5-2005

Mass Spectrometry-Based Proteomics for Studying Microbial Physiology from Isolates to Communities

Nathan Christopher VerBerkmoes
University of Tennessee - Knoxville

To the Graduate Council:

I am submitting herewith a dissertation written by Nathan Christopher VerBerkmoes entitled "Mass Spectrometry-Based Proteomics for Studying Microbial Physiology from Isolates to Communities." 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, Frank W. Larimer, Major Professor

We have read this dissertation and recommend its acceptance:

Jeffrey Becker, Loren Hauser, Steven W. Wilhelm

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 Nathan Christopher VerBerkmoes entitled "Mass Spectrometry-Based Proteomics for Studying Microbial Physiology from Isolates to Communities." I have examined the final electronic copy of this dissertation for form and content and recommended that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Life Sciences.

Robert L. Hettich___
Major Professor

Frank W. Larimer___
Major Professor

We have read this dissertation
and recommended its acceptance:

Jeffrey Becker_____

Loren Hauser_____

Steven W. Wilhelm_____

Accepted for the Council:

____Anne Mayhew_____
Vice Chancellor and
Dean of Graduate Studies

Original signatures are on file with official student records.

MASS SPECTROMETRY-BASED PROTEOMICS FOR STUDYING MICROBIAL PHYSIOLOGY FROM ISOLATES TO COMMUNITIES

A Dissertation Presented for the Doctor of Philosophy Degree
The University of Tennessee, Knoxville

Nathan Christopher VerBerkmoes
May 2005

DEDICATION

I dedicate this dissertation to all who have helped me get to where I am, but especially to my wife, Kris-for believing in me and for seeing to things so I had the freedom to pursue and achieve my goals; to my children, Briana, Alex, and Gabriel-for being my inspiration; and to my grandparents, Henry and Vickie VerBerkmoes-for raising me and instilling in me the importance of education.

ACKNOWLEDGMENTS

I would like to thank the many people who assisted me in the completion of the research presented in this dissertation. I would first like to thank my graduate advisors Dr. Robert Hettich and Dr. Frank Larimer for their guidance in my doctoral studies. I would like to thank Dr. Jeffrey Becker and Dr. Melinda Hauser for first introducing me to the wonderful world of biological research. I would like to thank Dr. James Stephenson Jr. for introducing me to biological mass spectrometry and all it has to offer and for teaching me to look at my data and others' data with a critical eye. I would like to thank Dr. Loren Hauser, Dr. Jeffery Becker and Dr. Steve Wilhelm for serving on my doctoral committee and always insisting that I search for biological relevance in my research.

I would like to thank all staff of the Organic and Biological Mass Spectrometry Group, including Dr. Doug Goeringer, Dr. Gary Van Berkel, Dr. Gregory B. Hurst, and especially Dr. Hayes McDonald for teaching me the MudPIT technique which was instrumental in a number of studies in this dissertation. I sincerely thank Becky R. Maggard of the Organic and Biological Mass Spectrometry Group for secretarial assistance in the preparation of many of the manuscripts that make up this dissertation and for assistance with the dissertation as a whole. I would like to thank all of the students and post-docs of the Organic and Biological Mass Spectrometry Group especially Melissa Thompson, Christine Shook and Dr. Brad Strader. A special thank you goes to my wife, Kris, for patiently proofreading and editing this entire dissertation.

I would especially like to thank Manesh Shah for his continued collaborative efforts with me on all aspects of proteome informatics. His continued help and discussions have made the difficult challenge of dealing with very large datasets and the large number of collaborators involved in this dissertation seem much less daunting. Much of the work presented here could not have reached the final format, which could easily be transferred to collaborators and the scientific field as a whole, without the collaborative effort between Manesh and myself. I would like to thank Dr. David Tabb (ORNL) and the Yates Proteomics Laboratory at Scripps Research Institute for

DTASelect/Contrast software, and the Institute for Systems Biology for proteome bioinformatics tools used in analysis of the MS data.

I would like to thank Dr. Dorothea Thompson for collaborative efforts on the *Shewanella oneidensis* proteome project. I thank Dr. Jillian Banfield, Dr. Rachna Ram and Dr. Michael Thelen for collaborative efforts on the Acid Mine Drainage Community proteome project and I would like to thank Patricia Lankford, Dr. Dale Pelletier, Dr. Tom Beatty, Dr. Caroline Harwood, and Dr. Robert Tabita for collaborative efforts on the *Rhodopseudomonas palustris* proteome project. I thank Dr. Brian Davison and Dr. Abhijeet Borole for collaborative efforts on the mixed cultures proteome project. I would also like to thank Dr. Goran Mitulovic, Mark van Gils, and LC Packings for providing the 2D-LC-MS/MS system used for many of the studies in this dissertation.

Lastly, I would like to acknowledge support from the University of Tennessee (Knoxville)-ORNL Graduate School of Genome Science and Technology. Much of the research presented here was funded by the U.S. Department of Energy (Office of Biological and Environmental Research, Office of Science) grants from the Genomes To Life and Microbial Genome Programs. Without continued financial support from the three institutes, none of this research would have been possible.

ABSTRACT

With the advent of whole genome sequencing, a new era of biology was ushered in allowing for "systems-biology" approaches to characterizing microbial systems. The field of systems biology aims to catalogue and understand all of the biological components, their functions, and all of their interactions in a living system as well as communities of living systems. Systems biology can be considered an attempt to measure all of the components of a living system and then produce a data-driven model of the system. This model can then be used to generate hypotheses about how the system will respond to perturbations, which can be tested experimentally. The first step in the process is the determination of a microbial genome. This process has, to a large extent, been fully developed, with hundreds of microbial genome sequences completed and hundreds more being characterized at a breathtaking pace. The developments of technologies to use this information and to further probe the functional components of microbes at a global level are currently being developed. The field of gene expression analysis at the transcript level is one example; it is now possible to simultaneously measure and compare the expression of thousands of mRNA products in a single experiment. The natural extension of these experiments is to simultaneously measure and compare the expression of all the proteins present in a microbial system. This is the field of proteomics.

With the development of electrospray ionization, rapid tandem mass spectrometry and database-searching algorithms, mass spectrometry (MS) has become the leader in the attempts to decipher proteomes. This research effort is very young and many challenges still exist. The goal of the work described here was to build a state-of-the-art robust MS-based proteomics platform for the characterization of microbial proteomes from isolates to communities. The research presented here describes the successes and challenges of this objective. Proteome analyses of the metal-reducing bacteria *Shewanella oneidensis* and the metabolically versatile bacteria *Rhodopseudomonas palustris* are given as examples of the power of this technology to elucidate proteins important to different metabolic states at a global level. The analysis of microbial proteomes from isolates is

only the first step of the challenge. In nature, microbial species do not act alone but are always found in mixtures with other species where their intricate interactions are critical for survival. These studies conclude with some of the first efforts to develop methodologies to measure proteomes of simple controlled mixtures of microbial species and then present the first attempt at measuring the proteome of a natural microbial community, a biofilm from an acid mine drainage system. This microbial system illustrates life at the extreme of nature where life not only exists but flourishes in very acidic conditions with high metal concentrations and high temperatures. The technologies developed through these studies were applied to the first deep characterization of a microbial community proteome, the deciphering of the expressed proteome of the acid mine drainage biofilm.

The research presented here has led to development of a state-of-the-art robust proteome pipeline, which can now be applied to the proteome analysis of any microbial isolate for a sequenced species. The first steps have also been made toward developing methodologies to characterize microbial proteomes in their natural environments. These developments are key to integrating proteome technologies with genome and transcriptome technologies for global characterizations of microbial species at the systems level. This will lead to understanding of microbial physiology from a global view where instead of analyzing one gene or protein at a time, hundreds of genes/proteins will be interrogated in microbial species as the adapt and survive in the environment.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

AAC	Amino acid composition
AMD	Acid mine drainage
AMT	Accurate mass tag
BCA	Bicinchoninic acid solution
CAD	Collisional activated dissociation
Capp	ES flow rates over 1 ul/min with LC
CNBr	Cyanogen bromide
COG	Cluster of orthologous groups
DTT	Dithiothreitol
EDTA	Ethylenediaminetetraacetic acid
EM	Electron multiplier
ES	Electrospray ionization
ESI	Electrospray ionization
FA	Formic acid
FAB	Fast atom bombardment
FISH	Fluorescent *in-situ* hybridization
FPLC	Fast protein liquid chromatography
FT-ICR	Fourier transform ion cyclotron resonance
FT-MS	Fourier transform mass spectrometry
Fur	Ferric uptake regulator
H/D	Hydrogen deuterium exchange
HPLC	High performance liquid chromatography
IAA	Iodoacetamide
ICAT	Isotope coded affinity tags
i.d.	Internal diameter
JGI	Joint Genome Institute
KEGG	Kyoto Encyclopedia of Genes and Genomes
LC-MS	Liquid chromatography-mass spectrometry
LC-MS/MS	Liquid chromatography-tandem mass spectrometry
LCQ	Thermo Finnigan ES quadrupole ion trap
LhaA	Light harvesting apparatus assembly protein
LIT	Linear ion trap
LTQ	Thermo Finnigan ES linear ion trap
MALDI	Matrix assisted laser desorption
MASPIC	DBDigger scorer
MMS	Multiple mass range scanning
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
MudPIT	Multidimensional Protein Identification Technology
MW	Molecular weight
Nano	ES flow rates less than 1 ul/min with LC
ORF	Open reading frame
ORNL	Oak Ridge National Laboratory

PCR	Polymerase chain reaction
PMF	Peptide mass fingerprint
PPM	Parts per million
PTM	Post translational modification
QIT	Quadrupole ion trap
RP	Reverse phase
RPLC	Reverse phase liquid chromatography
SAX	Strong anion exchange
SCX	Strong cation exchange
SDS-PAGE	Sodium dodecyl sulphate polyacrylamide gel electrophoresis
TAP	Tandem affinity purification
TCA	Trichloroacetic acid
TFA	Trifluoroacetic acid
TIC	Total ion chromatogram
TIGR	The Institute for Genome Research
TOF	Time of flight
WT	Wild-type
Xcorr	SEQUEST cross-correlation score
2D-PAGE	Two-dimensional polyacrylamide gel electrophoresis

Chapter 1

Introduction to Mass Spectrometry-Based Proteome Analysis of Microbial Systems

Some of the text presented below has been published as Nathan C. VerBerkmoes, Heather Connelly, Chongle Pan, and Robert L. Hettich, Mass Spectrometric Approaches to Characterizing Bacterial Proteomes. *Expert Review in Proteomics* (2004), 1, 433-445. & Nathan C. VerBerkmoes, Joshua Sharp, and Robert L. Hettich, Mass Spectrometry. Book Chapter in *Microbial Functional Genomics by J. Zhou, D.K. Thompson, Ying Xu, and James M. Tiedje,* 2004, John Wiley & Sons, Inc. 241-283.

Of all the life forms on planet Earth, bacteria and archaea are often the most overlooked, even though they are the most abundant by far. These single-celled organisms represent some of the most simplistic forms of life, teaching us how life can survive and even thrive at the very basic level. The adaptability of these organisms to extreme growth conditions, a process which still is very poorly understood, is a testimony to their viability and resilience. Within the last ten years, a revolution in the biological studies of microbial species has occurred, fueled primarily by the availability of complete genome sequences for many microbial species. This genetic information reveals the blueprint for life in that it includes all information about the genes and gene products used by the organism for all of its life functions. This level of global genome information about an organism now makes it possible to begin to pursue a "systems-biology" approach to understanding how these organisms live and function *by cataloging and understanding all of the biological components, their functions and all of their interactions in a living system and communities of living systems* (reviewed Ideker, 2001). Systems biology can be considered as an attempt to measure all of the components of a living system and then trying to produce a data-driven model of the system (Figure 1.1), which can then generate hypothesis about how the system will respond to perturbations, which can be tested experimentally. For microbial systems, that generally means measurements of the genome (the genetic blueprint of the species), the transcriptome (the mRNA transcripts being produced at any given time point), the proteome (the proteins being produced at any given time point, which are the machinery of a cell involving structural, catalytic and signaling processes), and the metabolome (the

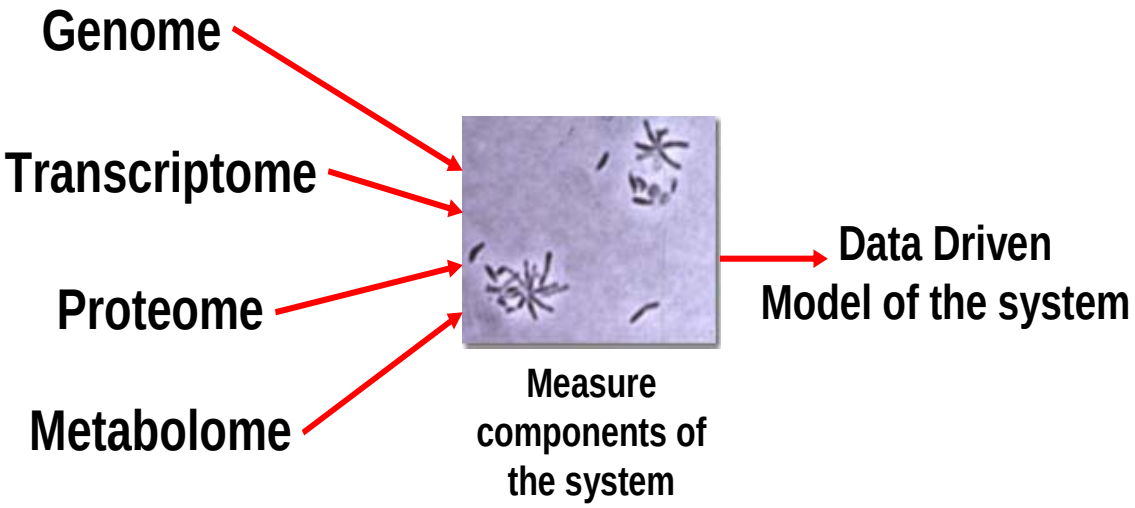


Figure 1.1: Major components of systems biology.
The goal of systems biology is to integrate whole genome, transcriptome, proteome, and metabolome measurements of a microbial system into a data-driven model of that system.

large collection of small molecules, which act as energy sources, building blocks, signaling molecules and many other diverse functions). While this may be simply stated in practice, it is a very daunting challenge. Only because of major technological advances in molecular biology, computing, physics, chemistry and high-end instrumentation have such measurements became possible over the last ten years. Furthermore, the biggest challenge is the integration of the technologies and the people with the expertise in the field. As shown in Figure 1.2, the challenge is the seamless integration of biology, analytical technologies and computational tools in single laboratories as well as coalitions of laboratories and collaborators. This dissertation will detail the development of only one of the measurement types, *mass-spectrometry based proteomics for the measurement of microbial isolates to microbial communities*.

The field of systems biology as it relates to microbial systems can be considered to have gotten its start with the sequencing of whole genomes such as *Haemophilus influenzae* (Fleishmann, 1995), *Mycoplasma genitalium* (Fraser, 1995), *Saccharomyces cerevisiae* (Dujon, 1996), *Methanococcus jannaschii* (Bult, 1996), and *Escherichia coli* K-12 (Blattner, 1997). Since then, microbial species genomes are sequenced at a massive pace at large-scale sequencing facilities such as the Department of Energy's Joint Genome Institute (http://www.jgi.doe.gov/) and The Institute for Genomic Research (http://www.tigr.org/). Over 200 microbial species have been sequenced and annotated to date and strains of microbes are being sequenced for comparative analysis to the original sequenced strain. Furthermore, over 600 microbial sequencing projects are currently under way. Recently, whole genome sequencing of microbial communities have been attempted (Tyson, 2004; Venter, 2004) paving the way for systems biology studies of microbial systems where they matter most, in their natural habitats (see Chapters 6 and 7).

