12-2006

An Integrated Experimental and Computational Approach to Proteomics: Scaling from High Resolution Qualitative Analysis to Quantitative Measurements with Confidence Evaluation

Chongle Pan
University of Tennessee - Knoxville

To the Graduate Council:

I am submitting herewith a dissertation written by Chongle Pan entitled "An Integrated Experimental and Computational Approach to Proteomics: Scaling from High Resolution Qualitative Analysis to Quantitative Measurements with Confidence Evaluation." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Life Sciences.

Robert L. Hettich, Major Professor

We have read this dissertation and recommend its acceptance:

Nagiza F. Samatova, Arnold M. Saxton, Dale A. Pelletier, W. Hayes McDonald

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

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

To the Graduate Council:

I am submitting herewith a dissertation written by Chongle Pan entitled "An Integrated Experimental and Computational Approach to Proteomics: Scaling from High Resolution Qualitative Analysis to Quantitative Measurements with Confidence Evaluation." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Life Sciences.

Robert L. Hettich
Major Professor

We have read this dissertation
and recommend its acceptance:

Nagiza F. Samatova

Arnold M. Saxton

Dale A. Pelletier

W. Hayes McDonald

Accepted for the Council:

Linda Painter
Interim Dean of Graduate Studies

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

An Integrated Experimental and Computational Approach to Proteomics: Scaling from High Resolution Qualitative Analysis to Quantitative Measurements with Confidence Evaluation

A Dissertation Presented for the Doctor of Philosophy Degree

The University of Tennessee, Knoxville

Chongle Pan

December 2006

DEDICATION

I dedicate this dissertation to my parents, Gongqiang Pan and Siping Liang, for their love and support for me.

ACKNOWLEDGMENTS

I would first like to thank my advisors, Dr. Robert L. Hettich and Dr. Nagiza F. Samatova, for guiding my research, supporting me with great confidence, and giving me wide latitude to pursue my scientific interests. I would also like to thank my committee members, Dr. Dale A. Pelletier, Dr. W. Hayes McDonald, and Dr. Arnold M. Saxton, who have advised me not only in my committee meetings but also in our numerous discussions along the course of my research. I would like to thank Dr. Gregory B. Hurst, Dr. Nathan C. VerBerkmoes, and Dr. David L. Tabb, from whom I have learned great amount of scientific knowledge and experimental skills. I would like to thank Praveen Chandramohan and Guruprasad H. Kora. Without their help, it would be impossible for me to develop any of the computer algorithms described in this dissertation. All my research work has been the results of collaborative efforts and I would like to thank my coworkers and collaborators: Dr. Brad Strader, Dr. Yasuhiro Oda, Patricia K. Lankford, Dr. Ying Xu, Dr. Bo Yan, Dr. Victor N. Olman, Dr. Caroline S. Harwood, Dr. Byung H. Park, and Dr. Bing Zhang. I would like to thank my fellow graduate students: Judson Hervey, Jun Wu, Melissa Thompson, Heather Connelly, Christine Shook, Alon Savidor, and Carlee McClintock. It has been a pleasure to study and work with them. Finally, I would like to thank my parents, Gongqiang Pan and Siping Liang, and my friend, Yang Wang.

.

ABSTRACT

As a component of systems biology, proteomics aims to characterize the entire protein complement of an organism, including qualitative identification of protein types and quantitative measurement of protein abundance changes as a function of different cellular states. This dissertation presents an integrated experimental and computational approach to improve proteomic measurements, including qualitative measurements using Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR-MS) and quantitative measurements with statistically derived confidence evaluation.

Although FT-ICR-MS provides high-performance mass measurements, its potential has not yet been fully explored for proteomics applications. A novel tandem mass spectrometry method was developed for FT-ICR-MS to obtain sequence tag information directly from intact proteins in a mixture. The interpretation of FT-ICR tandem mass spectra for sequence tagging was facilitated with a new graph-theoretical algorithm for separation of y- and b-ions. To scale FT-ICR-MS for general proteomic characterizations, low flow-rate liquid chromatography was integrated with FT-ICR-MS. The high-performance MS greatly enhanced the depth and quality of the proteomics measurements. In total, these studies demonstrated that FT-ICR-MS is of practical value for proteomic measurements, and that additional experimental and computational developments could make this into a robust and automated approach.

Quantitative proteomics based on stable isotope labeling enables global gene expression profiling at the protein level. However, major challenges remain for extracting reliable protein quantification information from noisy mass spectrometric data. A principal component analysis algorithm was developed to accurately estimate peptide abundance ratios and to provide rigorous scores for their estimation variability and bias. The peptide quantification results were then processed by a novel profile likelihood algorithm to estimate protein abundance ratios with confidence interval evaluation. These algorithms were integrated into a computer program, ProRata, for automated data analysis. Quantitative proteomic measurements were conducted using ProRata, and integrated with transcriptomic analysis to study the anaerobic metabolism of *p*-coumarate in *Rhodopseudomonas palustris*. This study yielded a putative cellular pathway for *p*-coumarate catabolism.

In the research described here, a substantial advancement in both qualitative and quantitative proteomic measurements was achieved using an integrated experiment and computational approach. The improved proteomic measurements can help elucidate a range of biological processes.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

.

LIST OF SYMBOLS AND ABBREVIATIONS

BCA	Bicinchoninic acid solution
CAD	Collision-activated dissociation
Da	Dalton
DIGE	Difference gel electrophoresis
DNA	Deoxyribonucleic acid
DTT	Dithiothreitol
EDTA	Ethylenediaminetetraacetic acid
ESI	Electrospray ionization
FA	Formic acid
FT-ICR	Fourier transform ion cyclotron resonance
FT-MS	Fourier transform mass spectrometry
HPLC	High performance liquid chromatography
ICAT	Isotope coded affinity tags
i.d.	Internal diameter
LC	Liquid chromatography
LC-MS	Liquid chromatography-mass spectrometry
LC-MS/MS	Liquid chromatography-tandem mass spectrometry
LCQ	Thermo Finnigan ES quadrupole ion trap
LTQ	Thermo Finnigan ES linear ion trap
MALDI	Matrix assisted laser desorption
mRNA	Messenger ribonucleic acid

MS	Mass spectrometry
MSAD	Multipole-storage assisted dissociation
MS/MS	Tandem mass spectrometry
MudPIT	Multidimensional Protein Identification Technology
MW	Molecular weight
m/z	Mass-to-charge ratio
PAGE	Polyacrylamide gel electrophoresis
PCR	Polymerase chain reaction
PPM	Parts per million
PTM	Post-translational modification
QIT	Quadrupole ion trap
RP	Reverse phase
RNA	Ribonucleic acid
SCX	Strong cation exchange
SDS	Sodium dodecyl sulfate
S/N	Signal-to-noise ratio
TAP	Tandem affinity purification
TIC	Total ion chromatogram
TOF	Time-of-flight
Xcorr	SEQUEST cross-correlation score

Chapter 1

Introduction to Proteomics and its Role in Systems Biology

One of the greatest advances in biology is the establishment of the central dogma of molecular biology (Figure 1.1) (Astbury, 1961). It was discovered that, for almost all life forms on the Earth, complete genetic information is encoded in deoxyribonucleic acid (DNA) molecules (Benfey, 2004). This genetic information can be transmitted horizontally between different organisms and vertically from parents to progeny through DNA replication. The functional units of the genetic information are genes. To express a gene, the DNA sequence of the gene is transcribed to a messenger RNA (mRNA), which is then translated into a protein. In most cases, the cellular function is carried out by the final product of gene expression, the protein.

Understanding biological processes at the molecular level has ushered in a new research paradigm for biology. Previously, biology was largely a descriptive science, where general principles were deducted from observational data. For example, Charles Darwin conceived the theory of natural selection from his five-year voyage on the Beagle ship (Darwin, 1859; Browne, 1996). With the emergence of molecular biology, biologists have become capable, for the first time, of manipulating life in a directed manner. This has helped transform biology into a so-called "hypothesis-driven" discipline. A biologist can propose a hypothesis and then design experiments to test the hypothesis. This approach has resulted in elucidation of the functions and regulation of thousands of genes,

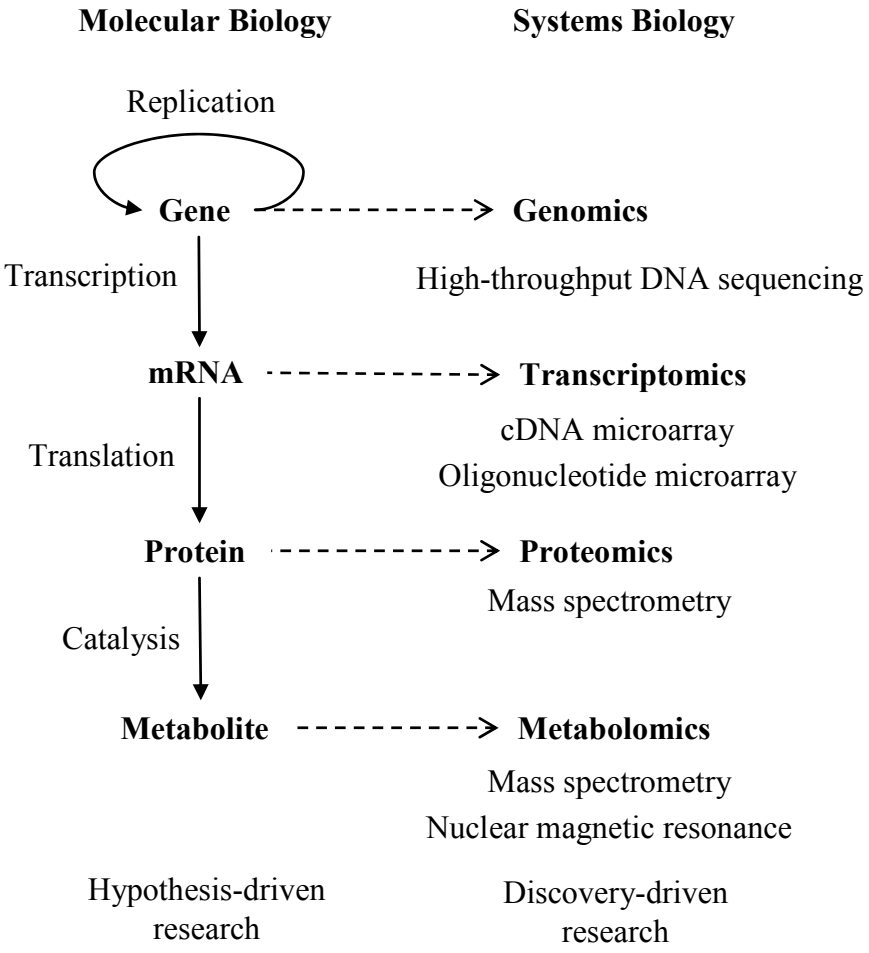


Figure 1.1: Research paradigm shift for biology. Molecular biology based on the central dogma paved the way to systems biology that takes advantage of different "-omics" technologies

and has impacted all areas of biology: cell biology, developmental biology, evolutionary biology, e*tc*.

However, there are at least two challenges in this hypothesis-driven research paradigm. First, the research is limited by the hypothesis to be tested (Goodman, 1999). The research documented in scientific journals is often the validation of a reasonable hypothesis; whereas rejecting a "wrong" hypothesis is rarely considered as scientific progress. To avoid the risk of formulating a wrong hypothesis in the first place, the hypotheses to be tested are often logical incremental extensions of known facts. However, pushing the envelope of knowledge from the known has precluded many biologists from leaping into uncharted territories in biology.

Second, most conventional molecular biology experiments focus on one gene, one protein complex, or one pathway at a time. This provides a very narrow view of biological systems. In a living organism, there are convoluted protein interactions, complex regulatory networks, and numerous cellular pathways, which require biologists to seek answers to their questions in the context of a complex inter-connected system (Ideker, 2001). A targeted experimental approach becomes insufficient to provide such an unbiased global perspective about biological systems.

In recent years, the development of high-throughput technologies and corresponding data analysis algorithms has enabled a variety of "omics" research, including genomics (Cole, 1994; McKusick, 1997), transcriptomics (Bednar, 2000; Harrington, 2000), proteomics

(Porubleva, 2000; Yates, 2000), and metabolomics (Fiehn, 2002; Reo, 2002) (Figure 1.1). Among them, genomics was first to emerge, due to the advent of high-throughput DNA sequencing technology (Salser, 1974; Martin, 1989). Genomics focuses on determining and studying the genome of an organism, which encodes for the entire genetic information of the organism.

The availability of complete genome information enabled the development of transcriptomics, which is based on high-density microarray technology (Kurian, 1999). The complete set of all mRNA molecules, or "transcripts", produced in an organism is its transcriptome. Because the synthesis of a protein requires the presence of its mRNA, the mRNA abundance change can be used as a surrogate for the protein abundance change. Thus, the global gene expression profiling became possible at mRNA level with transcriptomics.

As the next level of the “omics” measurements, proteomics aims to characterize the entire protein complement of an organism (Blackstock, 1999). Compared with genomics and transcriptomics, proteomic measurements still demand much effort in methodology development, due to heterogeneous characteristics of proteins (size, hydrophobicity, structure, *etc*). Additionally, the technologies for genomics and transcriptomics are based on molecular biology tools such as polymerase chain reaction (Erlich, 1989), reverse transcription, hybridization, *etc*. However, there are no equivalent tools to manipulate and amplify proteins.

A natural extension of proteomics is metabolomics, which focuses on characterization of the entire metabolite complement of an organism. Metabolomic measurements are even more technologically challenging and their methodology has not yet been well defined (Bino, 2004).

The advent of these "-omics" measurements leads to a new research paradigm, termed "discovery-driven" or "data-driven" research, which is different from the previous hypothesis-driven research (Goodman, 1999). A biological system can be characterized by genomics, transcriptomics, proteomics, or any combination of such. Genomics reveals the "screenplay" of an organism. Transcriptomics and proteomics present molecular "shows" inside the organism under different scenes to biologists. With these new technologies, modern biologists can study life by making discoveries from large sets of data. Free from the confinement of a pre-defined hypothesis, discoveries can now shed light on the uncharted territories of biology and reveal unknown or even unexpected mechanisms of biological processes. It has become evident that deducing discoveries from observational "-omics" data will complement the conventional approach of validating induced hypotheses (Lastowski, 2000).

Such -omics research has another characteristic distinct from the hypothesis-driven research – the global view of a biological system. If one considers that the expression of a gene represents the vertical information flow from one level to another, then genomics, transcriptomics, proteomics, and metabolomics all aspire to profile the entire horizontal information landscape at each level. Such a wide-angle perspective on the life processes

of an organism allows the development of an understandable model for a biological system in its entirety. This endeavor has been termed "systems biology" (Anderson, 2000; Ideker, 2001; Regnier, 2002).

A genome can be completely characterized by determining nucleic acid sequences of all DNA molecules. The transcript of a gene is the messenger that passes the genetic information to protein synthesis machinery. The abundance of a transcript is actively regulated to control the production of a protein. Hence, the measurement of a transcriptome can be achieved by quantifying the abundance of the transcripts for all genes. Compared with DNA and mRNA, proteins are, however, dynamic and diverse in their abundances, chemical modifications, tertiary structures, cellular locations, and physical interactions. Due to multifaceted characteristics of proteins, the characterization of the proteome requires the combination of different measurements that focus on different properties of proteins or different categories of proteins. The different types of proteomics measurements include, but are not limited to:

- *Structural proteomics*: characterization of protein secondary or higher order structures (Schmid, 2002);
- *Interactomics*: identification of physical interactions between proteins (Cesareni, 2005);
- *Amino acid sequence variations*: detection of nucleotide polymorphisms, signal peptide cleavage, and sequence database errors (Gatlin, 2000);

- *Post-translational modifications (PTMs)*: determination of chemical modifications to proteins, including phosphorylation, acetylation, methylation, glycosylation, *etc* (Cantin, 2004; Jensen, 2004);
- *Sub-cellular proteomics*: cataloguing of all proteins in an organelle or a cellular compartment (Brunet, 2003; Huber, 2003);
- *Whole-cell proteomics*: characterization of the entire protein complement of a cell (Washburn, 2000).

There are multiple challenges in realizing the promises of proteomics (Marko-Varga, 2004; Reinders, 2004; Bertone, 2005). The first challenge is the complexity of proteome samples. There are thousands of different proteins in a proteome sample. Each of the proteins may have multiple modification types. The second challenge is the enormous dynamic range between proteins. Dynamic range is the concentration difference between the most abundant proteins and the least abundant ones. The dynamic range can reach up to 10^6 in a bacterial proteome (Corthals, 2000). The third challenge is the difficulty of measuring membrane proteins (Santoni, 2000). Membrane proteins play crucial roles in many cellular activities. But due to their high hydrophobicity and tight association with lipids, membrane proteins are not very amenable to many experimental methods. In addition, there are other experimental challenges, such as measurement throughput, reproducibility, *etc*.

Different methodologies have been developed to address these challenges in proteomics measurements. The three main methodologies are two-dimensional gel electrophoresis

(Rabilloud, 2002; Watt, 2003), shotgun proteomics (Wolters, 2001; McDonald, 2003), and top-down proteomics (Kelleher, 2004) (Figure 1.2). They all combine high peak-capacity separation and mass spectrometry (MS) to handle the sample complexity and dynamic range of proteome samples.

Two-dimensional gel electrophoresis was the first methodology capable of large-scale proteome measurement. The extracted proteins are first separated by their isoelectric points with isoelectric focusing and then by their molecular weights with denaturing sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) (Figure 1.2) (Bernard, 2004). The 2-dimensional separation resolves most proteins into individual spots in the gel. The gel spots are excised and digested with protease. The digestion products are then measured with mass spectrometry, and the proteins are generally identified by peptide mass fingerprinting (Cottrell, 1994; Pappin, 1997). Two-dimensional gel electrophoresis also allows quantitative measurements based on signal intensities of the gel spots (Watt, 2003). Recently, quantitative proteomics has also been performed with fluorescence 2-D difference gel electrophoresis (DIGE) (Unlu, 1997), where proteins from different samples are labeled with different fluorescent dyes and then mixed together for separation on the same gel.

Two-dimensional gel electrophoresis has been used for a variety of proteomic applications, including cataloging proteins in proteome samples, measuring protein abundance changes between cellular conditions, detecting and quantifying post-translational modifications, *etc*. For example, many types of PTMs, such as

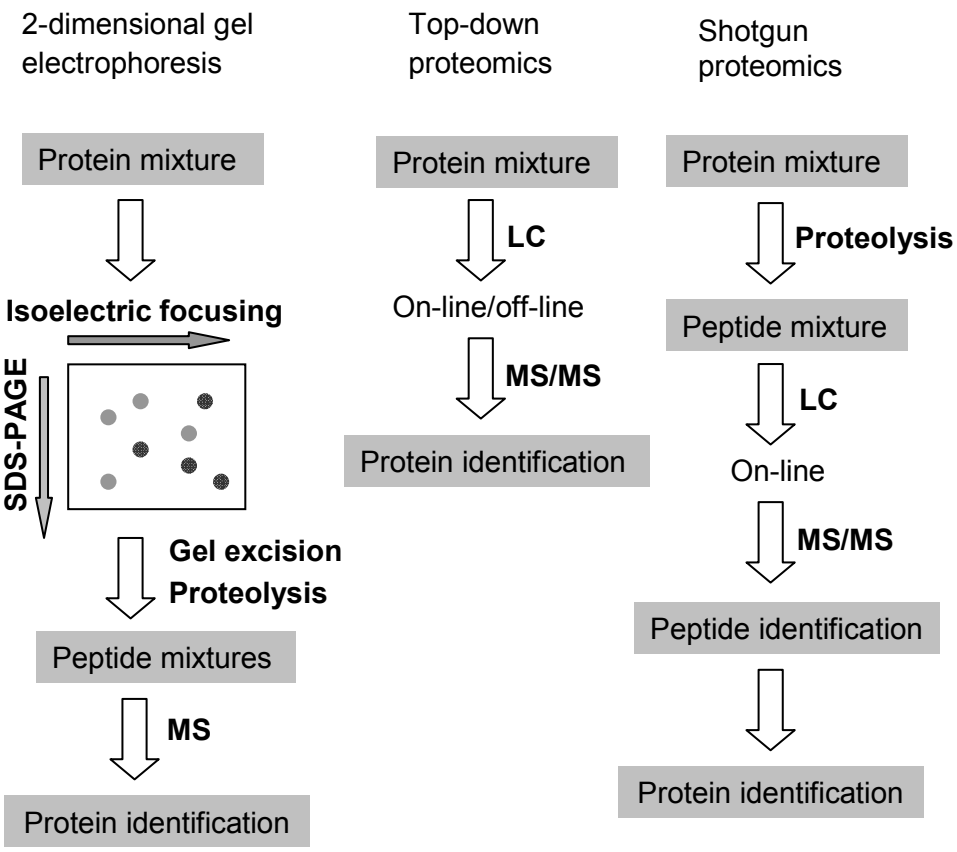


Figure 1.2: Schemes of three qualitative proteomics methodologies. The general procedure for protein identification involves three steps: separation, mass spectrometry measurement and data analysis.

phosphorylation and glycosylation, can change the isoelectric point of a protein (Banks, 2000). The horizontal migration of the protein's gel spot to a new pI position would suggest the presence of these types of PTMs. The comparison between two gel spots of the protein can yield quantitative information on what percentage of the protein molecules are modified.

There are certain disadvantages in employing two-dimensional gel electrophoresis (VerBerkmoes, 2004), including:

- Bias against membrane proteins and large proteins;
- Poor recovery rate of low-abundance proteins from the gel;
- Limited reproducibility;
- Limited automation available.

There are also at least two challenges for protein quantification with DIGE. First, a gel spot can have more than one protein, and the fluorescence intensity ratio measured in that gel spot is a weighted average of abundance ratios of all proteins in that spot. Second, each protein is quantified with a single data point in one DIGE measurement, which necessitates extensive replication for reliable error estimation.

Top-down proteomics separates intact proteins with liquid chromatography (LC), – usually ion exchange LC, C4 reverse phase LC, or their combination (VerBerkmoes, 2002; Wang, 2005). Liquid chromatographic separation obviates the problem of recovering intact proteins from the gel in two-dimensional gel electrophoresis. The LC elution can be coupled online with mass spectrometry or offline by fraction collection.

Generally, top-down proteomics requires mass spectrometers with high mass measurement performance to analyze intact proteins (VerBerkmoes, 2002; Bogdanov, 2005). The accurate mass measurement of intact proteins often provides definitive information on protein identity. Tandem mass spectrometry (MS/MS) can also be employed to measure the gas-phase fragmentation product ions of the intact proteins to provide sequence information for more definitive protein identification (Kelleher, 2004).

The main characteristic of top-down proteomics is the direct mass spectrometry measurement of intact proteins. Two-dimensional gel electrophoresis yields approximate isoelectric points and masses of the intact proteins from their gel spot positions. However, mass spectrometry measurement can measure intact protein masses at accuracies that are orders of magnitude higher than SDS-PAGE, and tandem mass spectrometry provides protein sequence information. Such measurements make top-down proteomics uniquely advantageous for characterizing post-translational modifications of proteins. Virtually all PTMs alter the masses of proteins, and the measured mass shift for a protein allows the inference of the number and the types of all PTMs on the protein. Top-down proteomics has been used for whole-cell proteomics to detect the presence of methylation, acetylation, phosphorylation, N-terminal truncation, *etc* (VerBerkmoes, 2002; Kelleher, 2004; Strader, 2004).

Although significant progress has been made in top-down proteomics, this methodology has the following limitations to be addressed:

- Molecular mass cannot be measured accurately for proteins larger than 100,000 Daltons;
- Liquid chromatography is not straightforward for large proteins or membrane proteins;
- The gas-phase fragmentation of intact proteins is not very robust or extensive;
- Many low-abundance proteins are not detected.

Overcoming these limitations requires further development of intact protein liquid chromatography, high performance mass spectrometry instrumentation, and data analysis algorithms.

Both two-dimensional gel electrophoresis and top-down proteomics process proteins from a proteome sample in their intact forms. Many aforementioned disadvantages of the two methodologies stem from the challenges in the separation and the MS analysis of intact proteins. Instead, shotgun proteomics analyzes proteolysis-derived peptides (Wolters, 2001). A protein mixture is first treated with protease, which cleaves all proteins into peptides. The peptides are then measured with liquid chromatography-tandem mass spectrometry (LC-MS/MS). Shotgun proteomics has been shown to be the methodology most comprehensive in cataloging proteins in a complex protein mixture (VerBerkmoes, 2004). With shotgun proteomics, thousands of proteins have been identified in whole-cell lysates from a wide range of organisms (Jungblut, 1999; Cash, 2003; Lilley, 2003). Mitochondria, lysosomes and chloroplasts have been purified, and their protein constituents have been characterized with shotgun proteomics (Brunet, 2003; Huber, 2003). Proteins carrying specific PTMs, such as phosphorylation (Gronborg, 2002;

MacCoss, 2002) and ubiquitination (Peng, 2003), have been enriched with biochemical methods and identified with shotgun proteomics.

Shotgun proteomics has also been coupled with a variety of stable isotope labeling techniques for measuring relative abundances of proteins between different proteome samples (Tao, 2003). The combination of LC-MS measurements with stable isotope labeling for quantification has been used in isotope dilution mass spectrometry (IDMS) for decades (Bjorkhem, 1980). An isotopic analogue of identical structure as the analyte to be quantified is synthesized and mixed with the sample. The analogue becomes the internal standard in known abundance; and the ratio between the ion currents of the analyte and the internal standard measured by LC-MS reflects the ratio of their abundances. The same strategy is employed in quantitative shotgun proteomics. As it is impractical to synthesize isotopic analogue for every protein, a wide range of stable isotope labeling techniques have been developed, including ^{15}N or ^{13}C metabolic labeling (Oda, 1999), stable isotope labeling with amino acids in cell culture (SILAC) (Ong, 2002), $H_2{}^{18}O$ digestion (Yao, 2001), and isotope-coded affinity tags (ICAT) (Gygi, 1999). With these stable isotope labeling techniques, quantitative shotgun proteomics allows accurate relative quantification of thousands of proteins in a high-throughput manner.

Compared with two-dimensional gel electrophoresis and top-down proteomics, a disadvantage of current shotgun proteomics measurements is the lack of information on the full sequence of a protein. Proteins are digested into peptides during the first step of shotgun proteomics. The existing LC-MS/MS technology can generally measure a subset

of peptides in the peptide mixture, and the measured peptides for the majority of proteins can only cover a portion of amino acid sequences. Although the detection of a portion of a protein generally presents a sufficient evidence for the presence of the intact protein, there is little information about the undetected portion of the protein sequence, such as the presence or absence of an amino acid sequence variation, *N*-terminus truncation, chemical modification, *etc*.

The field of proteomics rests on four cornerstones, namely, protein chemistry, mass spectrometry, informatics, and biology (Figure 1.3):

- Protein chemistry: Extraction of proteins from cells and preparation of proteome samples.
- Mass spectrometry: Comprehensive measurement of the proteome samples with suitable analytical platforms.
- Informatics: Processing of mass spectral data to yield information, such as protein identification, chemical modification detection, and quantification.
- Biology: Interpretation of proteomic results to generate knowledge to further the understanding of a biological system.

The methodology development for proteomics often requires devoting coordinated effort in the four cornerstones.

The major goal of this dissertation is to develop new proteomics methodologies for confident protein identification and quantification. Two main research directions have been pursued: The first direction is development and prototyping of various high-

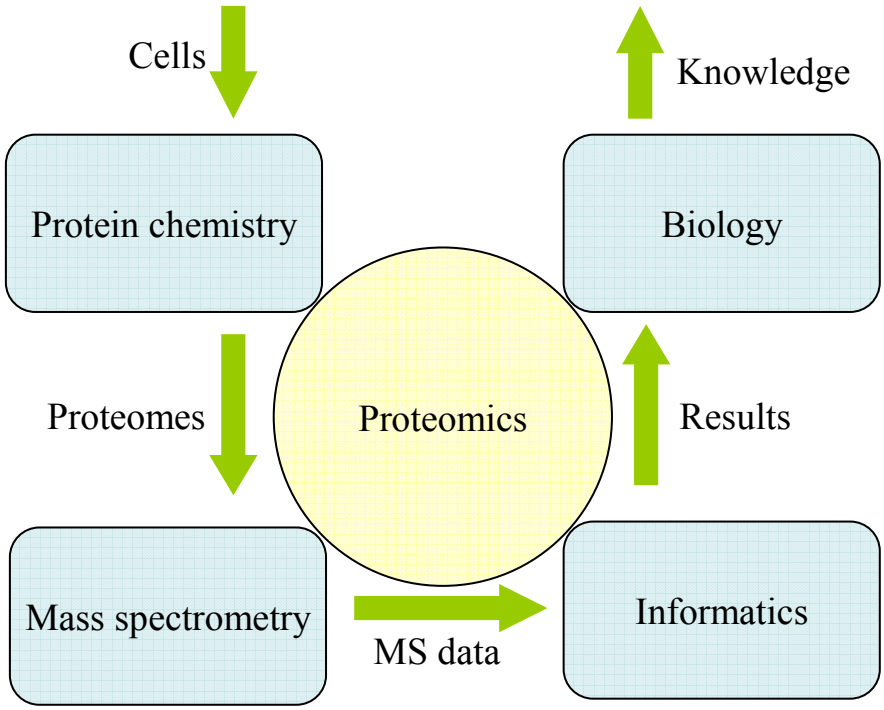


Figure 1.3: Cornerstones of proteomics. Proteomics is based upon four cornerstones: protein chemistry, mass spectrometry, informatics and biology.

performance methodologies based on Fourier transform ion cyclotron resonance (FT-ICR) mass spectrometry (Hendrickson, 1999). The conventional analytical platform for shotgun proteomics is primarily based on quadrupole ion trap (QIT) MS, because QIT-MS is a robust medium-cost instrument with exquisite capabilities for high-throughput tandem mass spectrometry measurement (Brancia, 2006). However, QIT-MS provides mass measurements of only moderate accuracy and resolution. We believe that the next-generation analytical platform for proteomics will be based on high-performance mass spectrometers, such as FT-ICR. The exceptional accuracy, resolution, and dynamic range in mass measurement offered by FT-ICR could revolutionize proteomics. However, there was insufficient prior research work on methodology development based on FT-ICR. A focus of this dissertation is to explore the potential of FT-ICR for various proteomics measurements, including direct sequence tagging from intact proteins, *de novo* sequencing of peptides, and high confidence peptide identification. Each of these studies demonstrated the unique advantages of high performance MS instruments for proteomics. Although not yet suitable for proteomics laboratories operating in a "pipeline" measurement mode, these prototype methodologies represented pioneering development of proteomics methodology. In fact, with the recent introduction of two commercial hybrid high-performance instruments, the ThermoFinnigan LTQ-FTMS (Peterman, 2005) and LTQ-Orbitrap (Erickson, 2006), we believe that the transition to high-performance instruments has began, and the value of our pioneering work in this area will become more evident.

The second research direction focuses on quantitative shotgun proteomics with advanced data analysis algorithms. Shotgun proteomics was developed initially to catalogue proteins in a proteome. However, the detection of a protein's presence in a proteome is not as informative about its function as quantification of the protein's abundance change between different cellular states. Quantitative proteomics has been developed to determine the abundance changes of thousands of proteins. Quantitative proteomics results are equivalent to transcriptomics results: both provide global gene expression profiles, one at mRNA level and the other at protein level. For quantitative shotgun proteomics measurement, the critical step in protein chemistry is stable isotope labeling of proteins, and the key in mass spectrometry analysis is acquisition of high-quality full scan mass spectra in LC-MS/MS measurement.

A variety of stable isotope labeling methods have been developed for quantitative proteomics. However, the informatics for quantitative proteomics has greatly lagged behind. Algorithms are required to estimate the relative abundances of peptides from full scan data and then to estimate the abundance ratios of proteins by assembling their peptides together. As quantitative shotgun proteomics measurements yield selected ion chromatograms at highly variable signal-to-noise ratios for tens of thousands of peptides, we developed algorithms that not only robustly estimate the abundance ratios of different peptides but also rigorously score each abundance ratio for the expected estimation bias and variability. A profile likelihood algorithm was then used for maximum likelihood point estimation and profile likelihood confidence interval estimation of protein

abundance ratios. The confidence interval estimation provides an "error bar" for each protein abundance ratio that reflects its estimation precision and statistical uncertainty.

Echoing a systems biology theme of integrating multiple "-omics" technologies, we have combined genomics, transcriptomics, and quantitative proteomics to study anaerobic *p*-coumarate degradation in *Rhodopseudomonas palustris*. Multiple genes in the hypothesized *p*-coumarate pathway were identified by sequence similarity and expression change. The global gene expression profiles at both mRNA level and protein level showed the coordinated responses from a multitude of related cellular pathways to *p*-coumarate degradation. The discoveries made in this study have showcased how systems biology can further our understanding of a biological system in a high-throughput and comprehensive manner.

The research conducted under this dissertation project is discussed by first highlighting the fundamental experimental and computational work and then applying them to explore biological applications. Thus, the dissertation is organized as follows: Chapter 2 describes the mass spectrometry technology for proteomic measurements. Chapter 3 introduces a novel MS^3 tandem mass spectrometry analysis with FT-ICR for deriving sequence tags from intact proteins. Chapter 4 details a graph-theoretical algorithm for interpreting the high-resolution FT-ICR tandem mass spectra. Chapter 5 describes our effort in interfacing FT-ICR with an LC system. The next three chapters focus on quantitative shotgun proteomics. Chapter 6 describes novel algorithms for peptide quantification. Chapter 7 details both a protein quantification algorithm and the

benchmark results of quantitative proteomics measurements. Chapter 8 shows a study of *p*-coumarate degradation in *Rhodopseudomonas palustris*. Chapter 9 concludes the described research work and discusses the future directions of proteomics and systems biology.

Chapter 2

Mass Spectrometry as a Foundation Technology for Proteomics

BACKGROUND

This chapter describes general technology background for proteomics measurements. The studies described in Chapters 3, 4, and 5 are based on FT-ICR technology for prototyping high performance proteomics methodologies. The studies presented in Chapters 6, 7, and 8 are based on shotgun proteomics with quadrupole ion trap technology for quantitative proteomics development and application. Although the exact methods used in these studies vary to some extent and are detailed in the MATERIALS AND METHODS section of each chapter, this chapter presents a general background on the fundamental mass spectrometry technology, the shotgun proteomics methodology, and the biological system of interest.

Principles of mass spectrometric ion manipulation and measurement

All mass spectrometers have three fundamental components, namely the ion source, the mass analyzer, and the detector (Hoffmann, 2001). In the ion source, the analytes in a sample are transformed into gas-phase ions, which are then transferred through a series of ion optics into the mass analyzer. The process of generating gas-phase ions of the analytes is called ionization. The ions are then resolved according to their mass-to-charge

ratio (*m/z*) in the mass analyzer. Finally, the detector measures the signal intensities of ion species occurring at different mass-to-charge ratios.

A variety of mass analyzers have been developed, including three-dimensional quadrupole ion traps, linear quadrupole ion traps, triple quadrupoles, time-of-flight (TOF), sectors, Fourier transform ion cyclotron resonance (FT-ICR), and Orbitrap (Hoffmann, 2001; Erickson, 2006). The performance of these different mass analyzers can be measured in terms of different figures-of-merit, including:

- Mass accuracy: the mass error divided by the expected mass. For modern mass spectrometers, this is generally much less than 1% and is commonly measured in parts per million (ppm).
- Resolution: the ability to separate a mass spectral peak from other closely-positioned ones. This is measured by the expected mass divided by the peak width at half height.
- Dynamic range: the ion signal ratio between the most abundant detected species and the least abundant detected species.
- Sensitivity: the increase of mass spectrometric signal intensity from a unit increase of the analyte concentration.
- Detection limit: the minimum concentration of the analyte that can be detected with a signal-to-noise ratio of three.
- Upper mass limit: the maximum mass-to-charge ratio that can be measured.

The figures-of-merit for different mass analyzers are compared in Table 2.1. The mass spectrometers with higher performance, such as Orbitrap and FT-ICR, are also more expensive than the other mass spectrometers.

Ionization methods for biomolecules

Two soft ionization methods used in the ion source are matrix assisted laser desorption ionization (MALDI) (Hillenkamp, 1990) and electrospray ionization (ESI) (Fenn, 1989). MALDI ionizes samples from a crystallized form. An aqueous sample is mixed with a matrix solution, spotted to an arrayed plate, and crystallized by air-drying. Ions from the sample can be generated by bombarding the crystals at the spot with a focused laser beam. The matrix assists the ionization of analyte molecules and protects them from the disruptive energy of the laser. In contrast, ESI transports preformed ions directly from the liquid phase to the gas phase. A continuous stream of sample solution is sprayed through a needle to a counter-electrode, which is electrically biased with a few thousands of volts potential difference. The analyte ions are released from the droplets in the spray plume and desolvated by the heated gas. Because the continuous solution stream can be the eluent of an LC system, ESI enables placing mass spectrometer online with LC. The integrated platform is called LC-MS. Both top-down proteomics and shotgun proteomics employ ESI in their LC-MS measurement.

During electrospray ionization, a sample is pumped at a low flow rate (10 nL/min–10 μL/min) through a capillary needle biased to a positive high potential relative to the

Table 2.1: Figures-of-merit of different mass analyzers.

Mass analyzer	Mass accuracy	Resolution	Dynamic range
QIT [1]	20 ppm	1k — 2k	100
TOF [2]	2 — 5 ppm	2k — 10k	1000
Sectors	< 1 ppm	5k — 100k	10,000
Orbitrap	2 — 5 ppm	5k — 50k	5,000
FT-ICR [3]	< 1 ppm	5k — 1000k	10,000

[1] Quadrupole ion trap

[2] Time-of-flight

[3] Fourier transform ion cyclotron resonance

orifice of mass spectrometer in the positive ion mode (Figure 2.1). The capillary needle, whose inner diameter should match with the flow rate, is referred to as ESI emitter. Often ESI is referred to as nanospray if its flow rate is below a few hundred nL/min and as microspray if its flow rate is between 1 μL/min and 10 μL/min. As the solution reaches the capillary tip, where a strong electric field is present, the liquid emerges out of the tip to form Taylor cone and then streams out as a jet, which is subsequently dispersed into a plume of charged fine droplets. Note that stream and plume are formed as a result of the strong electric field, not the fluidic pressure. Through the orifice and a differential pumping system, these droplets are transferred from atmospheric pressure into the vacuum of the mass analyzer, and free ions are released from the shrinking droplets.

The process of transferring analytes from liquid phase to gas phase is thermodynamically unfavorable in itself, due to the loss of their solvation energy. Currently the electrospray process is thought to consist of three steps: (1) generation of charged droplets; (2) a cascade of uneven fission of charged droplets; and (3) release of ions into gas phase from fine charged droplets (Hager, 1994; Kebarle, 1999). The strong electric field at the needle tip attracts cations to the surface of Taylor cone, while repulsing anions away to high-voltage anode to be oxidized to neutral species. The electric force of drawing excess cations away from anions in bulk liquid can eventually overcome the liquid surface tension and a fine jet of liquid is ejected. This jet subsequently disintegrates into charged droplets with a surplus of cations. Due to solvent evaporation from heating, the charged droplets shrink in size, approaching the so-called “Rayleigh limit”, which is the maximum charge a spherical droplet can hold before the Coulomb repulsion overcomes

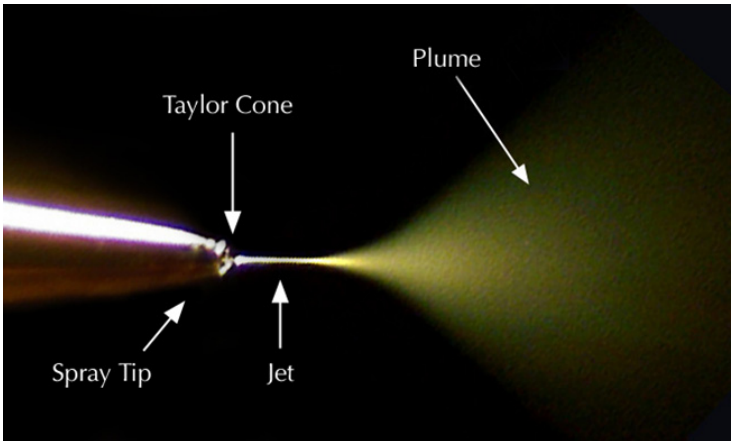


Figure 2.1: Image of a stable electrospray. A stable electrospray consists of Taylor cone, jet and plume. A portion of the charged fine droplets in the plume are admitted into mass spectrometer through an orifice in the right (not shown). (Image from www.NewObjective.com.)

the surface tension (Taflin, 1989). Once the Rayleigh limit is reached, the droplets eject about 20 offspring droplets, which carry away about 15% of the charge (excess cations) and about 2% of the mass from the initial droplet. This process is called uneven fission. The mechanism for the third step is still under debate. The charge residue model assumes that cycles of solvent evaporation and uneven fission continue to occur in the offspring droplets until the droplets contain only one ion, which is released after the solvent evaporates (Dole, 1968). A competing model, the ion evaporation model, assumes that ions "evaporate" directly from offspring droplets of 10-nm diameter (Iribarne, 1976).

Ion Trapping Mass Spectrometry

FT-ICR-MS has been the most widely used mass spectrometer for top down proteomics, because of its high performance in mass measurement (VerBerkmoes, 2002; Kelleher, 2004). In FT-ICR MS, the mass-to-charge ratios of ions are determined based on their cyclotron frequencies in a magnetic field (Marshall, 1998). The simplified relationship between the angular cyclotron frequency (ω_c) and the mass to charge ratio (m/z) is given by:

$$\omega_c = \frac{B}{m/z}, \tag{2.1}$$

where B is the magnetic field strength. Thus, the m/z of an ion can be determined by measuring its cyclotron frequency in a static magnetic field. For an m/z range of 100–2500 and a magnetic field of 9.4 Tesla, the cyclotron frequencies span the kHz to MHz radiofrequency range. Within this radiofrequency range, frequency can be measured with

excellent precision, and therefore the *m/z* of ions can be determined with ultrahigh resolution (Marshall, 1985).

The ions are trapped in the FT-ICR cell radially by the magnetic field and axially by two electrostatic trapping plates. The trapped ions are excited by a resonantly oscillating electric field which matches the ions natural cyclotron frequencies to generate a phase-coherent composite cyclotron motion at a large radius (Figure 2.2). The coherently orbiting ion packet induces an image current on the detection plates, which is recorded as a time-dependent transient. The detected signal is a superposition of sine waves from ion packets with different *m/z*, and can be converted to the frequency domain by fast Fourier transformation. The cyclotron frequencies of the ion packets can be used to calculate their mass-to-charge ratios according to Equation 2.1. Therefore, compared with quadrupole ion trap, time-of-flight and sector instruments, FT-ICR is unique in that all ions are simultaneously detected and then resolved in the frequency domain. The longer in time that the image current can be recorded, the better defined the cyclotron resonance frequency will be, and thus the higher resolution the mass measurement will be.

In the studies described here, all FT-ICR experiments were conducted with an IonSpec HiResESI FT-ICR instrument (IonSpec, Lake Forest, CA) equipped with a 9.4 Tesla magnet (Cryomagnetics Inc., Oak Ridge, TN). Samples were introduced to an electrospray ionization source (Analytica of Branford, CT). The continuously generated ions were accumulated in an external hexapole gated axially by the skimmer cone and an exit lens. At the end of the accumulation time period, the ion packet was transferred into

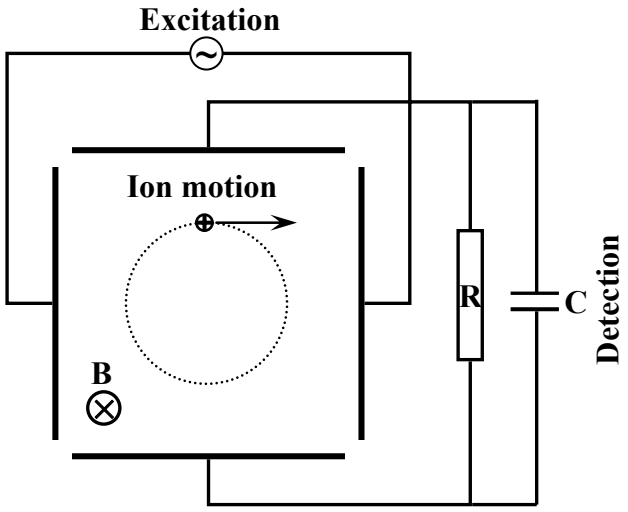


Figure 2.2: Ion cyclotron excitation and detection in FT-ICR cell. A radiofrequency electric field is generated by the two excitation plates to accelerate ions to a spatially coherent packet at detectable orbital radius. Then the ion cyclotron orbital motion is detected by measuring the image current induced in the two detection plates.

an rf-only quadrupole ion transfer device and down to the FT-ICR cell for mass analysis. Usually multiple transients were acquired and co-added to yield a mass spectrum. In sustained off resonance irradiation collision-activated dissociation (SORI-CAD) experiments (Gauthier, 1991), a stored-waveform inverse Fourier transform (SWIFT) pulse was used to isolate a parent ion species from the accumulated ion packet. The isolated parent ions were excited with an rf pulse (1-4 v p-p, 1s) at a frequency 1 kHz lower than the parent ion cyclotron frequency. At the same time, a pulse of nitrogen gas was admitted into FT-ICR cell as collisional gas (maximum pressure of ~ 3 x 10^{-6} Torr). After an 8 – 10 second pump-down delay to re-establish the pressure to ~ 3 x 10^{-10} Torr, the tandem mass spectrum was acquired. The FT-ICR instrument was routinely calibrated with ubiquitin for mass accuracies of ± 5 ppm and mass resolutions of 150,000 (FWHM).

Quadrupole ion traps (QIT) have become the most widely used mass spectrometers for shotgun proteomics measurements, because of their high-throughput and high-dynamic-range tandem mass spectrometry capabilities. QIT-MS is also an ion trapping instrument. However, the QIT-MS traps ions with a dynamic electric field but no magnetic field. QIT-MS has two configurations: three-dimensional form and two-dimensional form.

Three-dimensional QIT consists of two hyperbolic endcap electrodes and a hyperbolic ring electrode between the two endcap electrodes (Figure 2.3A) (Schwartz, 1996). The electric field between the three electrodes is composed of an rf field oscillating at ~1MHz (fundamental rf) and a static electric field. Under this dynamic electric field, the shape of

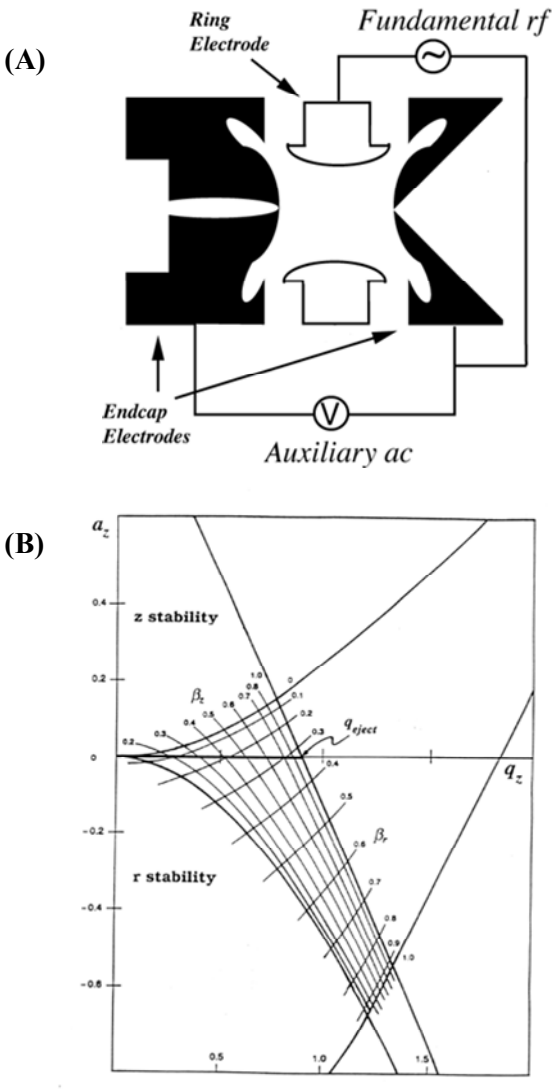


Figure 2.3: Three-dimensional quadrupole ion trap. (A) Schematic drawing of a quadrupole ion trap. The ions are trapped in the space surrounded by the ring electrode and the two endcap electrodes. (B) Stability diagram of ions. The shadowed area shows the theoretical region with both radical stability and axial stability. (Images from www.MatrixScience.com)

the ion ensemble oscillates between a spindle axially pulled towards the two endcap electrode and a disk radially pulled towards the ring electrode. To be trapped in the trap, ions must have axial stability (z stability) and radical stability (r stability) (Figure 2.3B). The complex ion trajectories inside the trap are determined by two parameters, a_z and q_z, which are the *x*-axis and *y*-axis, respectively, of the stability diagram. Ions with a_z and q_z values within the stability region are trapped inside QIT. The value of a_z is generally zero in most commercial QIT instruments. The value of q_z is determined by the mass-to-charge ratios of ions (*m/z*), the radial size of the ion trap (*r*), and the frequency of the fundamental rf (*w*) and the amplitude of the voltage on the ring electrode (*V*), as shown in equation 2.2:

$$q_z = \frac{4V}{\frac{m}{z} \cdot r^2 \cdot w^2}, \tag{2.2}$$

Ions can be manipulated by changing the electric field inside the ion trap. By scanning the amplitude of the fundamental rf voltage, ions can be ejected sequentially from low *m/z* to high *m/z* through the holes in the endcap electrode. Detection of the ejected ions with a conversion dynode and an electron multiplier system during the *m/z* scanning yields a mass spectrum. To perform collision-activated dissociation, ions within an *m/z* window are isolated by ejecting other ions in the trap, and then kinetic energy is deposited into the isolated ions by resonance excitation. The collisions between the excited ions and the helium gas molecules result in fragmentation.

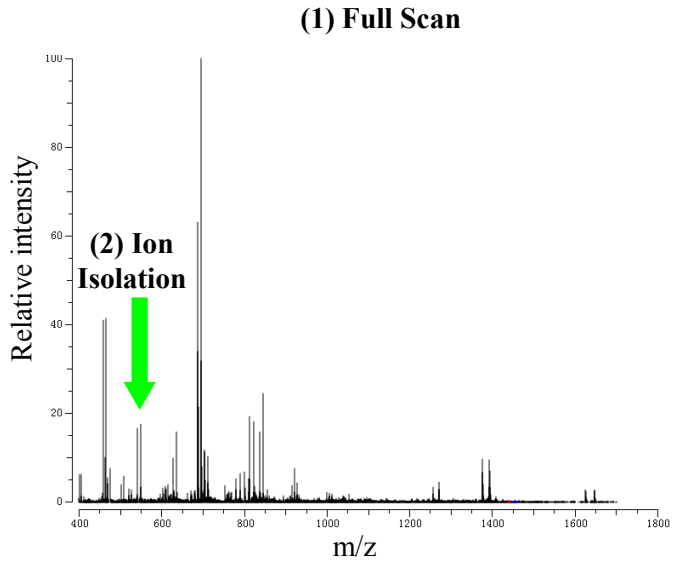
Two-dimensional QIT has a quadrupole made of four hyperbolic cross-sectional rods. The ions are trapped in an axial fashion in the quadrupole. Similar to three-dimensional

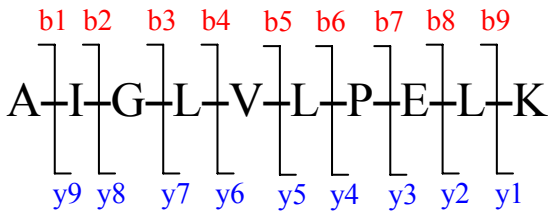
QIT, two-dimensional QIT acquires mass spectra by ejecting ions out sequentially by their *m/z*. A recent commercial implementation of two-dimensional QIT is ThermoFinnigan LTQ-MS. LTQ-MS achieved many functional enhancements over three-dimensional QIT, including 15X higher ion capacity, 3X faster scan rate, and 14X higher trapping efficiency (Schwartz, 2002).

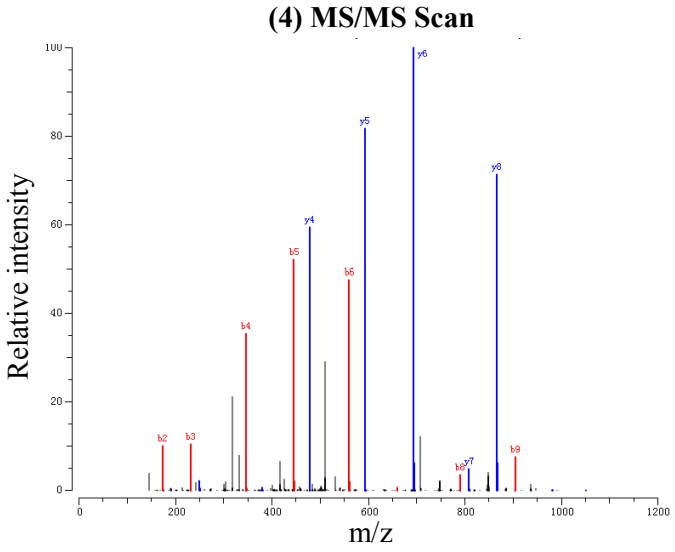
Tandem mass spectrometry

Besides measuring molecular masses, mass spectrometers are capable of interrogating the structure of an isolated analyte with gas-phase fragmentation. The typical procedure for tandem mass spectrometry measurement is illustrated in Figure 2.4. In a full scan mass spectrum, *m/z* and intensities of all ions species are measured. Different mass spectrometers can then use different mechanisms to isolate an ion species at a specific *m/z* window. The isolated ion species is called the parent ion. Fragmentation is then induced on the isolated ion ensemble by a variety of different techniques, such as collision-activated dissociation (CAD) (Senko, 1994), electron capture dissociation (Zubarev, 1998), electron transfer dissociation (Syka, 2004), infrared multiphoton dissociation (Little, 1994), black-body infrared radiative dissociation (Price, 1996), *etc*. Finally, the product ions are measured by the mass analyzer. As two successive stages of mass spectrometric analysis, the first one for the parent ions and the second one for the product ions, are used, this process is called tandem mass spectrometry (MS/MS). The product ion species are largely determined by the amino acid sequence of peptides or proteins. For example, when peptides are fragmented with CAD, the major product ions

Figure 2.4: Tandem mass spectrometry. The four steps for tandem mass spectrometry are (1) full scan; (2) ion isolation; (3) fragmentation; and (4) MS/MS scan. The green arrow in the full scan indicates the ion species to be isolated. When CAD is used for fragmentation, the probable product ions can be predicted from the peptide sequence. The cleavage usually occurs on the peptide bond between two residues. The two major ion series, y ions and b ions, are shown in blue and red, respectively. The product ions are measured in the MS/MS scan. The mass spectral peaks corresponding to the predicted y and b ions are highlighted in blue and red, respectively.


(1) Full Scan
(2) Ion Isolation
Relative intensity
m/z


(3) Fragmentation (CAD)
b1 b2 b3 b4 b5 b6 b7 b8 b9
A I G L V L P E L K
y9 y8 y7 y6 y5 y4 y3 y2 y1


(4) MS/MS Scan
Relative intensity
m/z

are y ions and b ions, generated from the cleavage of the peptide bond (Hunt, 1981). Therefore, the observed product ions of a peptide can be used to infer peptide sequence, often with a computer algorithm for high throughput identifications.

