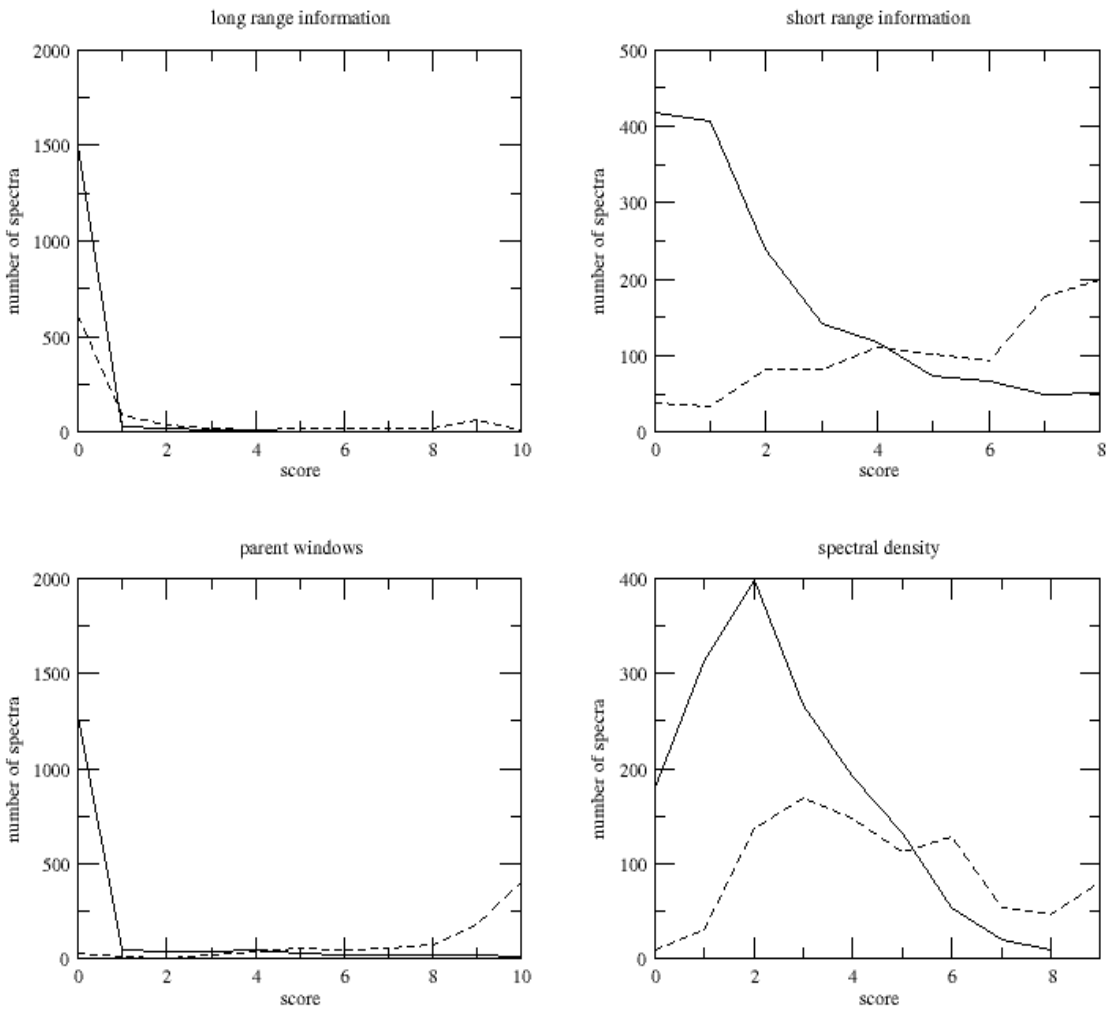

Figure 3.2 Significance of spectral parameters for charge state determination. The performance of charge separation by each of the four considered factors separately: a) long range information, b) short range information, c) parent windows, d) spectral density. The solid line stands for charge 2+ peptides, dashed line stands for charge 3+ peptides, X-axis denotes the neural network score, while Y-axis shows the number of spectra with the given score. The parent windows factor shows the greatest separation between charge 2+ and charge 3+ spectra.

information provides us with another three parameters (*ED1*, *ED2* and *ED3*). The short range information in itself also fails to provide a distinct differentiation between parent ions of charge 2 versus charge 3. Figure 3.2b demonstrates that the distributions of scores derived from the neural net analysis for charge 2 and charge 3 spectra are fairly broad, with considerable overlap. However, the short-range scores in Figure 3.2b provide a more marked distinction between parent ions of charge 2 versus charge 3 than the scores derived from amino acid differences (Figure 3.2a).

Parent windows

The concept of complementary ions is a well studied phenomenon in tandem mass spectrometry of peptides. When a peptide dissociates into two corresponding fragment ions, b and y, the sum of the masses of the two complementary fragment ions equals the mass of the parent peptide. This property is extensively used by charge determination algorithms such as those mentioned above: 2to3 (Sadygov, 2002), and Algorithm(K) and Algorithm(B) (Colinge, 2003). When the parent m/z is measured, the mass can be calculated with a particular assumed charge state and the summation of candidate pairs of complementary ions can be used to deduce the correct parent mass. The true parent mass "window" (a window of ±2 Da around given mass value) collects the sums of true complementary pairs of fragment ions, while other assumed parent mass windows collects random pair wise sums of m/z values. The comparison between the number of fragment pairs that produce the possible parent mass in summation is used to determine the likely charge state of the peptide. In our approach the procedure involves multiplying the normalized intensities of the two corresponding 'b' and 'y' fragment ions (their m/z values add up to a particular parent mass) and summing all the resulting products for a

particular parent mass to produce a final parent mass score (*PM*). The procedure is applied to all the considered parent masses, in the case of differentiating between charge 2+ and charge 3+ charge states, it involved computing the value for parent mass "window" of charge 2+ (*PM2*) and the value for parent mass "window" of charge 3+ (*PM3*), creating two parameters for the charge state determination algorithm. The potential capability of parent windows for determining the charge state is shown in Figure 3.2c. It is undoubtedly the most powerful of all the techniques that we apply to the charge state identification, however, it was reinforced by the addition of the other parameters.

Relative density

In case of parent ion of charge 1+, all of the valid fragment ions fall into mass range lower than that of the m/z of the parent ion while the mass range above the parent ion contains no peaks that belong to the parent peptide of charge 1+. Let us divide the spectrum into two parts: the first part with lower m/z range than the observed parent m/z and the second part with higher m/z range than the observed parent m/z. As previously described, in charge 1+ spectra, all the fragment ions have charge 1+, and they are concentrated in the first part of the spectrum, with no peaks due to authentic fragments of the parent ion in the second part of the spectrum. The approach currently used to identify charge 1+ spectra involves the comparison of composite intensities in the first (*D1*) and second (*D2*) parts of the spectrum. If the total intensity in the second part of the spectrum is below 5% of the total intensity of the spectrum, the spectrum is considered to be a charge 1+ spectrum. In case of a spectrum being produced from a parent peptide with charge 2+, the peaks located below the parent m/z hold either charge 1+ or charge 2+,

while all the peaks in the second part of the spectrum are of charge 1+. The reason for this model is based on the following: all of the fragments observed in the tandem mass spectrum must be less than or equal to the parent in mass; any fragment observed with higher m/z than the parent m/z must have a lesser charge than the parent ion, while the peaks below the parent m/z can have any charge from 1 up to the charge of the parent peptide ion. As the parent charge increases, the parent m/z becomes lower, and the fragment ions in the first region can carry higher charges. The ratio between the intensities in the first and second part of the spectrum changes depending on the peptide charge. A similar approach was used by Colinge, in his algorithm(N) (Colinge, 2003), however, here it is simplified to the division of the spectrum into only two parts (which is more realistic, considering mass to charge range), adding two parameters, *D1* and *D2* to the charge determination where they are used in combination with all the other parameters. The neural net scores for spectrum density are shown in Figure 3.2d.

Selecting cutoffs

The ten variables produced by the four contributing factors – the *AA1*, *AA2* and *AA3* (1+, 2+ and 3+ charge state parameters from the long range information), the *ED1*, *ED2* and *ED3* (1+, 2+ and 3+ from short range information), *PM2* and *PM3* (2+ and 3+ charged parent windows) and *D1* and *D2* (above and below parent mass densities) are used to describe the charge state of a given spectrum. Using the training sets described in "Training set" section, we trained and tested the separation between the charge 2+ and charge 3+ precursor ions based on the given parameters. The results of our charge determination for the *Seattle training dataset* are shown in Figure 3.3 and summarized in Table 3.1.

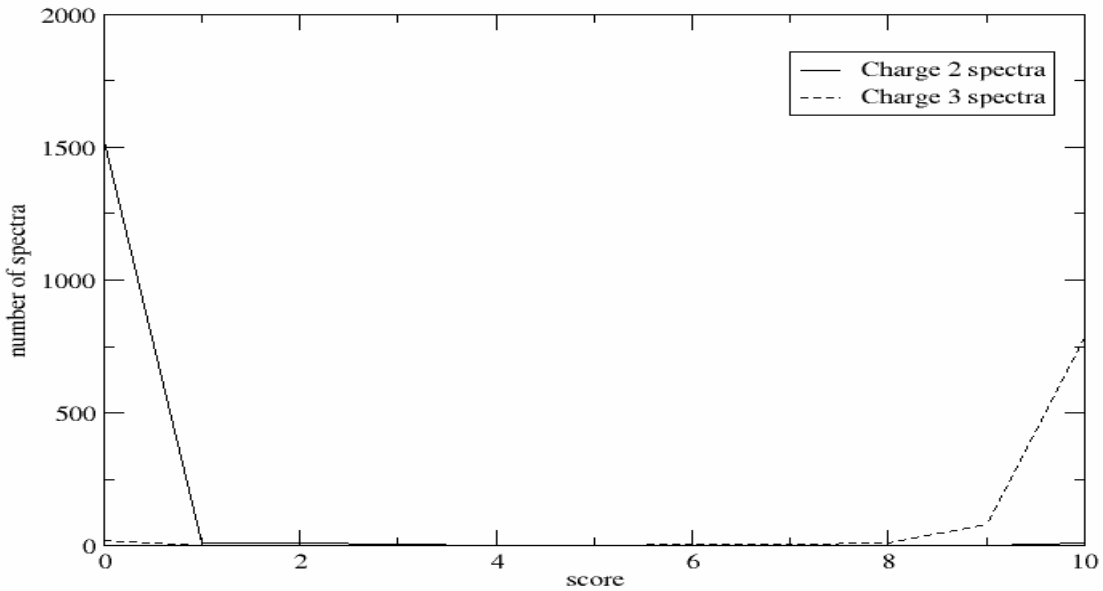

Figure 3.3 Charge separation results for the *Seattle dataset*. The distributions of charge 2 and charge 3 spectra by the neural network score. The charge 2 spectra are denoted by a solid line – most of the charge 2 spectra have a score below 1; charge 3 spectra are denoted by dashed line – most of charge 3 spectra have a score higher than 8.

Table 3.1 Charge state assignment results for neural net training with *Seattle dataset*

Parent ion charge	Correct*	Incorrect*	Undecided*
+2	1546	10	9
+3	878	26	10
Percent IDs	97.74%	1.45%	0.766%

The charge assignments for Seattle dataset *relative to the manually curated assignments (Keller, 2002)

Since the *Seattle dataset* was manually curated, all the charge states for the tandem MS in the dataset are considered correctly pre-assigned. The charge state discrimination presented in Figure 3.3 is significantly better defined than in any of the four contributing factors shown in the Figure 3.2. Most of the charge 2 spectra are shown to have a neural network score below 1, while most of charge 3 spectra have a score higher than 8. Based on the analysis which minimizes the erroneous assignments (increasing the number of undecided assignments), all the spectra with score less or equal to 1 are considered to be a product of charge 2+ peptides, while all the spectra with score greater or equal to 9 are considered to be a product of charge 3+ peptides. The peptides with scores between 1 and 9 are the "unassigned spectra", which will be subjected to SEQUEST analysis as both possible charge 2+ and charge 3+ spectra. The results show a rate of incorrect charge state assignment of 1.45% and unassignable spectra rate of 0.8% for *Seattle dataset* (36 spectra out of 2479 would have been incorrectly assigned and 19 spectra would have been considered as both charge 2 and charge 3 peptides). The absence of any single given parameter decreased the efficiency of charge separation in the training set.

Results and Discussion

The charge determination algorithm for low resolution quadrupole ion trap mass spectrometer promises an improvement to high throughput data interpretation in two main aspects: it is capable of significantly decreasing the number of false positive identifications and the increase of the speed of the analysis. It is difficult to ascertain the rate of false positive identifications made by search algorithm like SEQUEST, the only methods for such analysis are generally increasing the search space (by increasing the database size) and comparison to other algorithms. Here, we attempt to examine the SEQUEST's performance in a different manner – by comparing the performance of the described charge state determination algorithm to the performance of SEQUEST's charge state assignment under different X-correlation cutoffs. The performance of the described method is not affected by the changing X-correlation cutoffs, however, as they are increased the number of true positive SEQUEST's assignments increases and the true performance of the charge state algorithm can be seen. The true performance exhibited by our method for a real proteomic sample at 0.6% error and 10% unassigned spectra at highest X-correlation suggests that at the accepted X-correlations, the charge determination algorithm will be able to reduce the number of false positives from the 1.8% shown at X-correlation of 2.6 for charge 2+ to 0.6% while the number of unassigned spectra will be 17.7%.

One of the problems in measuring the error rate of bottom up data interpretation algorithms is that it is difficult to obtain a large set of correctly assigned peptides. If the dataset contains accepted 5% of incorrect identifications, the percent error of the tested method cannot be lower than 5%. In our test cases (the non-manually curated *ORNL*

dataset and *E. coli dataset)*, we had to rely on SEQUEST database search to produce correct peptide assignments for our training and test sets, however, there is no guarantee that all the assignments used in the analysis are correct. The peptide assignments are mainly based on particular set of cross-correlation (X-correlation) cutoffs derived from SEQUEST and different laboratories apply different cutoffs for their peptide identification. The high X-correlation cutoffs used for the training sets reduce the number of false identifications, however, with the increasing cutoffs the number of unidentified peptides also increases making the analysis unrealistic. In order to circumvent the problem of inflated error, we present a scheme to measure the performance of the charge determination algorithm independently of the static X-correlation cutoffs. The algorithm's performance is displayed in Figure 3.4 for different ranges of X-correlation cutoffs, to show the relationship between the accuracy of peptide sequence identification and charge state assignment. For charge 2+ peptides is X-correlation cutoffs are varied in the range from 1.0 to 3.8, in increments of 0.2. The charge 3+ X-correlation cutoffs range from 2.0 to 4.8 in increments of 0.2. As the X-correlation cutoff is increased, the number of false positive identifications decreases, the percent of SEQUEST's incorrect identifications is negligibly small with very high X-correlation cutoffs (as 3.5 for charge 2+ peptides and 4.5 for charge 3+ peptides). When the percent of incorrect identifications is reduced, the real performance of our algorithm can be observed. This performance measure should provide an accurate assessment of error rate of our algorithm, as well as provide the expectation of the algorithm's performance based on any set of cutoffs.

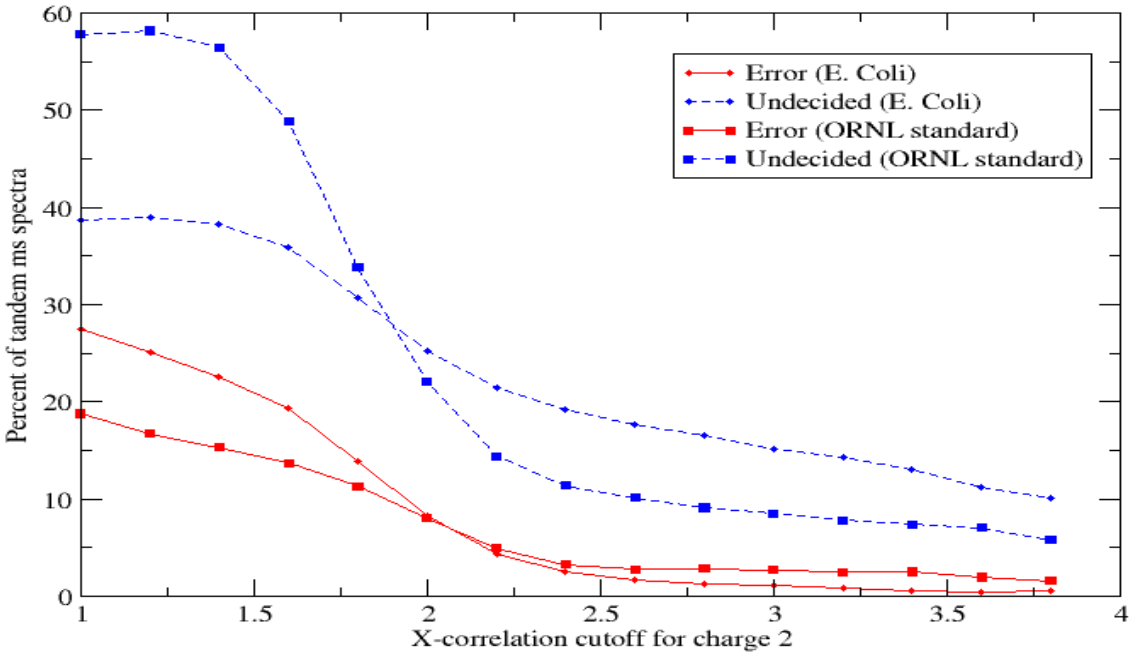

Figure 3.4 Estimation of error rates for charge state assignment in *E. coli* and *ORNL datasets*. The solid curves represent the percent of error made by the charge determining algorithm as comparing to the SEQUEST charge assignment, the dashed lines stand for the percent of spectra that the charge determining algorithm considered unassignable as comparing to the SEQUEST charge assignment**.**

There are two measures of error for the performance of a charge determination algorithm: the number of incorrectly assigned charge states, and the number of spectra for which charge could not be determined. The two characteristics are connected – it can generally be adjusted whether more spectra should be assigned (potentially producing more incorrect assignments), or more unassigned spectra can be allowed (reducing the number of erroneous assignments, but increasing the number of spectra with multiple charge states). The performance of our method was tested against two test sets, the *ORNL dataset* and the *E. coli dataset*, as shown in Figure 3.4 and summarized in Tables 3.2 and 3.3.

The percent of false positive assignments made by SEQUEST is expected to decrease with the increase of X-correlation cutoff. At the low X-correlation cutoffs (1, for charge 2+ peptides and 2 for charge 3+ peptides) both datasets show high error rates: 19% error for the *ORNL dataset* and 27% error for the *E. coli dataset*. However, as expected, when the X-correlation cutoffs are increased reducing the number of false identifications, the percent error of our charge state assignment method decreases dramatically. At the frequently used X-correlation cutoffs of 2.6 for charge 2 and 3.6 for charge 3, the percent errors in charge state determination are 2.8% and 1.8% for *ORNL dataset* and *E. coli dataset* respectively. At the higher X-correlation cutoffs, the decrease in charge state assignment error as a function of X-correlation levels off. At the highest X-correlation cutoff considered: X-correlation 3.8 and 4.8 for charge 2 and charge 3

Table 3.2 Charge assignments for the *ORNL dataset*.

X-correlation	%Error	%Undecided	Charge 2	Charge 3
2.2	5.0	14.4	528	193
2.4	3.3	11.4	475	165
2.6	2.8	10.1	431	141
2.8	2.8	9.1	395	131
3.0	2.7	8.5	363	118
3.2	2.5	7.8	331	104
3.4	2.5	7.4	303	89
3.6	2.0	7.08	275	77
3.8	1.6	5.8	248	62

The first column refers to the presented X-correlation cutoffs for charge 2+ peptides, corresponding X-correlation cutoffs for charge 3+ is greater by 1. Second column shows %error made by charge assignment method, SEQUEST assignments with given X-correlation cutoffs are correct. Third column displays the percent of unassigned spectra, while fourth and fifth columns show the number of SEQUEST assignments for charge 2 and charge 3 peptides in the database.

Table 3.3 Charge assignments for *E. coli dataset*.

X-correlation	%Error	%Undecided	Charge 2	Charge 3
2.2	4.4	21.5	2874	1515
2.4	2.6	19.2	2475	1326
2.6	1.8	17.7	2203	1149
2.8	1.3	16.6	1963	1001
3.0	1.1	15.1	1740	862
3.2	0.8	14.3	1501	740
3.4	0.6	13.1	1273	634
3.6	0.5	11.3	1066	520
3.8	0.6	10.1	875	438

The first column refers to the presented X-correlation cutoffs for charge 2+ peptides, corresponding X-correlation cutoffs for charge 3+ is greater by 1. Second column shows %error made by charge assignment method, SEQUEST assignments with given X-correlation cutoffs are correct. Third column displays the percent of unassigned spectra, while fourth and fifth columns show the number of SEQUEST assignments for charge 2 and charge 3 peptides in the dataset.

peptides respectively, *ORNL dataset* shows 1.6% error, while *E. coli dataset* shows 0.6% error. These final numbers approximate the accuracy of our charge determination as compared to X-correlation cutoffs, since as X-correlation decreases, the less SEQUEST assignment contributes to the percent of error in the analysis. Thus, for the *ORNL dataset*, the charge assignment's error is approximately 1.6% (based on the highest X-correlation cutoffs), and for the *E. coli dataset*, the error is approximately 0.6%.

The percent of unassigned spectra behaves similarly to the percent error, as the X-correlation cutoffs increase. For relatively low X-correlation cutoffs of 1 and 2 for charge 2+ and charge 3+ peptides respectively, the *ORNL dataset* shows 19% unassigned spectra, while the *E. coli dataset* displays 38% unassigned spectra. At the X-correlation cutoffs of 2.6 for charge 2+ and 3.6 for charge 3+, 10% of spectra are unassigned in the *ORNL dataset*, while the *E. coli dataset* shows 18% unassigned spectra. At the X-correlation cutoffs of 3.8 for charge 2+ and 4.8 for charge 3+, the *ORNL dataset* has roughly 6% of spectra with unassigned charge state, and the *E. coli dataset* shows 10% of spectra with unassigned charge states. The spectra that fall under the classification of being undecided will be searched as either charge 2+ and charge 3+ spectra, therefore they will not be lost as identifications as the ones whose charge was assigned incorrectly. The disadvantages to having unassigned spectra are that they increase the search time of the algorithm and that they cause incorrect assignments on the peptide levels. However, it is important not to loose possible identifications by making incorrect charge state assignments, thus as long as the number of unassigned spectra is not overly large (greater than 20%), it is beneficial to open the windows for a few more unassigned spectra rather then cause increase in error rate, leading to unidentified spectra.

Some of the differences between the our method's performance for the two datasets can be explained by the nature of the data: *ORNL dataset*, like *Seattle dataset* is very limited in number of peptides, while *E. Coli* dataset is a real life complex mixture with significantly larger number of peptides in every range of scoring. As shown above, the differences between the datasets are displayed by both the error percent and the percent of unassigned peptides (*ORNL dataset*: error 1.8%, unassigned 5.8%; *E. coli dataset*: error 0.6%, unassigned 10%). These differences point at the different distributions of charge 2+ and charge 3+ spectra as a function of the neural network score (Figure 3.3), which affects the percent error and percent undecided. It is possible to change the neural network score cutoffs for charge 2+ spectra and charge 3+ spectra to produce similar percent of errors and undecided for both datasets.

The algorithm has been trained on the data received from two ion trap instruments from different laboratories (ISB and ORNL). It appears that the performance is best if the neural network is partially trained on the data available from the instrument where the data is analyzed, it is at this time unknown whether it is due to the instrumental differences or the differences in the experiment or the quality of data. At this time, the best performance shown by the charge determination method was in application to real proteomic data rather than the standard datasets where the protein content is assumed as known. In part, the reason for this phenomenon might be caused by the practice of accepting most of SEQUEST's identifications which correspond to the expected proteins, even though such identification might still be coincidental.

The neural network charge state determination method combines an extensive study of features of tandem MS spectrum with the pattern recognition of neural networks.

It embraces a combination of known and new observations in the MS/MS spectra such as examining the number of peaks consistent with parent masses, the spectral densities and the long and short range spectral information. The added features and the new intensity based schemes increase the sensitivity and accuracy of charge state determination, and neural network allows for the normalized score estimating the likelihood of a charge state for a tandem MS spectrum. In the future, this algorithm is easily extensible to the higher charge state models using trivial addition to the feature parameters. This new approach is a promising new method for charge state determination for low resolution mass spectrometry, which can be used to improve the specificity and time of high-throughput mass spectrometric data analysis.

Chapter 4

Computational Identification of Post-Translational Modifications from Shotgun Mass Spectrometry Data.