A natural extension of *genomics* (the study of the complete set of genes for an organism) research is the characterization of the gene products, most of which are proteins. This latter research area is defined as *proteomics* (the study of the entire suite of proteins from a genome). Proteome analyses, whether in simple microbes, yeast, or higher organisms, present a much greater challenge than the genomics sequencing efforts.

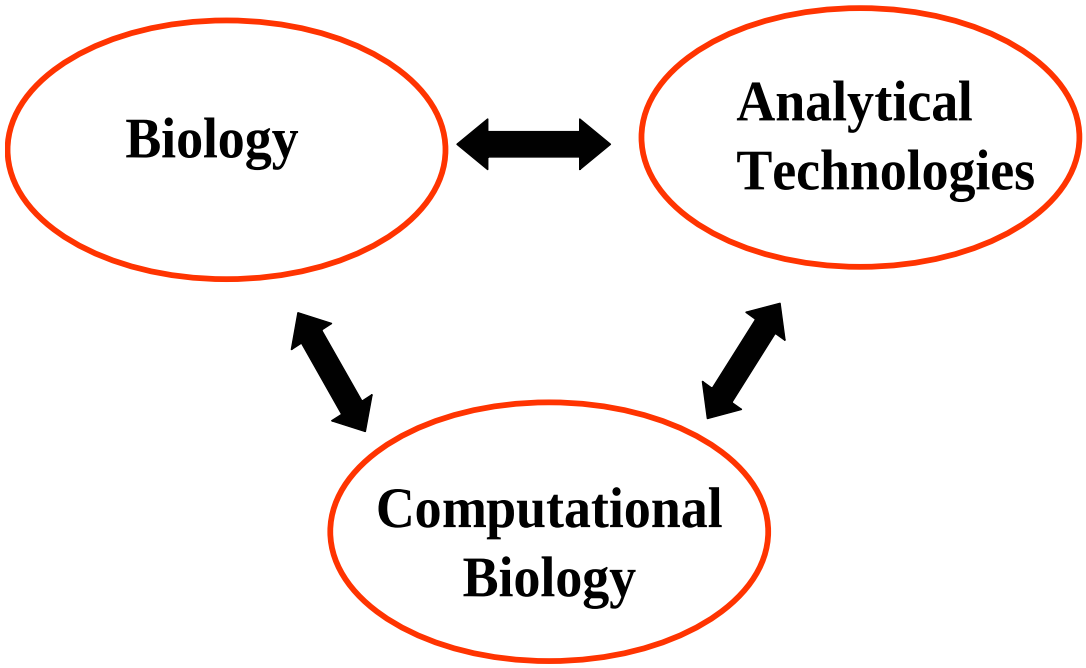


Figure 1.2: Integration of three main branches of science in systems biology. For systems biology studies to be effective it is necessary to integrate three diverse fields of science: Biology, specifically biochemistry and molecular biology; Analytical Technologies, specifically high-end instrumentation such as DNA sequencing, electron microscopy, x-ray crystallography, nuclear magnetic resonance, mass spectrometry, and many others; Computational Biology, all systems biology studies create large datasets that need to be processed, filtered, sorted, and compared.

While the genome is relatively static, the proteome is very dynamic. The genome generally contains a set number of copies of every gene; however, proteins in the proteome can be expressed in a wide concentration range, varying from only a few copies per cell for regulatory proteins to many thousands per cell for ribosomal subunits. The genome for microbes is generally the same under any given metabolic state and environmental situation. Microbes are continually changing and modifying their proteome to adapt to their surroundings, thus a true understanding of a microbial proteome requires many measurements under many metabolic states and environmental conditions. Furthermore, proteins can be highly decorated with any number of post-translational modifications (PTMs); more than three hundred have been recorded (James, 2001). These modifications can be static or dynamic, and may be present in multiple places on a protein. Finally, in the current state of proteomics technologies, there is no amplification technique for proteins similar to the polymerase chain reaction (PCR) that has become so important for oligonucleotide studies. Thus, the proteome of even simple microbes presents a much greater analytical challenge than the corresponding genome (or even transcriptome) analysis. Even in light of these difficulties, a complete understanding of microbes and microbial communities necessitates the development of analytical techniques for rapid and accurate analysis of whole proteomes and protein complexes.

The field of proteomics as it currently exists is diversified and complex with no single measurement platform, experimental approach or desired result. Figure 1.3 illustrates the major objectives in proteomics today. These include:

1) Protein cataloging: the major goal in these studies is to simply measure as many proteins that are being expressed in a given organism at a given point in time.

2) Differential analysis: the major goal of these studies is to compare one or more metabolic states and to determine those proteins that are differentially expressed.

3) Protein localization: the goal of these studies is to determine where proteins are located within a cell, for example, the cytoplasm, the membrane, or the periplasmic space.

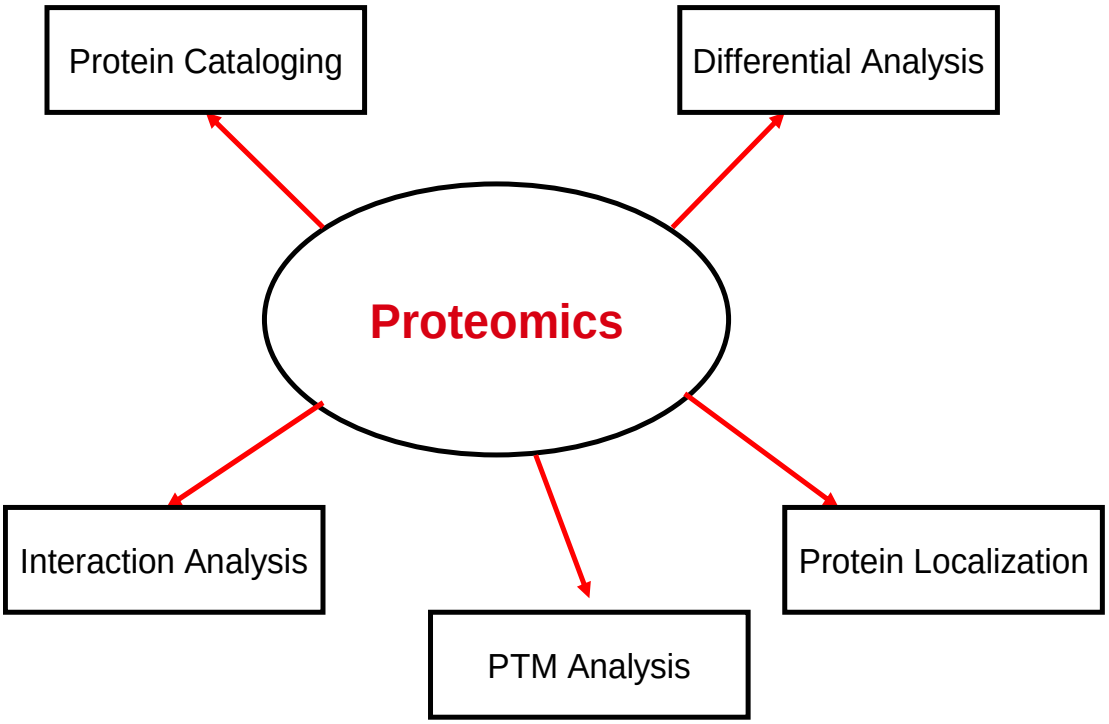


Figure 1.3: Major applications of proteomics.
Currently, proteomics can be broken down into five major objectives. They are all fundamentally different in the way the measurements are made and the type of information that is obtained.

4) PTM analysis: the goal of these studies is to determine what post translational modifications are on proteins and where their specific locations are.

5) Interaction analysis: proteins generally do not function as individual units, rather they function in conjunction with other proteins as macromolecular machines or protein complexes.

The methodologies and the information obtained from these diverse experimental objectives are unique adding to the challenges and the excitement found in proteomics today.

The start of proteomics can probably be most accurately tied to the invention of 2D-PAGE or two-dimensional polyacrylamide gel electrophoresis (O'Farrell, 1975; Klose, 1975). This technique allows for the separation and visualization of 1,000-2,000 proteins on a single gel. Proteins are first separated by their isoelectric point in the first dimension and then by their molecular weight (MW) in the second dimension. While most abundant proteins from a simple bacterium such as *E. coli* can be visualized on a 2D-PAGE gel, the identification of the proteins proved to be much more difficult. At the time, the only technique that had consistent results for protein identification was N-terminal Edman sequencing developed in 1950 (Edman, 1950) and improved upon in 1957 (Edman and Begg, 1957). The development of stable membranes, which the proteins could be transferred to, helped to automate the process (Aebersold, 1986). Still, the evaluations of entire proteomes were massive projects, taking many years with many collaborators. One of the first major efforts in the bacterial world was the Neidhardt *E. coli* gene-protein index, which attempted to define all the proteins visualized on 2D-PAGE gels of *E. coli* grown under numerous conditions (reviewed Neidhardt, 1987). Developments in a seemingly completely unrelated field, termed mass spectrometry, in the 1980's and the 1990's would revolutionize the field of proteomics, creating a completely new approach to the field termed *MS-based proteomics*.

Mass spectrometry (MS) is a family of structural biology tools that have in common the measurement of *ions* of intact and fragmented molecules. A common misconception about MS is that it only provides molecular mass information. In reality, MS is not only powerful for molecular mass measurement, but also provides ion

manipulation capabilities for obtaining detailed structural information at the isomeric level. Mass spectrometry is a well-established technique that historically was first important in particle physics. Mass spectrometry made its commercial impact in the petroleum and pharmaceutical industry where the main emphasis was measurements of small molecules. In the 1980's, the potential for MS for the characterization of biological molecules (especially peptides and proteins) began to be realized.

In the laboratories of Dr. Donald Hunt and Dr. Klaus Biemann a new idea was developing that mass spectrometers might be able to sequence peptides at a much faster rate, at higher sensitivities and from more complex mixtures than was possible by common Edman sequencing techniques (Hunt, 1981; Hunt, 1986, Biemann, 1986; Biemann, 1988). This was accomplished by a process called tandem mass spectrometry or mass spectrometry/mass spectrometry (MS/MS) or collisional activated dissociation (CAD). In this process, a sample of peptides or proteins of similar m/z are isolated in the gas phase inside the mass spectrometer; generally collisional energy is added to the peptide which causes breaks along the peptide backbone in a uniform and predictable manner. Fragment ions are created which correspond to the partial breaking of the peptides from both the N- and C-terminus (Figure 1.4). The fragment ions are then measured by the MS and the sequence information can be reconstructed (though not directly) through database searching or other methods (see discussions below). Because the cleavage can occur at multiple sites, a systematic alphabetic code is used, as shown in Figure 1.4. Fragment ions, which retain the charge on the N-terminus end of the original peptide are designated as "a, b, or c" type ions, depending on the cleavage site. Fragment ions that retain the charge on the C-terminus end of the original peptide are designated as "x, y, or z" type ions, depending on their cleavage site (Roepstorff, 1984; Biemann, 1988). The most common fragment ions observed for peptides and proteins are usually b- and y-type ions. It should be noted that a fragment must retain a charge for it to be detected by the mass spectrometer. This ability to fragment peptides in mass spectrometers to obtain sequence information has become the heart of many proteomic efforts and is central to this entire dissertation. After 1986, MS started to become the dominant method for sequencing peptides and thus identifying proteins.

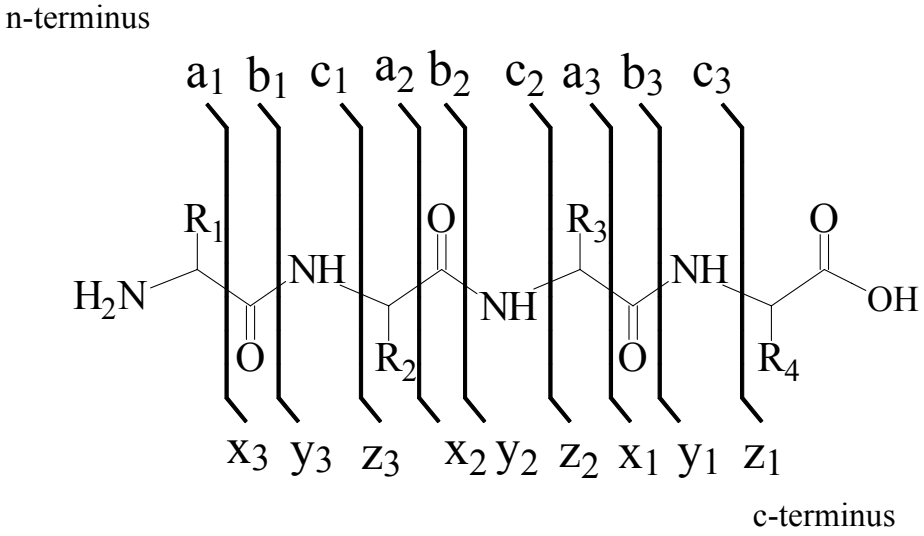


Figure 1.4: Nomenclature for the fragmentation of peptides by MS/MS.
Illustrates the typical fragmentation of a peptide or protein by low energy CAD. The most common fragment ions observed for peptides and proteins are usually b- and y-type ions.

While the advent of MS/MS for obtaining peptide sequence information was a tremendous step, there were still some major fundamental limitations to analyzing peptides and proteins with mass spectrometry. The biggest fundamental challenge was the desorption and ionization of peptides and proteins into the gas phase, which is required for any MS analysis. At that time, the ionization sources, such as fast atom bombardment (FAB), were not very sensitive or robust. The development of two new ionization methods was key; electrospray ionization (ESI or ES) allowed for the direct desorption and ionization of peptides and proteins into the gas phase from a liquid matrix (Fenn, 1989), matrix assisted laser desorption ionization (MALDI) allowed for the direct desorption and ionization of peptides and proteins into the gas phase from a solid matrix (Hillenkamp, 1991; Nakanishi, 1994). Electrospray ionization utilizes a high voltage needle (typically about 2,000-4,000 V) to transfer preformed peptide or protein ions from a flowing liquid solution phase into the gas phase. The resulting mass spectrum usually consists of a range of multiply-charged ions, such as $(M+nH)^{n+}$. The range of multiply-charged ions will depend on the length, basicity and higher order structure of the peptide or protein. In MALDI, the experiment is conducted by using a pulsed laser to desorb peptides and proteins that have been imbedded in a spectrally-absorbing matrix compound (typically a small organic acid). The resulting mass spectrum consists primarily of singly-charged species, such as $(M+H)^{+}$, although some higher charged species, especially doubly-charged ions, are observed in some cases. Typical matrix compounds for proteins and peptides include sinapinic acid and alpha-cyano-4-hydroxycinnamic acid. Both of these techniques have revolutionized the analysis of peptides and proteins by mass spectrometry allowing for sensitive, reproducible and robust methods for their desorption and ionization into the gas phase.