Although the quadrupole ion traps have limited performance in mass measurement (i.e. moderate mass accuracy and resolution), this technique excels in fast and automated tandem mass spectrometry. A three-dimensional quadrupole ion trap can acquire ~15 MS2 scans in a minute, and a linear quadrupole ion trap can acquire ~60 MS2 scans in a minute. The fast scan rate of MS2 allows examination of more peptides at a given retention time window in HPLC-MS measurements. The tandem mass spectrometry in QIT is also highly automated for data-dependent LC-MS/MS analysis; 3 – 5 most abundant peaks on a full scan are selected for MS2 analysis and then pushed into a rolling exclusion list to prevent repeated MS2 scans within their chromatographic peaks. The dynamic gain control feature of QIT also gives this instrument superior sensitivity and dynamic range for tandem mass spectrometry. To achieve this feature, the QIT varies the accumulation time of the isolated ions to fill the ion trap with the same amount of ions for every fragmentation. As the less abundant ions are accumulated for a longer time, high quality MS2 scans can be acquired from the most abundant ions to the least abundant ions in a full scan. This is critical for handling the wide dynamic range of a proteome sample. The large number of MS2 scans acquired in a proteomics measurement necessitates an automated data analysis procedure. The two representative algorithms for peptide identification from tandem mass spectrometry are SEQUEST (Eng, 1994) and MASCOT (Perkins, 1999), which match experimental tandem mass spectra against the theoretical tandem mass spectra predicted from peptide sequence databases.

Shotgun proteomics measurement

Shotgun proteomics is the foundation for the studies presented in Chapters 6, 7 and 8. A shotgun proteomics measurement can be described as a pipeline shown in Figure 2.5. A proteome is extracted from the pellet of cells grown under conditions of interest. The proteome can be fractionated to reduce the sample complexity. The obtained protein mixture is treated with protease, which cleaves all proteins into peptides. The resulting peptide mixture is then analyzed with liquid chromatography-tandem mass spectrometry (LC-MS/MS). Trypsin is the most commonly used protease. Trypsin specifically cleaves at the carboxyl side of the basic amino acids, lysine, and arginine, except when these two residues are followed by proline. Tryptic peptides generally carry a positive charge on both the *N*-terminus and *C*-terminus and can be ionized efficiently by electrospray ionization (ESI) and be fragmented readily by tandem mass spectrometry to yield rich sequence information. Tryptic peptides can also be readily separated at very high resolution by a variety of liquid chromatography techniques.

As proteolysis greatly increases the sample complexity by turning every protein into multiple peptides, two-dimensional separation is generally employed in shotgun proteomics. A number of other liquid chromatography methods have also been coupled with reverse phase LC to generate two-dimensional liquid chromatography separation for the peptides. As strong cation exchange (SCX) LC separates peptides based on their charge and reverse phase (RP) LC separates peptides based on their hydrophobicity, the

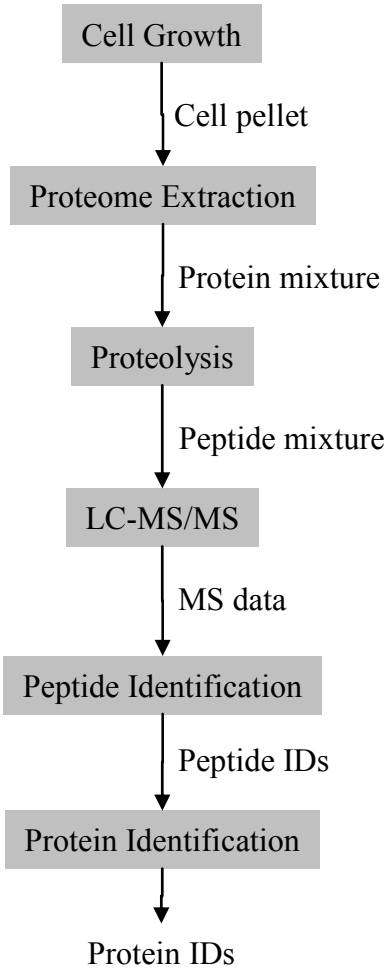


Figure 2.5: Process pipeline of shotgun proteomic measurement. Each step of the pipeline is shown in the grey blocks. The first four steps are experimental steps and the last two steps are computational steps.

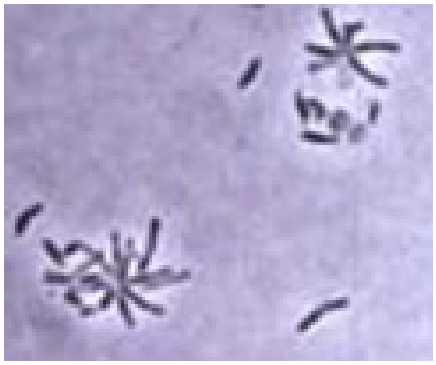
common two-dimensional separation is SCX LC as the first dimension separation and RP LC as the second dimension separation. In this dissertation, the multidimensional protein identification technology (MudPIT) has been used for 2-dimensional LC separation (Link, 1999; MacCoss, 2002; McDonald, 2002).

Reverse phase LC is generally interfaced directly with a mass spectrometer via electrospray ionization in order to minimize post-column peak broadening and to increase analysis throughput. Peptides are automatically selected for the tandem mass spectrometry measurement, and the fragmentation product ions of peptides can yield sequence information needed for computer algorithms to definitively identify peptides. Finally, the identified peptides are computationally assembled into proteins, using a protein sequence database. Here, the computer programs, SEQUEST (Eng, 1994) and DTASelect (Tabb, 2002), are used for peptide identification and protein identification, respectively.

The metabolically versatile bacterium *Rhodopseudomonas palustris*

R. palustris is a purple non-sulfur phototrophic bacterium (Figure 2.6). It has been recognized as one of the most metabolically versatile bacteria. *R. palustris* is capable of utilizing three energy sources (light, inorganic compounds, and organic compounds), two carbon sources (wood-derived compounds and carbon dioxide), and three electron donor sources (oxygen, carbon, and nitrogen) in response to different growth conditions (Larimer, 2004). Because of its extraordinary adaptability to different environmental

(A)



(B)

Figure 2.6: ***Rhodopseudomonas palustris*****.** (A) The clustered bacterial cells under microscope. (B) batch growth in laboratory under the anaerobic photoheterotrophic state. Figure courtesy of Dr. Dale A. Pelletier.

conditions, *R. palustris* is widely distributed in nature, including soil, aquifers, aquatic sediments, underground water, pond water, *etc* (Oda, 2003). This bacterium grows under all the four types of metabolism: photoautotrophic (energy from light and carbon from carbon dioxide), photoheterotrophic (energy from light and carbon from organic compounds), chemoheterotrophic (energy and carbon from organic compounds), and chemoautotrophic (energy from inorganic compounds and carbon from carbon dioxide) (Larimer, 2004). The respiration mode of *R. palustris* can be switched between anaerobic growth and aerobic growth according to oxygen availability (Harwood, 1988). Hence, *R. palustris* is used as a model organism to study how a biological system responds to changes in carbon, nitrogen, electron, and energy sources by adjusting its gene expression profile and metabolic network.

R. palustris also has the remarkable capability to degrade diverse aromatic compounds under anoxic environments (Dutton, 1967). Biodegradation of aromatic compounds is a vital link in the carbon cycle of our ecological system. Recycling of lignin, perhaps the second most abundant carbon polymer on Earth, requires degradation of its phenolic monomers (Kirk, 1984). The large quantities of industrially-generated aromatic contaminants in the environment necessitates remediation, and one possible remedy appears to be biodegradation (Xu, 1996). These aromatic compounds are often released into anoxic environments. Although many bacteria can catabolize aromatic compounds with oxygen (Zylstra, 1991), *R. palustris* is one of a few bacteria capable of disrupting benzene rings by a reduction reaction without using oxygen (Harwood, 1999). Thus, *R.*

palustris is a model bacterium to understand anaerobic aromatic compound degradation pathways and their regulation.

R. palustris also draws tremendous interest because of its potential to be engineered for generating hydrogen as biofuel (Barbosa, 2001). Although there are other bacteria that can produce hydrogen using hydrogenase, the reaction is reversible, which prevents significant hydrogen accumulation in a closed chamber (Hilhorst, 1982). *R. palustris* possesses three nitrogenases for fixing dinitrogen gas and generating hydrogen (Oda, 2005). Although the nitrogenase activity requires large amount of ATP, the reaction is largely irreversible. Thanks to the metabolic diversity of *R. palustris*, it can collect ample energy from sunlight through photosynthesis and derive copious reducing equivalents through aromatic compound degradation.

Because of these interesting features of *R. palustris*, its genome was sequenced in 2004 (Larimer, 2004). The genome consists of a circular chromosome with 5,459,213 base pairs and a plasmid with 8,427 base pairs. A total of 4,836 genes are predicted from the *R. palustris* chromosome. Its metabolic versatility is conferred by the large number of genes involved in cellular metabolism, which account for 31% of the predicted genes. *R. palustris* harbors 451 potential regulatory and signaling genes to sense the environment for different resources and regulate the metabolism genes for optimal growth. While most bacteria devote 5–6% of genes in their genome to transportation, *R. palustris* has 325 transport systems comprising at least 700 genes, which sum up to about 15% of the genome. Many of the transport systems are hypothesized to be responsible for aromatic

compound trans-membrane transportation. All the studies described in this dissertation focus on *R. palustris* as a model organism for methodology development and biological application.

Chapter 3

Multipole-Storage Assisted Dissociation for Characterization of Large Proteins and Protein Mixtures

All of the data presented below has been published as

C. Pan, R.L. Hettich. Multipole-Storage-Assisted Dissociation for the Characterization of Large Proteins and Simple Protein Mixtures by ESI-FTICR-MS. *Analytical Chemistry* 2005, 78, 3072-3082.

C. Pan's primary contributions include experimental design, FT-ICR measurement and data interpretation.

INTRODUCTION

High-resolution MS experiments with Fourier transform ion cyclotron resonance (FT-ICR) mass spectrometry provide exquisite information about the molecular masses of intact proteins. However, for unambiguous identifications, it is advantageous to supplement these measurements with ion fragmentation experiments to obtain in-depth information on protein sequence, post-translational modifications, and even higher order structure (Kelleher, 1999). For example, a wide range of dissociation techniques for intact proteins have been implemented in the analyzer cell of FT-ICR mass spectrometers, including sustained off resonance irradiation collision-activated dissociation (SORI-CAD) (Senko, 1994), electron capture dissociation (ECD) (Zubarev, 1998), infrared multiphoton dissociation (IRMPD) (Little, 1994), and black-body infrared radiative dissociation (BIRD) (Price, 1996). The most common procedure for conducting these ion

dissociation methods involves isolating an ensemble of parent ions at a given mass-to-charge ratio (*m/z*) inside the analyzer cell, and then activating the trapped parent ions by these different methods to achieve fragmentation. While these dissociation methods have proven to be quite valuable, they have severe limitations for very large proteins and for the high-throughput investigation of protein mixtures.

Most current FT-ICR mass spectrometers also utilize a second ion-trapping/ accumulation device: an external rf-only hexapole or octapole bounded by electrostatic elements (Senko, 1997). Electrosprayed ions traverse the skimmer cone and are accumulated in the rf-only linear multipole storage trap by employing a dc-controllable gate electrode at the exit end of the multipole, as illustrated in Figure 3.1. The voltage and timing of this gate provides the ability to accumulate ions for a desired period of time, after which they can be transported out of the multipole and down to the FTICR analyzer cell for mass/charge measurement. Because electrospray ionization (ESI) is continuous, the multipole functions as a linear ion trap to admit and accumulate a sufficient ion population for eventual FT-ICR ion detection. New fragmentation techniques for FTICR-MS have been developed by exploiting ion dissociation in this linear ion trap, with either gas phase collisional activation accomplished with multipole-storage assisted dissociation or MSAD (Sannes-Lowery, 1998; Hakansson, 2000; Palmblad, 2000; Sannes-Lowery, 2000; McDonnell, 2002; Keller, 2004), "ion thrashing" (McFarland, 2004) or photon-induced dissociation (termed external IRMPD) (Hofstadler, 1999; Hofstadler, 2003) While these approaches may seem to be a minor variation of the established CAD techniques listed above, in fact these multipole dissociation methods afford a number of

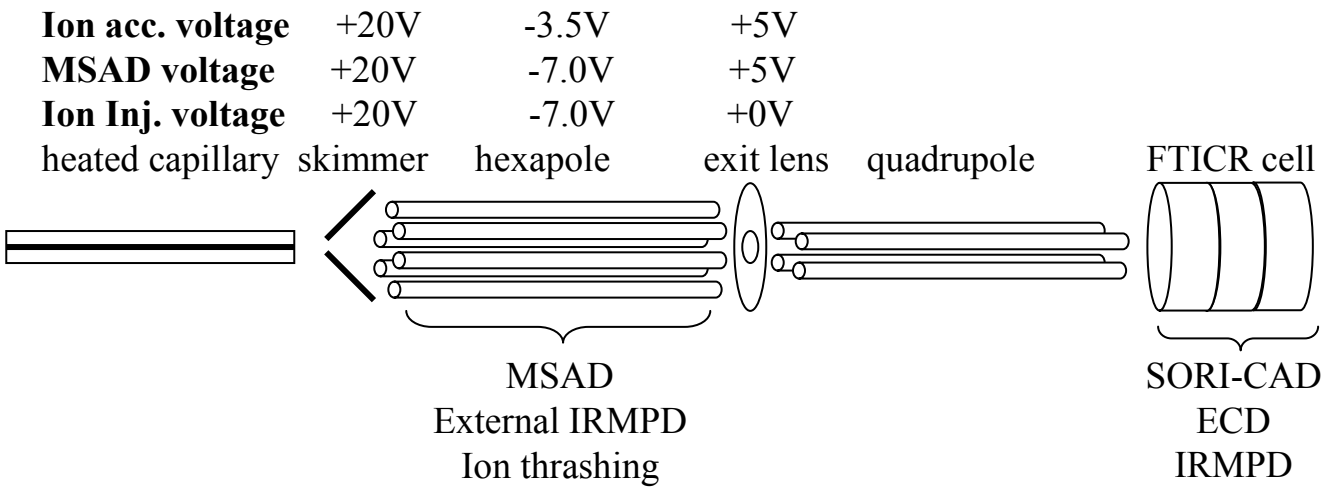


Figure 3.1: Ion optics of FT-ICR mass spectrometer (not drawn to scale). During the ion accumulation stage, ions flow through skimmer and are trapped in hexapole, confined radially by the rf voltage of the rods and axially by the skimmer voltage, exit lens voltage and hexapole dc offset voltage. MSAD, ion threshing, and external IRMPD can be accomplished in this stage at the location indicated. During ion injection stage, ions are transferred through quadrupole and are trapped in FT-ICR cell primarily by the magnetic field. Conventional CAD, IRMPD, ECD etc. can be induced at this stage inside FTICR cell. Voltage settings are listed directly above the appropriate electrostatic component.

advantages, including eliminating the need for a collision gas in the high vacuum region of the FTICR instrument, and the ability to conduct multiplexed fragmentation at relatively high energies.

MSAD was first observed by accumulating ions in the multipole for an extended timeframe (Sannes-Lowery, 1998). It was postulated that once the ion density reaches the space charge limit in the multipole, the Coulomb force will push the ion ensemble to spread out radially, enabling the ions to oscillate at higher amplitude. This would allow coupling of the rf energy in the hexapole rods to the ions, effectively accelerating them to higher kinetic energy (Hakansson, 2000; Sannes-Lowery, 2000; Belov, 2001). Fragmentation then would result from the collisions of excited ions with the background gas molecules in the hexapole (typically air at ~10^{-5} Torr), and thus is generally regarded as a form of CAD. Like nozzle-skimmer collisional activated dissociation, MSAD is also an in-source fragmentation. Compared with SORI-CAD, MSAD obviates the need for introduction of collisional gas into the analyzer cell and subsequent pump-down. However, in an rf-only multipole, no parent ion selection is possible; thus MSAD fragments all species present, which limits its use in a targeted fragmentation experiment.

We have undertaken a systematic investigation of ways to control the collision energy and fragmentation pattern for intact proteins to evaluate this MSAD process. In particular, we have focused on examination of the hexapole dc offset voltage and accumulation time, which are the two key parameters in controlling the ion population. Seven representative proteins covering a molecular mass range of 8-116 kDa were

employed to study the fragmentation pattern of intact proteins under a variety of MSAD conditions. In addition, the ability to conduct MSAD experiments on protein mixtures was also investigated.

To extend the capabilities of MSAD, we have devised an experimental method in which selected MSAD fragment ions were subjected to a further stage of tandem mass spectrometry in the FTICR analyzer cell. This MS^3 type experiment enables coupling of the efficient, relatively high-energy MSAD process with the more selective SORI-CAD. The goal of this approach was to generate sequence tag information by dissociation of the MSAD fragment peptide for protein identification, in a manner analogous to generating sequence tags by dissociation of peptides from enzymatic digestion (Mann, 1994). McLafferty and coworkers have employed a similar approach to directly generate sequence tag from intact proteins (Mortz, 1996; Horn, 2000a). They have shown that a sequence tag and an intact protein mass were sufficient to identify a protein from a protein database. Although this approach appears to be quite promising, there are at least three major challenges that complicate this method. First, the fragments of intact proteins are usually very large and exhibit a wide isotopic package. When comparing the masses of two adjacent fragment ions in an effort to identify the residual amino acid, the difficulty in accurately choosing the correct isotopic mass in each packet can lead to the so-called '1 Da error' (Horn, 2000b). Because these high-resolution measurements do not directly determine the average molecular masses, transposing the measured isotopic masses into an average value has some inherent uncertainty due to the variation in peak height abundances (which can skew the calculated average mass value and thus degrade

the resolution of the mass measurement). Either way of calculating the mass difference will compromise the reliability of obtaining sequence tag information from an unknown protein. Second, due to the large size of intact proteins and their residual tertiary structures, it is very difficult to establish a standard dissociation energy that can induce substantial fragmentation at multiple consecutive peptide bonds, which prevents implementing this approach in a robust fashion to most proteins. Third, the standard dissociation techniques (SORI-CAD, ECD, IRMPD) are virtually ineffective for dissociating very large proteins (Mr > 100 kDa). In contrast, obtaining sequence tag information from peptides that are generated by proteolytic digestion is relatively straight-forward to measure and interpret. In fact, the MS3 approach consisting of an in-source dissociation step and a conventional dissociation step has been used for deriving sequence information from oligonucleotides, oligosaccharides, peptides, and intact proteins (Chen, 2001a; Raska, 2002; Suckau, 2003; Ginter, 2004). Recently, a new sequence tagging approach for intact protein with in-source dissociation has been shown by using a class of "mass defect" tags incorporating the element ^{35}Br (Hall, 2003). While the objectives of this study are similar to those aforementioned techniques, we feel that the capabilities of the MSAD technique, in particular for efficient high-energy dissociation, make this uniquely suited for this approach. We propose that MSAD can be used to efficiently generate small fragment ions from intact proteins with molecular masses exceeding 100 kDa, and these fragment ions can be further dissociated to give sequence tag information.

MATERIALS AND METHODS

All protein standards were acquired from Sigma-Aldrich (St. Louis, MO) and used as received with no additional purification. Samples for mass spectrometry were prepared at formal concentrations of 520 μM in 50:50 (v/v) acetonitrile:water, with 0.1% acetic acid added. All mass spectrometry experiments were conducted with a HiResESI Fourier-transform ion cyclotron resonance mass spectrometer (IonSpec, Lake Forest, CA) equipped with a 9.4T magnet (Cryomagnetics Inc., Oak Ridge, TN). Samples were introduced to an electrospray source (Analytica of Branford, CT) by direct infusion at 2–3 μl/min. Ions were accumulated in an external hexapole situated between the skimmer cone on one end and an exit lens and mechanical shutter (Figure 3.1) on the other. The static pressure in this region of the instrument was typically around 2 x 10^{-5} Torr. At the end of the accumulation time period, the exit lens voltage was dropped to zero and a mechanical shutter was pulsed open to allow ion transfer into an rf-only quadrupole ion transfer device and down to the ICR cell. In SORI-CAD experiments, ion accumulation (typically 0.5 to 3 s) was followed by ion isolation, which was accomplished with a SWIFT pulse. Off-resonance ion excitation was achieved with an rf pulse (1-4 v p-p, 1s) at a frequency 1 kHz lower than the parent ion cyclotron frequency, in the presence of nitrogen which was admitted with a pulsed valve to a transient pressure of 5 x 10^{-6} Torr. An 8-10 s pump-down delay was inserted to allow the base pressure to re-establish (~ 3 x 10^{-10} Torr) prior to ion detection. For normal ESI-FTICR-MS experiments, ion accumulation was usually performed for 0.5 to 3 s at a hexapole dc offset voltage of –3.5 v, as shown in the top line of Figure 3.1. This yielded multiply-charged molecular ions

with virtually no fragmentation. To achieve MSAD, ion accumulation/ activation was accomplished by lengthening the accumulation times (2–8 sec) and adjusting the hexapole dc offset voltage (–7 to –12 v). This condition creates a deeper axial potential well than the standard offset setting (–3.5 v) and promotes ion fragmentation during the accumulation period. Discrete parent ion isolation and collision gas pump-down delay times were not necessary, so overall scan times for MSAD were determined solely by the accumulation times (~2–6 s per scan). Each spectrum was comprised of ten co-added scans acquired at 512K data points/transient, and external calibration was performed with ubiquitin; these conditions typically result in mass accuracy of ± 5 ppm and resolutions of 150,000 (FWHM) for intact proteins. Product ion spectra were deconvoluted to zero charge state with the IonSpec software deconvolution tool. Sequence tags were identified by manual inspection of the deconvoluted spectra in following steps. First, the mass difference between two fragment ions masses and between fragment ion and parent ion mass were calculated and an amino acid was assigned if this mass difference corresponded to an amino acid mass, denoted by |← →| in the tandem mass spectra. Second, the mass differences between the parent ion mass and the sum of two fragment ion masses were calculated. If the two fragments are complementary ion types (e.g. y and b ion type) and have an amino acid between them, then the mass difference calculated would give the identity of this amino acid, denoted by →| |← in tandem mass spectra. If this mass difference corresponded to an amino acid mass, these two fragments may have arisen from two complementary ion fragment species (e.g. y or b ions) with this amino acid situated between them. Third, contiguous identified amino acids constituted a sequence tag.

RESULTS AND DISCUSSION

Experimental Parameters for Controlling MSAD in a Hexapole Storage Trap

In order to optimize the MSAD technique, a systematic examination of the experimental parameters governing this dissociation method was undertaken. The two key factors involved in MSAD (at similar protein concentrations) were observed to be the ion accumulation time in the hexapole and dc offset voltage. The dependencies on accumulation time, rf amplitude, skimmer and exit lens voltages, and target gas pressure in hexapole have been discussed previously (Hakansson, 2000; Sannes-Lowery, 2000), but the effect of the dc offset voltage has not been reported in literature.

The dc offset voltage controls the depth of the electrostatic axial well. To probe the effect of dc offset voltage on fragmentation, we examined the MSAD of the protein apomyoglobin with the accumulation time maintained at 4 seconds and all other parameters kept constant. When dc voltage is between -3.5 v and -6 v, mass spectra revealed no fragmentation for most protein ions, as shown in Figure 3.2A for apomyoglobin. A sharp threshold for dissociation is observed at dc offset voltages between –6 v and –7 v. For example, at –6.5 v, apomyoglobin dissociates into two types of fragments; a few abundant multiply-charged fragments and many low-abundance singly-charged fragments (Figure 3.2B). At –7 v, low-mass, singly-charged fragments dominate the mass spectra (Figure 3.2C). From –7 v to –11 v, no noticeable differences in

Figure 3.2: Apomyoglobin MS^2 from MSAD. The ion accumulation time and hexapole dc offset voltage offset respectively are at 4 sec and –6.0 v (A), 4 sec and –6.5 v (B), 4 sec and –7 v (C) and 1.4 sec and –10 v (D). A sharp threshold of hexapole dc offset voltage for MSAD from no dissociation (A) to extensive sequential dissociation (C) was observed. Similar fragmentation can be achieved by long accumulation time and high dc offset voltage (A) or short accumulation time and very high dc offset voltage (D). The intermediate condition for MSAD yielded two distinctive population of fragments, singly charged small fragments and highly charged large fragments (B).

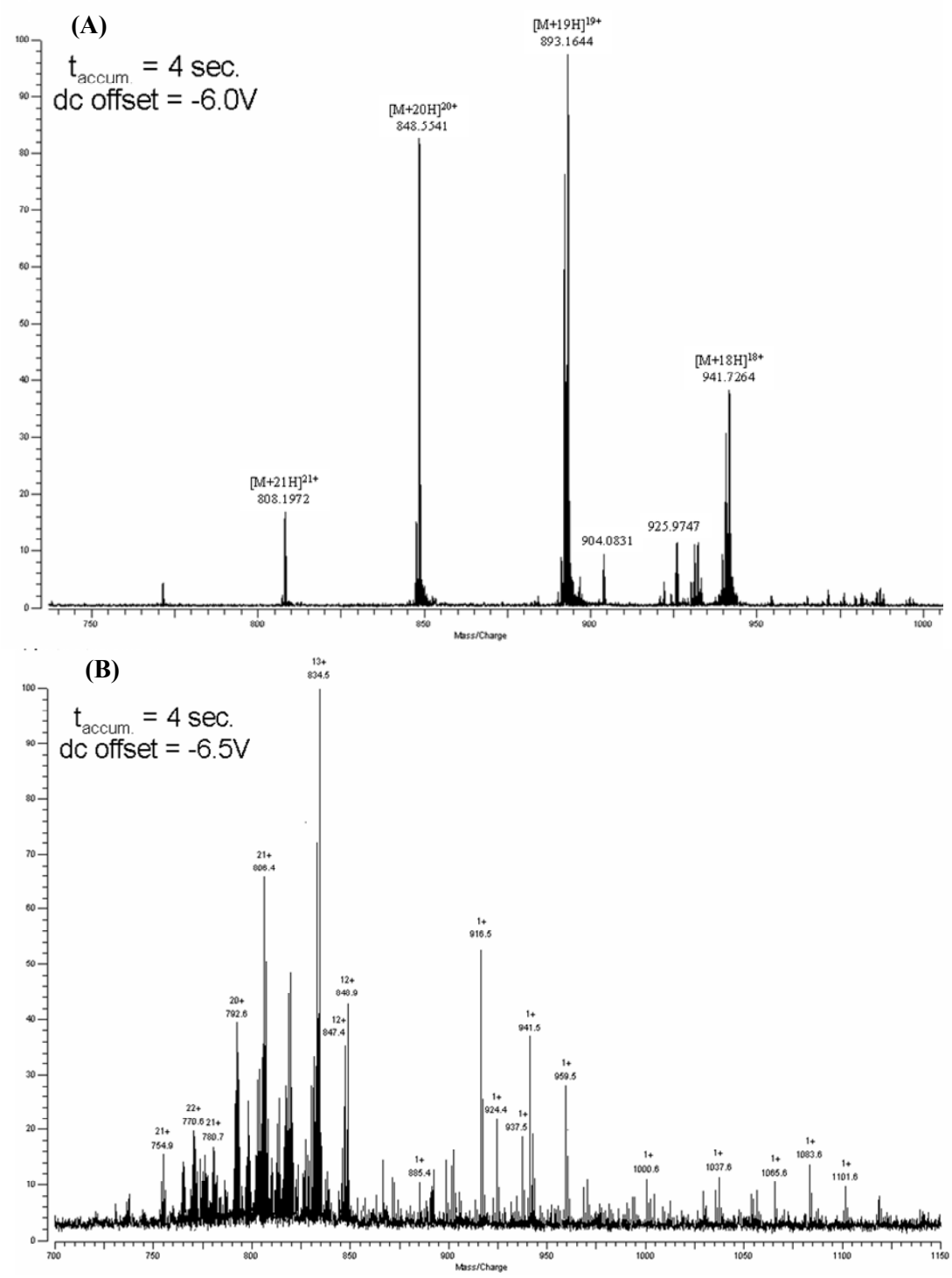


Figure 3.2: Continued.

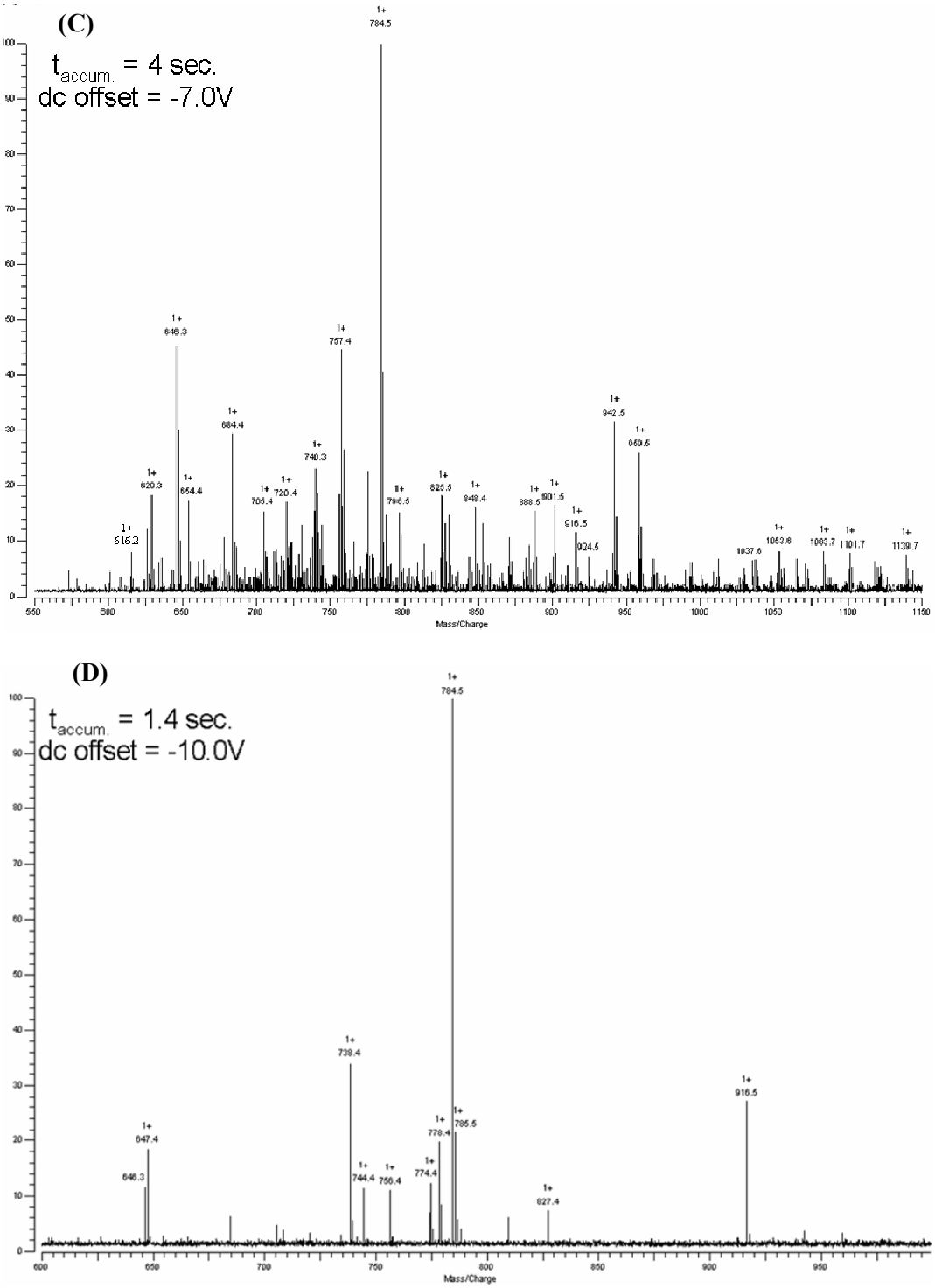


Figure 3.2: Continued.

fragment ion species are observed. However, a more negative dc voltage will induce fragmentation at a much shorter accumulation times (Figure 3.2D). These results indicate that even at a fixed ion accumulation time in the hexapole, the magnitude of the dc offset voltage has a dramatic effect on ion fragmentation.

The information obtained above suggests that a combination of dc offset voltage and ion accumulation time can be used to effectively control the degree of fragmentation in a MSAD experiment. Empirically, the lower boundary for fragmentation to occur involves accumulation times of at least 1200 ms and for dc voltages of at least -6.5 v for the protein samples (~uM concentration) that were examined in this study.

The ion storage/accumulation capability of a hexapole is controlled by the confining forces of the multipole device. In particular, the electrostatic potential created by the rf-only mode of operation of a hexapole provides extensive ion confinement in the x-y direction (i.e. perpendicular to the hexapole rods), and somewhat more limited ion confinement in the z-direction (parallel to the hexapole rods). By employing electrostatic voltage confinement at the ends of the hexapole, it is possible to accumulate and store ions for an extended period of time in the hexapole device. Previous reports have suggested that extended ion accumulation results in a sufficiently large ion population for which space charge pushes the ions outward radially and allows energy coupling with the rf-only hexapole rods. However, our experiments on the dc offset voltage and previous reports on the skimmer and exit lens voltage (Sannes-Lowery, 2000; Belov, 2001) have revealed that the depth of axial potential well is critically important, and can induce

fragmentation. These results suggest an alternative fragmentation process. Because the voltage at the entrance of the hexapole is static (usually held at 25 v in our experiments), the lowering of the dc offset on the hexapole to more negative voltages will induce a translational energy component to the ions as they enter the hexapole. Since the ions are confined in the hexapole in a multiple pass configuration along z-direction, even a modest amount of translational energy added as the ions enter the multipole device will result in substantial fragmentation. Note that the higher charged parent ions will pick up a proportionately higher translational energy as they enter the multipole. Therefore, while we cannot rule out the possibility of rf-coupling with the hexapole rods as the energy source for fragmentation, we believe that we have identified an additional MSAD fragmentation mechanism, in which a translation energy component can be exploited to produce substantial ion fragmentation in the multipole device. Note that MSAD is a single excitation process; whereas ion thrashing is multiple excitation process that can be tuned somewhat (McFarland, 2004).

Although variation of sample concentration, rf amplitude, skimmer and exit lens voltages and collision gas pressure in the hexapole undoubtedly would also affect the MSAD experiment, those parameters were not examined in this study. Control of the collisional energy was based only on the dc offset voltage and accumulation time.

Single Protein MSAD: Identification of Fragmentation Extent and Ion Types

A range of proteins were examined with MSAD, to determine the general utility of the technique as well as investigate any sequence dependent fragmentation. Previously, only limited research had been conducted on pure small to medium sized proteins over a narrow low energy range. In this study, we have conducted MSAD over a wide collisional energy range on seven individual proteins whose molecular masses range from 8-115 kDa (ubiquitin, lysozyme, apomyoglobin, ß-lactoglobulin B, carbonic dehydrogenase, serum albumin, and ß-galatosidase). The proteins examined in this study exhibit substantial diversity in their amino acid sequence, molecular weight, and number of disulfide bonds present, and thus should represent a general case for other proteins. All protein samples were prepared by directly solubilizing the protein into the ESI solution (see MATERIALS AND METHODS section). Due to the preservation of disulfide bonds and the gentle experimental conditions, these proteins may have a large amount of residual tertiary structure.

It has been demonstrated that proteins have similar fragmentation behavior in low energy MSAD experiments and in SORI-CAD experiments (Hakansson, 2000). We also have verified this trend under our experimental conditions (Uchiki, 2002; Keller, 2004). In our study, MSAD with accumulation times ranging from 2000 – 3000 ms, and dc voltages ranging from -6.0 to -6.5 v were regarded as low energy conditions. This rather empirical range is defined as low energy based on the experimental observation of a small amount of fairly large fragment ions for most of the proteins examined. In these typical low

energy MSAD experiments, the mass spectra consisted of a few large y- and b-type ions derived from the parent molecular species. The observed fragmentation is less extensive in low energy MSAD experiments than in SORI-CAD experiments. However, the types of fragment ions common to MSAD (i.e, y- and b-type species) are quite similar to SORI-CAD experiments and differ substantially from ECD (which is predominantly c- and z-type ions). In particular, both MSAD and SORI-CAD not only reveal similar types of fragment ions (Keller, 2004), but also a common preference for dissociation at residues such as proline, asparatic acid, and glutamic acid in some cases. These results verify that MSAD is a gas-phase collisional activated dissociation process, and may serve as a higher duty cycle experiment than SORI-CAD (provided that ion isolation is not required). A large amount of undissociated parent ion is present in the low energy MSAD experiment, which indicates there is no clear cutoff between the normal MS experiment and the low energy MSAD experiment.

Even though low energy MSAD is attractive due to its similar fragmentation with SORI-CAD, high-energy dissociation makes MSAD unique among SORI-CAD and nozzle-skimmer CAD in terms of the amount of collisional energy that can be put into protein ions. High-energy collisions can be achieved simply by elongating the accumulation time and/or adjusting the magnitude of the dc offset voltage. In this study, we found 4-sec accumulation time and -7 v dc offset is a generic high energy MSAD condition that can be employed to dissociate most proteins.

This MSAD process technique is illustrated for the protein beta-lactoglobulin B in Figure 3.3. "Normal" electrospray mass spectra can be acquired easily with a modest accumulation time (2 sec) and dc offset (-3.5 v), and reveal multiply-charged ions corresponding to the protonated molecule with no fragmentation (Figure 3.3A). The inset reveals the isotopic resolution of the deconvoluted molecular ion region, illustrating the high resolution capabilities of the FTICR-MS technique. By altering the hexapole conditions to those listed above for high-energy MSAD (i.e. 4 sec accumulation time with -7 v dc offset), it was possible to completely dissociate the protein into small, singly-charged fragment ions, as shown in Figure 3.3B. Similar fragmentation results from high-energy MSAD experiment were observed in all other examined proteins. The MSAD fragments are generally small, singly charged, abundant, and quite distinct for different proteins. When longer accumulation times and more negative dc offset voltages were used, the types of fragment ions remain basically the same, although their relative abundances vary and the overall signal/noise for the spectra decreases. The small size of the fragments suggests they may come from sequential fragmentation. This is further supported by the identity of fragments determined by the sequence tag technique, as will be discussed below. Thus these MSAD fragments correspond to not only classical terminal fragment species such as y- and b- ions, but also internal fragment species such as y/b ions from parent ion. In our proposed MSAD mechanism, the sequential fragmentation may be a result of ions' multiple pass in the z-direction in hexapole. Alternatively, this could be explained with the space-charge repulsion mechanism by considering the continuous deposition of energy through rf-coupling with the hexapole. Thus high-energy MSAD tandem mass spectra for large proteins, while complex, are not

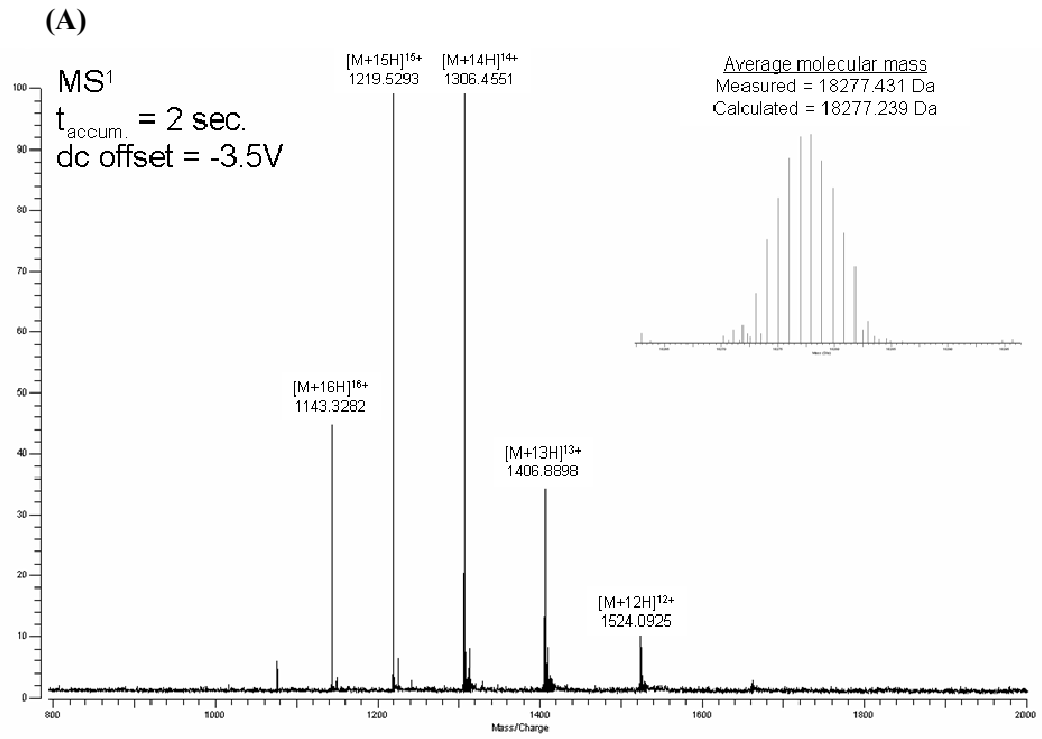


Figure 3.3: Examination of ß-lactoglobulin B MS³ by MSAD/SORI-CAD. A sequence tagging experiment consists of normal MS (A) for determining molecular weight of intact protein, MS² (B) from MSAD for identifying fragments and MS³ (C) from MSAD/SORI-CAD for acquiring sequence tag from a MSAD fragment. The deconvoluted mass spectrum and protein sequence is show in (A) inset. The MSAD fragment indicated with an arrow in (B) is isolated and fragmented, as shown in (C). This fragment is highlighted in the protein sequence with sequence tag underlined. Most ions in MS³ can be identified using the general rules of peptide CAD fragmentation.

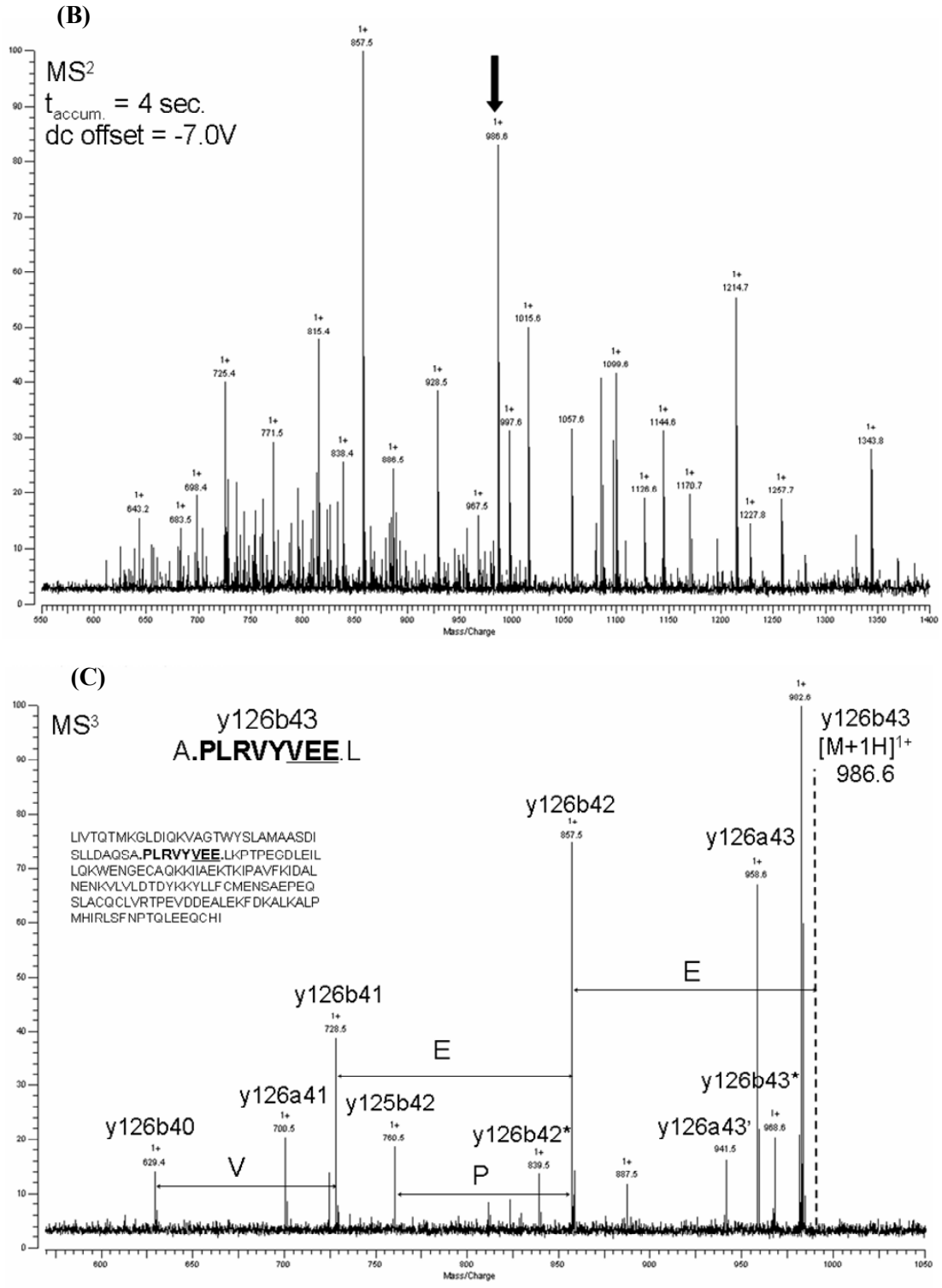


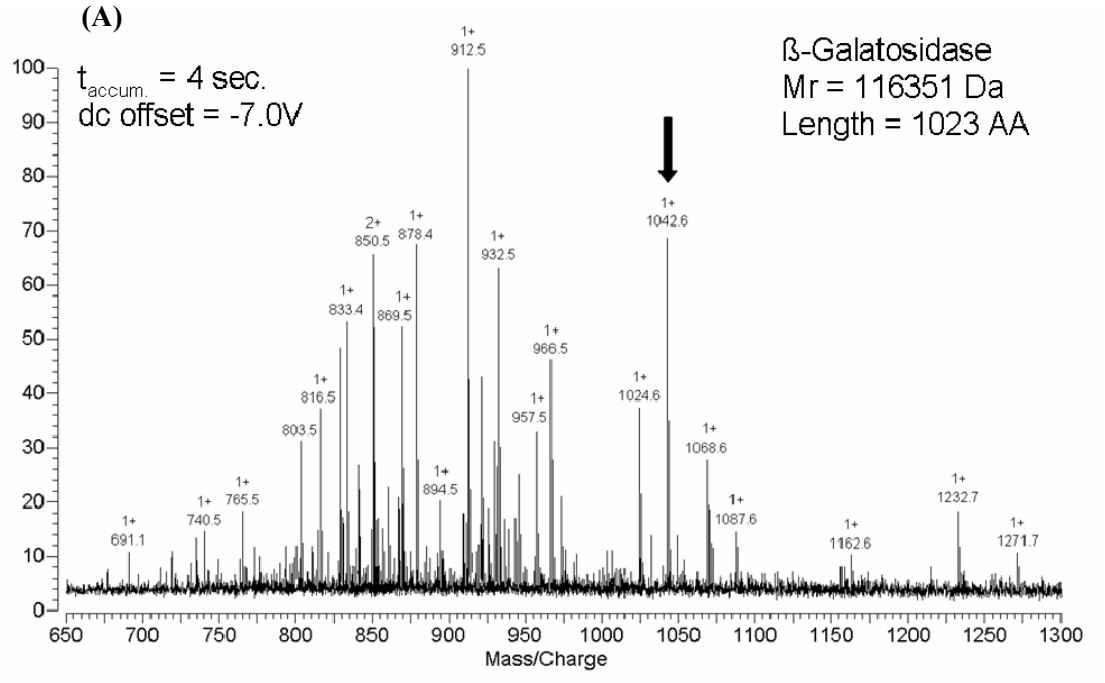
Figure 3.3: Continued.

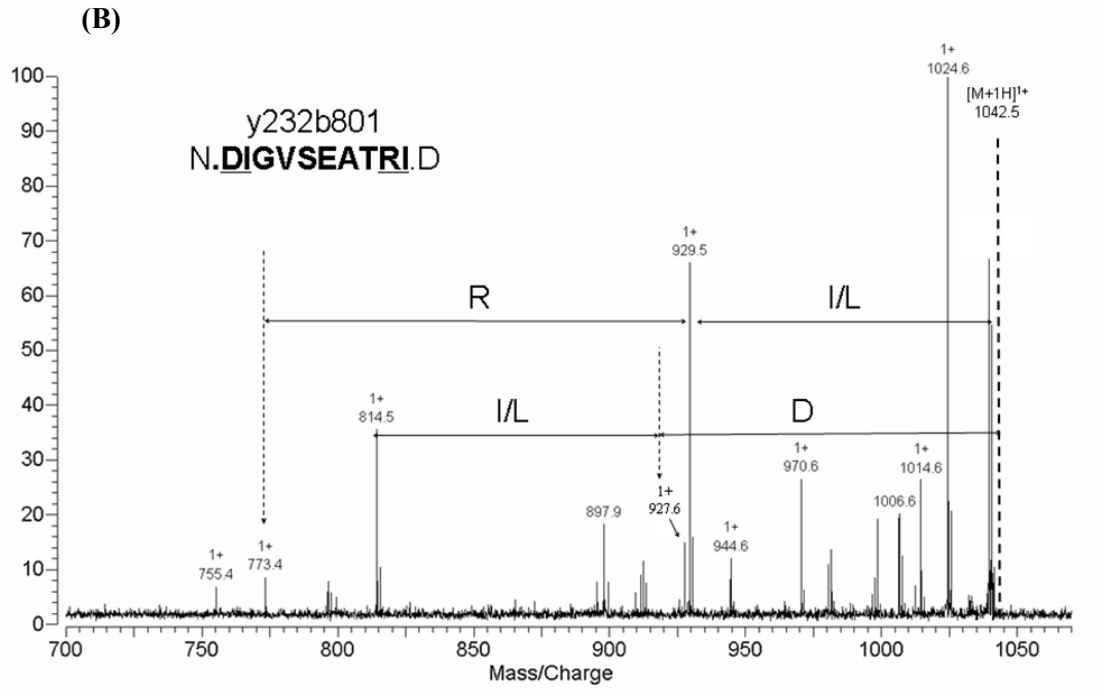
completely intractable. Many of the initial fragment ions are not stable enough to survive the multiple high-energy collisions. The most stable fragments, corresponding to the abundant peaks in the spectrum, are undoubtedly dictated by their sequence, the sequence surrounding them, the residual protein tertiary structure, and the distribution of positive charges. Because the stable fragments are fairly characteristic for each protein (due to the complex factors involved), we refer to them as "MSAD signature", which potentially could provide identification of a protein.

As a remarkable demonstration of the high collisional energy, MSAD was used to fragment ß-galatosidase, which has a molecular mass of 116,351 Da (Figure 3.4A). Note that because the translational energy gained scales proportionally to the protein's charge state, the accumulation time and dc offset voltage of the MSAD experiments for ß-galatosidase are identical with those for smaller proteins. To our knowledge this may be the most extensive fragmentation of a protein whose molecular mass is over 100 kDa. With conventional CAD, IRMPD or ECD experiments, even if fragmentation could be achieved, the fragments of this protein most likely would still be too large to be easily resolved in FTICR-MS. In contrast, the sequential fragmentation under MSAD was able to dissociate the intact protein to fragment ions that are easily measured.

Under "normal" hexapole ion accumulation conditions, no fragmentation occurs. In low energy MSAD, limited dissociation occurs to generate large fragment ions. In high energy MSAD, sequential fragmentation occurs until only small singly- or doubly-charged fragments remain. This leaves one to speculate about a protein's fragmentation

Figure 3.4: Examination of ß-galatosidase MS^3 by MSAD/SORI-CAD. Despite this protein's large size (~116 kDa), MSAD with 4 sec. ion accumulation and –7.0 v dc offset voltage generated a complex pattern of small fragments (A). Its MSAD condition and fragmentation pattern were similar to other medium-sized proteins. Sequence tag information was then derived (B) with MSAD/SORI-CAD from a MSAD fragment. This fragment is a y/b ion resulted from sequential fragmentation during MSAD.


(A)
$t_{accum.}$ = 4 sec.
dc offset = -7.0V
ß-Galatosidase
Mr = 116351 Da
Length = 1023 AA
1+ 912.5
1+ 1042.6
2+ 850.5
1+ 878.4
1+ 932.5
1+ 833.4
1+ 869.5
1+ 966.5
1+ 816.5
1+ 1024.6
1+ 957.5
803.5
1+ 1068.6
1+ 894.5
1+ 765.5
1+ 740.5
1+ 1087.6
1+ 1232.7
1+ 691.1
1+ 1162.6
1+ 1271.7
Mass/Charge


(B)
y232b801
N.DIGVSEATRI.D
1+ 1024.6
[M+1H]1+ 1042.5
1+ 929.5
R
I/L
1+ 814.5
I/L
D
1+ 970.6
1+ 1014.6
1006.6
897.9
1+ 927.6
1+ 944.6
1+ 755.4
1+ 773.4
Mass/Charge

behavior in the intermediate energy MSAD experiment. By definition, intermediate energy MSAD should give medium size fragments with multiple charges (more than three). When using experimental conditions intermediate between high energy MSAD and low energy MSAD, the simultaneous coexistence of large fragments and small fragments, instead of medium size fragments, was observed (Figure 3.2B). Such a sharp transition between high and low energy MSAD suggests that intermediate energy MSAD may be difficult to achieve. This is probably because once sufficient collisional energy conditions are achieved in an MSAD experiment, the sequential fragmentation process will continue to reduce mid-sized fragment ions to the smaller, more stable species.

Sequencing MSAD Fragment Ions by Subsequent SORI-CAD

High-energy MSAD is a very efficient way to generate small-sized, singly charged peptides. When subjected to SORI-CAD experiment, many of these fragment peptides produce an easily interpretable tandem mass spectrum that often yields a sequence tag for the protein. This MS^3 experiment is illustrated with ß-lactoglobulin B in Figure 3.3. The accurate mass of the multiply charged intact protein is first measured with a normal mass spectrum (Figure 3.3A). Then the accumulation time is extended to 4 s and the dc offset voltage is decreased to –7 v to acquire this protein's MSAD tandem mass spectrum (Figure 3.3B). Now the high complexity of the fragmentation products is actually advantageous for offering a wide range of peptide fragments for SORI-CAD interrogation. The SORI-CAD tandem mass spectrum of an abundant fragment ion species of *m/z* 986.6 is shown in Figure 3.3C. A sequence tag [VEE] can be found and the

parent ion of *m/z* 986.6 can be identified from ß-lactoglobulin B sequence as $y_{126}b_{43}$, which is shown in boldface in protein sequence in Figure 3.3C's inset with the sequence tag underlined. This is direct evidence of the identities of high-energy MSAD fragments to be internal fragments. Most of the fragment ions of the $y_{126}b_{43}$ parent ion can be readily attributed to common y-, b- and a-type ions, along with internal fragments and ions resulting from loss of water or ammonia. MS^3 experiments from MSAD/SORI-CAD have a comparable S/N level to MS^2 experiments from SORI-CAD alone. This is probably because the dissociation and fragment ion collection efficiencies in the MSAD experiment are very high.

Identification of large intact proteins (>100kDa) has been a challenge, as large proteins can neither be easily measured in mass nor be dissociated to give informative tandem mass spectra. We have shown the extensive dissociation of ß-galatosidase with high energy MSAD (Figure 3.4A). Once large proteins are dissociated into small peptide fragments, these ions are no different from those of small proteins. Thus we conducted MSAD/SORI-CAD experiment on the ß-galatosidase MSAD fragment ion of *m/z* 1042.6 with identical MSAD and SORI-CAD conditions as ones for ß-lactoglobulin B. Two sequence tags from both ends of this fragment are identified (Figure 3.4B), verifying that this fragment ion is a $y_{232}b_{801}$ ion. Although only two-residues long, the two sequence tags along their parent peptide's mass provide enough information to identify this protein from SWISS-PROT database. Thus MSAD/SORI-CAD presents a unique way to characterizing large proteins.