Some of the data presented below has been presented as Razumovskaya J., VerBerkmoes N., Hurst G., Uberbacher E., poster presentation at ASMS 2005

Introduction

The proteome can be defined as the set of all expressed proteins in a cell, tissue or organism. Proteomic analysis is the product of the need to understand the function of proteins. Proteomics gives us insight into the interactions between proteins, allowing us learn more about the complex network of molecular interactions. Now, as more and more is revealed about proteins, protein structure and function, we find ourselves looking for a deeper knowledge of protein-protein interactions and protein pathways, which gives us a clue to the mechanics of life. The task of identifying and modeling protein pathways is challenging and introduces a great degree of complexity to the studies, as it is a multi-parametric dynamic system. The protein pathways, such as kinase signaling pathway, involve multiple proteins interacting at different times. Additionally, it often involves protein regulation with post-translational modifications, which is essential to the process. Post-translational modifications (PTMs) are covalent processing events that change the properties of a protein by proteolytic cleavage or by addition of a modifying group to one or more amino acids. Far from being mere protein decorations, PTMs of a protein can determine its activity state, localization, turnover and interactions with other proteins. For example, "kinase cascades are turned on and off by the reversible additions and removal of phosphate groups, and in the cell cycle ubiquitination marks cyclins for

destruction at defined time points." (Mann, 2003). The study of PTMs of proteins is, therefore, absolutely essential to facilitate our understanding of cellular processes.

Post-translational modifications

A post-translational modification (PTM) is a modification to a protein which occurs after its translation, causing the protein to appear in altered form from the one originally suggested by its DNA sequence. PTMs can be in a form of proteolytic cleavage such as a signal peptide cleavage, or a covalent addition of various chemical groups to one or more amino acid residues. PTMs are important for protein function: they can control the protein's activity, be related to protein's localization, and have an effect on protein-protein interactions (Mann, 2003). At this time, a large number of different PTMs has been observed in eukaryotic and prokaryotic organisms. RESID, one of the available databases of post translational modifications, reports 330 confirmed PTMs (Garavelli, 2004), but it is expected that the actual number is significantly greater. Some of the more commonly observed post-translational modifications include 1)phosphorylation which is involved in regulation of enzyme activity and signaling; 2)acetylation which affects the protein stability (protection of N-terminus) and regulation of protein-DNA interactions, 3)methylation, regulating of gene expression, 4)acylation, cellular localization and targeting signals, membrane tethering, mediator of protein-protein interactions; 5)hydroxyproline, protein stability and protein-ligand interactions, 6)sulfation, modulator of protein-protein and receptor-ligand interactions; 7)deamidation, possible regulator of protein-ligand and protein-protein interactions, also a common chemical artifact; 8)glycosylation, cell-cell recognition/signaling, reversible, regulatory functions (Mann, 2003). Other PTMs, like disulphide-bond formation seem to be only

involved in protein structure stabilization, or like GPI anchor, in membrane tethering. The main PTMs found in prokaryotes include phosphorylation, methylation, loss of the first methionine, and acetylation. Some PTMs have only been so far observed in eukaryotes, which might suggest their later evolution, or is just related to our current inability to perform a whole proteome PTM analysis.

Methods to measure PTMs

Several approaches have been used to attempt the study of PTMs on the proteomics scale. Some PTMs can be predicted from DNA sequence by computational methods, like signal peptide cleavage, and some by homology to the previously observed proteins (like kinase cascades) in different organisms or pathways. The unknown PTMs are very difficult to detect; many of them have been uncovered by accident during studies of particular proteins, or specific pathways. The study of post-translational modifications in an organism is made difficult by the nature of PTMs: generally, they can only be found on the protein level, the DNA and mRNA do not carry the information about most PTMs. Edman degradation (Edman, 1950) and various mass spectrometry methods have been the most successful to detection. However, Edman degradation involves analyzing pure proteins, which prohibits a high throughput analysis; additionally, the candidates for Edman degradation must come from some prior analysis. Therefore, while the Edman degradation technique is useful for identifying and localizing PTMs on specific proteins of interest or as a confirmation technique, it cannot be practically applied to a whole genome study.

Post-translational modifications are post-processing events that generally change the mass of the proteins from the original mass prescribed by the DNA sequence as is

illustrated in the figure 4.1. Since PTMs affect the mass of a protein, such techniques as 2-D PAGE and mass spectrometry that measure protein mass can be applied to detecting the mass change and characterize present post-translational modification. While the combination of isoelectric point information and molecular weight provided by 2-D PAGE separations has been shown useful for detection of post-translational modifications, 2-D PAGE separations coupled to various mass spectrometric methods provides significant additional improvements (Wilkins, 1999).

Despite the presence of many other mass spectrometric techniques, shotgun bottom up mass spectrometry is one of techniques that are most widely used for a high throughput whole proteome analysis (Pandley, 2000). Using this technology it is possible to analyze the whole proteome under varying growth conditions and at different stages of development to be able to monitor the changes in the dynamics of the proteome. Bottom up proteomics is a method that can be used to detect PTMs in a comparatively high-throughput fashion in a whole proteome data. However, mass spectrometry requires computational algorithms to evaluate and interpret the measured information and detection of post-translationally modified proteins is a challenging problem in terms of current computational technology.

The three types of software tools that are currently applied to MS data interpretation and are used for PTM analysis are *de novo* and hybrid algorithms and database searches. At the moment, the *de novo* and hybrid algorithms are not up to the standards of high-throughput proteomic data interpretation, in general, exhibiting high rates of incorrect identifications. In many cases, these approaches are limited by the

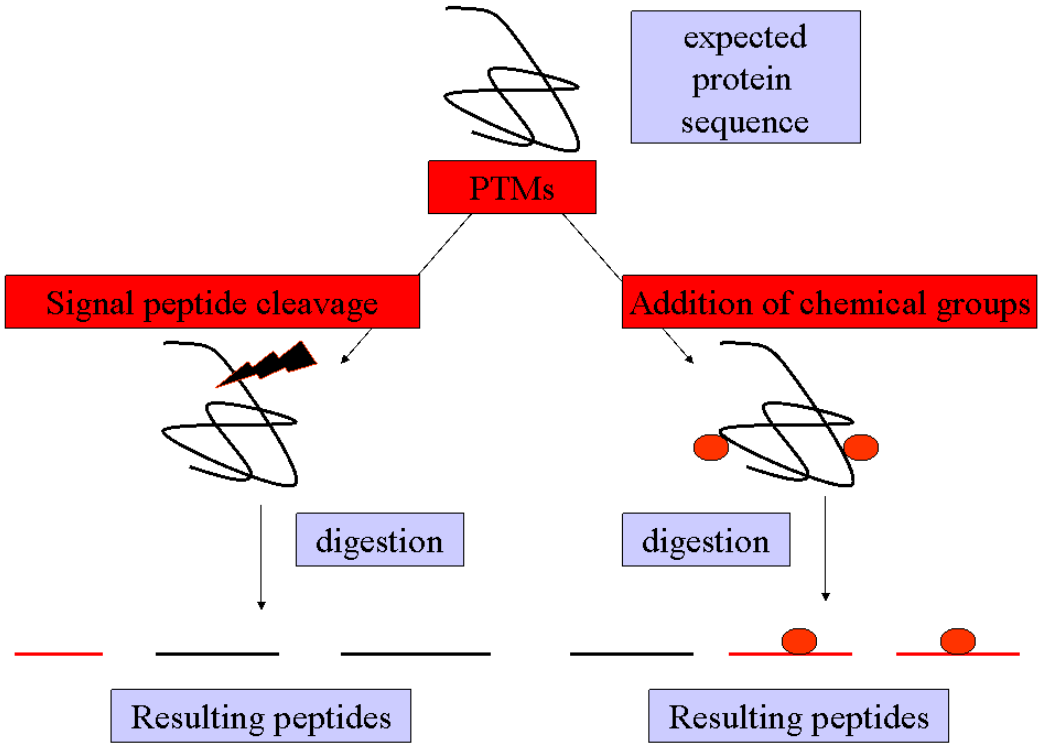

Figure 4.1 Post translational modifications. Example of the effect of PTM on protein sequence and it's impact on bottom up mass spectrometry analysis. The peptides marked in red contain change in mass that can only be detected by special analysis considering PTMs.

quality of data produced by MS instruments as they are more sensitive to the quality of data than the database searches. Thus, though they are not inherently suited for PTM detection, for proteomic experiments on ion trap instruments, database searches currently remain the method of choice.

As previously mentioned, the peptide identifications made by database search algorithms are based on comparisons between the fragmentation patterns of theoretical and experimental peptides, where theoretical fragmentation patterns are derived *in silico* from the available database and experimental fragmentation patterns are measured by

mass spectrometer from a peptide found in a sample. Database searches are at the moment considered the most reliable of the MS data interpretation tools for the shotgun proteomic experiments, however, they are extremely sensitive to database deviations, any peptides inconsistent with the database by mass or sequence are unidentifiable by this method. Since PTMs change the mass of the peptides, these deviations from the database are generally fatal, as they often prevent placing the appropriate peptide into the candidate list formed by parent mass. Such inconsistency between the database mass and the peptide present in the sample is illustrated above in the figure 4.1 which outlines two avenues for PTM occurrences: the simple signal peptide cleavage and the addition of two covalent modifications to the sequence. In either of these cases, the masses of affected (modified) peptides are inconsistent with the masses present in the protein database causing errors in identifications. It is noted that with the current methods of peptide identification only 10-30% of all tandem MS spectra in a proteomics sample are identified (the rest are either discarded, or receive such low scores that they cannot be considered to be correct answers). While some of the spectra could be of inferior quality or carry single amino acid substitutions, it is currently unknown what percentage of the remaining 70% of peptide spectra may contain post translational modifications. The attempt to analyze PTMs is built into database search algorithms. As proposed by Yates et al (Yates, 1995), the PTM analysis performed by database searches involves placing all PTMs of interest into all possible places in peptide sequences. This approach to analyze PTMs leads to a large combinatorial problem, even for a small number of PTMs; additionally Yates et al. indicated that "extending this approach to a larger set of modifications is an open problem". This approach is not up to the task in terms of speed

and accuracy of the identifications; in addition it relies entirely on the user to provide a limited list of PTMs to search during each run.

The database search algorithms are currently considered the most robust system for peptide identification, but they are not inherently suited to solving the problem of identification of PTMs, as they have to rely on the database sequences to be exact representations of the parent peptide. The only way to circumvent the problem of PTM detection by database search lays in the organism specific database annotation. As of now, with an exception of ProSight PTM (LeDuc, 2004) designed for intact protein analysis, there is no organism specific PTM database annotation, and without a coherent strategy of PTM identifications, the database size quickly becomes unmanageable, increasing the time of analysis and the number of false identifications. In addition 40-60% of genes in current genome annotations are hypothetical proteins (Blattner, 1997; Fraser, 1995; Heidelberg, 2002; Larimer, 2004) which may never be expressed in the organism, but can add an enormous number of false identifications. In many cases PTMs identified by database searches must be manually confirmed as to accuracy and biological significance of the identification (how likely a particular protein is to carry a certain type of post-translational modification based on prior studies). While further experiments must be performed to corroborate the presence of the PTM detected by shotgun bottom up mass spectrometry, an additional insight to the accuracy of detection is invaluable to target further studies.

One of the major sources for incorrect identifications of peptides by database search algorithms is due to the increasing of the database size, which causes an increase in the number of candidate sequences. As the result of many similar candidate sequences

present in the database, many of them might produce comparable scores. A thorough search for PTMs increases the database sizes dramatically: for example, a possible phosphorylation, which in principle may affect any tyrosine, serine or threonine (and in case of prokaryotic organism histidine) in the sequence, can easily increase a database hundreds of times, since every peptide containing any of these common residues may be modified once, or multiple times depending on the number of these residues. In addition to the increasing number of false positives, such thorough search can lead to days of computational time and all the results should be manually confirmed to ensure that they are biologically sound. All of these factors severely limit the ability of current database search algorithms to perform a comprehensive all proteome PTM detection, leaving us with a capability of detecting no more then a few PTMs at a time while the likelihood of correct detection is frequently uncertain. The algorithmic approach presented in this chapter guarantees that all the detected PTMs are biologically sound, by utilizing all the current knowledge about post-translational modifications. This approach provides a new way for PTM driven database annotation, which includes the combination of the prior knowledge of PTM carrying domains in multiple organisms, homology inference and basic PTM prediction software.

Rhodopseudomonas Palustris is a prokaryotic microbial organism commonly found in water and soil. It is equipped to endure an extensive range of growth conditions: aerobic to anaerobic, as well as dark to light as illustrated in the Figure 4.2.

In the light conditions, the bacteria are capable to convert light into cellular energy. In the absence of oxygen it converts atmospheric CO_2 into biomass. It is also

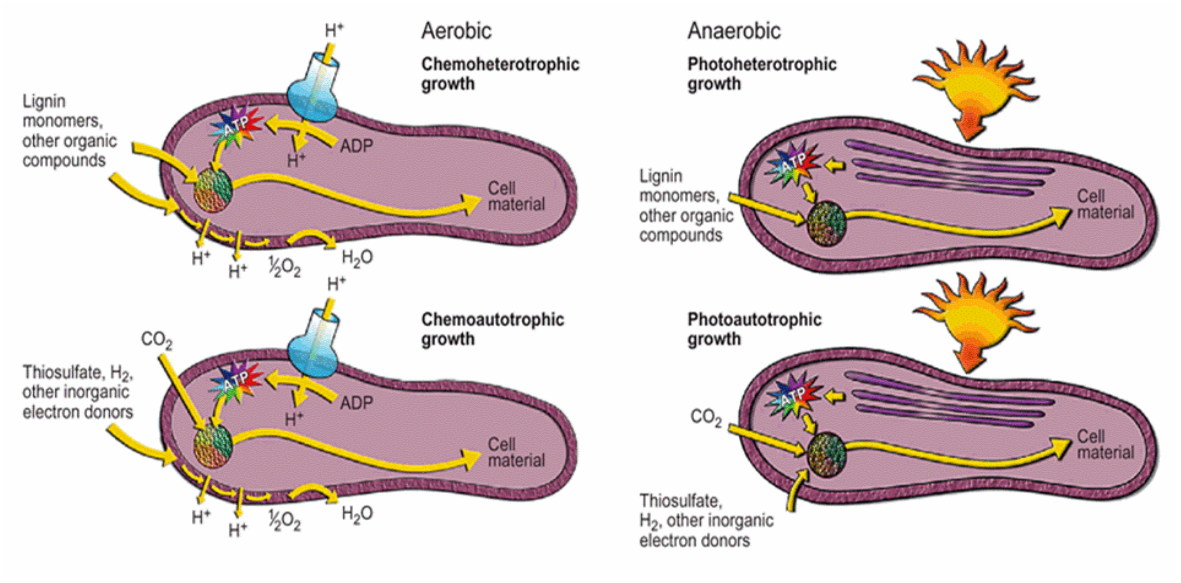

Figure 4.2 ***R. palustris*** **growth conditions.** The four metabolic growth conditions for *R. palustris* are aerobic, anaerobic, chemoautotrophic and photoautotrophic. The PTM analysis was applied to each of these growth conditions. Adapted from Larimer et al. 2004.

capable of degrading organic compounds including toxins like 3-chlorobenzoate building blocks. In the presence of oxygen, *R. palustris* generates energy by degrading carbon containing compounds (such as sugars, lignin-monomers, etc) and carrying out respiration. *R. palustris* is one of the most metabolically versatile bacteria described at this time. *R. palustris* was recently sequenced (Larimer, 2004), revealing the genome sequence of the 5,459,213 base pair circular chromosome, with 4,836 predicted genes. The careful study of the predicted genes showed that 31% of genome is devoted to energy metabolism and cellular processes, 14.5% of genome to transport, 4.7% to signal transduction. About 3.5% of the genome may carry PTMs as showed by the preliminary analysis based on COGs (Tatusov, 1997) homology comparison of functional domains with multiple organisms. Due to its ability to degrade toxic compounds, *R. palustris* became an organism of interest for Department of Energy (DOE) and multitude of mass spectrometric data is being collected for the analysis of the organism's proteome (VerBerkmoes et al., 2005) (http://compbio.ornl.gov/rpal_proteome/) and protein complexes.

In this chapter a new algorithmic approach to PTM detection is introduced and applied to analyzing *R. palustris* under a variety of metabolic growth conditions. The new PTM discovery driven approach for shotgun bottom up MS data interpretation is focused on utilizing all of the current knowledge about post-translationally modified proteins to address the limitation of current database search algorithms in terms of PTM detection. The methodology for designing PTM annotated protein domain library is an extension to the one proposed in the ProSight PTM introduced by LeDuc in (LeDuc, 2004) for intact protein analysis ("top down" analysis (McLuckey, 1998)), which was

successfully applied to the yeast PTM analysis by Meng et al (Meng, 2004), here applied to high-throughput bottom up MS data. Our new PTM detection approach uses all the current knowledge of PTM proteins and protein domains to annotate *R. palustris* sequence database using the newly developed algorithmic approach PTMsearch performs proteomic PTM analysis for bottom up MS of *R. palustris*.

Materials and Methods

Rhodopseudomonas Palustris sample preparation

All datasets were kindly provided by VerBerkmoes et al (VerBerkmoes et al, 2005) and their generation is briefly described below.

Cell growth, production of protein fractions and proteome digestion

R. palustris strain CGA010, a hydrogen-utilizing derivative of the sequenced strain (unpublished C.S. Harwood) and referred to here as the wild-type strain, was grown under the six conditions (photoheterotrophic, chemoheterotrophic, photoautotrophic, photoheterotrophic with nitrogen fixation, photoheterotrophic with benzoate as a carbon source). All cultures were grown anaerobically in light or aerobically in dark, with shaking in 1.5 liters of defined mineral medium at 30°C to mid-log phase ($OD_{660nm} = 0.6$). All anaerobic cultures were illuminated with 40 or 60 W incandescent light bulbs. Carbon sources were added to a final concentration of 10 mM succinate (for all growth modes except benzoate and photoautotrophic), 3 mM benzoate (benzoate growth) or 10 mM sodium bicarbonate with H_2 gas in the head space (photoautotrophic growth). For the photoheterotrophic N_2 fixing cultures, ammonium sulfate was replaced by sodium sulfate in the culture medium and N_2 gas was supplied in the head space.

Cell extracts were prepared as follows: cells were harvested by centrifugation, washed twice with ice-cold wash buffer (50 mM Tris-HCl buffer (pH 7.5) with 10 mM EDTA) and resuspended in ice-cold wash buffer. Cells were then lysed by sonication and unbroken cells were removed with low-speed centrifugation (5,000 g x 10 min). Four proteome fractions were created from this cellular extract by ultracentrifugation (100,000 g for 1 h led to membrane and crude fractions; this supernatant was then further centrifuged at 100,000 x g for 18 h leading to pellet and cleared fractions). All four proteome fractions were analyzed as below.

Proteome fractions from each growth state were processed by the same protocol: Briefly, proteome fractions were denatured, reduced, digested with sequencing grade trypsin and de-salted by solid phase extraction.

LC-ES-MS/MS analysis

The four proteome fractions from each growth state were analyzed in duplicate via multiple one-dimensional LC-ES-MS/MS experiments performed with an Ultimate HPLC (LC Packings, a division of Dionex, San Francisco, CA) coupled to an LCQ-DECA or LCQ-DECA^{XPplus} quadrupole ion trap mass spectrometer (Thermo Finnigan, San Jose, CA). To increase dynamic range in the 1D-LC-ES-MS/MS analysis, separate injections were made with a total of 8 overlapping segmented m/z ranges scanned (referred to as gas phase fractionation or multiple mass range scanning). These entire datasets were used in this study.

PTM fragmentation

The PTMs that modify a protein by covalent additions of chemical groups can be classified into three basic categories based on the binding strength of PTM to the peptide and their impact on the MS fragmentation pattern. The modifications can be "extremely labile", "labile" and "stable"(Figure 4.3), based on their behavior during the analysis as mentioned by Mann et al. While is frequently difficult to predict the full impact of modification on the tandem MS fragmentation, the figure illustrates how different types of modifications might affect the tandem MS fragmentation pattern of a peptide, making the analysis of modified peptides more difficult due to the inconsistencies between the experimental tandem MS spectrum of a modified peptide and the theoretical spectrum of a candidate peptide.

1. The *"extremely labile" PTMs* (such as serine/threonine phosphorylations) have a propensity to fall off the peptide during CID very easily, creating a dominating fragment ion in the tandem MS spectrum. The spectra with this type of modification generally take the form of a single major peak at Precursor mass – PTM, as illustrated in Figure 4.3a.
2. The *"labile" PTMs* are significantly more stable during the fragmentation than the "extremely labile" PTMs, only some proportion of the PTM falls off during fragmentation of parent ion, causing a potential fragment ion at the mass of Precursor mass – PTM . Additionally the 'b' and 'y' fragment ions also proceed to lose some proportion of the PTM causing satellite peaks to the 'b' and 'y' ions in the form of 'b' ion - PTM and 'y' ion - PTM ions in the fragmentation pattern, as illustrated at the Figure 4.3.b.

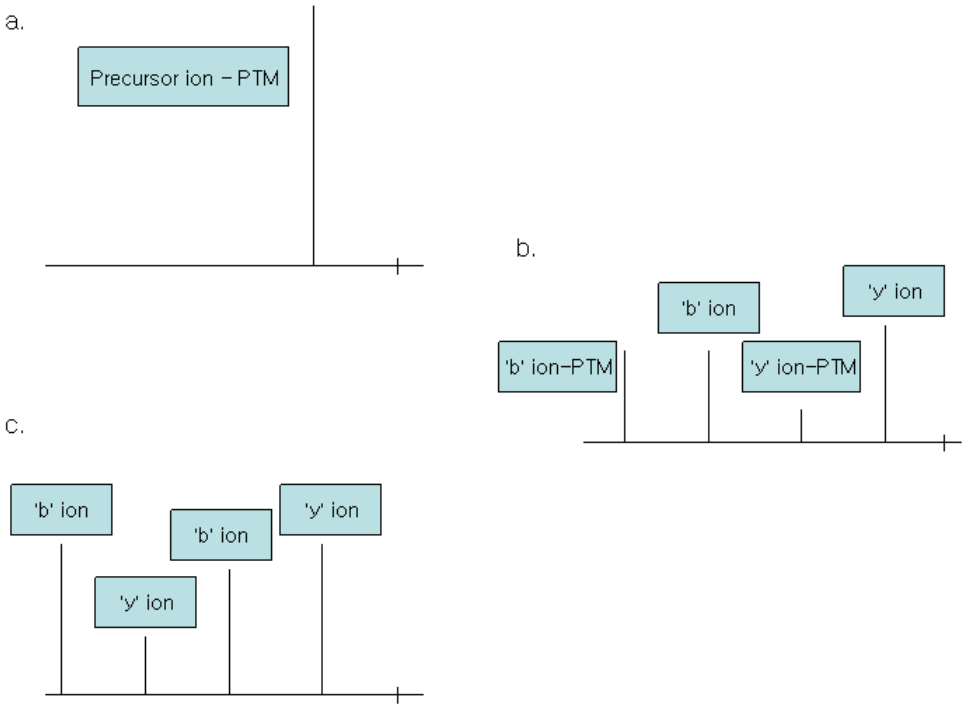

Figure 4.3 Effect of post-translational modifications on fragmentation patterns. The examples of tandem MS spectra for the three types of PTMs: a. "extremely labile" PTM, b. "labile" PTM, and c. "stable" PTM.

3. The *"stable" PTMs* tend to be firmly attached to the peptide. The bond between PTM and side chain is not easily breakable. As shown at the Figure 4.3.c, the spectrum does not have any indication of the presence of a post translational modification as in previous 2 cases since the PTM does not fall off the peptide during the fragmentation. The tandem spectrum offers the same type of sequencing information that is present in spectra with no PTMs.

For some of the PTMs, the fragmentation behavior is already known, and some have not yet been observed through mass spectrometry. These different types of PTM behavior tend to significantly alter the characteristics of the theoretical mass spectrum, and can

create considerable difficulties for PTM detection with conventional database scoring schemes.

Building PTM library

Over the past decades, the relationships between protein function and post translational modifications have been observed in different organisms. The fact that presence of a PTM in a protein is expected to have functional significance, leads to a conclusion that homologous proteins are likely to have the same PTMs. Based on this assumption, we use the homology between proteins to predict a presence and localization of a PTM as it is done for protein functional annotation. In order to detect biologically important post translational modifications, we created a comprehensive library of proteins that have been documented to carry PTMs in multiple organisms – PTMLib. RESID database (Garavelli, 2004) is a comprehensive collection of annotations and structures for protein modifications including amino-terminal, carboxyl-terminal, peptide chain cross-link and many other PTM types. Currently RESID database contains over 330 residues either predicted or observed in proteins arising through natural modifications of encoded amino acids, which include N-formyl methionine, selenocysteine and pyrrolysine. The creators of RESID database focused on creating a database which documents experimentally detected post translational modification, the protein sequences where they were found and the literature references.

The PTMs documented in the database also contain references to PIR and/or Swiss-Prot sequences where they were discovered. This database is continuously increasing as more data is being collected, the new experimentally verified entries of PTMs are added at an average of 15 per year. Each RESID database entry, as shown in

Figure 4.4 presents a chemically unique modification and shows how that modification is currently annotated in the protein sequence databases, Swiss-Prot (Farriol-Mathis, 2004) and the Protein Information Resource (PIR). The database is becoming an invaluable tool for further studies of post translational modifications. The RESID database is available at http://pir.georgetown.edu/pirwww/dbinfo/resid.html.