Currently, there are two major methods for analyzing proteins, protein complexes and proteomes by mass spectrometry. The *top-down* method involves measuring intact proteins, either with or without MS/MS of these intact proteins. This method was first introduced with electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry or ES-FT-ICR-MS (Little, 1994; Mortz, 1996; Kelleher, 1998) and expanded to ion traps with novel ion-ion reactions (McLuckey, 1998). In the *bottom-up,*

or *shotgun* method, intact proteins, protein complexes or proteomes are digested with a protease such as trypsin, Glu-C or chemical such as cyanogen bromide (CNBr), and the resulting peptide mixtures are analyzed by MS or MS/MS. It should be noted that in this definition it does not matter whether the initial separations are performed on intact proteins or peptides; rather, the experiment type is defined by the species measured by the MS. Thus, 2D-PAGE of intact proteins followed by in-gel digestion and MS analysis is considered a bottom-up approach. The actual development of the bottom-up methodology cannot be traced to a single lab, but rather evolved from multiple labs using very different techniques including gel-based (Hess, 1993; Mortz, 1994; Shevchenko, 1996; Wilm, 1996; Gatlin, 1998) and solution-based separations (McCormack, 1997; Martin, 2000; Shen, 2001) followed by MS or MS/MS for peptide identifications. These two general approaches can be summarized as follows:

Bottom-up proteomics: Complex protein mixtures (from cell lysate or protein complexes) are proteolytically digested (usually with trypsin), and the resulting peptide mixture is examined by mass spectrometry. The MS data are used to query a peptide database from the specific organism to identify the protein components of the original mixture. *This method is excellent for determining protein identities, but provides very limited information about the molecular form of the intact proteins.*

Top-down proteomics: Complex protein mixtures from cell lysates or protein complexes are examined directly by on-line or off-line MS. No digest is conducted; rather the intact proteins are measured with MS and MS/MS. *This method provides fewer protein identities, but does give detailed information about the intact molecular forms of the proteins, including post-translational processing [small molecule additions (PTMs), truncation, mutations, and signal peptides].*

Both techniques have advantages and disadvantages. Bottom-up proteomics is by far the more widely-used method, mainly because it is much simpler to conduct and does not require high performance MS instrumentation. The progress in the field of bottom-up proteomics has been staggering. It has now become possible (if not routine) to measure ~1000-1500 proteins from a microbe under a given growth condition with a high degree of confidence in a 1-3 day period, depending on the technology used. Furthermore, if

enough mass spectrometers are assembled, this analysis can be rapidly repeated for protein identification for an organism under a variety of different growth conditions.

On the other hand, top-down proteomics has moved along at a slower pace. This is primarily due to the following factors: liquid-based separations of intact proteins are more difficult than peptides, MS and MS/MS analyses of intact proteins are more difficult to conduct and interpret than peptides, the high performance MS instruments capable of adequate analysis of intact proteins from complex mixtures are fairly expensive and have not been designed for routine operation in most cases, and the algorithms to analyze MS/MS of intact proteins are not as well developed or commercially available. Even with these experimental challenges, top-down proteomics provides a level of information that the bottom-up technique does not, which is the *intact state of the protein*. This is critical, as proteins function as intact molecular species, not as a combination of simple, small peptides. Thus, a full understanding of the intact state of proteins (PTMs, truncation, mutations, signal peptides) is necessary, suggesting that an integrated top-down, bottom-up proteomics method would be the most comprehensive (this method is discussed in detail in Chapter 3, the rest of this dissertation focuses on bottom-up or "shotgun" proteomics methodologies).

Protein analysis by MS-based methodologies can be broken down into the following three general areas. While each is unique, they all can be addressed with similar MS technology.

Individual Protein Analysis: This involves the analysis of purified proteins for quality control in structural or biochemical experiments as a means of studying post-translational processing, or in structural analysis of individual proteins by techniques such as H/D exchange (Dharmasiri, 1996), cross-linking (Young, 2000) and surface labeling (Bennett, 2000; Sharp, 2003; Sharp, 2004). While these methods of structural analysis by MS all show great promise for the future, a detailed discussion is beyond the scope of this chapter. Protein detection over a wide dynamic range is not much of an issue here, but sensitivity may be, for example, if the goal is to purify a low-copy number protein with a transient modification.

Protein Complex Analysis: This involves the analysis of purified protein complexes or protein machines, which may also include the analysis of signaling pathways in which the protein complexes are generally more transient and difficult to analyze. Protein complexes are typically purified by centrifugation/sucrose gradient techniques, although more recent approaches employ immunoprecipitation or tandem affinity purification tags (TAP) (Puig, 2001). The greatest challenge is in the biochemical purification and the sensitivity required in the analysis, since typically 100-1000 ng of the complex can be prepared, although preparation of protein complexes from microbes can often yield much greater quantities. Dynamic range in the MS detection is not as much of a problem but may be important if a transient protein is associating with a large complex or if one is searching for PTMs on this complex. A key element of these measurements is the need to purify these complexes from cell lysates. This field is developing rapidly at present, as evidenced by attempts to characterize protein-protein interactions at the proteome level (Gavin, 2002; Ho, 2002).

Whole Proteome Analysis: This typically refers to the analysis of whole cell lysate, organelle preparations, or crude fractions obtained by affinity chromatography or centrifugation, such as membrane preps, cleared serum, etc. The greatest challenge in this analysis is dynamic range because medium to high abundance proteins mask most low abundance proteins. Sensitivity is generally not as much of a problem for whole cell lysates due to large quantities of starting material, but may be an issue for organelle preparations, affinity purified/cleaned fractions or microbial proteomes from the natural environment.

The focus of this dissertation is the analysis of whole microbial proteomes and proteome fractions, though Chapter 3 will contain discussion on all three sub-disciplines. The most complex of these procedures is the examination of a whole proteome. The experimental approach for such a measurement can be broken down into four separate steps. The first step of sample preparation (i.e., cell growth or sample isolation) is followed by the second step of protein fractionation or separation. This leads to the third step of MS characterization, which is followed by the final step of computational analysis of the data (proteome informatics). For the analyses of protein complexes and whole

proteomes, the most important aspect is the dynamic range of the measurement. The demands on the MS can be relaxed somewhat by incorporating separation technologies (such as either off-line or on-line fractionations of proteins/peptides) prior to mass spectrometric detection. These fractionations can be very crude or highly specific, depending on the nature of the application. Methodologies for rapid, deep, sensitive and reproducible characterization of whole microbial proteomes are the central focus of this dissertation.

In the 1990's, the coupling of 2D-PAGE with mass spectrometry (2D-PAGE-MS) was the dominant methodology for analyzing microbial proteomes (Shevchenko, 1996; Wasinger, 2000; Langen, 2000; Fulda, 2000; Grunenfelder, 2001; Hernechova, 2001; Bumann, 2001; Molloy, 2001; Wagner, 2002). The detailed process of 2D-PAGE-MS is beyond the scope of this chapter but has been reviewed in great depth (Jungblut, 1997; Jensen, 1998). The basic process involves separating the proteins by their isoelectric point followed by their molecular weight in a SDS-PAGE gel. Proteins are then excised from the gel, in-gel digested with trypsin, de-salted, and analyzed by a variety of MS techniques. The review by Jungblut et al. (1997), gives a very detailed schematic of all the possible modes of protein identification, and how they are related. The most common is a technique called peptide mass fingerprinting (PMF), where the peptides from the excised spots are measured for their intact masses by MALDI-TOF. The masses are searched against a database of proteins to find the best candidate protein that matches the largest amount of measured peptide masses. For most cases where this methodology does not work, the peptides will be loaded into a nanospray needle and analyzed by static nanospray-MS/MS where sequence information can be obtained from the individual peptides for a more robust search. 2D-PAGE followed by MS analysis has been established as the gold standard for proteome analysis, especially for microbial species. But within the last few years, there has been a noticeable migration away from this methodology toward pure liquid-based approaches. This movement is mainly due to the inherent weaknesses in the 2D-PAGE-MS methodology. The advantages and disadvantages of 2D-PAGE-MS as they relate to the liquid-based methods are highlighted below. Some of these inherent weaknesses are currently being addressed

with new techniques; however, the fundamental fact remains that at the current time, the depth of analysis of protein complexes or whole proteomes by 2D-PAGE does not compare well with the emerging liquid-based methodologies. 2D-PAGE is still routinely used in many labs around the world and undoubtedly will continue.

Advantages of 2D-PAGE for proteome analysis:

i) Gold standard, widely used and understood
ii) Very high resolving power
iii) Sensitive staining methodologies available
iv) Commercial software is available for automated quantitation

Disadvantages of 2D-PAGE for proteome analysis:

i) Poor reproducibility
ii) Limited recovery of low abundance proteins
iii) Limited pI and MW ranges
iv) Time consuming
v) Membrane proteins do not enter the second dimension effectively
vi) Coupling with MS is an indirect process
vii) Intact protein analysis is very difficult

The coupling of liquid chromatography with mass spectrometry (LC-MS) is one of the most promising approaches to overcoming some of the limitations of 2D-PAGE-MS discussed above. The advent of ES has provided a natural way to interface liquid chromatography directly to MS, since ES involves dynamic introduction of a flowing liquid stream directly into a mass spectrometer. It is reasonable to propose to connect a liquid chromatography system to the electrospray source, so that the benefits of liquid-based separation can be combined with high-resolution molecular mass (and MS/MS) measurements. While some work has been done on chromatography of intact proteins in conjunction with mass spectrometry, the majority of effort has focused on chromatography of enzymatically-generated peptides in conjunction with mass spectrometry (bottom-up or "shotgun" method for proteomics). This is primarily due to the fact that peptides are much easier to handle, separate, and analyze than intact proteins. It is somewhat counterintuitive that it is desirable to take a complex protein mixture and make it more complex by digesting the proteins into representative peptides. For

example, each averaged size protein can generate ~20 peptide fragments. So, proteolytic digestion of a sample containing 1,000 proteins will generate a new sample that contains ~20,000 peptides. While this appears to be a poor choice, in practice liquid chromatography and mass spectrometry of peptides are currently well-developed and robust, even for very complex peptide mixtures. The advantages of coupling liquid chromatography with mass spectrometry are obvious when one considers some of the necessities of proteomics, as highlighted below:

1) *Dynamic Range*. The need for multiple dimensions of separation has become most apparent in the use of LC-MS for proteome analysis, due to the large dynamic range necessary for measuring a whole proteome. While 2D-PAGE offers a very high resolving power for proteins, this methodology is currently limited with respect to the types of proteins it can analyze and the number of quality identifications that can be made from any one gel. The coupling of multiple dimensions of chromatography with mass spectrometry offers a solution to this problem. This can be easily noted from the fact that while ~1,000 proteins have been visualized on a 2D-PAGE gel, there have been no published reports of more than a few hundred proteins being identified from a single gel. It has now become routine for the accurate identification of 500-1,500 proteins from a single sample in 20-30 hours on a LC-MS/MS system operated in automated mode.

2) *Sensitivity.* While 2D-PAGE gels have very sensitive staining methodologies for *observing* spots, the ability to *identify* proteins from these spots by mass spectrometry has not matched this level. This is primarily due to the large sample losses in the in-gel digestion step. Pure LC-MS methodologies promise to be more sensitive, due to the reduced overall sample handling while keeping the sample in the liquid phase.

3) *Quantitation.* Differential analysis between two or more sample types is a primary need for successful proteome applications. One of the reasons that 2D-PAGE-MS has remained the gold standard is that it currently is inherently better at quantitation than liquid-based methodologies, due to the fact that LC-MS

suffers from matrix effects that make comparison of run-to-run peak intensities very difficult. This is primarily due to the electrospray ionization methodology and not the mass spectrometry. Furthermore, the apparent accuracy of 2D-PAGE for quantitation became questionable when it was realized that many spots on any 2D-PAGE gels contained more than one protein. The use of stable isotopes for peptide labeling (see quantitation discussion below) and methods in semi-quantitation (see chapters 4 and 5) have proven that quantitation can be accomplished by LC-MS.

4) *Protein Diversity.* A major advantage of liquid-based methodologies in comparison with 2D-PAGE-MS is the diversity of proteins that can be analyzed. Virtually any protein that can be subjected to either chemical or enzymatic digestion can be analyzed. This includes membrane proteins, proteins of high and low pI values, and proteins of high and low molecular mass.

5) *Throughput*. In the field of proteomics, one of the biggest concerns is sample throughput, including not only how fast samples are analyzed, but how well they can be characterized in a short period of time. This is just as important in bacterial proteomics as mammalian or plant proteomics. Currently, at least 200 microbes have been fully sequenced and annotated and hundreds more are under way. Researchers need to be able to analyze the proteomes from these organisms under many different growth conditions and with many different mutants for a systems biology approach to be truly effective. For sample throughput, LC-MS has already been well-developed in the pharmaceutical industry, where thousands of samples are processed by large numbers of mass spectrometers in hundreds of laboratories every year.

The trend towards liquid chromatography methods for proteome analysis can be made clear by examination of Table 1.1. Over the last few years, all of the large-scale proteome analyses have been accomplished with some form of LC-MS methodology. The analysis of *Oryza sativa* provides the best example for a direct comparison of 2D-PAGE analysis and LC/LC-MS/MS analysis for whole proteomes (Koller, 2002). In this

Table 1.1: Landmark papers in large-scale proteome analysis by MS.

Author	Year	Species	Separation Methods	MS Methods	# of IDs
Shevchenko et al.	1996	*S. cerevisiae*	2D-PAGE	Nano-ES-MS/MS	150
Wasinger et al.	2000	*M. genitalium*	2D-PAGE	MALDI PMF	158
Langen et al.	2000	*H. influenzae*	2D-PAGE	MALDI PMF	502
Washburn et al.	2001	*S. cerevisiae*	MudPIT	Nano-ES-MS/MS	1484
VerBerkmoes et al.	2002	*S. oneidensis*	2D SCX-RP/1D MMS	Capp-ES-MS/MS	868
Lipton et al.	2002	*D. radiodurans*	2D SCX-RP MMS	Nano-ES-MS AMT	1910
Koller et al.	2002	*O. Sativa*	2D-PAGE-RP/MudPIT	Nano-ES-MS/MS	2528
Mawuenyega et al.	2002	*C. elegans*	2D SAX-RP	Capp-ES-MS/MS	1616
Peng et al.	2002	*S. cerevisiae*	Offline 2D SCX-RP	Nano-ES-MS/MS	1504

study, the rice plant was broken into three fractions - leaves, roots, and seed tissues. Proteins were isolated from each fraction and analyzed by 2D-PAGE, followed by automated Nano-LC-MS/MS or by multidimensional protein identification technology, or "MudPIT" (Washburn, 2001). The analysis of all three fractions by 2D-PAGE-MS resulted in 556 non-redundant identifications, while the analysis of all three fractions by MudPIT resulted in 2,363 non-redundant identifications. There was no mention from the authors about the length of time each analysis consumed, but we estimate the MudPIT analysis could have been accomplished in 5 days on a single mass spectrometer once the system was optimized. This throughput, as well as the enhanced dynamic range, is why the liquid-based methodologies are having the greatest impact on proteome analysis. For further reading on the advancement of LC-MS in proteomics, see these excellent reviews (Peng, 2001; Mann, 2001; Liu, 2002).