Table 3.1 summarizes the information on the sequence tags and their parent MSAD fragments from different proteins. Most of those MSAD fragments are small singly-charged y/b ions with high abundance in MSAD tandem mass spectra. They are often the fragmentation products of cleavage next to P or D. The sequence tags generally arise from the cleavage of two or three peptide bonds next to the peptide termini. For all of these sequence tag measurements, a 4-second accumulation time and -7-v dc offset voltage were used for MSAD, and 3.7 v excitation voltage was used for SORI-CAD. SORI-CAD tandem mass spectra were interpreted and sequenced manually as described in experimental section. High confidence sequence tag determination can be achieved from the accurate mass measurement of FT-ICRMS, the simple isotopic envelope, the small size of parent ion, and the sparse fragment ions in the tandem mass spectra. In addition to peaks that contribute to the sequence tag, there are other, less informative peaks coming from loss of water, ammonia, a-type ions, or other internal fragmentation. A fraction of MSAD fragments do not generate sequence tags because of limited fragmentation of parent ion due to their small size or amino acid sequences. Because the standardized MSAD/CAD conditions yield a rich diversity of fragment ions, many MSAD ions can be evaluated with reasonable effort and time. In this study, SORI-CAD was conducted only on a few major MSAD fragments. In our initial survey, we were unable to obtain sequence tags from major MSAD fragments of lysozyme and carbonic dehydrogenase; however, there may be other less abundant MSAD fragment ions from these proteins that could yield such information.

Table 3.1: Summary of the proteins examined with MSAD/SORI-CAD.

Protein[a]	MW	Sequence tag	Parent ion	Identity	Sequence
Ubiquitin	8564	X_2 (I/L)(K/Q)(K/Q) E X_n	699.8^{3+}	y18	D.**YN<u>IQKE</u>STLHLVLRLRGG**
Myoglobin	16,951	(I/L) H V (I/L) H X_2	784.5^{1+}	y42u118	I.**<u>IHVLH</u>SK**.H
Lactoglobulin	18,281	X_n V E E	986.6^{1+}	y126b43	A.**PLRVY<u>VEE</u>**.L
Serum albumin	66,433	(I/L) P (K/Q)(I/L)(K/Q) P D	792.5^{1+}	y472b118	D.**<u>LPKLKPD</u>**.P
Galatosidase	116,351	D (I/L) X_n R (I/L)	1042.5^{1+}	y232b801	N.**<u>DI</u>GVSEAT<u>RI</u>**.D

a. Lysozyme (14kDa) and carbonic dehydrogenase (35kDa) were not included since we were unable to generate sequence tags from the major fragment ions of these two proteins.

b. The cleavage type at the C-terminus of this peptide is nonstandard and is denoted by "u" to indicate a fragment 31.0 Da less than that from b-ion type cleavage, possibly due to side chain cleavages.

There are three unique advantages for this MSAD/SORI-CAD method that make it most promising to be applied in a high-throughput manner for obtaining sequence tags. First, unlike other fragmentation methods, MSAD/SORI-CAD has no discrimination against large proteins or certain types of proteins. As demonstrated with several different representative proteins here, MSAD dissociates proteins regardless of their heterogeneity into uniform small peptides, which can then be fragmented to yield sequence tags. Secondly, generic experimental conditions for MSAD/SORI-CAD experiments work for most proteins. The automation of tandem mass spectra processing is also straightforward due to the high interpretability of the peptide tandem mass spectra and the continued development of computational tools for processing peptide CAD spectra. Third, MSAD of any protein yields a large number of MSAD fragments that can be surveyed by an additional step of SORI-CAD. This might provide a versatile method for investigation of sequence tag information from proteins.

The main disadvantage of this MSAD/SORI-CAD method is that the success of obtaining sequence tag from a given MSAD fragment is variable. This is probably because both MSAD and SORI-CAD are induced by collisions with gas molecules. The fragments that could survive MSAD are less likely to be fragmented easily by SORI-CAD again. A possible solution to this is to employ other dissociation methods such as IRMPD to dissociate MSAD fragments.

Simple Protein Mixture MSAD: Identification of Protein Components

Despite the inability of MSAD to isolate a parent ion for subsequent fragmentation, two methods were attempted to identify proteins from mixtures up to four components: inspection of the MSAD signature for each protein and MS^3 to obtain sequence tag information. Experimental approaches that provide multiplexed tandem mass spectrometry capabilities are beginning to appear (Masselon, 2000; Meng, 2001; Purvine, 2003) and provide the potential for high-throughput measurements of complex mixtures. Even without parent ion selection, MSAD can be employed for generating multiplexed tandem mass spectra for protein mixtures. The challenge for the application of MSAD in this case, much like the other multiplexed CAD techniques, is the interpretation of the complex fragmentation patterns. Because of the high degree of internal fragmentation for MSAD, it appears that the spectra interpretation for this method will be formidable, especially for *de novo* sequencing or correlation-related computational techniques to identify proteins. However, the abundant MSAD fragments are quite characteristic for each protein, and, at high resolution and high mass accuracy, could be used to generate a protein's MSAD signature. Thus a protein's presence in a mixture could be suggested by its high-resolution, accurate fragment ion mass MSAD signature. A protein mixture with four components at equal ratio was examined with high-energy MSAD. The tandem mass spectrum was then compared with the previously acquired MSAD tandem mass spectra of each of the proteins. Most of major fragments were confidently attributed to one and only one component protein due to the high mass accuracy and the uniqueness of their MSAD signature (Figure 3.5A). Thus, ubiquitin, apomyoglobin, and serum albumin can

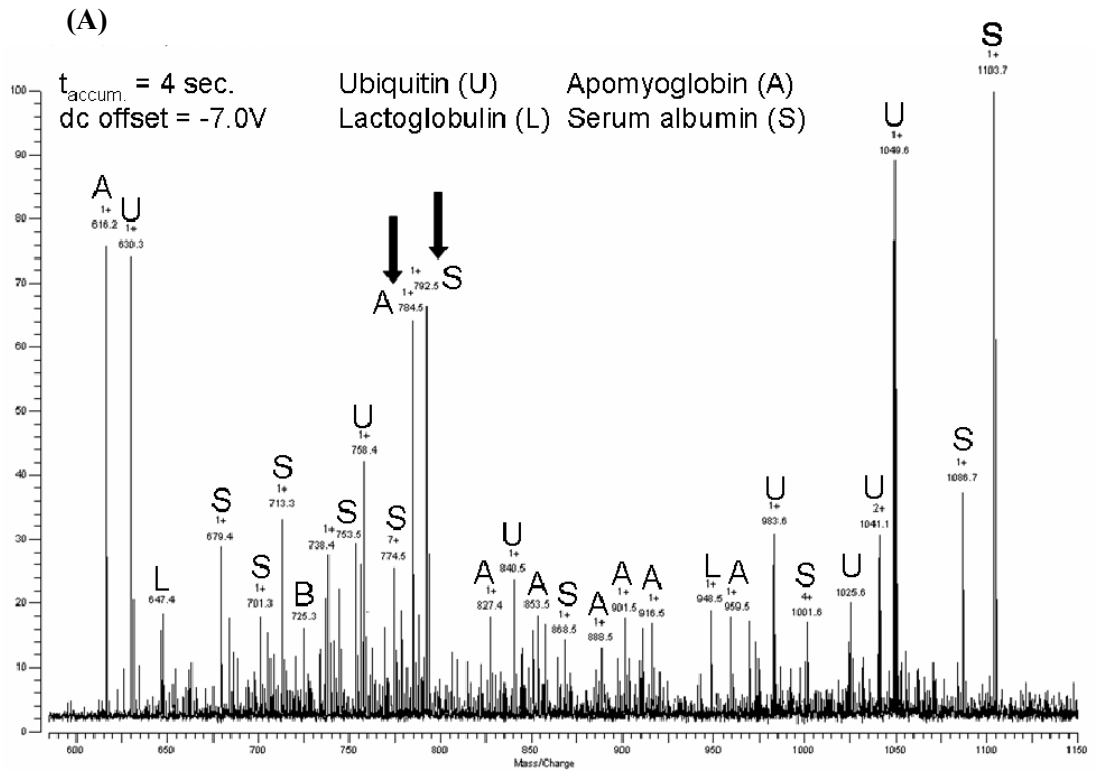


Figure 3.5: MSAD MS³ of a four-protein equimolar mixture. The mixture consists of ubiquitin, Apomyoglobin, Lactoglobulin and serum albumin (A). The origin of fragments were determined by matching with individual protein's MSAD fragments and labeled with protein name's initial in the spectrum. The MSAD fragments can be then used to generate sequence tag of the component proteins (B, C). Note that a pair of single-headed arrows pointing each other (→ ←) denotes that the mass differences between the parent ion mass and the sum of two fragment ion masses corresponds to an amino acid, which wraps sequence tag from y ion series to b ion series or vice versa. .

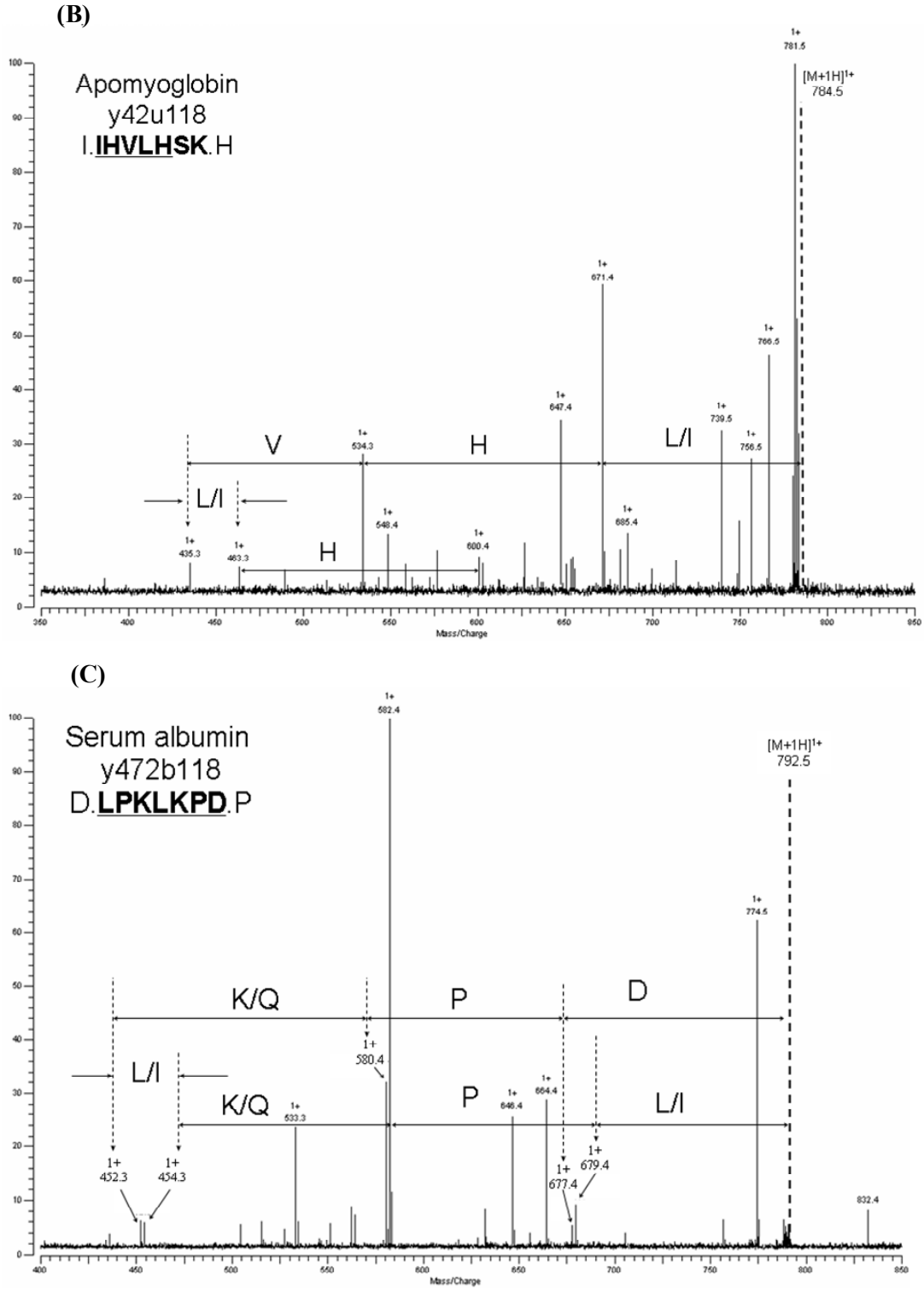


Figure 3.5: Continued.

be identified by their MSAD signature. However, lactoglobin's MSAD signature is not discernible in this mixture, probably due to the discrimination against lactoglubin by electrospray ionization and/or MSAD. In general, we estimate that the dynamic range for protein identification by MSAD in mixtures might range to at least 1:10. This should be tempered with the knowledge that ionization suppression in the ESI source may suppress the observed signal of some species, even before the MSAD process is conducted.

The second identification method attempted is sequence tagging of the component proteins by MSAD/SORI-CAD experiments. Similar to what enzymatic digestion does in solution, MSAD transforms a gas-phase protein mixture into a more complex gas-phase peptide mixture. Yet complex peptide mixtures are amenable for sequential examination of constituents with ion isolation. Peptides from MSAD fragmentation of 4-component mixture were surveyed (Figure 3.5A) and sequentially examined with SORI-CAD. Tandem mass spectra of fragment $y_{42}u_{118}$(*) from apomyoglobin and fragment $y_{472}b_{118}$ from serum albumin are shown in Figure 3.5B and 3.5C respectively, both of which yield sequence tags for the originating proteins. (* the cleavage type at the C-terminus of this peptide is non-standard and is denoted by "u" to indicate a fragment 31.0 Da less than that from b-ion type cleavage possibly due to side-chain cleavages). For comparison we attempted direct SORI-CAD on the intact proteins in an effort to obtain sequence tag information. However we were unable to successfully dissociate large proteins such as serum albumin to give informative fragment ions, much less sequence tag information.

The results obtained above indicate that proteins in mixture can be identified by these two methods. Identification of a protein by its MSAD signature is fast and parallel, but

requires prior information on the protein's MSAD signature, which would be suitable for prompt detection of targeted proteins. This can be done in a selected reaction monitor (SRM) mode in an LC-MS experiment. However, identification by sequence tagging would be suitable for singling out proteins from a sequence database in the application such as top-down proteomics. Exploiting MSAD's capability of dissociating large proteins enables both methods to identify large proteins from mixture in a robust and standardized manner.

CONCLUSIONS

MSAD is a new in-source fragmentation method, initially attributed to hexapole rf-coupling induced by extended accumulation time. In this report, we provide evidence that dc offset voltage of the hexapole can be used to induced fragmentation even at short accumulation times, thus providing a method of conducting MSAD on more rapid timescales, such as those compatible with on-line chromatography FTICRMS. The data obtained with the dc offset voltage suggested an alternate dissociation mechanism based on ion kinetic energy excitation due to the increased potential difference between skimmer and hexapole.

A variety of proteins were examined with MSAD at a range of collisional energies measured by accumulation time and dc offset voltage. While low-energy MSAD yields y and b ions similar to SORI-CAD, high-energy MSAD can fragment proteins up to 116kDa into small singly charged ions, which are characteristic to the protein and can be

referred to as its MSAD signature. This MSAD signature was used as an identification method to identify a protein's presence in a mixture.

The protein fragments from high-energy MSAD can be dissociated further by SORI-CAD (MSAD/SORI-CAD), analogous to enzymatic digestion followed by CAD. From such MS^3 spectra, sequence tags were obtained for five proteins whose sizes range from 8kDa to 116kDa. This sequence tagging technique was extended to proteins in a mixture, showing the potential of being applied in top-down proteomics. Its advantages include standardized experimental conditions, readily interpretable peptide fragmentation mass spectra, and applicability to large proteins.

Chapter 4

Graph-theoretical Approach to Separation of y- and b- Ions in High Resolution Tandem Mass Spectra

All of the data presented below has been published as

B. Yan, C. Pan, V.N. Olman, R.L. Hettich and Y. Xu. A graph-theoretic approach for the separation of b and y ions in tandem mass spectra. *Bioinformatics* 2005 21(5):563-574

As co-first author, C. Pan's primary contributions include problem formulation, graph theoretical algorithm development, and FT-ICR data acquisition. B. Yan is responsible for dynamic programming and simulation results (data not shown).

INTRODUCTION

Tandem mass spectrometry (MS/MS) has become a dominant technique for proteomics due to its ability to identify peptides in a high-throughput manner (Aebersold, 2003). In a typical liquid chromatography (LC)/MS/MS experiment, a protein mixture of interest is digested with proteases, and the resulting peptides are separated by one- or multi-dimensional LC. When eluted from the LC column, peptides are transported into the gas phase as positively charged ions using electrospray ionization and then introduced into a mass spectrometer. After measuring the mass/charge ratios (*m/z*) of all ions, MS can precisely isolate each peptide by its *m/z* and fragment the peptide through collisional-activated dissociation (CAD). The resultant fragments from this peptide are then measured. This process involves two sequential mass spectrometric measurements for a peptide (the full scan and the MS/MS scan), thus called tandem mass spectrometry

(MS/MS). Modern mass spectrometers can acquire thousands of high-resolution MS/MS spectra per day. Interpretation of such high-throughput mass spectral data in a reliable and efficient manner represents a highly challenging computational problem.

MS/MS spectra are informative about the composition and the order of amino acids in a peptide sequence, as it can reveal the molecular masses of the peptide's fragment products. Several bonds along the backbone of a peptide can be broken with the gas-phase collisional process. If the charge is retained on the N-terminal fragment, the ion is classified as a, b or c, and if the charge is retained on the C-terminal, the ion is classified as x, y or z. It has been observed that in a typical MS/MS spectral dataset, the majority of the N- and C-terminal ions are b and y ions, respectively, and each of these ion types contains the derivatives of neutral loss of water or ammonia (Dancik, 1999; Tabb, 2003). Note that the sum of any two complementary ion masses should be equal to the mass of the parent ion and the mass difference between two consecutive ions of the same type is exactly the mass of an amino acid residue.

There are two popular approaches to interpret tandem mass spectra data for protein identification: the 'database search' and '*de novo* sequencing' methods. The database search method compares experimental tandem spectra with theoretical tandem mass spectra of each peptide derived from a protein sequence database, such as Swiss-Prot (Boeckmann, 2003) and reports the best match or matches (Eng, 1994; Mann, 1994; Clauser, 1999; Perkins, 1999; Fenyo, 2000) assuming that the query peptides exist in the protein sequence database. This approach is highly effective and has been used

successfully in several proteomics projects on organisms with well-studied genomes (Washburn, 2001; Ho, 2002; Lasonder, 2002; Andersen, 2003). However, it is not applicable when a target sequence is not present in the protein database. This can happen for a number of reasons, including novel proteins, protein mutations, post-translational modifications (PTMs), and protein sequence database errors. Since the database search method usually generates high-false-positive recognition rates, it remains an open problem as how to validate database search results (Nesvizhskii, 2004).

De novo sequencing methods attempt to derive a protein sequence directly from tandem mass spectra (Taylor, 1997; Dancik, 1999; Pevzner, 2000; Chen, 2001b; Taylor, 2001; Lu, 2003; Ma, 2003). Theoretically, a full-length peptide sequence could be derived from an ideal tandem mass spectrum by computing the differences in mass between adjacent fragment ions of the same ion type in a complete series of fragment ions. The *de novo* methodology generally employs a graph-theoretic approach in which the tandem mass spectrum is represented as a graph. Each spectral peak is represented as a vertex, and a pair of spectral peaks that differ precisely by the mass of one amino acid is represented as an edge, which is referred to as type-1 edge in this study. The partial sequence of a target peptide is predicted by finding one or a set of longest directed paths in the spectrum graph. In this method, no special attempt was made to distinguish ion types and only the information on type-1 edges was used. However as shown in Figure 4.1, type-1 edges could be created to connect different types of ions by mistake. That is, although the difference in mass between any two ions of the same type is always equal to the total mass of an amino acid, the converse is not necessarily true. Since a tandem mass

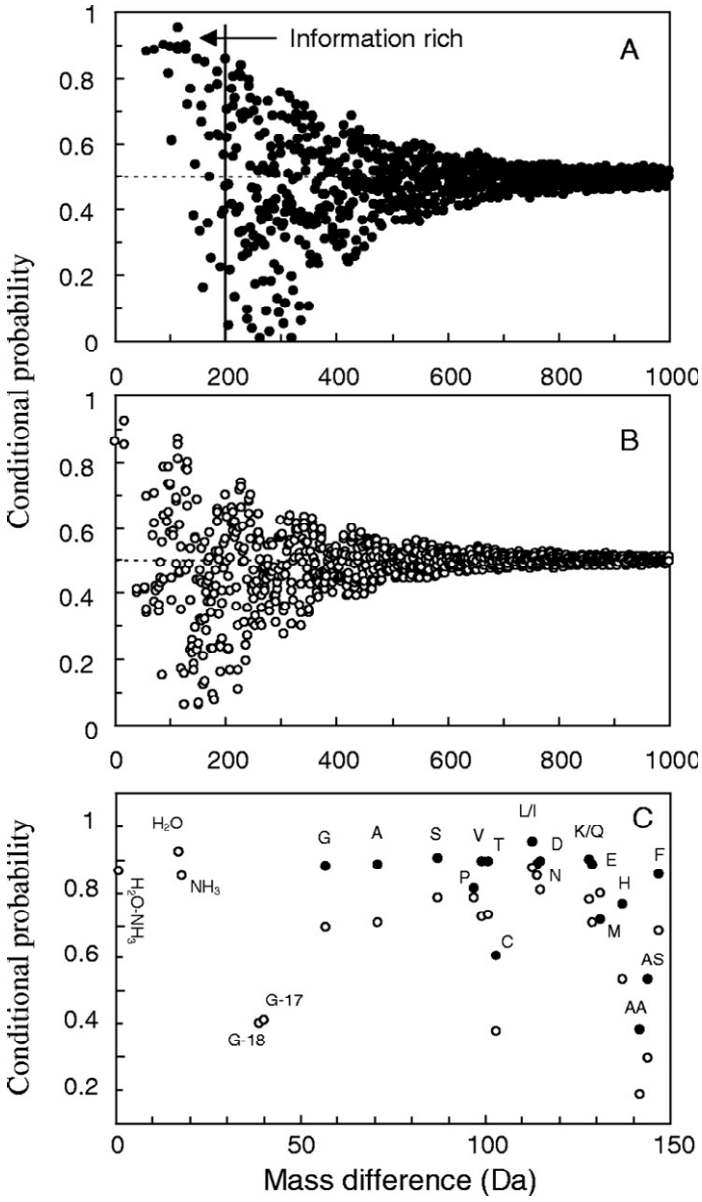


Figure 4.1: Conditional probability profile showing two ions being of the same type at a given mass difference. The values were derived from the statistical analysis of the simulated tandem mass spectra of all tryptic peptides with a mass range of [800 Da, 4000 Da], digested from proteins in Yeast genome. (A) Only b and y ions were considered. (B) b, y ions and their loss of water or ammonia were considered. (C) The combination of (A) and (B) in the information-rich zone. Only the points of interest were plotted and labeled.

spectrum generally mixes up various ions of different types, such a kind of misconnection decreases the accuracy of *de novo* sequencing. Furthermore, currently available *de novo* sequencing programs are computationally intensive and require high-quality MS/MS data. Owing to these difficulties, the *de novo* sequencing approach has not widely been used. In this study, we solve the problem of ion-type identification separately from the problem of *de novo* sequencing. We developed a novel graph-theoretic approach to the identification of ion types in a set of high-quality MS/MS data. Since the majority of MS/MS spectral peaks are either b or y ions [e.g. as observed in ion trap instruments (Dancik, 1999; Tabb, 2003)], our algorithm attempted to separate a set of MS/MS peaks into: (1) b ions and their variants (i.e. loss of water or ammonia), (2) y ions and their variants, and (3) the other ion types. We used a spectrum graph to represent a given set of spectral peaks in a similar fashion to some of the previous works (Taylor, 1997; Dancik, 1999; Pevzner, 2000; Chen, 2001b; Taylor, 2001; Lu, 2003; Ma, 2003). The main difference in our graph representation is that we considered two types of edges, one representing the connection between a pair of peaks suspected to be of the same ion type (type-1 edge) and the other representing the connection between a pair of peaks suspected to be of different ion types (type-2 edge). This is based on the observations that the mass difference between any two ions of the same type must be equal to the combination of some amino acids; and if the mass difference is not equal to the mass of any amino acid, it must arise from different type ions. Edge weights were assigned based on the estimated probabilities of whether the edges truly connect ions of the same or different types of ions. We formulated the ion-type identification problem as a graph partition problem, which is to partition the graph into three subgraphs, B, Y and U, respectively, to maximize the total weight of type-1

edges while minimizing the total weight of type-2 edges within each subgroup. This problem has been rigorously and efficiently solved using a dynamic programming algorithm, with a running time of $O(\sum_{i=1}^{L} 3^{|S_{i-1}|+|S_i|})$ in the worst case, where i is the distance from the root in a breadth-first tree (BFT) of the spectrum graph, L is the depth of the BFT and $|S_i|$ is the number of vertices on the i-th level of the BFT. For a spectrum graph, $|S_i|$ is generally small in value. We have systematically tested our algorithm on 114,851 simulated tandem mass spectra data derived from tryptic digested peptides of proteins in the Yeast genome, and have achieved an accuracy level of >90% for the separation of b and y ions. The tests on 19 sets of high-quality experimental Fourier transform ion cyclotron resonance (FT-ICR) tandem mass spectra indicate an average accuracy of 88%. On a typical dataset with 40 spectral peaks, our identification program PRIME (PaRtition of Ion types in tandem Mass spEctra) generally finds the globally optimal partition within 1 second, on a Dell Workstation with Pentium 4 (2.1 GHz).

MATERIALS AND METHODS

Spectrum graph representation.

Let $S = \{s_1, s_2, \cdots s_k\}$ be a set of tandem mass spectral data with k peaks $s_j = \{M_j, I_j\}$, where M_j and I_j denote the neutral mass and intensity of the peak s_j. Throughout this Chapter, we assume that ion masses are already derived from their *m/z* ratios based on the analysis of the isotope peaks. We define the zero mass peak as $s_0 = \{0, I_{k+1}\}$ and parent

mass peak as $s_{k+1} = \{M, I_{k+1}\}$, where $I_{k+1} = \max\{I_j, 1 \le j \le k\}$ and M is the neutral mass of the parent peptide. Theoretically, each ion has a complementary ion (e.g. the complementary ion of a b ion is a y ion) in the same spectrum. That is for any ion with a mass X, there should be an ion with a mass Y such that $X + Y = M$. In an experimental spectrum dataset, some ions may have their complementary ions missing because of various reasons. In such cases, we add their complementary ions back. That is for any j, $1 \le j \le k$, if there is no peak $\{M_i, I_i\}$ with $M_i + M_j = M$, we add a peak $\{M - M_j, I_j\}$ to the spectrum. Note that the masses of expanded spectrum S are distributed symmetrically around half of the parent mass.

We shall proceed to examine the mass differences between pairs of spectra peaks. First, we note that the mass difference between two ions of the same type (b or y or others) is always equal to the combination of some amino acids, whereas if the mass difference is not equal to the combination of any amino acid, it must be from different type ions. Our algorithm separates ions into different ion types primarily based on this property. However, two ions with a mass difference of the combined mass of some amino acids do not necessarily belong to the same ion type. To estimate the probability of a given mass difference to be from two ions of the same type, we conducted a simulation of tandem mass spectrometry experiment *in silico* using tryptic digested peptides from proteins derived from the Yeast genome in Swiss-Prot database (version of March 14, 2004). Each peptide was theoretically fragmented into b, y ions and their chemical variants (i.e. loss of water for the first present residue S, T, E, D or loss of ammonia for the first present

residue R, K, Q, N). The mass differences between the ions of the same type and between ions of different types were calculated and tabulated (we treated b ions and their variants as the same group and so for y ions and their variants). The conditional probability that a given mass difference δ is from two ions of the same type was then estimated by the ratio of the counts of δ between two ions of the same type to the total occurrences of δ. The results are shown in Figure 4.1. Apparently, mass differences that do not correspond to the total mass of some amino acids must arise from ions of different types, while mass differences arising from one or two amino acids are highly probable to come from ions of the same type (see the information rich zone, 0–200 Da in Figure 4.1).

We used the following procedure to construct the spectrum graph. Each peak of tandem mass spectrum is represented as a vertex. To capture two distinct relationships between two vertices, two kinds of edges, type-1 and type-2 edges are considered. A pair of peaks is connected by a type-1 edge if their mass difference is the same as the mass of a single amino acid (suggesting that they are most probably of the same ion type), or by a type-2 edge if their mass difference is not equal to the combination of any amino acids (indicating that they definitely belong to different ion types). In the interest of reducing the complexity of the graph, we currently use the following more conservative definition that keeps the number of type-2 edges small. A pair of peaks is connected by a type-2 edge only if their mass difference is ≤35 Da (which is smaller than the mass of any amino acid) and not equal to 1, 17 or 18 Da (which correspond to masses of H_2O–NH_3, water loss and ammonia loss, respectively). We named this threshold as δ_2. To make the

calculations more efficient, if a vertex had only type-2 edges, we discarded all these type-2 edges. We call this graph $G = (V, E)$ a spectrum graph, with V and E being the vertex and edge set, respectively.

Each type-1 edge has an assigned weight, representing the possibility that the two peaks involved are of the same ion type. The weight of a type-1 edge $E(V_m, V_n)$ is defined as follows:

$$W_1(E) = \ln(I_m \cdot I_n) + \ln[\Pr(\delta_{mn})] - \alpha \cdot |\, m(aa_i) - \delta_{mn} \,| + \ln(F_i), \quad 1 \le i \le 20 \ , \tag{4.1}$$

where I_m, I_n in (0, 1] are the relative intensities of ions m and n; Pr (δ_{mn}) is the conditional probability that ions m and n are of the same type, given the mass difference $\delta_{mn} = |M_n - M_m|$; aa_i is an amino acid type that has the smallest $|m(aa_i) - \delta_{mn}|$ value and satisfies the condition $|m(aa_i) - \delta_{mn}| \le \delta_1$ (δ_1= 0.01 Da for simulated tandem mass spectra or 0.05 Da for experimental FT-ICR data), and $m(aa_i)$ is the mass of aa_i; F_i is the relative frequency of amino acid, aa_i, in the target genome. The scaling factor $\alpha > 0$ is determined empirically. This definition validates the observation that a good type-1 edge usually has a small mass deviance, connects two peaks with high intensities, and has a high probability to arise from the same type ions. Since the type-2 edges generally connect different type ions based on the above conservative definition, the weight of a type-2 edge $E(V_m, V_n)$ is simply defined as:

$$W_2(E) = \ln(I_m \cdot I_n) \tag{4.2}$$

Problem formulation.

If we imagine that type-1 edges carry attractive force and type-2 edges repulsive force, then the vertices of the same ion type in a spectrum graph should naturally cluster together, whereas vertices of different ion types repel away. Noise can be disconnected or attached by false type-1 edges to b or y ions. The separation of b and y ions can then be achieved by optimally cutting all the type-2 edges and false type-1 edges.

Let Ω be the set of all possible tri-partitions of vertices set V of a spectrum graph G. Without loss of generality, we assume that G is a connected graph; otherwise G will be an individual connected component. We define the scoring function *Score(P,G)* for any tri-partition $P = \{V_B, V_Y, V_U\} \in \Omega$ as:

$$\begin{aligned} Score(P,G) &= Q_1 \cdot [W_1(V_Y,V_Y) + W_1(V_B,V_B) - W_1(V_Y,V_B)] \\ &+ Q \cdot [W_2(V_Y,V_B) - W_2(V_Y,V_Y) - W_2(V_B,V_B)], \end{aligned} \tag{4.3}$$

where $W_i(A,B)$ represents the total weight of type-i edges between vertices of subsets A, $B \subseteq V$, Q_i is a positive factor, $i = 1, 2$. Our goal is to find the optimal partition $P^{opt} \in \Omega$ such that

$$Score(P^{opt},G) = \max\{Score(P,G) \mid P \in \Omega\}, \tag{4.4}$$

Algorithm.

We now present a dynamic programming algorithm for solving the optimization problem defined above. For a carefully chosen vertex $v_0 \in V$ (see below), we construct a BFT(v_0)

of the spectral graph G, with v_0 being the root (Cormen *et al*., 2001). Let S_i be a set of vertices at the i-th level of BFT(v_0) (hence, their distance to the root is i), where i = 0, 1, 2, ..., L and L is the length of the longest path. We have $S_0 = \{v_0\}$ and $V = U_{i=0}^{L} S_i$. We define a set of subgraphs $G_i = (V_i, E_i)$ such that $V_i = U_{j=0}^{i} S_j$ consists of edges connecting vertices of V_i, i = 0, 1, 2, ...,L. Note that in BFT(v_0), there is no edge between V_i and S_j for any $j > i$+1. The definition is illustrated in Figure 4.2. For each partition P_i of vertices of S_i, let Ω/P_i be a subset of Ω satisfying P_i. We define the conditional optimal partition of vertices V_i of G_i, $P^{opt}(P_i) \in \Omega/P_i$, as

$$Score[P^{opt}(P_i), G_i] = \max\{Score(P, G_i), P \in \Omega / P_i\}, \tag{4.5}$$

The following theorem relates the conditional optimal partition of the whole graph with the conditional optimal partitions of its subgraphs. It enables us to divide the problem into smaller segments, to tackle one by one, without compromising the global optimality of the final solution.

THEOREM: For any subgraph G_i of G and any partition $P_i \in \Omega_i$ where Ω_i is a set of all possible partitions of S_i, the conditional optimal partition of vertices V_i of G_i can be decomposed into a conditional optimal partition of vertices V_{i-1} of G_{i-1} and a particular partition of S_i, i = 1, 2, ...,L:

$$Score[P^{opt}(P_i), G_i] = \max\{Score[P^{opt}(P_{i-1}), G_{i-1}] + Score(P_{i-1} \otimes P_i, g_i), P_{i-1} \in \Omega_{i-1}\}, \tag{4.6}$$

where g_i is a subgraph of G formed by a vertex set $S_i \cup S_{i-1}$ connected by edges $E_i - E_{i-1}$, and $P_{i-1} \otimes P_i$ presents the virtual union of P_{i-1} and P_i (Figure 4.2).

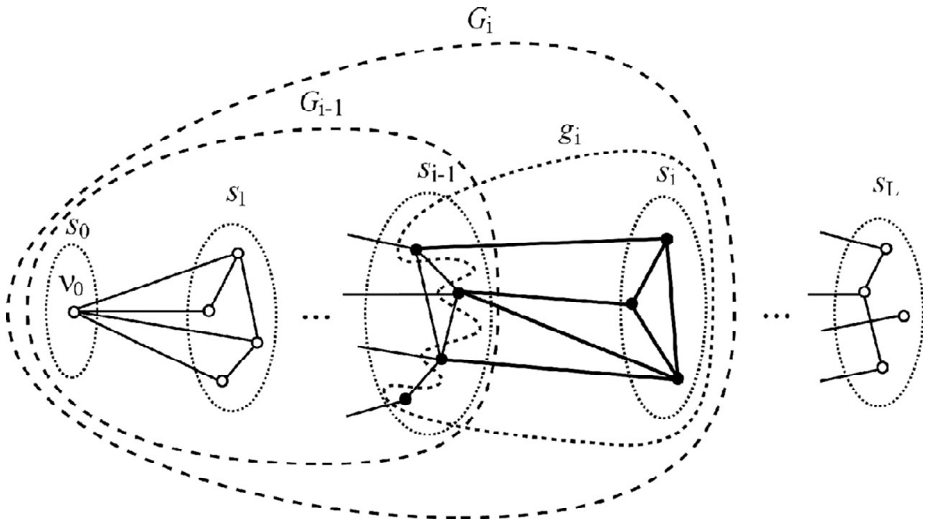


Figure 4.2: Scheme of the graph partition algorithm. A breadth-first search tree is constructed from the spectrum graph. A dynamic programming algorithm is then used to identify the optimal partition of vertices.

PROOF: Because the scoring function (Equation 4.3) is additive in terms of the weights of edges, and subgraphs G_{i-1} and g_i do not have any common edges based on their definitions, we always have:

$$Score(P, G_i) = Score(P, G_{i-1}) + Score(P, g_i), \tag{4.7}$$

By combining Equations 4.5 and 4.7 with the decomposition $\Omega / P_i = U\, \Omega/(P_{i-1} \otimes P_i)$, $\Omega / P_i = U\, \Omega/(P_{i-1} \otimes P_i)$, $P_{i-1} \in \Omega_{i-1}$, we have:

$$\begin{aligned} & Score[P^{opt}(P_i), G_i] \\ & = \max\{Score(P, G_i), P \in \Omega / P_i\} \\ & = \max\{Score(P, G_{i-1}) + Score(P, g_i), P \in \Omega / P_i\} \\ & = \max\{Score(P, G_{i-1}) + Score(P, g_i), P \in U\{\Omega / P_{i-1} \otimes P_i, P_{i-1} \in \Omega_{i-1}\}\} \end{aligned} \tag{4.8}$$

However, (P, G_{i-1}) is independent of P_i, and Score(P, g_i) depends only on P_{i-1} and P_i. Thus, we have

$$\begin{aligned} & \max\{Score(P, G_{i-1}) + Score(P_{i-1} \otimes P_i, g_i), P \in \Omega / P_{i-1}, P_{i-1} \in \Omega_{i-1}\} \\ & = \max\{Score(P^{opt}(P_{i-1}), G_{i-1}) + Score(P_{i-1} \otimes P_i, g_i), P_{i-1} \in \Omega_{i-1}\} \end{aligned}, \tag{4.9}$$

This completes the proof of the theorem.

Implementation.

The following pseudo-code describes the dynamic programming algorithm for solving the optimal partition problem defined by Equation 4.6.

1. Select v_0 from G based on a specific rule (see below);
2. Build a $BFT(v_0)$ of G;

3. **For** each partition $P_0 \in \Omega_0$ of S_0 **Do** Score($P^{opt}(P_0)$,G_0)←0;

4. **For** i = 1 **To** L **Step** 1 **Do**

5. **For** each tri-partition $P_i \in \Omega_i$ of S_i **Do**

6. max_score←0;

7. **For** each tri-partition $P_{i-1} \in \Omega_{i-1}$ of S_{i-1} **Do**

8. **If** max_score>*Score* < ($P^{opt}(P_{i-1})$, G_{i-1}) + ($P_{i-1} \otimes P_i$,g_i) **Do**

9. max_score←($P^{opt}(P_{i-1})$, G_{i-1})+ ($P_{i-1} \otimes P_i$, g_i);

10. $P^{opt}(P_i) \leftarrow P^{opt}(P_{i-1}) \otimes P_i$;

11. Select the partition $P^{opt}(P_L)$ with the maximal ($P^{opt}(P_L)$,G_L) value as the final optimal partition of G;

12. Determine the ion types of the partitioned vertices as follows.

The dynamic programming algorithm classifies all the spectrum peaks into three classes, B, Y and U. These three groups are really two large groups of ions of the same type (i.e. B or Y), plus a smaller set for the remaining ions (i.e. U). In the algorithm, we did not specifically use properties that are directly associated with b ions or y ions. Therefore, B may actually be y ions and Y may be b ions. To further determine whether B set contains b ions or y ions (the same for Y set), we need additional information. We used two properties of tandem mass spectra to decide which group contains the b ion set or the y ion set. (1) The b ion group should include a vertex with a mass of 0 and a vertex with a mass of peptide parent mass -18 Da (i.e. [M-H_2O]), while the y ion group should include a vertex with a mass of 18 Da (the complementary ion of [M-H_2O]) and a vertex with parent mass. (2) Statistical analysis of tandem mass spectra has shown that the average

intensity of y ions is typically more than twice as much as that of b ions, although the numbers of ions are almost the same (Dancik, 1999; Tabb, 2003). Based on such information, PRIME reports the ion type of each ion as output.

Computational complexity.

Let $C(G)$ be the computational complexity for calculating the optimal partition of a spectrum graph G defined by Equation 4.6 and define $C(G)$ as the number of function (P,G) calls. The computational complexity of our dynamic programming algorithm can be derived from the lines 4, 5 and 7 of the pseudo-code directly,

$$C(G) \le O(\sum_{i=1}^{L} 3^{|S_{i-1}|+|S_i|}), \tag{4.10}$$

where i is the distance from the root v_0 of the (v_0), L is the length of the longest path of the (v_0) and $|S_i|$ is the number of vertices on the i-th level of (v_0). To make $C(G)$ as small as possible, we exhaustively searched through all vertices to find the root v_0 that gave the smallest argument of O in (10) before starting the partition procedure.

Generalized algorithm—considering chemical variants.

Neutral losses from fragment ions are common in tandem mass spectra. A loss of water or ammonia will reduce fragment ion masses by 18 or 17 Da. In addition, protein PTMs, a common and important phenomenon in cell functioning, also change the patterns of mass spectra by shifting a portion of peaks by specific masses. To detect and deal with such

variants, we introduced a new type of pseudo amino acid with the specific mass for each suspected chemical variant, and treated the resulting ions the same way as we did with b or y ions. For example, for water loss, we added a pseudo amino acid, namely 'water', with a mass of –18.011 Da to the amino acid library. Since it is difficult to estimate the frequency of each possible pseudo amino acid occurring in tandem mass spectra in advance, we used a simplified Equation 4.1 to calculate the weight of type-1 edge involving these new residue types,

$$W_1(E) = \ln(I_m \cdot I_n) - \alpha \cdot | m(aa_i) - \delta_{mn} |, \qquad (4.11)$$

By doing so, virtually no change is needed in our partition algorithm for dealing with chemical variants. In addition, by introducing a pseudo amino acid for each suspected PTM, we prevented the exhaustive combinatorial search for all possible mass modifications, in which even a small set of modification types will lead to a large combinatorial problem (Yates *et al*., 1995). Finally, the generalized algorithm can deal with the special case where the neutral loss ions of a spectrum are so prominent, and the base b or y ions are sufficiently small that only the neutral loss ions survive the spectral pre-processing procedure. Since we treated b ions and their variants as the same group and dealt with them independently (similar approach for y ions and their variants), the absence of base b or y ions does not cause major problems.

RESULTS

Our algorithm was implemented as a computer program PRIME using C++ programming language. PRIME was tested on 19 sets of high-quality experimental FT-ICR-MS data

from synthetic peptides as well as tryptic peptides obtained from digestion of horse myoglobin, bovine serum albumin and lysozyme. All mass spectra were acquired with an IonSpec (Lake Forest, CA) 9.4-Tesla HiRes electrospray ESI-FTICR-MS. Ions were generated with a low-flow rate dynamic electrospray source (flow rate of ~400–500 nl/min), accumulated in an external hexapole, transferred into the high-vacuum region with a quadrupole lens system, and then detected in the cylindrical analyzer cell of the mass spectrometer. Because ion detection was achieved in an ultra-low vacuum regime (~2 x 10^{-10} Torr), broadband mass resolutions of about 200,000 (full width of peak at half maximum) at *m/z* 1,000 were possible. Calibration was accomplished by using standard proteins (such as ubiquitin and myoglobin), and provided mass accuracies within a few milli-mass units. Low-energy fragmentation via ion collisional dissociation was accomplished with an FT-ICR-MS instrument. This method of collisional fragmentation was conducted by isolating an ion of interest (either a peptide or a protein) within the analyzer cell of the mass spectrometer, and then accelerating the ion into a nitrogen target gas under sustained off-resonance irradiation collision-activated dissociation (SORI-CAD). Energetic collisions between the accelerated ion and the target gas generated ionic fragment ions and could be measured at high resolution.

Data preprocessing

For each precursor mass, the raw profile data of its corresponding mass spectrum was loaded using the manufacturer's software IonSpec99. The molecular mass/intensity list was then tabulated using the View function and saved into a text file (noise was filtered

with the default threshold). The resulting file was fed to an in-house software for further isotopic peak reduction. We found that a few isotopic peaks were not identified and consequently their charges were not determined correctly by IonSpec99. This problem was fixed by re-analyzing isotopic peaks carefully, i.e. checking the consecutive peaks with uniform intervals (1.000 Da for +1, 0.500 Da for +2 and 0.333 Da for +3 and so on, with the mass tolerance of 0.005 Da). Such isotopic peaks were removed and their intensities were added to the monoisotopic peaks. This pre-processing resulted in a number of ions in each spectrum which varied from 22 to 50, covering 25–90% of hypothetical b and y ions. PRIME was then used to determine the ion type of each individual peak. The thresholds δ_1=0.05 Da and δ_2=15 Da were used to build type-1 edges and type-2 edges, respectively. In most cases, the calculations were done within 1 second on a Dell workstation with Pentium 4 (2.1 GHz). The prediction results are summarized in Table 4.1.

Overall performance.

As shown in Table 4.1 for each of the 19 test spectra, our partition program identified most of the b and y ions and their chemical variants (i.e. loss of water or ammonia) correctly. Among the 19 datasets, 7 were identified with 100% accuracy. The partition accuracy for each individual spectrum ranged from 57 to 100%, with an average accuracy of 88%. In the 'sequence' column of Table 4.1 the b and y ions that were correctly differentiated were labeled with underlines (b ions) and overlines (y ions), respectively. We observed that the experimental tandem mass spectra with good coverage of b and y

Table 4.1: The test results on 19 sets of experimental FT-ICR tandem mass spectra

No.	Sequence[a]	Source	b ions[b]	y ions[b]	CPU[c]
1	VEADIAGHGQEVLIR	Horse myoglobin	18/18	15/15	0.81
2	HGTVVLTALGGILK	Horse myoglobin	7/7	10/10	0.02
3	HGTVVLTALGGILKK	Horse myoglobin	10/10	6/6	0.03
4	YLEFISDAIIHVLHSKHPGDFGADAQGAMTK[d]	Horse myoglobin	4/5	15/16	0.02
5	VEADIAGHGQEVLIR[d]	Horse myoglobin	5/5	13/13	0.02
6	KGHHEAELKPLAQSHATK[d]	Horse myoglobin	6/6	2/4	0.02
7	LFTGHPETLEK	Horse myoglobin	2/7	6/7	0.01
8	Acetyl-LVFFAEDVGSNK[e]	Synthesis	6/6	6/6	0.02
9	GKAKVTGRWK	Synthesis	11/13	2/3	0.04
10	DAFLGSFLYEYSR	BSA	17/17	11/11	0.03
11	LVNELTEFAK	BSA	2/4	7/7	0.02
12	TVMENFVAFVDK	BSA	7/7	3/6	0.03
13	LGEYGFQNALIVR	BSA	6/7	7/7	0.02
14	GLVLIAFSQYLQQCPFDEHVK[d]	BSA	4/5	5/6	0.01
15	HLVDEPQNLIKQNCDQFEK[d]	BSA	2/3	5/6	0.01
16	SLHTLFGDELCK	BSA	5/5	6/7	0.02
17	GYSLGNWVCAAK	Lysozyme	7/7	7/7	0.04
18	KIVSDGNGMNAWVAWR	Lysozyme	3/3	7/11	0.04
19	NLCNIPCSALLSSDITASVNCAK	Lysozyme	4/4	10/12	0.10

[a] b, y ions correctly identified were labeled with underline (b ions) or overline (y ions), respectively.

[b] The numerator indicates the number of the correctly identified b or y ions and their chemical variants, i.e. loss of water or ammonia, while the denominator represents that of observed in experimental spectrum.

[c] The unit of CPU time measured in seconds.

[d] The precursor ions of these peptides were triply charged, while the rest were doubly charged.

[e] The synthesized peptide LVFFAEDVGSNK was acetylated.

ions always had high identification accuracy (e.g. the peptide with 100% accuracy), while those with poor fragmentation resulted in relatively lower partition accuracy. We also observed that most induced sequences covered only a portion of the target peptide sequence. However, those accurately determined subsequences with enough length might be appropriate to perform homology-based sequence database search through homology search tools like PSI-BLAST (Altschul, 1997; Taylor, 1997).

The part A and part C of Figure 4.3 show the FT-ICR tandem mass spectrum of two peptides and their partition results. Several interesting results can be observed from Figure 4.3. (1) All the b, y ions and their variants (loss of water here is denoted by X) were identified correctly. (2) The full-length peptide sequences except for the first two residues can be readily derived from either b ion series or y ion series (no distinction between the amino acids L and I). The first di-peptide was assigned to tryptophan, since neither b1 nor y12 was observed in the spectrum and coincidently the mass of Trp is equal to the sum of Asp and Ala. It was also clearly shown that adding back of the complementary ions (denoted by open symbols) greatly improved the completeness of spectra for *de novo* sequencing. (3) There existed a second continuous mass ladder (572.27, 719.34, 832.43, 995.50, 1124.55, 1287.63, 1374.67 and 1530.74) that apparently corresponds to b-H_2O ion series starting from position 6 of the peptide. This evidence strongly indicates that Ser-6 lost water during fragmentation. This kind of secondary mass ladder information should be highly useful in future applications for detecting the types and sites of protein PTMs, since X could be any other specified neutral mass losses of PTMs, for example, –98 Da for loss of H_3PO_4 from phosphopeptides. (4) Three false

Figure 4.3: Partition of two FT-ICR tandem mass spectra. The b ions (red circles) and y ions (blue squares) were partitioned into two subgraphs, where vertices were connected through type-1 edges (thin red or blue lines) within each subgraph and through type-2 edges (thick black lines) between the two subgraphs. Thin dashed lines represent the false type-1 edges while the thick dashed lines denote the discarded type-2 edges. Noises were labeled with diamonds. The closed symbols represent the ions observed in experimental spectrum while the open symbols denote the added complementary ions.

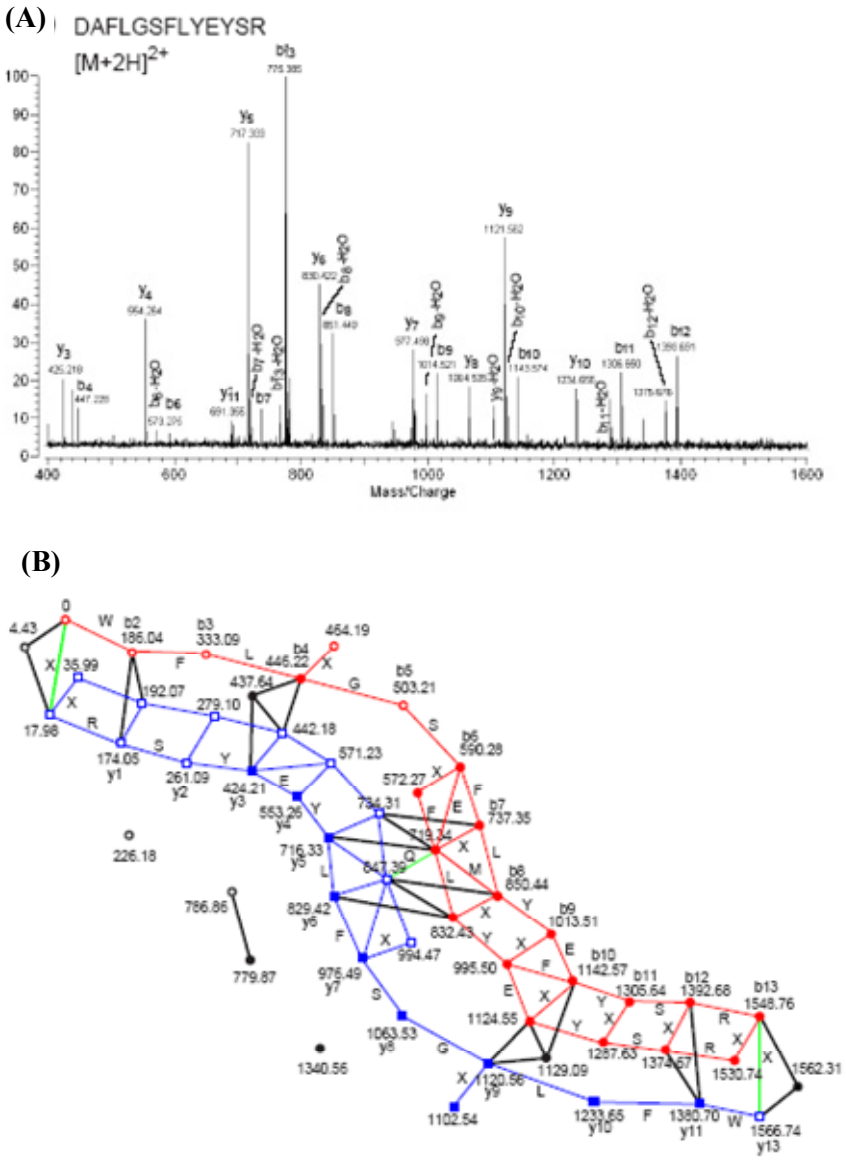


Figure 4.3: Continued.

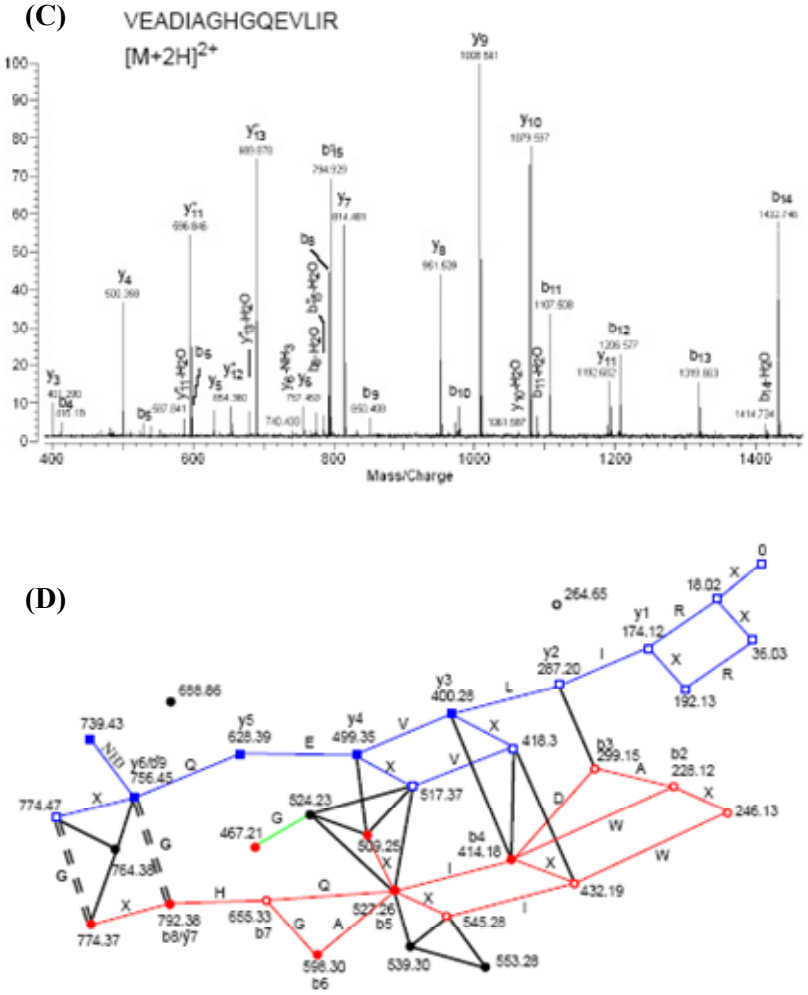


Figure 4.3: Continued.

type-1 edges (labeled with thick teal lines) were formed between b and y ions. Our algorithm recognized all of them correctly based on the global optimization.