The post-translationally modified proteins documented in the RESID database, the information about their corresponding PTMs and the site specificities were extracted from RESID database and used for the creation of global PTM annotated protein library PTMLib. Each entry in the PTMLib database specifies a protein, the post translational modification it can carry and the annotated possible PTM target sites. The final PTMLib database includes 223 types of post-translational modifications in 326 different proteins.

PTM annotation of *R. palustris* was performed based on the protein similarity allowing location of potentially modified sites for a large number of post-translational modifications with experimental proof. The annotation of *R. palustris* for putative post-translationally modified proteins was achieved by running BLAST (Altschul, 1990; Altschul, 1997) searches between the proteins present in PTMLib and the sequences found in the *R. palustris* genome. All sequences with BLAST E-score $< 10^{-3}$ were included into the candidate PTM sequences. The PTM annotated *R. palustris* database contains 287 proteins annotated for 220 different post-translational modifications.
The PTM target sites were annotated by the PROSITE pattern search (Hofmann, 1999; Falquet, 2002). PROSITE is a search tool, which using a large collection of biologically meaningful PTM signatures is designed to detect short PTM patterns in the given

AA0036	3'-phospho-L-histidine	L-histidine	Fc=217.12, Fp=217.0252, Cc=79.98, Cp=79.9663,	phosphohistidine phosphoprotein	PIR:Active site: His (phosphohistidine intermediate) PIR:Binding site: phosphate (His) (covalent) PIR:Binding site: phosphate (His) (covalent) (by ...) SP:ACT_SITE PHOSPHOHISTIDINE INTERMEDIATE SP:MOD_RES PHOSPHORYLATION SP:MOD_RES PHOSPHORYLATION (AUTO-)
AA0037	O-phospho-L-serine	L-serine	Fc=167.06, Fp=166.9983, Cc=79.98, Cp=79.9663,	phosphoprotein	PIR:Active site: Ser (phosphoserine intermediate) PIR:Binding site: phosphate (Ser) (covalent) PIR:Binding site: phosphate (Ser) (covalent) (by ...) SP:ACT_SITE PHOSPHOSERINE INTERMEDIATE SP:MOD_RES PHOSPHORYLATION SP:MOD_RES PHOSPHORYLATION (AUTO-)
AA0038	O-phospho-L-threonine	L-threonine	Fc=181.08, Fp=181.0140, Cc=79.98, Cp=79.9663,	phosphoprotein	PIR:Binding site: phosphate (Thr) (covalent) PIR:Binding site: phosphate (Thr) (covalent) (by ...) SP:MOD_RES PHOSPHORYLATION SP:MOD_RES PHOSPHORYLATION (AUTO-)
AA0039	O4'-phospho-L-tyrosine	L-tyrosine	Fc=243.15, Fp=243.0296, Cc=79.98, Cp=79.9663,	phosphoprotein	PIR:Binding site: phosphate (Tyr) (covalent) PIR:Binding site: phosphate (Tyr) (covalent) (by ...) SP:MOD_RES PHOSPHORYLATION SP:MOD_RES PHOSPHORYLATION (AUTO-)
AA0040	2'-[3-carboxamido-3-(trimethylammonio)propyl]-L-histidine	L-histidine	Fc=280.35, Fp=280.1773, Cc=143.21, Cp=143.1184,	diphthamide	PIR:Modified site: 2'-[3-carboxamido-3-(trimethylammonio)propyl]histidine (His) SP:MOD_RES DIPHTHAMIDE
AA0041	N-acetyl-L-alanine	L-alanine	Fc=114.12, Fp=114.0555, Cc=42.04, Cp=42.0106,	acetylated amino end	PIR:Modified site: acetylated amino end (Ala) PIR:Modified site: acetylated amino end (Ala) (in mature form) SP:MOD_RES ACETYLATION
AA0042	N-acetyl-L-aspartic acid	L-aspartic acid	Fc=158.13, Fp=158.0453, Cc=42.04, Cp=42.0106,	acetylated amino end	PIR:Modified site: acetylated amino end (Asp) PIR:Modified site: acetylated amino end (Asp) (in mature form) SP:MOD_RES ACETYLATION

Figure 4.4 Example entries in the RESID database.
The database entry shows the name, molecular weight of the modification, keyword under which they are frequently found, and reference to the protein(s) where PTM was found. It also contains the protein sequences, the PTM binding sites and references to the journal articles describing the evidence of the PTM's presence.
http://pir.georgetown.edu/pirwww/dbinfo/resid.html.

sequence**.** While in many cases, PROSITE tends to give a greater number of possible modification sites than can be expected, when the database is limited to the proteins with expected PTMs and the PTM types are known, the number of PROSITE predictions is quite tractable. PROSITE is freely available at: http://pir.georgetown.edu/pirwww/search/patmatch.html. The new *R. palustris* PTM annotated database improves searches against multiple post-translational modifications both in terms of search time and management of number of false positive identifications.

The PTM detection database search approach

The new approach to PTM analysis involved building a search engine that would allow searching for any number of post-translational modifications (number of modifications searched for in *R. palustris* was 220) at a time. A new database search approach (PTMsearch) was developed in order to perform the searches for post-translational modifications in *R. palustris*. The PTMsearch was modeled after the widely used database search algorithm, SEQUEST (Eng, 1994), with the basic peptide parent mass used as a filter for candidate peptides and the use of cross-correlation scoring scheme based on the published SEQUEST's scoring scheme, X-correlation. The PTMsearch differs from SEQUEST by allowing only annotated PTMs in a given sequence as specified in the database, and allowing only one expected PTM in a peptide sequence at one time. In addition, PTMsearch is extended to detect PTMs receiving relatively low scores, since some of PTMs can significantly alter the appearance of a fragmentation pattern. Thus, PTMsearch is capable of incorporating in the further analysis all the candidate peptides including the ones that do not appear with the top ten scores, as it is done in SEQUEST, which improves the sensitivity of the method for PTM

detection. The use of PTMsearch greatly reduces the number of runs that have to be made by SEQUEST to perform the analysis since only SEQUEST run allows for a maximum of 3 post-translational modification at one time, it reduces the number of false identifications made by SEQUEST when any number of modifications is allowed in one peptide and allows to detect peptides with lower scores.

PTMsearch approach ensures that all the detected PTMs are found in expected sequences, however in a novel approach, additional measures had to be taken to improve the likelihood of the identifications. Therefore, an additional filtering approach was introduced to increase the confidence of the detected post-translational modifications. The filtering scheme for PTM detection in *R. palustris* by PTMsearch was set up to involve a set of conditions, which are used to accept or reject an identification made by PTMsearch. The conditions are established to reduce the number of false positive results while retaining most of the reliable PTM detections. The conditions for accepted PTM detection include: 1) appropriate scoring cutoffs, 2) growth conditions, 3) number of occurrences.

The scoring cutoff condition refers to the range of PTMsearch scores that can be accepted for peptide identification. Though the scoring scheme used in the PTMsearch is modeled after the X-correlation, it is not an exact replica of X-correlation and while there is a direct correlation between X-correlation and the PTMsearch scoring scheme, the scoring cutoffs for confident identification are not as well explored as for X-correlation. In addition to this consideration, as mentioned above in a section on "PTM fragmentation", the "labile" and the "extremely labile" PTMs produce a different fragmentation pattern than the expected fragmentation pattern described in the

"Fragmentation pattern" section of introduction. Because of these factors, the scoring scheme behaves in an unpredictable manner when detecting post-translational modifications and the score cutoff conditions were lowered to include most of the identifications, including the identifications which are not the top candidate peptide.

The growth condition cutoff refers to identifying PTMs only under appropriate *R. palustris* growth conditions when they are known. As an example, uridylation is a post-translational modification that is known to be present during the nitrogen-fixing conditions in P-II proteins. The only accepted identifications containing uridylation in P-II proteins were made under nitrogen-fixing growth conditions. Unfortunately, in many cases, the conditions in which a PTM is expected to be present are not definitively known or they have only been detected in a limited set of conditions. This condition is only used if the number of detected PTMs is unmanageably large since it can be extremely limiting. It is suggested to be used only as a final confirmation of the PTM presence after additional experiments have been performed.

The number of occurrences condition refers to the number of times the peptide with the PTM has been detected by PTMsearch. The repetitive detection of a peptide increases the chances of the peptide's presence. This condition was set to 3 or more occurrences to be required for accepted peptide detection. In general at least 2 of the occurrences are expected to be made under the repetition of the same growth condition. This condition does not ensure that peptide identification has been made correctly since the same fragmentation pattern can be misidentified several times. However, it does ensure that the stable fragmentation pattern has been detected multiple times, which decreases the likelihood of detection of a noise spectrum.

Results and Discussion

Post-translational modifications frequently serve as regulatory switches to protein activities, they are capable of changing protein properties to influence transcription, translation, ligand-binding interaction and many other cell processes, thus playing a crucial part in the life of organisms and are extremely important to our understanding of biological processes. While there have been many studies of post-translational modifications in eukaryotic organisms and many interesting PTMs were discovered by a variety of methods, there have not been as many attempts for the whole organism study of PTMs in prokaryotes. Mass spectrometric instrumentation is uniquely qualified for high-throughput PTM detection both because of its capability to measure the cell proteome content and its ability to detect mass differences which generally accompany post translational modifications. In this chapter, PTMsearch, a new algorithmic approach for high-throughput PTM detection by bottom up mass spectrometry is introduced. In addition, a whole proteome study of post-translational modifications in *R. palustris* is performed with the use of available biological information.

R. palustris is a metabolically versatile prokaryotic organism it is expected to be highly regulated on the proteomic level. The availability of *R. palustris* MS proteomic data for a range of different growth conditions inspired an effort to attempt the detection of a range of post-translational modifications in the organism. For the purpose of this experiment *R. palustris* was grown under five different metabolic conditions. The growth conditions included in the study were chemoheterotrophic, photoheterotrophic, photoautotrophic, photoheterotrophic grown in benzoate medium and photoheterotrophic nitrogen fixation; each of the conditions were analyzed at least twice to ensure the quality

of data. In part, the project was driven by the hope to uncover some of the eukaryotic post-translational modifications which could potentially be present in *R. palustris* and have never previously been detected in prokaryotic organisms. As the result of the computational analysis, 228 different peptides with post-translational modifications were detected in 85 proteins (not all of the peptides were unique to one protein), however, all of the computational data must be subjected to a set of experiments in order to be confirmed. Thus, most of the results of this study are potential candidates for further biological studies. However, so far two of the PTMs detected by this methodology have been confirmed by a separate top down MS experiment.

A PTMLib database was built to incorporate all the protein sequences with documented post-translational modifications presented in Garavelli's RESID database. Based on RESID, 223 different types of PTMs and 326 protein sequences were included in the final PTMLib database. With the use of the PTMLib database, *R. palustris* sequence database was pre-annotated for the possible 223 types of PTMs using BLAST similarity search. All the *R. palustris* sequences with a BLAST's E-value greater than 10^{-3} were included in the *R. palustris* PTM database which resulted in 287 proteins with 220 possible PTMs. The target sites were then annotated with the use of PROSITE software. All of the collected MS data for *R. palustris* was then searched against the annotated *R. palustris* database using new PTMsearch algorithm in order to detect the presence of post-translational modifications.

PTMsearch is a generic database search algorithm based on SEQUEST's X-correlation scoring scheme, coupled with filtering procedures designed to decrease the amount of false positive identifications. As a result of analyzing the data from

VerBerkmoes et al. 2005 with an exception of the stationary growth phase and lhaA mutant growth phase, 228 unique peptides were identified as potential PTM carriers, with 29 distinct types of post translational modifications. As required by the filtering procedure each of these peptides had to be detected at least 3 times during the analysis before they were included in the final list of modifications. The detected post-translational modifications include such PTMs as phospho-uridylylation, biotinyl, lipoyl, acetyl, methyl, dehydroalanine, and carboethyl modifications and many others. The list of all peptides with candidate PTMs can be found in the appendix 1, and the list of all the proteins can be found in the appendix 2. However, as previously stated, even though, these modifications can be considered biological sound in terms of their protein localization, there are still many factors that could have caused incorrect identifications. These detected peptides should, therefore, be experimentally verified, while they can now be considered only as potential candidates for further studies rather than definite identifications. The potential verifications can be done with the use of multiple enzyme cleavage followed by LC-MS-MS (MacCoss, 2002), top down mass spectrometry, or Edman degradation.

The spectra for two of the detected post translational modifications are presented below. These modifications are uridylation and lipoylation. Uridylation was represented in three tandem MS spectra under one growth condition, while lipoylation appears in seven spectra and is found under a variety of different growth conditions. Both of these modifications are not typically searched for in high-throughput proteomic data by typical database search engines.

An uridylation of P-II protein is an important regulatory signal modification which serves as a regulator of the nitrogen metabolism in many organisms such as *Rhodospirillum rublum*, *Escherichia coli*, *Rhodopseudomonas palustris*. Uridylation is a reversible modification: under the conditions of nitrogen excess P-II proteins are unmodified, while when the nitrogen concentration is low P-II proteins become uridylated. The modified form of P-II is considered to be a signal of nitrogen starvation (Atkinson, 1994). Thus, the presence of uridylated P-II protein are only expected during nitrogen limiting conditions, while in all other growth conditions P-II protein are expected to be unmodified.

The PTMLib derived from RESID database contains a sequence documented to carry uridylation. The sequence blasted against *R. palustris* database yielded three protein sequences: RPA2066 glnB nitrogen regulatory protein P-II 3360442:3360780 forward with E-value of 9e-37 (NREF entry number is NF01165177), RPA0272 glnK1 GlnK, nitrogen regulatory protein P-II 300253:300591 forward with E-value of 1e-34 (NREF entry number is NF01528520); RPA0274: glnK2 GlnK, nitrogen regulatory protein P-II 302307:302645 forward with E-value of 2e-33 (NREF entry number is NF01528171). The tandem MS spectrum of the uridylated version of peptide GAEYAVSFLPK from RPA0274 protein is shown in the figure 4.5. The uridylation site predicted by PROSITE for RPA0274 sequence was Y**!**RGAEY**!**, with tyrosine as the potential modification binding site, "**!**" sign in the sequence denotes all the possible modification sites. As it is shown in the figure 4.5, the detected modifications site is EY**!**A. However, only one of the proteins was detected in the proteomic data, the peptide with expected modification site appearing in both modified and unmodified form.

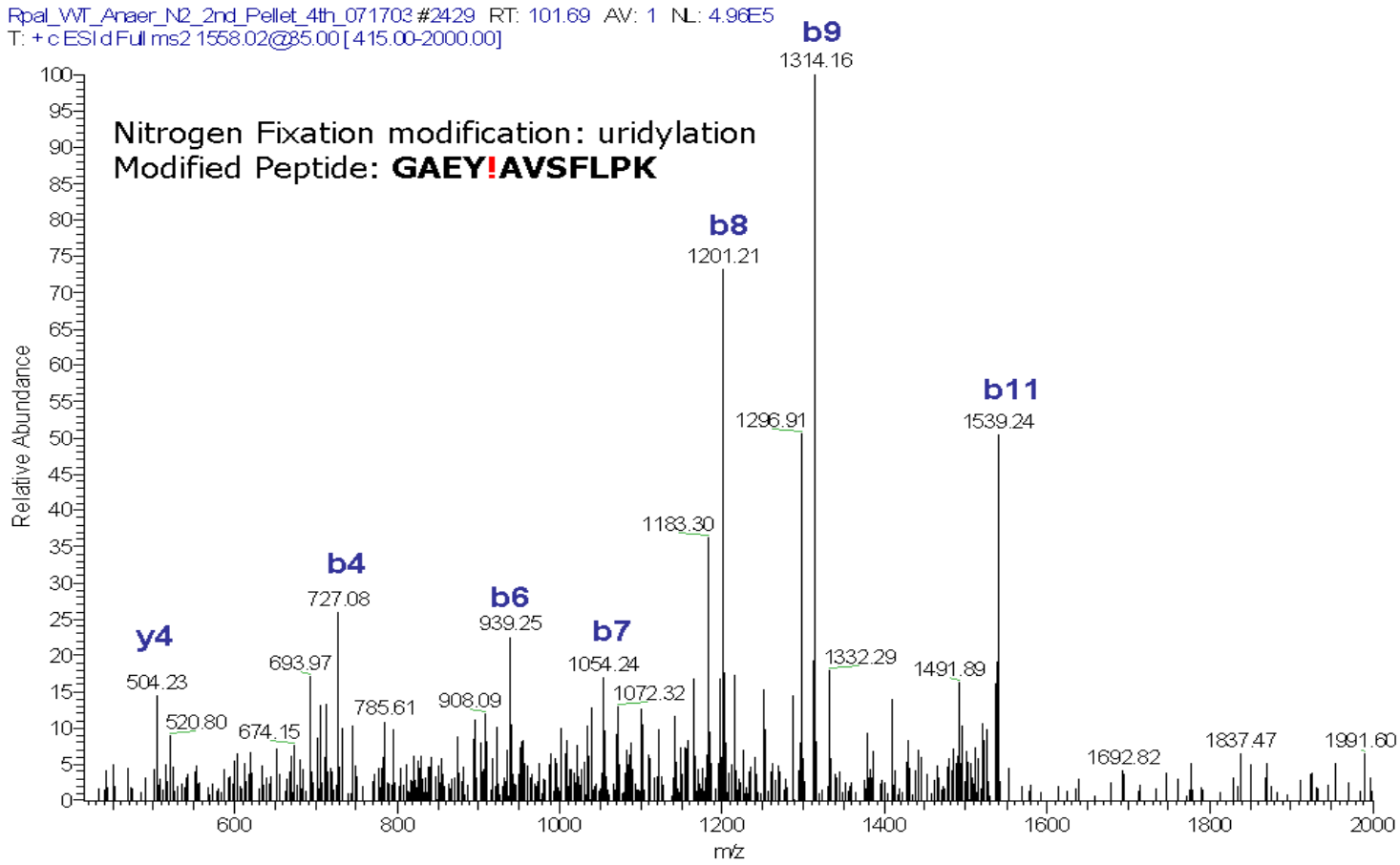

Figure 4.5 The tandem MS spectrum of uridylated peptide. A representative spectrum of modified peptide GAEYAVSFLPK is presented, showing the labeling of consecutive 'b' ions.

As expected the modified version of the peptide appeared under the nitrogen limiting conditions, while no unmodified peptide version was detected. No modified version of the peptide was detected in other growth conditions.

One of the other detected post-translational modifications is lipoylation, which mediates the transfer of electrons and activated acyl groups resulting from the decarboxylation and oxidation of α-keto acids within the complexes. It is expected to be present in the pyruvate dehydrogenase complex, which consists of three proteins: a pyruvate decarboxylase, a dihydrolipoyl acetyltransacetylase, which contains α-lipoic acid covalently bonded through amide linkage with lysine residue, and a dihydrolipoyl dehydrogenase. (most of the information is adapted from Zubay, Biochemistry IIIrd edition.)

The PTMLib contains four sequences homologous to lipoyl carrying proteins documented in RESID database. The proteins are RPA2864 dihydrolipoamide acetyltransferase 3241258:3242649 reverse MW:48330 (NF01529250), RPA3849 glycine cleavage system protein H 4348444:4349967 forward MW:12927, (NF01528984), RPA0188 sucB dihydrolipoamide succinyl transferase 208123:209376 reverse MW:111339 (NF01530219), and RPA2866 pyruvate dehydrogenase E1 beta subunit 3242963:3242958 reverse MW:11042 (NF01532208). However, the only protein with lipoylated peptide detected was RPA2864, the peptide sequence being SGDVIAEIETDK**!**ATMEVEAADEGTLAK, one of the tandem MS spectra for the peptide is shown in the figure 4.6. The lipoylated sequence motif predicted by PROSITE is: GDK**!**VK**!**SGDVIAEIETDK**!**ATMEVEAADEGTL, with lysine as specified binding

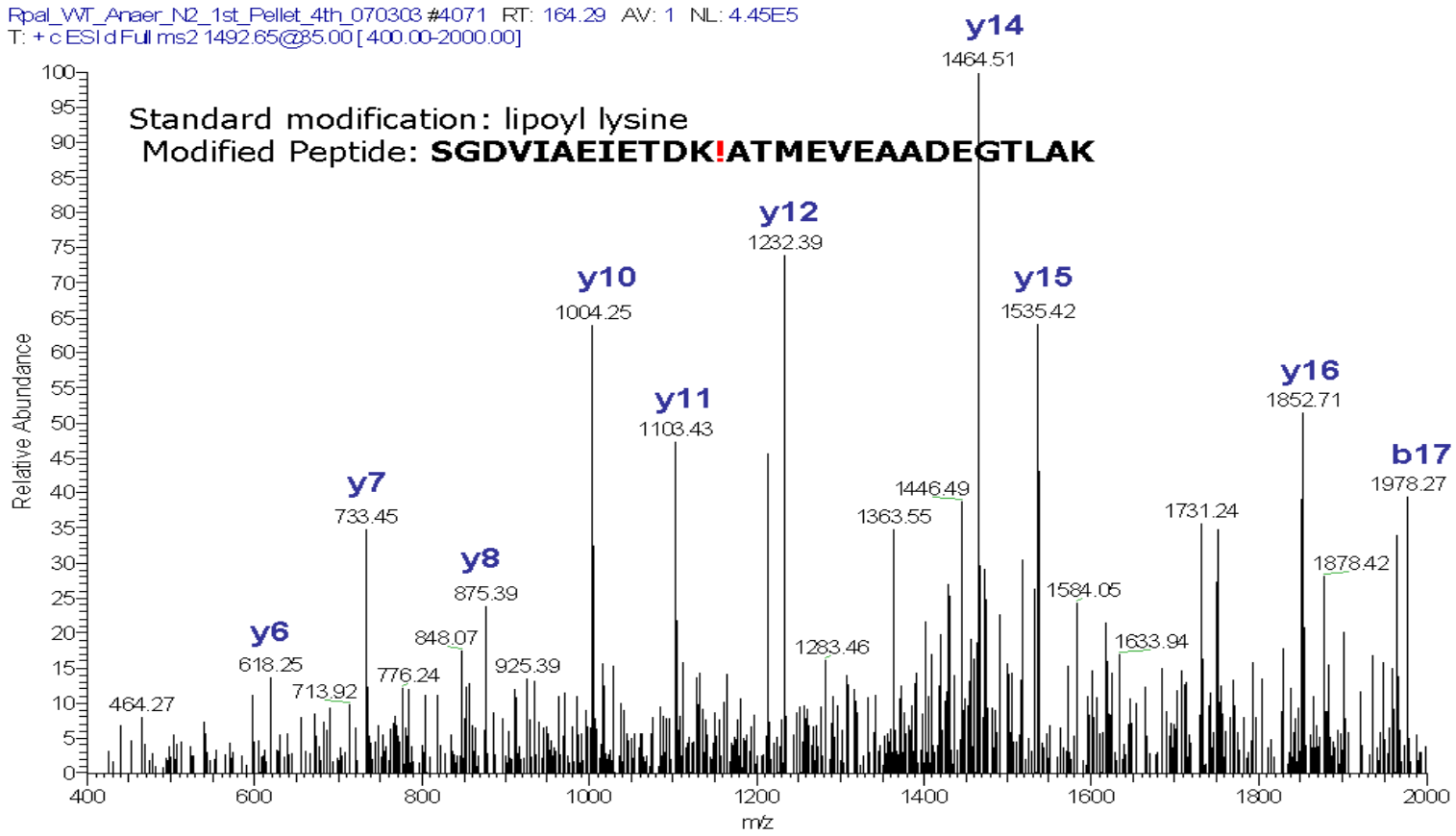

Figure 4.6 The tandem MS spectrum of lipoylated peptide. A representative spectrum of SGDVIAEIETDKATMEVEAADEGTLAK modified peptide is presented, showing the labeling of consecutive 'y' ions.

site (all the lysines in the sequence are marked as possible modification sites). The peptide was detected seven times under various growth conditions, as is expected from the type of modification, since it's presence is expected at all metabolic conditions.

PTMs are post-processing events that occur after protein translation and can rarely be detected at the DNA sequence level. The detection of post-translational modifications is very important for our understanding of biological processes as they modulate the activity of many proteins during the lifetime of an organism. And while they are notoriously difficult to analyze, their characterization can provide an immense insight into biological function. While there are many studies of proteins and pathways regulated by post-translational modifications they are generally performed on case by case basis. These experiments are tremendously important for the advance of the biological knowledge, however, they tend to be extremely slow and frequently must be based upon already known facts. And while the goal of detecting and characterizing all post-translational modifications in the cell is currently unattainable, it is desirable to be able to elucidate the targets for studies of post-translational modifications with greater efficiency as well as perform detections of already known PTMs.