The entire process of an LC-MS/MS experiment for shotgun proteomics is detailed in Figure 1.5. Sample preparation, different versions of LC-MS procedures, and quantitation will then be explained in detail below and in Chapter 2. The shotgun proteomics technique begins with enzymatic digestion of a microbial proteome sample and analysis of the resulting peptide mixture by automated LC-MS/MS or LC/LC-MS/MS in a data-dependent manner (top, Figure 1.5 details this process). Bear in mind that the sample that is injected onto the chromatographic system consists of a mixture of thousands of distinct peptides. These peptides are separated physically over a period of time by their hydrophobicity or net charge, and are sequentially ionized by ES and injected into the mass spectrometer. At any given point in time, 20-200 peptides can be entering the MS depending on sample complexity and length of separation. By using the mass spectrometer to record the overall ion intensity as a function of time, it is possible to obtain a Total Ion Chromatogram (TIC) much like a UV chromatogram (top left, Figure 1.5). During the entire chromatographic run, the mass spectrometer is oscillating between full scan mode, where it is acquiring m/z values of peptides entering the mass spectrometer at that time point (top right, Figure 1.5), and subsequent MS/MS mode, which examines the fragmentation of the most abundant peptides (generally 3-5) as they elute from the column (bottom left, Figure 1.5). This latter mode is accomplished

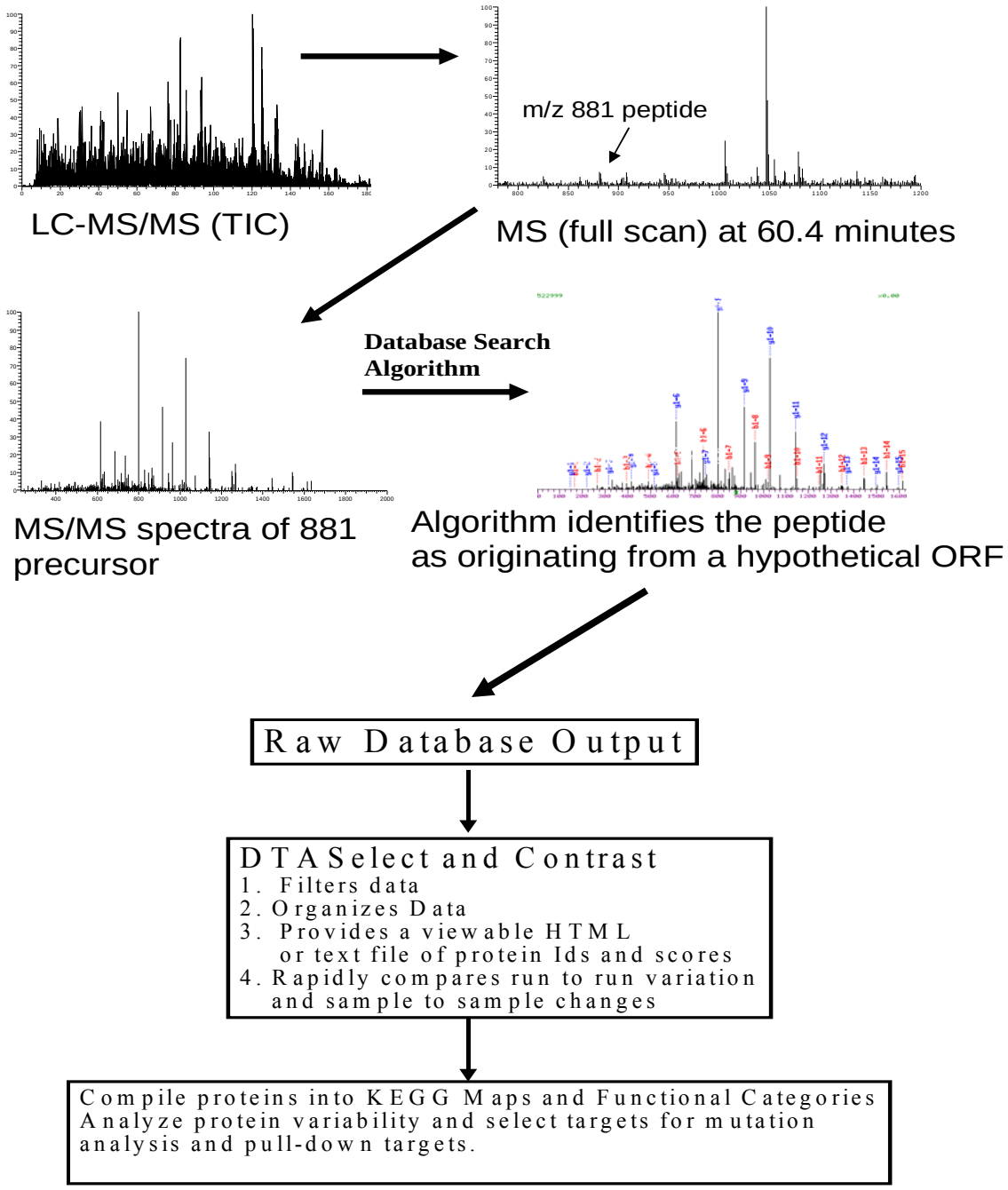


Figure 1.5: Flow diagram for typical LC-MS/MS experiment.
Depicts the typical flow path of a LC-MS/MS experiment from LC-MS/MS analysis, to database searching, to data filtering and comparison and final biological output. TIC-total ion chromatogram.

by gas-phase isolation of individual peptides, followed by collisional activated dissociation and detection of fragment ions. The mass spectrometer records the fragment ions and the mass of the precursor ion. To increase dynamic range, most methodologies employ some type of dynamic exclusion so that peptides that have already been fragmented are not fragmented again. The precursor masses and the fragmentation patterns are then submitted to search algorithms such as SEQUEST (Eng, 1994) and MASCOT (Perkins, 1999), which can query thousands of MS/MS spectra against protein or nucleotide databases (bottom right, Figure 1.5). It should be noted that sequence information cannot easily be directly interpreted from the MS/MS spectrum due to the complexity of the fragmentation processes. Instead, the search algorithms perform cross-correlation (SEQUEST) or probability (MASCOT) comparisons between the observed spectrum and computationally derived spectra from protein and nucleotide databases. The parent mass of the peptide provides a look-up function to find candidate peptide sequences within the potential mass window of the observed parent peptide. The observed MS/MS spectrum is then directly compared to hundreds of potential candidate MS/MS spectra and a best scoring candidate match is made. This by no means guarantees the peptide is the correct identification; it is just the best match to that given spectrum from that given database.

The final stage of the process is illustrated in the bottom of Figure 1.5. Typically, a single LC-MS/MS experiment produces tens of thousands of MS/MS spectra. The identification from these spectra must be filtered and sorted in order to extract useful information from them. Filtering and sorting software, such as DTASelect (Tabb, 2002), are used to extract and sort positive identifications, whereas the program Contrast (Tabb, 2002) is used to compare run-to-run variations and sample-to-sample changes. Correct filtering of MS/MS identifications from SEQUEST and MASCOT outputs is critical and discussed in detail in Chapter 2. The protein identifications can then be compiled into KEGG maps and functional categories for rapid viewing of metabolic and signaling pathways that are activated. This information allows targets to be designed for mutations, gene knockouts, and protein-protein interaction studies (see Chapters 4 and 5).

The analysis of whole proteomes and protein complexes by mass spectrometry can provide very useful *qualitative* information, but one of the most interesting areas of proteomics is the *quantitative* comparison between different growth conditions, or mutants, for a given organism. For example, quantitative analysis of microbial proteomes has been dominated by 2D-PAGE-MS, but this is shifting towards pure liquid-based methods due to the deficiencies in the 2D-PAGE methodology described above. This is primarily due to the fact that the pure liquid-based methods are inherently higher throughput, and are not biased against any protein type. There are difficulties with the liquid-based methods for quantitation, and technologies to address these are only now being developed and implemented. Due to matrix effects associated with both ES and MALDI, direct comparisons of ion peak heights or area for given peptides or proteins eluting from LC columns into the MS should only be used as approximations for abundance levels, and are most likely not accurate for absolute quantitation. Recent developments in stable isotope labeling has allowed for accurate, relative quantitation of proteins in two different samples, such as *E. coli* grown under high salt and low salt conditions. In these experiments, a given protein(s) can be compared with its counterpart from a different growth condition to obtain a relative expression level of up- or down-regulation, but an absolute level of protein expression is still very difficult to determine. Three main methodologies that employ stable isotopes for relative quantitation have developed over the last few years. Each has advantages and disadvantages, which are highlighted below:

Isotope Coded Affinity Tags (ICAT): ICAT was originally developed in 1999 (Gygi, 1999), and has become commercially available through Applied Biosystems. The methodology has since been applied to the analysis of protein expression and comparison with microarray data in *Saccharomyces cerevisiae* (Ideker, 2001; Griffin, 2002), as well as the analysis of human cell line HL-60 microsomal proteins (Han, 2001). This technique entails the use of an isotope encoded affinity tag. The proteins in a sample are mixed with the ICAT reagent, which specifically reacts with cysteine residues. The reagent has an isotopic label (either a "light" version containing either hydrogen atoms on the aliphatic chain or a "heavy" version containing eight deuterium atoms in the same

location) and a biotin affinity tag for isolation of cysteine-containing peptides. This technique is applied to protein samples by labeling one sample with the light reagent and the other sample with the heavy reagent. The samples are then combined, digested with trypsin, and passed over an avidin column to enrich the cysteine-containing peptides. LC-MS/MS methodologies described above can then be used to analyze the complex peptide samples to obtain peptide identifications as well as quantitative information by comparing the peak heights of heavy and light versions of the same peptides. The major disadvantages of this technique are the commercial price of the reagent and the fact that the current version of the commercial reagent only labels cysteine residues. This latter point presents a serious problem for some bacterial species. The average number of cysteines per protein is much lower in bacterial species, as compared with some common eukaryotic species. Furthermore, a large percentage (50-60%) of proteins in bacterial species contain either 0, 1, or 2 cysteine residues, as shown in Table 1.2, which either prevents quantitation, or requires that the quantitation be based on one or two data points.

^{18}O Water Labeling: This methodology was recently introduced as an alternative to ICAT for accurate protein quantification (Yao, 2001). In this technique, one protein sample is digested with trypsin in the presence of ultrapure ^{18}O water, while the other sample is digested in normal water. The samples are then pooled and analyzed by LC-MS/MS or MALDI methodologies on high-resolution mass spectrometers. The results clearly demonstrated that the carboxy termini of the tryptic fragments digested in ^{18}O water are fully labeled with ^{18}O, and this label is stable. Thus, all tryptic peptides from the $H_2{}^{18}O$ sample have an increase in mass of 4 Daltons (two incorporated oxygen atoms on the C-terminus of each peptide). The peptides can then be quantitated by comparing peak areas of co-eluting peptides separated by 4 Daltons. The main disadvantage of this technique is the price and availability of $H_2{}^{18}O$, the fact that the labeling process is so far down stream in the methodology (allows for errors to be introduced) and the need for a high-resolution mass spectrometer to analyze the peptides with such small mass differences.

Nitrogen Labeling: In this methodology, the microbe of interest is grown under a defined media with either normal media (containing naturally-occurring isotopic

Table 1.2: Average number of cysteines per protein for typical prokaryotic and eukaryotic species.

S. oneidnesis (bacterium)

Average # Cysteine residues per protein= 3.159

Cysteines per protein	Proteins	%
0	1000	19%
1	950	18%
2	830	16%
3	646	12%
4 or more	1751	31%
Total Proteins	**5177**	

S. cerevisiae (yeast)

Average # Cysteine residues per protein=6.268

Cysteines per protein	Proteins	%
0	627	9%
1	663	9%
2	668	10%
3	645	9%
4 or more	4332	62%
Total Proteins	**6935**	

R. palustris (bacterium)

Average # Cysteine residues per protein=2.795

Cysteines per protein	Proteins	%
0	946	20%
1	904	19%
2	854	18%
3	614	13%
4 or more	1489	31%
Total Proteins	**4807**	

A. Thaliana (plant)

Average # Cysteine residues per protein=7.892

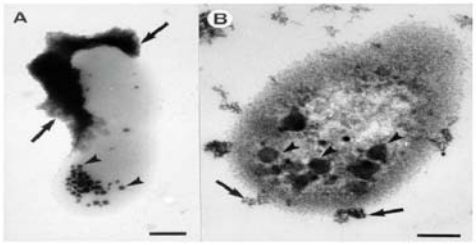
Cysteines per protein	Proteins	%
0	1328	5%
1	1794	7%
2	2080	8%
3	2188	8%
4 or more	18426	71%
Total Proteins	**25816**	

abundances) or isotopically enriched or depleted media (Oda, 1999; Pasa-Tolic, 1999). The most common method is to grow the microbe in defined media without amino acids with only ammonium sulfate as a nitrogen source. The ammonium sulfate can then be either ammonium-^{15}N sulfate or normal ammonium sulfate. The microbe will incorporate the stable heavy isotope into its proteins. The normal and heavy-labeled samples can then be grown under the desired conditions, combined, lysed and digested with a protease. The peptides will have heavy and light pairs that should elute at the same time in a LC-MS/MS experiment, and again the peptides can be quantified by comparing peak areas. This methodology has recently been employed for the largest quantitative proteome analysis to date of the yeast proteome by nitrogen labeling, followed by MudPIT analysis (Washburn, 2002). The major advantages of this technology are the low cost and the ability to quantitate any type of protein that can be digested with either chemical or enzymatic methods. Furthermore, the samples are mixed immediately after growth so that any changes in sample preparation affect both samples in the same manner. The major disadvantage is that this technique can only be used for species whose growth conditions can be exquisitely controlled. This is a severe limitation for many microbial systems and completely impossible for experiments involving microbial systems analyzed directly from their natural environments (see Chapter 7).

All of the above listed techniques for accurate relative quantification have shortcomings and none are very useful for comparing large numbers of metabolic states. There has been recent effort in the field to analyze proteome samples by semi-quantitative techniques both in microbial systems (Gao, 2003) and in human proteome projects (Chelius, 2002; Wiener, 2004). It has been shown that semi-quantitative comparisons of proteome datasets based on the % sequence coverage, # of identified peptides, and the repeat count for a protein (how many MS/MS sequencing events are acquired per protein) are all indicators of protein abundance (Liu, 2004). One of the major goals of this dissertation was to develop and test methods for semi-quantitation of microbial species with gene knockouts (Chapter 4) and microbial species grown under various metabolic states (Chapter 5).

Our main focus for these studies was microbes with potential for carbon sequestration, bioremediation, long-term energy production and potential to survive at the extremes of the natural environment (extremophiles). The bacteria we choose to study in this dissertation clearly fall into those categories. The individual bacteria and the natural community that were the focus of these studies are shown in Figure 1.6. *S. oneidensis* is a facultatively anaerobic γ-proteobacterium which possesses remarkably diverse respiratory capacities that have important implications with regard to the potential for bioremediation of metal contaminants in the environment. In addition to utilizing oxygen as a terminal electron acceptor during aerobic respiration, *S. oneidensis* can anaerobically respire various organic and inorganic substrates [i.e., fumarate, nitrate, chromium, thiosulfate, trimethylamine N-oxide (TMAO), Fe(III), and Mn(III)]. *Rhodopseudomonas palustris* is a purple nonsulfur anoxygenic phototrophic bacterium in the α-proteobacteria family. *R. palustris* is of great interest due to its high metabolic diversity and ability to degrade simple aromatic hydrocarbons (lignin monomers). It has exceptional metabolic versatility in its modes of energy generation and carbon metabolism. *R. palustris* is capable of producing hydrogen gas as a by-product of nitrogen fixation making it a potential biofuel producer. *R. palustris* also has the potential to act as a greenhouse gas sink by converting CO_2 into cellular material. Since most of these metabolic states can easily be produced in laboratory settings, it makes *R. palustris* a model system for the study of diverse metabolic modes and their control. The acid mine drainage (AMD) communities provide an excellent model system for studying life at the extremes of the natural environment. These communities of bacteria not only survive but thrive in acidic streams (pH <1.0), with molar metal concentrations and high temperatures. The goal was to determine the cellular localization of expressed proteins, provided clues to protein function, and yielded information about the physiological challenges faced by a self-sustaining, chemolithoautotrophic microbial community. These three microbial systems were the core test subjects for the application of MS-based proteomics to attempt and gain better understanding of microbial systems with potential for carbon sequestration, bioremediation, long-term energy production and potential to survive at the extremes of the natural environment.