DISCUSSION

Methods for accurate identification of ion types provided the basis for many mass spectrometry data interpretation problems, including (1) *de novo* sequencing, (2) identification of PTMs, and (3) validation of database search results. Compared to previous *de novo* sequencing methods (Taylor, 1997; Dancik, 1999; Pevzner, 2000; Chen, 2001b; Taylor, 2001; Lu, 2003; Ma, 2003) the uniqueness of our approach is that we treat the problem of ion-type identification separately from the problem of *de novo* sequencing. By decoupling the two problems, we arrived at a conceptually clearer framework for solving the two problems separately rather than having them tangled together.

In addition, among these published papers on *de novo* sequencing algorithms and applications (Taylor, 1997; Dancik, 1999; Pevzner, 2000; Chen, 2001b; Taylor, 2001; Lu, 2003; Ma, 2003) only the information similar to what we call type-1 edges was utilized. As shown in Figure 4.1, false type-1 edges could arise from different types of ions by accident. To recognize this kind of false connections and to partition ions into correct classes, we introduced a new type of connection, the type-2 edges, which connect two peaks of possibly different ion types; and include the probability that a given mass difference between two ions corresponds to an amino acid into the objective function. Since we use a very conservative definition in the construction of type-2 edges, the

probability of having a false type-2 edge is intrinsically much lower than that of having a false type-1 edge. These facts imply that our algorithm does utilize more informational context of a given MS/MS experimental data and might describe the real spectrum more completely. The systematic tests of our approach on 114,851 simulated tandem mass spectra derived from Yeast genome and on the 19 sets of experimental FT-ICR data showed that the type-2 edges were powerful. The inclusion of a few sparse type-2 edges can make the partition correctly.

Protein PTMs generate tremendous diversity, complexity and heterogeneity of gene products, and are involved in many important cell functioning and regulation processes (Gooley, 1997). To date, there are over 340 entries reported in Delta Mass (http://www.abrf.org/index.cfm/dm.home?AvgMass=all), a database of protein PTMs. Therefore, the verification of PTMs raises a major challenging problem after the human genome was completely sequenced (Mann, 2003). The capacity for detecting the types and sites of PTMs efficiently should be a key feature of a good *de novo* sequencing algorithm. We have shown that our algorithm can be easily extended to consider chemical variants, i.e. loss of water or ammonia, or PTMs, by introducing a new type of pseudo amino acid with the specified mass for each suspected chemical variants, without increasing the computational complexity.

Several *de novo* sequencing algorithms and software packages, such as Lutefisk (Taylor, 1997; Taylor, 2001) and PEAKS (Ma, 2003) were reported to perform *de novo* sequencing successfully, and the latter also employed a dynamic programming algorithm.

However as we discussed above, the basis of these approaches are quite different from ours: no special attempt was made to distinguish ion types, and solely the information of type-1 edges was used to construct the spectrum graph representation. Since our effort in this study has been to identify the ion type of each individual peak (a first stage of *de novo* sequencing), no comparison was made to evaluate the performance of our algorithm with others. However, the tests on the simulated and experimental tandem mass spectra showed that PRIME achieved an accuracy of ~90% for the separation of b and y ions. Once the ions are partitioned into correct classes, the *de novo* sequencing should be easily derivable from one single ion series. Work on extending this algorithm to deal with the *de novo* sequencing problems by using more robust data and exploring its potential applications in detecting PTMs is on-going and will be presented in the future publications.

CONCLUSIONS

Ion-type identification is a fundamental problem in computational proteomics. Methods for accurate identification of ion types provide the basis for many mass spectrometry data interpretation problems, including (a) *de novo* sequencing, (b) identification of post-translational modifications and mutations, and (c) validation of database search results.

Here we present a novel graph-theoretic approach to solve the problem of separating b ions from y ions in a set of tandem mass spectra. We represent each spectral peak as a node and consider two types of edges: type-1 edge connecting two peaks probably of the

same ion types and type-2 edge connecting two peaks probably of different ion types. The problem of ion-separation is formulated and solved as a graph partition problem, which is to partition the graph into three subgraphs, representing b, y and others ions, respectively, through maximizing the total weight of type-1 edges while minimizing the total weight of type-2 edges within each partitioned subgraph. We have developed a dynamic programming algorithm for rigorously solving this graph partition problem and implemented it as a computer program PRIME (PaRtition of Ion types in tandem Mass spEctra).

Chapter 5

Integration of Nanoscale Liquid Chromatography with Fourier Transform Ion Cyclotron Resonance Mass Spectrometer

Part of the data presented below has been published as

M.B. Strader, D.L. Tabb, W.J. Hervey, C. Pan, and G.B. Hurst. Efficient and Specific Trypsin Digestion of Microgram to Nanogram Quantities of Proteins in Organic-Aqueous Solvent Systems. *Analytical Chemistry* 2006, 78, 125 -134,

C. Pan's contributions to this manuscript include all LC-FT-MS measurements, nanospray ionization optimization, and associated data analysis.

INTRODUCTION

FT-ICR mass spectrometers combine high mass accuracy, superb resolution, and excellent dynamic range with versatile tandem mass spectrometry capability (Marshall, 1998). In the previous chapters, we have demonstrated novel uses of FT-ICR with standard samples in direct infusion mode. However, MS measurements in direct infusion mode are not sufficient for analysis of a typical sample from shotgun proteomics, because of at least the following reasons.

First, many peptides in a typical proteome digest cannot be completely resolved by their mass-to-charge ratios. It is common for a bacterial proteome to contain more than 3000 proteins, which are transformed into greater than 30,000 peptides by proteolysis in a shotgun proteomics measurement (Hochstrasser, 2002). Most of the tryptic peptides have

a mass-to-charge ratio between 400 and 3000 and a charge state between +1 and +3. In such a dense mass spectrum, it is difficult to completely resolve all the peptides from each other, even using a high-resolution mass spectrometer.

Second, the dynamic range of a typical proteome digest is far larger than the dynamic range of FT-ICR. The proteins in a proteome exist in vastly different abundance. The concentration difference between the high-abundance proteins and the low-abundance proteins can reach 10^6 (Corthals, 2000); whereas the dynamic range of FT-ICR is ~ 10,000. The orders of magnitude of difference between the desired dynamic range and the attainable dynamic range in FT-ICR would result in the selected detection of only the most abundant proteins in a proteome.

Third, electrospray of a large diversity of peptides in a single sample leads to the limited ionization of only a subset of peptides. In the electrospray ionization process, peptides have to compete for a fixed number of positive charges. The characteristics of a peptide, including proton affinity, hydrophobicity, size, *etc.*, determine how well the peptide can be ionized (Avery, 2003). The ionization of a peptide can be partially or even completely suppressed in the presence of many other competing peptides. This ionization suppression effect leads to the selected detection of a few peptides that can be well ionized from a complex peptide mixture.

To address these challenges, mass spectrometry has been coupled with liquid chromatography as an integrated analytical platform (Link, 1999; Washburn, 2001). The

peptides are separated in time with liquid chromatography, and the LC eluent is directly electrosprayed into mass spectrometers. The mass spectrometer becomes an online detector of the LC by measuring the *m/z* and ion intensity of the co-eluting peptides. Since LC retention time and *m/z* value are two orthogonal properties of peptides, the LC-MS system multiplies the peak capacity of mass spectrometer with the peak capacity of liquid chromatograph. Therefore, the complex peptide mixture can be well resolved with LC-MS. The dynamic range limitation of mass spectrometers and ionization suppression of the electrospray are also mitigated, because different sets of peptides are eluted off sequentially, and, at a given retention time, only a small set of peptides are electrosprayed and analyzed by the mass spectrometer.

In this chapter, we describe the development of an LC-MS system integrating reverse phase LC with FT-ICR MS. Our FT-ICR instrument is equipped with an electrospray source from Analytica of Branford Inc. This electrospray source was originally designed for microspray (flow rate between 1 μL/min and 10 μL/min) in the direct infusion mode, as demonstrated in the previous chapters. Here we have focused a significant amount of effort on incorporating nanospray (flow ratio between 50 nL/min and 300 nL/min) with the Analytica source, because of the following two advantages of nanospray ionization.

First, the electrospray of aqueous solvent is more stable at lower flow rates in nL/min range. As the LC eluent is directly electrosprayed into the mass spectrometer, the electrospray solvent changes from highly aqueous (95% H_2O and 5% ACN) to moderately organic (50% H_2O and 50% ACN) with the progression of the LC gradient.

Microspray often fails to form a stable electrospray plume with highly aqueous solvent, which compromises the MS measurement of peptides eluting off early in the LC gradient. Nanospray is able to more effectively ionize peptides throughout the LC gradient.

Second, nanospray can be coupled with the nanoscale LC to provide higher sensitivity than capillary LC. Nanoscale LC typically uses a column with inner diameter less than 100 μm. The flow rate is 100–300 nL/min for both LC elution and electrospray. Ultimately, the use of the nanoscale LC reduces the amount of proteome sample needed for an LC-MS measurement.

With the implementation of nanospray on our FT-ICR instrument, the high-performance LC-MS measurements were demonstrated with bacterial proteome digest samples. However, since our FT-ICR MS lacked the capability to perform data-dependent tandem mass spectrometry, we proposed to combine the accurate mass measurement of FT-ICR and the automated tandem mass spectrometry of a three-dimensional quadrupole ion trap (QIT) to achieve confident identification of peptides. By using an identical LC setup for LC-FT-MS and LC-QIT-MS, the peptides measured in the two LC-MS platforms were correlated computationally by their retention time. Here, we present development of the LC-FT-MS platform and the computational tools for data integration.

MATERIALS AND METHODS

Materials.

All proteins, salts, buffers, dithiothreitol (DTT), guanidine hydrochloride, trifluoroacetic acid, diethyl pyrocarbonate, phenylmethanesulfonyl fluoride (PMSF), sucrose, and RNase-free DNase I were obtained from Sigma Chemical Co. (St. Louis, MO). RNase Away was obtained from Molecular BioProducts (San Diego, CA). Sequencing-grade trypsin was purchased from Promega (Madison, WI). Formic acid was obtained from EM Science (affiliate of Merck KgaA, Darmstadt, Germany). HPLC grade acetonitrile and water were used for all LC-MS analyses (Burdick & Jackson, Muskegon, MI). Ultrapure 18-MΩ water obtained from a Millipore Milli-Q system (Bedford, MA) was used for sample buffers. Fused-silica capillary tubing was purchased from Polymicro Technologies (Phoenix, AZ). BCA assay reagent and standards were obtained from Pierce Chemical Co. (Rockford, IL).

Construction of Protein Standard Mixture.

Protein standard mixtures were generated using six proteins: bovine serum albumin (MW 69 kDa), yeast alcohol dehydrogenase I (MW 37 kDa), bovine carbonic anhydrase II (MW 29 kDa), horse myoglobin (MW 17 kDa), bovine hemoglobin (MW 15 kDa), and chicken egg lysozyme C (MW 14 kDa). Hemoglobin includes α- and β- polypeptides, and the isomer yeast alcohol dehydrogenase II was found to be a component of yeast alcohol

dehydrogenase I, giving a total of eight polypeptides in the mixture. Mixtures contained equal concentration of each protein. The proteins were dissolved in 50 mM Tris-HCl/10 mM $CaCl_2$ (pH 7.6) and then combined in equal concentration.

Proteolytic Digestion.

For the trypsin digestion study, the protein mixture digestion was performed using two different protocols for comparison: (A) overnight digestion in aqueous buffer; and (B) 1-hour digestion in 80% acetonitrile buffer. In both digestion experiments, 200 ng of trypsin were added to the 1-μg sample to yield an enzyme-to-substrate ratio (w/w) of 1:5, and the digestion was performed at 37°C with shaking. After digestion, all peptide samples were treated with DTT (20 mM) for 1 h at 37°C to reduce disulfide bonds. The reduced peptides were lyophilized and resuspended in 100 μL of 95% H_2O/5% acetonitrile/0.1% formic acid. To further inhibit trypsin activity, 2 μL of 10% formic acid was added to each resuspended sample.

For the nanoLC-FT-MS platform development, the protein standard mixture sample and an *Rhodopseudomonas palustris* proteome sample were digested using the manufacturer's protocol for the trypsin used in these experiments, which included denaturation in 6 M guanidine hydrochloride/50 mM Tris-HCl/10 mM $CaCl_2$ (pH 7.6) for 45 min followed by dilution to 0.5 M guanidine hydrochloride and overnight digestion with 200 ng of trypsin at 37°C. The samples were then treated with 20 mM DTT for 1 hour at 60°C as a final reduction step. The resultant peptides from this control digestion were desalted using

solid-phase extraction (C_{18} Zip-Tip, Millipore, Billerica, MA) and solvent exchanged into 0.1% formic acid in water by centrifugal evaporation.

1D capillary LC-FT-MS Analysis.

1D capillary LC-FT-MS experiments were performed with an Ultimate HPLC system (Dionex) coupled with a HiResESI Fourier transform ion cyclotron resonance mass spectrometer equipped with a 9.4-T magnet. Samples were separated with a Vydac C18 column (300 μm i.d. × 15 cm, 300-Å pore size, 5-μm particles) at a flow rate of 4 μL/min and directly introduced to the FTICR MS with an electrospray source (Analytica, Branford, CT).

1D nanoLC-FT-MS Analysis.

For nanoLC-FT-MS analysis, samples were separated using a nanobore Vydac C18 column (750 μm i.d. × 15 cm, 300-Å pore size, 5-μm particles). The LC operation was automated with a Famos/Switchos/Ultimate HPLC System (Dionex, Sunnyvale, CA) (Figure 5.1). The sample is automatically loaded into a 50-μL loop from a vial with the auto-sampler Famos and then pumped through a pre-concentration column (300 μm i.d. × 5 mm C18 PepMap) on the switching system Switchos. After the pre-concentration column is switched online with the analytical nanobore C18 column, the peptides are resolved and eluted at the flow rate of 200 nL/min. The LC eluent ran through a grounded metal union and electrosprayed into FTICR MS using a 10 μm i.d. emitter (PicoTip, New

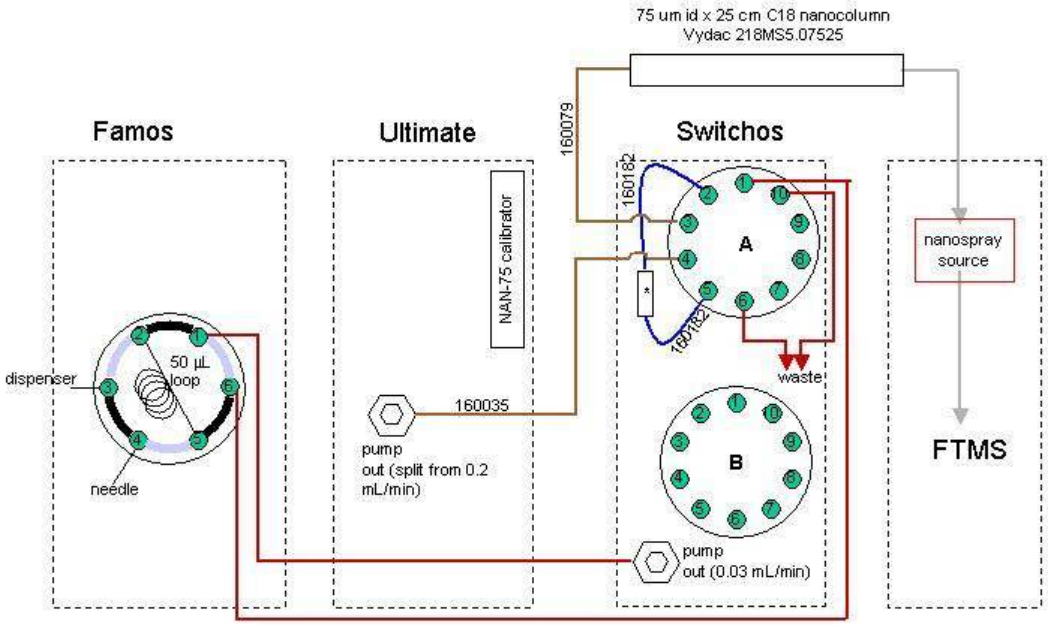


Figure 5.1: Famos/Switchos/Ultimate nanobore HPLC system. This system consists of an auto-sampler, Famos, a switching system, Switchos, and an HPLC pump, Ultimate. The sample is injected into the pre-concentration column on the Switchos and then resolved on the 25-cm nanobore column. (Figure courtesy of Dr. Gregory B. Hurst)

Objective, Woburn, MA). The emitter and the counter-electrode were 3 mm apart with 1800 volts potential difference. The FT-ICR MS measurement used 2 second hexapole ion accumulation, 256K data points analog-to-digital conversion at 1Mhz, and 2-scan signal averaging.

1D LC-QIT-MS/MS Analysis.

For all peptide samples, one-dimensional LC-QIT-MS/MS experiments were performed with a Famos/Switchos/Ultimate HPLC System coupled to an LCQ-DECA XP Plus quadrupole ion trap mass spectrometer (Thermo Finnigan, San Jose, CA) equipped with a nanospray source as previously described. The reverse-phase HPLC was run at a flow rate of 200 nL/min using a Vydac C18 column (75 μm i.d. × 15 cm, 300-Å pore size, 5-μm particles). For all 1D LC-MS/MS data acquisition, the LCQ was operated in the data-dependent mode with dynamic exclusion enabled (repeat count 2), where the four most abundant peaks in every MS scan were subjected to MS-MS analysis. Data-dependent LC-MS/MS was performed over a parent *m/z* range of 400-2000.

Protein Identification from QIT MS/MS Data Analysis.

The SEQUEST algorithm (Eng, 1994) was used to match experimental MS-MS spectra with their counterparts predicted from a protein sequence database. An unconstrained database search was employed so that peptides resulting from cleavage at residues other than lysine or arginine at one end (semitryptic peptides) or both ends (nontryptic) could

be identified. The sequence database used for searches in this study consisted of two major elements. The 4833 ORFs of the published *R. palustris* protein sequence database (Larimer, 2004) were search targets for the *R. palustris* proteome searches but acted as distractors (indicators of false positive identifications) during the protein standard mixture searches. Sequences for the eight proteins in the standard mixture were also included in the database; we added the sequence for alcohol dehydrogenase II, because this protein was observed as a component in the alcohol dehydrogenase I standard.

DTASelect was used to assemble, filter, and compare the identifications from SEQUEST searches on all data sets (Tabb, 2002). This software sorts peptide identifications by the proteins that contain them. A protein in the mixture was considered successfully identified if at least two component peptides passed DTASelect's default SEQUEST score cutoffs. Spectra from singly charged peptides were required to exceed 1.8 in the SEQUEST parameter XCorr, while XCorr values for doubly- and triply-charged peptides were required to exceed 2.5 and 3.5, respectively. The best matching sequence for each spectrum was required to have an XCorr at least 8% greater than the second best (DeltCN > 0.08).

Correlation of LC-QIT-MS/MS Data and LC-FT-MS Data.

The LC-FT-MS data were processed with the IonSpec FTdoc program. The processing steps include charge state determination, deconvolution of isotopic clusters into

monoisotopic peaks, and filtering of noise peaks. The monoisotopic masses, retention times, and intensities were exported to ACSII files.

The peptide data points from the LC-FT-MS measurement were correlated with the peptide identification from the LC-QIT-MS measurement. The retention time of the LC-QIT-MS measurement was normalized as follows. The confident peptide identifications from the QIT data were obtained by filtering the SEQUEST results with the conservative Xcorr cutoff as described above. These peptides were matched against all peptide data points measured in the FT-ICR data without retention time constraint. If the mass error between a QIT peptide identification and a FT-MS data point was less than 0.05 Da, this peptide was considered as positively correlated between the two measurements. A linear regression model was constructed between the two retention time points for all correlated data points: one from LC-QIT-MS measurement and the other one from LC-FT-MS measurement. The retention time for the peptide data points from LC-FT-MS was normalized with this linear function.

The LC-FT-MS measurement was then used to validate peptide identifications with lower Xcorr scores. Filtering of peptide identifications were performed with Xcorr cutoff (Xcorr (+1) > 1.3, (+2) > 2.0, (+3) > 3.0). A peptide identification is considered as being validated by the LC-FT-MS measurement if there is a peptide data point found with less than 0.05 Da mass measurement error and 3 minutes retention time shift. Each validated peptide identification was marked by a flag in the DTASelect result files. The data correlation and peptide identification validation were performed with Perl scripts.

RESULTS AND DISCUSSION

Comparison of the Proteolysis Results of Two Digestion Protocols with capillary LC-FT-MS

Protein complexes can be isolated with tandem affinity purification (TAP) in a high-throughput manner. The subunits of an isolated protein complex can then be identified with shotgun proteomics measurements. However, the isolation of protein complexes by TAP often yields samples with less than 100 ng of total protein. This complicates enzymatic digestion and peptide identification, as protein concentrations less than 10 ng/μL are unsuitable for efficient enzymatic digestion with the conventional protocols. In this study, a new digestion protocol was proposed that uses 80% acetonitrile in the digestion buffer. This 80% acetonitrile digestion was compared with conventional aqueous buffer for trypsin digestion. The digestion products with the two protocols were measured with capillary LC-FT-MS (Figure 5.2).

FT-ICR-MS allows isotopic resolution of multiply charged ions, allowing the determination of molecular mass to better than 10 ppm accuracy. By knowing the accurate molecular masses, we were able to characterize the dense patch of late-eluting unidentified ions in the LC-QIT-MS/MS results. Most of those ions are intact undigested proteins or partially digested protein fragments with charge states too high to be identified by SEQUEST. Figure 5.2 shows the LC-FTICR total ion chromatograms of the mixture digested overnight in aqueous buffer and digested for 1 h in 80% acetonitrile.

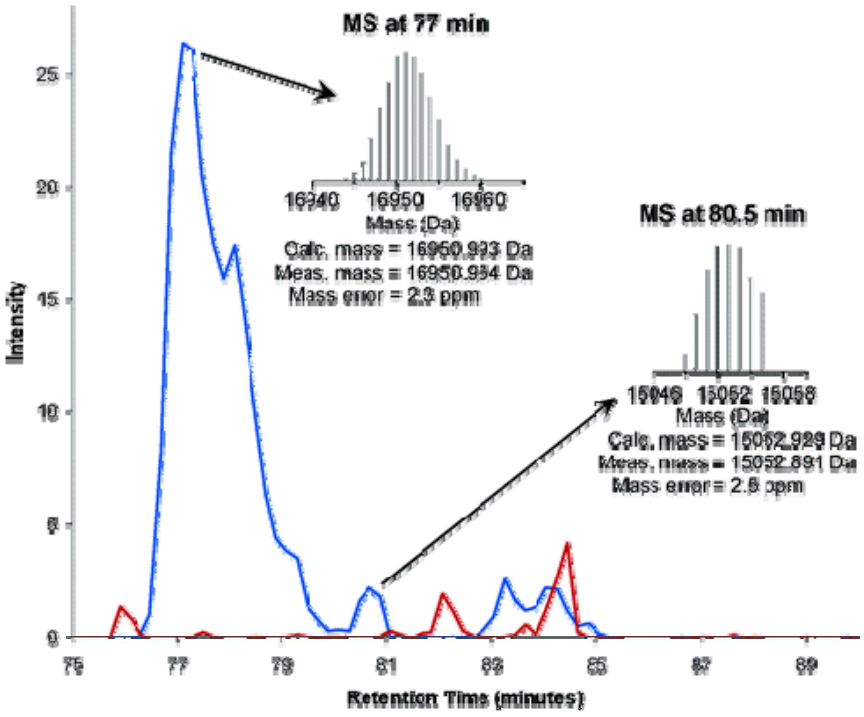


Figure 5.2: Capillary LC-FT-MS total ion chromatograms of the 1-*u*g mixture. The blue trace is from overnight digestion in aqueous buffer; the red trace is from a 1-h digestion in 80% acetonitrile. The insets illustrate the isotopic resolution of nominal masses representing undigested myoglobin and hemoglobin chain.

The C18 reverse-phase column, LC instrumentation, and gradients used for these experiments were identical to those used for the LC-QIT-MS/MS analyses. As illustrated in the insets in Figure 5.2, intact myoglobin (most abundant isotope mass [MAIM] measured 16,950.954 Da, calculated MAIM 16,950.993 Da, mass error 2.3 ppm) and hemoglobin α chain (measured MAIM 15,052.891 Da, calculated MAIM 15,052.929 Da, mass error 2.5 ppm) were identified in the 75–82-min retention time window for the overnight aqueous digestion. In addition, several unidentified species, probably partially digested protein fragments, in the 27–29 kDa range were observed (data not shown). We did not identify any intact proteins in the 75–82-min retention time window for the 80% acetonitrile digestion shown in Figure 5.2, and overall ion signal was lower in this retention time window. These data show that trypsin cuts more efficiently in the 80% acetonitrile digestion.

Development of nanoLC-FT-MS Platform for Shotgun Proteomics Measurement

The original ionization source on our FT-ICR instrument was designed for microspray at the flow rate between 1 μl/min to 10 μl/min. The electrospray emitter is a grounded metal needle with ~ 100 μm i.d. To perform nanospray, the metal needle was replaced by a commercial nanospray emitter – PicoTip from New Objective. Since the PicoTip is made of fused silica, it had to be connected to a metal union to provide electric contact for the nanospray. The main factors that can affect the ionization efficiency and stability of nanospray include:

- The i.d. of the nanospray emitter;

- The potential difference between the emitter and the counter-electrode; and
- The distance from the emitter to the counter-electrode.

We optimized the three factors with an aqueous solution of 1 μM ubiquitin in direct infusion mode. The combination of a 10 μm i.d. emitter, 1800 volt potential difference, and 3 mm distance was found to provide a stable and intense signal.

This nanospray setup was then connected to a Famos/Switchos/Ultimate HPLC System. This system fully automated sample injection and gradient elution for the HPLC analysis. During LC elution, the full scan mass spectra were continuously acquired at a rate of 6 scans per min with the FT-ICR instrument. The LC-FT-MS platform was first tested with a protein standard mixture digest (Figure 5.3). A linear gradient (shown as red lines in Figure 5.3A) was used for the reverse phase LC elution. The total ion chromatogram shows many resolved chromatographic peaks (Figure 5.3A). The peak width of these chromatographic peaks indicates high LC separation resolution. The MS measurement with FT-ICR gives accurate mass measurement of peptides. A full scan acquired at a retention time 40 minute is shown in Figure 5.3B. The isotopic envelope of individual peptides was resolved, which allowed the determination of the peptides' charge states. The high sensitivity and excellent dynamic range of FT-ICR enabled the detection of many low-abundance peptides. The low-abundance peptides measured in the 614–632 *m/z* window at this full scan are shown in the inset of Figure 5.3B.

The LC-FT-MS platform was then used to measure an *R. palustris* proteome sample (Figure 5.4). As opposed to the protein standard mixture that contains 6 proteins at equal

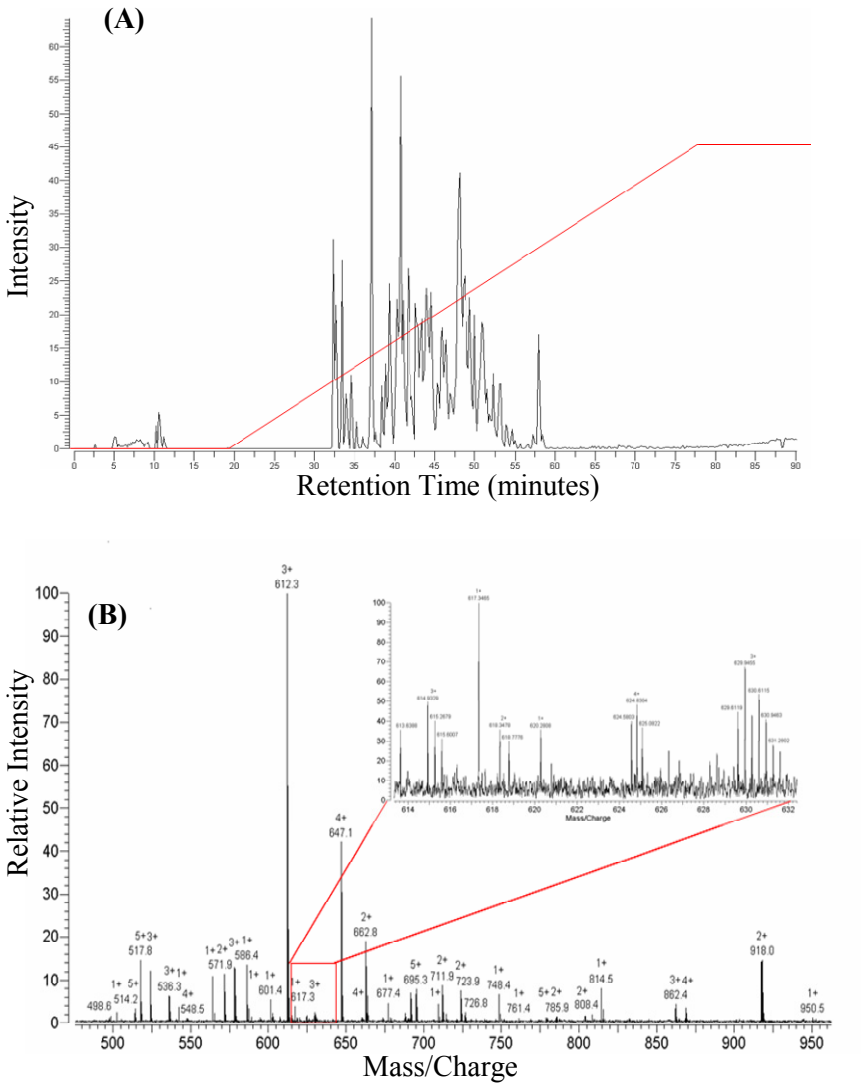


Figure 5.3: nanoLC-FT-MS measurement of a protein standard mixture digest. (A) Total ion chromatogram. The red line shows the LC gradient. (B) The full scan at the retention time 40 minute. The inset shows the zoom-in of an m/z region.

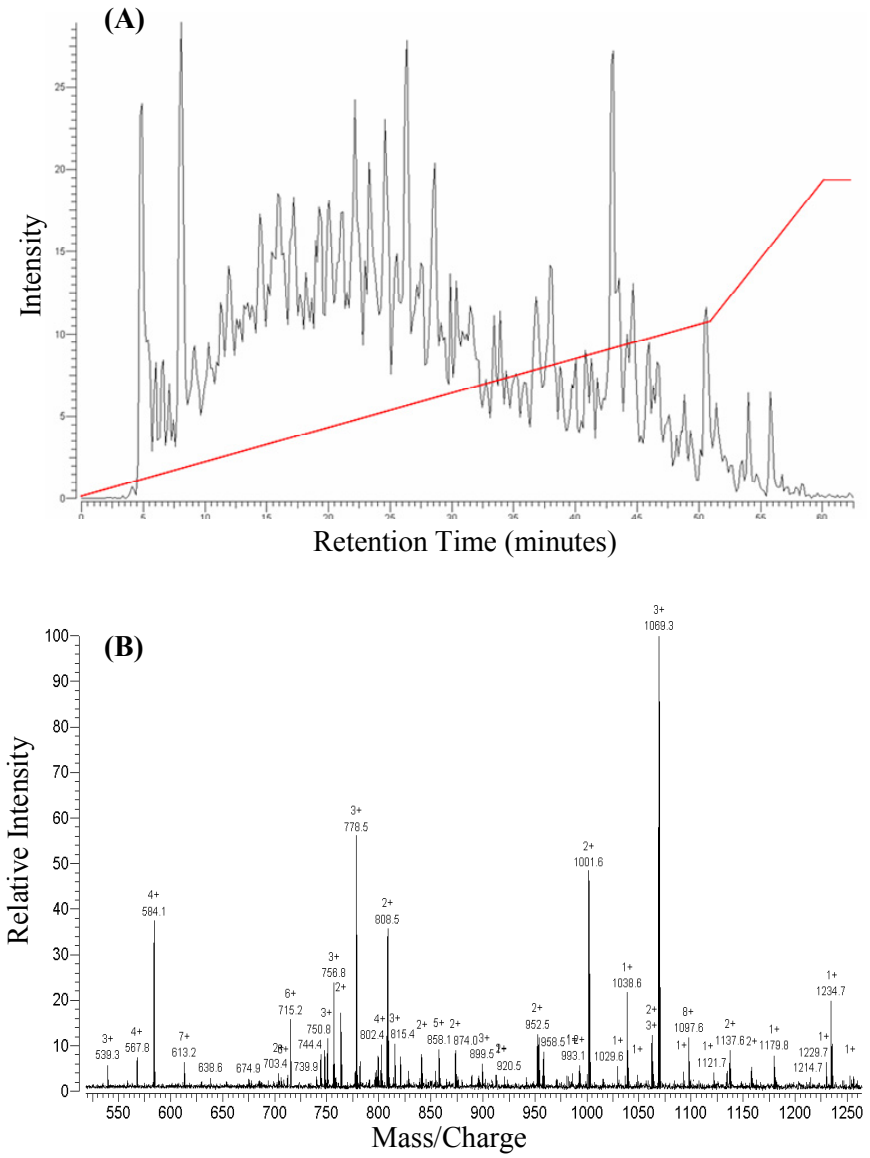


Figure 5.4: nanoLC-FT-MS measurement of an *R. palustris* proteome. (A) Total ion chromatogram. The red line shows the LC gradient. (B) The full scan at the retention time 21 minute.

abundance, the proteome sample contained thousands of proteins at vastly different abundance. Due to this daunting complexity, many peptides were eluted off the LC column at any retention time point and no individual chromatographic peak can be identified in the total ion chromatogram. A full scan at the retention time of 21 minute is shown in Figure 5.4B, which yielded the accurate mass measurement for many peptides at different intensities. Although trypsin digestion is expected to generate peptides with one, two or three positive charges, many peptides with charge state higher than three were observed. These peptides are generally ignored in the data analysis of LC-QIT- MS measurement, as all peptides are assumed to be singly, doubly or triply charged for MS/MS database searching. As the charge states of peptides were experimentally determined in the FT-ICR MS measurement, the mass spectra showing mass-to-charge ratios of peptides can be deconvoluted into the mass spectra showing the neutral mass of those peptides. The deconvolution can greatly simplify the data analysis procedure for LC-FT-MS measurement.

Integration of LC-FT-MS with LC-QIT-MS/MS for Confident Peptide Identification

Shotgun proteomics generally uses the MS/MS data from LC-QIT-MS/MS measurement for peptide identification. The best match between an MS/MS scan and a peptide sequence is scored and, if the score exceeds a threshold, the peptide is considered being identified from this MS/MS scan. However, many true peptide identifications can score lower than the threshold and, conversely, many spurious peptide identifications can score

higher than the threshold. Here, we propose to use the accurate mass measurement from LC-FT-MS to improve both the false positive rate and the false negative rate of peptide identification in shotgun proteomics by measuring the protein standard mixture digest with both LC-FT-MS and LC-QIT-MS/MS. The peptide identifications were filtered with a reduced threshold, which reduced the false negative rate at the expense of increasing the false positive rate. Then, the peptide identifications were filtered with the LC-FT-MS data to remove the false peptide identifications that passed the threshold.

To filter the peptide identifications with LC-FT-MS data, we first normalized the retention time of LC-FT-MS measurement (Figure 5.5). 5496 ion species were observed in the LC-FT-MS experiment, which were displayed in a 2-dimensional ion map with their retention time and measured monoisotopic mass (Figure 5.5A). 301 peptides were identified by the LC-QIT-MS/MS experiment, which were also displayed in a 2-dimensional ion map with their retention time and monoisotopic mass calculated from the peptides' sequence (Figure 5.5B). Data points in the two ion maps were matched if their monoisotopic masses have less than 0.05 Da difference. All matched data points were plotted into a scatter-plot with the two observed retention times: one from the LC-FT-MS experiment and the other from the LC-QIT-MS/MS experiment (Figure 5.5C). A linear model was fitted into the data points for retention time normalization:

$$RT_{QIT} = 0.9669 \cdot RT_{FT} - 11.5872 ,$$

where RT_{QIT} is the retention time of a peptide in the LC-QIT-MS/MS experiment and RT_{FT} is the retention time of this peptide in the LC-FT-MS experiment. The intercept of

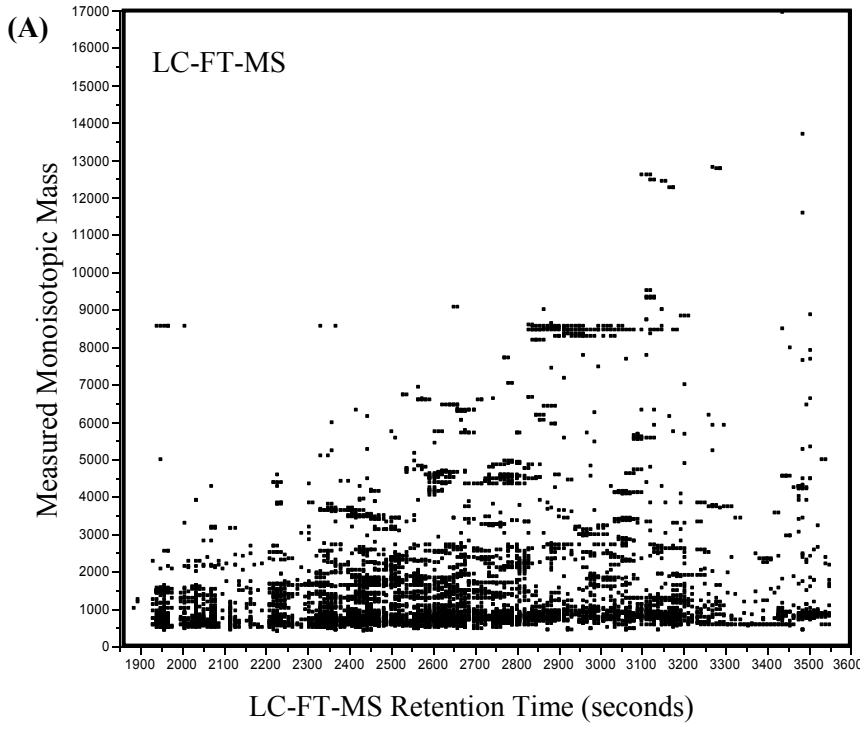


Figure 5.5: Integration of LC-FT-MS data and LC-QIT-MS/MS data from the protein standard mixture digest. (A) Ion map from LC-FT-MS measurement. (B) Ion map from LC-QIT-MS/MS measurement. The monoisotopic masses were calculated from peptide sequences. (C) Correlation of retention times between LC-QIT-MS/MS and LC-FT-MS

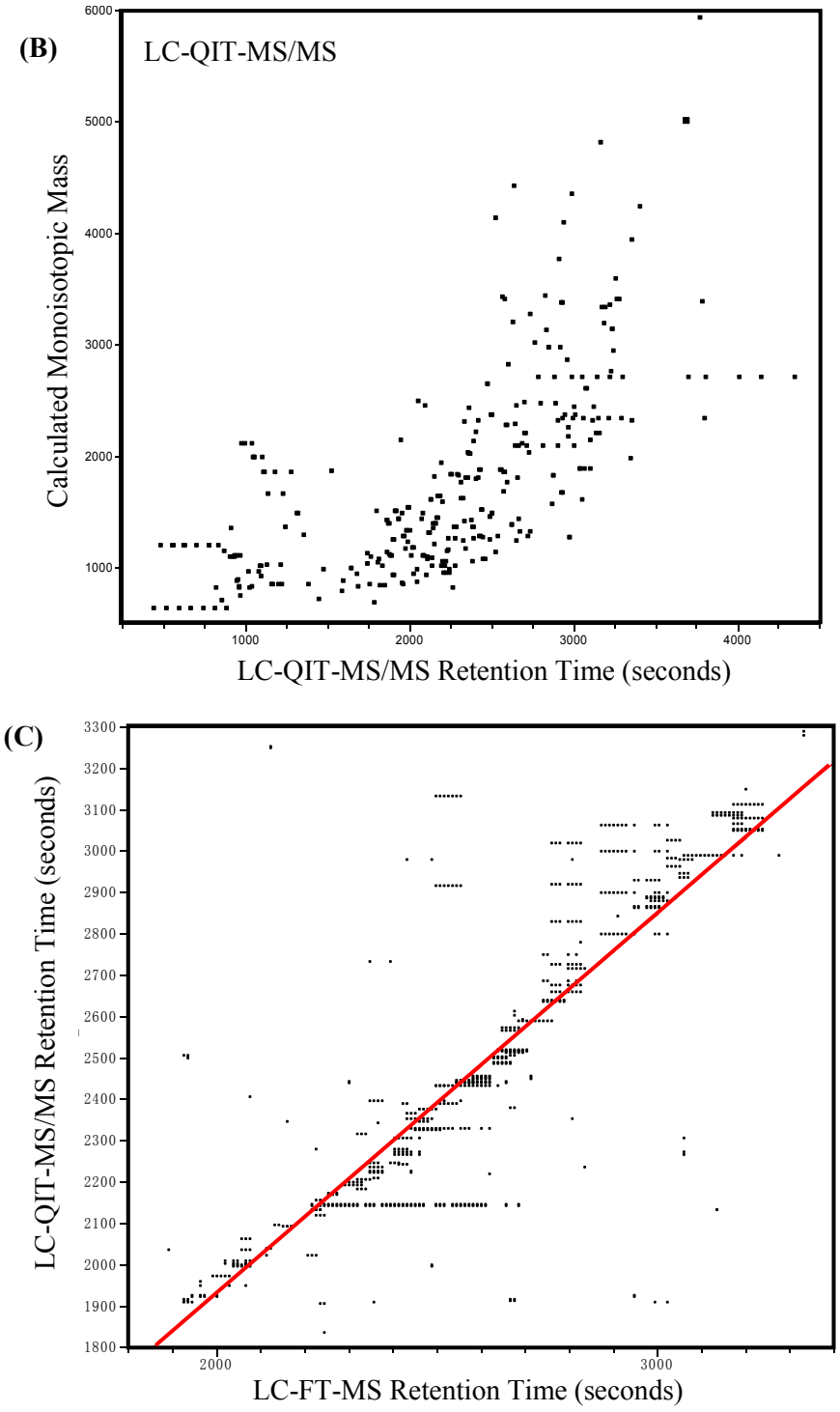
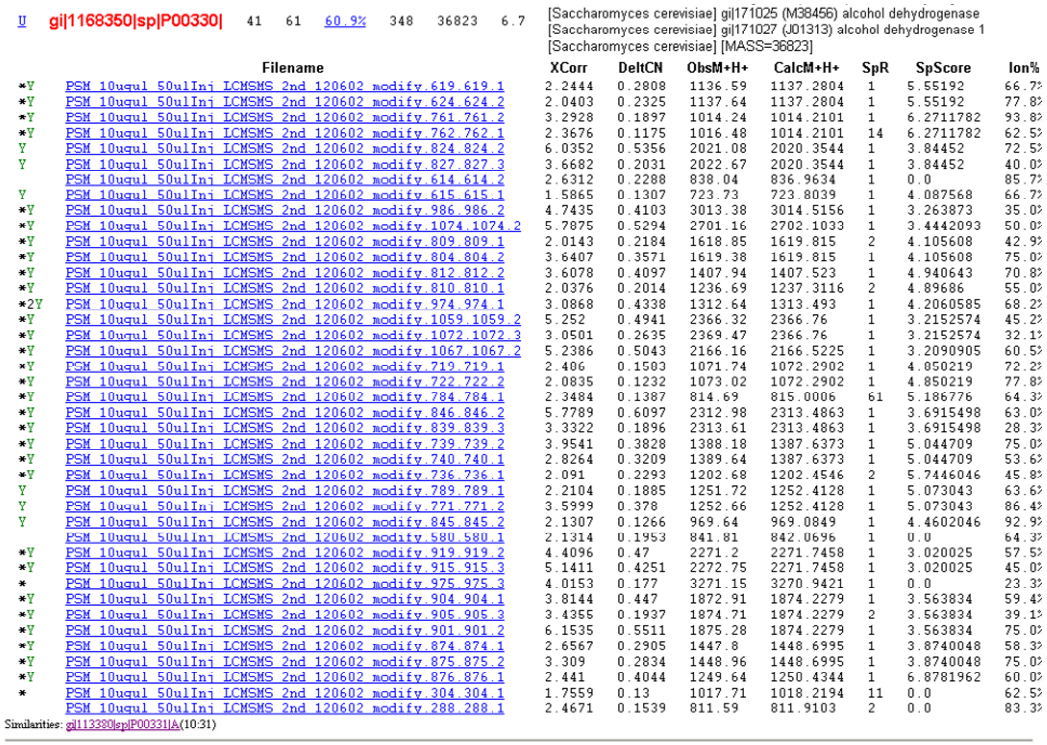


Figure 5.5: Continued.

this linear function indicates an 11.6-second retention time shift between the two LC elutions, which can arise from different sample injection times or different dead volumes. The slope of this linear function is approximately 1, which indicates that the retention time scale is not stretched or compressed between the two LC elutions. The correlation coefficient of the data points was 0.94, indicating good reproducibility of the peptides' retention time. The retention time for all data points from the LC-FT-MS experiments was normalized with this linear function.

Next, the peptide identification results from the LC-QIT-MS/MS experiment were re-filtered with a lowered threshold using DTASelect. A peptide identification is marked as being validated by the accurate mass measurement, if a mass spectral peak is observed in the LC-FT-MS experiment with less than 0.05 Da mass difference and less than 3 min retention time difference. To facilitate result interpretation, the manual validation flag in the DTASelect result file was used to indicate the validation by the LC-FT-MS data. Figure 5.6 shows the identification results of the alcohol dehydrogenase. Note that many true peptides that would have been discarded according to their Xcorr score were validated by the accurate mass measurement using the LC-FT-MS data. This indicates that integration of LC-FT-MS measurement and LC-QIT-MS/MS measurement can improve both the sensitivity and the confidence of peptide identification in shotgun proteomics.



U gi|1168350|sp|P00330| 41 61 60.9% 348 36823 6.7 [Saccharomyces cerevisiae] gi|171025 (M38456) alcohol dehydrogenase [Saccharomyces cerevisiae] gi|171027 (J01313) alcohol dehydrogenase 1 [Saccharomyces cerevisiae] [MASS=36823]

	Filename	XCorr	DeltCN	ObsM+H+	CalcM+H+	SpR	SpScore	Ion%
*Y	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.619.619.1	2.2444	0.2808	1136.59	1137.2804	1	5.55192	66.7%
*Y	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.624.624.2	2.0403	0.2325	1137.64	1137.2804	1	5.55192	77.8%
*Y	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.761.761.2	3.2928	0.1897	1014.24	1014.2101	1	6.2711782	93.8%
*Y	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.762.762.1	2.3676	0.1175	1016.48	1014.2101	14	6.2711782	62.5%
Y	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.824.824.2	6.0352	0.5356	2021.08	2020.3544	1	3.84452	72.5%
Y	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.827.827.3	3.6682	0.2031	2022.67	2020.3544	1	3.84452	40.0%
	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.614.614.2	2.6312	0.2288	838.04	836.9634	1	0.0	85.7%
Y	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.615.615.1	1.5865	0.1307	723.73	723.8039	1	4.087568	66.7%
*Y	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.986.986.2	4.7435	0.4103	3013.38	3014.5156	1	3.263873	35.0%
*Y	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.1074.1074.2	5.7875	0.5294	2701.16	2702.1033	1	3.4442093	50.0%
*Y	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.809.809.1	2.0143	0.2184	1618.85	1619.815	2	4.105608	42.9%
*Y	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.804.804.2	3.6407	0.3571	1619.38	1619.815	1	4.105608	75.0%
*Y	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.812.812.2	3.6078	0.4097	1407.94	1407.523	1	4.940643	70.8%
*Y	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.810.810.1	2.0376	0.2014	1236.69	1237.3116	2	4.89686	55.0%
*2Y	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.974.974.1	3.0868	0.4338	1312.64	1313.493	1	4.2060585	68.2%
*Y	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.1059.1059.2	5.252	0.4941	2366.32	2366.76	1	3.2152574	45.2%
*Y	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.1072.1072.3	3.0501	0.2635	2369.47	2366.76	1	3.2152574	32.1%
*Y	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.1067.1067.2	5.2386	0.5043	2166.16	2166.5225	1	3.2090905	60.5%
*Y	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.719.719.1	2.406	0.1503	1071.74	1072.2902	1	4.050219	72.2%
*Y	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.722.722.2	2.0835	0.1232	1073.02	1072.2902	1	4.850219	77.8%
*Y	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.784.784.1	2.3484	0.1387	814.69	815.0006	61	5.186776	64.3%
*Y	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.846.846.2	5.7789	0.6097	2312.98	2313.4863	1	3.6915498	63.0%
*Y	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.839.839.3	3.3322	0.1896	2313.61	2313.4863	1	3.6915498	28.3%
*Y	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.739.739.2	3.9541	0.3828	1388.18	1387.6373	1	5.044709	75.0%
*Y	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.740.740.1	2.8264	0.3209	1389.64	1387.6373	1	5.044709	53.6%
*Y	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.736.736.1	2.091	0.2293	1202.68	1202.4546	2	5.7446046	45.8%
Y	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.789.789.1	2.2104	0.1885	1251.72	1252.4128	1	5.073043	63.6%
Y	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.771.771.2	3.5999	0.378	1252.66	1252.4128	1	5.073043	86.4%
Y	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.845.845.2	2.1307	0.1266	969.64	969.0849	1	4.4602046	92.9%
	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.580.580.1	2.1314	0.1953	841.81	842.0696	1	0.0	64.3%
*Y	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.919.919.2	4.4096	0.47	2271.2	2271.7458	1	3.020025	57.5%
*Y	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.915.915.3	5.1411	0.4251	2272.75	2271.7458	1	3.020025	45.0%
*	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.975.975.3	4.0153	0.177	3271.15	3270.9421	1	0.0	23.3%
*Y	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.904.904.1	3.8144	0.447	1872.91	1874.2279	1	3.563834	59.4%
*Y	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.905.905.3	3.4355	0.1937	1874.71	1874.2279	2	3.563834	39.1%
*Y	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.901.901.2	6.1535	0.5511	1875.28	1874.2279	1	3.563834	75.0%
*Y	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.874.874.1	2.6567	0.2905	1447.8	1448.6995	1	3.8740048	58.3%
*Y	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.875.875.2	3.309	0.2834	1448.96	1448.6995	1	3.8740048	75.0%
*Y	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.876.876.1	2.441	0.4044	1249.64	1250.4344	1	6.8781962	60.0%
*	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.304.304.1	1.7559	0.13	1017.71	1018.2194	11	0.0	62.5%
	PSM_10ugul_50ulInj_LCMSMS_2nd_120602_modify.288.288.1	2.4671	0.1539	811.59	811.9103	2	0.0	83.3%

Similarities: gi|113380|sp|P00331|A(10:31)

Figure 5.6: Validation of MS/MS peptide identifications with LC-FT-MS data. The peptide identifications validated by the LC-FT-MS data are marked with a green "Y" in the first column.

CONCLUSIONS

FT-ICR mass spectrometry has to be coupled with liquid chromatography to provide adequate peak capacity and dynamic range for analyzing a proteome digest sample. With an integrated LC-FT-MS system, peptides are first separated by liquid chromatography and then analyzed by the FT-ICR mass spectrometry. We first demonstrated the use of a capillary LC-FT-MS system to compare the digestion results of two proteolysis protocols. Undigested proteins were detected with the conventional aqueous digestion protocol and were not detected with the new 80% acetonitrile digestion protocol, which indicates the superior digestion efficiency of the new protocol. To provide better ionization of aqueous solution and higher measurement sensitivity, a nanospray ionization source was developed for our FT-ICR instrument, and a fully automated nanoLC-FT-MS system was constructed. The nanoLC-FT-MS was tested with the protein standard mixture and a *Rhodopseudomonas palustris* proteome. Finally, the accurate mass measurement with nanoLC-FT-MS was integrated with the tandem mass spectrometry measurement with LC-QIT-MS/MS. The integration improved the confidence and sensitivity of peptide identification for shotgun proteomics.

Chapter 6

Robust Estimation of Peptide Abundance Ratios and Rigorous Scoring of Their Variability and Bias in Quantitative Shotgun Proteomics

All of the data presented below has been published as

C. Pan, G. Kora, D.L. Tabb, D.A. Pelletier, W.H. McDonald, G.B. Hurst, R.L. Hettich, and N.F. Samatova1, Robust Estimation of Peptide Abundance Ratios and Rigorous Scoring of Their Variability and Bias in Quantitative Shotgun Proteomics *Analytical Chemistry* 2006 (In press)

As first author, C. Pan's contributions to this article include algorithm development and MS data acquisition and interpretation.

INTRODUCTION

In quantitative shotgun proteomics, proteolysis-derived peptides are measured with liquid chromatography–tandem mass spectrometry (LC–MS/MS) and used as surrogates of their parent proteins for relative quantification. In a label free approach, the proteomes under comparison are analyzed separately in standardized LC-MS/MS runs. Alternatively, by employing stable isotope labeling, the proteomes under comparison are mixed and analyzed in one LC-MS/MS run, which eliminates the variability in sample processing steps after mixing and LC-MS/MS analysis. The common stable isotope labeling methods include ^{15}N or ^{13}C metabolic labeling (Oda, 1999), SILAC (Ong, 2002), $H_2{}^{18}O$ digestion (Yao, 2001) and ICAT (Gygi, 1999). Each peptide in the mixture of two isotopically labeled proteomes has two mass-different isotopic variants, the light isotopologue from one proteome and the heavy isotopologue from the other. Here we consider algorithms

for peptide relative quantification using mass-different stable isotope labeling. Note that the algorithms discussed here are not applicable to the isobaric labeling method iTRAQ (Ross, 2004), which generates specific reporter ions in tandem mass spectra for quantification.

Figure 6.1 illustrates the general computational procedure for estimating the abundance ratio between the light isotopologue and the heavy isotopologue of a peptide. The sequence of the peptide is identified from an MS/MS scan of one of its isotopologues (Figure 6.1A). The full scan that triggered this MS/MS scan is shown in Figure 6.1B, in which the mass spectral peaks of the two isotopologues are highlighted. Selected ion chromatograms for the two isotopologues are then extracted, and peak detection is performed to define front and back boundaries of the two isotopologues' chromatographic peaks (Figure 6.1C). Finally, the abundance ratio of the peptide is evaluated from the two chromatographic peaks. In this study, we developed novel algorithms for peak detection and for peptide abundance ratio evaluation.

Normally, peak detection is performed in the selected ion chromatograms. However, a large fraction of chromatographic peaks have a very low chromatographic signal-to-noise ratio (S/N) that results in incorrect assignments of their peak boundaries. We have improved the robustness of peak detection by employing a parallel paired covariance algorithm, which was developed based on sequential paired covariance algorithm originally devised for rapid component identification from ion electropherograms (Muddiman, 1995; Muddiman, 1997). The parallel paired covariance algorithm integrates

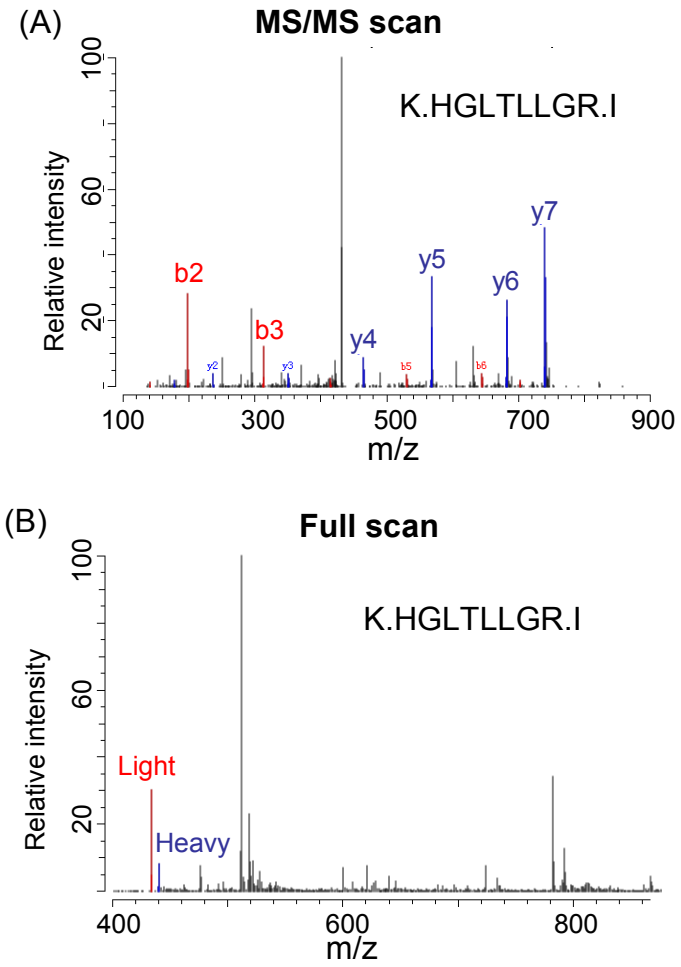


Figure 6.1: Estimation of peptide abundance ratios in quantitative shotgun proteomics. (A) The sequence of a peptide is identified from an MS/MS scan. (B) The peak pair in the full scan for the two isotopologues of this peptide is identified. (C) The selected ion chromatograms of the light isotopologue (red) and the heavy isotopologue (blue) are extracted from the full scans. The brown vertical line indicates the MS/MS scan. The chromatographic peaks of the isotopologue pair are detected as between the two vertical green lines. The abundance ratio between the two isotopologues can be estimated as the ratio of the peak areas. (D) The abundance ratio is estimated from a peak profile. The blue data points represent the ion intensities of the two isotopologues measured in the full scans within the chromatographic peak. The red line has the minimum total squared perpendicular offset to the data points, whose slope is an abundance ratio estimator.