Mass spectrometry is frequently used for detecting proteins in complex samples, with its use it is possible to gain insight into the appearance of a mature protein which can frequently be different from the expected based on the DNA sequence. It is a technique that is well qualified for PTM analysis. While top down mass spectrometry appears more desirable for analysis of post-translational modifications it is currently limited by the technology limitations and is not generally used for high-throughput analysis. The shotgun bottom up MS is a technique that is capable of characterizing

proteins in complete proteomic samples. In this chapter, a new algorithmic method for analysis of post-translational modifications by shotgun mass spectrometry was presented and applied to the detection of post-translational modifications in *R. palustris*. While prokaryotic organisms are rarely studied for PTMs, it was expected that an organism with such metabolic variability might be highly regulated on the PTM level. However, while shotgun mass spectrometry can be used to detect post-translational modifications, further experiments must be considered to verify the made identifications. With the use of the new algorithm for PTM detection, we have uncovered a number of peptides that are likely to be post-translationally modified. There are several experiments that can be used for verification of these findings. One strategy that could be applied to the PTM presence verification is multiple digest strategy. It involves digesting the sample with different types of enzymes and analyzing them by bottom up MS. The presence of the PTM in several overlapping peptides from different digests is a strong indication for the correct identification. The other strategies can involve isolating the proteins of interest under the specified conditions and analyzing them one by one. In this case, the purified proteins can be subjected either to Edman degradation to determine the presence of the post-translational modification or they be analyzed with top down MS using ES-FT-ICR, which can measure the accurate mass of the intact protein, providing the information on the mass shift. To be conservative, an extensive analysis is necessary to confirm the presence of a post-translational modification after it has been detected by bottom up MS method. However, the shotgun bottom up PTM detection can be used to produce candidate proteins for the thorough analysis providing new and interesting avenues in the research of post-translational modifications.

Chapter 5

Computational Simulations for Mass Spectrometry-Based Identification of Biological Agents

Some of the data presented below has been presented as Razumovskaya J., Fridman T., Day R., Borziak A., VerBerkmoes N., Hettich R., Uberbacher E., Gorin A., poster presentation at ASMS 2004

Introduction

Current political events and acts of terrorism have elevated the demand for suitable instrumentation to detect and identify potentially threatening biological agents, such as bacteria, viruses, toxins and chemical agents. This heightened demand for a robust instrument with the capability to simultaneously identify all possible threats within a narrow timeframe exceeds current technology. In order to develop novel instrumentation with such capabilities, it is necessary to probe the threshold of current instrumentation using computational simulations. This chapter describes a computational simulation of organism detection in a complex biological background using top down and bottom up mass spectrometry. The focus of this chapter is to explore the differences between these two approaches and the acceptable instrumental parameters for each of the methods.

Detection of an organism in a complex biological sample is a two-sided problem of sensitivity and specificity of detection. From the side of sensitivity, it is necessary to be able to detect the organism of interest, however, it is also essential to be sufficiently specific to recognize the absence of an organism in the complex mixture. Both sensitivity and specificity of a method are equally important. Without high sensitivity many biological agents may not be detected, while without high specificity, there will be

a very high rate of incorrect detection, rendering the method similarly ineffectual. Sensitivity of organism detection by mass spectrometry is significantly dependent upon the differences between the concentrations of the organism of interest and the background as mass spectrometry measures protein abundance. Specificity is only affected by the measurements of the proteins present in the sample, focusing on whether background proteins could be erroneously accepted as the evidence of the presence of the organism of interest. In the presented theoretical simulations, the data is fully computer generated, the real experimental conditions are only approximated and noise potentially present in the spectra is ignored. Thus, the sensitivity is not in question – a protein from the organism in question can always be detected. However, there is an uncertainty as to the specificity of the method: whether a protein from the background could be misidentified as belonging to the organism of interest. Therefore, in this chapter, we evaluate only the specificity of detection of an organism of interest in a complex background by computer simulations.

Current methods of organism characterization

Basic PCR detection method: Polymerase chain reaction (PCR) based organism detection is an old and well established technique. It is based on the concept of DNA hybridization – a set of oligonucleotide primers from a particular organism is added to the DNA sample, if the primers complementary to the DNA in the sample, they hybridize, and the organism is detected. If there is no hybridization, it is assumed that the organism's DNA is not present in the sample.

Possible downfalls/problems:

1. DNA can be changed through genetic manipulations while it is significantly more difficult to alter proteins.
2. Have to be able to make template DNA from the sample in field in order to detect any new organisms.
3. All the components of PCR must be made stable under long periods of time.

Antibody based detection method: Antibodies are developed to recognize particular proteins. When the proteins are present in the sample, the antibodies will recognize them and give out a signal.

Possible downfalls/problems:

1. The development of new antibodies is difficult and takes a long time. This constraint, therefore, limits it's usefulness for the detection of new organisms.
2. It is difficult to select a combination of proteins which will be specific only to one specific organism strain.
3. The antibodies may not be very specific to an organism in the truly complex background. When more organisms are introduced into the sample the chances that another organism possesses the protein for which the antibody was designed increases.

Mass spectrometry: Mass spectrometry is a technique uniquely qualified to quickly characterize a complex proteomic mixture by measuring the masses and fragmentation patterns of the proteins in the mixture. This capability can be used for detection of one or more organisms within a large proteomic sample. Frequently mass spectrometry is used

to characterize an organism's proteomic content, therefore it can be proposed that it also possible to characterize protein content of multiple organisms.

Top down versus bottom up

In this chapter we discuss two main mass spectrometric techniques used for protein characterization: the top down and bottom up methods as the techniques of organism detection. The top down method is used to directly characterize protein content and the bottom up characterizes the protein content by the analysis of protein pieces -- peptides. The top down method involves measuring the m/z ratios of intact proteins, while the bottom up method involves proteolytic digestion, cutting the intact proteins into shorter amino acid stretches (peptides) with one or more proteases (such as trypsin, pepsin, GLU-C, etc) and measuring m/z ratios of the resulting peptides. Either of these methods can be coupled with tandem MS analysis, providing the sequence information (fingerprint) for the analyzed protein.

The advantages of using top down analysis for the organism detection lie in the potential speed of measurement and the reduced complexity of the mixture leading to the reduced complexity in the identification process. The potential speed of measurement is due to the fact that top down technique does not require protein digestion period -- protein masses are measured intact. The reasons for reduced complexity of the identification (as well as the speed of analysis) are a) intact protein masses are generally more distinctive then the peptide masses and b) each protein in the organism corresponds to a single measurement (assuming there is only one form of every protein, and not taking into account the isotopic packet). The disadvantages of this technique involve a) not fully developed instrumentation (expensive and not designed for routine operation

devises), b) difficulty of protein mixture separation by liquid chromatography, c) difficulty in deconvoluting the intact protein spectra, d) difficulty in performing tandem MS on intact proteins cause incomplete tandem MS patterns.

For the past decade bottom up has been the main method for high-throughput proteomic experiments, instrumentation becoming well developed, robust and routine to operate. Another advantage of bottom up technology is that it provides nearly complete peptide tandem MS fragmentation. The disadvantages of bottom up are the increased time of measurement and the identification complexity. The increased time of measurement is due to the time spent performing the proteolytic digestion which can range from a few hours to an overnight digest (which is a common practice during proteomic experiments). The increase in the identification complexity is caused by the amount of peptides to be interpreted, as each protein can correspond to 20 peptides, in most cases any single peptide identification being inconclusive as to the presence of the parent protein.

The described computational simulation is designed for two purposes: a) to examine the performance of the two techniques for detection of a potential biological agent in an environmental sample and b) to evaluate the instrumental parameters which will be necessary for the task for each of the techniques. The performances of the top down and bottom up methods are measured by a series of computational studies which not only make it possible to easily create a controlled experiment, they also allow for straightforward way to vary the instrumental parameters. Even though computational simulations lack many elements of real life situations such as noise, concentration detection limits, multiple parent ions in the fingerprint pattern and others it can serve as

an indication of best case scenarios for these experiments. Using the guidelines shown by the simulation, it will be possible to select the best experiments to pursue the strategies for organism detection. The simulation involves choosing an organism of interest which will represent the biological agent and a realistic organism background which can approximate the complexity of an environmental sample. Then the specificity of organism of interest detection is tested by both top down and bottom up methods with varying instrumental parameters. The specificity performances of the two methods can then be compared and the best instrument and experiment can be described.

Materials and Methods

Simulation design

Simulation of biological agent

In order to simulate the detection problem, an organism of interest was selected to represent a biological agent and the complex organism background to represent an environmental sample.

The organism of interest chosen for detection is the widely studied gram negative prokaryotic organism *Escherichia coli (E. coli)* K-12. One of the best studied model organisms in molecular biology and biochemical genetics, *E. coli* K-12 was one of the earliest candidates for full genome sequencing, its complete genome sequence published in 1997 (Blattner, 1997). *E. coli* K-12 is a generally harmless bacteria frequently found in mammalian intestinal tracks. *E. coli* genome consists of 4237 number of genes, each of which could potentially form a gene product that can be measured by mass spectrometry. However, some of the predicted genes may not be coding for proteins and not all of the proteins predicted by the genome database are expressed at all times in the

proteome; in addition many of them are not in the abundance to be successfully detected by mass spectrometry. In order to successfully detect an organism by mass spectrometry all of these points must be taken into an account and only detectable proteins can be used to represent the organism of interest. Thus, rather tha

n include all of the possible genes into the analysis, only a subset of *E. coli* genome was selected to represent it as the target organism in the simulation. The representative proteins were selected based on experiments, where the cell proteome was repeatedly measured by mass spectrometry and analyzed by database searching algorithms. The 376 proteins repeatedly detected in the proteome became the "signature proteins" (from now will be referred to as "signature proteins") for *E. coli* detection.

Simulation of environmental sample

The representation of environmental sample consists of twelve distinct organisms, including prokaryotic and eukaryotic bacteria, plant and fungi (Figure 5.1). The proteomes of some of the organisms are similar to *E. coli* proteome, increasing the complexity of the analysis. Many of the organisms are commonly found in soil and water samples. The twelve organisms are expected to be a reasonable representation of the complexity of an environmental sample. Here are short descriptions of the selected organisms. *Bacillus anthracis* is gram positive spore forming prokaryotic organism widely distributed in nature and a potential biological weapon, *Deinococcus radiodurans* is a gram positive prokaryote and the most radiation resistant organism known, *Burkholderia xenovorans* formally known as *fungorium* and is a gram negative organism widely found in soils, *Geobacter metallireducens* is a gram negative microorganism

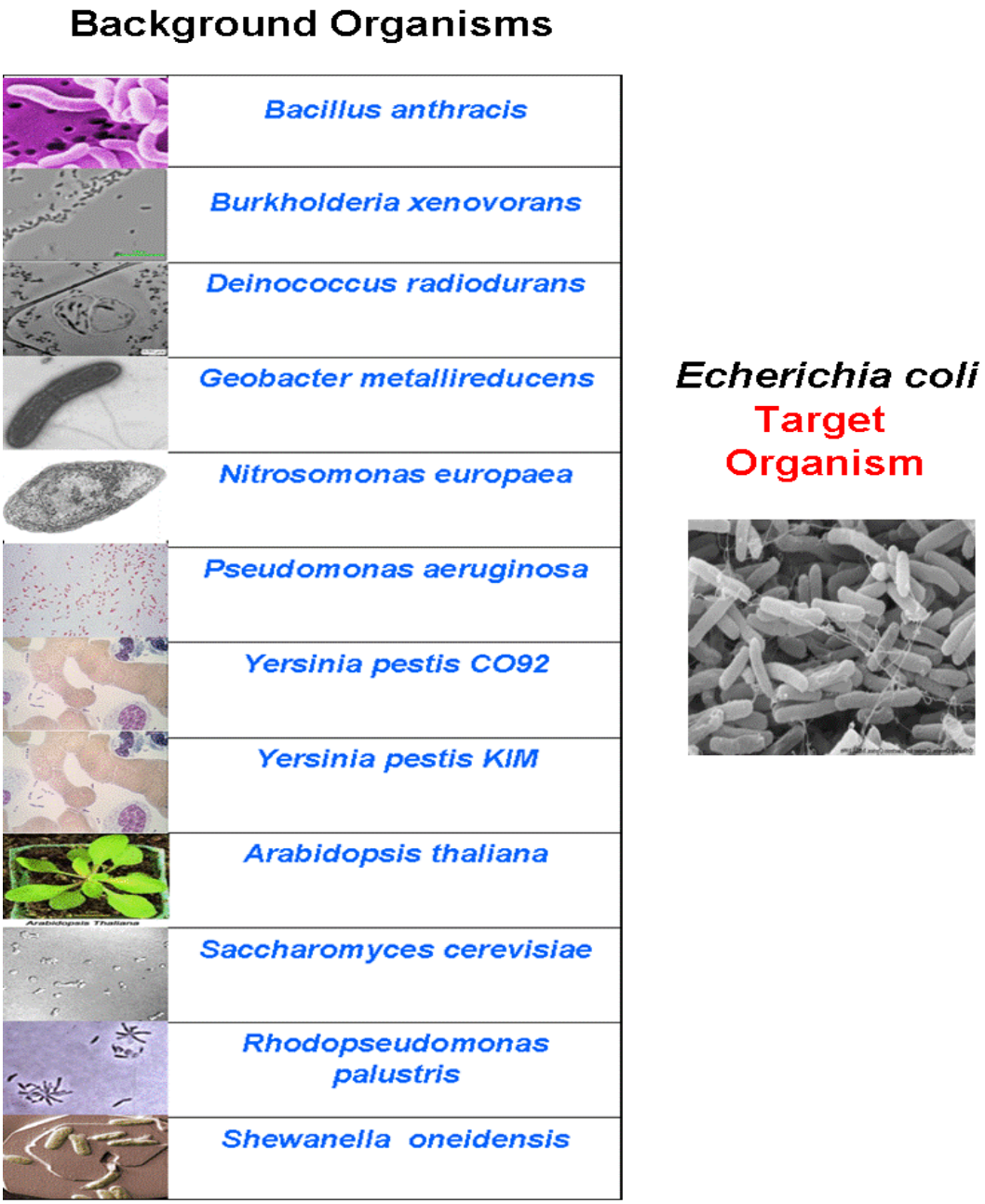

Figure 5.1 Component organisms used in the simulation.
The representation of background sample consisting of in total 12 organisms and the target organism (*Escherichia coli*)

capable of reducing metals generally found in sedimentary environments, *Nitrosomonas europaea* is a gram negative ammonium oxidizing bacteria commonly found in the soil, and *Pseudomonas aeruginosa*, is a gram negative opportunistic pathogen widely distributed in soil and water. The two strains of *Yersinia pestis* CO92 and KIM are gram negative organisms known to cause plague, commonly found in nature, *Arabidopsis thaliana* is the first plant to be completely sequenced, *Saccaharomyces cerevisiae* is a eukaryotic fungus and is widely utilized by humans in food production industry, *Rhodopseudomonas palustris* is a gram negative prokaryote commonly found in sedimentary environments and *Shewanella oneidensis* is a gram negative prokaryotic organism found in water and soil. The majority of the information about the organisms was adapted from Margulis and Scwartz's "Five Kingdoms", 2001. Together, the proteomes of these twelve organisms represent the environmental background for this simulation. The environmental background protein mixture consists of 83,777 proteins, which will later be referred to as "background proteins".

In order to assure that the simulation is realistic we compared the mass distributions of the 376 "signature proteins" found in *E. coli* to the 83,777 "background proteins" as shown in the figure 5.2. The mass distribution of the "signature proteins" is similar to the mass distribution of "background proteins"; all of the masses of "signature proteins" lay in the same region as the "background protein" masses, which simulates the worst case scenario since none of the "signature protein" masses are significantly different from the "background protein" masses. The inset shows an enlarged view of mass distribution of *E. coli* proteins in the window of 10,000 Da. The number of *E. coli* proteins in the window is small comparing to the number of the background proteins.

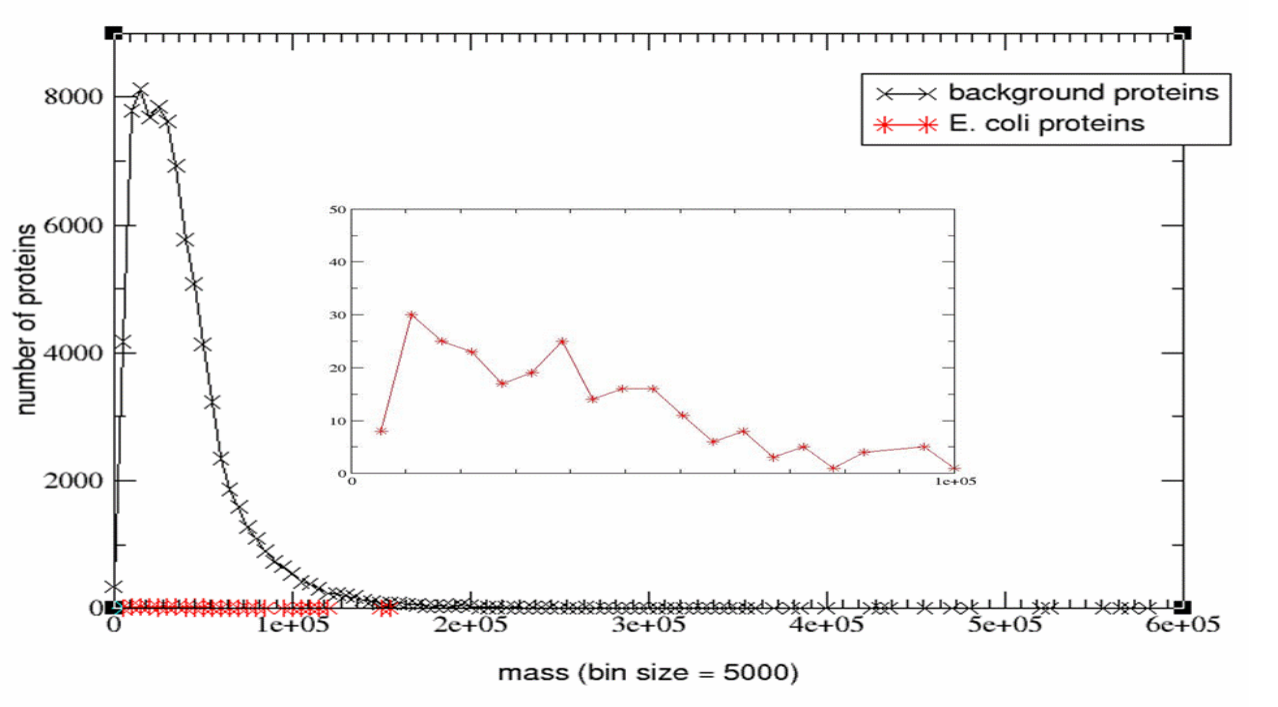

Figure 5.2 Protein mass distributions for the background and *E. coli* proteins. Illustrates the mass distributions of background proteins and the *E. coli* "signature proteins". The inset demonstrates the blown up area of "signature proteins" in the mass range of 0 – 10,000 Da. Each "signature protein" is found in a bin with about 400 times more background proteins.

Scoring schemes

The second step to performing the study was to formulate a feasible scoring system for protein and peptide identifications and organism detection. The scoring schemes consist of following two parts: a) score for tandem MS spectra, which are used to measure similarity between two tandem MS spectra and b) score for the organism detection, which is used to determine presence or absence of an organism within the mixture.

Fingerprint scoring schemes

The tandem MS scoring scheme for both top down and bottom up data analysis was adapted from the widely used mass spectrometry data interpretation software package, SEQUEST. It is a general cross-correlation score used for spectral comparison. As described in the original SEQUEST paper (Eng, 1994), the cross-correlation function between two spectra can be calculated using following formula:

$F(\tau) = \sum_{i=1}^{n} x[i] * y[i+\tau]$, where x[i] represents a spectrum from *E. coli* "signature proteins", y[i] represents a spectrum from the background protein and τ is the displacement value by which the mass index is offset. The cross-correlation then is computed as follows: $cross-correlation = \frac{F(0) - \sum_{i=-10}^{10} F(i)}{20}$. As shown in the formula, in this study, the displacement τ is varied in an interval of [-10;10] in increments of 1. If two spectra are the same, the cross-correlation function should be maximized at the displacement 0, thus the comparison between the *F(0)* and average of the displacements in the [-10;10] interval reflects the similarity between the two spectra. The cross

correlation function can then be normalized by dividing by the best cross-correlation achievable: auto-correlation. Then the final normalized cross-correlation formula applied in the simulation can be described as follows:

$$Score(pattern1, pattern2) = \frac{F(0) - \frac{\sum_{i=-10}^{10} F(i)}{20}}{Score(pattern1, pattern1)}$$, where *pattern1* stands a spectrum (fingerprint pattern) of a "signature protein" and *pattern2* stands for the spectrum (fingerprint) of a background protein.

Organism scoring schemes

The score used for the assessment of organism detection (later referred to as *OrganismScore*) was developed to reflect the likelihood of organism's presence, based on the detection of "signature proteins" found in the mixture. Though it appears that the occurrence of each of the "signature proteins" in the mixture should increase the probability for the presence of an organism in the sample, it cannot be expected that a) all of the "signature proteins" will be reliably detected and b) all of the "signature proteins" are unique to the organism of interest and have no duplicates within the background. Therefore, a concept of *OrganismScore* was introduced in order to evaluate the reliability of organism detection. The *OrganismScore* should involve two factors for each "signature protein": *factor1*, how reliably the protein was detected and *factor2* how unique the protein is to the organism of interest versus the background mixture proteins.

The reliability of protein detection, *factor1*, can be estimated by any scoring scheme used for spectral comparison (comparison between the fingerprints of the "signature protein" and the measured fingerprint) such as cross-correlation used in this

simulation. However, since this simulation does not involve modeling noise, the cross-correlation score for a perfect match always results in a score of 1 (where 1 is the highest possible score). Therefore, for this simulation purposes, the reliability term of protein detection has been made binary – a protein is either detected (*factor1* is assigned to 1) or it is not (*factor1* is assigned to 0). In further studies it is possible to model noise and improve detection by assigning weights to *factor1*.

The assessment of protein's uniqueness to the organism of interest in a background mixture is *factor2* for the *OrganismScore*. In case of a particular "signature protein" being present only in an organism of interest, and not in the background mixture, the protein would be unique to the organism of interest (here, *E. coli*) and thus, a reliable detection of such protein would lead to the detection of the organism. On another hand, if several "signature proteins" could also be found in the organisms present in the background sample, their detection would not mean the presence of the organism of interest. Thus, knowledge whether "signature proteins" are unique to the organism of interest is very important, however, since the identity of organisms present in the real samples is unknown, we would have to assess the likelihood that the protein is unique to only one organism in nature.