Shewanella oneidensis ***Rhodopseudomonas palustris***



Glausser et al. *Science* **2002** *295*, 117-119.

Acid Mine Drainage Community
Life at pH less than 1.0

Figure 1.6: The microbial systems which were the focus of this research.
The three microbial systems which were the focus of this dissertation. *Shewanella oneidensis* a bacteria with great potential for bioremediation. *Rhodopseudomonas palustris* a metabolically versatile bacterium with potential for energy production, carbon sequestration and bioremediation. The acid mine drainage community a model system for studying a microbial community which can survive at the extreme of nature (extremenly low pH, high metal concentrations and high temperatures).

The major goal of this dissertation was to build a robust high-throughput platform for the analysis of complex protein mixtures, and then evaluate and extend this platform to the characterization of microbial isolates and microbial community proteomes discussed above with the goal of gaining greater biological insight into their complex systems. At the start of this dissertation, proteomics was only beginning to be developed in numerous laboratories. Thus a major effort was needed to develop the necessary biological, analytical, and computational tools to addresses this daunting technical challenge. Hopefully the research presented here has helped to bring us one step closer to achieving that goal.

The following is an outline of that effort. Chapter 2 details the current ORNL "shotgun" proteomics pipeline for microbial proteomics, which was developed primarily through efforts of this dissertation. Chapter 3 details a new methodology of integration of the top-down and bottom-up techniques for the analysis of individual proteins, protein complexes and whole proteomes. Chapter 4 illustrates our first major report on qualitative comparisons between a global regulator knockout and the wild-type (WT) strain in the bacterium *Shewanella oneidensis*. Chapter 5 further illustrates the effectiveness of semi-qualitative comparisons with a large-scale analysis of the major metabolic modes of the bacterium *Rhodopseudomonas palustris*. Chapter 6 introduces methodology development and testing for characterizing microbial communities focusing on artificially prepared microbial mixtures and Chapter 7 concludes with the application of "shotgun" proteomics for the first characterization of a natural microbial proteome with the global characterization of the proteome from an acid mine drainage biofilm. This dissertation is the culmination of years of effort to develop a global proteomics platform for the characterization of microbial proteomes from isolates to communities.

Chapter 2

Experimental Platform for Global Analysis of Microbial Proteomes

Introduction

This chapter describes the experimental platform for global analysis of microbial proteomes extracted from either microbial isolates or natural communities that was developed through the course of this dissertation. While a common experimental thread of analyzing bacterial proteome by liquid chromatography in conjunction with tandem mass spectrometry (LC-MS/MS or "shotgun" proteomics) can be found in all following chapters, the exact methods vary to some degree. This chapter breaks each part of the process down and explains variations and advantages and disadvantages of the various methods. The ORNL proteomics platform is illustrated in Figure 2.1. The major parts include cell growth, protein extraction/sample preparation, liquid chromatography, mass spectrometry, proteome informatics and biological information extraction. Each of these subtasks is detailed below.

Cell Growth

For all studies presented in this dissertation (except Chapter 7) bacteria were grown from stock solutions in batch format. Generally, glycerol stock solutions of the WT strain or a mutant strain are kept at -80^{o}C. For the *S. oneidensis* studies presented in Chapter 4, the strains were obtained from Dr. Dorothea Thompson in the Environmental Science Division at ORNL. For the *R. palustris* studies in Chapter 5, the wild-type (WT) strain CGA0010 and LhaA knockout mutant were gifts from Dr. Caroline Harwood at the University of Washington and can be obtained from Dr. Dale Pelletier in the Life Science Division at ORNL. The *Escherichia coli* and *Saccharomyces cerevisiae* strains that were used as background samples in Chapter 6 were supplied by Dr. Brian Davison in the Life Science Division at ORNL. The acid mine drainage (AMD) biofilm samples discussed in Chapter 7 were the only samples that were not grown from isolates. These are naturally occurring microbial biofilms that were collected from the Iron Mountain Mine, Redding, California, and were supplied by Dr. Jillian Banfield, University of California, Berkeley.

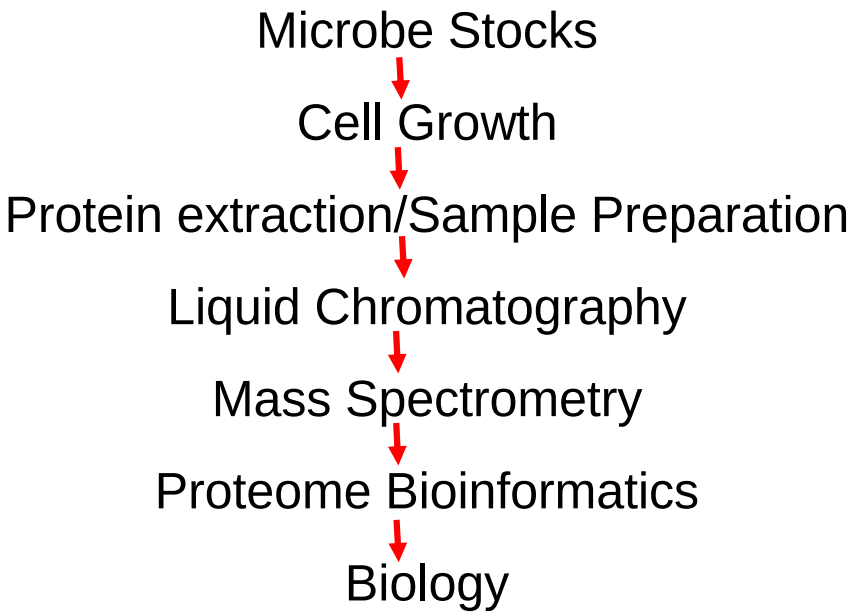


Figure 2.1: Major steps in ORNL proteomics pipeline.
Illustrates each major step in the ORNL proteomics pipeline for the analysis of single microbial isolates, mixtures of isolates, and natural community samples. It should be noted that natural community samples are not grown from stocks but rather collected directly from the environment.

Generally, for growth of isolates, the stock solution was aliquoted into a 1 L solution of growth media, which is generally in a 2 L flask. The choice of growth media was different for each microbial system and dependent upon the metabolic state required. For aerobic growth (growth in the presence of oxygen), the flask was agitated on a shaker to allow for full aeration. Sometimes, oxygen or air can also be sparged through the system. For anaerobic growth, the entire flask was completely closed. For growth requiring photosynthesis, cells were fully illuminated with a light source. Figure 2.2 illustrates *R. palustris* growing anaerobically with a light source.

For most studies in this dissertation (unless otherwise noted), microbial growth was run into mid-log phase. For comparative studies, this is generally the best place to harvest cells, since most cells in the culture will be at an equal state of metabolism. Figure 2.3 illustrates the general growth pattern of a microbial system in culture. Immediately after inoculation, the system is in lag phase, where the microbes are adapting to their new environment and exponential growth has not yet begun. Individual cell mass increases, but the cells are not dividing rapidly. The microbial system then moves into log phase, where exponential growth is occurring; there are plenty of nutrients and the cells are not overly stressed. During this interval of growth, cell numbers are doubling at some regular interval, which is determined by the species type, its doubling time, available nutrients and environmental factors. Once nutrients start to become limiting, the cell culture will enter a stationary phase where cell death and new cell development is relatively equal, and there will be no great changes in overall cell concentrations. Cells are starting to become stressed in this situation as nutrients have become limiting, toxins build up and environmental changes such as pH shifts occur. After stationary phase, some microbial systems will enter a decline, or death phase, where nutrients are very limiting and cell death is much greater than generation of new cells. Overall viable cell numbers start a rapid decline and cells are very stressed and major morphological changes can occur. The length of times for each of these stages varies from microbe to microbe and from metabolic state to metabolic state. Thus, very careful initial experiments must be performed to determine the general rate of cell growth and progression through this cycle. It is very important when comparing two metabolic

Figure 2.2: *R. palustris* under batch photoheterotrophic growth.
Illustrates a standard growth chamber for *R. palustris*, in this case the bacteria is growing under the photoheterotrophic state, where it is fully anaerobic and light is provided as the energy source.
Figure courtesy of Dale Pelletier.

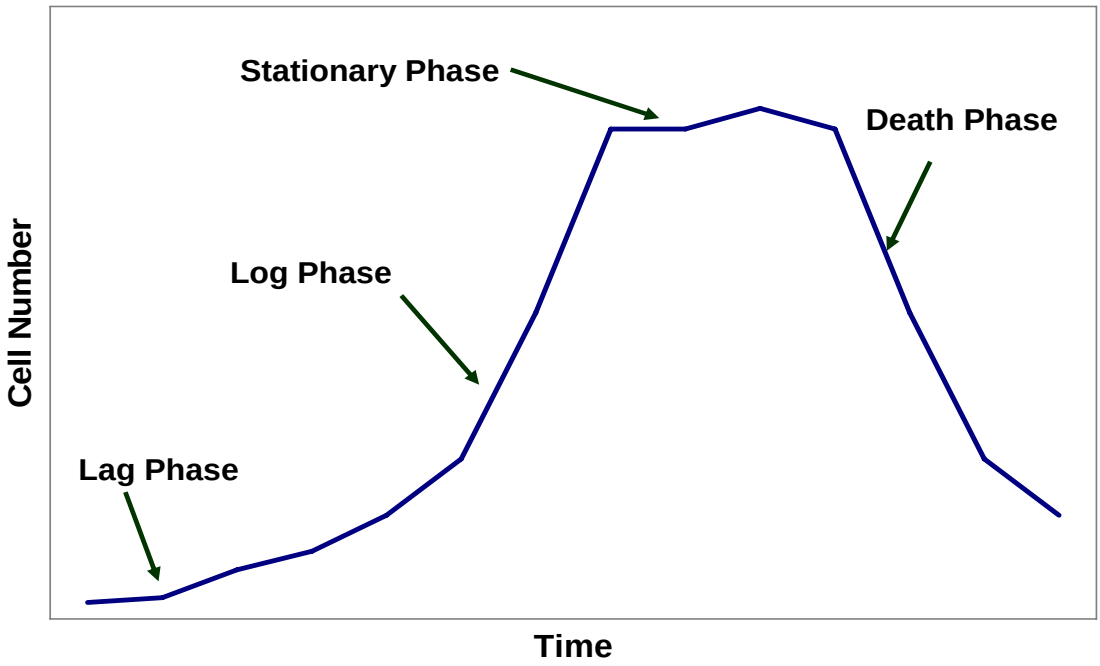


Figure 2.3: Standard bacterial growth curve.
Illustrates a standard bacterial growth curve in batch culture where the microbial system moves from lag to log phase, then into stationary phase and finally a death or decline phase as nutrients become severely limiting.

states, that the cells are harvested at close to the same point in the growth curves as possible, otherwise one is not only comparing the metabolic state but also the point in the growth states which can have very dramatic effects on the expressed proteome. Once the microbial culture has reached its desired state, it is rapidly harvested by low-speed centrifugation (5000 g x 10 min). The cell pellet is then resuspended in 50 mM Tris (pH 7.5) and centrifuged a second time. This is done to remove residual media from the cell pellet. This is always done on ice as rapid as possible to minimize metabolic changes to the proteome. Care should be taken to treat the cells gently to avoid mechanical lyses. After the second centrifugation step, the cell pellet can either immediately be processed as below or stored indefinitely at -80^0C.

Protein Extraction/Sample Preparation

The next stage of the proteomics pipeline involves proteome extraction and sample preparation. The microbial cell pellet was resuspended in 50 mM Tris/10 mM EDTA (pH 7.5). The EDTA was kept in all buffers from this point forward to help inhibit metalloproteases, by chelating away the necessary metal ion co-factors. From this point forward, all steps should be done as rapidly as possible and kept on ice. This was extremely important since after cell lysis, endogenous protease activity will no longer be controlled by the cell and protein degradation will occur. It should be noted that for bottom-up or "shotgun" proteomics experiments, protease inhibitor cocktails were not used. The reason for this is two-fold; first the protease inhibitor cocktails all contain trypsin-like serine proteases inhibitors and since the proteome will be digested with trypsin, adding an inhibitor is not a good idea. Second, most protease inhibitor cocktails contain small peptides/protein inhibitors, such as Aprotinin. These inhibitors generally co-purify with the tryptic peptides through the sample preparation scheme and thus cause major problems in the LC-MS/MS experiment in that they are very concentrated and can mask large portions of the LC-MS analysis. The volume of buffer used depends on the amount of cell pellet to be lysed. Generally, we used 2 g of wet cell pellet and 10 ml of buffer for cell resuspension prior to lyses. While there are many ways to lyse microbial cells, such as bead beating, French press and sonication, all of the experiments in this

dissertation used sonication. The main reason for this is the simplicity and speed of the sonication processes. The resuspended cell material was homogenized and transferred into a large test tube (25-50 ml). The tube was packed into a beaker full of ice and put into the sonicator. Sonication was typically conducted for 5 minutes, with 20 second bursts followed by 20 second cooling periods. It is important not to go too long in either the burst time or the total length of time because thermal denaturation and degradation of proteins can occur. While all cells will not be lysed by this method, the majority is, and for proteomic applications, usually much more protein is extracted from a 1 L culture of cells than is ever needed for the proteome experiments.