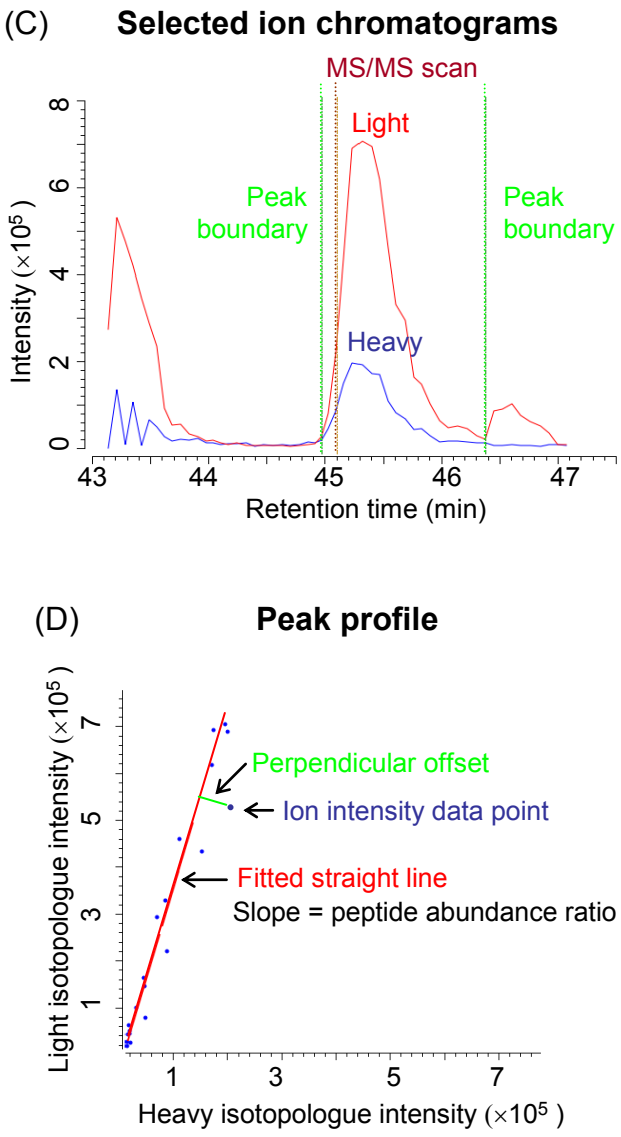


Figure 6.1: Continued.

the two selected ion chromatograms and reconstructs a covariance chromatogram of improved chromatographic S/N for peak detection.

Estimation of the peptide abundance ratios from the detected chromatographic peaks can be accomplished with several existing algorithms. A common algorithm is based on peak area. First, the selected ion chromatograms are smoothed to remove random noise. Then, the background is subtracted from the chromatograms. Finally, the area under the two chromatographic peaks is calculated by integration. The ratio between the two peak areas is considered as the abundance ratio of a peptide. The peak area algorithm has been used in quantitative proteomics programs XPRESS (Han, 2001), ASAPratio (Li, 2003), and MSQuant (Schulze, 2004). The accuracy of the peak area calculation highly depends on two empirical steps – chromatogram smoothing and background subtraction. MacCoss *et al* argued that background subtraction is difficult to optimize for thousands of different chromatographic peaks measured in a proteomics experiment and leads to less reliable abundance ratio estimation (MacCoss, 2003). In their program RelEx, a correlation algorithm based on peak profiles is used to calculate peptide abundance ratios (Lawson, 1980; MacCoss, 2003). A *peak profile* is a scatter plot of ion intensities of the two isotopologues detected in each full scan within the chromatographic peaks (Figure 6.1D). The correlation algorithm fits a straight line that has the minimum total squared perpendicular offset to the data points in the peak profile (Figure 6.1D). The slope of the fitted line is an estimator of the peptide abundance ratio.

The existing algorithms that evaluate peptide abundance ratios do not formally "score" the abundance ratio estimates for their expected bias and variability. In quantitative shotgun proteomics, the abundance ratios for tens of thousands of identified peptides can be estimated, but with dramatically varying error. We propose to use a principal component analysis algorithm to not only estimate the abundance ratio of a peptide from its peak profile but also score the estimation with a signal-to-noise ratio measure of its peak profile (*profile S/N*). We show that the profile S/N is inversely correlated with both the standard deviation and the bias of the abundance ratio estimation. Thus, the profile S/N allows stratification of the peptide abundance ratios into those with greater or lesser estimation accuracy and precision. As a result, it becomes possible to statistically evaluate every peptide abundance ratio in the subsequent protein abundance ratio estimation (Pan, 2006a).

Here we describe both the parallel paired covariance algorithm for peak detection and the principal component analysis algorithm for peptide abundance ratio estimation and scoring. These two algorithms have been assembled into a computer program, termed ProRata (Pan, 2006a). ProRata automates the entire data analysis pipeline for quantitative proteomics with stable isotope labeling, incorporating selected ion chromatogram extraction and peptide abundance ratio evaluation for the ultimate goal of protein abundance ratio estimation. The graphical user interface of ProRata also allows manual data interrogation for result validation. ProRata is freely available at www.MSProRata.org.

MATERIALS AND METHODS

Standard Isotopically Labeled Proteome Mixture Preparation.

Wild type *Rhodopseudomonas palustris* CGA0010 strain was grown anaerobically in light on defined minimal growth media to mid-log phase at 30°C. $(NH_4)_2SO_4$ was the only nitrogen source available for bacterial assimilation, provided as $(^{14}NH_4)_2SO_4$ for the unlabeled culture and as $(^{15}NH_4)_2SO_4$ for the ^{15}N-labeled culture (>98 atom percentage excess, Sigma-Aldrich, St. Louis, MO). The ^{15}N-enriched nitrogen from $(NH_4)_2SO_4$ was incorporated into proteins through metabolism in the ^{15}N-labeled culture. Except for the different isotopologues of $(NH_4)_2SO_4$ in the growth media, the two cultures were otherwise identically prepared. Cells were harvested by centrifugation and washed twice with ice-cold wash buffer (50 mM Tris-HCl buffer at pH 7.5 with 10 mM EDTA). Cells were then lysed by sonication in ice-cold wash buffer and unbroken cells were removed with low-speed centrifugation (5000 g for 10 minutes). The obtained cell lysates were fractionated by ultracentrifugation at 100,000 g for 1 hour and the supernatants from the unlabeled and ^{15}N-labeled cell lysates were labeled as the ^{14}N proteome and ^{15}N proteome, respectively. Protein concentration in the two proteomes was determined with Lowry's analysis (Lowry, 1951). Standard mixtures were prepared by mixing the two proteomes at ^{14}N:^{15}N ratios of 10:1, 5:1, 1:1, 1:5, and 1:10 by their total protein mass.

Shotgun Proteomics Measurement.

The proteins in the standard mixtures were denatured and reduced with treatment of 6 M guanidine and 10 mM dithiothreitol (DTT) (Sigma Chemical Co. St. Louis, MO) at 60 °C for 1 hour. After six-fold dilution with 50 mM Tris-HCl/10 mM $CaCl_2$ (pH 7.8), the proteins were digested at 37°C with sequencing grade trypsin (Promega, Madison, WI). The samples were then reduced with 20 mM DTT for 1 hour at 60 °C and were desalted using C18 solid-phase extraction (Sep-Pak Plus, Waters, Milford, MA). The protein digests were examined with LC-MS/MS using twelve-step split-phase MudPIT (MacCoss, 2002; McDonald, 2002). The samples were loaded via a pressure bomb (New Objective, Woburn, MA) onto a 250-*u*m-I.D. front column packed with 2 cm strong cation exchange resin (Luna, Phenomonex) and 2 cm C18 reverse-phase resin (Aqua, Phenomonex). A 100-*u*m-I.D. PicoFrit column (New Objective, Woburn, MA) was packed with 15 cm C18 reverse-phase resin. The front column was connected with the PicoFrit column and then placed in-line with a Surveyor quaternary HPLC (ThermoFinnigan, San Jose, CA). The composition of the aqueous solvent was 95% H_2O (Burdick & Jackson, Muskegon, MI), 5% ACN (Burdick & Jackson, Muskegon, MI), and 0.1% formic acid (EM Science, Darmstadt, Germany) and the composition of the organic solvent was 30% H_2O, 70% ACN and 0.1% formic acid. Two-dimensional LC separation was performed with twelve salt pulses (0 mM, 35 mM, 50 mM, 60 mM, 75 mM, 100 mM, 125 mM, 150 mM, 200 mM, 250 mM, 300 mM and 500 mM ammonium acetate (Sigma Chemical Co., St. Louis, MO) in the aqueous solvent). Each salt pulse was followed by a 2-hour reverse-phase gradient from 100% aqueous solvent to 50% aqueous solvent and 50% organic solvent.

LC-MS/MS analysis was performed on an LTQ linear ion trap instrument (ThermoFinnigan, San Jose, CA) with dynamic exclusion enabled. Each full scan (400-1700 *m/z*) was followed by five data-dependent MS/MS scans at 35% normalized collision energy. All scans were averaged from two microscans.

Peptide and Protein Identification.

All MS/MS scans were searched with the SEQUEST program (Eng, 1994) against an *R. palustris* protein sequence database (Larimer, 2004). The light isotopologues of peptides were identified using normal amino acid masses in the SEQUEST parameter file and the heavy isotopologues were identified using ^{15}N-labeled amino acid masses. OUT files were converted to SQT files using UNITEMARE program (generously provided by Dr. John R. Yates' laboratory). DTASelect (Tabb, 2002) was used to filter the peptide identifications based on Xcorr and delCN (Xcorr > 1.8 (+1), > 2.5 (+2), and > 3.5 (+3); delCN > 0.08). The peptides were assembled into proteins, retaining duplicate MS/MS scans of a peptide (DTASelect option: `-t 0`).

Selected Ion Chromatogram Extraction.

The Xcalibur RAW files were converted into the mzXML format with the ReAdW program (Pedrioli, 2004). An mzXML parser, RAMP, was used to access the mzXML files. The selected ion chromatograms were extracted in the following steps. 1) The peptide identifications were parsed out from DTASelect-filter.txt, including their amino

acid sequence, charge state, and protein locus. 2) The *m/z* windows were calculated for the two isotopologues of each peptide identification. Theoretical isotope distributions for both isotopologues were calculated based on the sequence, the user-defined isotopic compositions of the atoms and the user-defined atomic compositions of all residues and their modifications. In this study, the nitrogen atoms in the heavy isotopologues were specified to be 98%-enriched ^{15}N. The *m/z* windows for an isotopic distribution are configured to be its major isotopes' *m/z* values plus and minus the *m/z* tolerance; the major isotopes were specified to be the isotopes with a relative abundance of more than 10%, and the *m/z* tolerance was defined to be 0.5 in this study. 3) The peptide identifications with the same sequence and charge state were grouped, if their MS/MS scans were acquired within a 2-minute interval. Redundant identifications from a single chromatographic peak of a peptide were often found for either or both isotopologues of the peptide. 4) A pair of selected ion chromatograms was extracted for the light and heavy isotopologues of each peptide identification group. The retention time window for both selected ion chromatograms was defined as from 2 minutes before the first MS/MS scan to 2 minutes after the last MS/MS scan of the grouped peptide identifications.

Chromatographic Peak Detection.

The covariance chromatogram of an isotopologue pair was reconstructed from its selected ion chromatograms with the parallel paired covariance algorithm. The parallel paired covariance at a full scan in the covariance chromatogram is the product of the

background-subtracted ion intensities at that full scan in the two selected ion chromatograms:

$$C_k = (I_k^L - I_{BG}^L) \cdot (I_k^H - I_{BG}^H) \quad d \le k \le h\,, \tag{6.1}$$

where I_{BG}^L and I_{BG}^H are the background ion intensities of the selected ion chromatograms for the light and heavy isotopologues, respectively; I_k^L and I_k^H are the ion intensities at full scan k in the m/z windows for the light and heavy isotopologues, respectively; and C_k is the covariance of the two isotopologues' background-subtracted intensities at full scan k. The background ion intensities I_{BG}^L and I_{BG}^H were defined to be the minimum ion intensities in the selected ion chromatograms for the light and heavy isotopologues, respectively. Scan d and scan h are the first and last full scans, respectively, in the selected ion chromatograms. The time series of I_k^L and I_k^H form the selected ion chromatograms for the two isotopologues and, likewise, the time series of C_k forms the covariance chromatogram of the isotopologue pair as illustrated in Figure 6.2. The covariance chromatogram was then smoothed with 7-point quadratic Savitsky-Golay filter (Press, 2002). The chromatographic peak for a peptide was defined as between scan a and scan b ($d \le a < b \le h$) that are two local covariance minima with the smallest interval that includes all MS/MS scans matched to this peptide (Figure 6.2). A local covariance minimum was a scan with the lowest covariance within a seven-point symmetric window surrounding this scan in the covariance chromatogram. Scans a and b were the peak boundaries as labeled in Figure 6.2. The parallel paired covariance algorithm is also capable of determining the retention time shift between the two isotopologues. For this process, the two selected ion chromatograms would be shifted

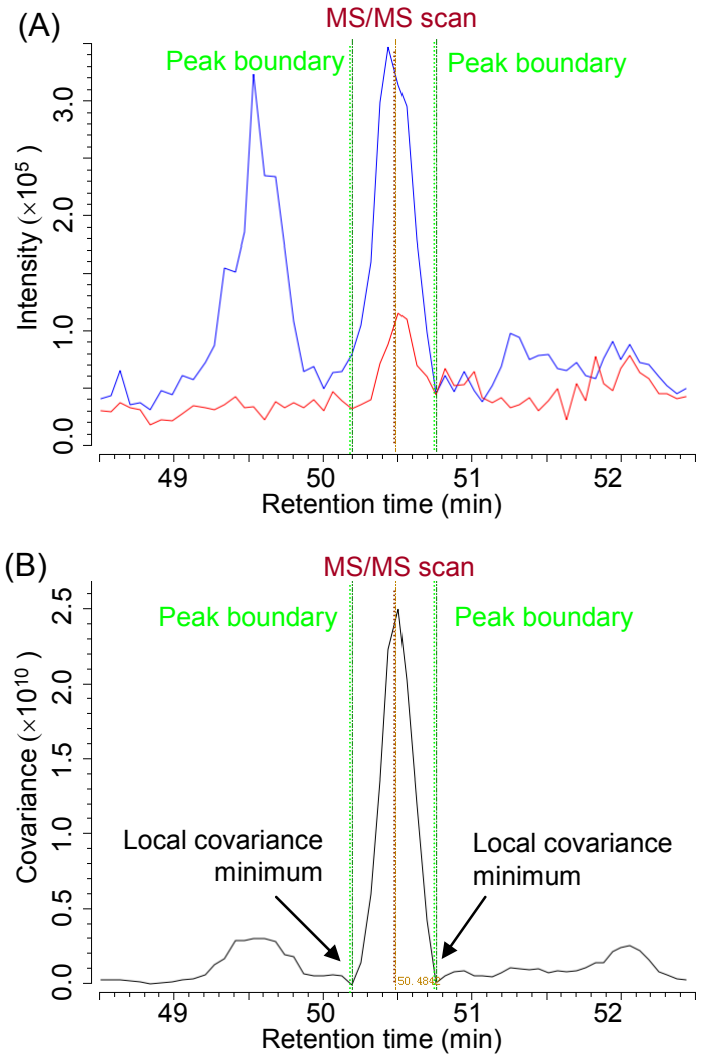


Figure 6.2: Selected ion chromatograms and parallel-paired covariance chromatogram. The selected ion chromatograms for the light isotopologue (red) and heavy isotopologue (blue) of a peptide are shown in part A and the covariance chromatogram is shown in part B. The peak boundaries (vertical green lines) are determined in the covariance chromatogram and transferred to the selected ion chromatograms.

relative to each other scan by scan until the peak height of the peptide in the covariance chromatogram is maximized.

Peptide Abundance Ratio Estimation and Scoring.

The peptide abundance ratios and the profile S/Ns were estimated with principal component analysis of the peak profiles. Note that the background subtraction was only performed for peak detection. The peak profiles were constructed from the originally extracted selected ion chromatograms. Here, we present the equations used in the estimation and their derivation from a set of definitions and assumptions.

The detected ion intensities, I_i^L and I_i^H, of the light and heavy isotopologues at full scan *i* are composed of their true signal (denoted by S_i^L and S_i^H, respectively) corrupted by the random noise (denoted by N_i^L and N_i^H, respectively) and superimposed on the backgrounds (denoted by B^L and B^H, respectively) (Lawson, 1980):

$$\begin{cases} I_i^L = S_i^L + N_i^L + B^L \\ I_i^H = S_i^H + N_i^H + B^H \end{cases} \quad a \le i \le b\,, \tag{6.2}$$

where *a* and *b* are the chromatographic peak boundaries. Let us assume that: i) the backgrounds, B^L and B^H, hold constant across full scans; ii) the random noises, N_i^L and N_i^H, have zero-mean; and iii) the ratio between the two signals, S_i^L and S_i^H, is constant across scans, and defines the peptide abundance ratio, *R*. The third assumption, expressed as Equation 6.3, is based on the exact co-elution of the two isotopologues:

$$R = S_i^L / S_i^H \quad a \le i \le b\,. \tag{6.3}$$

The constant background, B^L and B^H, can be eliminated from our consideration by centering the intensities and the signals on their means. Let I'^L_i and I'^H_i denote the mean-centered intensities and let S'^L_i and S'^H_i denote the mean-centered signals. Then Equations 6.2 and 6.3 can be transformed, respectively, to:

$$\begin{cases} I'^L_i = S'^L_i + N^L_i \\ I'^H_i = S'^H_i + N^H_i \end{cases} \quad a \le i \le b \tag{6.4}$$

$$R = S'^L_i / S'^H_i \tag{6.5}$$

This transformation is the reason why the peak-profile-based algorithms can obviate the background subtraction step and, therefore, eliminate the probable error in this step.

Principal component analysis is generally applied to a set of vectors. Let us define the following vectors: the ion intensity vector $\mathbf{I}_i = (I'^H_i, I'^L_i)$, the signal vector $\mathbf{S}_i = (S'^H_i, S'^L_i)$ and the noise vector $\mathbf{N}_i = (N^H_i, N^L_i)$. Therefore, Equations 6.4 and 6.5 can be transformed to a vector form:

$$\mathbf{I}_i = \mathbf{S}_i + \mathbf{N}_i\,, \tag{6.6}$$

$$R = \tan(\theta_{\mathbf{S}_i})\,, \tag{6.7}$$

where $\theta_{\mathbf{S}_i}$ is the direction angle of the signal vector $\mathbf{S}_i$ (Figure 6.3). The vectors $\mathbf{I}_i$ are known from the measurement, but the vectors $\mathbf{S}_i$ and $\mathbf{N}_i$ are unknown and need to be determined for calculating the abundance ratio, R, and the profile signal-to-noise ratio.

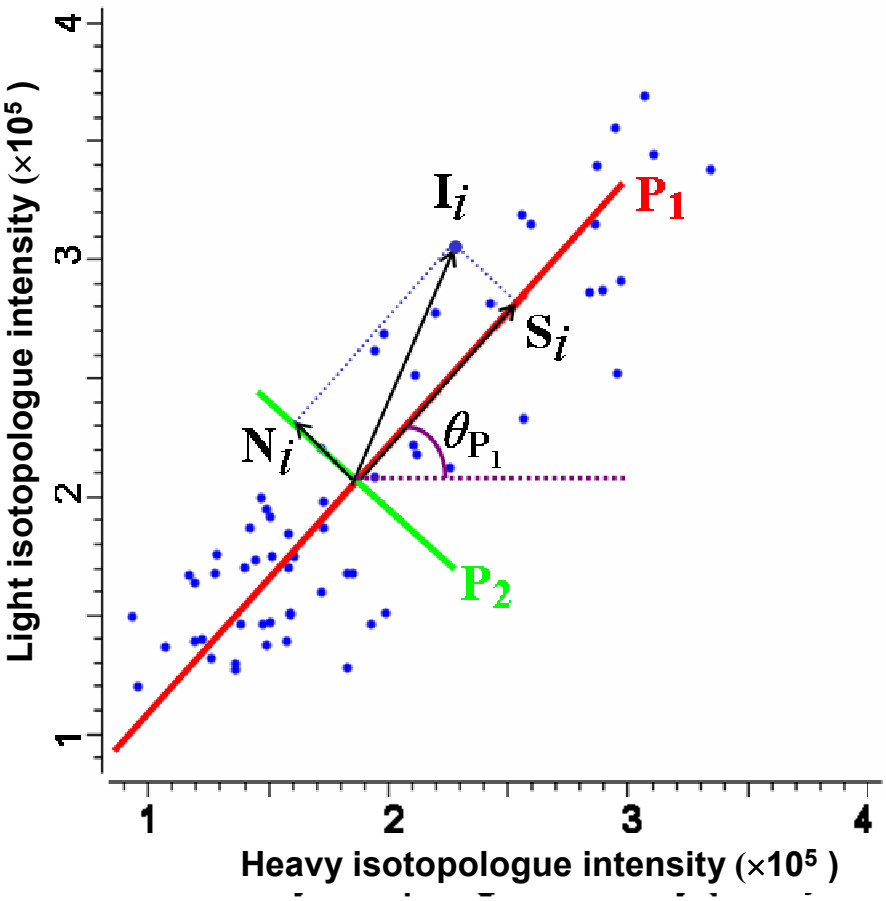


Figure 6.3: Estimation of peptide abundance ratio with the principal component analysis algorithm. Two principal components are represented with two lines, the red line for the first principal component P1 and the green line for the second principal component P2. The ratio between the lengths of the two lines is plotted to be equal to the profile S/N, which captures how elliptical the ensemble of the data points in the peak profile is. The ion intensity vector for each data point (Ii) can be decomposed to the signal vector (Si, the projection of Ii on P1) and the noise vector (Ni, the projection of Ii on P2). The slope of P1 (tan(θ P1)) is an estimator of the peptide abundance ratio.

Obviously, $\mathbf{S}_i$ and $\mathbf{N}_i$ cannot be solved analytically from Equation 6.6. Instead, determining $\mathbf{S}_i$ and $\mathbf{N}_i$ is formulated as an optimization problem of finding such vectors $\mathbf{S}_i$ and $\mathbf{N}_i$ that the variance of the norm of the noise vectors, $\sigma^2(|\mathbf{N}_i|)$, is minimized. This optimization problem can be solved by principal component analysis of the peak profile, as shown in Equation 6.8,

$$\begin{cases} \mathbf{S}_i = (\mathbf{I}_i \cdot \mathbf{P}_1)\mathbf{P}_1 \\ \mathbf{N}_i = (\mathbf{I}_i \cdot \mathbf{P}_2)\mathbf{P}_2 \end{cases}, \tag{6.8}$$

where $\mathbf{P_1}$ and $\mathbf{P_2}$ are the corresponding first and second principal components of the intensity vectors $\mathbf{I}_i$. This means that the direction of all signal vectors is the direction of the first principal component and the length of a signal vector is the dot product between the intensity vector and the first principal component. The direction and length of the noise vectors are determined, likewise, with the second principal component. Geometrically, a signal vector and a noise vector are the projections of their intensity vector on the first principal component and the second principal component, respectively as illustrated in Figure 6.3. Principal component analysis of the intensity vectors $\mathbf{I}_i$ in the peak profile calculates the principal components $\mathbf{P_1}$ and $\mathbf{P_2}$ and their associated eigenvalues λ_1 and λ_2. The principal components and eigenvalues provide the estimators for the peptide abundance ratios and profile S/Ns, as described below.

The abundance ratio is the tangent of the direction angle of the first principal component:

$$R = \tan(\theta_{\mathbf{P}1}). \tag{6.9}$$

The abundance ratio estimated with principal component analysis is exactly the same as the abundance ratio estimated with linear correlation. This is because the direction of the first principal component is exactly the same as the direction of the straight line with minimum total squared perpendicular offset (Jolliffe, 2002), both of which are estimators of the peptide abundance ratio.

Let us define the signal-to-noise ratio for the ion intensity vectors in the peak profile as the ratio between the standard deviation of the length of the signal vectors and that of the length of the noise vectors:

$$S/N_{profile} \equiv \frac{\sigma(|\mathbf{S}_i|)}{\sigma(|\mathbf{N}_i|)}. \tag{6.10}$$

We refer to this signal-to-noise ratio as the *profile signal-to-noise ratio* to distinguish it from the chromatographic S/N, since the profile S/N is based on the peak profile, while the chromatographic S/N is based on the ion chromatogram. The first eigenvalue, λ_1, is the variance of the projection of the intensity vectors on the first principal component, which is the variance of the length of the signal vectors. Likewise, the second eigenvalue, λ_2 for the second principal component is the variance of the length of the noise vectors. The profile S/N is calculated as the square root of the ratio between λ_1 and λ_2:

$$S/N_{profile} = \sqrt{\frac{\lambda_1}{\lambda_2}}. \tag{6.11}$$

The calculation of the profile S/N from the eigenvalues is the key feature of the principal component analysis algorithm that distinguishes it from the linear correlation algorithm.

For comparison, the peptide abundance ratios were also estimated with the peak area algorithm. To calculate the peak area, selected ion chromatograms were smoothed with a 7-point quadratic Savitsky-Golay filter (Press, 2002). The background of a selected ion chromatogram was set to be the straight line connecting the two ion intensities at the peak boundaries. The peak area is the total background-subtracted intensities of a peak.

RESULTS AND DISCUSSION

In this study, standard mixtures of isotopically labeled proteomes were used to test the proposed algorithms. Abundance ratios between the light and heavy isotopologues of all peptides in the standard mixtures were expected to be approximately the same as the mixing ratio of the ^{14}N proteome and the ^{15}N proteome. This does assume that the protein abundance profiles should be the same for the two proteomes extracted from the identically grown cells. Six datasets were acquired from standard mixtures, including 1:1a, 1:1b, 5:1, 1:5, 10:1 and 1:10. The 1:1a and 1:1b datasets are from duplicate measurements of the 1:1 standard mixture.

Ion Chromatogram Extraction with Re-organized Peptide Identifications

Both the light and heavy isotopologues of peptides were considered when searching the MS/MS scans (Table 6.1). We consider every chromatographic peak as an independent measurement of the peptide abundance ratio. If a peptide is identified in multiple chromatographic peaks, all identifications are retained and used for extracting the

Table 6.1 The peptide quantification results from the six standard mixture datasets.

Standard Mixtures		Peptide Counts				Log-ratio		Log-profile-S/N	
^{14}N:^{15}N	Log-ratio	^{14}N ID	^{15}N ID	SIC[#]	Quantified	Median	AAD *	Median	AAD *
1:1a	0.0	13,766	11,665	17,574	11,919	-0.17	0.67	2.17	0.58
1:1b	0.0	13,975	12,230	18,472	12,958	-0.16	0.69	2.15	0.56
5:1	2.3	23,527	5,122	23,256	14,583	1.66	1.04	2.29	0.73
1:5	-2.3	5,676	18,037	18,855	12,453	-1.99	0.94	2.44	0.78
10:1	3.3	24,257	2,725	22,770	12,914	2.20	1.37	2.31	0.85
1:10	-3.3	3,167	21,396	20,945	12,910	-2.80	1.39	2.46	0.91
Average		---	---	20,312	12,956	---	1.02	2.30	0.74

* AAD: Absolute average deviation from the median

[#] SIC: Selected ion chromatogram

selected ion chromatograms. However, if a peptide is identified in the same charge state at multiple retention time points across a single chromatographic peak, the different identifications are combined and used to extract a single selected ion chromatogram pair.

The *m/z* windows for extracting ion chromatograms were calculated from the isotopic distributions of a peptide's isotopologues. The heavy isotopologues had a theoretical isotopic distribution skewed by the incomplete enrichment of the heavy stable isotope. Our ion chromatogram extraction algorithm has the capability of handling mass spectral data of varying resolution. Normally, ion chromatograms are extracted from a single *m/z* window for an ion species. To allow high-resolution ion chromatogram extraction, an *m/z* window is opened for each major isotope in the isotopic distribution. The width of the m*/z* windows can be configured to fit the measurement resolution of the mass spectrometer. In this way the background noise between two isotopes can be notched out, if a high-resolution mass spectrometer is used. In this study, as a linear ion trap instrument was used, the mass tolerance was set to be ±0.5 Da and the *m/z* windows for individual isotopes were merged into one *m/z* window per isotopologue. Our algorithm can also be configured to extract ion chromatograms for other isotope labeling techniques, such as SILAC, ICAT and $H_2{}^{18}O$ digestion.

Chromatographic Peak Detection with Parallel Paired Covariance

The selected ion chromatograms of a peptide were extracted for a user-defined retention time window around the MS/MS scans of the peptide. Then, the exact retention time

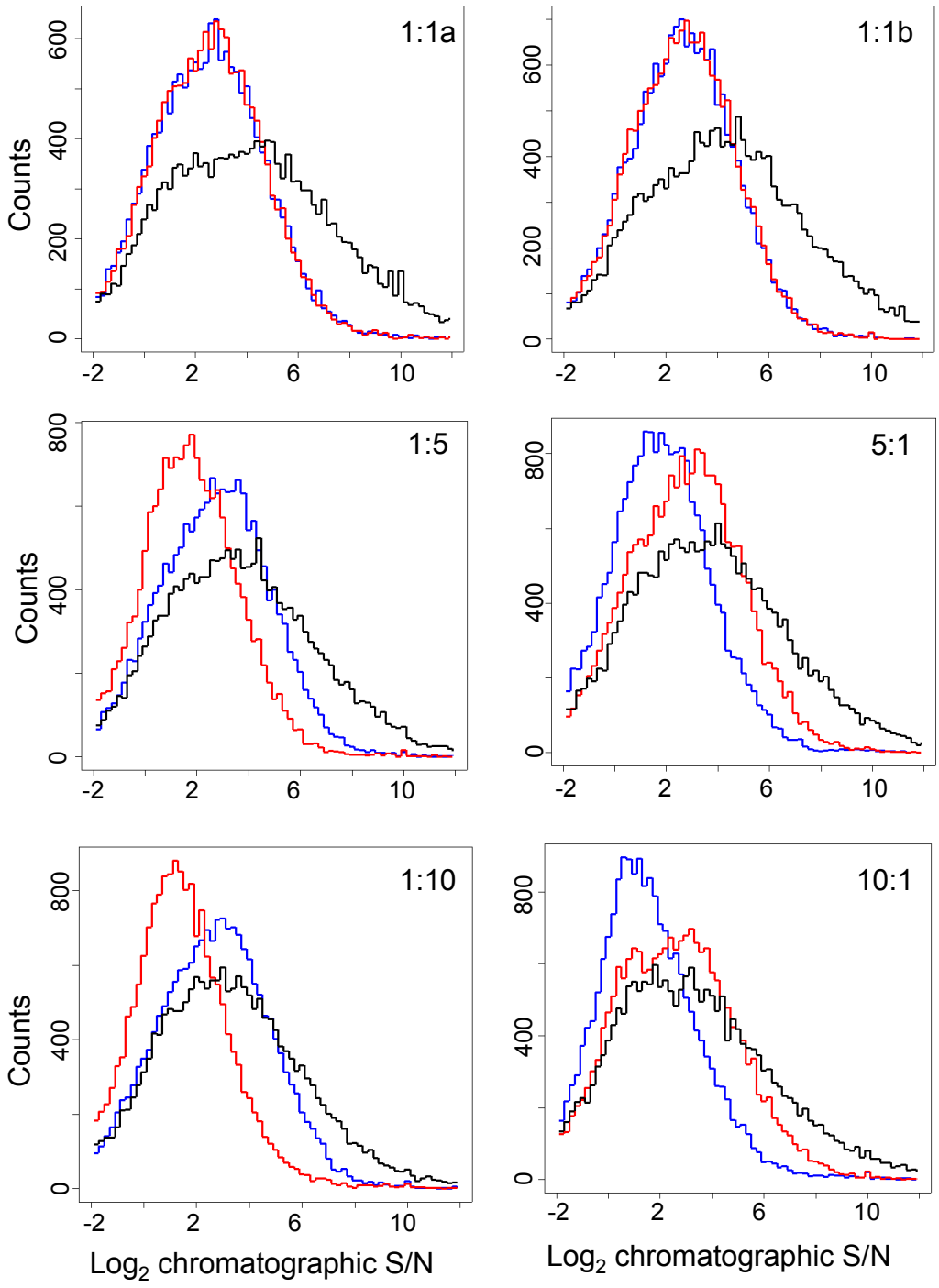
boundaries of the two isotopologues' chromatographic peaks were determined by performing peak detection in the covariance chromatogram. The covariance chromatogram was constructed by combining the two selected ion chromatograms (Figure 6.2). The two co-eluting peaks in selected ion chromatograms are represented by one greatly enhanced peak in the covariance chromatogram, whereas the noise peaks appearing in only one selected ion chromatogram are suppressed in the covariance chromatogram (Figure 6.2). This indicates that the parallel paired covariance algorithm multiplies the signal of the two chromatographic peaks and effectively reduces the uncorrelated noise in the selected ion chromatograms. As a result, the signal-to-noise ratio of the covariance chromatogram is greater than either one of the selected ion chromatograms, and the peak representing the elution of the isotopologue pair can be detected with greater accuracy.

A practical advantage of using the parallel paired covariance algorithm is obviation of the need for peak detection in *both* selected ion chromatograms. In Figure 6.2, the peak detection for the light isotopologues (in the red chromatogram) is difficult due to the low peak height and high noise fluctuation. Peak detection was found to be virtually impossible for the less abundant isotopologue of many peptides in the 1:10 and 10:1 standard mixtures. An alternative method is to determine peak boundaries only in the selected ion chromatogram of the more abundant isotopologue, but this method ignores the signal of the other isotopologue and requires the knowledge of which isotopologue is more abundant *prior to* estimating the abundance ratio between the two isotopologues.

As the accuracy of the peak detection is directly related to the signal-to-noise ratio of the chromatogram, we constructed the histograms of the signal-to-noise ratio for the covariance chromatograms (the black histogram) and the selected ion chromatograms (the blue and red histograms) (Figure 6.4). Virtually all peptides have an improved signal-to-noise ratio in the covariance chromatogram than in the selected ion chromatogram of the more abundant isotopologue. The degree of improvement is related to the signal-to-noise ratio of the less abundant isotopologue. In accord with general chromatography protocols, a signal-to-noise ratio of 3 or greater is generally required for a chromatographic peak to be accurately defined in its chromatogram. More peptides exceed this signal-to-noise ratio threshold with their covariance chromatogram than with their individual selected ion chromatogram; therefore, more peptides can have correctly assigned peak boundaries by using the parallel paired covariance algorithm.

The parallel paired covariance algorithm is based on the assumption of the co-elution of the two isotopologues. While this is generally true for most peptides labeled with ^{13}C, ^{18}O, and ^{15}N, the peptides labeled with ^{2}H can show a retention time shift between the two isotopologues. The retention time shift can be computationally offset by shifting one selected ion chromatogram relative to the other one. However, this extra computational step can add additional probable error to the peptide quantification process.

Figure 6.4: Distribution of $\log_2$ chromatographic S/N for the six standard mixture datasets. Histograms of $\log_2$ chromatographic S/N are shown for the two selected ion chromatograms (blue for the light isotopologue and red for the heavy isotopologue) and their covariance chromatogram (black). The covariance chromatogram has a higher average signal-to-noise ratio than either of the two selected ion chromatograms. The mixing ratio between the ^{14}N proteome and the ^{15}N proteome for a standard mixture is shown in each histogram.


1:1a
1:1b
1:5
5:1
1:10
10:1
Counts
Log2 chromatographic S/N
0
200
400
600
800
-2
2
6
10

Evaluation of the Peptide Abundance Ratio Estimation Accuracy

The peptide abundance ratios were then estimated by principal component analysis of the peak profile (Figure 6.3). As the isotopologues begin eluting off the column, the trace of the data points starts from close to the origin and moves upwards and to the right. After the two isotopologues reach the top of their chromatographic peaks, the trace of the data points regresses back toward the origin. Ideally, the trace of the data points should form *a straight line* whose slope is the peptide abundance ratio. However, the random noise component of the ion intensities will make the trace "wiggle" along the straight line. Essentially, the purpose of principal component analysis is to separate the noise component and the signal component of the ion intensities.

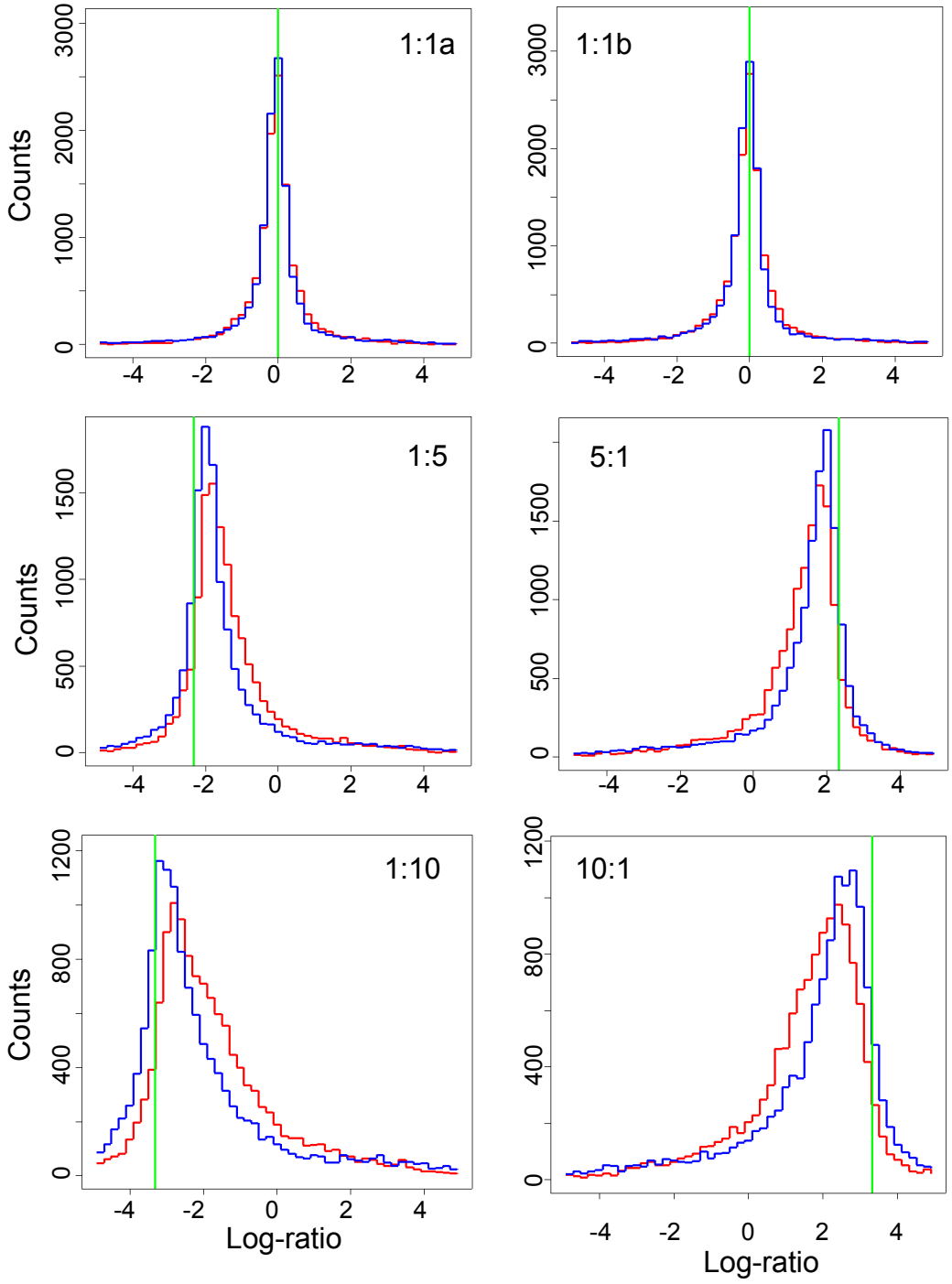
The accuracy of estimating peptide abundance ratios was benchmarked with the six standard mixture datasets. The principal component analysis algorithm was compared with a more commonly used algorithm based on peak area calculation. On average, in each dataset, the selected ion chromatograms were extracted for ~20,000 peptides and, after filtering the peptides with a profile S/N cutoff of 2.0, the total number of quantified peptides was ~13,000 (Table 6.1). The filtering effectively removes the peptides that cannot be reliably quantified by either algorithm. This profile S/N cutoff is discussed in the next section.

The estimated abundance ratios were transformed to logarithm base-2, abbreviated as *log-ratio*, and histograms of the log-ratios estimated with the principal component

analysis algorithm and the peak area ratio algorithm were constructed (Figure 6.5). Although all peptides in a standard mixture should have the same abundance ratio as the mixing ratio, the log-ratio distributions spread around a center, which can be attributed to the random error of the estimation with both methods and, probably to a less extent, to the biological variability of the ^{14}N and ^{15}N cultures. For the four standard mixtures of uneven mixing ratios, their log-ratio distributions have a center slightly shifted towards zero from the log mixing ratio and have a heavier shoulder in the side toward zero. This suggests a systematic error of the estimation with both methods. However, the log-ratio distributions from the principal component analysis method (the blue histogram) is located closer to the log mixing ratio with less spread around the center and a lighter shoulder in the side toward zero, which indicates the less random and systematic errors in the log-ratio estimation with the principal component analysis algorithm.

Two features of the principal component analysis algorithm, which are also shared with the correlation algorithm, can contribute to the improved peptide abundance ratio estimation (Thorne, 1984; Thorne, 1986; MacCoss, 2003). The first feature is the obviation of background subtraction by assuming a constant background within the chromatographic peak. In the peak area method, however, the backgrounds of the two chromatograms have to be subtracted from their peak area. Since the automatic routine for background estimation can be error-prone for many peptides, the errors in the background estimation translate directly into errors of abundance ratio estimation. The second feature is a built-in mechanism for removing random noise. The signal component

Figure 6.5: Distribution of peptide log-ratio estimates for the six standard mixture datasets. The histograms of the log-ratios estimated with the principal component analysis algorithm (blue histograms) and the peak area algorithm (red histograms) are shown for all standard mixture datasets. Only the peptides with profile S/Ns greater than two are considered. The $\log_2$ mixing ratio is marked with a green vertical line in each histogram.


1:1a
1:1b
1:5
5:1
1:10
10:1
Counts
Log-ratio
0
1000
2000
3000
500
1500
400
800
1200
-4
-2
0
2
4

and the random noise component are separated into the first principal component and the second principal component, respectively, and the first principal component is used to estimate the abundance ratio. The peak area method can only rely on chromatogram smoothing to remove the random noise. As the selection of a routine and its parameters for chromatogram smoothing is fairly subjective, it would be difficult to obtain optimized chromatogram smoothing across thousands of selected ion chromatograms with varying peak shapes.

Compared with the peak area method, a disadvantage of the peak profile-based algorithms is their limited applicability to the isotopologue pairs with retention time shift. Although the retention time shift can be offset computationally, the offset can be incorrectly estimated, which might lead to the error in the abundance ratio estimation.

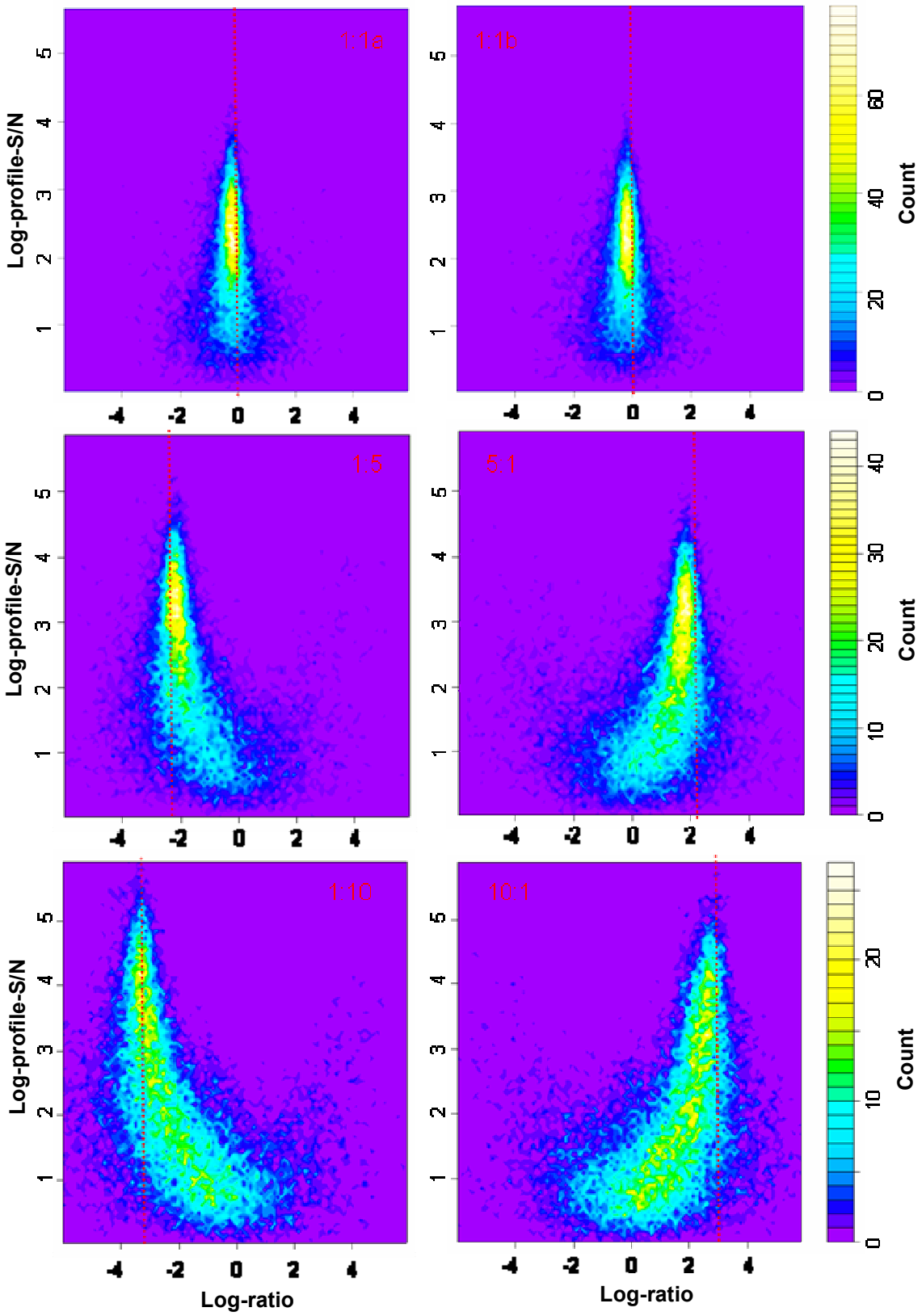
Scoring of Peptide Abundance Ratios for Estimation Variability and Bias

The principal component analysis algorithm scores each estimated peptide abundance ratio with a profile S/N. Note that profile S/N is a signal-to-noise ratio measure of the peak profile, which is different from chromatographic S/N. Chromatographic S/N directly impacts peak area calculation accuracy and, therefore, estimation accuracy of peak area ratio for a peptide should be related to the two chromatographic S/Ns for the light and heavy isotopologues. On the other hand, the principal component analysis algorithm estimates peptide abundance ratios from peak profiles and, therefore, its estimation variability and bias are expected to be directly related to profile S/N.

The peptides were separated into bins by the logarithm base-2 of their profile S/N (log-profile-S/N). The bins were evenly spaced by 0.1 units between log-profile-S/Ns of 0 and 6. The log-ratio distributions were constructed in all log-profile-S/N bins. To present the series of the log-ratio distributions, a two-dimensional heatmap histogram was plotted for each standard mixture dataset (Figure 6.6). Each horizontal band of the two-dimensional histogram represents a log-ratio distribution, in which the peptide count is color-coded. The histograms show that the log-ratio distribution changes in a consistent manner with the log-profile-S/N. At high log-profile-S/N region, the estimated log-ratios tightly cluster close to the $\log_2$ mixing ratios (the dashed red lines in Figure 6.6), indicating accurate and precise log-ratio estimation for peptides with high log-profile-S/N. As the log-profile-S/N decreases, the spread of the log-ratio distribution increases, which suggests the elevating variability of log-ratio estimation.

When the log-profile-S/N decreases below a threshold, the log-ratio distribution gradually regresses away from the log mixing ratio and approaches the log-ratio of zero. This shows the higher bias in the log-ratio estimation for peptides with lower log-profile-S/N. Also note that the log-profile-S/N threshold for the onset of log-ratio estimation bias is higher in the standard mixtures with larger $\log_2$ mixing ratios. This supports the previous observation that the abundance ratio estimation is often biased for the low-concentration peptides with large abundance difference between their isotopologues (Ong, 2003). The less abundant isotopologue often receives a higher percentage of ion intensity from the noise than the more abundant isotopologue. As a result, the abundance

Figure 6.6: Two-dimensional heatmap histograms of log-ratio and log-profile-S/N for the six standard mixture datasets. The color keys for the frequency of peptides are shown in the right of the histograms. The log2 mixing ratios are marked with the red dotted lines. The two-dimensional heatmap histograms show the log-ratio distributions at different log-profile-S/N levels. Log-profile-S/N is inversely related to the spread of the log-ratio distribution and its deviation from the log2 mixing ratio.


1:1a
1:1b
1:5
5:1
1:10
10:1
Log-profile-S/N
Log-ratio
Count

ratio estimate becomes biased towards 1:1. For peptides with log-profile-S/Ns close to zero probably due to the fact that both isotopologues are "buried" in the background noise, the abundance ratios will most likely be estimated as 1:1, regardless of their true value.

The variability of the log-ratio estimation can be measured with the standard deviation of the log-ratio distribution. The standard deviations of the log-ratio distributions were plotted against log-profile-S/N (Figure 6.7A). There is an apparent inverse linear correlation between the standard deviation and the log-profile-S/N. A linear regression model was constructed between the log-profile-S/N and the standard deviation:

$$\sigma = 1.2 - 0.2 \cdot V ,$$

where σ is the standard deviation and V is the log-profile-S/N. The coefficient of determination of the linear regression model was 0.766. The majority of residuals in the linear regression model arise from the stratification of the data points by the mixing ratios.

The log-ratio estimation bias can be quantified with the absolute values of the averages of the log-ratio distributions, which were plotted against the log-profile-S/N of the peptide bin (Figure 6.7B). At the high log-profile-S/N range, the absolute averages are close to, but slightly below, the absolute value of the $\log_2$ mixing ratios of the datasets. As the log-profile-S/N decreases, the average also regresses back to zero. A zero-intercept straight line was fit into the data points on the track of regressing to zero and, in conjunction with the largely unbiased average at the high log-profile-S/N region, a linear regression model can be obtained:

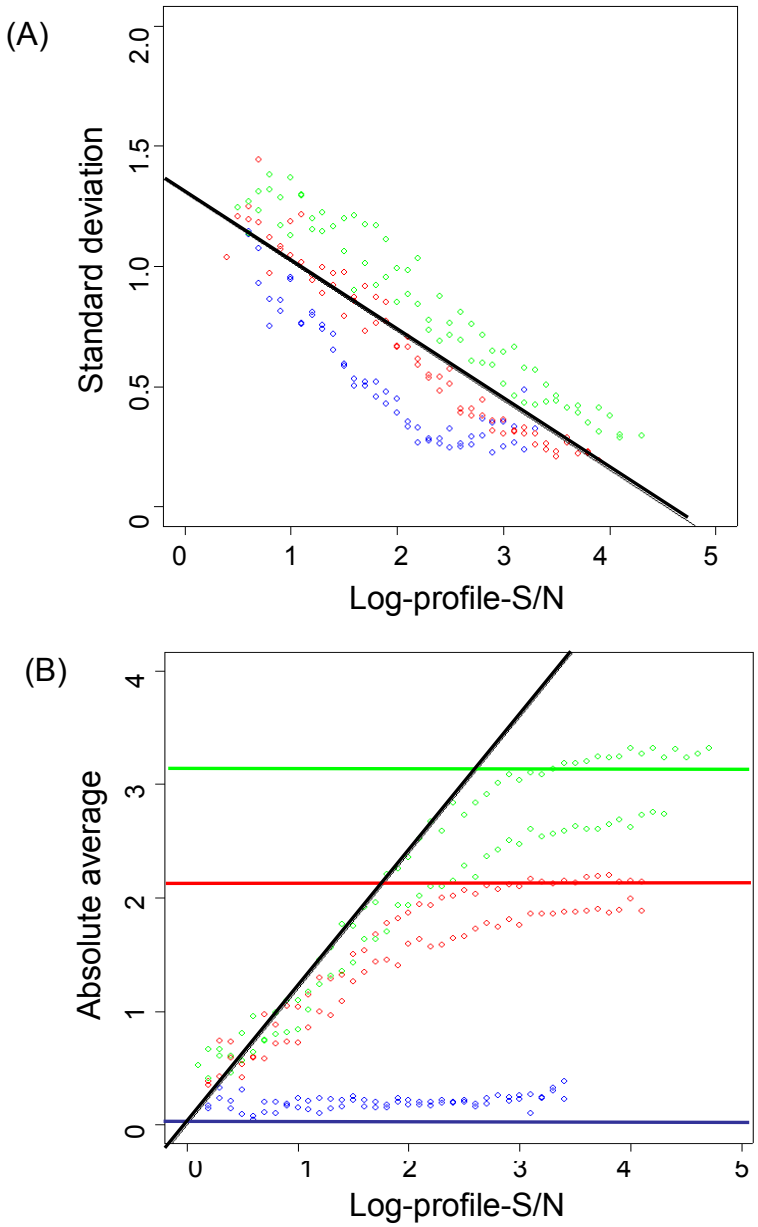


Figure 6.7: Linear models for the standard deviation and absolute average of the log-ratio distribution. Parts A and B show the standard deviations and absolute average of the log-ratio distributions at different log-profile-S/N levels, respectively. The data points from the two 1:1 datasets are shown in blue, those from the 5:1 and 1:5 datasets in red and those from the 10:1 and 1:10 datasets in green. In part A, the standard deviations are modeled with a linear model of log-profile-S/N (the black straight line). In part B, the biased absolute averages are also modeled with a linear model of log-profile-S/N (the black straight line). At the high log-profile-S/N region, the absolute averages are largely consistent with the log2 mixing ratios (marked with the horizontal lines).

$$|\mu| = \begin{cases} 1.2 \cdot V, & 1.2 \cdot V < |H| \\ |H|, & 1.2 \cdot V \geq |H| \end{cases},$$

where |μ| is the absolute value of the average, V is the log-profile-S/N, and |H| is the absolute value of the log mixing ratio or the true log-ratio. The linear regression model of the average shows that when log-profile-S/N is zero, the average is zero; as the log-ratio increases, the average increases proportionally until it reaches the true log-ratio and levels off afterwards.