A protein with a distinctive amino acid sequence (the protein mass can coincide with masses of other proteins but their sequences must be different) can be considered unique, and ideally can be identified by mass spectrometry by a combination of mass and fingerprint information. A presence of a homologous protein with a 98% sequence similarity (a slightly different sequence) in a background mixture generally won't lead to an incorrect identification with top down mass spectrometry due to the differences in

overall protein masses. However, since not all the organisms are yet sequenced and our knowledge about genes and proteins is still limited it is not a simple task to assess the uniqueness of a protein in nature. In general, it is expected that protein sequence is conserved if it is crucially important for its functionality. The answer of protein sequence conservation lies in observing its family members – the homologous proteins from other organisms. It is frequently observed that functionally important amino acids are conserved within sequences in a protein family, while parts of a sequence may differ; the more amino acids are functionally important the more sequence is shared between homologous proteins. Protein family based profiles are used to assess the uniqueness of a protein. Using BLAST searches against nr (All non-redundant GenBank CDS translations, RefSeq Proteins, PDB, SwissProt, PIR and PRF) database available from NCBI, all of the homologous proteins from different sequenced organisms present in the mentioned databases can be found, it is then possible to calculate the position dependent frequency of amino acid conservation (how frequently an amino acid in a particular position is conserved within the family); an example of such matrix is shown in the figure 5.3. It is then possible to compute the likelihood that the exact copy of the protein can be present in another organism using the amino acid conservation frequency information by combining the conservation frequencies of amino acids in each position. The probability is computed in the form of $factor2(i) = 1 - \prod_{k=1}^{M} FM(a_k, k)$, where M is number of amino acids present in the ith protein, FM is the family profile frequency matrix, and $FM(a_k,k)$ stands for frequency of conservation of amino acid a, in kth position in the sequence. As

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	
M	0.04974	0.02240	0.01583	0.01552	0.01065	0.02399	0.02376	0.02218	0.00873	0.13086	0.13467	0.02600	0.16363	0.03029	0.01847	0.03353	0.04124	0.
K	0.04434	0.28323	0.03317	0.02972	0.00520	0.05185	0.05833	0.03134	0.01544	0.01586	0.03361	0.24133	0.01153	0.01155	0.02513	0.04956	0.03524	0.
Y	0.04074	0.03914	0.02030	0.01959	0.00764	0.02926	0.03430	0.02475	0.01557	0.04097	0.14521	0.12337	0.06667	0.10649	0.01941	0.03637	0.03227	0.
K	0.11698	0.06087	0.03043	0.03033	0.00833	0.04189	0.05359	0.04014	0.01233	0.02675	0.09646	0.24184	0.01550	0.01571	0.02871	0.09689	0.04347	0.
H	0.05295	0.02570	0.03141	0.02549	0.00748	0.02611	0.03281	0.23122	0.12199	0.02247	0.09075	0.02962	0.01346	0.02156	0.01998	0.07884	0.03247	0.
L	0.03695	0.01662	0.01087	0.01225	0.00922	0.01412	0.01722	0.01611	0.00634	0.14463	0.33833	0.01791	0.02883	0.06707	0.01459	0.02382	0.03089	0.
I	0.11693	0.01724	0.01229	0.01366	0.00956	0.01583	0.02033	0.02418	0.00679	0.11658	0.21914	0.02022	0.02265	0.05238	0.01688	0.03335	0.03274	0.
L	0.03575	0.01776	0.01211	0.01263	0.00888	0.01560	0.01823	0.01650	0.00627	0.09659	0.42518	0.01906	0.04806	0.02917	0.01490	0.02753	0.06941	0.
S	0.19681	0.02191	0.02378	0.02278	0.00961	0.02148	0.03051	0.06266	0.00953	0.02708	0.09638	0.02867	0.01382	0.10194	0.02274	0.18578	0.06698	0.
L	0.16100	0.01928	0.01552	0.01708	0.02864	0.01788	0.02439	0.03018	0.00880	0.07131	0.18441	0.02368	0.01897	0.07826	0.06437	0.05510	0.03404	0.
S	0.24873	0.02052	0.01930	0.01949	0.03514	0.02052	0.02754	0.09031	0.00817	0.04608	0.13800	0.02698	0.03812	0.01744	0.02213	0.10476	0.03882	0.
L	0.17904	0.01819	0.01386	0.01492	0.00957	0.01667	0.02219	0.05523	0.00734	0.04922	0.23885	0.02197	0.02110	0.07700	0.01835	0.04010	0.03275	0.
I	0.10292	0.01787	0.01301	0.01414	0.02406	0.01648	0.02094	0.02286	0.00654	0.06894	0.29498	0.02097	0.03478	0.02445	0.04614	0.04696	0.03401	0.
M	0.08498	0.01739	0.01147	0.01227	0.02513	0.01579	0.01891	0.02089	0.00672	0.05421	0.32653	0.01914	0.04503	0.05587	0.01528	0.02903	0.02971	0.
L	0.12400	0.02116	0.01983	0.01947	0.00952	0.03525	0.02774	0.07754	0.00796	0.04522	0.20875	0.02640	0.01937	0.02003	0.02021	0.10471	0.07483	0.
G	0.18273	0.02435	0.02468	0.02632	0.00994	0.02363	0.03371	0.14795	0.00991	0.02732	0.10133	0.03218	0.01370	0.01677	0.13929	0.09575	0.06129	0.
P	0.08290	0.02559	0.02606	0.02904	0.00880	0.03820	0.03731	0.11522	0.01052	0.03721	0.04228	0.03462	0.02126	0.01495	0.30374	0.10609	0.05892	0.
L	0.08563	0.02721	0.02609	0.02731	0.01133	0.02548	0.03538	0.03825	0.01066	0.04828	0.15345	0.03293	0.03492	0.02299	0.05456	0.13562	0.08057	0.
A	0.35292	0.02305	0.02088	0.02269	0.07282	0.02268	0.03221	0.09955	0.00941	0.02734	0.05580	0.03090	0.01264	0.01526	0.02771	0.06566	0.03993	0.
H	0.09600	0.04342	0.03388	0.11852	0.00754	0.04576	0.08630	0.06753	0.10042	0.02145	0.04762	0.03763	0.01079	0.01448	0.06774	0.11517	0.03782	0.
A	0.37073	0.08023	0.02299	0.02293	0.01075	0.02603	0.03542	0.04926	0.00981	0.02181	0.03703	0.03854	0.01130	0.01252	0.02671	0.11318	0.07852	0.
E	0.04511	0.03466	0.03229	0.09971	0.00474	0.08765	0.40177	0.02878	0.01461	0.01281	0.02457	0.05107	0.00849	0.00940	0.02521	0.09686	0.03505	0.

Figure 5.3 Frequency matrix based on family profile information. Shows an example of family based profile frequency matrix, where the vertical sequence represents the protein sequence and the horizontal sequence represents the twenty amino acids. Each position in the matrix symbolizes the frequency of an amino acid substitution in the sequence based on the protein family.

it can be noticed, multiplication of numbers below 1 over a large *M* would produce a 0, therefore, in real calculations log scales were employed to compute *factor2*.

For the purposes of the simulation, *factor1*, the presence of a protein in the mixture, is defined in binary terms dependant on the accepted cross-correlation score cutoff (given a score cutoff, all the proteins with a cross-correlation lower than the cutoff are considered absent: *factor1* = 0, and all the proteins with a cross-correlation above the cutoff are considered present: *factor1* = 1). The *factor2*, protein uniqueness scores, are normalized, so that the sum of all *factor2* scores for the "signature proteins" is equal to 1 $\left(\sum_{i=1}^{N} factor2(i)\right) = 1$, where *N* stands for the number of "signature proteins" in the organism of interest (here, 376). The definition of *OrganismScore* is then the normalized sum of the multiplied protein presence and protein uniqueness over all the "signature proteins" of an organism: $OrganismScore = \sum_{i=1}^{N} factor1(i) * factor2(i)$, where if all of the "signature proteins" are present in the sample, *OrganismScore* is 1, and each of the present proteins makes a contribution to the *OrganismScore* according to its likelihood of uniqueness in nature, *factor2*. Since detection of all "signature proteins" in the sample yields an *OrganismScore* of 1, it is easy to compare the probability of organism detection between different samples and different organisms of interest where background and "signature proteins" are different.

Results and Discussion

The described simulation involved exploring the efficiency of organism detection by top down and bottom up mass spectrometry approaches with varying instrumental parameters. Both approaches were performed in two modes: MS mode and fingerprint

mode. MS mode refers to using only MS information (measurement of protein or peptide mass), and fingerprint mode refers to using both MS and MS/MS information (using measurement of mass and sequence fingerprint).

The MS mode was used to determine the fractions of "signature" proteins and peptides which could be separated from the "background" proteins and peptides based on their mass alone for varying accuracy measurements. The accuracy measurements for the MS mode top down experiment were varied in the interval of 0 to ±20 Da, with an incremental step of ±1 Da; for the bottom up experiment, the accuracy measurements were varied between 0 and ±5 Da, with ±0.001 Da and ±0.1 Da in the range of 0 to 1 and an incremental step of ±1 Da in the range of 1 to 5.

The fingerprint mode was used to evaluate the specificity of *E. coli* detection in a background of 12 other organisms. The top down method simulations were arranged to explore varying parameters for the measurement accuracy and the fragmentation efficiency. Since fragmentation of intact proteins is not easily achieved and is generally very limited, the fragmentation efficiency for top down experiment was explored. The results shown for the top down method are for the measurement accuracy of 20 Da, the fragmentation efficiencies are varied between 1 – 150 fragments per protein. In case of bottom up experiments, the peptide fragmentation is significantly more efficient than the protein fragmentation and the bottom up data interpretation programs, such as SEQUEST, assume complete peptide fragmentation. The bottom up method simulations were performed with ±3 Da measurement accuracy, while the fragmentation efficiency for peptides is considered to be complete.

Top down MS mode

The MS mode for top down simulation shows the dependency between the fraction of unique "signature protein" masses and the measurement uncertainty. In order to establish the fraction of unique "signature protein" masses, mass of each "signature protein" was computed and compared to the computed masses of the proteins from the "background mixture" with varying measurement uncertainty in the interval of 0 to ±20 Da. Measurement uncertainty applies to the delta error allowed for the matching between observed and expected masses. As expected, the higher is the measurement uncertainty (higher measurement accuracy), the less there are "signature proteins" with unique masses. At an extremely low measurement uncertainty of 10^{-4}Da, the fraction of unique protein masses approaches 1 as shown in the figure 5.4. Most of the proteins possess unique mass at such accurate mass measurement, using an instrument with such accuracy and resolution, it might be possible to perform specific organism detection in a top down MS mode unless there are present background proteins with the same amino acid composition and different sequence. However, the fraction of unique proteins is then sharply reduced to 0.21 at the measurement uncertainty of ±1 Da, decreasing even further to 0.03 at the measurement uncertainty of ±10 Da. The top down MS mode can only be used for organism detection when the instrumental accuracy and resolution are extremely high, rendering measurement uncertainty to being almost negligible.

Bottom up MS mode

A similar analysis to measure the specificity of detection was performed for bottom up MS mode as previously mentioned for the top down MS mode. In order to

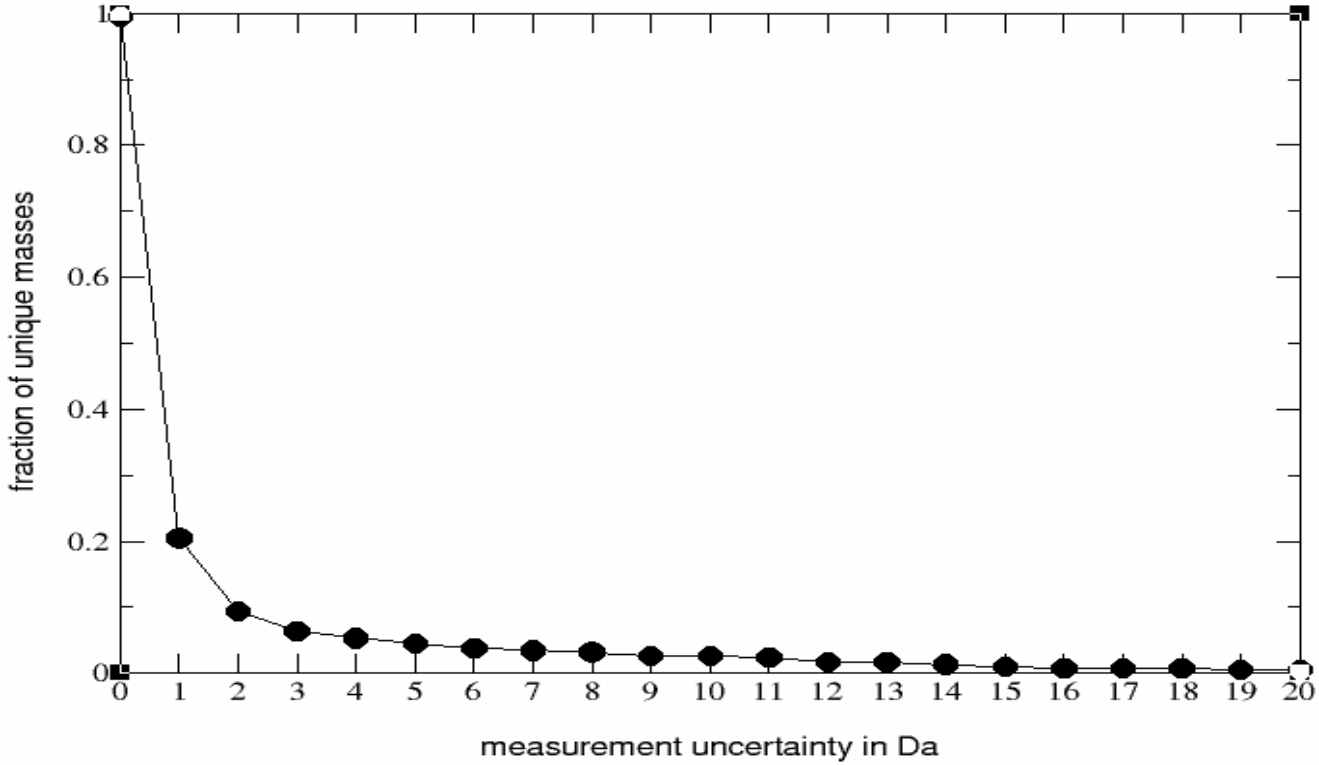

Figure 5.4 Unique "signature protein" masses as comparing to the background set as a function of measurement error. The plot of intact protein masses that are unique to the *E. coli* "signature proteins" at varying measurement uncertainties.

perform bottom up simulation, the "signature proteins" firstly have been *in silico* digested by trypsin to produce "signature peptides". Trypsin digest leads to cleavages along the peptide backbone after Lysine (K) and Arginine (R) residues. During an experiment, due to insufficient digest time, incomplete peptide denaturation and other causes trypsin digest is often incomplete leaving peptides with internal K and R residues. In an attempt to model a realistic tryptic digest, up to four missed cleavages has been allowed in the *in silico* digest process (the accepted settings for bottom up data interpretation programs). The digestion produced 70,000 "signature peptides" (only peptides with different sequences were considered), and 125 million of unique background peptides. The MS mode for bottom up simulation shows significantly reduced number of unique masses for the "signature peptides". As illustrated in the figure 5.5, the fraction of unique peptide masses at the measurement uncertainty of 10^{-4}Da is less than 0.24, while at the measurement uncertainty of ±0.01 Da, the fraction of unique peptides is 0.08. It would be impossible to specifically detect *E. coli* in this background based exclusively on the peptide masses with measurement uncertainty higher than 10^{-4}Da and even then, the number of non-unique peptides if very high. In this simulation, MS mode for bottom up does not appear to be a successful approach for a specific organism detection.

Top down fingerprint mode

The fingerprint mode for top down simulation shows the specificity of *E. coli* detection in the complex background sample under varying measurement parameters. The first parameter is previously examined measurement uncertainty and the second is

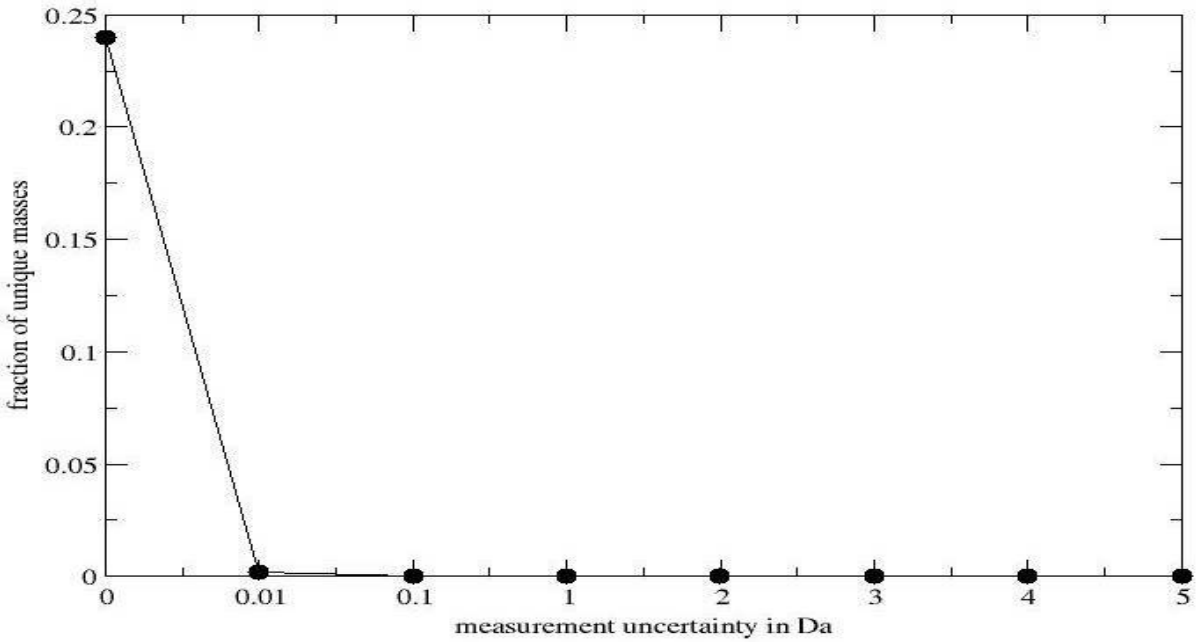

Figure 5.5 Unique "signature peptide" masses as comparing to the background set as a function of measurement error. The plot of peptide masses unique to the *E. coli* "signature peptides" at varying measurement uncertainties.

the tandem MS fragmentation efficiency. The experiments involving measurement uncertainty have been addressed in the MS mode section, showing the impact of delta error on the protein/peptide detection accuracy. The primary focus of this part of the experiment is to observe the changes in detection specificity as a function of fragmentation efficiency. The fragmentation efficiency refers to the capacity of the instrument to break a biological molecule into smaller fragments creating its fingerprint pattern (sequence information). As previously mentioned, the fragmentation of intact proteins is frequently incomplete, providing only a few fragments while the core of the protein often remains whole. The greater is the amount of fragmentation, the more information is contained in the fingerprint which then provides a more specific identification.

As noted above in the top down MS mode section, measurement uncertainty applies to the delta error allowed in matching between observed and expected masses. In the fingerprint mode it was fixed at the reasonable values, when the parent mass alone can not be used as a detection factor. In case of top down experiment, it is expected that increased measurement error will not greatly affect the resulting specificity since at ±10 Da error a large increase in measurement error causes comparably small changes in the number of matching background proteins. The range of 0-1 Da of measurement uncertainty where the parent mass for top down experiment can be used for protein detection was not considered in this study as an unrealistic measurement constraint creating a greater background dependency (more proteins with similar masses could be found within a different background).

In order to establish fragmentation efficiency necessary for confident and specific organism detection, the detection of *E. coli* proteins was attempted in the background mixture with varying fragmentation efficiencies for the background proteins. As previously noted, greater fragmentation efficiency (more resulting fragments) provides a more informative fingerprint – the greater the number of fragments the lower is the chance that protein will be identified incorrectly. Separate simulations were performed for eighteen fragmentation efficiencies. Fragmentation efficiency was modeled by creating incomplete fingerprints of length N for each of the background proteins, where N stands for the number of fragments allowed per protein and remains the same for all proteins in a given simulation. The procedure involves creating all the fragment ions possible for a protein (for an average protein of length 300 amino acids, the number of all expected fragments is ~600), and randomly selecting N of them to create a fingerprint. Each protein fingerprint from *E. coli* "signature proteins" was compared using cross-correlation to the incomplete fingerprints of all of the background proteins that match the "signature protein" by parent mass within the delta mass window. The plot of "signature protein's" scores for a fragmentation efficiency of $N = 10$ is shown in the Figure 5.6. The black line ("second hit") representing the highest cross-correlation score from the background protein as comparing to each single "signature protein". In the absence of noise, the cross correlation of "signature protein's" fingerprint to itself would produce a score of 1. The red line ("average hit") shows the average of cross-correlation scores for all the background proteins matching each single "signature protein". Only in one case a best score for a background protein achieves a cross correlation of 1 (full score for "signature protein"), all other background protein top scores ranging from 0.7 to 0; the

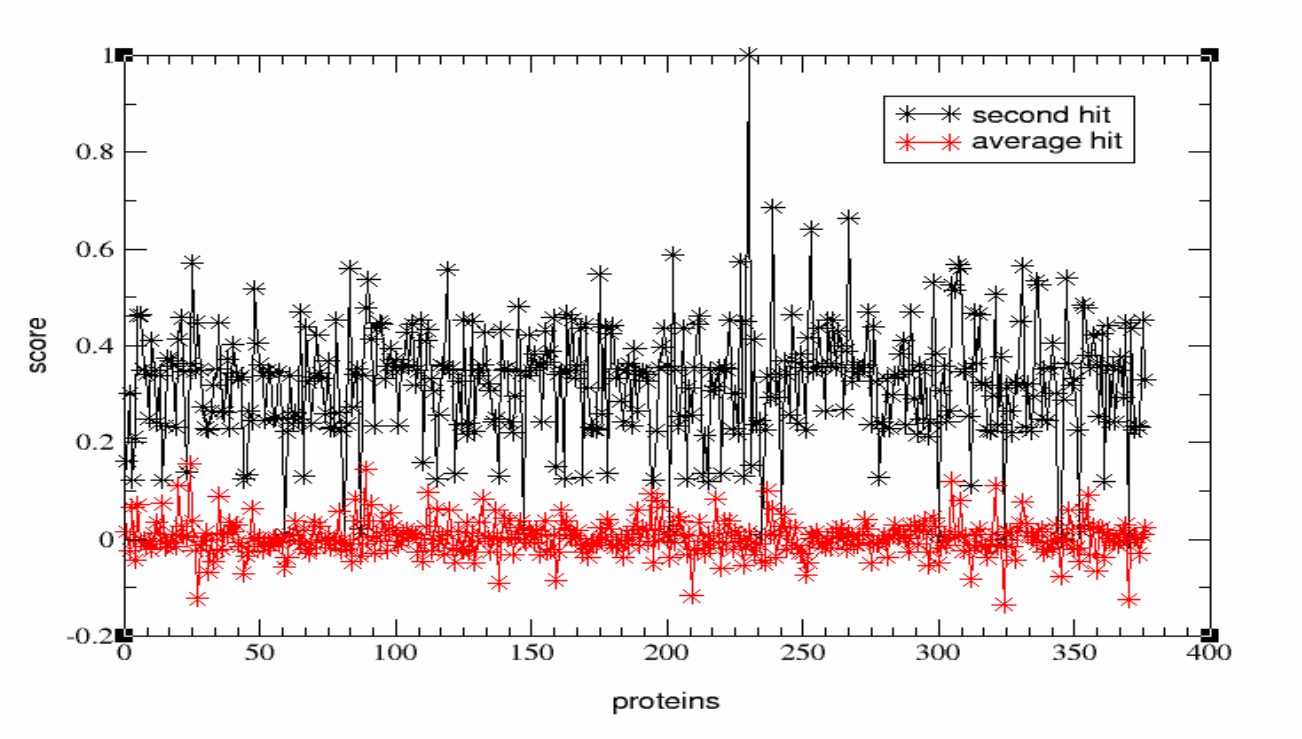

Figure 5.6 Protein detection with fragmentation efficiency *N* = 10.
Difference in scores for "signature proteins" fragmentation patterns as comparing to the best and average matches in the background. The "x" axis represents the number of a protein in the "signature" set from 0 to 376 and "y" axis represents normalized cross correlation score.

average of top background protein scores are 0.38. The average scores of all background proteins show how far the top background protein score is from the average. In this figure, the average of cross correlation of background proteins to the "signature proteins" is shown to be 0.

The specificity of the method in this case can be measured by the percentage of incorrect identifications as a function of cross-correlation score cutoff. Score cutoffs are set in order to make a decision whether a protein is detected. It is not expected to have "signature proteins" in the background sample, therefore all of the proteins "detected" in this simulation are in fact incorrect identifications, which can lead to unspecific detection of an organism of interest. The specificity of protein detection for the fragmentation efficiency of $N = 10$ is shown on the Figure 5.7. At the cross-correlation score cutoff of 0.4, 21.8% of "signature proteins" are detected in the background sample. It is sharply reduced with increasing score, with cross-correlation cutoff of 0.5, less then 10% of incorrect identifications remain, and at score cutoff of 0.9, there is only one "signature protein" that is still detected in the background mixture. Upon examination, it was found that one of the background proteins has an identical amino acid sequence as the "signature protein", making it impossible to differentiate between them. Therefore, at least one *E. coli* "signature protein" will always be incorrectly "detected" in this background mixture.