After cell lyses, the resultant protein solution was centrifuged at low speed (5000 g x 15 min) to remove unbroken cells and cellular debris. At this point, the proteome can be processed directly but in many cases further fractionation by centrifugation was applied. The choice of centrifugation steps is dependent on the amount of finer fractionation needed and the type of information that is sought. The main reason for fractionation is to increase the dynamic range of the proteome measurement. The proteome of even a simple microbe is very complex and beyond the current analytical capabilities of even the best LC-MS/MS systems. Centrifugation techniques are the simplest and quickest methods to simplify the proteome, and information about cellular localization can sometimes be inferred. Figure 2.4 illustrates the common centrifugation techniques used for the studies in this dissertation. The first centrifugation step was at 100,000 g for 1 hour and creates a first pellet and a soluble proteome. This first pellet loosely termed the membrane fraction, was enriched in membrane-bound proteins but should not be considered a complete membrane preparation. This pellet was then washed once with the lyses buffer and sonicated briefly to help get the proteins back into solution. For many experiments, this was the only level of centrifugation that is applied and it was the most important. The observed proteomes for the crude supernatant and the first pellet were very different as illustrated in Table 2.1 for a representative *S. oneidensis* proteome. Only 23.3% of the observed proteome was redundant between the two fractions, while 32.3% of the total proteins were only found in the membrane proteome

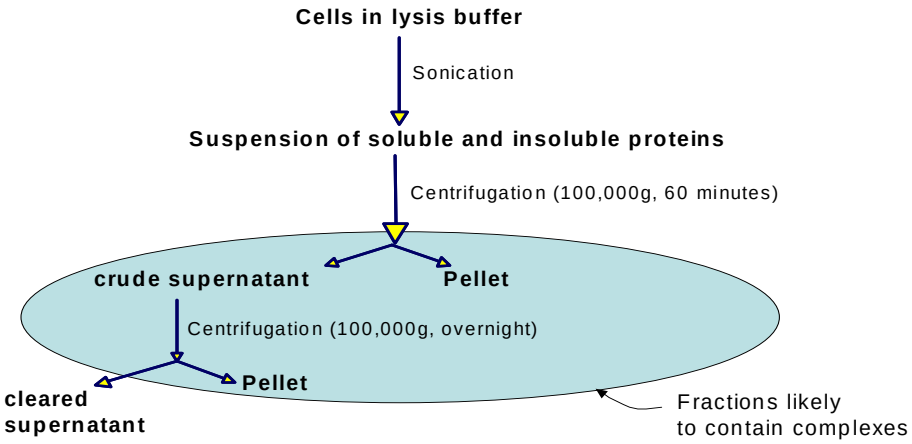


Figure 2.4: General fractionation scheme by ultracentrifugation.
Illustrates the different levels of fractionation possible by centrifugation techniques. Proteome analysis can be conducted after any step of this process; at the very top on a suspension of soluble and insoluble proteins, after the first centrifugation step on the crude supernatant and the 1st pellet or after finer separation on all four fractions including the cleared supernatant and 2nd pellet.

Table 2.1: Proteome fractionation of *Shewanella oneidensis.*

Soluble	Membrane	Percent of Total	Total Proteins	
X	X	23.3	172	
	X	32.3	239	
X		44.4	328	
500	**411**		**739**	**NR Totals**

and 44.4% were only found in the soluble proteome. This level of difference was definitively due to the fractionation and not to the lack of replication between two runs since that is known to run at 70-80% for replicate analyses (see below and Chapter 5).

For the *R. palustris* proteome study in Chapter 5, a finer level of fractionation was attempted. The main reason for this was previous findings of differential fractionation of protein complexes and an increased dynamic range afforded by the extra centrifugation step. For this study, the crude soluble proteome from the first high-speed spin was split in half; half was aliquoted and frozen at -80^{0}C and the other half was centrifuged at 100,000 g for 16 hours to create a second pellet and a cleared fraction. The cleared fraction was generally found to be devoid of proteins known to be found in protein complexes. Many soluble low-abundance proteins, such as periplasmic binding proteins and transcription factors, can be observed that were masked by other high-abundance proteins in the first soluble fraction. It should be noted that this level of fractionation creates two more samples for analysis, while not providing the massive increase in dynamic range found in the first fractionation. Table 2.2 illustrates the differences and overlaps between each fraction from a typical analysis of an *R. palustris* proteome.

For all cases, after the fractionation process, the protein solutions were aliquoted into 1 ml aliquots and immediately frozen at -80^{0}C. One aliquot from each fraction was then quantitated with the BCA protein assay reagent (Pierce Biotechnology, Inc., Rockford, IL) to determine approximate total protein quantity. For most proteome fractions in our studies, the protein concentration falls between 1 mg/ml and 10 mg/ml. This is an important step because it gives the necessary information for the amount of trypsin to be added and roughly how much to concentrate the peptides after digestion to achieve an optimal working concentration (discussed below). In some cases (Chapter 7 AMD study), there was not enough protein available to do the BCA assay (generally this takes ~1-2 mg of total protein). In these cases, the amount of total protein was roughly estimated.

The next step involves protein denaturation and reduction. This step was necessary to completely denature and reduce the proteins, making them completely accessible for protein digestion. There are many variations of how to accomplish this.

Table 2.2: Deeper proteome fractionation of *R. palustris* proteome.

Cleared	Soluble	Pellet 2	Pellet 1	Percent	Total Proteins	
X	X	X	X	6.10 %	58	
	X	X	X	3.60 %	34	
X		X	X	2.00 %	2	
X	X		X	0.70 %	7	
X	X	X		25.90 %	247	
		X	X	1.40 %	13	
	X		X	0.30 %	3	
X			X	0.20 %	2	
	X	X		9.50 %	91	
X		X		1.20 %	11	
X	X			12.70 %	121	
			X	17.50 %	167	
		X		7.50 %	72	
	X			2.70 %	26	
X				10.60 %	101	
549	**587**	**528**	**286**		**955**	**NR Totals**

We have found that denaturation with 6 M Guanidine and reduction with 10 mM DTT at 60^0C for 1 hour was a very effective method. For most cases, 2-5 mg of total protein was diluted in 2 ml of 50 mM Tris/10 mM $CaCl_2$ (pH 7.6) with 6 M Guanidine-HCl/10 mM DTT. We have also found that it is essential to rotate the proteome fractions end over end to avoid protein settling during this process. Many protocols call for the labeling of cysteine residues with a reagent such as iodoacetamide (IAA). This reagent permanently labels cysteine residues so disulfide bonds cannot reform. For all our studies, we have omitted this step. This is due to two major reasons: 1) bacterial species in general do not have a large number of cysteine residues [see Chapter 1]; and 2) if the reagent is not used with the utmost care and at the exact concentration for the exact amount of time, non-specific labeling of other residues is possible. We have found through work on protein standards (mainly bovine serum albumin) that this step can be skipped if the peptide solution after digestion is again fully reduced before sample clean-up (VerBerkmoes, unpublished work)

After the denaturation/reduction step, individual fractions were digested with trypsin. For all studies, Promega Modified Sequencing Grade Trypsin (Promega, Madison, WI) was used. The rational for using trypsin is multi-fold. First, it is an inexpensive, high quality enzyme that is easy to obtain in very large quantities. Second, it is very specific, cleaving on the C-terminal side of lysine and arginine residues (except when proline is the next residue), which are very prevalent in most proteins. Third, since it cleaves C-terminal to lysine and arginine residues, it creates peptides that can carry a positive charge on the N-terminus and C-terminus, giving a large abundance of peptides that can carry a +2 charge. These peptides electrospray very well and fragment well via the MS/MS processes. 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). It is important to at least dilute the guanidine by 6-fold from 6 M to 1 M guanidine. Trypsin is effective up to ~1 M guanidine concentrations. The addition of $CaCl_2$ is thought to be an essential co-factor for trypsin activity. The correct pH of the solution is also an important consideration. Trypsin as an optimal activity between 7-8 pH, and its activity dramatically decreases at lower pH. The pH of the proteome solution

should be checked after dilution and before addition of trypsin by pipetting a small aliquot onto pH paper. After addition of trypsin, the solution was incubated with gentle end-over-end rocking for overnight or at least 4-5 hours at 37^0C. After the first digestion period, a second aliquot of trypsin was added again at 1:100 (wt:wt) and digestion should proceed another 4-5 hours. This is necessary because trypsin loses activity after 4-5 hours and the second addition helps to obtain complete digestion. After the final digestion step, solid DTT was added to the solution to obtain a final concentration of 20 mM DTT. The final reduction step is allowed to proceed for 1 hour with gentle end-over-end rocking at 37^0C. This step is absolutely essential to break disulfide bonds that have randomly reformed during the digestion process. The peptide solution was then spun at 5000 g x 5 min to remove undigested protein material and aggregated DNA/lipids. This is essential to avoid clogging of the extraction cartridges in the next step. At this point the peptide solutions can be stored at -80^0C or one can proceed directly to sample clean up.

The next step is sample clean up. This is necessary to remove the excess salts and guanidine before sample analysis. While this step can be done directly on-line or off-line on the actual chromatographic columns, we have found the processes to be more reproducible and robust to de-salt samples prior to loading onto LC-MS systems or columns. If total protein sample is limited, then this step should be omitted and some method of on-line clean up should be attempted. For all studies presented in this dissertation, the total protein quantity was not limiting, so sample clean up was done by the described procedure. C_{18} Sep-Pak (Waters, Milford, MA) were used to de-salt all samples. Briefly, the Sep-Pak was conditioned with an organic such as acetonitrile and washed with H_2O/0.1% FA before sample loading. The proteome fraction was loaded onto the Sep-Pak and washed with ~10 ml of H_2O/0.1% FA. The sample was then eluted with 4-5 ml of acetonitrile/0.1% FA. The peptide solution was now de-salted but was too dilute and remains in an incompatible high organic solvent for most chromatographic loading purposes. The peptide sample was concentrated using a centrifugal evaporator (Savant Instruments, Holbrook, NY) to ~10 μg/μl starting material and solvent exchanged by the addition of at least 1 ml of H_2O/0.1% FA. It is very important in this

process to never completely dry the sample, since major peptide loss will occur. After the proteome sample reaches ~10 µg/µl starting material, the sample should be filtered through 45 µm filters (Millipore, Bedford, MA) to remove any particulates. We have found that ~10 µg/µl is an optimal concentration for loading onto any of the LC-MS/MS systems described below. Lower concentrations require much larger sample injections which are sometimes impractical and high concentrations cannot be obtained without major peptide loss through aggregation. Samples were then aliquoted and frozen at -80^{0}C until LC-MS/MS analysis.

Liquid Chromatography/Mass Spectrometry

The studies presented in this dissertation employed a variety of LC-MS/MS methodologies, which are discussed in detail below. Each has advantages and disadvantages, which will also be discussed below. One general theme was that all "shotgun" studies employed liquid chromatography in conjunction with an electrospray ion trap mass spectrometer operated in data-dependent mode, as described in chapter 1. Three major forms of LC-MS/MS were employed: 1-dimensional LC-MS/MS with multiple mass range scanning, 2-dimensional switching LC/LC-MS/MS, and 2-dimensional MudPIT LC/LC-MS/MS. Two types of ion trap mass spectrometers were employed-the quadrupole ion trap mass spectrometer (reviewed Stafford, 2002) and the linear ion trap mass spectrometer (Schwartz, 2002). These two ion trap mass spectrometers will be explained in detail and contrasted below.

1-dimensional LC-MS/MS with multiple mass range scanning

Of the three methodologies, this is the simplest and easiest to implement. It requires only three major instruments, a low-flow HPLC pump, an autosampler, and the ES-MS (Figure 2.5). As stated repeatedly, the biggest challenge in MS-based proteomics is dynamic range. The peptide mixtures obtained from the digestions of proteome fractions are very complex with thousands of peptides. Current mass spectrometers are simply not fast enough perform MS/MS on that many peptides in the general time of normal separation space (2-4 hours for a single dimension). This can be addressed by

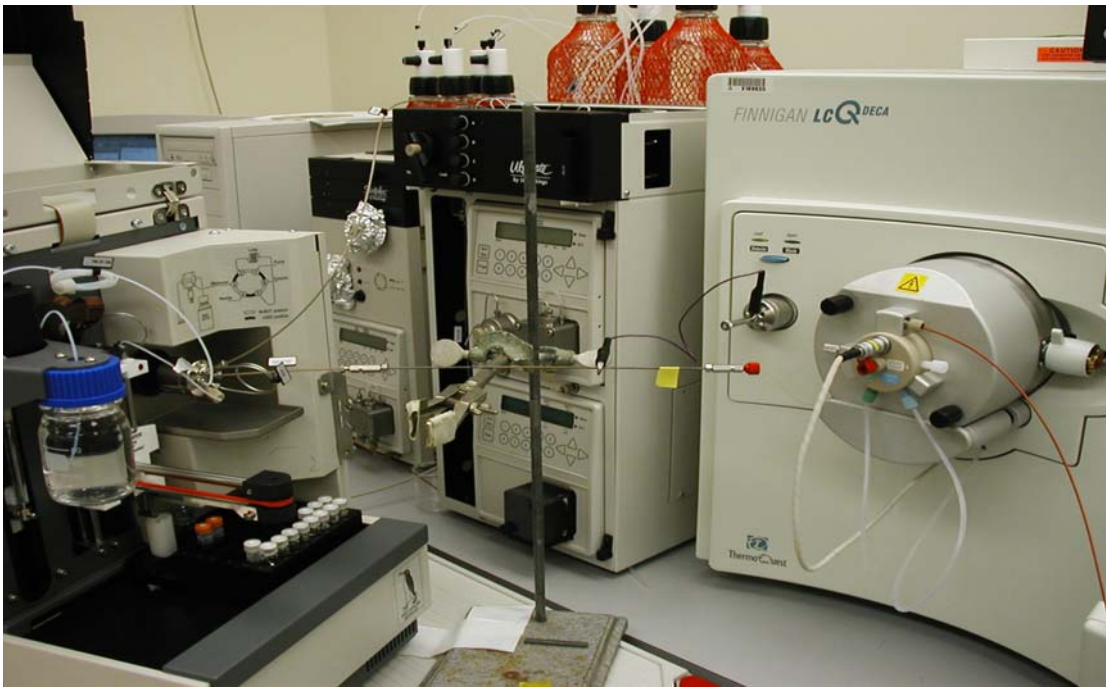


Figure 2.5: System for 1-D LC-MS/MS with multiple mass range scanning. Basic system for 1D LC-MS/MS with multiple mass range scanning. Consists of low-flow HPLC pump (back left with red solvent bottles), connected to autosampler (front-left) which makes automated injections onto the capillary column (middle with yellow tape) which is connected to electrospray source on an ion trap mass spectrometer (right side of photo).

two possible methods: 1) peptides can be resolved by multi-dimensional techniques to relax the complexity of peptides the MS sees at any point or time, 2) the MS can be used as a separation device to minimize the number of peptides it sees at any given point in time. This latter technique is termed multiple mass range scanning or gas phase fractionation and was concurrently developed and optimized at numerous laboratories and institutes, including ORNL, Amgen, Celera, and University of Colorado (Spahr, 2001; Davis, 2001; VerBerkmoes, 2002). For this technique, peptides are generally separated by a one-dimensional HPLC but multiple injections and separations of the same sample are made. For each subsequent injection and separation, the MS is set to scan a narrow m/z region, thus limiting the number of peptides the instrument sees and is required to attempt MS/MS scans on (Figure 2.6). In our experience, we have found that eight overlapping m/z regions are sufficient to obtain quality MS/MS spectra on most detectable peptides in a complex proteome mixture. As shown in Figure 2.6, the use of large overlapping m/z regions was found to be more useful than many very small m/z regions. While not completely understood, we have found the performance of ion trap mass spectrometers to diminish significantly when regions of less than 100-150 m/z units are scanned (VerBerkmoes, unpublished data). We have also found that this technique is most effective with electrospray (2-5 µL/min), where the observed peptides are spread over the entire m/z window of 400-2000 m/z, while in nanospray applications (100-300 nL/min) peptides are mainly observed between 400-1,200 m/z.