The bias and the variability of the abundance ratio estimation are indispensable for making the statistical inference on the estimated abundance ratio. In isotope dilution mass spectrometry, the variability and bias of the abundance ratio estimation are determined experimentally by replicate measurements of the sample and calibration with standard solutions (Sargent, 2002). These two experimental routines are of limited practicality in quantitative proteomics, where thousands of peptides are quantified in each experiment. These two linear regression models shown in Figure 6.7 enable the prediction of the variability and the bias of the abundance ratio estimation from the profile S/N for quantitative proteomic experiments. This paves the way for statistically evaluating every peptide abundance ratio using its predicted estimation variability and bias in protein abundance ratio estimation process (Pan, 2006a).

CONCLUSIONS

Here we presented two algorithms for peptide quantification in quantitative shotgun proteomics. The parallel paired covariance algorithm was developed to improve the accuracy of assigning peak boundaries of the chromatographic peaks from a peptide. This algorithm integrates the two selected ion chromatograms into one covariance chromatogram. We showed that covariance chromatograms generally have a better signal-to-noise ratio than either selected ion chromatogram and result in better peak detection accuracy. We then used a principal component analysis algorithm to estimate the peptide abundance ratios from peak profiles. The estimation accuracy was shown to be better than using a peak area algorithm. More importantly, for each peptide abundance ratio estimate, the principal component analysis algorithm provides a signal-to-noise ratio measure of the peak profile, called profile signal-to-noise ratio. The profile signal-to-noise ratio is inversely correlated with the variability and bias of the peptide abundance ratio estimation.

Chapter 7

ProRata: a Quantitative Proteomics Program for Accurate Protein Abundance Ratio Estimation with Confidence Interval Evaluation

All of the data presented below has been published as

C. Pan, G. Kora, W.H. McDonald, D.L. Tabb, N.C. VerBerkmoes, G.B. Hurst, D.A. Pelletier, N.F. Samatova, and R.L. Hettich. ProRata: A quantitative proteomics program for accurate protein abundance ratio estimation with confidence interval evaluation. *Analytical Chemistry.* 2006 (In press)

As first author, C. Pan's contributions to this article include algorithm development and MS data acquisition and interpretation

INTRODUCTION

Organisms often respond to environmental or physiological stimuli by adjusting the type and abundance of proteins in their cells. Measurement of the relative abundances of proteins in treatment cells subjected to stimuli, compared to that in the reference cells, provides valuable insights about protein function and regulation. Quantitative shotgun proteomics has recently emerged as a high-throughput technique for measuring the relative abundances of thousands of proteins between two cellular conditions (Ong, 2005a). The reference and treatment proteomes are labeled with different stable isotope tags (Gygi, 1999; Oda, 1999; Yao, 2001; Ong, 2002) and then mixed in equivalent amounts. In such a proteome mixture, each protein has two mass-different isotopic variants: the *light isotopologue* and the *heavy isotopologue* (Muller, 1994), or a *protein isotopologue pair*. Finally, the proteome mixture is digested and then analyzed with

liquid chromatography–tandem mass spectrometry (LC–MS/MS) (Link, 1999). The proteolysis turns each protein isotopologue pair into multiple peptide isotopologue pairs. Each of the peptide isotopologue pairs is expected to have the same abundance ratio as the protein isotopologue pair. Although multiplication of isotopologue pairs in the mixture by proteolysis increases the complexity of the sample for LC-MS/MS analysis, it provides multiple indirect measurements of a protein's abundance ratio, derived from the abundance ratios of its peptide isotopologue pairs.

To evaluate protein abundance ratios from quantitative proteomics measurements, two types of statistical estimation should be employed: *point estimation* and *interval estimation*. The point estimation gives an abundance ratio for every quantified protein, which "best" approximates the true abundance ratio. Unfortunately, the point estimation provides no information about protein quantification precision, which can significantly vary across different proteins. Generally, a protein should have better quantification precision if it has more proteolytic peptides quantified from mass spectral data of higher signal-to-noise ratio. It is misleading in quantitative proteomics to treat all proteins' abundance ratios identically, regardless of their estimation precision.

The interval estimation complements the point estimation by providing *confidence intervals* for protein abundance ratios. If 90% of quantified proteins have confidence intervals that contain their true abundance ratios, then confidence intervals are estimated at 90% *confidence level*. The confidence level for the interval estimation in quantitative proteomics is analogous to the true positive rate for protein identification in qualitative

proteomics. More importantly, at a given confidence level, the confidence interval intuitively reflects the quantification precision for each protein as an "error bar" of the abundance ratio estimate.

In quantitative shotgun proteomics, each protein's abundance ratio is estimated by "combining" multiple peptide abundance ratios measured with LC-MS/MS. Several computer programs have been developed for the point estimation of protein abundance ratios, including XPRESS (Han, 2001), ASAPratio (Li, 2003), MSQuant (Schulze, 2004), and RelEx (MacCoss, 2003). The first three programs calculate peptide abundance ratios from ratios of peak areas in selected ion chromatograms. RelEx, on the other hand, estimates peptide abundance ratios using a linear correlation algorithm, which reduces the peptide abundance ratio estimation error. The average of peptide abundance ratios is then used by RelEx, XPRESS and MSQuant to estimate protein abundance ratios. To improve the accuracy of protein abundance ratio estimation, ASAPratio weighs the peptide abundance ratios with their peak area and uses the weighted average to estimate the protein abundance ratios. In these programs, the standard deviation of the peptide abundance ratios is used as a measure of the variability of the protein abundance ratio estimation. However, without assuming the normality of the peptide abundance ratio distribution, the standard deviation is not directly related to the confidence interval of the protein abundance ratio.

In the previous chapter, we scored every peptide's abundance ratio with a *profile signal-to-noise ratio* for the peptide's mass spectral data (Pan, 2006b). It was observed that the

estimation variability and bias of peptide abundance ratios in $\log_2$ scale (*peptide log-ratios*) were linearly correlated with profile signal-to-noise ratios in $\log_2$ scale (*log-profile-S/N*s). This was illustrated with a standard mixture of isotopically labeled proteomes, in which all peptides were expected to have an abundance ratio of 1:5 or a log-ratio of –2.3. The two-dimensional heatmap histogram of peptide log-ratio versus log-profile-S/N shows a comet-like distribution from this standard mixture (Figure 7.1). In the high log-profile-S/N region, the peptide log-ratio distributions are tight with a center at the expected log-ratio. As the log-profile-S/N level decreases, the horizontal spread of the log-ratio distributions increases, indicating the elevation of variability in peptide log-ratio estimation. At the same time, the distributions also deviate more from the true log-ratio, indicating the increase of bias in peptide log-ratio estimation. The change of variability and bias of the log-ratio distributions with log-profile-S/N makes it unsuitable to treat peptides with different log-profile-S/N identically and to assume normality for the aggregated peptide log-ratio distribution.

In this chapter, we describe a profile likelihood algorithm that yields both maximum likelihood point estimation (Eliason, 1993) and profile likelihood confidence interval estimation (Venzon, 1988) of protein abundance ratios. This likelihood-based approach allows us take into account the changing estimation variability and bias of peptide abundance ratios in the process of protein abundance ratio estimation. This improves the accuracy of the point estimation and the precision and confidence level of the interval estimation, as benchmarked with standard mixtures of isotopically labeled proteomes.

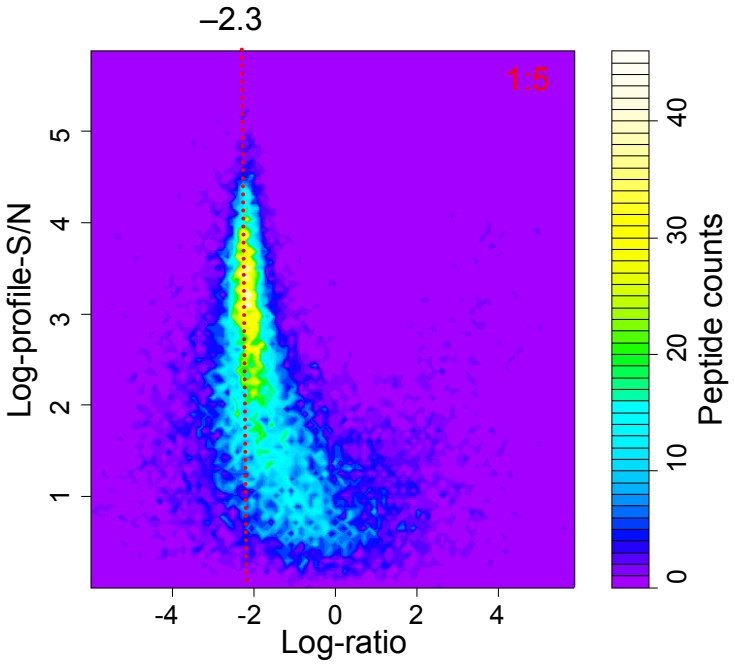


Figure 7.1: A two-dimensional heatmap histogram of peptide log-ratio versus log-profile-S/N. The color-scale (shown on the right) represents the number of peptides at a given log-ratio and log-profile-S/N location. The expected log-ratio for all peptides in this 1:5 standard mixture is –2.3, indicated with the red dotted vertical line. As the log-profile-S/N level lowers, the horizontal distribution of the peptide log-ratios spreads wider and deviates more from the line of expected log-ratio.

The profile likelihood algorithm is part of a computer program called *ProRata*, which automates the entire data analysis process for quantitative shotgun proteomics. ProRata is applicable to a variety of stable-isotope labeling techniques, including $^{15}N/^{13}C$ metabolic labeling (Oda, 1999), SILAC (Ong, 2002), ICAT (Gygi, 1999) and $H_2{}^{18}O$ proteolysis (Yao, 2001). Figure 7.2 shows the data analysis flowchart of ProRata. ProRata extracts selected ion chromatograms for peptide isotopologue pairs and detects their chromatographic peaks with a parallel paired covariance algorithm (Pan, 2006b). Principal component analysis is then used to calculate peptide abundance ratios and profile signal-to-noise ratios. Finally, protein abundance ratios and their confidence intervals are estimated with the profile likelihood algorithm.

MATERIALS AND METHODS

Chemicals and Reagents.

HPLC-grade water and acetonitrile were obtained from Burdick & Jackson (Muskegon, MI), and the 98% formic acid from EM Science (Darmstadt, Germany). All other chemicals were purchased from Sigma-Aldrich (St. Louis, MO) unless noted otherwise.

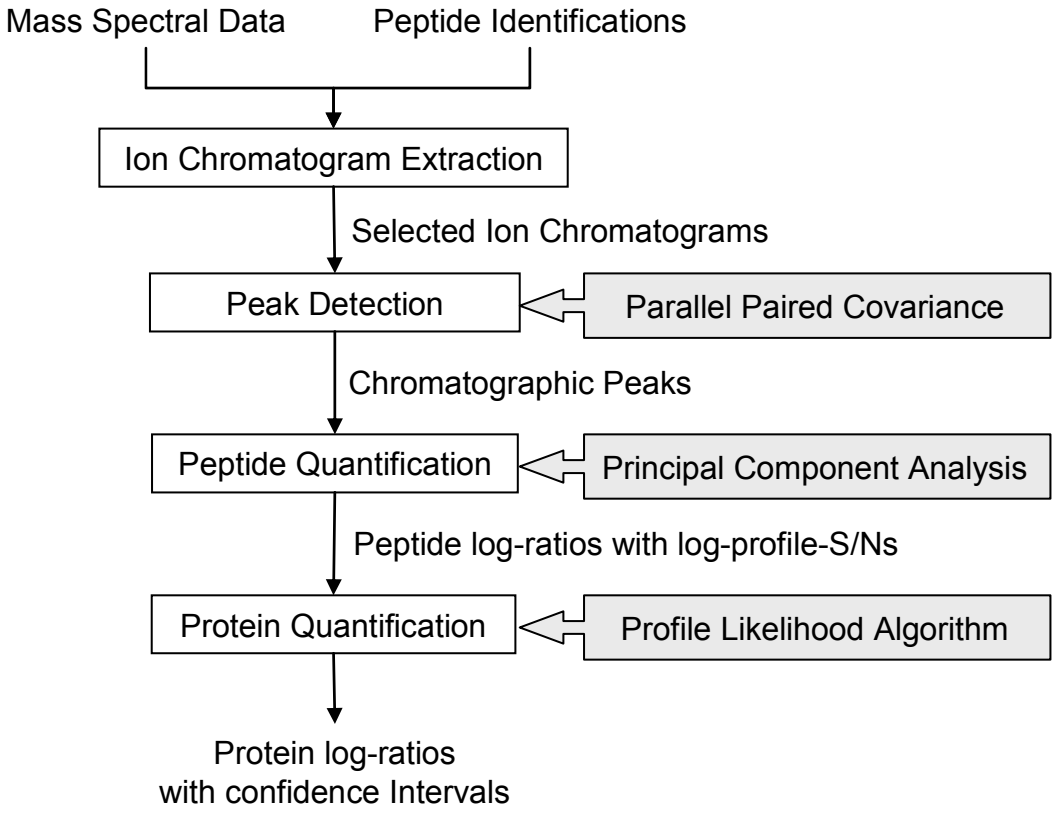


Figure 7.2: Data processing flowchart of ProRata. ProRata consists of four modules, shown as blocks. The algorithm used in each module is specified on the right. The data flow from one module to the next is shown with the solid arrows, starting from the input data: mass spectral data and peptide identification results.

Standard Isotopically Labeled Proteome Mixture Preparation.

Rhodopseudomonas palustris CGA0010 strain was grown anaerobically in light at 30°C to mid-log phase. The defined minimal growth media supplies $(NH_4)_2SO_4$ as the only nitrogen source for bacterial growth. The unlabeled culture was grown with $(^{14}NH_4)_2SO_4$. The ^{15}N-labeled culture was grown identically with $(^{15}NH_4)_2SO_4$ (>98 atom percentage excess, from Sigma-Aldrich, St. Louis, MO). The ^{14}N proteome and the ^{15}N proteome were prepared from the unlabeled culture and the ^{15}N-labeled culture, respectively, as described in Chapter 6. The total protein concentration of the two proteomes was quantified with Lowry's analysis (Lowry, 1951). Six standard mixtures were prepared by mixing the ^{14}N proteome and the ^{15}N proteome at the ratios of 10:1, 5:1, 1:1, 1:5, and 1:10 by total protein mass. An aliquot of the ^{14}N proteome was also retained for shotgun proteomics measurement.

Shotgun Proteomics Measurements.

The proteome samples were processed by the described procedure (Pan, 2006b). Briefly, after disulfide bond reduction and protein denaturation with 10 mM DTT and 6 M guanidine, the proteome samples were digested with sequencing grade trypsin (Promega, Madison, WI). The samples were then treated with 20 mM DTT for 1 hour at 60 °C as a final reduction step. The samples were immediately desalted with Sep-Pak Plus C18 solid-phase extraction (Waters, Milford, MA) and solvent exchanged into 0.1% formic acid in water by centrifugal evaporation. The protein digests were examined with the

twelve-step split-phase MudPIT technique, as described previously (MacCoss, 2002; McDonald, 2002). Briefly, the samples were first separated by twelve-step strong cation ion exchange liquid chromatography and then by two-hour continuous gradient reverse phase liquid chromatography. Eluted peptides were electrosprayed at 2-kV distal electrospray voltage into an LTQ mass spectrometer (Thermo Finnigan, San Jose, CA). Tandem mass spectrometry analysis was performed with each full scan (400-1700 *m/z*) followed by five data-dependent MS/MS scans at 35% normalized collision energy. Dynamic exclusion was enabled. All scans were averaged from two microscans.

Peptide and Protein Identification.

All MS/MS scans were searched in two iterations against a FASTA database containing all annotated *Rhodopseudomonas palustris* proteins (Larimer, 2004) using the SEQUEST program (Eng, 1994). In the first iteration, the unmodified amino acids were used and in the second iteration the ^{15}N-labeled amino acids were used. The peptide identifications from the two iterations were merged. The DTASelect program (Tabb, 2002) was used to filter the peptide identifications and to assemble the peptides into proteins using the following parameters: retain the duplicate MS/MS scans for each peptide sequence (DTASelect option: `-t 0`), consider fully tryptic peptides only, filter with a delCN of at least 0.08 and cross-correlation scores (Xcorrs) of at least 1.8 (+1), 2.5 (+2), and 3.5 (+3).

Ion Chromatogram Extraction.

For each identified peptide, two selected ion chromatograms were extracted for the two peptide isotopologues. The *m/z* window for the light isotopologue was calculated from the natural isotopic envelope of the peptide. The *m/z* window for the heavy isotopologue was calculated by using 98%-enriched ^{15}N for all nitrogen atoms. The retention time window of the selected ion chromatograms was defined as from 2 minutes before the MS/MS scans to 2 minutes after the MS/MS scans.

Peptide Abundance Ratio Estimation.

The chromatographic peaks of the peptide isotopologue pairs were detected with a parallel paired covariance algorithm. The abundance ratio and the profile S/N of a peptide were calculated by analyzing its peak profile with principal component analysis as described in Chapter 6. Briefly, the peak profile was constructed as a scatter-plot with ion intensities of the two isotopologues as its coordinates (Lawson, 1980). Principal component analysis of the peak profile generated two principal components and their associated eigenvalues. The peptide abundance ratio was estimated as the slope of the first principal component. The profile S/N was calculated as the square root of the ratio between the first eigenvalue and the second eigenvalue. Peptides with a profile S/N below 2.0 were removed, due to the large estimation error of their abundance ratios. Peptides shared among multiple proteins were also discarded, because the abundance

ratio of a shared peptide will be a weighted average of the abundance ratios of multiple proteins and thus cannot be used for any of the individual proteins.

Protein Abundance Ratio Estimation.

Quantified peptides were assembled into proteins. Proteins with more than two quantified peptides were selected for abundance ratio estimation. For the point estimation and confidence interval estimation, the profile likelihood algorithm solves a likelihood function of protein log-ratio with a numerical method, in the absence of an analytical method. The profile likelihood algorithm has three steps:

1. *Log-ratio enumeration step:* Enumerate all protein log-ratios to be considered through discretization of a continuous log-ratio interval.
2. *Likelihood calculation step*: Calculate the likelihood for each considered log-ratio to be the true log-ratio of a protein, given the abundance ratios and profile S/Ns of multiple peptides from this protein;
3. *Point and interval estimation step*: Select a log-ratio with the maximum likelihood as the maximum likelihood estimate; select two log-ratios on a likelihood threshold as the lower and upper limits of profile likelihood confidence interval.

Below we describe the three steps in detail and the rationale for the procedures.

Log-ratio enumeration step: The continuous interval of protein log-ratio, [–7.0, 7.0], is discretized at the precision of 0.1 to create a discrete set of protein log-ratios, i.e. –7.0, –

6.9, –6.8, … 6.8, 6.9, and 7.0. This set, denoted as *G*, enumerates all protein log-ratios that the profile likelihood algorithm will consider for every protein. The minimum and maximum protein log-ratios and the discretization precision are configurable in ProRata. The maximum protein log-ratio of 7 and the minimum protein log-ratio of –7 correspond to 128-fold up-regulation and 128-fold down-regulation in protein abundance, respectively. These maximum and minimum log-ratios sufficiently encompass the practical dynamic range of our instruments for quantification. The discretization precision of 0.1 is also the protein quantification precision that can be realistically achieved in quantitative proteomics measurements. The discretization of protein log-ratio solution space allows for an efficient and correct numerical solution of the likelihood function of protein log-ratio by brute force enumeration.

Likelihood calculation step: The likelihood for each protein log-ratio in *G* to be the true log-ratio of a protein is calculated. Assume that the protein has *n* peptides with log-ratios of R_1, R_2 … R_n and log-profile-S/Ns of V_1, V_2 … V_n. Let *H* be an arbitrary log-ratio from *G.* Then the likelihood for *H* to be the true log-ratio of a protein given the log-ratios and the log-profile-S/Ns of its peptides equals the probability of observing these peptide log-ratios given their log-profile-S/Ns and the protein log-ratio of *H*:

$$\mathrm{L}\left(H \mid R_1, R_2 \cdots R_n, V_1, V_2 \cdots V_n\right) = \mathrm{P}\left(R_1, R_2 \cdots R_n \mid V_1, V_2 \cdots V_n, H\right), \qquad (7.1)$$

where L() is the likelihood function and P() is the probability function. Assume that the protein's peptides are measured independently. Then the probability of observing these *n*

peptides together is the product of individual probabilities of observing each of these peptides independently:

$$\begin{aligned} \mathrm{P}(R_1, R_2 \cdots R_n \mid V_1, V_2 \cdots V_n, H) &= \mathrm{P}(R_1 \mid V_1, H) \cdot \mathrm{P}(R_2 \mid V_2, H) \cdots \mathrm{P}(R_n \mid V_n, H) \\ &= \prod_{i=1}^{n} \mathrm{P}(R_i \mid V_i, H). \end{aligned} \tag{7.2}$$

Theoretically, the log-ratio of a peptide is expected to be equal to the log-ratio of its protein. However, due to the changing variability and bias of peptide log-ratio estimation, the probability distribution of peptide log-ratio is modeled with a mixture model of normal distribution and uniform distribution:

$$\mathrm{P}(\mathrm{R_i} \mid \mathrm{V_i}, \mathrm{H}) = 85\% \cdot \mathrm{P_{normal}}(\mathrm{R_i} \mid \mu_\mathrm{i}, \sigma_\mathrm{i}) + 15\% \cdot \mathrm{P_{uniform}}, \tag{7.3}$$

where $\mathrm{P_{normal}}()$ is a normal probability function and $\mathrm{P_{uniform}}()$ is a uniform probability function. The mixture model is employed, because the uniform distribution with a weight of 15% models approximately 15% of outlier peptides that are not well captured by the normal distribution.

The absolute value of the mean ($|\mu_\mathrm{i}|$) and the standard deviation (σ_i) of the normal distribution in the mixture model are approximated with two linear functions of the log-profile-S/N (V_i) and the protein log-ratio (H):

$$|\mu_\mathrm{i}| = \begin{cases} 1.2 \cdot V_i, & 1.2 \cdot V_i < |H|, \\ |H|, & 1.2 \cdot V_i \geq |H|, \end{cases} \tag{7.4}$$

$$\sigma_\mathrm{i} = 1.2 - 0.2 \cdot V_i \tag{7.5}$$

The sign of the mean (μ_i) is the same as that of the protein log-ratio (H). The two linear functions and their coefficients were estimated from experimental data as described in Chapter 6. Briefly, two-dimensional heatmap histograms were constructed for six different standard mixtures of isotopically labeled proteomes as shown in Figure 7.1. The means and standard deviations of the peptide log-ratio distributions at different log-profile-S/N levels were calculated, from which the two linear functions were then derived. The absolute mean linear function (Equation 7.4) captures that the bias of peptide log-ratio estimation deceases to zero with an increase of log-profile-S/N. The negative slope of the standard deviation linear function (Equation 7.5) models the decrease of peptide log-ratio estimation variability with an increase of log-profile-S/N.

In summary, the likelihood function of protein log-ratio is constructed as:

$$\begin{aligned} \mathrm{L}\left(H \mid R_1, R_2 \cdots R_n, V_1, V_2 \cdots V_n\right) &= \prod_{i=1}^{n} \mathrm{P}(\mathrm{R_i} \mid \mathrm{V_i}, \mathrm{H}) \\ &= \prod_{i=1}^{n} (0.85 \cdot \mathrm{P}_{\mathrm{normal}}(\mathrm{R_i} \mid \mu_i, \sigma_i) + 0.15 \cdot \mathrm{P}_{\mathrm{uniform}}) \end{aligned} \tag{7.6}.$$

The likelihood H is calculated for each protein log-ratio in the set G and transformed to the natural logarithm scale, denoted as *ln-likelihood*. The calculation result of the likelihood function can be represented graphically with *profile likelihood curve*. Let the x-axis and the y-axis be log-ratio and ln-likelihood, respectively, and plot all considered protein log-ratios with their ln-likelihood as points. The profile likelihood curve is constructed by connecting the points adjacent along the log-ratio axis (the blue curves in Figure 7.3). For manual data analysis, the data points representing the peptides of a

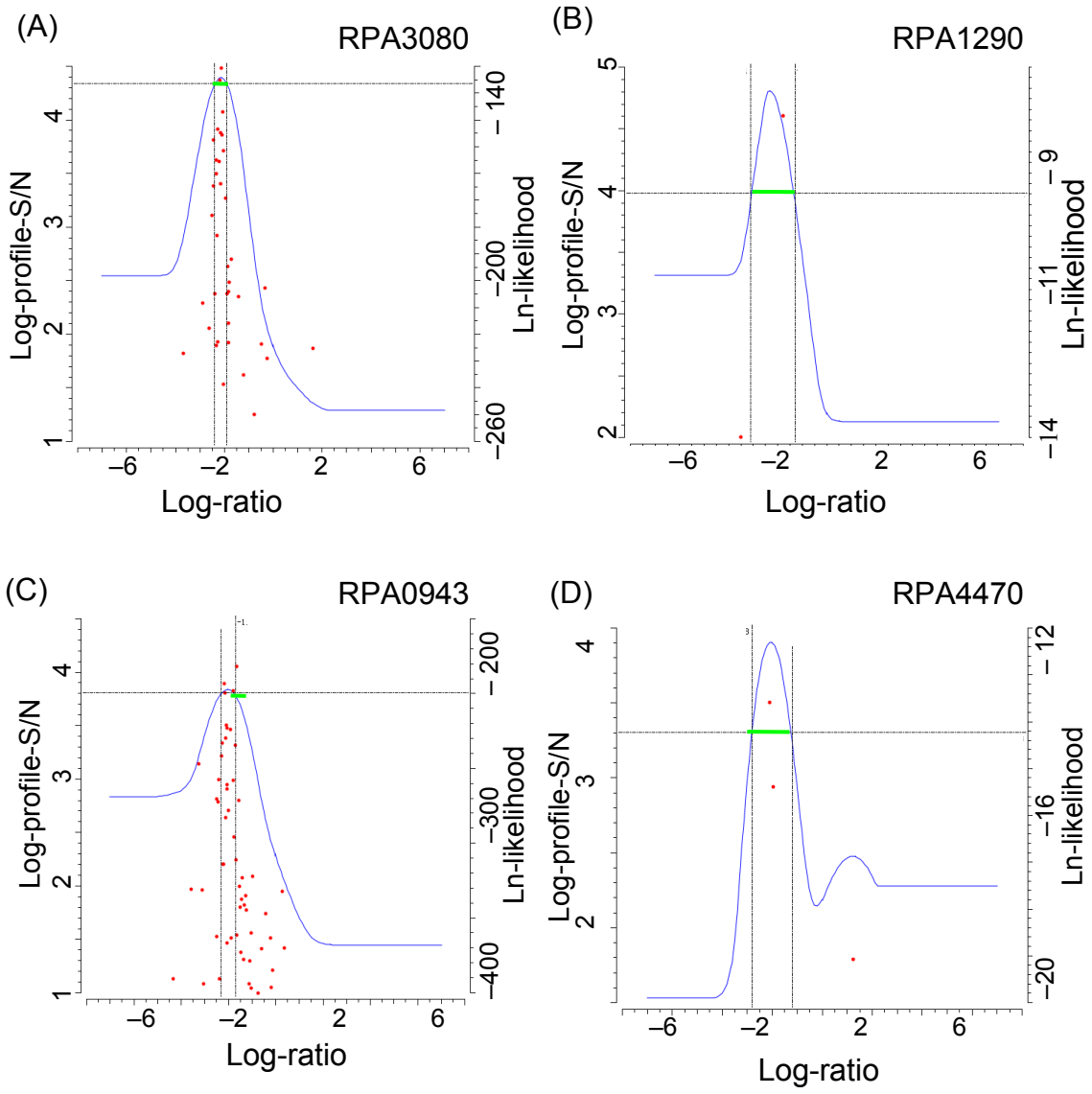


Figure 7.3: Estimation of protein log-ratios with profile likelihood curves. All four proteins (locus shown in the upper right corner) are expected to have a log-ratio of –2.3. Profile likelihood curves (the blue curves) plot the ln-likelihood (y-axis on the right) for the log-ratios (x-axis) of proteins. Note the different ln-likelihood ranges for different proteins. Peptide data points (the red dots) represent the log-ratio (x-axis) and the log-profile-S/N (y-axis on the left) of the quantified peptides. The confidence intervals are shown as the green bars.

protein (the red points in Figure 7.3) are overlaid with the protein's profile likelihood curve.

Point and interval estimation step: Both maximum likelihood estimate and profile likelihood confidence interval of a protein's log-ratio are estimated from the protein's profile likelihood curve. The maximum likelihood estimate of protein log-ratio is the log-ratio with the maximum likelihood. The confidence interval is calculated by assuming Chi-square distribution for the likelihood ratio test as described in the standard methodology for profile likelihood confidence interval estimation (Venzon DJ, 1988). Let L_{max} be the maximum likelihood. The confidence interval with a nominal confidence level of $(1 - \alpha) \cdot 100\%$ includes all the log-ratios that have a ln-likelihood exceeding the threshold of $\ln(L_{max}) - 0.5 \cdot \chi^2_{1,\alpha}$. The ln-likelihood threshold is $\ln(L_{max}) - 1.96$ for the confidence interval of 95% nominal confidence level ($0.5 \cdot \chi^2_{1,\ 0.05} = 1.96$). The lower and upper limits of the confidence interval are the minimum and maximum log-ratios with ln-likelihood exceeding the threshold, respectively. We discarded proteins with confidence intervals that are wider than 7.

Protein abundance ratios were also estimated with the RelEx program (MacCoss, 2003) for comparison purpose. The DTASelect result files and the Xcalibur data files were input to the RelEx program. Data smoothing, ratio correction and chromatogram filtering were enabled with the default settings. The protein filter of minimum peptide number of two was also applied. RelEx was executed in two iterations, one using the light isotopologue

for peak detection and the other using the heavy isotopologue. The results of the two iterations are combined. The abundance ratios of the proteins quantified in both iterations are assigned as the average of the abundance ratios estimated in the two iterations.

Software Development.

ProRata was written in C++ and compiled with the MinGW g++ compiler. The graphical user interface was implemented using Qt library. ProRata used the RAMP (random access minimal parser) library to access mzXML files. The histograms were constructed with R scripts.

RESULTS AND DISCUSSION

The features of the profile likelihood algorithm were evaluated initially with individual proteins. Then the aggregate performance metrics were determined, including the accuracy of the point estimation as well as the confidence level and median width of the confidence interval estimation. Finally ProRata was equipped with a graphical user interface to enable manual interrogation of the quantification result for any given protein. Note the following abbreviations used in this study: (a) *log-ratio* for the abundance ratio in $\log_2$ scale, (b) *log-profile-S/N* for the profile S/N in $\log_2$ scale, and (c) *ln-likelihood* for the natural logarithm of the likelihood for a log-ratio to be true for a protein. The $\log_2$ transformation for the abundance ratio is to treat up- and down-regulation of protein abundance symmetrically and to replace a multiplication operation on the abundance

ratios with the addition operation on the log-ratios.

Point Estimation and Confidence Interval Estimation of Protein Abundance Ratios with Profile Likelihood Curves

The maximum likelihood point estimation and the profile likelihood confidence interval estimation of a protein's log-ratio is based on the protein's profile likelihood curve, which is constructed from the quantified peptides of the protein. Figure 7.3 shows profile likelihood curves (blue curves) together with peptide data points (red points) for four proteins in the 1:5 standard mixtures.

The maximum likelihood point estimate is the protein log-ratio with the maximum likelihood, i.e. the log-ratio position of the highest peak in the profile likelihood curve. Conceptually, the sharper the peak is, the more precise the maximum likelihood estimate is, and, therefore, the narrower the confidence interval should be. This intuition is captured by how the profile likelihood confidence interval is estimated. The confidence interval at the nominal confidence level of 95% is the log-ratio range of the curve segment above the ln-likelihood threshold (the horizontal lines shown in Figure 7.3) at 1.96 units ($0.5 \cdot \chi^2_{1,\ 0.05}$) below the peak top. A profile likelihood confidence interval can be asymmetric, with different distances from the lower and upper interval limits to the point estimate.

The shape of the profile likelihood curve is determined by peptide data points in a number of ways, as illustrated in Figure 7.3 showing four proteins with an expected log-ratio of –2.3. First, a profile likelihood curve forms a peak at the log-ratio location with largest density of peptide data points of high log-profile-S/N. To illustrate this, Figure 7.3A shows the profile likelihood curve of 50S ribosomal protein L9 (Locus: RPA3080). This protein has many peptides with high log-profile-S/Ns and consistent log-ratios. This leads to a high and sharp profile likelihood peak in the ln-likelihood range of [–260, –140] and a narrow confidence interval of [–2.4, –1.9].

Second, the log-ratio location of a profile likelihood peak is largely determined by the peptide data points with higher log-profile-S/N. Figure 7.3B shows a putative oxidoreductase (Locus: RPA1290) with only two quantified peptides. The peptide with higher log-profile-S/N has a log-ratio of –1.8 and the other peptide has a log-ratio of –3.5. The log-ratio position of the profile likelihood peak top, i.e. the maximum likelihood estimate, is –2.3, which is closer to the log-ratio of the peptide with higher log-profile-S/N.

Third, the protein log-ratio estimation accounts for the log-ratio estimation bias in peptides with low log-profile-S/N. Figure 7.3C shows a phosphoglycerate kinase (Locus: RPA0943) that has a large fraction of peptide data points with poor log-profile-S/Ns and log-ratios considerably biased towards 0. The prevalence of biased peptide log-ratios in the low log-profile-S/N region is also shown in Figure 7.1. A simple average of all

peptide log-ratios would give a biased estimation of the protein log-ratio. In contrast, the profile likelihood peak is located in the log-ratio region containing the peptides with high log-profile-S/N, which yields an unbiased estimation of protein log-ratio. The peptides with low log-profile-S/N were used to suppress the ln-likelihood of the protein log-ratios on the right side of the biased peptide log-ratios.

Fourth, the profile likelihood peak excludes the peptide data points that are outliers in the log-ratio axis. Figure 7.3D shows the profile likelihood curve for a hypothetical protein (Locus: RPA4470). Only three peptides are quantified and one of them is likely to be an outlier with an erroneous log-ratio. The outlier creates a small profile likelihood peak, but has no effect on the large profile likelihood peak used for protein log-ratio estimation.

In summary, the point estimate and the confidence interval of a protein log-ratio is calculated with peptide log-ratio weighting, bias suppression, and outlier exclusion. All of these are achieved using the likelihood function of protein log-ratio (Equation 7.6). A weakness of this algorithm is the need to set the following parameters in the likelihood function: (a) the proportion between the normal and uniform distributions in the mixture model (Equation 7.3), which sets the tolerance to outliers modeled by the uniform distribution; and (b) the parameters in the linear models for inferring the standard deviation and the mean of peptide log-ratio distributions (Equations 7.4 and 7.5), which set the relative weights and the expected biases of peptide log-ratios with their log-profile-S/Ns. These parameters were estimated from experimental data as described in

Chapter 6. This weakness might be alleviated in the future by improving the likelihood function of our algorithm or by employing other related algorithms from data fusion (Hall, 2004), pattern recognition (Marques, 2001), data mining (Larose, 2006), *etc*.

Benchmark of Protein Abundance Ratio Estimation Performance using Standard Mixtures of Isotopically Labeled Proteomes

The profile likelihood algorithm was tested as a part of the program ProRata using the standard mixtures. A widely recognized challenge in proteomics is the enormous dynamic range between different proteins. Quantitative proteomics presents another type of the dynamic range challenge: the potentially large abundance difference between the two isotopologues of a protein. We refer to the former dynamic range as *protein dynamic range* and to the latter one as *isotopologue dynamic range*. Various standard mixtures of metabolically labeled *R. palustris* proteomes were prepared with different mixing ratios, which represent the isotopologue dynamic range. The following five mixing ratios between the ^{14}N and ^{15}N proteomes were used in this study: 10:1, 10:1, 5:1, 1:5, and 1:1. The 1:1 mixture was analyzed in duplicate, and the two data sets are designated as 1:1a and 1:1b. The protein quantification results are summarized in Table 7.1.

On average, 1,362 proteins were identified in a standard mixture (Table 7.1 Approximately 200 fewer proteins were identified from these standard mixtures than from the proteome sample before mixing. This reduction is probably because the mixing essentially doubled the sample complexity. Full scans from the standard mixtures contain

Table 7.1 Summary of protein quantification results from the standard mixtures of isotopically labeled proteomes.

Standard mixture		Protein count		Log-ratio point estimation		Confidence interval estimation		Hypothesis testing	
^{14}N:^{15}N	Log-ratio	Identified	Quantified	Median	AAD *	Median width	Confidence level	Significance	Power
1:1a	0.0	1,392	1,117	-0.2	0.318	1.4	93%	8%	---
1:1b	0.0	1,348	1,071	-0.2	0.361	1.4	92%	10%	---
5:1	2.3	1,384	1,054	1.8	0.481	1.4	90%	---	94%
1:5	-2.3	1,263	1,024	-2.1	0.390	1.4	92%	---	97%
10:1	3.3	1,475	1,096	2.5	0.561	1.6	88%	---	96%
1:10	-3.3	1,312	1,000	-3.1	0.639	1.6	87%	---	98%
Average		1,362	1,060	---	0.458	1.5	90%	9%	96%

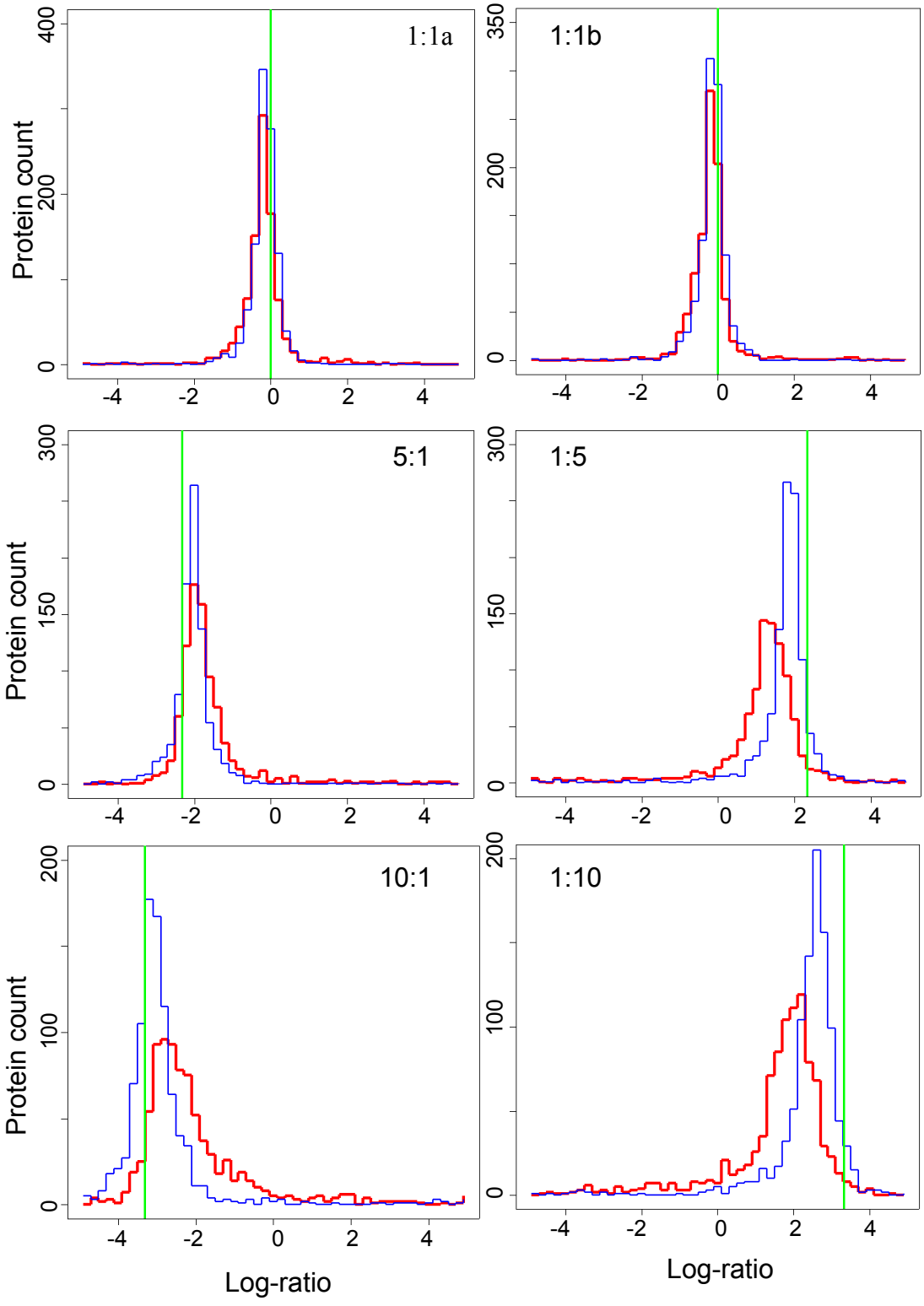
* AAD: Absolute average deviation from the median

"doublet" peaks from the two isotopologues of peptides. Many MS/MS scans were targeted to different isotopologues of the same peptide, rather than to new peptides. On average, 1,060 proteins were quantified out of the 1,362 identified proteins. Not every identified protein can be quantified, as quantification of a protein requires at least two quantified peptides with relatively high profile S/Ns and consistent log-ratios.

We compared RelEx's and ProRata's point estimation of protein abundance ratios. RelEx was used for comparison, because both RelEx and ProRata take the identification results from the DTASelect program and they employ a similar strategy for calculating peptide abundance ratios. RelEx, unlike ProRata, uses the average of peptide abundance ratios as the protein abundance ratio estimate. To evaluate the protein quantification results, the histogram of protein log-ratio estimates were constructed for each standard mixture. The protein log-ratio estimates should have the same true value in a standard mixture. Hence, the spread of the log-ratio distribution would reflect the random estimation error and the difference between the distribution center and the true log-ratio would reflect the systematic estimation error.

The protein log-ratio distributions for the two 1:1 standard mixtures are similar between ProRata (blue) and RelEx (red) (Figure 7.4, top). For the other mixtures, the protein log-ratio distributions from ProRata (blue) are closer to the true log-ratio (green line) and tighter than those from RelEx (red) (Figure 7.4, middle and bottom). This indicates that the profile likelihood algorithm gives a more accurate and precise point estimation of protein log-ratio than averaging. The median and the average absolute deviation of

Figure 7.4: Comparison of protein log-ratio point estimation with RelEx and ProRata. The histograms of the protein log-ratios estimated with the two programs (blue for ProRata and red for RelEx) are constructed for six standard mixtures. The mixing ratios in log2 scale are represented by the vertical green lines.

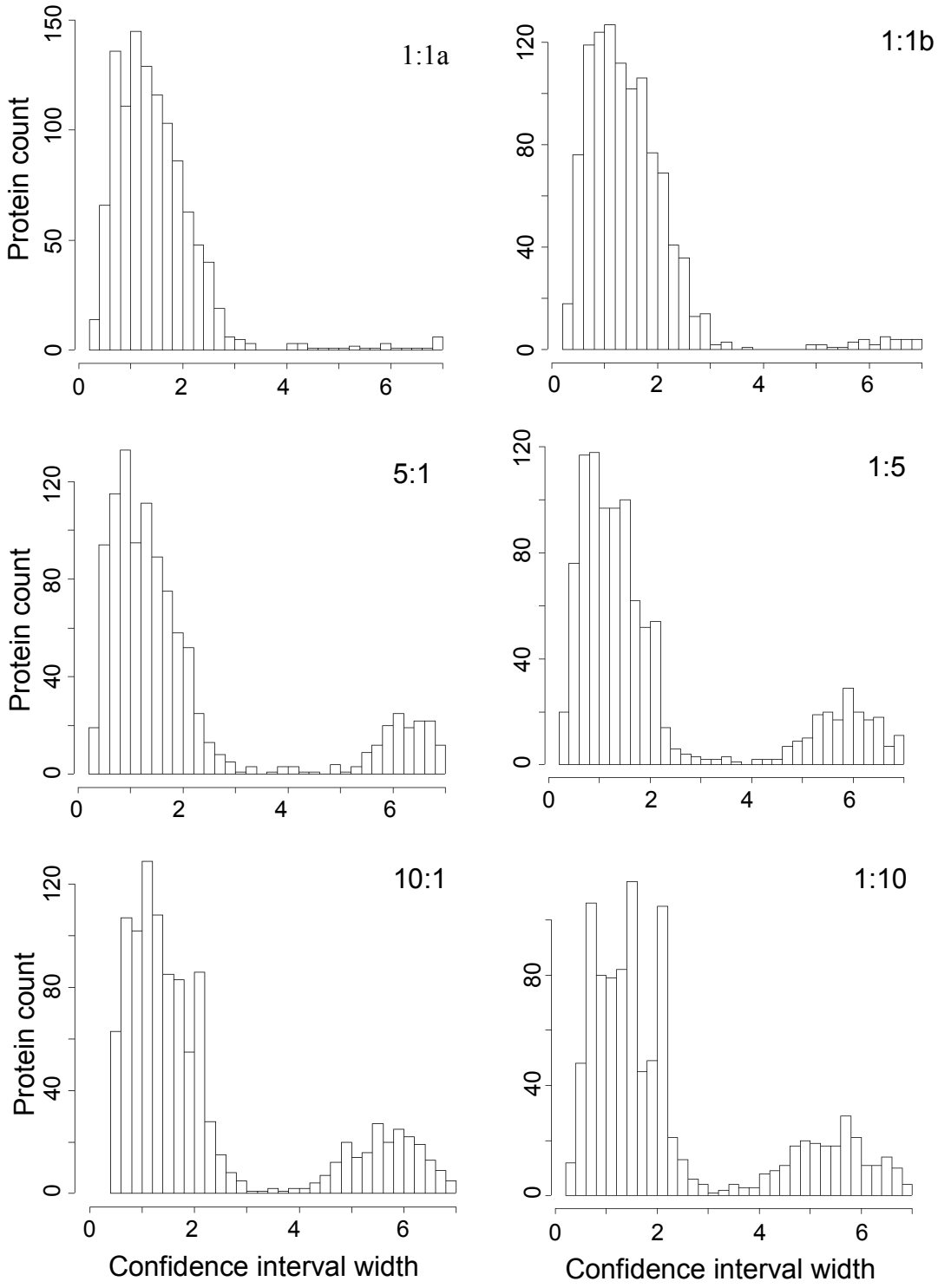

1:1a
1:1b
5:1
1:5
10:1
1:10
Protein count
Log-ratio

ProRata's point estimation are shown in Table 7.1. The average absolute deviation is the average difference of the point estimates from their median, which indicates the spread of the distribution.

The profile likelihood confidence intervals were also estimated for the quantified proteins. The confidence interval width for the majority of proteins was distributed between 0 and 3 (Figure 7.5). We observed that confidence intervals are generally smaller for high-abundance proteins, such as ribosomal proteins, than for low-abundance proteins, such as DNA polymerase proteins. In the 5:1, 1:5, 1:10 and 10:1 standard mixtures, there was a distinct, small distribution of confidence intervals that are wider than 4. This distribution largely stems from highly asymmetric confidence intervals that have a lower limit extending to the minimum log-ratio of –7.0 or an upper limit to the maximum log-ratio of 7.0. For example, consider a dehydrogenase protein (Locus: RPA4259) in the 10:1 standard mixture. The point estimate for this protein's log-ratio was 2.1 and the confidence interval was [1.0, 7.0]. The profile likelihood algorithm only determined that the log-ratio is greater than 1.0 and extended the upper limit to the maximum log-ratio, 7.0, which gave rise to a wide confidence interval.

The confidence level was benchmarked as the percentage of the true confidence intervals. In a standard mixture, a confidence interval was determined to be true, if it contains the median of the protein log-ratio point estimates. On average, 955 proteins out of the 1,060 quantified proteins in a standard mixture had true confidence intervals. This means that, although the confidence interval estimation has a nominal confidence level of 95%, only

Figure 7.5: Histograms of the width of protein log-ratio confidence intervals. The distribution of confidence interval width reflects the varying quantification precision of proteins in a quantitative proteomics measurement.


1:1a
1:1b
5:1
1:5
10:1
1:10
Protein count
Confidence interval width

an average confidence level of 90% was obtained in the standard mixtures (Table 7.1). The decrease of the observed confidence level from the nominal one is probably because the peptide log-ratio probability model (Equations 7.3 – 7.5) is only an approximation to the true distribution. The confidence level of the interval estimation can be increased at the expense of widening the confidence intervals. This can be achieved by increasing the value of α in the ln-likelihood threshold, $\ln(L_{max}) - 0.5 \cdot \chi^2_{1,\alpha}$, which lowers the ln-likelihood threshold (the horizontal dashed line in the profile likelihood curves shown in Figure 7.3).

Confidence interval estimation enables hypothesis testing on the abundance change of a protein. The hypothesis testing can be used to filter the quantified proteins and select those with significant abundance change for further examination. Since most of the proteins in a proteome are not affected by a treatment, the null hypothesis is that there is no statistically significant difference in the abundance of a protein between two proteomes. The alternative hypothesis is that there is such a difference. The null hypothesis can be rejected for a protein if the protein's confidence interval does not include zero.

Two performance characteristics of a hypothesis testing method are its power and significance. A *post hoc* significance test was performed using the two 1:1 standard mixtures, in which the null hypothesis should hold for all proteins (Table 7.1). The average significance was 9%, which means that 9% of the proteins with no abundance

change are falsely asserted to have significant change. A *post hoc* power analysis was then performed using the standard mixtures with the other mixing ratios. The average power was 96%, which means that 96% of the proteins with abundance change are correctly identified. The abundance change of those proteins with accepted alternative hypothesis is only *statistically* significant, and the proteins should then be selected by the biological significance of their abundance change.

Testing of Abundance Ratio Estimation for Proteins with Extremely Large Abundance Change

The performance of the profile likelihood algorithm is sensitive to the isotopologue dynamic range. As the abundance difference between the two isotopologues increases among different standard mixtures, the average absolute deviation of the log-ratio point estimation increases and the confidence level of the interval estimation drops (Table 7.1). As the performance decrease is not very significant, we believe that the isotopologue dynamic range can reach 10-fold abundance difference with the LTQ-MS instrument and the ProRata program.

However, real-world biological samples might have proteins with extremely large abundance change, such as present in one proteome and absent in the other. We tested ProRata with an unlabeled proteome sample. In this case, all proteins only have the light isotopologue, and their log-ratios between the two isotopologues are expected to be infinity. Ideally, all protein log-ratios estimated by ProRata should be the maximum

considered log-ratio of 7. A total of 961 proteins were quantified. The histogram of their log-ratio estimates is shown in Figure 7.6. About 250 proteins have a log-ratio estimate next to the maximum log-ratio and the log-ratio estimates for most other proteins are evenly distributed between 2.0 and 6.0. The under-estimation of protein log-ratios is derived from the under-estimation of peptide log-ratios, which stems from noise fluctuations that falsely define the abundance of the non-existent heavy isotopologue.

Confidence interval estimation and hypothesis testing were also tested. The confidence intervals for half of the quantified proteins have an upper limit at the maximum log-ratio, which means only the lower bound can be estimated for those protein log-ratios. In this unlabeled proteome sample, the alternative hypothesis should be accepted for all proteins because of their change from present to absent. We found it to be the case for 97% of the quantified proteins. This percentage agrees with the observed power of the hypothesis testing in other standard mixtures. Therefore, although many of these proteins with very large abundance change have considerable error in their log-ratio point estimates, most of them can be correctly identified to have significant abundance change.

Manual Inspection of Protein Abundance Ratio Estimation through ProRata Graphical User Interface

Point estimation and confidence interval estimation of protein log-ratios were performed with ProRata automatically. However, there were a small percentage of proteins with spurious estimation results. One effective way to reducing the uncertainty is to manually

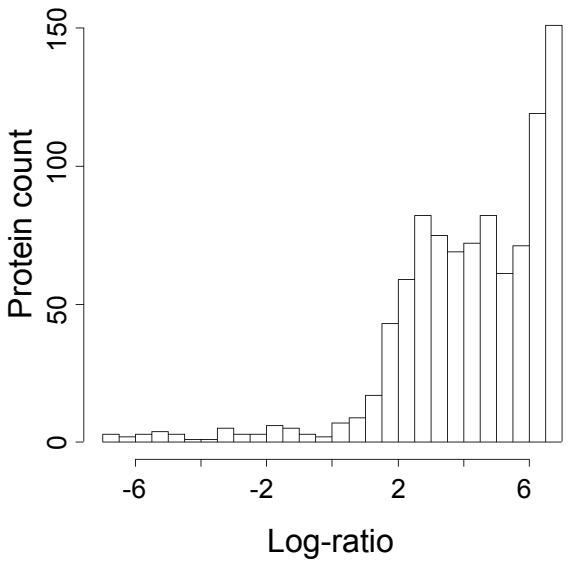


Figure 7.6: Histogram of the log-ratio estimates for proteins with extremely large abundance change. All proteins are expected to have a very large abundance ratio between the light isotopologue and the non-existent heavy isotopologue. 26% of the estimated protein abundance ratios are greater than 64:1 and 90% are greater than 4:1.

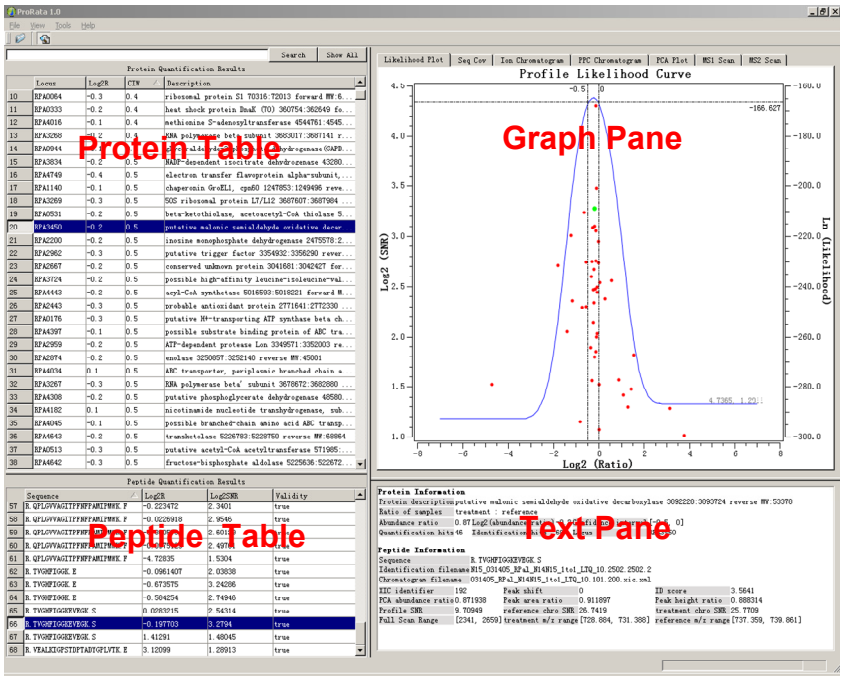
validate the proteins of interest. ProRata is equipped with a graphical user interface to enable interactive data interrogation and facilitate manual result validation.