In the eighteen top down organism detection simulations, the fragmentation efficiency ranged from $N = 10$ to $N = 150$ fragments with an increment of 10 fragments; in the window 0-10, additional points of $N = 1$ and $N = 5$ fragments were included. The dependency of "signature protein" detection as a function of

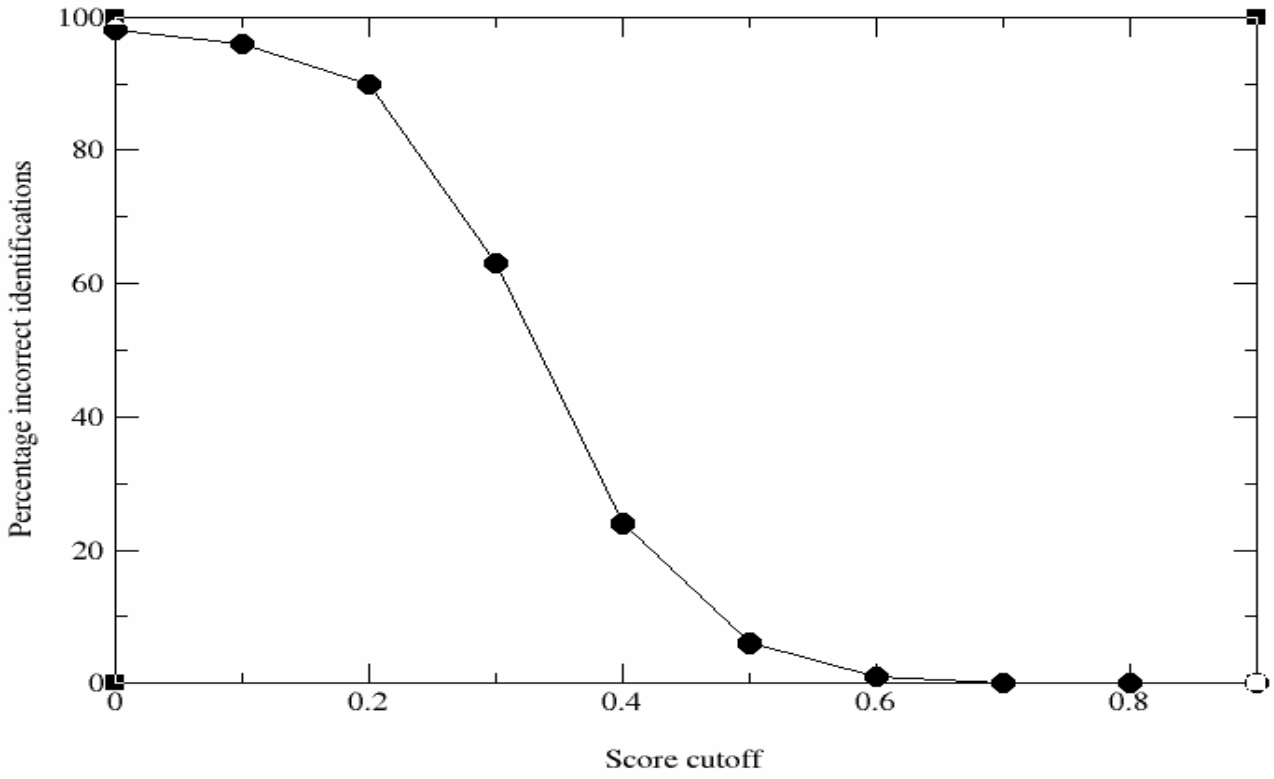

Figure 5.7 Specificity of protein detection.
Percent of background proteins incorrectly matching to *E. coli* "signature proteins" as a function of the cross correlation score cutoff.

fragmentation efficiency for the range of N = [0, 150], is shown in figure 5.8. The black line shows the average of cross-correlation top scores between all the "signature proteins" and the corresponding background proteins for each N. The red line refers to the average of the mean cross-correlation scores of all the "signature proteins" to the background proteins. As expected, Figure 5.8 shows that there is a relationship between the fragmentation efficiency (N) and the specificity of protein detection: as the fragmentation efficiency increases, the correlation between the background proteins and the "signature proteins" decreases, improving the specificity of the detection. While the average top background score for N = 1 was 0.8, the score for N = 20 is 0.23, for N = 60 is 0.18 and for the greatest considered fragmentation efficiency of N = 150 is 0.16. The lower is the score, the less is the likelihood of incorrect detection of a "signature protein" in the background sample, and the greater is the noise tolerance in the detection scheme.

Bottom up fingerprint mode

The fragmentation efficiency that is considered a substantial factor for protein identification in the top down experiments does not play a significant part in the bottom up experiments. Smaller peptides tend to fragment more efficiently than the large proteins, in general providing good fragmentation coverage. In this simulation, therefore, the fragmentation efficiency of peptides was considered as full fragmentation – all of the possible background peptide fragments were included in the fingerprint. The measurement uncertainty used for the bottom up simulations was accepted as the ±3 Da. As previously shown in the MS mode section, the peptide mass can not be used for reliable peptide detection, the delta error of ±3 Da yielding to 3 peptides unique to the set of "signature peptides" in the background peptide mixture. In addition, ±3 Da is

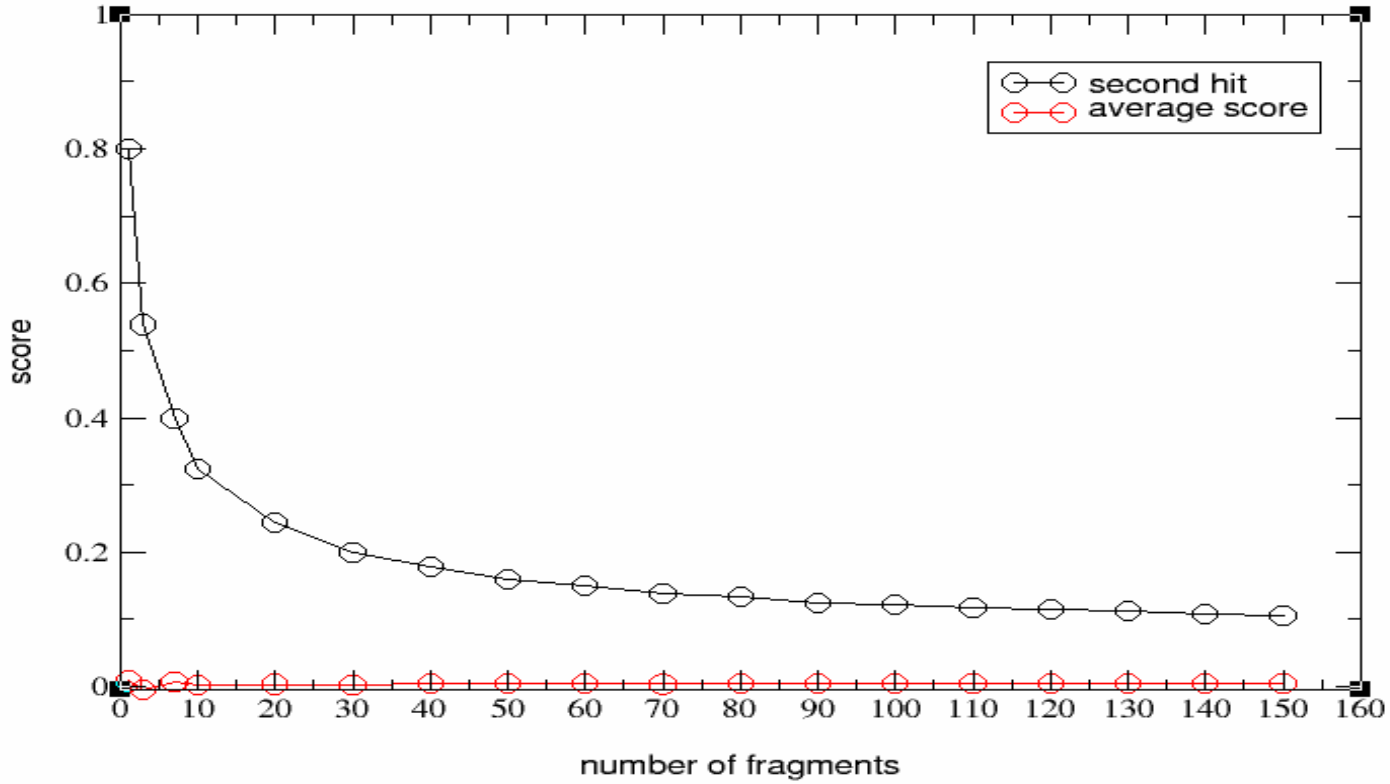

Figure 5.8 Dependency of correct detection on the number of fingerprint fragments. The greater is the fragmentation efficiency, the higher is the specificity of organism detection.

considered one of the accepted delta error settings during real database search procedures and expected to be a reasonable error for peptide parent mass matching.

The results of the peptide detection simulation in terms of top and average background peptide cross-correlation scores are shown in the figure 5.9. The peptides are arranged by masses, as it is shown, the smaller are the peptide (the lower is the parent mass) the higher are the scores of background peptide as comparing to the "signature peptides". In fact, there are many background peptides, for which cross-correlation scores are higher than 0.9, approaching 1. Such background peptides would likely produce erroneous peptide detection when the parent protein is not present in the background sample. It can be seen that at higher peptide masses the cross-correlation scores between the "signature peptides" and background peptides are significantly lower than at the lower masses. The reason behind this phenomenon can be easily explained, a short amino acid sequence is generally less specific than the longer one: the peptides with lower masses have a short amino acid sequence since there is a direct relationship between the mass and length of a peptide and a number of possible combinations of amino acids in a short sequence with matching parent mass is significantly smaller that in the long one, allowing for a greater chance of the same sequence occurring in both "signature peptides" and the background samples. The "signature peptides" with mass greater than 10,000 Da are as unique as the "signature proteins". Unfortunately, there are limitations to mass detection with the instrument used to perform the bottom up analysis, as this study is focused on ion trap MS which can detect a m/z range of 200-2000 Da. Even though the instrument detects mass to charge ratio (increasing the scope of the mass

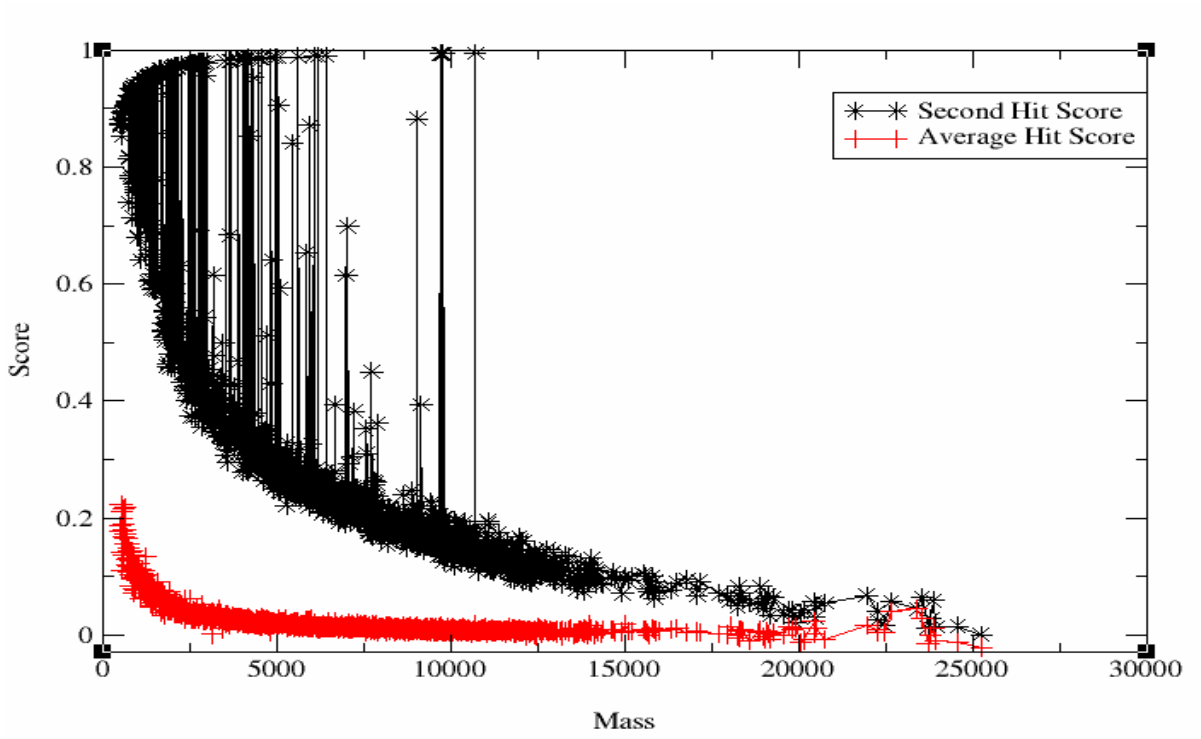

Figure 5.9 Peptide detection. Difference in scores for "signature peptides" fragmentation patterns as comparing to the best and average matches in the background. The "x" axis represents peptide mass and "y" axis represents normalized cross correlation score. With increasing mass, the cross correlation scores decrease (the specificity increases).

detection) there is no hope to detect peptides with very large masses. Therefore, for the purpose of this simulation, only peptides with masses less than or equal than 6,000 Da were used for further analysis. This mass cutoff has been judged extremely generous, since on average the peptides detected with the technology are of 2,000 – 4,000 Da. The specificity of cross-correlation for the bottom up method for peptides with mass less than 6,000 Da is shown in figure 5.10 as the percent of incorrect peptide detection as a function of the cross-correlation score cutoff. At the cross-correlation score cutoff of 0.4, 57% of "signature peptides" have been detected in the background mixture incorrectly, while with the score cutoff of 0.9, 19% of "signature peptides" are incorrectly detected. The percentage of incorrect identifications in bottom up analysis at a score cutoff of 0.4 is incomparable to that of top down (21.8%). Additionally, in bottom up simulation, while the number of incorrect identifications is reduced with increasing score cutoffs, it never reaches less than 19%. It is expected that there is a large overlap between the "signature peptides" and background peptides as can be seen in 5.9 (the background peptides which received the cross correlation score of 1 have the same sequence as some of the "signature peptides"). Because of this fact, at the highest score cutoffs there will still be some background peptides with high enough scores to be identified the "signature peptides".

Top down and bottom up comparison

The specificity of detecting "signature proteins" is severely dependant on fragmentation efficiency. Sparse fragmentation can frequently causes incorrect or inconclusive identifications since any of the fragments in one spectrum can match

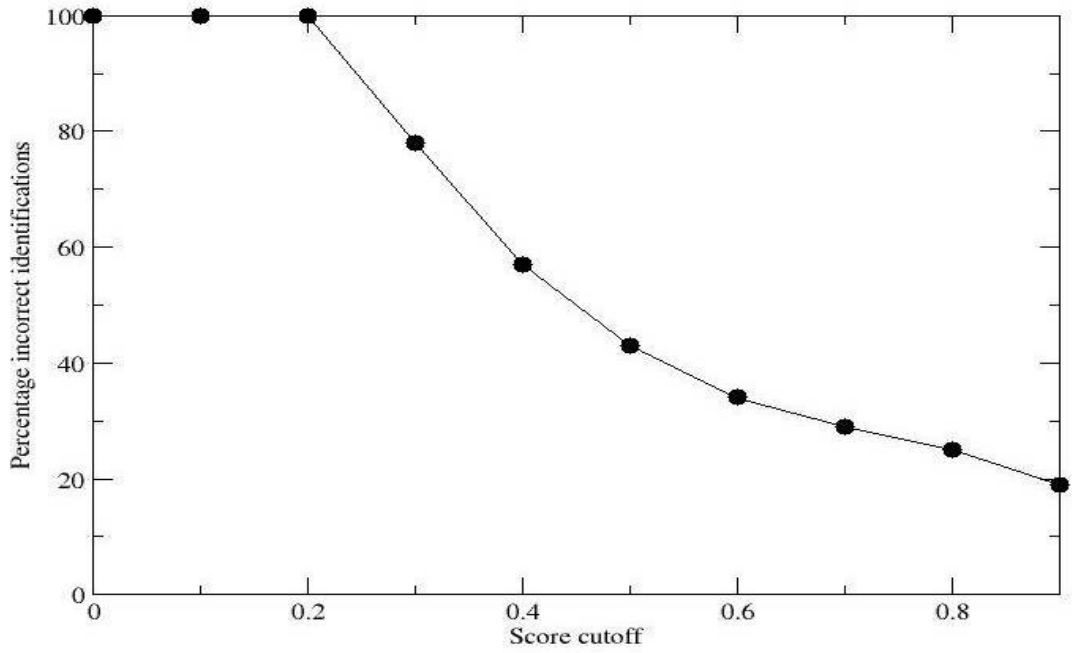

Figure 5.10 Specificity of peptide detection. Percent of background peptides matching to incorrectly to "signature peptides" with masses less than or equal to 6,000 Da as a function of the cross correlation score cutoff.

another spectrum by accident. The likelihood of correct identification increases with the number of matches while only when all major fragments between two spectra match (and none of the major fragments are mismatched), the identification can truly be considered confident. In case of top down tandem MS, the complete fragmentation cannot be expected, while the bottom up mass spectrometry is generally expected to provide complete fragmentation. To explore the advantages and disadvantages of the two methods a comparison must be made in terms of their specificity of protein detection. Since the fragmentation efficiency is a significant factor in the top down mass spectrometry, several fragmentation efficiencies must be considered for the comparison. As shown in the figure 5.11, the percent of incorrect identifications was compared between bottom up method (complete fragmentation efficiency) and top down method with three different fragmentation efficiencies.

The top down fragmentation efficiencies include the worst case scenario, where fragmentation efficiency is very low: $N = 5$, the mid-case scenario, where the fragmentation efficiency is somewhat efficient: $N = 20$, and the best case scenario, where the fragmentation is closest to being complete (in the scope of this experiment): $N = 150$, where, as previously mentioned, N stands for the number of allowed fragments per protein. There is a significant difference between the number of incorrect identifications in the three scenarios at the lower score cutoffs – the best case scenario performs dramatically better than both worst and mid-case up to cutoff of 0.3, at fragmentation of N = 150 when at the cutoff of 0.2, the number of incorrect identifications is already nearly negligible (around 5%), the efficient fragmentation ensuring the efficiency of the scoring. The mid-case fragmentation efficiency ($N = 20$) becomes comparable to the best

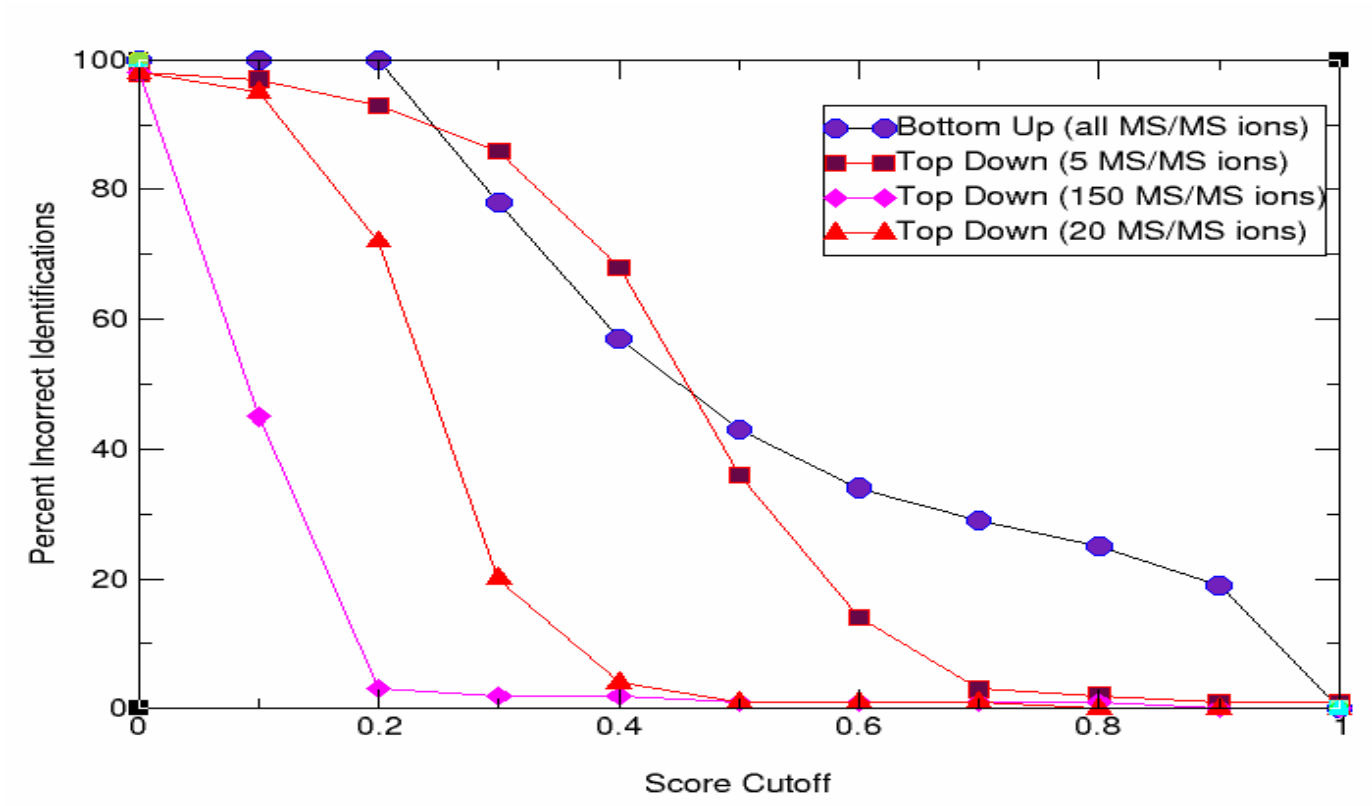

Figure 5.11 Comparison between top down and bottom up detection specificity. Percent of incorrect identifications as a function of cross correlation score cutoff. The results are shown for the bottom up simulation and for three fragmentation efficiency scenarios of top down simulation. The fragmentation efficiencies include the best ($N = 150$), worst ($N = 5$) and mid ($N = 20$) cases.

case at the score of 0.4, getting closer and closer to the best case performance at the increasing score cutoffs. The worst case scenario ($N = 5$) starts to show similar performance to the other cases only after a score cutoff of 0.7, showing 3%, where as the others two cases show 1% of incorrect identifications (in either case the percentage can be considered negligible). However, though the worst case scenario is still capable of showing a good performance at higher score cutoffs, the simulation contains no noise modeling which would necessarily decrease the performance of all three scenarios, and in likely case, would make the worst case scenario inadequate for protein identification.

As shown in the figure, bottom up's performance in the simulation is significantly inferior to top down in both the mid and the best case scenarios. Top down with fragmentation efficiency of $N = 20$, at cross-correlation cutoff of 0.4 shows less than 10% of incorrect identifications, while bottom up at the same cutoff shows 57%. The gap between performances is not reduced at the higher score cutoffs. The bottom up method is only comparable to the top down worst case scenario ($N = 5$). It shows a slightly higher specificity (lower percent of incorrect identifications) than the worst case scenario for top down method in the score cutoff region of 0.2 – 0.45, while performing slightly lower for all the other score cutoffs. This result suggests that in case of this simulation, bottom up technology could not be used to perform organism detection in the complex background due to the great overlap between the "signature peptides" and the background peptides. Even though, it is likely that bottom up method would be sensitive enough to detect peptides it lacks the necessary specificity to be useful for real detection.

OrganismScore

Specificity of protein detection plays a very important part in the detection of the organism of interest. However, the detection of "signature proteins" in the background mixture does not necessarily signify the presence of the organism represented by these "signature proteins". As the background sample becomes more complex (more organisms and proteins are included in the sample) the chances that some of the "signature proteins" appear in background organisms increase. While some of the "signature proteins" can be fairly distinctive of the organism of interest, others can be quite common in nature. Limiting the "signature proteins" to the proteins unique to the organism of interest is very desirable, however there are two negative factors to reducing the number of "signature proteins". Firstly, the proteome content is known only for a small fraction of existing organisms, making it difficult to differentiate between unique and non-unique proteins and secondly it can greatly reduce the number of "signature proteins" and where one unique protein cannot be used for positive organism detection, a detection of a combination of non-unique proteins can be more conclusive.

OrganismScore scheme, described in the Scoring Schemes section, utilizes protein uniqueness information as well as the number of detected proteins.

We estimate the uniqueness of each of the *E. coli* "signature protein" in nature using positional frequency matrix created using family based profiles. Each of the "signature proteins" received a weight coefficient (likelihood of a protein uniqueness), which determines its value to the organism detection. We then were able to calculate *OrganismScore* for each protein cross-correlation score cutoff. The *OrganismScores* are shown for four fragmentation efficiencies for top down mass spectrometry in Figure 5.12.

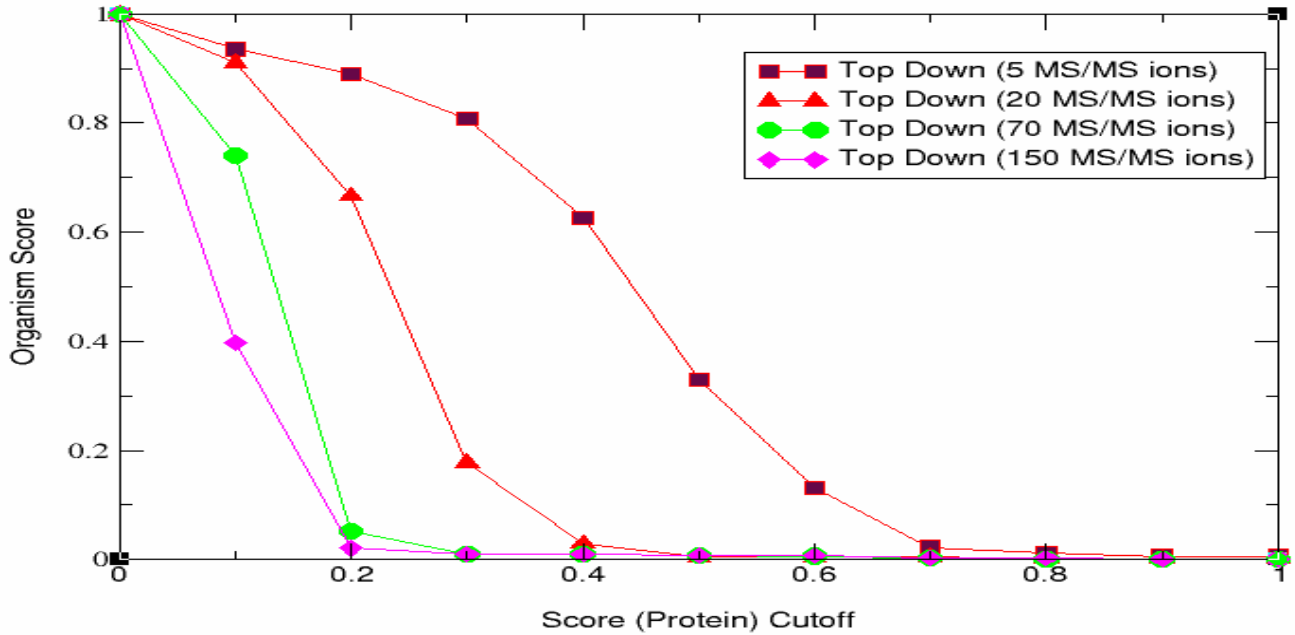

Figure 5.12 Specificity of organism detection with *OrganismScore.* The top down organism detection with OrganismScore is shown as a function of cross-correlation score cutoff. Top down fragmentation efficiencies include the best ($N = 150$), worst ($N = 5$) and two mid ($N = 20$, $N = 70$) cases.