The complete experimental procedure involves loading an autosampler with enough material to make the 8 necessary injections. For our studies, each injection required 60 µL of sample, so a total of 500 µL of peptide solution was needed at a concentration of ~10 µg/µL starting material, thus requiring ~5 mg of starting protein material for each analysis. The autosampler makes automated injections onto a C18 column (300 µm id × 25 cm, 300 Å with 5 µm particles) at a flow rate of 4 µl/min and peptides were separated over 240 minutes with a gradient separation of 95% H_2O/ 5% ACN/ 0.5% FA to 30% H_2O/ 70% ACN/ 0.5% FA. Peptides were eluted directly into an electrospray source (Thermo Finnigan) with 100 µm i.d. fused silica. During the entire separation process, the mass spectrometer oscillates between full scans and MS/MS scans

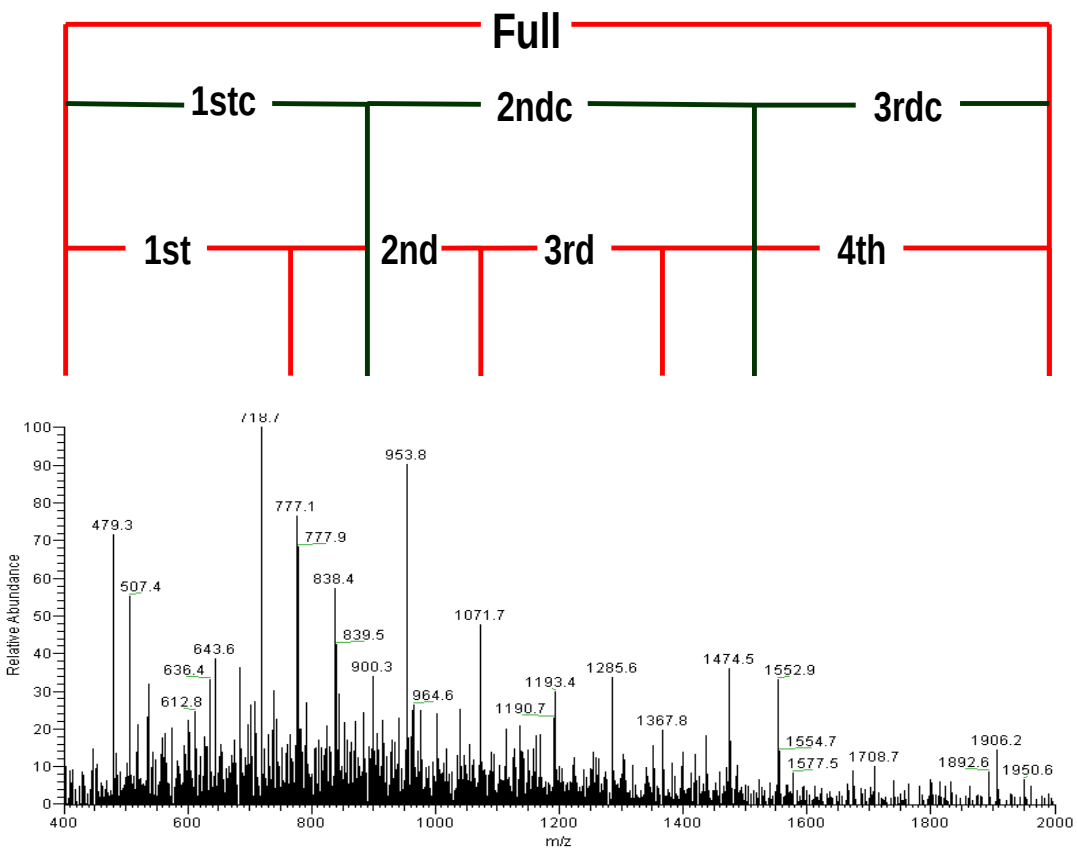


Figure 2.6: Multiple mass range scanning.
The general concept of multiple mass range scanning-the full scan spectra is divided into seven narrow overlapping m/z regions reducing the complexity the MS sees for any given analysis. The designations, Full, 1st, 1stc, 2nd, etc. indicate the segment of the m/z range being scanned in the file names.

in a data-dependent mode as discussed in Chapter 1. After the separation and MS analysis, the column was re-equilibrated to 95% H_2O/ 5% ACN/ 0.5% formic acid and the next injection and m/z range was applied. This entire process is repeated until all m/z ranges have been analyzed. The process was completely automated under the control of the Xcalibur software (Thermo Finnigan) and the user only needs to add more sample to the autosampler daily and add solvents to the HPLC. We have run the LC-MS/MS system in this mode for two weeks straight many times. Every two-three weeks, the MS must be vented and the source region cleaned. This methodology was employed for months straight on a single LC-MS/MS system for the large-scale analysis of the *R. palustris* proteome under major metabolic states as described in Chapter 5.

The major advantages of this methodology are simplicity, robustness, and ease of use. Once the system is operational, it can be kept running with very minimal human intervention (~15 minutes per day is all that is needed). We have also found the system to have very good reproducibility with very little downtime due to system failures. The major disadvantage of the system is primarily the sensitivity regarding the total amount of sample that is needed. Since it is most effective with electrospray sources, it inherently needs much more starting material than the other methods described below which employ nanospray sources. Furthermore, since multiple injections are being made of the same sample but only part of the mass range is being scanned, much of the observable peptides are lost. Generally, at least 5 mg of each proteome fraction is needed to obtain quality proteome fraction coverage (300-800 proteins depending on the fraction) while the two alternative techniques listed below can obtain the same results from 200-500 μg starting material. This technique is best used when a large number of samples need to be analyzed with minimal user involvement and when plenty of protein material is available.

2-dimensional switching LC/LC-MS/MS

After the development of the 1D-multiple mass range scanning technique discussed above, it became clear that a 2-dimensional separation platform was needed to address the shortcomings of the 1D method, especially the lack of sensitivity. The main

difficulties in designing a 2D separation methodology for peptides are the choice of separation phases, the ability to couple the phases together and the ease in automating the analysis. Most all 1D separation techniques use C18 reverse phase (RP) separations as the major mode of peptide separation. There are multiple reasons for this but mainly that this separation technique is very robust with high resolution for peptides and it can be directly connected with electrospray or nanospray source (the solvent systems are completely compatible). At the time, only three successful papers had been written illustrating functional 2D systems (Washburn, 2001; VerBerkmoes, 2002; Peng, 2002). We had tested two methodologies at the time, including strong anion exchange (SAX) of the intact proteins followed by digestion of the fractions and reverse phase separation LC-MS/MS. This methodology was found to be less than adequate for global "shotgun" proteomics due to the difficulties in separating intact proteins by SAX. This technique is useful in the combined top-down bottom-up analysis discussed in Chapter 3 and will continue to be employed in those experiments. The second method we attempted was direct connection of a commercial strong cation exchange (SCX) and a C18 RP column. Separation was achieved by similar method as Washburn et al. 2001 and described below. This method was found to have limited reproducibility, low robustness and was abandoned. Methodologies proposed by Washburn and Peng both relied on the capability to pack HPLC columns in-house, a capability which ORNL did not have at the time.

At that time, LC Packings (a division of Dionex, San Francisco, CA) introduced a new 2D LC system they designed based on switching column technology. We initiated a collaboration for testing, fully integration with the Thermo Finnigan LCQs and established working protocols for the analysis of bacterial, yeast and plant proteomes (http://www.lcpackings.com/, application notes, proteomics #10). The system layout for the columns is illustrated in Figure 2.7. It consists of three separate functional units: the Famos autosampler, the Switchos loading pump/column switching unit, and the UltiMate low flow HPLC (all LC Packings). The basic operational methodology is as follows: a peptide solution from a proteome fraction (generally 50 μL of a 10 μg/μL solution) was injected by the Famos autosampler onto the SCX cartridge (500 μm x 15 mm) on the

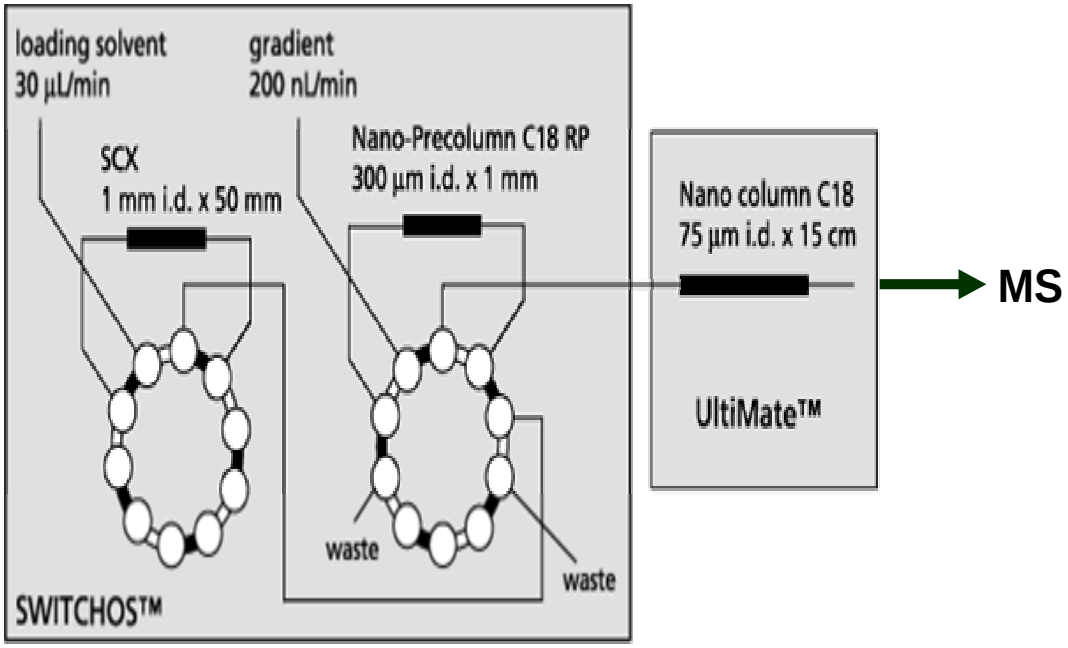


Figure 2.7: 2D switching LC/LC-MS/MS system.
The column layout for the switching 2D LC system which includes an SCX column for the 1st dimension (far left), an RP trapping cartridge (middle), and an RP nano-resolving column (far left) which is directly connected to the nanospray source on an ion trap MS. Figure courtesy of LC Packings.

Switchos valve one. This injection was made at a high flow rate of 30 µL/minute (100% H_2O/ 0.1% FA) allowing for complete sample loading in under 10 minutes. The peptides which are not caught by the SCX cartridge are caught by an RP trap (C18, 300 µm x 5 mm) which resides on valve 2. After loading of the sample, the SCX cartridge was moved out of line so no flow is going over the SCX cartridge and the RP cartridge is moved in-line with the nano resolving column (C18, 75 µm x 15 cm) which rests between the Switchos valve 2 and the nanospray source on an LCQ. A linear gradient of 95% H_2O/ 5% ACN/ 0.1% FA to 30% H_2O/ 70% ACN/ 0.1% FA provided by the UltiMate low flow HPLC (~200 nL/min) was back-flushed over the RP cartridge eluting peptides onto the nano resolving column where they are resolved by a 2-hour RP gradient into the nanospray source, ionized, and analyzed by data-dependent MS/MS on the ion trap MS. After the completion of the reverse phase gradient and the analysis of the unbound injection peptides, all the trap cartridges on the Switchos system were flipped back in line with the Switchos high flow and equilibrated to 100% H_2O/ 0.1% FA. The Famos autosampler then made an injection of 20 mM ammonium acetate from autosampler vials onto the SCX cartridge, which elutes peptides from this cartridge to the RP cartridge. The peptides are caught on this cartridge and completely de-salted for 10 minutes. Again, the RP cartridge was flipped in line with the nano-resolving column and peptides are again eluted by an RP gradient for 2 hours into the nanospray source and ion trap MS. This entire processes was repeated with injections of 50 mM, 100 mM, 200 mM, 400 mM, 600 mM, 800 mM, 1000 mM, and 2000 mM from autosampler vials giving a 10-cycle (including injection cycle) 2-dimensional analysis which takes ~24 hours. The entire process was fully automated and under control of the Xcalibur software system. The user only needs to put sample vials and vials with correct ammonium acetate concentrations into the autosampler, prepare the method in Xcalibur and start the system. After the 24-hour analysis, the columns can be thoroughly cleaned (high organic wash) and the next sample started.

This methodology was found to be fairly robust and sensitive and provides good dynamic range for protein complexes and proteomes. It was widely-used in our laboratory for two years, resulting in numerous publications (Strader, 2004; Wan, 2004;

VerBerkmoes, 2005). The main advantage of this system is a complete commercial 2D system which was fully automated and easy to implement. The system offers great flexibility in potential separation modes, though this has not been fully explored. The nano-resolving columns and the nano-spray tips never see the ammonium acetate, which increases their lifetimes. The main disadvantages of the system, compared with the 2-dimensional system described below, are the loss of sample on switching valves (overall sensitivity), overall dynamic range and lack of long-term stability of the system (the system required large amounts of maintenance and had extended down times due to component failures). In the end, this system has been entirely replaced in our laboratories for 2D separations of proteomes by the system described below but a modification of the system is routinely used for 1D analysis of protein complexes. Modifications are planned for the future to test and improve upon the system for 2D separations of proteomes.

2-dimensional MudPIT LC/LC-MS/MS

The integrated nano 2-dimensional LC system or multidimensional protein identification technology (MudPIT) technique was developed in the laboratory of Dr. John Yates Jr., Scripps Institute, San Diego, CA in 2001-2002 (Washburn, 2001; Wolter, 2001) and has since been applied by many laboratories for "shotgun" proteomics applications. The main principle of this methodology is an integrated nanocolumn of SCX and RP material. The main reason this technique was not immediately employed at ORNL was the expertise needed for the procedure and the long load times required for loading whole proteomes onto the system. The load time problem was solved with the development of the split-phase MudPIT columns in 2002 (McDonald, 2002). Dr. Hayes McDonald brought the technology to ORNL in spring of 2004. This technology, for reasons discussed below, has been found to be superior to the two methods listed above and has completely replaced them in the proteomics pipeline for the analysis of microbial proteomes.

The basic concept of the integrated split-phase biphasic nano column is illustrated in Figure 2.8. The split-phase columns were generally constructed as follows: the back column was packed with approximately 3.5 cm of strong cation exchange resin (Luna

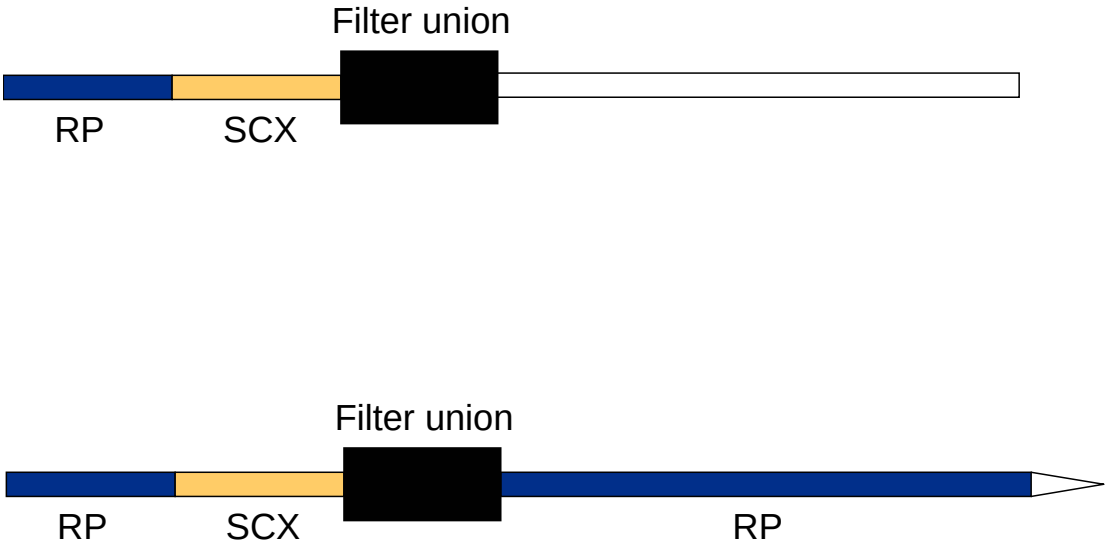


Figure 2.8: Split-phase MudPIT column.
The design of the split-phase MudPIT column (McDonald, 2002). Top is the back column which is packed with SCX followed by C18 RP material and then loaded with sample. Back column is then positioned behind C18 front column which is packed with C18 resin.