ProRata's graphical user interface has four panes in its main window, including a *Protein Table*, a *Peptide Table*, a *Text Pane*, and a *Graph Pane* (Figure 7.7A). It was designed to give users a hierarchical view of their proteomics measurements. All quantified proteins are listed in the Protein Table. When a protein of interest is selected from the Protein Table, its profile likelihood curve and sequence coverage (Figure 7.7B) are displayed in the Graph Pane and its peptides are listed in the Peptide Table. Then a peptide from this protein can be selected to show its selected ion chromatograms (Figure 7.7C) and MS/MS scans (Figure 7.7D). Furthermore, the full scan at a retention time point in the selected ion chromatograms can be viewed with the mass spectral peaks for the two isotopologues highlighted (Figure 7.7E).

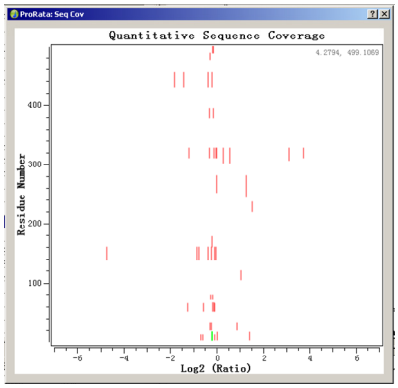
We have examined the proteins with erroneous point estimation and/or false confidence interval estimation. These proteins usually have less than three reliably quantified peptides. The reliably quantified peptides can generally be ascertained by inspecting their MS/MS scans, full scans, selected ion chromatograms and peak profiles. Therefore manual validation of the automated estimation results can provide an additional safeguard in quantitative proteomics against yielding false information.

Figure 7.7: Graphical user interface of ProRata. The main window of ProRata has four panes: Protein Table, Peptide Table, Graph Pane, and Text Pane (Part A). The Graph Pane contains two protein plots (sequence coverage plot (Part B) and profile likelihood curve plot), three peptide plots (selected ion chromatograms (Part C), parallel paired covariance chromatogram, and principal component analysis of peak profile), and two types of mass spectra (MS/MS scans (Part D) and full scans (Part E)). In the sequence coverage plot of a protein, a peptide is represented with a vertical segment indicating its log-ratio and its location on the protein sequence (Part B).

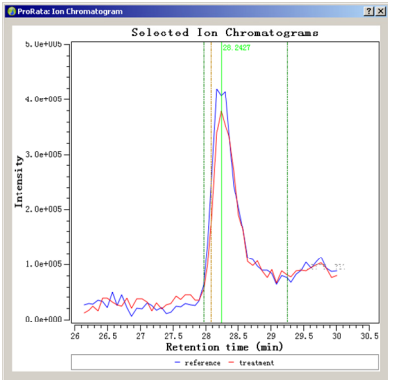
(A)


Protein Table
Graph Pane
Peptide Table
Text Pane
Profile Likelihood Curve
Log2 (Ratio)
Log2 (SNR)
Ln (Likelihood)

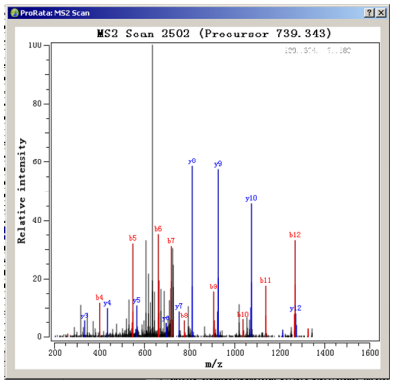
(B)


Quantitative Sequence Coverage
Residue Number
Log2 (Ratio)

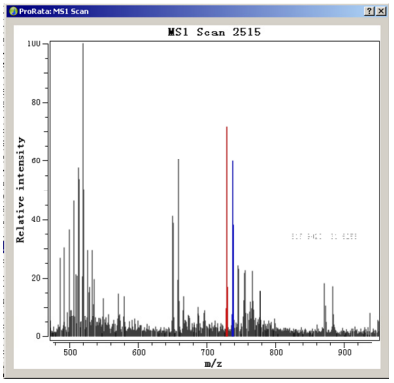
(C)


Selected Ion Chromatograms
Intensity
Retention time (min)
reference
treatment

(D)


MS2 Scan 2502 (Precursor 739.343)
Relative intensity
m/z

(E)


MS1 Scan 2515
Relative intensity
m/z

CONCLUSIONS

In this study, we applied maximum likelihood point estimation and profile likelihood confidence interval estimation for protein abundance ratio evaluation in quantitative shotgun proteomics with a profile likelihood algorithm. This algorithm is able to weight peptide abundance ratios by their estimation variability, account for peptide abundance ratio estimation bias, and suppress contribution from outliers. The algorithm was tested with standard mixtures of isotopically labeled proteomes at various mixing ratios. We demonstrated that the point estimation accuracy was improved using maximum likelihood estimation. The confidence intervals were estimated at the observed confidence level of 90%. With confidence interval estimation, hypothesis testing was performed on protein abundance change, which was benchmarked to have a significance of 9% and a power of 96%. The profile likelihood algorithm was also tested with an unlabeled proteome sample to show its ability to analyze proteins with extremely large abundance change. The profile likelihood algorithm was built into a computer program, ProRata, which automates the entire data analysis procedure for quantitative shotgun proteomics. ProRata's graphical user interface allows for manual validation of protein quantification results.

Chapter 8

Characterization of Anaerobic Catabolism of *p*-Coumarate in *Rhodopseudomonas palustris* by Integrating Quantitative Proteomics and Microarray

All of the data presented below is in preparation for submission

C. Pan, Y. Oda, D.A. Pelletier, B. Zhang, P.K. Lankford, N.F. Samatova, C.S. Harwood; Robert L. Hettich. Characterization of Anaerobic Catabolism of p-Coumarate in Rhodopseudomonas palustris by Integrating Quantitative Proteomics and Microarray. *Journal of Bacteriology* (2006), (In preparation)

As first author, C. Pan's contributions to this article include all proteomics data acquisition and interpretation.

INTRODUCTION

Lignin constitutes almost one third of all plant dry mass, which makes lignin the second most abundant organic compound on earth, after cellulose. Biodegradation of lignin during decay of plant residue in natural environment is a massive biological process within the global carbon cycle (Kirk, 1984). Lignin biodegradation is also of great practical significance because of its application to biological treatment and reuse of agricultural wastes. Lignin is a polymer of phenylpropanoid units, and its biodegradation involves depolymerization and then catabolism of the derived phenolic monomers (Sarkanen, 1971). *p*-Coumarate, or 4-hydroxy-cinnamic acid, is one of the main phenolic monomers (Hartley, 1989). The degradation of *p*-coumarate can occur in anoxic environments, such as aquifers, aquatic sediments, and submerged soils.

The purple nonsulfur phototrophic bacterium *Rhodopseudomonas palustris* is one of a few known microorganisms capable of anaerobic catabolism of diverse aromatic compounds, such as benzoate, *p*-coumarate, cinnamate, and vanillate (Harwood, 1997; Diaz, 2004). A central intermediate, benzoyl-CoA, is used in degradation of many of these aromatic compounds in *R. palustris* (Figure 8.1) (Harwood, 1999). Peripheral pathways transform different aromatic substrates to benzoyl-CoA. The central benzoyl-CoA pathway then degrades benzoyl-CoA to acetyl-CoA (Elder, 1994; Harwood, 1999). We hypothesize that *p*-coumarate is transformed to benzoyl-CoA via a peripheral pathway consisting of CoA ligation, ß-oxidation, and dehydroxylation (Figure 8.2). The generated benzoyl-CoA is catabolized through the known benzoyl-CoA pathway.

In this study, the gene expression profile of *R. palustris* grown with *p*-coumarate as the sole organic carbon source was compared to those of *R. palustris* grown with succinate or benzoate. As succinate is a simple dicarboxylic acid, the succinate growth condition provides the gene expression profile of *R. palustris* without aromatic degradation activity. On the other hand, the aromatic degradation activity is induced in both benzoate and coumarate growth conditions. Benzoat-CoA ligase (RPA0661) transforms benzoate directly to the central intermediate benzoyl-CoA. Therefore, the comparison between the benzoate growth condition and the coumarate growth condition was expected to shed light on the peripheral pathway from coumarate to benzoyl-CoA (Figure 8.2).

Microarray analysis has been established as a high-throughput method for gene expression profiling at the mRNA level (Harrington, 2000). On the other hand,

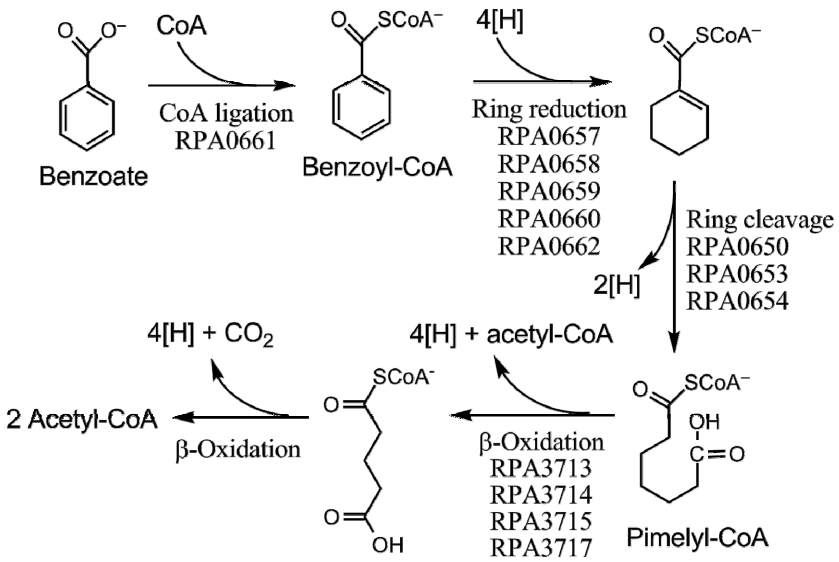


Figure 8.1: ***R. palustris*** **benzoyl-CoA pathway.** The degradation of benzoyl-CoA involves three steps: ring reduction, ring cleavage and ß–oxidation (Harwood, 1999). The identified genes are shown for each step.

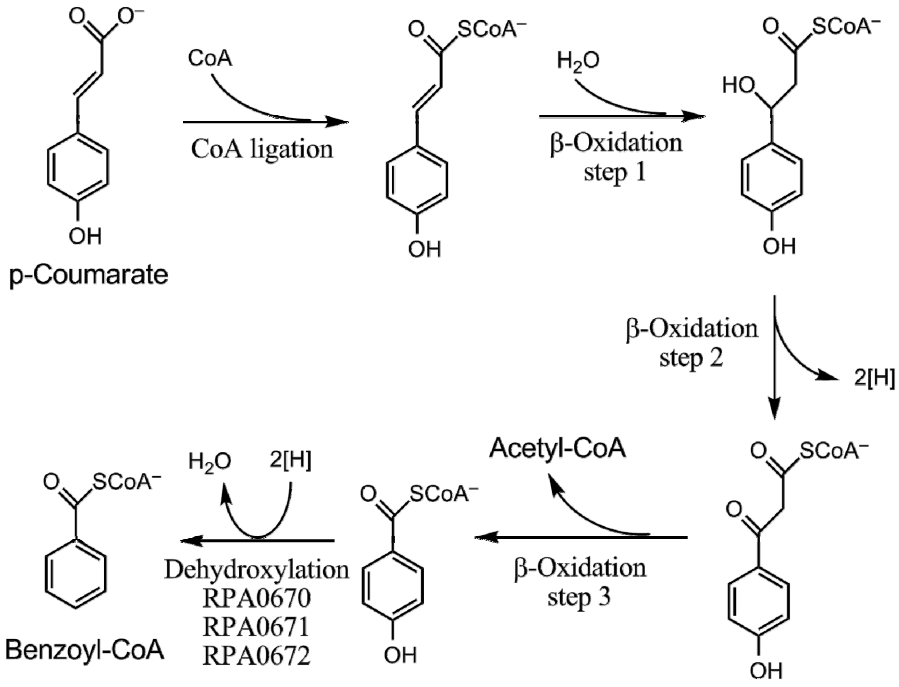


Figure 8.2: Proposed *R. palustris p*-coumarate degradation pathway. The catabolism of *p*-coumarate is hypothesized to proceed through the benzoyl-CoA pathway. The side chain of *p*-coumarate is removed by ß–oxidation.

proteomics is a burgeoning technology for characterization of the protein complement of a cell. Earlier studies with proteomics have largely been qualitative, focusing on cataloging proteins (Washburn, 2000). With recent technological development, proteomics has turned quantitative, capable of measuring the abundance ratios of thousands of proteins between two proteomes (Ong, 2005b). It has now become possible to ascertain the relative expression activities of a large set of genes at both mRNA level and protein level by integrating microarray and quantitative proteomics data.

In this study, we demonstrate the value of such integrated gene expression profiling data. First, the gene expression profiles at both mRNA level and protein level allow the identification of both transcriptional and post-transcriptional regulation of many genes. Second, greater confidence was attained for many genes of interest, when they showed consistent expression changes measured by the two independent methodologies.

MATERIALS AND METHODS

Bacterial Growth and Metabolic Stable Isotope Labeling.

Wild type *Rhodopseudomonas palustris* CGA0010 was grown anaerobically on different defined mineral growth media in sealed tubes with a nitrogen gas headspace at 30 °C with ample incandescent light illumination. $(NH_4)_2SO_4$ was the only nitrogen source available for bacterial assimilation, and was provided as $(^{14}NH_4)_2SO_4$ for the unlabeled culture and as $(^{15}NH_4)_2SO_4$ for the ^{15}N-labeled culture (>98 atom percentage excess,

Sigma-Aldrich, St. Louis, MO). 3 mM *p*-coumarate was supplied as the sole organic carbon source for the unlabeled coumarate culture. 3 mM benzoate and 10 mM succinate were supplied as the sole organic carbon sources for the ^{15}N-labeled benzoate and succinate cultures, respectively. Duplicate cultures were prepared for each of the three growth conditions. Cell growth was monitored spectrophotometrically at 660 nm and cells were harvested in mid-log phase at $OD_{660\ nm}$ of 0.6 by centrifugation and washed twice with ice-cold wash buffer (50 mM Tris-HCl buffer at pH 7.5 with 10 mM EDTA). The harvested cell pellet from each culture was divided for quantitative shotgun measurements and microarray analysis.

Proteome Sample Preparation.

Duplicate cell mixtures of the unlabeled *p*-coumarate culture and the ^{15}N-labeled succinate culture were prepared by mixing equal weight of cell pellets from duplicate cultures. Duplicate cell mixtures of the unlabeled *p*-coumarate culture and the ^{15}N-labeled benzoate culture were prepared similarly. The cell mixtures were lysed by sonication in the ice-cold wash buffer, and unbroken cells were removed by centrifugation at 5000g for 10 min. The obtained cell lysates were fractionated by ultracentrifugation at 100,000g for 1 h. The resulting supernatants were labeled as the soluble protein fraction. The pellets were re-suspended by sonication and labeled as the membrane protein fraction. Protein concentration for each sample was determined with Lowry's analysis. The two fractions from each cell mixtures were digested using the following protocol. The proteins were denatured and reduced with 6 M guanidine and 10 mM dithiothreitol (DTT) (Sigma

Chemical Co. St. Louis, MO) at 60°C for 1 h. The denatured proteome fractions were diluted 6-fold with 50 mM Tris/10 mM $CaCl_2$ (pH 7.6), and sequencing grade trypsin was added at 1:100 (wt:wt). The first digestion was run overnight at 37°C and, after adding additional trypsin, the second digestion was run for 5 hrs at 37°C. The samples were then reduced with 20 mM DTT for 1 h at 60°C and were desalted using C18 solid-phase extraction (Sep-Pak Plus, Waters, Milford, MA).

Quantitative Proteomics Measurement.

The protein digests were examined with LC-MS/MS using twelve-step split-phase MudPIT (MacCoss, 2002; McDonald, 2002) in duplicate. The samples were loaded via a pressure bomb (New Objective, Woburn, MA) onto a 250-*u*m-I.D. fused silica front column fritted into an M-520 filter union (Upchurch Scientific). The column packing consisted of 2 cm strong cation exchange resin (Luna, Phenomonex) and 2 cm C18 reverse-phase resin (Aqua, Phenomonex). A 100-*u*m-I.D. PicoFrit column (New Objective, Woburn, MA) was packed with 15 cm C18 reverse-phase resin. The front column was connected with the PicoFrit column and then placed in-line with a Dionex Ultimate quaternary HPLC. Two-dimensional LC separation was performed with twelve salt pulses, each of which was followed by a 2-h reverse-phase gradient. MS/MS analysis was performed on an LTQ linear ion trap instrument (ThermoFinnigan, San Jose, CA) with dynamic exclusion enabled. Each full scan (400-1700 *m/z*) was followed by three data-dependent MS/MS scans at 35% normalized collision energy. The full scans were

averaged from five microscans and the MS/MS scans were averaged from two microscans.

Quantitative Proteomics Data Analysis.

All MS/MS scans were searched in two iterations against the FASTA database containing all annotated *Rhodopseudomonas palustris* proteins (Larimer, 2004) using the SEQUEST program (Eng, 1994). In the first iteration, the unmodified amino acids were used, and, in the second iteration, the modified amino acids with ^{15}N-labeling were used. The peptide identifications from the two iterations were merged. The DTASelect program (Tabb, 2002) was used to filter the peptide identifications and to assemble the peptides into proteins using the following parameters: retaining the duplicate MS/MS spectra for each peptide sequence (DTASelect option: -t 0), fully tryptic peptides only, with a delCN of at least 0.08 and cross-correlation scores (Xcorrs) of at least 1.8 (+1), 2.5 (+2), and 3.5 (+3). Selected ion chromatogram extraction, peptide abundance ratio estimation and protein abundance ratio estimation were completed with the ProRata program as described in previous chapters.

RESULTS AND DISCUSSION

Experimental Design and Measurement Result Overview

Both microarray and quantitative proteomics measure the abundance changes of gene expression products between a treatment condition and a reference condition. The treatment condition in this study was the anaerobic photosynthetic cell growth with *p*-coumarate as the sole organic carbon source. To identify the genes activated for *p*-coumarate catabolism, two reference conditions were selected, in which succinate or benzoate replaced *p*-coumarate as the sole organic carbon source. Comparison of the *p*-coumarate condition to the benzoate condition could yield information on the pathway from *p*-coumarate to benzoyl-CoA (Figure 8.2). As succinate can be readily used to generate acetyl-CoA through the citric acid cycle, comparison of the *p*-coumarate condition to the succinate condition should help elucidate the entire pathway for *p*-coumarate catabolism.

The experimental scheme is shown in Figure 8.3. At the cell culture step, *R. palustris* was grown under three conditions with ^{14}N or ^{15}N stable isotope labeling. Two biological replicates were prepared for each condition and analyzed independently to capture both biological and technical variability. Each biological replicate was divided for microarray and quantitative proteomics analysis. We found that the correlation between microarray results and quantitative proteomics results was greatly improved by splitting the same

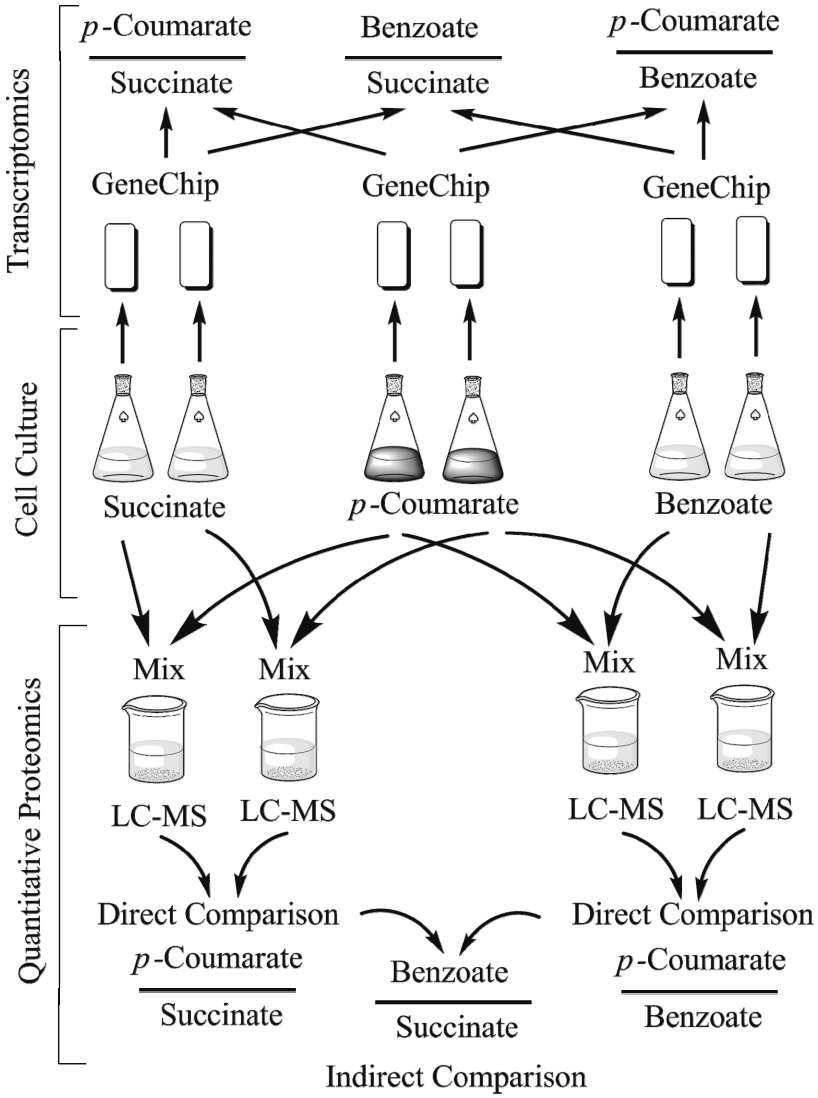


Figure 8.3: Experimental scheme of integrated gene expression profiling. The *p*-coumarate, succinate and benzoate cultures were prepared in biological duplicate with metabolic stable isotope labeling. The cell pellets were divided for quantitative proteomics measurements and transcriptomics measurements, which yielded the relative gene expression profiles among the three growth states.

sample for the two measurements, as compared to using separate samples for the two measurements (Data not shown).

Microarray analysis was performed in Dr. Caroline S. Harwood's group (University of Washington, Seattle) with a custom-designed GeneChip, which contained probes for all 4836 predicted genes and 3190 non-coding regions in the *R. palustris* genome. Each biological replicate was analyzed with a GeneChip (Figure 8.3). RNA molecules from the succinate and benzoate conditions contained ^{15}N-enriched nitrogen, as a result of ^{15}N metabolic labeling. The ^{15}N labeling does not appear to interfere with the microarray analysis. The reproducibility of the signal intensity measurement between biological replicates was calculated using correlation coefficient, which exceeded 0.98 for all three conditions (Data not shown). The relative abundances of mRNAs were derived by taking the ratios of the signal intensities from two conditions.

The ^{14}N-labeled *p*-coumarate cell pellet was mixed with the ^{15}N-labeled succinate cell pellet or benzoate cell pellet for quantitative proteomics analysis (Figure 8.3). The measurement of the two biological replicates yielded the direct comparison of protein abundance between the *p*-coumarate condition and each of the two reference conditions: succinate and benzoate. The two direct comparisons were then combined to form the indirect comparison between the two reference conditions. The numbers of identified proteins and quantified proteins in each comparison were summarized in Table 8.1. While microarray analysis can be used to determine mRNA levels for most genes in a genome, quantitative proteomics can only quantify protein levels for a subset of genes.

Table 8.1: Summary of quantitative proteomics results

Comparison	Coumarate : Succinate			Coumarate : Benzoate			Benzoate : Succinate
Number of Proteins	Biological Replicate 1	Biological Replicate 2	Direct Comparison	Biological Replicate 1	Biological Replicate 2	Direct Comparison	Indirect Comparison
Identified	2385	2172	2012	2058	1979	1801	1500
Quantified	1859	1790	1539	1785	1730	1627	1151

The missed proteins are generally membrane proteins, low-abundance proteins, or proteins unexpressed in either condition. The reproducibility of quantitative proteomics between biological replicates was measured by correlation coefficients of protein abundance ratios in $\log_2$ scale. The correlation coefficients were 0.87 for the direct comparison of *p*-coumarate and succinate and 0.84 for the direct comparison of *p*-coumarate and benzoate (Figure 8.4). The level of reproducibility in the $\log_2$ abundance ratio scale was comparable between quantitative proteomics and microarray analysis.

Integration of mRNA Abundance Profiles and Protein Abundance Profiles.

The transcriptomics data and the proteomics data were cross-matched by gene locus. Due to the incomplete coverage of the proteome, "Not Available" (N/A) values were assigned to the absent protein log-ratios. The correlations between the mRNA log-ratios and the protein log-ratios of the cross-matched genes are shown in Figure 8.5, which all havepositive Pearson correlation coefficients (r^2). The majority of data points were tightly clustered around the center between the log-ratio interval of [-1, 1] along the mRNA log-ratio axis and the protein log-ratio axis. This shows that, at the expression fluctuation of less than 2 fold, the mRNA change and the protein change have little correlation. The relatively lower correlation coefficient of the benzoate-coumarate comparison is a result of the lower number of genes with large expression changes.

Histograms of differences between the mRNA log-ratios and the protein log-ratio (Δ log-ratio) were also constructed for each comparison (Figure 8.5). The discrepancy between

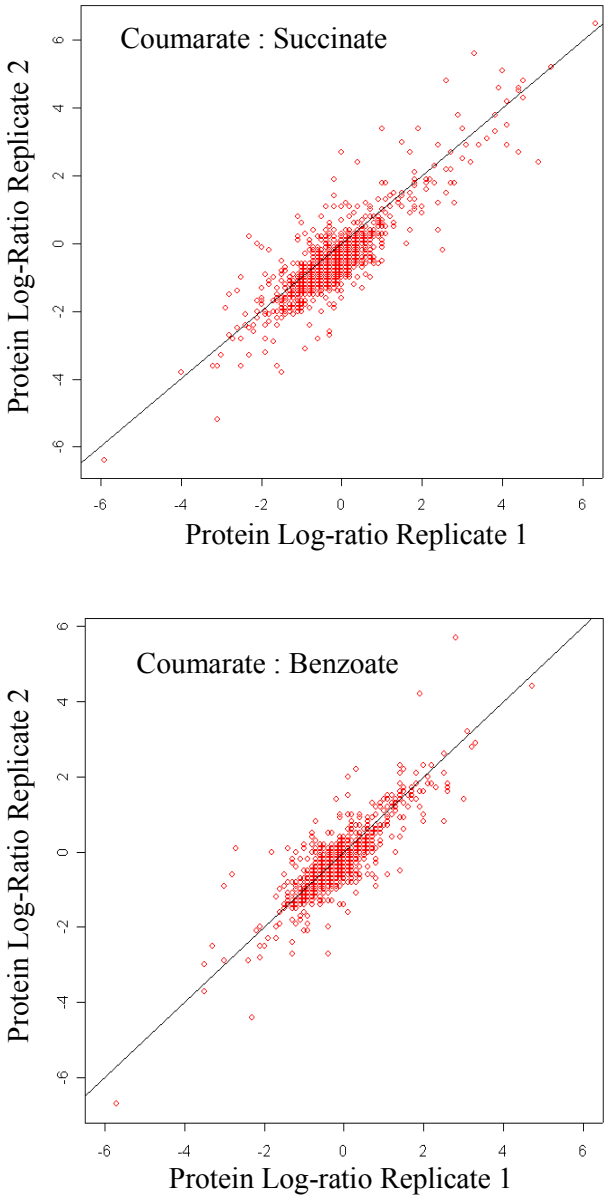
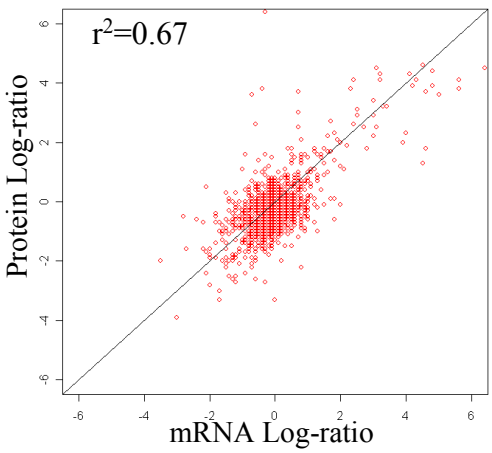
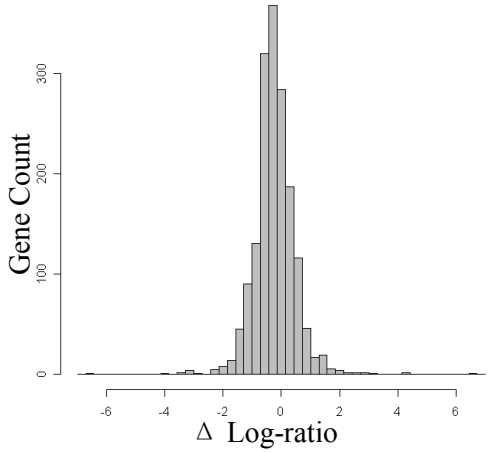


Figure 8.4: Reproducibility of quantitative proteomics results. The protein log-ratios measured from the two biological replicates were compared to benchmark the reproducibility. The proteins with consistent log-ratios coincide with the solid line (y = x).

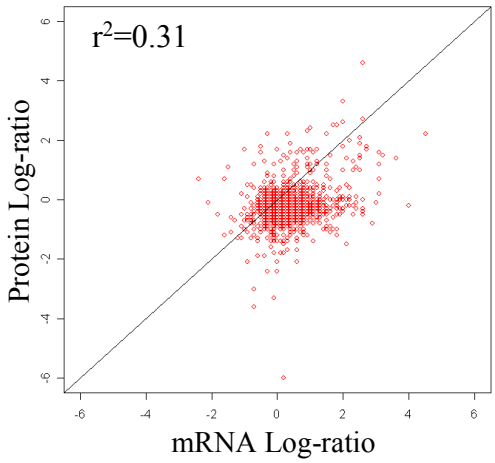
Figure 8.5: Comparison of mRNA log-ratios and protein log-ratios. The protein log-ratios and mRNA log-ratios of quantified genes are shown as scatter-plots for the three comparisons. And the histograms of the differences between the protein log-ratio and the mRNA log-ratio (Δ log-ratio) were constructed for the three comparisons.

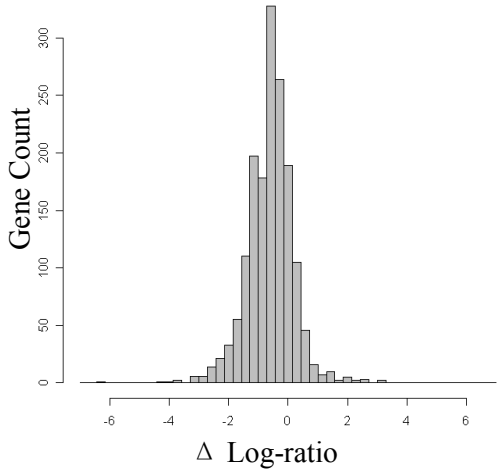
Coumarate : Succinate



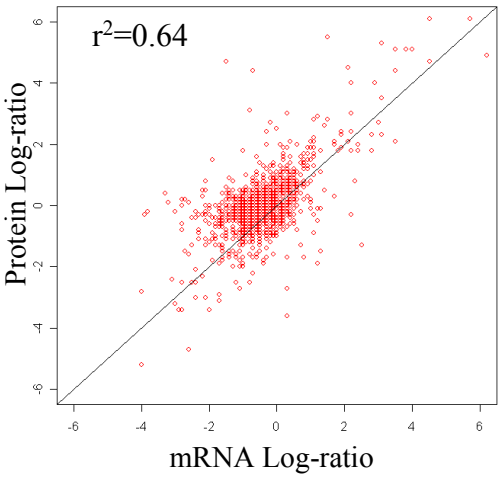


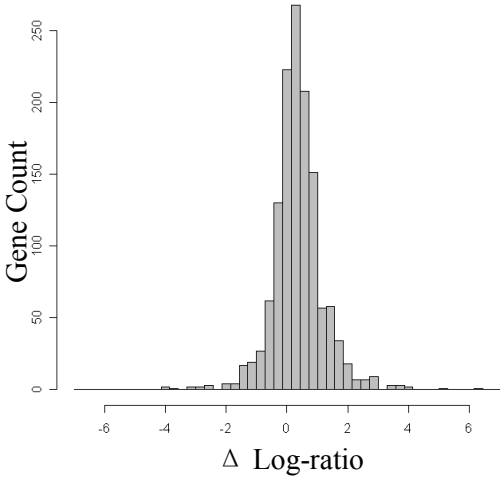
Coumarate : Benzoate





Benzoate : Succinate





mRNA log-ratio and protein log-ratio of a gene can stem from the combination of following effects (Julka, 2004):

- Measurement errors. Both mRNA log-ratio and protein log-ratio of a gene contain random errors from global high-throughput measurements. Strictly speaking, only the deviation between the *confidence intervals* of mRNA log-ratios and protein log-ratios is statistically significant discrepancy, which can be attributed to factors other than measurement errors.
- Post-transcriptional regulation. The mRNA abundance change is not directly linked to the protein abundance change. Two additional steps can be regulated: the protein synthesis rate from a unit amount of mRNA and the protein degradational rate, which can together be termed as post-transcriptional regulation.
- Sustained protein presence from transient transcriptional regulation. Generally, mRNA molecules have a very short half-life time. Once the transcription induced by a stimulus ends, the mRNA level would quickly return to baseline. However, as proteins are very stable biomolecules, the proteins synthesized from this pulse of mRNA transcripts can exist for an extended time frame. The snapshot measurement of the mRNA level and the protein level after the transcriptional induction can show a baseline mRNA level and an enhanced protein level.

The integrated results of transcriptomics and quantitative proteomics are summarized in Figure 8.6. The tables show the number of genes categorized by the regulation directions at mRNA level and protein level. All categories were color-coded to illustrate the complementarity between transcriptomics and proteomics. The categories in yellow show

Coumarate : Succinate

Protein	mRNA: Down	mRNA: Null	mRNA: Up
N/A	149	2757	224
Up	1	36	62
Null	46	1301	21
Down	46	166	1

Coumarate : Benzoate

Protein	mRNA: Down	mRNA: Null	mRNA: Up
N/A	528	2743	215
Up	1	34	38
Null	181	884	11
Down	4	78	7

Benzoate : Succinate

Protein	mRNA: Down	mRNA: Null	mRNA: Up
N/A	528	2743	215
Up	2	73	50
Null	181	884	11
Down	67	54	3

Figure 8.6: Summary of gene expression profiling results. The expression changes of genes were categorized into up-, null- and down-regulation at the protein level and the mRNA level. The genes not quantified by proteomics were put into the N/A group of protein abundance regulation. The complementarities between proteomics and transcriptomics were color-coded in the tables. Genes in the yellow cells only have transcriptomics data. The genes in the green cells requires proteomics data to reflect its true activity. The genes in the red cells have consistent regulation observed by both measurements.

the advantage of transcriptomics, which is measurement of the genes missed by proteomics. The categories in green show the advantage of proteomics. Significant change at the protein level can be associated with an insignificant change at the mRNA level, and conversely a significant change at the mRNA level does not necessarily lead to a significant change at the protein level. The categories in red contain genes with concordant mRNA change and protein change. The consistent result from two independent measurements gives enhanced confidence in the expression regulation of those genes and alleviates the need for validating these genes' expression individually with RT-PCR, northern blotting, western blotting, *etc*.

Identification of cellular pathways for aromatic compound catabolism

As both coumarate and benzoate are aromatic compounds, genes with up-regulated expression under both aromatic degradation conditions were selected and grouped into known cellular pathways. The benzoyl-CoA pathway is the most significantly up-regulated pathway (Table 8.2). The activation of this pathway in the coumarate condition supported our hypothesis that catabolism of coumarate proceeds through the benzoyl-CoA pathway (Figure 8.1). However, the benzoyl-CoA pathway is less activated in the coumarate condition than in the benzoate condition, probably because there is a rate limiting step in generating benzoyl-CoA from coumarate. Also note the anti-correlation between the mRNA regulation and the protein regulation of this pathway in the coumarate-benzoate comparison. The genes in pimelyl-CoA β-oxidization were less up-regulated than the genes up-stream in the benzoyl-CoA pathway. A number of genes in

Table 8.2: Expression change of the genes in the benzoyl-CoA pathway.

Reation	Locus	Coumarate : Succinate				Coumarate : Benzoate				Benzoate : Succinate				Gene description
		mRNA*		Protein		mRNA*		Protein		mRNA*		Protein		
		Log-ratio	P-value	Log-ratio	CI	Log-ratio	P-value	Log-ratio	CI	Log-ratio	P-value	Log-ratio	CI	
CoA ligation	RPA0661	3.0	5.E-08	2.2	[1.9, 2.4]	1.5	7.E-06	-0.7	[-0.8, -0.5]	1.5	2.E-06	2.8	[2.6, 3.1]	Benzoate-CoA ligase
Ring reduction	RPA0657	3.2	6.E-10	4.3	[4.0, 4.7]	1.8	6.E-08	-1.2	[-1.3, -1.1]	1.5	4.E-06	5.5	[5.2, 2.9]	Benzoyl-CoA reductase subunit badD
	RPA0658	4.2	1.E-12	3.9	[3.6, 4.2]	1.1	1.E-06	-1.4	[-1.5, -1.1]	3.1	1.E-07	5.3	[5.0, 2.6]	Benzoyl-CoA reductase subunit badE
	RPA0659	4.8	5.E-13	4.4	[4.2, 4.6]	0.3	4.E-02	-1.7	[-1.9, -1.3]	4.5	3.E-09	6.1	[5.9, 6.3]	Benzoyl-CoA reductase subunit badF
	RPA0660	6.4	9.E-09	4.5	[4.2, 4.7]	0.6	7.E-03	-1.7	[-1.8, -1.5]	5.7	5.E-10	6.1	[5.9, 6.4]	Benzoyl-CoA reductase subunit badG
	RPA0662	3.4	3.E-08	3.2	[2.8, 3.6]	1.3	4.E-05	-0.8	[-1.2, -0.5]	2.2	7.E-08	4.0	[3.5, 4.5]	Ferredoxin
Ring cleavage	RPA0650	3.9	2.E-07	N/A	N/A	0.6	1.E-02	-1.3	[-1.5, -1.2]	3.2	2.E-10	N/A	N/A	Cyclohex-1-ene-1-carboxyl-CoA hydratase
	RPA0653	4.8	5.E-10	3.9	[3.7, 4.0]	0.9	5.E-04	-1.2	[-1.3, -1.1]	3.8	5.E-06	5.1	[4.9, 5.2]	2-ketocyclohexanecarboxyl-CoA hydrolase
	RPA0654	4.6	1.E-08	3.7	[3.5, 4.0]	0.6	2.E-02	-1.4	[-1.5, -1.3]	4.0	7.E-07	5.1	[4.9, 5.4]	2-hydroxycyclohexanecarboxyl-CoA dehydrogenase
β-Oxidation	RPA3713	1.4	2.E-04	1.6	[1.5, 1.8]	0.3	2.E-01	0.1	[0.0, 0.2]	1.0	8.E-05	1.5	[1.3, 1.7]	Pimeloyl-CoA dehydrogenase small subunit
	RPA3714	1.2	7.E-04	1.8	[1.6, 1.9]	0.2	5.E-01	0.1	[0.0, 0.2]	1.1	8.E-05	1.7	[1.5, 1.8]	Pimeloyl-CoA dehydrogenase large subunit
	RPA3715	1.3	2.E-04	1.6	[1.4, 1.7]	0.2	4.E-01	0.0	[-0.2, 0.1]	1.1	2.E-05	1.6	[1.4, 1.8]	Acetyl-CoA acetyltransferase
	RPA3717	1.8	5.E-05	1.4	[1.3, 1.6]	2.2	1.E-05	0.0	[-0.1, 0.2]	-0.5	3.E-02	1.4	[1.2, 1.6]	Enoyl-CoA hydratase

* The mRNA data were provided by Dr. Caroline S. Harwood's group (University of Washington, Seattle)

fatty acid metabolism were measured to have an expression level changed by the coumarate and benzoate catabolism. Probably a portion of pimelate molecules were diverted to other fatty acid metabolism pathways, rather than being degraded to acetyl-CoA

The Calvin cycle was activated in both coumarate and benzoate conditions (Elder, 1994). As shown in Figure 8.1, six reducing equivalents [H] are generated when a molecule of benzoyl-CoA is degraded into acetyl-CoA. In an anaerobic condition, the reducing equivalents cannot be oxidized with oxygen to produce ATP. As a result, the Calvin cycle was turned on as a reducing equivalent sink and as a supplement carbon supply source in addition to aromatic compound degradation (Elder, 1994). As the Calvin cycle and the benzoyl-CoA pathway are coupled by the reducing equivalent, the Calvin cycle was also less activated in the coumarate condition than in the benzoate condition.

Chemotaxis systems were also induced in both aromatic degradation conditions. Many plant-derived aromatic compounds have been found to be chemo-attractants to bacteria capable of those compounds' degradation (Parales, 2002). As there was no aromatic compound gradient in the laboratory culture condition, the induction of the chemotaxis system was probably due to coordinate regulation with the aromatic degradation process.

Characterization of cellular pathways for coumarate catabolism

A hypothesis to be tested in this study is that coumarate is degraded to 4-hydroxyl-benzoyl-CoA and subsequently to benzoyl-CoA as shown in Figure 8.2. If the hypothesis is true, genes required for this β-oxidization should be up-regulated in the coumarate condition compared to both the benzoate and succinate conditions. All genes with up-regulated protein level in the coumarate condition are listed in Table 8.3.

RPA1787 has approximately 50% amino acid sequence identity to known feruloyl-CoA synthetases in *Pseudomonas sp* and *Pseudomonas fluorescens* (Overhage, 1999). The up-regulated locus RPA1787 and the down-regulated locus RPA1707 are two putative feruloyl-CoA synthetases in *R. palustris* genome. Considering the structural similarity between ferulate and coumarate and the specific up-regulation of RPA1787 expression, RPA1787 is likely the CoA ligase for coumarate catabolism (Figure 8.7). The up-regulated RPA1786 was annotated as a putative 3-hydroxybutyryl-CoA dehydratase, but it has equally strong sequence similarity to enoyl-CoA hydratases. As coumaroyl-CoA is an aromatic enoyl-CoA, up-regulation of RPA1786 expression suggests that it is the enoyl-CoA hydratases for coumaroyl-CoA in β-oxidization (Figure 8.7). However, there was no highly probable dehydrogenase and acyl-CoA thiolase for the β-oxidization from the list of genes with up-regulated mRNA level or protein level.

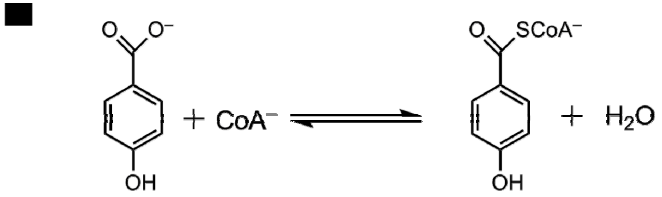
Interestingly, a 4-hydroxybenzoyl-CoA thioesterase (RPA1788) was up-regulated, which would turn 4-hydroxybenzoyl-CoA to 4-hydroxybenzoate. The induction of this

Table 8.3: Genes with up-regulated protein level in comparisons of coumarate with succinate and benzoate.

Locus	Coumarate : Succinate				Coumarate : Benzoate				Benzoate : Succinate				Gene description
	mRNA		Protein		mRNA		Protein		mRNA		Protein		
	Log-ratio	P-value	Log-ratio	CI	Log-ratio	P-value	Log-ratio	CI	Log-ratio	P-value	Log-ratio	CI	
RPA0665	3.0	2.E-06	2.9	[2.7, 3.2]	3.1	2.E-06	1.6	[1.3, 1.9]	-0.1	7.E-01	1.4	[1, 1.7.0]	Putative ABC transporter subunit, ATP-binding component
RPA0668	2.4	2.E-07	2.6	[2.4, 2.7]	2.0	9.E-07	1.8	[1.7, 1.9]	0.5	5.E-02	0.8	[0.6, 0.9]	Putative ABC transporter subunit, substrate-binding component
RPA0669	5.6	8.E-10	4.1	[3.8, 4.4]	2.2	6.E-08	2.0	[1.9, 2.1]	3.5	3.E-09	2.1	[1.8, 2.4]	4-hydroxybenzoate-CoA ligase ■
RPA0670	6.3	4.E-08	3.1	[2.6, 3.7]	2.5	7.E-07	1.0	[0.7, 1.3]	3.9	5.E-09	2.2	[1.6, 2.7]	4-hydroxybenzoyl-CoA reductase, first subunits
RPA0671	5.6	2.E-10	3.8	[3.4, 4.2]	2.5	4.E-09	1.5	[1.3, 1.7]	3.1	4.E-09	2.3	[1.9, 2.7]	4-hydroxybenzoyl-CoA reductase,second subunits
RPA0672	4.5	2.E-08	4.6	[4.3, 5.0]	2.6	1.E-07	2.7	[2.2, 3.4]	1.9	5.E-09	1.9	[1.2, 2.6]	4-hydroxybenzoyl-CoA reductase, third subunits
RPA1009	3.3	3.E-11	3.2	[3.1, 3.4]	2.0	9.E-10	3.3	[3.2, 3.4]	1.2	8.E-05	-0.1	[-0.2, 0.1]	Possible cytochrome P450
RPA1206	-0.5	3.E-01	1.4	[1.2, 1.5]	-1.0	5.E-02	1.7	[1.5, 1.8]	0.5	3.E-01	-0.3	[-0.5, -0.1]	Aldehyde dehydrogenase
RPA1414	-1.0	1.E-04	1.1	[0.9, 1.4]	-0.7	2.E-03	1.2	[0.9, 1.4]	-0.3	2.E-01	0.0	[-0.3, 0.3]	MaoC-like dehydratase
RPA1782	1.2	4.E-04	1.7	[1.5, 1.9]	0.3	2.E-01	1.4	[1.2, 1.6]	0.9	6.E-05	0.3	[0.0, 0.5]	C4-dicarboxylate periplasmic binding protein, dctP subunit,
RPA1786	4.3	7.E-06	4.1	[3.7, 4.4]	1.5	6.E-04	2.2	[2.1, 2.4]	2.8	1.E-08	1.8	[1.5, 2.2]	Putative 3-hydroxybutyryl-CoA dehydratase ▲
RPA1787	4.1	2.E-05	4.3	[4.1, 4.5]	1.7	6.E-04	2.5	[2.4, 2.6]	2.4	7.E-07	1.8	[1.6, 2.0]	Putative feruloyl-CoA synthetase ‡
RPA1788	2.4	1.E-03	4.1	[3.8, 4.4]	2.7	8.E-04	1.7	[1.4, 2.0]	-0.3	2.E-01	2.4	[2.0, 2.8]	Possible 4-hydroxybenzoyl-CoA thioesterase ■
RPA1789	0.4	2.E-01	1.8	[1.7, 1.9]	2.5	6.E-05	2.1	[2.0, 2.2]	-2.1	8.E-06	-0.3	[-0.4, -0.2]	Putative branched-chain A.A. transporter, substrate-binding protein
RPA1791	1.9	2.E-04	2.1	[1.8, 2.3]	2.1	2.E-04	1.4	[1.1, 1.6]	-0.1	5.E-01	0.7	[0.4, 1.0]	Putative branched-chain A.A. transporter, ATP-binding protein
RPA1792	2.2	1.E-04	1.9	[1.7, 2.2]	2.5	7.E-05	1.4	[1.1, 1.6]	-0.3	8.E-02	0.6	[0.3, 0.9]	Putative branched-chain A.A. transporter, ATP-binding protein
RPA3011	2.5	2.E-08	2.9	[2.6, 3.3]	2.6	1.E-08	4.6	[4.3, 4.9]	-0.1	7.E-01	-1.6	[-2.1, -1.1]	Unknown protein
RPA3014	0.8	5.E-03	2.1	[1.6, 2.5]	4.5	7.E-09	2.2	[1.9, 2.6]	-3.8	4.E-06	-0.2	[-0.7, 0.4]	Two-component transcriptional regulator, LuxR family
RPA3101	-0.1	9.E-01	1.5	[1.4, 1.7]	0.2	5.E-01	1.5	[1.4, 1.6]	-0.3	3.E-01	0.0	[-0.1, 0.2]	Unknown protein
RPA3423	0.8	5.E-04	1.6	[1.3, 1.9]	0.4	2.E-02	1.6	[1.3, 2.0]	0.4	9.E-02	0.0	[-0.4, 0.4]	Unknown protein
RPA3893	4.0	4.E-07	2.3	[2.0, 2.6]	3.6	6.E-07	1.4	[1.1, 1.8]	0.4	6.E-01	0.9	[0.4, 1.3]	Putative carboxylesterase
RPA4092	2.8	5.E-08	1.8	[1.7, 2.0]	2.7	9.E-08	1.8	[1.6, 2.2]	0.1	8.E-01	0.0	[-0.4, 0.3]	Unknown protein
RPA4096	2.3	7.E-06	2.6	[2.0, 4.1]	2.0	2.E-05	3.1	[2.7, 3.6]	0.3	5.E-01	-0.5	[-1.3, 1.0]	Possible multidrug efflux membrane fusion protein mexE
RPA4198	2.3	5.E-05	3.8	[3.5, 4.0]	0.9	1.E-02	1.2	[1.1, 1.3]	1.4	3.E-06	2.6	[2.3, 2.8]	Amidohydrolase 2

Table 8.3: Continued

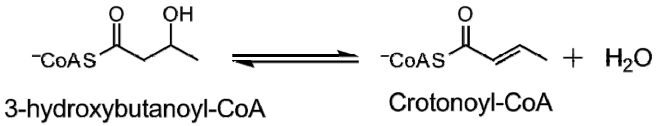
■ 4-hydroxybenzoyl-CoA thioesterase



4-hydroxybenzoate 4-hydroxybenzoyl-CoA

4-hydroxybenzoate-CoA ligase

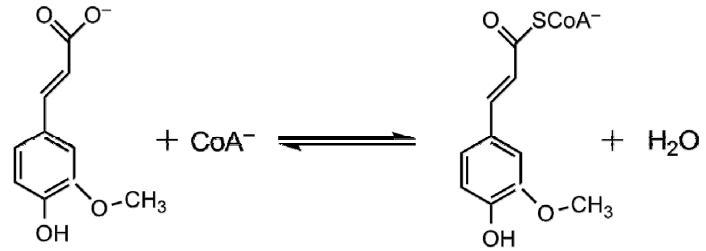
▲ Enoyl-CoA hydratase



3-hydroxybutanoyl-CoA Crotonoyl-CoA

3-hydroxybutyryl-CoA dehydratase

‡



Ferulate Feruloyl-CoA

Feruloyl-CoA synthetase

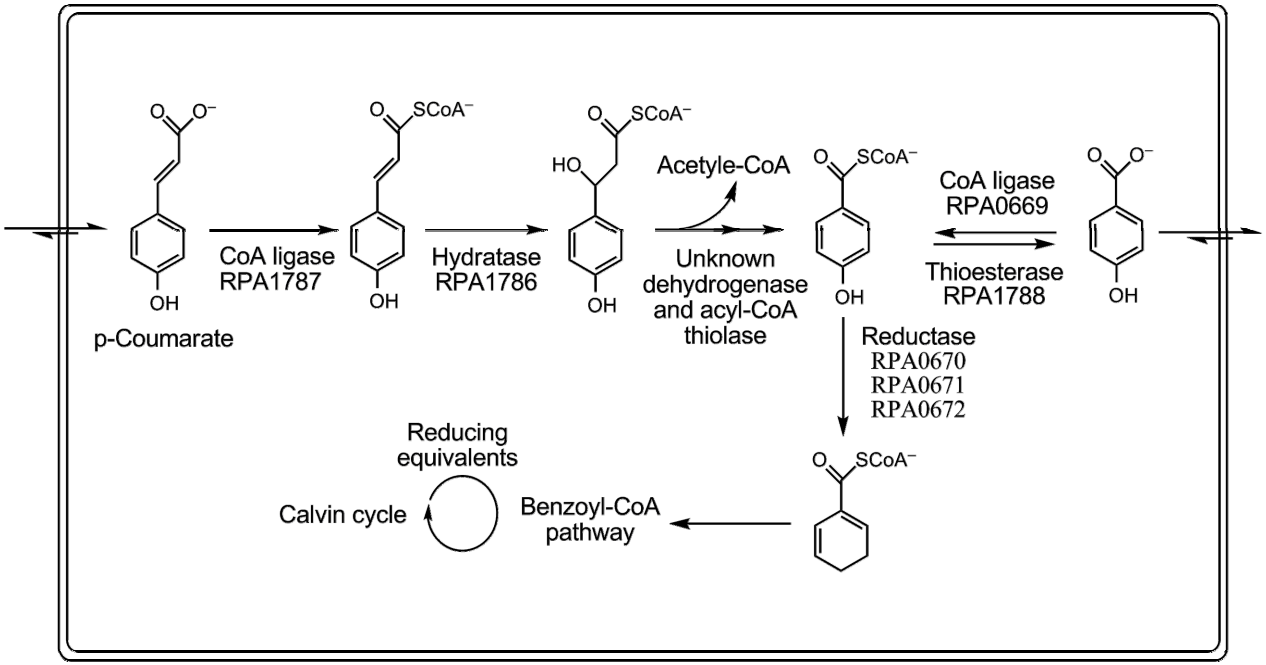


Figure 8.7: Cellular pathways for coumarate catabolism. Coumarate is converted to hydroxyl-benzoyl-CoA, which is either excreted to growth medium after removing CoA or completely degraded to acetyl-CoA through the benzoyl-CoA pathway.

thioesterase activity appeared to be balanced by the almost equivalent induction of the opposing 4-hydroxybenzoate-CoA ligase activity (Table 8.3, Figure 8.7). This supports a previously proposed hypothesis for aromatic degradation: the metabolisable aromatic intermediates are excreted into the growth medium and subsequently absorbed for further degradation. This hypothesis was mainly supported by the transitory presence of aromatic intermediates in the growth medium (Sasikala, 1994). In this study, we measured the coordinated up-regulation of both CoA ligase and thioesterase for 4-hydroxybenzoate, which suggests that 4-hydroxybenzoate is transiently excreted during coumarate catabolism. The cross-membrane transportation of these aromatic compounds is probably facilitated by those up-regulated transporter proteins (Table 8.3).

CONCLUSIONS:

In this study, transcriptomics and quantitative proteomics were combined to identify the genes probably responsible for coumarate degradation and characterize the response of the global metabolic network to the utilization of coumarate as the sole carbon source in *R. palustris*. Gene expression profiles at both mRNA level and protein level were measured in *R. palustris* grown with succinate, benzoate and p-coumarate as the carbon source. 1000 – 2000 genes were quantified by both methods in each binary comparison sample. The induction of the benzoyl-CoA pathway suggests that coumarate catabolism proceeds through benzoyl-CoA as an intermediate metabolite. Conversion of coumarate to benzoyl-CoA was hypothesized to consist of the following steps: CoA ligation, β-oxidization and dehydroxylation. Global gene expression profiling enabled the discovery

of the coumaryol-CoA ligase (RPA1787), the coumaryol-CoA hydratase (RPA1786) and 4-hydroxybenzoate-CoA thioesterase (RPA1788). In addition, two destinations for the intermediate 4-hydroxybenzoate-CoA were suggested: complete degradation through the benzoyl-CoA pathway and transitory excretion to the growth media. It is possible that at the beginning of the culture growth the bacteria mainly derive carbon from β-oxidization of the side chain of coumarate and excrete 4-hydroxybenzoate that is more difficult to metabolize. With the decline of coumarate concentration, bacteria gradually switch to deriving carbon from the aromatic ring.