The fragmentation efficiencies shown are $N = 5$, 20, 70 and 150. The specificity of *E. coli* detection in the background mixture using the *OrganismScore* shows improvement from the simple count of number of proteins – for example, the protein count for $N = 5$ fragmentation efficiency with score cutoff of 0.4 would produce a score of 0.69 (69% of incorrect identifications as shown in the figure 5.11), while the *OrganismScore* for the same setting produces a score of 0.62. The *OrganismScore* for score cutoff of 0.8 for $N = 20$, $N = 70$ and $N = 150$ all show the same results of 0.001 (where as protein count would have shown 0.1) and for $N = 5$, shows 0.0062 (where as protein count would have shown 0.2).

The described simulations of detecting *E. coli* in a complex background were fashioned to approximate a real life situation of detecting a biological agent in an environmental sample. The biological agent was represented by a set of 376 *E. coli* "signature proteins", which are readily detectable by mass spectrometry, while the environmental sample was represented by a mixture of twelve organisms, containing 83,777 background proteins. The detections were performed using two mass spectrometric approaches: the top down and bottom methods. The *in-silico* experiments were focused upon establishing the specificity of detecting a biological agent in a complex sample and exploring a set of instrumental parameters needed for such detection.

The bottom up method is one of the most established methods for organism's characterization. The instrumentation and procedures are optimized and robust making it a desirable approach for organism detection. However, there are a few expected drawbacks to the bottom up methodology in terms of speed: a) bottom up method

requires protein digestion before the analysis, to convert the proteins to shorter peptides, b) the analysis itself is very complex due the great number of peptides present in the mixture. In addition, the short peptides are generally not unique to a protein or an organism, frequently making peptide detection not indicative of an organism's presence. Since the instrumentation is well established, the set of explored instrumental parameters was limited. Here, the simulations only addressed the impact of parent mass accuracy measurements on specificity of organism detection, the fragmentation efficiency was considered complete basing on the fragmentation efficiency of the bottom up instruments.

The top down method is less established as a high throughput method, the instrumentation and procedures being significantly less developed than that of bottom up method. However, one of the goals of this simulation is to probe the thresholds of the current instrumentation and to reach a conclusion which of these two methods is more suited to solve the problem of biological agent detection. The current drawbacks of top down methodology are mostly involved with analyzing proteins with mass spectrometry, which is currently a difficult process in terms of protein separation, introducing proteins into mass spectrometer and protein fragmentation. Some of these concerns were addressed in the simulation with varying parameters in mass accuracy and fragmentation efficiency. It was noted that changes in mass accuracy have a comparably insignificant effect on the detection specificity after mass accuracy of ±5 Da for top down (Figure 5.4) and ±1 Da for bottom up (Figure 5.5), therefore, the simulations were focused on reasonable mass accuracy ranges instead of performing the simulations with all mass accuracies.

According to the *in-silico* simulations the top down methodology is inherently better suited to the organism detection than the bottom up method. The comparisons of specificity of *E. coli* detection by the two methods is shown in figure 5.11. The analysis suggests that if top down fragmentation efficiency is greater than $N = 5$ (there are at least 5 fragment ions per protein) while the accuracy of protein mass measurement is no lower than ±10 Da, it will be more specific in organism detection than the bottom up with complete fragmentation efficiency and accuracy of peptide mass measurement of ±3 Da. However, if there is no fragmentation in the top down method (only parent mass is available) or the fragmentation efficiency is low (less than $N = 5$), top down method loses its specificity as comparing to the bottom up. Even though, as illustrated in the Figure 5.4, when accurate parent mass is available (measurement accuracy is 10^{-4} Da) almost all protein masses are unique to the *E. coli* "signature proteins", the background sample might become more complex rendering protein masses less specific. For the top down experiment the preferred fragmentation efficiency based on this study is $N = 20$, a point, where the specificity of detection does not change significantly with the increasing number of fragments. It must also be noted that the protein fragmentation efficiency was modeled by randomly choosing N ions from all the fragments produced from complete protein fragmentation. In reality, the fragmentation is likely to follow different pattern, and protein fingerprint might be less meaningful.

The last issue examined in the study was establishing a reasonable scoring scheme for organism detection. *OrganismScore* was designed to provide the assessment of likelihood of organism detection and allow for comparison between detections of different organisms. *OrganismScore* is based upon the concept of uniqueness of a given

protein in nature. The protein uniqueness is measured with the use of family based profile of the protein. In essence, it is a measurement of amino acid conservation in the protein sequence. Using a family based position dependent frequency matrix example of such a matrix shown in the Figure 5.3, derived from the family based profile, it is possible to compute the probability of all amino acids in the sequence remaining the same in a different organisms. *OrganismScore* is normalized to 1, which allows for easy assessment of organism detection as well as the comparison between *OrganismScores* for different organisms in different environmental samples.

Mass spectrometry is an analytical tool that can detect proteomic signatures in complex samples. As shown in this simulation, the complex background does not overly interfere with the specificity of signature detection, while sensitivity of detection will have to be explored by real experimental studies. One of the advantages of top down mass spectrometry over the other techniques for organism detection is that the organism does not have to be well characterized or sequenced before it can be detected by top down mass spectrometry. The detection can be done based on a proteomic signature pattern of unknown proteins as long as it has been measured once before. Indeed, top down mass spectrometry can be used to detect changes in the environment, by separating the signatures of organisms present in the background from any new organism signature which appears in the sample. This detection is easily achievable by subtracting the previously detected background signal from the measured spectrum of the environmental sample leaving only the newly emerging signatures. If the new signatures are judged harmless, they can be added to the background signal and won't be considered in the new

analysis. This approach can potentially make bio-organism signature detection completely automatic and highly efficient.

LIST OF REFERENCES

Aebersold R., Mann M., *Nature*, 2003, 422(6928):198-207.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J., *J.Mol.Biol*.,1990, 215:403-410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D., *J. Nucleic Acids Res*.,1997, 25:3389-3402.

Atkinson M.R., Kamberov E.S., Weiss R.L., Ninfa A.J., *J.Biol.Chem.*, 1994, Nov 11;269(45):28288-93.

Biemann K., *Anal Chem*., 1986 Nov;58(13):1288A-1300A.

Biemann K., *Biomed. Env. Mass Spec*., 1988, 16, 99.

Blattner F.R., Plunkett G. IIIrd, Bloch C.A., Perna N.T., Burland V., Riley M., Collado-Vides J., Glasner J.D., Rode C.K., Mayhew G.F., Gregor J., Davis N.W., Kirkpatrick H.A., Goeden M.A., Rose D.J., Mau B., Shao Y., *Science*, 1997, Sep 5; 277(5331):1453-74.

Clauser K. R., Baker P. R. and Burlingame A. L., *Anal. Chem*., 1999,71, 14, 2871.

Colinge J., Magnin J., Dessingy T., Giron M., Masselot A., *Proteomics*., 2003, Aug; 3(8):1434-40.

Dancik V., Addona T, Clauser K., Vath J., Pevzner P., *J. Comp Biol*., 1999, 6, 327-342.

Edman P., *Acta Chem. Scand*., 1950, 4, 283-293.

Eng, J., McCormack, A., Yates, J.R. III, J. *Am. Soc. Mass Spectrom*., 1994, 5, 976-989.

Eriksson, J., Chait, B. T., Fenyo, D., *Anal. Chem*., 2000, 72, 999-1005.

Falquet L., Pagni M., Bucher P., Hulo N., Sigrist C.J., Hofmann K., Bairoch A., *Nucleic Acids Res*., 2002 Jan 1;30(1):235-8.

Farriol-Mathis N., Garavelli J.S., Boeckmann B., Duvaud S., Gasteiger E., Gateau A., Veuthey A.L., Bairoch A., *Proteomics*., 2004 Jun;4(6):1537-50.

Fenn J.B., Mann M., Meng C.K., Wong S.F., Whitehouse C.M., *Science*, 1989, 246, 64-71.

Field, H.I., Fenyo, D., Beavis, R.C., *Proteomics*, 2002, 2, 36-47.

Fraser, C.M. et al., *Science*, 1995, 270, 397-403.

Fridman T., Day R., Razumovskaya J., Xu D., Gorin A., *CBS2003 conference proceedings*, Stanford University 11-14 August 2003, p.415-418.

Garavelli J.S., *Proteomics*, 2004 Jun;4(6):1527-33.

Gatlin C.L., Kleeman C.R., Hays L.G., Link A.J., Yates JR. III, *Anal Biochem*., 1998, 263, 93-101.

Gavin, A., Bosche, M., Krause, R., Grandi, P. et al., *Nature*, 2002, 415, 141-1147.

Geer L.Y., Markey S.P., Kowalak J.A., Wagner L., Xu M., Maynard D.M., Yang X., Shi W., Bryant S.H., *J. Proteome Res*., 2004, 5, 958-964.

Goodlett, D.R., Keller, A., Watts, J.D., Newitt, R., et al. *Rapid Commun. Mass Spectrom*., 2001, 15, 1214-1221.

Heidelberg, J.F. et al. *Nat. Biotechnol*., 2002, 20, 1118-1123.

Hess D., Covey T.C., Winz R., Brownsey, R.W., Aebersold R., *Protein Sci*., 1993, 2, 1342-1351.

Hillenkamp F., Karas M., Beavis R.C., Chait B.T., *Anal Chem*., 1991 Dec 15;63(24):1193A-1203A.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D. et al. *Nature*, 2002, 415, 180-183.

Hofmann K., Bucher P., Falquet L., Bairoch A., *Nucleic Acids Res*., 1999 Jan 1;27(1):215-9

Hunt D.F., Buko A.M., Ballard J.M., Shabanowitz J., Giordani A.B., *Biomed Mass Spectrom*., 1981 Sep;8(9):397-408.

Kelleher N.L., Taylor S.V., Grannis D., Kinsland C., Chiu H.J., Begley T.P., McLafferty F.W., *Protein Sci*., 1998 Aug;7(8):1796-801.

Keller, A., Nesvizhskii A., Kolker, E., Aebersold, R., *Anal.Chem*., 2002 74, 5383-5392.

Keller, A., Purvine, S., Nesvizhskii, A., Stolyar, S. et al., *OMICS J. Integrative Biology*, 2002, 6 2, 207-212 (Short Communication) *.

Larimer F.W., Chain P., Hauser L., Lamerdin J., Malfatti S., Do L., Land M.L., Pelletier D.A., Beatty J.T., Lang A.S., Tabita F.R., Gibson J.L., Hanson T.E., Bobst C., Torres J.L., Peres C., Harrison F.H., Gibson J., Harwood C.S., 2004, *Nat Biotechnol*., 2004 Jan;22(1):55-61.

LeDuc R., Taylor G., Kim Y., Januszyk T., Bynum L., Sola J., Graravelli J., Kelleher N., *Nucleic Acid Res*., 2004, (32), (Web Server issue).

Liebler, D., "Introduction to Proteomics", Humana Press, Totowa, New Jersey, 2001.

Loo J.A., Quinn J.P., Ryu S.I., Henry K. D., Senko M. W., McLafferty F. W., *Proc Natl Acad Sci USA*, 1992, Jan 1; 89(1)286-9.

Ma B., Zhang K., Hendrie C., Liang C., Li M., Doherty-Kirby A., Lajoie G. *Rapid Communication in Mass Spectrometry*,17(20): 2337-2342. 2003.

MacCoss M. J., McDonald W. H., Saraf A., Sadygov R., Clark J.M., Tasto J. J. Gould K.L., Wolters D. Washburn M., Weiss A., Clark J. I., Yates J.R.III, *PNAS*, 2002, 99, 12, 7900-5.

MacCoss, M. J., Wu, C. C., Yates J.R. III, *Anal. Chem*., 2002, 74,5593-5599.

Mann M., Wilm M.., *Anal Chem*., 1994 Dec 15;66(24):4390-9.

Mann M., Olsen J., *Nat. Biotechnol*. 2003 21(3):255-61.

Margulis and Scwartz, "Five Kingdoms" 3rd edition, W.H. Freeman and Co., 1998.

Martin, S.E., Shabanowitz., Hunt, D., Marto, J., *Anal. Chem*., 2000, 72, 4266-4274.

McCormack, A.L., Schieltz, D. M., Goode, B., Yang, S., Barnes, G., Drubin, D., Yates, J.R. III, *Anal. Chem*., 1997, 69, 767-776.

McLuckey, S., Stephenson, *J. Mass Spectrom. Rev*., 1998, 17, 369-407.

Meng F., Du Y., Miller L.M., Patrie S.M., Robinson D.E., Kelleher N.L.., *Anal Chem*., 2004 May 15;76(10):2852-8.

Mortz E., O'Connor P.B., Roepstorff P., Kelleher N.L., Wood T.D., McLafferty F.W., Mann M., *Proc Natl Acad Sci USA*, 1996, Aug 6;93(16):8264-7.

Pandley, A., Mann, M., *Nature*, 2000, 405, 837-846.

Perkins, D.N., Pappin, D.J., Creasy, D.M., Cottrell, J. S., *Electrophoresis*, 1999, 20, 3551-3567.

Razumovskaya J., Olman V., Xu D., Uberbacher E., Nathan Verberkmoes, Hettich R.L., Xu Y., *Proteomics*, 2004., Apr; 4(4):961-9.

Roepstorff P., Fohlman J., *Biomed. Mass Spec*., 11 (1984) 601.

Sadygov R.G., Eng J., Durr E., Saraf A., McDonald H., MacCoss M.,J., Yates JR III., *J. Proteome Res*., 2002, May-Jun; 1(3):211-5.

Shen, Y., Zhao, R., Belov, M., Conrads, T., Anderson, G., Tang, K., Pasa-Toli, L., Veenstra, T., Lipton, M., Udseth, H., Smith, R., *Anal Chem*., 2001, 73, 1766-1775.

Sunyaev S., Liska A., Golod A., Shevchenko A., Shevchenko A., *Anal. Chem*., 2003, 75, 1307-1315.

Tabb D.L., Saraf A., Yates JR. III., *Anal Chem*., 2003 Dec 1;75(23):6415-21.

Tabb, D.L., McDonald, W.H., Yates Jr. III, *J. Proteome Res*., 2002, 1:21-26.

Tabb, D.L., Narasimhan, C., Strader M.B., Hettich, R.L., *Anal Chem*., 2005 Apr 15;77(8):2464-74.

Tatusov R.L., Koonin E.V., Lipman D.J., *Science*, 1997 Oct 24;278(5338):631-7.

Taylor J.A., Johnson R.S., *Anal Chem*., 2001 Jun 1;73(11):2594-604.

Taylor, J.A., Johnson, R.S. *Rapid Commun. Mass Spectrom*.,1997, 11, 1067-1075.

VerBerkmoes N.C., et el. Submitted *J. Proteome Research*, 2005.

Wilkins M.R., Gasteiger E., Gooley A.A., Herbert B.R., Molloy M.P., Binz P.A., Ou K., Sanchez J.C., Bairoch A., Williams K.L., Hochstrasser D.F., *J. Mol Biol*., 1999, Jun;289(3):645-57.

Yates, JR., III, Eng, J. K., McCormack, A.L., *Anal. Chem*., 1995, 15, 3202-10.

Yates, JR., III, Eng, J. K., McCormack, A.L., Schieltz, D., *Anal. Chem*., 1995, 67 (8), 1426-1436. *

Zubay, "Biochemistry", 3rd edition, Wm. C. Brown Publishers, 1993.

APPENDIX 1

Modification Name	Peptide	# of Occurrence	MW of PTM
N6,N6,N6-trimethyl-L-lysine_	NYDPRAKIMQQVCHEVLAETGHHGDPLLK!	3	43.0548
O5-glycosyl-L-hydroxylysine_	GK!QGPVGKPGPQGK	12	178.0477+
trans-2,3-cis-3,4-dihydroxy-L-proline_	VAPP!PPPQHAAPRMAPPPAPVRAAPPPPHVAPPR	3	31.9898
L-3-oxoalanine_	QLPGKDFSS!VLTNPSSADIHAVR	3	-2.0156
trans-2,3-cis-3,4-dihydroxy-L-proline_	PSAPPTAAPAERPAAPP!PAAAPVRPPAPPAGEAPQR	4	31.9898
S-(L-isoglutamyl)-L-cysteine_	TLSRYQ!LSDLGNERGAK	5	-17.0265
N-acetyl-L-methionine_	AVFMTGHGGNEVIEVGDRPM!PQR	8	42.0106
N6-1-carboxyethyl-L-lysine_	VNAVNPGMVVTEGVK!	4	72.0211
N6-biotinyl-L-lysine_	EAASARWMK!EADK	6	226.0776
N6-lipoyl-L-lysine_	AAGAGWK!ASAGGAPSPQR	3	188.033
N6-biotinyl-L-lysine_	YPNDK!	12	226.0776
N5-methyl-L-glutamine_	VEEFRVSEDALLPVGAEIQADHFVVGQ!FVDVTGTSTGK	3	14.0157
trans-2,3-cis-3,4-dihydroxy-L-proline_	PAAPPP!AAAPVR	3	31.9898
N-acetyl-L-methionine_	LAM!AGFAAARALSTGFNDAPTKASR	3	42.0106
N6-1-carboxyethyl-L-lysine_	REALDALAAK!LGER	3	72.0211
trans-2,3-cis-3,4-dihydroxy-L-proline_	PSAPPTAAPAERPAAPPPAAAPVRPPAPP!AGEAPQR	5	31.9898
O4'-(phospho-5'-uridine)-L-tyrosine_	GAEY!IVDFLPK	3	306.0253
N5-methyl-L-glutamine_	VFTEAGEHIPVTVLKLGNCQ!VLGHRTK	3	14.0157
L-serine_	AS!ADVALLK	3	87.032
N-acetyl-L-proline_	FPEPYLAAFDGP!R	3	42.0106
trans-2,3-cis-3,4-dihydroxy-L-proline_	GLPPAPGVAARPGIP!SVAQPQPPGRPALGPGGPAAAR	3	31.9898
3'-(1'-L-histidyl)-L-tyrosine_	YVDY!PDAFAGWNLVSSIGSYISGFAVLVFLYGMTLAFIRKER	3	-2.0156
L-3-oxoalanine_	ITRS!DEAIAAK	4	-2.0156
N6-1-carboxyethyl-L-lysine_	K!TIPAPAQALDAEANR	4	72.0211
N6-1-carboxyethyl-L-lysine_	NAAEVDGAVAALK!	3	72.0211
N-acetyl-L-methionine_	RVVVTGM!GIVSSIGNNTQEVLASLHDAKSGISR	5	42.0106
N6,N6,N6-trimethyl-L-lysine_	LSVEAGSVKMFEIADRIEAVMHESK!	3	43.0548
N6-1-carboxyethyl-L-lysine_	GTDHKIGQLNPLK!R	3	72.0211

trans-2,3-cis-3,4-dihydroxy-L-proline_	AAPPPPHVAPPRPPAPPRAAPPP!R	3	31.9898
trans-2,3-cis-3,4-dihydroxy-L-proline_	VAPPPP!PQHAAPRMAPPPAPVRAAPPPPHVAPPR	4	31.9898
N6-1-carboxyethyl-L-lysine_	EGAAVVVNDLGGPRDGSGSDAGMAQQVVDAIK!	3	72.0211
N6-1-carboxyethyl-L-lysine_	GIGVEIALK!LAAEGAAVAVNYASSKQGADDVVDK	5	72.0211
N,N-dimethyl-L-proline_	NMVGP!ALGGVVGR	4	29.0391
trans-2,3-cis-3,4-dihydroxy-L-proline_	NGTVAPSAGSAP!KPLAGTPPAGGGPAVRPEAVR	3	31.9898
trans-2,3-cis-3,4-dihydroxy-L-proline_	LRPTPPVTAPARPAGPP!PAAAVDR	3	31.9898
N6-1-carboxyethyl-L-lysine_	VIGANLK!GAYFLATEVAR	3	72.0211
N-acetylglycine_	LLQTASVDQG!SK	6	42.0106
O5-glycosyl-L-hydroxylysine_	QGPVGK!PGPQGKAGPQGK	4	178.0477+
trans-2,3-cis-3,4-dihydroxy-L-proline_	DTEIP!TEGLR	4	31.9898
N-acetyl-L-alanine_	SVA!AR	3	42.0106
trans-2,3-cis-3,4-dihydroxy-L-proline_	MAPPPAPVRAAPPPPHVAPPRPP!APPR	3	31.9898
N-acetyl-L-proline_	FGRP!LLGATVK	6	42.0106
N6-biotinyl-L-lysine_	VAHTAEDLALAISTAGNEAKAAFGDASVYLEK!	4	226.0776
omega-N-(ADP-ribosyl)-L-arginine_	NVFIHGCDPKADSTR!LILGGK	3	541.061
3-hydroxy-L-proline_	P!GIAGKPGPDGKPGPIGPQGK	5	15.9949
N6-lipoyl-L-lysine_	SGDVIAEIETDK!ATMEVEAADEGTLAK	7	188.033
L-3-oxoalanine_	GS!LRSTYDGR	4	-2.0156
3-hydroxy-L-proline_	GPKGEAGAAGAP!GPAGPAGPAGPAGPAGPKGDAGPAGPAGPAGPAGPSGATGPAGPK	4	15.9949
trans-2,3-cis-3,4-dihydroxy-L-proline_	PAAPPPSPAGPP!AR	9	31.9898
N-acetyl-L-alanine_	EVPEA!IRKATESAK	3	42.0106
trans-2,3-cis-3,4-dihydroxy-L-proline_	GLP!PAPGVAARPGIPSVAQPQPPGRPALGPGGPAAR	5	31.9898
N6-1-carboxyethyl-L-lysine_	ALK!AAGYK	3	72.0211
trans-2,3-cis-3,4-dihydroxy-L-proline_	PSAPPTAAPAERPAAPPPAAAPVRPPAP!PAGEAPQR	3	31.9898
trans-2,3-cis-3,4-dihydroxy-L-proline_	PPAPPAGEAPQRRGPP!PGAVPPNAVPPNAAAPDAAK	4	31.9898
3-hydroxy-L-proline_	PGIAGKPGP!DGKPGPIGPQGK	4	15.9949
trans-2,3-cis-3,4-dihydroxy-L-proline_	PAAPPPSP!AGPPAR	3	31.9898
dehydroalanine_	AYRDILPESSPSELLIAVAGDYNY!VLPTLLVADR	3	-94.0419
N5-methyl-L-glutamine_	GQ!FAAAKVEPK	6	14.0157
trans-2,3-cis-3,4-dihydroxy-L-proline_	GLPPAPGVAARP!GIPSVAQPQPPGRPALGPGGPAAR	3	31.9898
N6-1-carboxyethyl-L-lysine_	LMMRK!K	3	72.0211
trans-2,3-cis-3,4-dihydroxy-L-proline_	P!EPKPAPGPLR	3	31.9898
trans-2,3-cis-3,4-dihydroxy-L-proline_	AEPAMPRP!PR	4	31.9898