Chapter 9

Conclusions

Tremendous progress has been made in the field of proteomics over the past five years. Thousands of proteins have been identified and quantified from whole-cell proteomes and sub-cellular proteomes of many organisms (Pandey, 2000; Brunet, 2003; Huber, 2003). A variety of post-translational modifications have been detected and mapped in proteins (Cantin, 2004; Jensen, 2004). Protein interaction networks have been reconstructed for several species (Ho, 2002; Cesareni, 2005). These achievements, together with the advances in genomics, transcriptomics and other "omics" research, have ushered a new era of biological research, in which hypothesis-driven research is combined with discovery-driven research, and targeted in-depth experiments are coupled with large-scale, high-throughput experiments.

The research work described in this dissertation was mainly devoted to development of advanced methodologies for proteomics. Two research directions were pursued: prototyping of various high-performance analytical platforms based on FT-ICR and development of quantitative shotgun proteomics with advanced data analysis algorithms enabling confidence interval estimation. Research progress was achieved with an interdisciplinary effort in three areas of proteomics: mass spectrometry methodology advancement, data analysis algorithm development, and biological applications.

FT-ICR mass spectrometry provides significantly higher mass accuracy, resolving power, and dynamic range in mass measurement than quadrupole ion trap, time-of-flight and triple quadrupole mass spectrometers (Bogdanov, 2005). We believe that the high performance instruments like FT-ICR hold the promise of revolutionizing the field of proteomics by providing deeper proteome coverage, more confident protein identifications, higher protein sequence coverages, and more accurate protein quantification. However, the potential of FT-ICR has not been fully explored for proteomics applications, due to the limited development of FT-ICR MS as a robust and high-throughput MS technology. Our laboratory is equipped with a 9.4 Tesla IonSpec FT-ICR instrument, which has been well maintained with almost 100% up-time. This allowed us to pioneer a variety of novel proteomics measurement methods with FT-ICR and demonstrate the great potential of high performance instruments.

Gas-phase fragmentation is an important method to interrogate the structure of an analyte with mass spectrometry. In FT-ICR, collisionally-activated dissociation (CAD) is typically accomplished within the analyzer cell. An alternative approach of multipole storage-assisted dissociation (MSAD) is afforded by inducing collisional fragmentation in the external multipole that is usually employed for ion accumulation. This MSAD method has the potential for very efficiently dissociating large proteins. To explore the utility of MSAD for interrogating large intact proteins (molecular masses exceeding 100 kDa) and protein mixtures in a multiplexed manner, we have investigated the means of controlling the collisional energy and the fragmentation patterns of seven intact proteins. With protein samples in the low micromolar concentration range, the two major

experimental parameters affecting MSAD in the hexapole region were found to be the dc offset voltage and accumulation time. While low-energy MSAD of intact proteins yields fragment ions similar to SORI-CAD, high-energy MSAD induces sequential fragmentation to yield a rich variety of singly-charged ions in the *m/z* 600-1200 Da region. Each of the proteins examined in this study exhibited their own characteristic MSAD fragmentation pattern, which could be used as a signature of the presence of a given protein, even in a mixture. In addition, any MSAD fragment can be isolated and dissociated further by SORI-CAD in an MS^3-type experiment inside the FTICR analyzer cell. This presents a novel way to interrogate the identities of these fragment ions as well as obtain amino acid sequence tag information that can be used to identify proteins from mixtures. Such MS^3 measurements on a high-performance instrument could potentially be employed in top-down proteomics to obtain sequence information directly from intact proteins, if the data acquisition and interpretation could be automated.

In shotgun proteomics, quadrupole ion traps are generally used to interrogate peptides with tandem mass spectrometry, and the peptides are identified by matching their MS/MS scans with all possible sequences using a database searching algorithm. We believe that high resolution tandem mass spectrometry can not only improve the confidence and sensitivity of peptide identification with database searching algorithms, but also enable discovery of amino acid substitutions, post-translational modifications, and novel peptides through the use of *de novo* sequencing algorithms. A key problem in interpreting a CAD-generated tandem mass spectrum is the separation of y and b ions from each other and from the noise peaks. We developed a graph-theoretic approach for separation of ion

types in high resolution tandem mass spectra. We represent each spectral peak as a node and consider two types of edges: type-1 edge connecting two peaks probably of the same ion types and type-2 edge connecting two peaks probably of different ion types. The problem of ion-separation is formulated and solved as a graph partitioning problem, which is to partition the graph into three subgraphs, representing b, y and others ions, respectively, through maximizing the total weight of type-1 edges, while minimizing the total weight of type-2 edges within each partitioned subgraph. A dynamic programming algorithm was developed to solve this graph partition problem. A set of high resolution peptide tandem mass spectra were acquired with an FT-ICR instrument to test the algorithm. An accuracy of ~90% was achieved for the separation of b and y ions.

The complexity and dynamic range of a proteome necessitate the coupling of mass spectrometry measurements with liquid chromatography separation. We built a nanoscale LC-FT-MS system for shotgun proteomics measurements. The LC was interfaced with FT-ICR with an optimized nanospray ionization source. The LC-FT-MS system has been tested with a protein standard mixture digest and an *R. palustris* proteome digest. We achieved 1–5 ppm mass accuracy, ~200,000 mass resolution, and >100 dynamic range. The high-performance mass measurement enabled accurate *m/z* determination, charge state calculation, isotopic envelope deconvolution, and detection of low-abundance peptides. As a result, monoisotopic neutral masses have been accurately determined for a large number of peptides. The LC-FT-MS measurement was integrated with LC-QIT-MS/MS measurement to enhance peptide identification confidence. Samples were measured with both LC-FT-MS and LC-QIT-MS/MS using the same LC settings.

Peptides observed in the two measurements are correlated by their retention times and masses. By combining accurate mass information from LC-FT-MS and MS/MS data from LC-QIT-MS/MS, more peptides were identified with greater confidence than with the conventional shotgun proteomics measurement employing LC-QIT-MS/MS alone.

In these three studies, we have explored different novel measurements with FT-ICR, including sequence tagging of intact proteins, ion type recognition from high resolution tandem mass spectra, and LC-FT-MS for shotgun proteomics. We demonstrated that high-performance mass spectrometry can help addressing multiple challenges in proteomics, such as amino acid sequencing with MS/MS, high-confidence peptide identification, deep proteome coverage, *etc*. However, we also found that conventional FT-ICR instruments need improvement in the following areas to deliver their promises for proteomics:

- Instrument control software. As FT-ICR has mainly been used for non-biological applications, the instrument control software for FT-ICR lacked many essential features for proteomics measurement, including data-dependent MS/MS, direct communication with LC systems, automated data acquisition sequence, *etc*.
- Tandem mass spectrometry. In LC-MS/MS analysis of a proteome digest, many peptides elute simultaneously; thus, it is essential to have a high MS/MS scan rate to examine as many peptides as possible. However, FT-ICR can only acquire ~20 scans per minute, compared with ~200 scans per minute for a two-dimensional quadrupole ion trap. And due to the lack of automatic gain control, FT-ICR has a limited dynamic range for tandem mass spectrometry analysis.

- Maintainability. FT-ICR requires ultra-high vacuum in the analyzer cell, high field superconductive magnets, and cryogenic temperatures for the superconductive magnet. As a result, the instrumentation for FT-ICR is much more sophisticated and less rugged than other types of mass spectrometers. This limits the use of FT-ICR in a production proteomics pipeline. Also, operation and maintenance of FT-ICR require more specialized expertise than quadrupole ion trap and time-of-flight instruments.
- Cost. The price of a high-field FT-ICR instrument is three times higher than that of a quadrupole ion trap. And FT-ICR requires routine purchase of liquid nitrogen and liquid helium for the superconductive magnet.

Recently, two commercial hybrid high-performance instruments have been developed: ThermoFinnigan LTQ-FTMS (Peterman, 2005) and LTQ-Orbitrap (Erickson, 2006). The two hybrid instruments are equipped with a linear quadrupole ion trap as the first stage and an FT-ICR or an Orbitrap as the final stage. These hybrid instruments have much improved instrument control software and tandem mass spectrometry capability, compared with conventional FT-ICR instruments. It is possible that the next generation of analytical platform for proteomics will be based on such hybrid high-performance instruments.

Quantitative proteomics is a proteomics field that attempts to determine the abundance changes of proteins between two proteomes. Establishment of quantitative shotgun proteomics can be attributed to the development of a variety of stable isotope labeling

techniques. The two proteomes under comparison can be labeled with different stable isotope tags at different stages of sample preparation: cell growth, protein processing, proteolysis, and peptide processing. The LC-MS/MS measurement and the data analysis for peptide and protein identification are largely the same in quantitative shotgun proteomics as in qualitative shotgun proteomics. However, quantitative shotgun proteomics needs new algorithms for peptide abundance ratio estimation and protein abundance ratio estimation.

The abundance ratio between the light and heavy isotopologues of an isotopically labeled peptide is estimated from their selected ion chromatograms. However, quantitative shotgun proteomics measurements yield selected ion chromatograms at highly variable signal-to-noise ratios for tens of thousands of peptides. This challenge calls for algorithms that not only robustly estimate the abundance ratios of different peptides but also rigorously score each abundance ratio for the expected estimation bias and variability. Scoring of the abundance ratios, much like scoring of sequence assignment for tandem mass spectra by peptide identification algorithms, enables filtering of unreliable peptide quantification and use of formal statistical inference in the subsequent protein abundance ratio estimation. We developed a parallel paired covariance algorithm for robust peak detection in selected ion chromatograms. A peak profile is generated for each peptide, which is a scatter-plot of ion intensities measured for the two isotopologues within their chromatographic peaks. Principal component analysis of the peak profile is proposed to estimate the peptide abundance ratio and to score the estimation with the signal-to-noise ratio of the peak profile (profile signal-to-noise ratio). We demonstrate

that the profile signal-to-noise ratio is inversely correlated with the variability and bias of peptide abundance ratio estimation.

We then developed a profile likelihood algorithm to infer the abundance ratios of *proteins* from the abundance ratios of isotopically labeled *peptides* derived from proteolysis. Given multiple quantified peptides for a protein, the profile likelihood algorithm probabilistically weighs the peptide abundance ratios by their inferred estimation variability, accounts for their expected estimation bias, and suppresses contribution from outliers. This algorithm yields maximum likelihood point estimation and profile likelihood confidence interval estimation of protein abundance ratios. This point estimator is more accurate than an estimator based on the average of peptide abundance ratios. The confidence interval estimation provides an "error bar" for each protein abundance ratio that reflects its estimation precision and statistical uncertainty. The accuracy of the point estimation and the precision and confidence level of the interval estimation were benchmarked with standard mixtures of isotopically labeled proteomes. The profile likelihood algorithm was integrated into a quantitative proteomics program, called *ProRata*, freely available to the public.

With the development of the ProRata program, we have significantly improved the data analysis procedure for quantitative shotgun proteomics. However, quantitative shotgun proteomics is still a relatively new field that requires further research efforts along the following directions:

- Stable isotope labeling. The ideal stable isotope labeling method is metabolic labeling, which occurs before any sample processing step and guarantees almost complete labeling efficiency. However, it has only been applied to a limited number of organisms and cell cultures. Although other enzymatic and chemical labeling methods have been developed downstream in the sample processing procedure, they all have certain disadvantages, such as incomplete labeling, side reactions, selected labeling of a subset of peptides, *etc*. Novel stable isotope labeling methods are needed to overcome these limitations.
- LC-MS/MS measurement. Quantitative shotgun proteomics uses the full scan mass spectra for quantification. This demands high dynamic range of the mass spectrometers to match the protein dynamic range and isotopologue dynamic range of an isotopically labeled proteome mixture. Therefore, a high-performance mass spectrometer is critical to obtain high-confidence protein quantification results in quantitative shotgun proteomics.
- Data analysis. The accuracy of point estimation and the precision and confidence of interval estimation for protein abundance ratios can be further improved with advanced algorithms.

Currently, stable isotope labeling is the basis for quantitative shotgun proteomics. Another route for quantitative shotgun proteomics is the label-free approach, in which the proteomes under comparison can be measured in separate LC-MS/MS runs. The label-free approach obviates the need for stable isotope labeling, but requires the construction

of a highly reproducible LC-MS/MS analysis platform and the development of data analysis algorithms for data quality control and measurement bias normalization.

As gene expression consists of transcription and translation, we believe that it is of great value to integrate quantitative proteomics and transcriptomics for global gene expression profiling. The abundance change of the mRNA product and the protein product of a gene can reveal the gene's regulation at the two expression levels. Combination of the results of two independent measurements can also minimize the false discovery rate and reduce the need for additional validation experiments.

We have demonstrated this integrated gene expression profiling with our study on anaerobic catabolism of *p*-coumarate by *R. palustris*. Coumarate is a major phenolic monomer resulting from lignin degradation. *R. palustris* is one of a few known bacteria capable of degrading coumarate under anoxic environments. However, the cellular pathway for coumarate catabolism was unknown in *R. palustris*. It was hypothesized that coumarate is degraded into benzoyl-CoA via β-oxidization, which then proceeds through the benzoyl-CoA pathway. In this study we attempted to identify the genes responsible for coumarate β-oxidization and characterize the impact of coumarate catabolism on the global metabolic network. Transcriptomics and quantitative proteomics were employed to measure the gene expression profiles of *R. palustris* grown with succinate, benzoate, or *p*-coumarate as the carbon source. Between 1000–2000 genes were quantified by both methods in each binary comparison sample. The induction of the benzoyl-CoA pathway supported our hypothesis that coumarate catabolism proceeds through benzoyl-CoA as an

intermediate metabolite. Probable genes for coumarate CoA ligase and coumaryl-CoA hydratase were identified. Interestingly, it was discovered that at least a portion of coumarate molecules are partially degraded into 4-hydroxyl-benzoate, which is likely excreted into the growth medium. Additionally, many other cellular pathways were found to be affected by coumarate catabolism.

The whole body of research work described in this dissertation represents our effort in multiple areas of proteomics: mass spectrometry, informatics, and biology. We have demonstrated the great potential of high-performance mass spectrometry for proteomics, developed a suite of much improved algorithms for quantitative proteomics data analysis, and demonstrated the biological impact of proteomics technology. Our progress was only one step forward in the field of proteomics, and we realize that there are many challenges needed to be addressed in the future, but the challenges also represent exciting research opportunities for chromatographers, mass spectrometrists, statisticians, computer scientists, and biologists. We expect that, in the next few years, proteomics will be truly realized for many organisms, *i.e.* comprehensive characterization of the primary structures, chemical modifications, tertiary structures, cellular locations, physical interactions, and abundances of all proteins under all cellular states of an organism. The wealth of proteomic data for an organism will assist in elucidating the cellular functions of all proteins encoded in the organism's genome.

Proteomics is an integral part of systems biology, together with genomics, transcriptomics, metabolomics and other forthcoming "omics" fields. We envision that

systems biology research will give us a mechanistic understanding of biological processes. Perhaps, based on this understanding, mathematical models can eventually be constructed for the life forms on Earth.

LIST OF REFERENCES

LIST OF REFERENCES

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

Altschul, S. F.; T. L. Madden; A. A. Schaffer; J. Zhang; Z. Zhang; W. Miller and D. J. Lipman (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic acids research* **25**(17): 3389-402.

Andersen, J. S.; C. J. Wilkinson; T. Mayor; P. Mortensen; E. A. Nigg and M. Mann (2003). "Proteomic characterization of the human centrosome by protein correlation profiling." *Nature* **426**(6966): 570-4.

Anderson, N. L.; A. D. Matheson and S. Steiner (2000). "Proteomics: applications in basic and applied biology." *Curr Opin Biotechnol* **11**(4): 408-12.

Astbury, W. T. (1961). "Molecular biology or ultrastructural biology?" *Nature* **190**: 1124.

Avery, M. J. (2003). "Quantitative characterization of differential ion suppression on liquid chromatography/atmospheric pressure ionization mass spectrometric bioanalytical methods." *Rapid Commun Mass Spectrom* **17**(3): 197-201.

Banks, R. E.; M. J. Dunn; D. F. Hochstrasser; J. C. Sanchez; W. Blackstock; D. J. Pappin and P. J. Selby (2000). "Proteomics: new perspectives, new biomedical opportunities." *Lancet* **356**(9243): 1749-56.

Barbosa, M. J.; J. M. Rocha; J. Tramper and R. H. Wijffels (2001). "Acetate as a carbon source for hydrogen production by photosynthetic bacteria." *Journal of biotechnology* **85**(1): 25-33.

Bednar, M. (2000). "DNA microarray technology and application." *Med Sci Monit* **6**(4): 796-800.

Belov, M. E.; M. V. Gorshkov; H. R. Udseth and R. D. Smith (2001). "Controlled ion fragmentation in a 2-D quadrupole ion trap for external ion accumulation in ESI FTICR mass spectrometry." *J Am Soc Mass Spectr* **12**(12): 1312-1319.

Benfey, P. and A. D. Protopapas (2004). *Essentials of Genomics*, Prentice Hall.

Bernard, K. R.; K. R. Jonscher; K. A. Resing and N. G. Ahn (2004). "Methods in functional proteomics: two-dimensional polyacrylamide gel electrophoresis with immobilized pH gradients, in-gel digestion and identification of proteins by mass spectrometry." *Methods in molecular biology (Clifton, N.J* **250**: 263-82.

Bertone, P. and M. Snyder (2005). "Prospects and challenges in proteomics." *Plant physiology* **138**(2): 560-2.

Bino, R. J.; R. D. Hall; O. Fiehn; J. Kopka; K. Saito; J. Draper; B. J. Nikolau; P. Mendes; U. Roessner-Tunali; M. H. Beale; R. N. Trethewey; B. M. Lange; E. S. Wurtele and L. W. Sumner (2004). "Potential of metabolomics as a functional genomics tool." *Trends in plant science* **9**(9): 418-25.

Bjorkhem, I.; R. Blomstrand; S. Eriksson; O. Falk; A. Kallner; L. Svensson and G. Ohman (1980). "Use of isotope dilution--mass spectrometry for accuracy control of different routine methods used in clinical chemistry." *Scandinavian journal of clinical and laboratory investigation* **40**(6): 529-34.

Blackstock, W. P. and M. P. Weir (1999). "Proteomics: quantitative and physical mapping of cellular proteins." *Trends in biotechnology* **17**(3): 121-7.

Boeckmann, B.; A. Bairoch; R. Apweiler; M. C. Blatter; A. Estreicher; E. Gasteiger; M. J. Martin; K. Michoud; C. O'Donovan; I. Phan; S. Pilbout and M. Schneider (2003). "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003." *Nucleic acids research* **31**(1): 365-70.

Bogdanov, B. and R. D. Smith (2005). "Proteomics by FTICR mass spectrometry: top down and bottom up." *Mass spectrometry reviews* **24**(2): 168-200.

Brancia, F. L. (2006). "Recent developments in ion-trap mass spectrometry and related technologies." *Expert review of proteomics* **3**(1): 143-51.

Browne, E. J. (1996). *Charles Darwin: Voyaging*, Princeton University Press.

Brunet, S.; P. Thibault; E. Gagnon; P. Kearney; J. J. Bergeron and M. Desjardins (2003). "Organelle proteomics: looking at less to see more." *Trends in cell biology* **13**(12): 629-38.

Cantin, G. T. and J. R. Yates, 3rd (2004). "Strategies for shotgun identification of post-translational modifications by mass spectrometry." *J Chromatogr A* **1053**(1-2): 7-14.

Cash, P. (2003). "Proteomics of bacterial pathogens." *Advances in biochemical engineering/biotechnology* **83**: 93-115.

Cesareni, G.; A. Ceol; C. Gavrila; L. M. Palazzi; M. Persico and M. V. Schneider (2005). "Comparative interactomics." *FEBS letters* **579**(8): 1828-33.

Chen, H.; K. Tabei and M. M. Siegel (2001a). "Biopolymer sequencing using a triple quadrupole mass spectrometer in the ESI nozzle-skimmer/precursor ion MS/MS mode." *J Am Soc Mass Spectrom* **12**(7): 846-52.

Chen, T.; M. Y. Kao; M. Tepel; J. Rush and G. M. Church (2001b). "A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry." *J Comput Biol* **8**(3): 325-37.

Clauser, K. R.; P. Baker and A. L. Burlingame (1999). "Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching." *Anal Chem* **71**(14): 2871-82.

Cole, S. T. and I. Saint Girons (1994). "Bacterial genomics." *FEMS microbiology reviews* **14**(2): 139-60.

Corthals, G. L.; V. C. Wasinger; D. F. Hochstrasser and J. C. Sanchez (2000). "The dynamic range of protein expression: a challenge for proteomic research." *Electrophoresis* **21**(6): 1104-15.

Cottrell, J. S. (1994). "Protein identification by peptide mass fingerprinting." *Peptide research* **7**(3): 115-24.

Dancik, V.; T. A. Addona; K. R. Clauser; J. E. Vath and P. A. Pevzner (1999). "De novo peptide sequencing via tandem mass spectrometry." *J Comput Biol* **6**(3-4): 327-42.

Darwin, C. R. (1859). *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. London.

Diaz, E. (2004). "Bacterial degradation of aromatic pollutants: a paradigm of metabolic versatility." *Int Microbiol* **7**(3): 173-80.

Dole, M.; L. L. Mack; R. L. Hines; R. C. Mobley; L. D. Ferguson and M. B. Alice (1968). "Molecular beams of macroions." *J Chemical Physics* **49**: 2240-2249.

Dutton, P. L. and W. C. Evans (1967). "Dissimilation of aromatic substrates by Rhodopseudomonas palustris." *The Biochemical journal* **104**(2): 30P-31P.

Elder, D. J. and D. J. Kelly (1994). "The bacterial degradation of benzoic acid and benzenoid compounds under anaerobic conditions: unifying trends and new perspectives." *FEMS microbiology reviews* **13**(4): 441-68.

Eliason, S. R. (1993). *Maximum Likelihood Estimation: Logic and Practice*. Newbury Park, CA, Sage Publications Inc.

Eng, J. K.; A. L. Mccormack and J. R. Yates (1994). "An Approach to Correlate Tandem Mass-Spectral Data of Peptides with Amino-Acid-Sequences in a Protein Database." *J Am Soc Mass Spectr* **5**(11): 976-989.

Erickson, B. (2006). "Linear ion trap/Orbitrap mass spectrometer." *Anal Chem* **78**(7): 2089.

Erlich, H. A. (1989). "Polymerase chain reaction." *Journal of clinical immunology* **9**(6): 437-47.

Fenn, J. B.; M. Mann; C. K. Meng; S. F. Wong and C. M. Whitehouse (1989). "Electrospray ionization for mass spectrometry of large biomolecules." *Science* **246**(4926): 64-71.

Fenyo, D. (2000). "Identifying the proteome: software tools." *Curr Opin Biotechnol* **11**(4): 391-5.

Fiehn, O. (2002). "Metabolomics--the link between genotypes and phenotypes." *Plant molecular biology* **48**(1-2): 155-71.

Gatlin, C. L.; J. K. Eng; S. T. Cross; J. C. Detter and J. R. Yates, 3rd (2000). "Automated identification of amino acid sequence variations in proteins by HPLC/microspray tandem mass spectrometry." *Anal Chem* **72**(4): 757-63.

Gauthier, J. W.; T. R. Trautman and D. B. Jacobson (1991). "Sustained Off-Resonance Irradiation for Collision-Activated Dissociation Involving Fourier-Transform Mass-Spectrometry - Collision-Activated Dissociation Technique That Emulates Infrared Multiphoton Dissociation." *Anal Chim Acta* **246**(1): 211-225.

Ginter, J. M.; F. Zhou and M. V. Johnston (2004). "Generating protein sequence tags by combining cone and conventional collision induced dissociation in a quadrupole time-of-flight mass spectrometer." *J Am Soc Mass Spectrom* **15**(10): 1478-86.

Goodman, L. (1999). "Hypothesis-limited research." *Genome research* **9**(8): 673-4.

Gooley, A. A. and N. H. Packer (1997). The importance of co- and post-translational modifications
in proteome projects. *Proteome Research: New Frontiers in Functional Genomics.* W. R. Wilkins, K. L. Williams, R. D. Appel and D. F. Hochstrasser. New York, Springer-Verlag**:** 65-91.

Gronborg, M.; T. Z. Kristiansen; A. Stensballe; J. S. Andersen; O. Ohara; M. Mann; O. N. Jensen and A. Pandey (2002). "A mass spectrometry-based proteomic approach for identification of serine/threonine-phosphorylated proteins by enrichment with phospho-specific antibodies: identification of a novel protein, Frigg, as a protein kinase A substrate." *Mol Cell Proteomics* **1**(7): 517-27.

Gygi, S. P.; B. Rist; S. A. Gerber; F. Turecek; M. H. Gelb and R. Aebersold (1999). "Quantitative analysis of complex protein mixtures using isotope-coded affinity tags." *Nature Biotechnology* **17**(10): 994-999.

Hager, D. B.; N. J. Dovichi; J. Klassen and P. Kebarle (1994). "Droplet electrospray mass-spectrometry." *Anal Chem* **66**: 3944-3949.

Hakansson, K.; J. Axelsson; M. Palmblad and P. Hakansson (2000). "Mechanistic studies of multipole storage assisted dissociation." *J Am Soc Mass Spectrom* **11**(3): 210-7.

Hall, D. L. and S. A. H. McMullen (2004). *Mathematical techniques in multisensor data fusion*, Artech House.

Hall, M. P.; S. Ashrafi; I. Obegi; R. Petesch; J. N. Peterson and L. V. Schneider (2003). ""Mass defect" tags for biomolecular mass spectrometry." *J Mass Spectrom* **38**(8): 809-16.

Han, D. K.; J. Eng; H. Zhou and R. Aebersold (2001). "Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry." *Nature Biotechnology* **19**(10): 946-51.

Harrington, C. A.; C. Rosenow and J. Retief (2000). "Monitoring gene expression using DNA microarrays." *Curr Opin Microbiol* **3**(3): 285-91.

Hartley, R. D. and C. W. Ford (1989). Phenolic constituents of plant cell walls and wall biodegradability. *Plant cell wall polymers: biogenesis and biodegradation*. N. G. Lewis and M. G. Paice. Washington, D.C., American Chemical Society**:** 137-145.

Harwood, C. S.; G. Burchhardt; H. Herrmann and G. Fuchs (1999). "Anaerobic metabolism of aromatic compounds via the benzoyl-CoA pathway." *FEMS Microbiology Reviews* **22**: 439-458.

Harwood, C. S. and J. Gibson (1988). "Anaerobic and aerobic metabolism of diverse aromatic compounds by the photosynthetic bacterium Rhodopseudomonas palustris." *Appl Environ Microbiol* **54**(3): 712-7.

Harwood, C. S. and J. Gibson (1997). "Shedding light on anaerobic benzene ring degradation: a process unique to prokaryotes?" *J Bacteriol* **179**(2): 301-9.

Hendrickson, C. L. and M. R. Emmett (1999). "Electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry." *Annual review of physical chemistry* **50**: 517-36.

Hilhorst, R.; C. Laane and C. Veeger (1982). "Photosensitized production of hydrogen by hydrogenase in reversed micelles." *Proceedings of the National Academy of Sciences of the United States of America* **79**(12): 3927-3930.

Hillenkamp, F. and M. Karas (1990). "Mass spectrometry of peptides and proteins by matrix-assisted ultraviolet laser desorption/ionization." *Methods in enzymology* **193**: 280-95.

Ho, Y.; A. Gruhler; A. Heilbut; G. D. Bader; L. Moore; S. L. Adams; A. Millar; P. Taylor; K. Bennett; K. Boutilier; L. Yang; C. Wolting; I. Donaldson; S. Schandorff; J. Shewnarane; M. Vo; J. Taggart; M. Goudreault; B. Muskat; C. Alfarano; D. Dewar; Z. Lin; K. Michalickova; A. R. Willems; H. Sassi; P. A. Nielsen; K. J. Rasmussen; J. R. Andersen; L. E. Johansen; L. H. Hansen; H. Jespersen; A. Podtelejnikov; E. Nielsen; J. Crawford; V. Poulsen; B. D. Sorensen; J. Matthiesen; R. C. Hendrickson; F. Gleeson; T. Pawson; M. F. Moran; D. Durocher; M. Mann; C. W. Hogue; D. Figeys and M. Tyers (2002). "Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry." *Nature* **415**(6868): 180-3.

Hochstrasser, D. F.; J. C. Sanchez and R. D. Appel (2002). "Proteomics and its trends facing nature's complexity." *Proteomics* **2**(7): 807-12.

Hoffmann, E. D. and V. Stroobant (2001). *Mass Spectrometry: Principles and Applications*, John Wiley & Sons.

Hofstadler, S. A.; J. J. Drader; H. Gaus; J. C. Hannis and K. A. Sannes-Lowery (2003). "Alternative approaches to infrared multiphoton dissociation in an external ion reservoir." *J Am Soc Mass Spectrom* **14**(12): 1413-23.

Hofstadler, S. A.; K. A. Sannes-Lowery and R. H. Griffey (1999). "Infrared multiphoton dissociation in an external ion reservoir." *Anal Chem* **71**(11): 2067-70.

Horn, D. M.; R. A. Zubarev and F. W. McLafferty (2000a). "Automated de novo sequencing of proteins by tandem high-resolution mass spectrometry." *P Natl Acad Sci USA* **97**(19): 10313-10317.

Horn, D. M.; R. A. Zubarev and F. W. McLafferty (2000b). "Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules." *J Am Soc Mass Spectr* **11**(4): 320-332.

Huber, L. A.; K. Pfaller and I. Vietor (2003). "Organelle proteomics: implications for subcellular fractionation in proteomics." *Circulation research* **92**(9): 962-8.

Hunt, D. F.; A. M. Buko; J. M. Ballard; J. Shabanowitz and A. B. Giordani (1981). "Sequence analysis of polypeptides by collision activated dissociation on a triple quadrupole mass spectrometer." *Biomed Mass Spectrom* **8**(9): 397-408.

Ideker, T.; T. Galitski and L. Hood (2001). "A new approach to decoding life: systems biology." *Annual review of genomics and human genetics* **2**: 343-72.

Iribarne, J. V. and B. A. Thomson (1976). "On the evaporation of charged ions from small droplets." *J Chemical Physics* **64**: 2287-2294.

Jensen, O. N. (2004). "Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry." *Current opinion in chemical biology* **8**(1): 33-41.

Jolliffe, I. T. (2002). Principal component analysis. *Springer series in statistics*, Springer**:** 34-36.

Julka, S. and F. Regnier (2004). "Quantification in proteomics through stable isotope coding: a review." *Journal of proteome research* **3**(3): 350-63.

Jungblut, P. R.; U. Zimny-Arndt; E. Zeindl-Eberhart; J. Stulik; K. Koupilova; K. P. Pleissner; A. Otto; E. C. Muller; W. Sokolowska-Kohler; G. Grabher and G. Stoffler (1999). "Proteomics in human disease: cancer, heart and infectious diseases." *Electrophoresis* **20**(10): 2100-10.

Kebarle, P. and M. Peschke (1999). "On the mechanisms by which the charged droplets produced by electrospray lead to gas phase ions." *Analytica Chemica Acta* **20070**: 1-25.

Kelleher, N. L. (2004). "Top-down proteomics." *Anal Chem* **76**(11): 197A-203A.

Kelleher, N. L.; H. Y. Lin; G. A. Valaskovic; D. J. Aaserud; E. K. Fridriksson and F. W. McLafferty (1999). "Top down versus bottom up protein characterization by tandem high-resolution mass spectrometry." *J Am Chem Soc* **121**(4): 806-812.

Keller, K. M.; J. S. Brodbelt; R. L. Hettich and G. J. Van Berkel (2004). "Comparison of sustained off-resonance irradiation collisionally activated dissociation and multipole storage-assisted dissociation for top-down protein analysis." *J Mass Spectrom* **39**(4): 402-11.

Kirk, T. K. (1984). Degradation of lignin. *Microbial degradation of organic compounds.* D. T. Gibson. New York, N.Y., Marcel Dekker, Inc.**:** 399-437.

Kurian, K. M.; C. J. Watson and A. H. Wyllie (1999). "DNA chip technology." *The Journal of pathology* **187**(3): 267-71.

Larimer, F. W.; P. Chain; L. Hauser; J. Lamerdin; S. Malfatti; L. Do; M. L. Land; D. A. Pelletier; J. T. Beatty; A. S. Lang; F. R. Tabita; J. L. Gibson; T. E. Hanson; C. Bobst; J. L. Torres; C. Peres; F. H. Harrison; J. Gibson and C. S. Harwood (2004). "Complete genome sequence of the metabolically versatile photosynthetic bacterium Rhodopseudomonas palustris." *Nat Biotechnol* **22**(1): 55-61.

Larose, D. T. (2006). *Data Mining Methods and Models*, Wiley-IEEE Press.

Lasonder, E.; Y. Ishihama; J. S. Andersen; A. M. Vermunt; A. Pain; R. W. Sauerwein; W. M. Eling; N. Hall; A. P. Waters; H. G. Stunnenberg and M. Mann (2002). "Analysis of the Plasmodium falciparum proteome by high-accuracy mass spectrometry." *Nature* **419**(6906): 537-42.

Lastowski, K. and W. Makalowski (2000). "Methodological function of hypotheses in science: old ideas in new cloth." *Genome research* **10**(3): 273-4.

Lawson, A. M.; C. K. Kim; W. Richmond; D. M. Samson; K. D. R. Setchell and A. C. S. Thomas (1980). *Isotope dilution mass spectrometry as a basis for accuracy in clinical chemistry*. Current Developments in the Clinical Applications of HPLC, GC and MS., Middlesex, UK, Academic Press, London.

Li, X. J.; H. Zhang; J. A. Ranish and R. Aebersold (2003). "Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry." *Analytical Chemistry* **75**(23): 6648-57.

Lilley, K. S. and D. R. Griffiths (2003). "Proteomics in Drosophila melanogaster." *Briefings in functional genomics & proteomics* **2**(2): 106-13.

Link, A. J.; J. Eng; D. M. Schieltz; E. Carmack; G. J. Mize; D. R. Morris; B. M. Garvik and J. R. Yates, 3rd (1999). "Direct analysis of protein complexes using mass spectrometry." *Nat Biotechnol* **17**(7): 676-82.

Little, D. P.; J. P. Speir; M. W. Senko; P. B. Oconnor and F. W. Mclafferty (1994). "Infrared Multiphoton Dissociation of Large Multiply-Charged Ions for Biomolecule Sequencing." *Analytical Chemistry* **66**(18): 2809-2815.

Lowry, O. H.; N. J. Rosebrough; A. L. Farr and R. J. Randall (1951). "Protein measurement with the Folin phenol reagent." *J Biol Chem* **193**(1): 265-75.

Lu, B. and T. Chen (2003). "A suboptimal algorithm for de novo peptide sequencing via tandem mass spectrometry." *J Comput Biol* **10**(1): 1-12.

Ma, B.; K. Zhang; C. Hendrie; C. Liang; M. Li; A. Doherty-Kirby and G. Lajoie (2003). "PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry." *Rapid Commun Mass Spectrom* **17**(20): 2337-42.

MacCoss, M. J.; W. H. McDonald; A. Saraf; R. Sadygov; J. M. Clark; J. J. Tasto; K. L. Gould; D. Wolters; M. Washburn; A. Weiss; J. I. Clark and J. R. Yates, 3rd (2002). "Shotgun identification of protein modifications from protein complexes and lens tissue." *Proceedings of the National Academy of Sciences of the United States of America* **99**(12): 7900-5.

MacCoss, M. J.; C. C. Wu; H. Liu; R. Sadygov and J. R. Yates, 3rd (2003). "A correlation algorithm for the automated quantitative analysis of shotgun proteomics data." *Analytical Chemistry* **75**(24): 6912-21.

Mann, M. and O. N. Jensen (2003). "Proteomic analysis of post-translational modifications." *Nat Biotechnol* **21**(3): 255-61.

Mann, M. and M. Wilm (1994). "Error Tolerant Identification of Peptides in Sequence Databases by Peptide Sequence Tags." *Analytical Chemistry* **66**(24): 4390-4399.

Marko-Varga, G. and T. E. Fehniger (2004). "Proteomics and disease--the challenges for technology and discovery." *Journal of proteome research* **3**(2): 167-78.

Marques, J. P. (2001). *Pattern Recognition: Concepts, Methods and Applications*, Springer.

Marshall, A. G. (1985). "Fourier transform ion cyclotron resonance mass spectrometry." *Accounts of Chemical Research* **18**: 316-322.

Marshall, A. G.; C. L. Hendrickson and G. S. Jackson (1998). "Fourier transform ion cyclotron resonance mass spectrometry: a primer." *Mass spectrometry reviews* **17**(1): 1-35.

Martin, W. J. (1989). "New technologies for large-genome sequencing." *Genome / National Research Council Canada = Genome / Conseil national de recherches Canada* **31**(2): 1073-80.

Masselon, C.; G. A. Anderson; R. Harkewicz; J. E. Bruce; L. Pasa-Tolic and R. D. Smith (2000). "Accurate mass multiplexed tandem mass spectrometry for high-throughput polypeptide identification from mixtures." *Anal Chem* **72**(8): 1918-24.

McDonald, W. H.; R. Ohi; D. T. Miyamoto; T. J. Mitchison and J. R. Yates (2002). "Comparison of three directly coupled HPLC MS/MS strategies for identification of proteins from complex mixtures: single-dimension LC-MS/MS, 2-phase MudPIT, and 3-phase MudPIT." *International Journal of Mass Spectrometry* **219**(1): 245-251.

McDonald, W. H. and J. R. Yates, 3rd (2003). "Shotgun proteomics: integrating technologies to answer biological questions." *Current opinion in molecular therapeutics* **5**(3): 302-9.

McDonnell, L. A.; A. E. Giannakopulos; P. J. Derrick; Y. O. Tsybin and P. Hakansson (2002). "A theoretical investigation of the kinetic energy of ions trapped in a radio-frequency hexapole ion trap." *Eur J Mass Spectrom* **8**(2): 181-189.

McFarland, M. A.; C. L. Hendrickson and A. G. Marshall (2004). "Ion "threshing": Collisionally activated dissociation in an external octopole ion trap by oscillation of an axial electric potential gradient." *Analytical Chemistry* **76**(6): 1545-1549.

McKusick, V. A. (1997). "Genomics: structural and functional studies of genomes." *Genomics* **45**(2): 244-9.

Meng, F.; B. J. Cargile; L. M. Miller; A. J. Forbes; J. R. Johnson and N. L. Kelleher (2001). "Informatics and multiplexing of intact protein identification in bacteria and the archaea." *Nat Biotechnol* **19**(10): 952-7.

Mortz, E.; P. B. O'Connor; P. Roepstorff; N. L. Kelleher; T. D. Wood; F. W. McLafferty and M. Mann (1996). "Sequence tag identification of intact proteins by matching tanden mass spectral data against sequence data bases." *Proc Natl Acad Sci U S A* **93**(16): 8264-7.

Muddiman, D. C.; B. M. Huang; G. A. Anderson; A. Rockwood; S. A. Hofstadler; M. S. WeirLipton; A. Proctor; Q. Y. Wu and R. D. Smith (1997). "Application of sequential paired covariance to liquid chromatography mass spectrometry data - Enhancements in both the signal-to-noise ratio and the resolution of analyte peaks in the chromatogram." *J Chromatogr A* **771**(1-2): 1-7.

Muddiman, D. C.; A. L. Rockwood; Q. Gao; J. C. Severs; H. R. Udseth; R. D. Smith and A. Proctor (1995). "Application of Sequential Paired Covariance to Capillary Electrophoresis Electrospray-Ionization Time-of-Flight Mass-Spectrometry - Unraveling the Signal from the Noise in the Electropherogram." *Analytical Chemistry* **67**(23): 4371-4375.

Muller, P. (1994). "Glossary of terms used in physical organic chemistry." *Pure and Applied Chemistry* **66**(5): 1132.

Nesvizhskii, A. I. and R. Aebersold (2004). "Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS." *Drug discovery today* **9**(4): 173-81.

Oda, Y.; K. Huang; F. R. Cross; D. Cowburn and B. T. Chait (1999). "Accurate quantitation of protein expression and site-specific phosphorylation." *P Natl Acad Sci USA* **96**(12): 6591-6596.

Oda, Y.; S. K. Samanta; F. E. Rey; L. Wu; X. Liu; T. Yan; J. Zhou and C. S. Harwood (2005). "Functional genomic analysis of three nitrogenase isozymes in the photosynthetic bacterium Rhodopseudomonas palustris." *J Bacteriol* **187**(22): 7784-94.

Oda, Y.; B. Star; L. A. Huisman; J. C. Gottschal and L. J. Forney (2003). "Biogeography of the purple nonsulfur bacterium Rhodopseudomonas palustris." *Appl Environ Microbiol* **69**(9): 5186-91.

Ong, S. E.; B. Blagoev; I. Kratchmarova; D. B. Kristensen; H. Steen; A. Pandey and M. Mann (2002). "Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics." *Mol Cell Proteomics* **1**(5): 376-386.

Ong, S. E.; I. Kratchmarova and M. Mann (2003). "Properties of C-13-substituted arginine in stable isotope labeling by amino acids in cell culture (SILAC)." *J Proteome Res* **2**(2): 173-181.

Ong, S. E. and M. Mann (2005a). "Mass spectrometry-based proteomics turns quantitative." *Nat Chem Biol* **1**(5): 252-262.

Ong, S. E. and M. Mann (2005b). "Mass spectrometry-based proteomics turns quantitative." *Nat Chem Biol* **1**(5): 252-62.

Overhage, J.; H. Priefert and A. Steinbuchel (1999). "Biochemical and genetic analyses of ferulic acid catabolism in Pseudomonas sp. Strain HR199." *Appl Environ Microbiol* **65**(11): 4837-47.

Palmblad, M.; K. Hakansson; P. Hakansson; X. D. Feng; H. J. Cooper; A. E. Giannakopulos; P. S. Green and P. J. Derrick (2000). "A 9.4 T Fourier transform ion cyclotron resonance mass spectrometer: description and performance." *Eur J Mass Spectrom* **6**(3): 267-275.

Pan, C.; G. Kora; W. H. McDonald; D. L. Tabb; G. B. Hurst; N. C. VerBerkmoes; D. A. Pelletier; N. F. Samatova and R. L. Hettich (2006a). "ProRata: A quantitative proteomics program for accurate protein abundance ratio estimation with confidence interval evaluation." *Anal Chem*.

Pan, C.; G. Kora; D. L. Tabb; D. A. Pelletier; W. H. McDonald; G. B. Hurst; R. L. Hettich and N. F. Samatova (2006b). "Robust Estimation of Peptide Abundance Ratios and Rigorous Scoring of Their Variability and Bias in Quantitative Shotgun Proteomics." *Anal Chem*.

Pandey, A. and M. Mann (2000). "Proteomics to study genes and genomes." *Nature* **405**(6788): 837-46.

Pappin, D. J. (1997). "Peptide mass fingerprinting using MALDI-TOF mass spectrometry." *Methods in molecular biology (Clifton, N.J* **64**: 165-73.

Parales, R. E. and C. S. Harwood (2002). "Bacterial chemotaxis to pollutants and plant-derived aromatic molecules." *Curr Opin Microbiol* **5**(3): 266-73.

Pedrioli, P. G. A.; J. K. Eng; R. Hubley; M. Vogelzang; E. W. Deutsch; B. Raught; B. Pratt; E. Nilsson; R. H. Angeletti; R. Apweiler; K. Cheung; C. E. Costello; H. Hermjakob; S. Huang; R. K. Julian; E. Kapp; M. E. McComb; S. G. Oliver; G. Omenn; N. W. Paton; R. Simpson; R. Smith; C. F. Taylor; W. M. Zhu and R. Aebersold (2004). "A common open representation of mass spectrometry data and its application to proteomics research." *Nature Biotechnology* **22**(11): 1459-1466.

Peng, J.; D. Schwartz; J. E. Elias; C. C. Thoreen; D. Cheng; G. Marsischky; J. Roelofs; D. Finley and S. P. Gygi (2003). "A proteomics approach to understanding protein ubiquitination." *Nat Biotechnol* **21**(8): 921-6.

Perkins, D. N.; D. J. Pappin; D. M. Creasy and J. S. Cottrell (1999). "Probability-based protein identification by searching sequence databases using mass spectrometry data." *Electrophoresis* **20**(18): 3551-67.

Peterman, S. M.; C. P. Dufresne and S. Horning (2005). "The use of a hybrid linear trap/FT-ICR mass spectrometer for on-line high resolution/high mass accuracy bottom-up sequencing." *J Biomol Tech* **16**(2): 112-24.

Pevzner, P. A.; V. Dancik and C. L. Tang (2000). "Mutation-tolerant protein identification by mass spectrometry." *J Comput Biol* **7**(6): 777-87.

Porubleva, L. and P. R. Chitnis (2000). "Proteomics: a powerful tool in the post-genomic era." *Indian journal of biochemistry & biophysics* **37**(6): 360-8.

Press, W. H.; B. P. Flannery; S. A. Teukolsky and W. T. Vetterling (2002). Numerical Recipes in C++, Cambridge University Press**:** 655-660.

Price, W. D.; P. D. Schnier and E. R. Williams (1996). "Tandem mass spectrometry of large biomolecule ions by blackbody infrared radiative dissociation." *Anal Chem* **68**(5): 859-866.

Purvine, S.; J. T. Eppel; E. C. Yi and D. R. Goodlett (2003). "Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer." *Proteomics* **3**(6): 847-50.

Rabilloud, T. (2002). "Two-dimensional gel electrophoresis in proteomics: old, old fashioned, but it still climbs up the mountains." *Proteomics* **2**(1): 3-10.

Raska, C. S.; C. E. Parker; C. Huang; J. Han; G. L. Glish; M. Pope and C. H. Borchers (2002). "Pseudo-MS3 in a MALDI orthogonal quadrupole-time of flight mass spectrometer." *J Am Soc Mass Spectrom* **13**(9): 1034-41.

Regnier, F. E.; L. Riggs; R. Zhang; L. Xiong; P. Liu; A. Chakraborty; E. Seeley; C. Sioma and R. A. Thompson (2002). "Comparative proteomics based on stable isotope labeling and affinity selection." *J Mass Spectrom* **37**(2): 133-45.

Reinders, J.; U. Lewandrowski; J. Moebius; Y. Wagner and A. Sickmann (2004). "Challenges in mass spectrometry-based proteomics." *Proteomics* **4**(12): 3686-703.

Reo, N. V. (2002). "NMR-based metabolomics." *Drug and chemical toxicology* **25**(4): 375-82.

Ross, P. L.; Y. N. Huang; J. N. Marchese; B. Williamson; K. Parker; S. Hattan; N. Khainovski; S. Pillai; S. Dey; S. Daniels; S. Purkayastha; P. Juhasz; S. Martin; M. Bartlet-Jones; F. He; A. Jacobson and D. J. Pappin (2004). "Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents." *Mol Cell Proteomics* **3**(12): 1154-69.

Salser, W. A. (1974). "DNA sequencing techniques." *Annual review of biochemistry* **43**(0): 923-65.

Sannes-Lowery, K.; R. H. Griffey; G. H. Kruppa; J. P. Speir and S. A. Hofstadler (1998). "Multipole storage assisted dissociation, a novel in-source dissociation technique for electrospray ionization generated ions." *Rapid Commun Mass Spectrom* **12**(23): 1957-61.

Sannes-Lowery, K. A. and S. A. Hofstadler (2000). "Characterization of multipole storage assisted dissociation: implications for electrospray ionization mass spectrometry characterization of biomolecules." *J Am Soc Mass Spectrom* **11**(1): 1-9.

Santoni, V.; M. Molloy and T. Rabilloud (2000). "Membrane proteins and proteomics: un amour impossible?" *Electrophoresis* **21**(6): 1054-70.

Sargent, M.; R. Harte and C. Harrington (2002). *Guidelines for achieving high accuracy in isotope dilution mass spectrometry (IDMS)*, Cambridge : Royal Society of Chemistry.

Sarkanen, K. V. and C. H. Ludwig (1971). Definition and nomenclature. *Lignins: occurence, formation, structure and reactions*. New York, John Wiley & Sons**:** 1-18.

Sasikala, C.; C. V. Ramana and P. R. Rao (1994). "Photometabolism of Heterocyclic Aromatic Compounds by Rhodopseudomonas palustris OU 11." *Appl Environ Microbiol* **60**(6): 2187-2190.

Schmid, M. B. (2002). "Structural proteomics: the potential of high-throughput structure determination." *Trends in microbiology* **10**(10 Suppl): S27-31.

Schulze, W. X. and M. Mann (2004). "A novel proteomic screen for peptide-protein interactions." *J Biol Chem* **279**(11): 10756-10764.

Schwartz, J. C. and I. Jardine (1996). "Quadrupole ion trap mass spectrometry." *Methods in enzymology* **270**: 552-86.

Schwartz, J. C.; M. W. Senko and J. E. Syka (2002). "A two-dimensional quadrupole ion trap mass spectrometer." *J Am Soc Mass Spectrom* **13**(6): 659-69.

Senko, M. W.; C. L. Hendrickson; M. R. Emmett; S. D. H. Shi and A. G. Marshall (1997). "External accumulation of ions for enhanced electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry." *J Am Soc Mass Spectr* **8**(9): 970-976.

Senko, M. W.; J. P. Speir and F. W. Mclafferty (1994). "Collisional Activation of Large Multiply-Charged Ions Using Fourier-Transform Mass-Spectrometry." *Analytical Chemistry* **66**(18): 2801-2808.

Strader, M. B.; N. C. Verberkmoes; D. L. Tabb; H. M. Connelly; J. W. Barton; B. D. Bruce; D. A. Pelletier; B. H. Davison; R. L. Hettich; F. W. Larimer and G. B. Hurst (2004). "Characterization of the 70S Ribosome from Rhodopseudomonas palustris using an integrated "top-down" and "bottom-up" mass spectrometric approach." *Journal of proteome research* **3**(5): 965-78.

Suckau, D. and A. Resemann (2003). "T3-sequencing: targeted characterization of the N- and C-termini of undigested proteins by mass spectrometry." *Anal Chem* **75**(21): 5817-24.

Syka, J. E.; J. J. Coon; M. J. Schroeder; J. Shabanowitz and D. F. Hunt (2004). "Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry." *Proceedings of the National Academy of Sciences of the United States of America* **101**(26): 9528-33.

Tabb, D. L.; W. H. McDonald and J. R. Yates (2002). "DTASelect and contrast: Tools for assembling and comparing protein identifications from shotgun proteomics." *J Proteome Res* **1**(1): 21-26.

Tabb, D. L.; L. L. Smith; L. A. Breci; V. H. Wysocki; D. Lin and J. R. Yates, 3rd (2003). "Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides." *Anal Chem* **75**(5): 1155-63.

Taflin, D. C.; T. L. Ward and E. J. Davis (1989). "Electrified droplet fission and the Rayleigh limit." *Langmuir* **5**: 376-384.

Tao, W. A. and R. Aebersold (2003). "Advances in quantitative proteomics via stable isotope tagging and mass spectrometry." *Curr Opin Biotechnol* **14**(1): 110-8.

Taylor, J. A. and R. S. Johnson (1997). "Sequence database searches via de novo peptide sequencing by tandem mass spectrometry." *Rapid Commun Mass Spectrom* **11**(9): 1067-75.

Taylor, J. A. and R. S. Johnson (2001). "Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry." *Anal Chem* **73**(11): 2594-604.

Thorne, G. C. G., Simon J. (1986). "Evaluation of smoothing routines for the optimization of selected ion monitoring data." *Biomedical & Environmental Mass Spectrometry* **13**(11): 605-9.

Thorne, G. C. G., Simon J.; Payne, Peter A. (1984). "Approaches to the improvement of quantitative precision in selected ion monitoring: high resolution applications." *Biomedical Mass Spectrometry* **11**(8): 415-20.

Uchiki, T.; R. Hettich; V. Gupta and C. Dealwis (2002). "Characterization of monomeric and dimeric forms of recombinant Sm1lp-histag protein by electrospray mass spectrometry." *Anal Biochem* **301**(1): 35-48.

Unlu, M.; M. E. Morgan and J. S. Minden (1997). "Difference gel electrophoresis: A single gel method for detecting changes in protein extracts." *Electrophoresis* **18**(11): 2071-2077.

Venzon DJ, M. A. (1988). "A method for computing profile-likelihood based confidence intervals." *Applied Statistics* **37**(1): 87-94.

Venzon, D. J. and A. H. Moolgavkar (1988). "A method for computing profile-likelihood based confidence intervals." *Applied Statistics* **37**(1): 87-94.

VerBerkmoes, N. C.; J. L. Bundy; L. Hauser; K. G. Asano; J. Razumovskaya; F. Larimer; R. L. Hettich and J. L. Stephenson, Jr. (2002). "Integrating 'top-down" and "bottom-up" mass spectrometric approaches for proteomic analysis of Shewanella oneidensis." *Journal of proteome research* **1**(3): 239-52.

VerBerkmoes, N. C.; H. M. Connelly; C. Pan and R. L. Hettich (2004). "Mass spectrometric approaches for characterizing bacterial proteomes." *Expert review of proteomics* **1**(4): 433-47.

Wang, Y.; B. M. Balgley; P. A. Rudnick and C. S. Lee (2005). "Effects of chromatography conditions on intact protein separations for top-down proteomics." *J Chromatogr A* **1073**(1-2): 35-41.

Washburn, M. P.; D. Wolters and J. R. Yates, 3rd (2001). "Large-scale analysis of the yeast proteome by multidimensional protein identification technology." *Nat Biotechnol* **19**(3): 242-7.

Washburn, M. P. and J. R. Yates, 3rd (2000). "Analysis of the microbial proteome." *Curr Opin Microbiol* **3**(3): 292-7.

Watt, S. A.; T. Patschkowski; J. Kalinowski and K. Niehaus (2003). "Qualitative and quantitative proteomics by two-dimensional gel electrophoresis, peptide mass fingerprint and a chemically-coded affinity tag (CCAT)." *Journal of biotechnology* **106**(2-3): 287-300.

Wolters, D. A.; M. P. Washburn and J. R. Yates, 3rd (2001). "An automated multidimensional protein identification technology for shotgun proteomics." *Anal Chem* **73**(23): 5683-90.

Xu, P.; X. M. Qian; Y. X. Wang and Y. B. Xu (1996). "Modelling for waste water treatment by Rhodopseudomonas palustris Y6 immobilized on fibre in a columnar bioreactor." *Applied microbiology and biotechnology* **44**(5): 676-82.

Yao, X. D.; A. Freas; J. Ramirez; P. A. Demirev and C. Fenselau (2001). "Proteolytic O-18 labeling for comparative proteomics: Model studies with two serotypes of adenovirus." *Analytical Chemistry* **73**(13): 2836-2842.

Yates, J. R., 3rd (2000). "Mass spectrometry. From genomics to proteomics." *Trends Genet* **16**(1): 5-8.

Zubarev, R. A.; N. L. Kelleher and F. W. McLafferty (1998). "Electron capture dissociation of multiply charged protein cations. A nonergodic process." *J Am Chem Soc* **120**(13): 3265-3266.

Zylstra, G. J. and D. T. Gibson (1991). "Aromatic hydrocarbon degradation: a molecular approach." *Genetic engineering* **13**: 183-203.

VITA

Chongle Pan was born in Nanchang, Jiangxi, China on January 2nd, 1980. He went to Songbo elementary school from 1986 to 1991, 28th junior high school from 1991 to 1994, and 2nd high school from 1994 to 1997 in Nanchang. From there, he went to East China Normal University in Shanghai, China, and received a B.S. in Biochemistry in 2001. He first enrolled in the Department of Biological Sciences of Vanderbilt University in 2001 and then transferred to the University of Tennessee–Oak Ridge National Laboratory Graduate School of Genome Science and Technology in 2002. He graduated with a Ph.D. in 2006 and accepted a staff scientist position at Oak Ridge National Laboratory.