N6,N6,N6-trimethyl-L-lysine_	GGAWTFDELNK!FLASPKGYIPGTAMSFAGVPNDK	3	43.0548
dehydroalanine_	AYRDILPESSPSELLIAVAGDY!NYVLPTLLVADR	3	-94.0419
N-acetyl-L-methionine_	MRAATVNRVDLYM!R	3	42.0106
N6-biotinyl-L-lysine_	EDNQLSDYLLGTLPELVPGDVKARYPNDK!	4	226.0776
S-methyl-L-cysteine_	DETQIHTHMC!YSEFNDIIDAIAAMDADVISIETSR	3	14.0157
trans-2,3-cis-3,4-dihydroxy-L-proline_	PSAP!PTAAPAERPAAPPPAAAPVRPPAPPAGEAPQR	4	31.9898
trans-2,3-cis-3,4-dihydroxy-L-proline_	APESAAPAAATPKP!AAPPPSPAGPPARR	5	31.9898
trans-2,3-cis-3,4-dihydroxy-L-proline_	PSAPP!TAAPAERPAAPPPAAAPVRPPAPPAGEAPQR	4	31.9898
N6-1-carboxyethyl-L-lysine_	SQSPRIVNIASTEALGATATHSPYSAAK!AGVTGLTR	3	72.0211
trans-2,3-cis-3,4-dihydroxy-L-proline_	APESAAPAAATPKPAAPPPSPAGPP!ARR	3	31.9898
trans-2,3-cis-3,4-dihydroxy-L-proline_	GLPP!APGVAARPGIPSVAQPQPPGRPALGPGGPAAAR	5	31.9898
N6-biotinyl-L-lysine_	LFGDKVAAK!ELAK	3	226.0776
N6-1-carboxyethyl-L-lysine_	TLALHGAQVVLVNLK!HESGEAAARAITNAGGDAR	7	72.0211
trans-2,3-cis-3,4-dihydroxy-L-proline_	GLPPAP!GVAARPGIPSVAQPQPPGRPALGPGGPAAAR	4	31.9898
N6-carboxy-L-lysine_	NVIDGRAMIASFLTLTIGNNQGMGDVEYAK!	3	43.9898
trans-2,3-cis-3,4-dihydroxy-L-proline_	TSVVVLDLAREQPFVP!GGSVASGLAMVEPNPK	3	31.9898
trans-2,3-cis-3,4-dihydroxy-L-proline_	GPPPGAPGTPPNATAP!GMTPPPGEAPRR	4	31.9898
N6-1-carboxyethyl-L-lysine_	ARGTIVNTASISGLFGDYGFAAYNAAK!GAVINLTR	7	72.0211
N6-biotinyl-L-lysine_	HVADLVEAAQQFDQPLIATDGADNRSSAAAAASK!	3	226.0776
trans-2,3-cis-3,4-dihydroxy-L-proline_	PPP!AAPRIQR	3	31.9898
trans-2,3-cis-3,4-dihydroxy-L-proline_	PEAP!AAEPNKGEAGAAPK	4	31.9898
N-acetyl-L-methionine_	M!RALTLVADR	3	42.0106
trans-2,3-cis-3,4-dihydroxy-L-proline_	AEPAMPRPP!R	3	31.9898
N-acetyl-L-methionine_	TEPPQEASSDQGGLHSVSM!ESKMSGDEVSKALIK	3	42.0106
N6,N6,N6-trimethyl-L-lysine_	LLQTASVDQGSKVAK!	3	43.0548
trans-2,3-cis-3,4-dihydroxy-L-proline_	PAPPTVSRPVPP!PMHVAPRVAPPPPPQHAAPR	3	31.9898
trans-2,3-cis-3,4-dihydroxy-L-proline_	APESAAP!AAATPKPAAPPPSPAGPPARR	3	31.9898
trans-2,3-cis-3,4-dihydroxy-L-proline_	EKPAQP!EAAKPEAAK	4	31.9898
trans-2,3-cis-3,4-dihydroxy-L-proline_	PPAPPAGEAPQRRGPPP!GAVPPNAVPPNAAAPDAAK	3	31.9898
N6-1-carboxyethyl-L-lysine_	VALVTGASKGIGVEIALKLAAEGAAVAVNYASSK!	3	72.0211
N6-1-carboxyethyl-L-lysine_	AHALGLAALGAK!	4	72.0211
N6-biotinyl-L-lysine_	VHVLAEAVEK!AK	4	226.0776
N-acetyl-L-methionine_	NIAESLDKMAAGM!LPVIDTEVPLDDVGAALKR	3	42.0106

N-acetyl-L-proline_	VTIAHPHGNFGAKIP!NLLSAVCGEGVFFSPGIPLIRLQDIR	3	42.0106
trans-2,3-cis-3,4-dihydroxy-L-proline_	AKEAP!PPGPRPAPAPPK	3	31.9898
3-(3'-L-histidyl)-L-tyrosine_	FMEGFGVH!TFRLVNADGESTFVK	3	-2.0156
trans-2,3-cis-3,4-dihydroxy-L-proline_	PEPPLQPLP!RTEPPMPRVEAPIMR	3	31.9898
3-hydroxy-L-proline_	GPKGEAGAAGAPGPAGPAGPAGP!AGPAGPKGDAGPAGPAGPAGPAGPSGATGPAGPK	3	15.9949
L-3-oxoalanine_	RDFLGLAMGAVAAGTSS!TVLGPTTAAAQAQPGGGSLPRK	3	-2.0156
trans-2,3-cis-3,4-dihydroxy-L-proline_	SRQDAPSEP!KQR	3	31.9898
S-(L-isoglutamyl)-L-cysteine_	ALTLCAGLALGLASAQ!AADKAFQRNELADAAIK	5	-17.0265
trans-2,3-cis-3,4-dihydroxy-L-proline_	PGIPSVAQPQPPGRPALGPGGP!AAARNGTVAPSAGSAPK	3	31.9898
N6-1-carboxyethyl-L-lysine_	IGQLNPLK!	4	72.0211
N6-carboxy-L-lysine_	EGLSVVMLMPMIVGLANFHLIAK!	6	43.9898
trans-2,3-cis-3,4-dihydroxy-L-proline_	RP!ADPFASLVPEPIAR	3	31.9898
N6-carboxy-L-lysine_	PK!AGFGNFIQTAAHFAAESSTGTNVEVSTTDDFTRGVDALVYEVDEANSLMK	8	43.9898
S-(L-isoglutamyl)-L-cysteine_	LLTTQSLQ!VK	3	-17.0265
N6-biotinyl-L-lysine_	AERDGTVKK!	3	226.0776
trans-2,3-cis-3,4-dihydroxy-L-proline_	QPPGERRGP!PPGAPGTPPNATAPGMTPPPGEAPR	5	31.9898
trans-2,3-cis-3,4-dihydroxy-L-proline_	GP!PPGAVPPNAVPPNAAAPDAAKPDAAK	3	31.9898
N-acetyl-L-proline_	P!KAGFGNFIQTAAHFAAESSTGTNVEVSTTDDFTRGVDALVYEVDEANSLMK	10	42.0106
trans-2,3-cis-3,4-dihydroxy-L-proline_	LTTPEPGQWEADTAAELPEPELPPLP!TRPLR	3	31.9898
N6-biotinyl-L-lysine_	EDNQLSDYLLGTLPELVPGDVK!ARYPNDK	3	226.0776
trans-2,3-cis-3,4-dihydroxy-L-proline_	APESAAPAAATP!KPAAPPPSPAGPPARR	4	31.9898
trans-2,3-cis-3,4-dihydroxy-L-proline_	APESAAPAAATPKPAAPPP!SPAGPPARR	3	31.9898
trans-2,3-cis-3,4-dihydroxy-L-proline_	QPPGERRGPP!PGAPGTPPNATAPGMTPPPGEAPR	5	31.9898
N6-lipoyl-L-lysine_	VK!SGDVIAEIETDKATMEVEAADEGTLAK	3	188.033
N6-methyl-L-lysine_	VGRKLFVK!	5	14.0157
N6-1-carboxyethyl-L-lysine_	VNVVAPGGARTPIWK!	3	72.0211
N-acetyl-L-alanine_	EVPEA!IR	3	42.0106
N6-1-carboxyethyl-L-lysine_	RLSPQGIERAFAINHLGPFLLTNLLLDLIK!	3	72.0211
N6-biotinyl-L-lysine_	LGIPVVPGSDGGVGPDDDAMAIAKEIGFPVLVK!	3	226.0776
trans-2,3-cis-3,4-dihydroxy-L-proline_	AAPSAPPPPP!AAAPPHVAPPPPPPAPPR	3	31.9898
N6-1-carboxyethyl-L-lysine_	TNLTAVFFTVQAALPYLNDGASIILNGSVISVLGNPGFAAYAASK!	7	72.0211
N6-1-carboxyethyl-L-lysine_	K!IR	4	72.0211

dehydroalanine_	IASIY!HGYPSK	3	-94.0419
N-acetyl-L-methionine_	ARVARM!QQMGPEK	3	42.0106
L-3-oxoalanine_	DFLGLAMGAVAAGTSSTVLGPTTAAAQAQPGGGS!LPR	3	-2.0156
N6-lipoyl-L-lysine_	SGDVIAEIETDK!ATMEVEAADEGTLGK	4	188.033
trans-2,3-cis-3,4-dihydroxy-L-proline_	VAPPP!PPQHAAPRMAPPPAPVRAAPPPPHVAPPR	3	31.9898
dehydroalanine_	FDY!ATPLTR	4	-94.0419
N5-methyl-L-glutamine_	SGVIAQ!K	4	14.0157
trans-2,3-cis-3,4-dihydroxy-L-proline_	PEPP!VLR	4	31.9898
N-acetyl-L-methionine_	FLGEGAAWNHVAM!EQAIADSGLEESEISNIR	4	42.0106
omega-N-(ADP-ribosyl)-L-arginine_	SLTVTQAELSGRTTIEAAPQSAQADVYRQLAR!	3	541.061
trans-2,3-cis-3,4-dihydroxy-L-proline_	PEP!PVLR	3	31.9898
N6-biotinyl-L-lysine_	IGFPLMLK!STAGGGGIGMQLCHDEATLRER	4	226.0776
trans-2,3-cis-3,4-dihydroxy-L-proline_	AAPPPRPQGP!AK	3	31.9898
N6-carboxy-L-lysine_	QLDIVRREGLSVVMLMPMIVGLANFHLIAK!	7	43.9898
N-acetyl-L-proline_	LFDGP!STTIKDLWR	6	42.0106
N5-methyl-L-glutamine_	VSEDALLPVGAEIQADHFVVGQ!FVDVTGTSTGKGFAGGMK	3	14.0157
N-acetyl-L-methionine_	AAGGRAVANTADISTM!AGGQSVFDDAIKHFGR	4	42.0106
N-acetyl-L-methionine_	LQPGETVLVFGVGGGVSLAAM!QIAAAAGARVLATSRSADK	3	42.0106
trans-2,3-cis-3,4-dihydroxy-L-proline_	PALGPGGPAAARNGTVAPSAGSAPKP!LAGTPPAGGGPAVR	4	31.9898
L-3-oxoalanine_	RDFLGLAMGAVAAGTSSTVLGPTTAAAQAQPGGGS!LPRK	5	-2.0156
N-acetyl-L-methionine_	ALGADAVIDAPADKIPAAVM!DLTSGR	3	42.0106
N6-1-carboxyethyl-L-lysine_	TLQGKVALVTGASKGIGVEIALK!	4	72.0211
trans-2,3-cis-3,4-dihydroxy-L-proline_	GPP!PGAVPPNAVPPNAAAPDAAKPDAAK	3	31.9898
3-hydroxy-L-proline_	GPKGEAGAAGAPGPAGPAGPAGPAGP!AGPKGDAGPAGPAGPAGPAGPSGATGPAGPK	5	15.9949
trans-2,3-cis-3,4-dihydroxy-L-proline_	GEAGAAP!KQGAGK	3	31.9898
N6-1-carboxyethyl-L-lysine_	VNGVAPGPVDTAMAK!QVHTADIRSDYR	3	72.0211
N6,N6,N6-trimethyl-L-lysine_	GYPIEQLAEK!	4	43.0548
N6-1-carboxyethyl-L-lysine_	ARGGGAIVNIGSRSSVNAYGGGAAYCASK!	3	72.0211
trans-2,3-cis-3,4-dihydroxy-L-proline_	P!EAPAAEPNKGEAGAAPK	3	31.9898
N-methyl-L-alanine_	KNIASGIAHVNSSFNNTTITITDAQGNA!IAWSSAGTMGFK	3	14.0157
N6-biotinyl-L-lysine_	LFGDK!VAAKELAK	7	226.0776
trans-2,3-cis-3,4-dihydroxy-L-proline_	GLPPAPGVAARPGIPSVAQP!QPPGRPALGPGGPAAAR	3	31.9898
N6-1-carboxyethyl-L-lysine_	NAVKNHAALATMANAPGK!	5	72.0211
N6-lipoyl-L-lysine_	AAAPAAAPAPAAPAPAAAPAAK!APPSDAPLAPSVRR	4	188.033

S-methyl-L-cysteine_	ETLGAEWRQYEDIFPAIDASSIQQVAVEC!R	5	14.0157
trans-2,3-cis-3,4-dihydroxy-L-proline_	APESAAPAAATPKP!AAPPPSPAGPPAR	4	31.9898
N6-biotinyl-L-lysine_	PEHLEAFGLK!HR	10	226.0776
N-acetyl-L-proline_	VEAIAP!IGETRFSVR	3	42.0106
O4'-(phospho-5'-adenosine)-L-tyrosine_	NIY!RAALQKLAAR	4	329.0525
trans-2,3-cis-3,4-dihydroxy-L-proline_	NGTVAPSAGSAP!K	4	31.9898
trans-2,3-cis-3,4-dihydroxy-L-proline_	PALGPGGPAAARNGTVAPSAGSAP!KPLAGTPPAGGGPAVR	4	31.9898
N6-carboxy-L-lysine_	PLLGATVK!PKLGLSGR	4	43.9898
N,N-dimethyl-L-proline_	ADKNMVGP!ALGGVVGRK	3	29.0391
O-phospho-L-threonine_	SSAQRVIAAT!NSWLHAETRR	3	79.9663
N6-biotinyl-L-lysine_	NLAALTAAPSTLGDLEFAAAVAAILRGEDEAAK!	3	226.0776
N6-1-carboxyethyl-L-lysine_	IVNIASIAGK!	3	72.0211
dehydroalanine_	IASIYHGY!PSK	9	-94.0419
3-hydroxy-L-proline_	GPKGEAGAAGAPGPAGPAGPAGPAGPAGP!KGDAGPAGPAGPAGPAGPSGATGPAGPK	4	15.9949
trans-2,3-cis-3,4-dihydroxy-L-proline_	PPAPPAGEAPQRRGP!PPGAVPPNAVPPNAAAPDAAK	3	31.9898
S-methyl-L-cysteine_	NDMVQYFGEQLSGFAFTKEGWVQSYGSRC!VR	3	14.0157
trans-2,3-cis-3,4-dihydroxy-L-proline_	AAPVEAEPP!AEAAAPAPGVEAQPTAAPEPEAKPTK	3	31.9898
trans-2,3-cis-3,4-dihydroxy-L-proline_	PLRPALAEEP!R	4	31.9898
N6-1-carboxyethyl-L-lysine_	QGADDVVDKITAQGGK!	3	72.0211
trans-2,3-cis-3,4-dihydroxy-L-proline_	PSAPPTAAPAERPAAPPPAAAPVRPP!APPAGEAPQR	4	31.9898
trans-2,3-cis-3,4-dihydroxy-L-proline_	QPPGERRGPPP!GAPGTPPNATAPGMTPPPGEAPR	4	31.9898
N6-acetyl-L-lysine_	AIASLIIDGK!	3	42.0106
trans-2,3-cis-3,4-dihydroxy-L-proline_	AEP!PIMRADPPILR	3	31.9898
N6-1-carboxyethyl-L-lysine_	AAMDATLK!	3	72.0211
N6-biotinyl-L-lysine_	PFGLIANNPK!HLGGAIDADAGDK	4	226.0776
trans-2,3-cis-3,4-dihydroxy-L-proline_	EKPAQP!EAAK	3	31.9898
N6-1-carboxyethyl-L-lysine_	IINNGSISAHAPRPFSAAYTATKHAISGLTK!	3	72.0211
3-hydroxy-L-proline_	GPKGEAGAAGAPGP!AGPAGPAGPAGPAGPKGDAGPAGPAGPAGPAGPSGATGPAGPK	3	15.9949
N-acetyl-L-methionine_	DGSGSDAGMAQQVVDAIKAAGGRAVANTADISTM!AGGQSVFDDAIK	3	42.0106
3-hydroxy-L-proline_	GPKGEAGAAGAPGPAGPAGPAGPAGPAGPKGDAGP!AGPAGPAGPAGPSGATGPAGPK	5	15.9949
N-acetyl-L-proline_	MDKFGRP!LLGATVK	3	42.0106
3'-(1'-L-histidyl)-L-tyrosine_	Y!VDYPDAFAGWNLVSSIGSYISGFAVLVFLYGMTLAFIRKER	3	-2.0156
trans-2,3-cis-3,4-dihydroxy-L-proline_	AAPVEAEP!PAEAAAPAPGVEAQPTAAPEPEAKPTK	3	31.9898
N6-carboxy-L-lysine_	VTIAHPHGNFGAK!	4	43.9898

N6-lipoyl-L-lysine_	SGGGLK!APASAPAGPAIAAAMSDQQIR	3	188.033
3-hydroxy-L-proline_	PGIAGKP!GPDGKPGPIGPQGK	3	15.9949
N-acetyl-L-methionine_	VGPFAVPKAM!SSTASATLATWFK	3	42.0106
trans-2,3-cis-3,4-dihydroxy-L-proline_	P!SAPPTAAPAERPAAPPPAAAPVRPPAPPAGEAPQR	4	31.9898
N6-lipoyl-L-lysine_	GLLK!AAIRDPNPVIFLEHEMLYGQHGEVPK	3	188.033
trans-2,3-cis-3,4-dihydroxy-L-proline_	NGTVAP!SAGSAPK	4	31.9898
trans-2,3-cis-3,4-dihydroxy-L-proline_	GEAGAAPKQGAGKPAAAPAAETPAHTDP!VPAVTPAPK	4	31.9898
trans-2,3-cis-3,4-dihydroxy-L-proline_	AAEP!ATEEPTADTSPAAGK	5	31.9898
L-histidine_	VAAH!PEFDMGAILGHRASADVALLKLAAPLPGK	3	137.0589
trans-2,3-cis-3,4-dihydroxy-L-proline_	RGPPPGAP!GTPPNATAPGMTPPPGEAPR	3	31.9898
trans-2,3-cis-3,4-dihydroxy-L-proline_	P!EPPVLR	3	31.9898
N6,N6,N6-trimethyl-L-lysine_	AQK!EDFDYR	3	43.0548
trans-2,3-cis-3,4-dihydroxy-L-proline_	PVPPPMHVAPRVAPPPPPQHAAP!RMAPPPAPVR	3	31.9898
trans-2,3-cis-3,4-dihydroxy-L-proline_	APESAAPAAATPKPAAP!PPSPAGPPARR	5	31.9898
N6-1-carboxyethyl-L-lysine_	AAVPHMKPGSAIINTASVNSDMPNPMLLAYATTK!	3	72.0211
N6-1-carboxyethyl-L-lysine_	LAAEGAAVAVNYASSKQGADDVVDKITAQGGK!	3	72.0211
N6-1-carboxyethyl-L-lysine_	RMIARQQGGNIVNIASVLGQSVLK!	3	72.0211
L-serine_	GNFCS!GTLIAPDLVLSAAHCVGPGADYK	3	87.032
N6-1-carboxyethyl-L-lysine_	VINIASIDGIFVNPLETYPYAASK!AGLIHLTR	3	72.0211
3-hydroxy-L-proline_	GPKGEAGAAGAPGPAGPAGP!AGPAGPAGPKGDAGPAGPAGPAGPAGPSGATGPAGPK	3	15.9949
N6-1-carboxyethyl-L-lysine_	EGNPNAAHYSASK!	3	72.0211
trans-2,3-cis-3,4-dihydroxy-L-proline_	VAP!PPPPQHAAPRMAPPPAPVRAAPPPPHVAPPR	3	31.9898
N-acetyl-L-methionine_	HWIARPAPVNLDISM!PVASAQGDSFPR	3	42.0106
N-acetyl-L-methionine_	PNVSHRLPLSQWAPAM!RLLIDR	4	42.0106
N6-biotinyl-L-lysine_	LVTTK!	7	226.0776
trans-2,3-cis-3,4-dihydroxy-L-proline_	P!PPAAPRIQR	3	31.9898
trans-2,3-cis-3,4-dihydroxy-L-proline_	AP!PPMPERRPADPFASLVPEPIAR	3	31.9898
3-hydroxy-L-proline_	GEAGEAAP!K	3	15.9949
N-acetyl-L-proline_	LTELHDVAVANGAGALLINAMP!VGLSAVRMLRK	3	42.0106
trans-2,3-cis-3,4-dihydroxy-L-proline_	AEPP!IMRADPPILR	3	31.9898
N6,N6,N6-trimethyl-L-lysine_	FLTDK!GKADQAVGVTK	4	43.0548
trans-2,3-cis-3,4-dihydroxy-L-proline_	VAPPPPP!QHAAPRMAPPPAPVRAAPPPPHVAPPR	4	31.9898
trans-2,3-cis-3,4-dihydroxy-L-proline_	PAAPPPSPAGP!PAR	9	31.9898

APPENDIX 2

3'-(1'-L-histidyl)-L-tyrosine	RPA0832
3-(3'-L-histidyl)-L-tyrosine	RPA3310
3-hydroxy-L-proline	RPA2801
3-hydroxy-L-proline	RPA3593
Dehydroalanine	RPA2627
Dehydroalanine	RPA3893
L-3-oxoalanine	RPA1365
L-3-oxoalanine	RPA1990
L-histidine	RPA1895
L-serine	RPA1895
N,N-dimethyl-L-proline	RPA1535
N5-methyl-L-glutamine	RPA3250
N6,N6,N6-trimethyl-L-lysine	RPA1535
N6,N6,N6-trimethyl-L-lysine	RPA2394
N6,N6,N6-trimethyl-L-lysine	RPA2907
N6,N6,N6-trimethyl-L-lysine	RPA3693
N6-1-carboxyethyl-L-lysine	RPA0109
N6-1-carboxyethyl-L-lysine	RPA0234
N6-1-carboxyethyl-L-lysine	RPA0532
N6-1-carboxyethyl-L-lysine	RPA0586
N6-1-carboxyethyl-L-lysine	RPA0895
N6-1-carboxyethyl-L-lysine	RPA1110
N6-1-carboxyethyl-L-lysine	RPA1684
N6-1-carboxyethyl-L-lysine	RPA1757
N6-1-carboxyethyl-L-lysine	RPA2073
N6-1-carboxyethyl-L-lysine	RPA2160
N6-1-carboxyethyl-L-lysine	RPA2172
N6-1-carboxyethyl-L-lysine	RPA2186
N6-1-carboxyethyl-L-lysine	RPA2417
N6-1-carboxyethyl-L-lysine	RPA3074
N6-1-carboxyethyl-L-lysine	RPA3191
N6-1-carboxyethyl-L-lysine	RPA3287
N6-1-carboxyethyl-L-lysine	RPA3339
N6-1-carboxyethyl-L-lysine	RPA3474
N6-1-carboxyethyl-L-lysine	RPA3551
N6-1-carboxyethyl-L-lysine	RPA3552
N6-1-carboxyethyl-L-lysine	RPA3631
N6-1-carboxyethyl-L-lysine	RPA4306
N6-1-carboxyethyl-L-lysine	RPA4464
N6-1-carboxyethyl-L-lysine	RPA4618
N6-1-carboxyethyl-L-lysine	RPA4786
N6-acetyl-L-lysine	RPA3764

N6-biotinyl-L-lysine	RPA1405
N6-biotinyl-L-lysine	RPA1450
N6-biotinyl-L-lysine	RPA2435
N6-biotinyl-L-lysine	RPA2539
N6-biotinyl-L-lysine	RPA3175
N6-biotinyl-L-lysine	RPA4071
N6-carboxy-L-lysine	RPA0262
N6-carboxy-L-lysine	RPA1559
N6-carboxy-L-lysine	RPA2169
N6-carboxy-L-lysine	RPA4641
N6-lipoyl-L-lysine	RPA0188
N6-lipoyl-L-lysine	RPA2864
N6-lipoyl-L-lysine	RPA2866
N6-methyl-L-lysine	RPA4257
N-acetylglycine	RPA3693
N-acetyl-L-alanine	RPA3233
N-acetyl-L-alanine	RPA3339
N-acetyl-L-methionine	RPA0426
N-acetyl-L-methionine	RPA0656
N-acetyl-L-methionine	RPA1775
N-acetyl-L-methionine	RPA2018
N-acetyl-L-methionine	RPA2184
N-acetyl-L-methionine	RPA3072
N-acetyl-L-methionine	RPA3191
N-acetyl-L-methionine	RPA3339
N-acetyl-L-proline	RPA0262
N-acetyl-L-proline	RPA1559
N-acetyl-L-proline	RPA2169
N-acetyl-L-proline	RPA4641
N-methyl-L-alanine	RPA3227
O4'-(phospho-5'-adenosine)-L-tyrosine	RPA0984
O4'-(phospho-5'-uridine)-L-tyrosine	RPA2966
O5-glycosyl-L-hydroxylysine	RPA3593
omega-N-(ADP-ribosyl)-L-arginine	RPA1438
omega-N-(ADP-ribosyl)-L-arginine	RPA2635
O-phospho-L-threonine	RPA3200
S-(L-isoglutamyl)-L-cysteine	RPA2553
S-methyl-L-cysteine	RPA2181
S-methyl-L-cysteine	RPA2397
trans-2,3-cis-3,4-dihydroxy-L-proline	RPA0213
trans-2,3-cis-3,4-dihydroxy-L-proline	RPA2923
trans-2,3-cis-3,4-dihydroxy-L-proline	RPA3081
trans-2,3-cis-3,4-dihydroxy-L-proline	RPA3889

VITA

Evgenia Razumovski was born in Moscow, Russia on September 25th, 1977. She moved to Tucson, AZ in 1992 where she graduated from Catalina High School. She then enrolled in University of Arizona in 1996 and graduated with a degree in Computer Science in December 2000.

She enrolled in Genome Science and Technology, a joint program between University of Tennessee and Oak Ridge National Laboratory to pursue her doctoral degree in the Life Sciences department with a focus in bioinformatics. Her Ph.D. dissertation was accepted in December 2005.