SCX 5 µm 100A Phenomenex, Torrance, CA) into a 100 µm fused silica via a pressure cell followed by 3.5 cm of C-18 reverse phase (RP) resin (Aqua C18 5 µm 200A Phenomenex) (top, Figure 2.8). The filter union acts as a frit to hold the packing material in the fused silica. The sample was then loaded off-line onto the dual phase column via the pressure cell. For most applications, ~200-500 µg of starting protein material was loaded onto the back dual phase column. Since there was no major impediment to flow in this design, samples can be loaded in ~30 minutes. Furthermore, samples can be directly de-salted on this system since most peptides are first caught by the RP material even in the presence of high salt.

The loaded RP-SCX columns were then directly connected behind a ~15 cm C18 RP column (Jupiter C18 5 µm 300A Phenomenex) also packed via pressure cell into Pico Frit tip (100 µm with 15 µm tip New Objective, Woburn, MA) (bottom, Figure 2.8). In this case, the tip acts as the frit holding the C18 resin in the fused silica column. The entire column system was then positioned into the nanospray source (Thermo Finnigan) on an ion trap mass spectrometer. The proteome samples were analyzed via a 2-dimensional separation of ammonium acetate pulses followed by reverse phase gradients (Washburn, 2001; Wolter, 2001; and Table 2.3). Cycle one was just a reverse phase gradient which moves peptides from the RP to the SCX material and elutes all peptides which do not bind to the SCX material into the nanospray source and ion trap MS. In the next cycle, the HPLC pump delivers a small pulse of ammonium acetate (7% of 500 mM ammonium acetate for 2 minutes). This moves another batch of peptides from the SCX material to the RP material. After a brief wash period, another RP gradient was run to elute peptides from the RP resin into the nanospray source and ion trap MS. This process was repeated as detailed in Table 2.3 for 24 hours until 100% of 500 mM ammonium acetate was reached and the run is finished. The back column was then disposed of and another column with loaded sample can be put onto the system.

This system has the best overall sensitivity of any of the systems and is the least expensive to implement once the necessary components have been purchased (pressure cell, packing material, fused silica, laser puller, etc.). Indeed, columns can be packed for less than $1 each, compared with $400-500 for the commercial columns used in the

Table 2.3: Cycles for typical 24-hour MudPIT experiment.

Cycle	Duration	Concentration
1	100	0
2	120	7
3	120	10
4	120	12
5	120	15
6	120	20
7	120	25
8	120	30
9	120	40
10	120	50
11	120	60
12	140	100

The Concentration is the percent of 500 mM ammonium acetate delivered by the HPLC.

above systems. Since the columns are so inexpensive, they can be disposed of after each analysis, eliminating the chance for cross-contamination between samples. The system also has better dynamic range than the above two systems, routinely identifying more proteins in 24 hours with less sample than is possible with either of the above two systems. The main disadvantage of this system is that some expertise is required in day to day operation of the systems and much more user input is needed (~1-2 hours is needed depending on user experience). We have observed similar reproducibility with this system as with the above two systems (~70-80% reproducibility in protein identification between replicate runs is common). For these reasons, we have primarily adopted the split-phase MudPIT technique for the characterization of microbial isolates or natural community proteomes (see examples in Chapters 6 and 7).

Mass Spectrometry

With the exception of Chapter 3, this entire dissertation employed the use of electrospray ion trap mass spectrometers for all MS analysis of "shotgun" proteomics experiments. The main reason for the use of electrospray ionization instead of MALDI is multifold. First, ES is straightforward to directly couple with liquid chromatography separations which were the core separation application for the studies presented here. For MALDI, this is not the case. While many efforts have been put forth to couple MALDI with LC separations, the process is not straightforward or routine. Second, ES provides better dynamic range for the analysis of mixtures than MALDI does. While all studies employed liquid separations prior to ES-MS analysis, the MS is very often scanning 20-200 peptides at any given point in time. ES is much more amenable to handling such complex mixtures than MALDI. Finally, ES produces more multiple charged peptide ions such as +2 and +3 parent ions. These ions are more amenable to sequencing by MS/MS than the +1 parent ions produced by MALDI.

Two types of ion trap mass spectrometers were employed for all studies-the quadrupole ion trap mass spectrometer (reviewed Stafford, 2002) and the linear ion trap mass spectrometer (Schwartz, 2002). The reason for primarily employing ion trap mass spectrometers instead of the myriad of other potential MS instruments is also multifold.

First, the ion traps are the most rugged instruments available for routine analysis of very complex mixtures. They routinely are used for 24/7 operations with very high up time and little down time for major maintenance. The ion trap mass spectrometers have been directly coupled to ES sources since the early 1990's, and the methodology for this coupling has been further optimized by the instrument companies. Ion trap mass spectrometers have good sensitivity and excellent dynamic range in the MS/MS mode. Finally, the Thermo Finnigan ion traps used in these studies have an excellent operating system (Xcalibur) that is best in its class for routine around the clock operations.

The quadrupole ion trap mass spectrometer is illustrated in Figure 2.9. At the start of this dissertation, this was the only major design for ion traps commercially available. While changes have been made in operating systems, electronics and ion transfer, the same basic design principle has been in use since the 1980's (reviewed, Stafford, 2002). As depicted in Figure 2.9, preformed solution phase peptide ions are sprayed through an electrospray or nanospray source on the front of the instrument into a heated capillary. The heated capillary is generally set at 150-250^{0}C and aids in the desolvation of the ions. The ions are then directed through a tube lens and through a skimmer. The skimmer acts to focus the ion beam and remove neutrals. The quadrupole (note the first octopole in the figure is actually a quadrupole on the LCQ MS) and octopole act strictly as ion beam guides to focus the ions into the ion trap. They are not used as storage devices or mass filters, as in some MS instruments. The ion beam enters the ion trap mass spectrometer through the end cap and is trapped inside by RF and DC potentials applied on the end caps and ring electrodes. All scan functions and ion manipulations occur inside the end cap and ring electrodes (this is the functional ion trap). Here, peptide ions are first trapped and scanned out by selectively destabilizing their orbital motion inside the ion trap. This is the full scan, and peptide ions are ejected from low m/z to high m/z from the ion trap out the back end cap and detected by the electron multiplier (EM). After the full scan, observed ions are selected by their m/z values for isolation and subsequent fragmentation. The ion of interest is selected by destabilizing and ejecting all other ions with a lower and higher m/z values essentially gas phase purifying the ion inside the mass spectrometer. The ion is then excited by

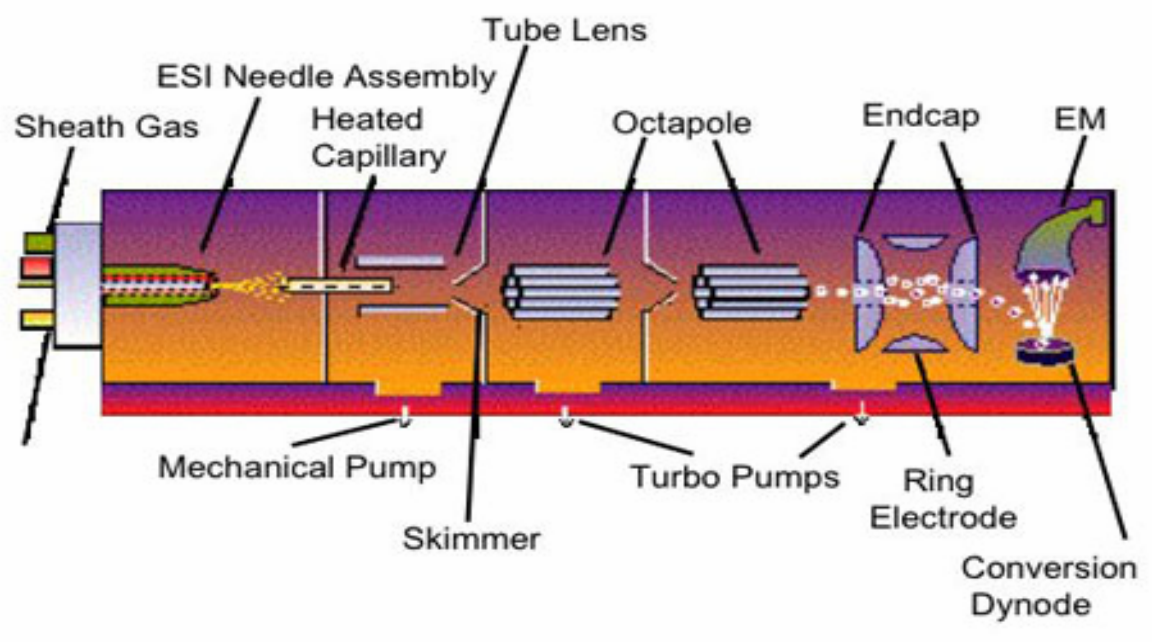


Figure 2.9: Quadrupole ion trap mass spectrometer (LCQ design).
Depicts the basic design of the Thermo Finnigan LCQ, a quadrupole ion trap mass spectrometer. EM-Electron multiplier.
Figure courtesy of Thermo Finnigan.

increasing its orbital frequency, which causes it to collide with the Helium bath gas which is always in the ion trap MS. These collisions cause collisional induced fragmentation. The fragment ions are still trapped within the ion trap MS. As above in the full scan, the fragment ions are selectively destabilized from low m/z to high m/z and injected through the end cap and detected by the EM. This process is repeated for three to four more ions and the MS returns to a full scan. This entire process is repeated through an entire chromatographic run, creating thousand of MS/MS spectra with full scan associated parent m/z measurements.

The linear ion trap mass spectrometer (Schwartz, 2002) was developed and commercially released in 2002. We obtained our first linear ion trap at ORNL in summer, 2004. This instrument is a major design improvement over the conventional quadrupole ion trap mass spectrometer and is depicted in Figure 2.10. The front-end instrumental design is very similar to the quadrupole ion trap mass spectrometer with an ion transfer tube replacing the heated capillary but performing the exact same function. A tube lens and skimmer still perform the same function of focusing the ion beam and removing neutrals. A small square quadrupole has been added for initial focus of the ion beam leaving the skimmer. The quadrupole and octopole still serve to focus the beam into the ion trap mass spectrometer. At this point, the major difference between the two instruments becomes apparent. While the quadrupole ion trap provides for trapping of ions in a three-dimensional field between two end caps and a ring electrode, the linear trap is much different. The linear trap is basically a long quadrupole with two lenses (the front lens and the back lens) which act to trap the ion packet in a two-dimensional field. The details of trapping and mass selective instability and resonance injections are outside the scope of this dissertation (see Schwartz, 2002 for details). Basically, the linear trap works the same as the quadrupole ion trap above except that the ions are ejected radially through slots in the rod by mass selective instability instead of axially through the end cap. Furthermore, two electron multipliers are positioned to detect ions on either side of the linear ion trap. The main advantage of the linear ion trap, when compared with the quadrupole ion trap, is that it can trap 10 times the number of ions without experiencing space charging effects. This leads to much better sensitivity and dynamic range,

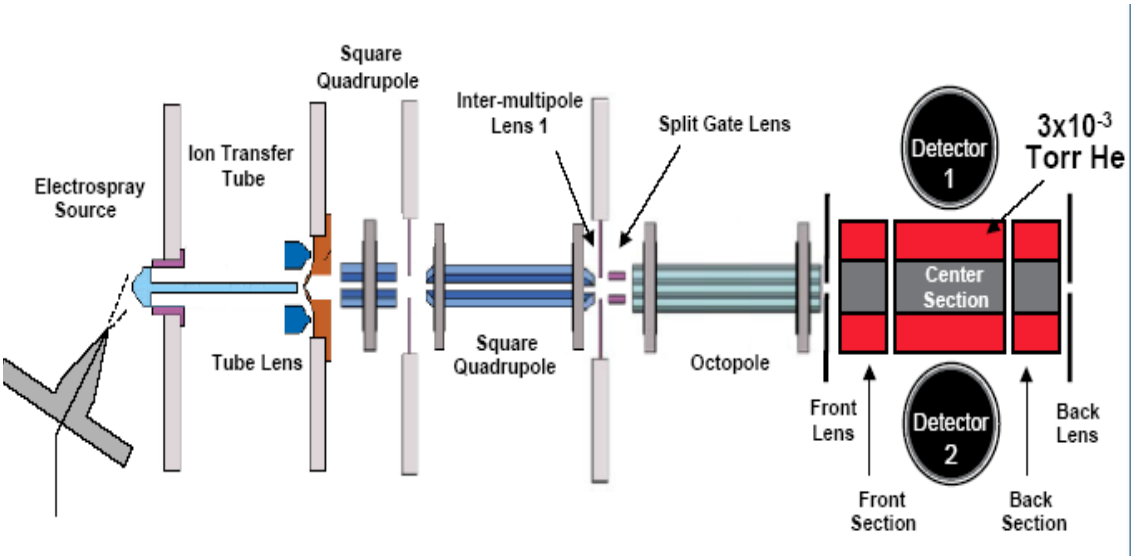


Figure 2.10: Linear ion trap mass spectrometer (LTQ design).
Depicts the basic design of the Thermo Finnigan LTQ, a linear ion trap mass spectrometer.
Figure courtesy of Thermo Finnigan.

especially in the MS/MS mode of operation. Also, the scan speed of this instrument is ~5 times faster than the quadrupole ion trap, as discussed above. The scan speed of the instrument is a critical parameter in analyzing the complexity of proteomes. In Chapter 7, the two instruments are compared with a similar complex proteome, illustrating the major enhancement for proteome analysis afforded by the linear ion trap mass spectrometer.

Proteome Informatics

One of the largest challenges in developing a proteomics platform was the development of a functional proteome informatics capability. At the start of this dissertation, the SEQUEST algorithm had been available for many years (Eng, 1994) for the analysis of MS/MS spectra against protein databases. But no major effort had been made to handle massive outputs generated from analyzing large datasets with SEQUEST. The SEQUEST algorithm had primarily been used to analyze individual MS/MS spectra from smaller experiments and not thousands of MS/MS spectra that are generated by a 24-hour "shotgun" proteomics experiment. At the time, the Finnigan Bioworks software package would generate an html output of the top identifications of all MS/MS spectra from a single individual LC-MS/MS analysis (2-4 hours worth of data). This output was not sorted by proteins' identifications and could only be filtered in a rudimentary way. Furthermore, there was no way to analyze the 8-12 LC-MS/MS files that generally make up a "shotgun" proteomics experiment. Thus it was necessary to export all identifications from all LC-MS/MS analyses into Excel, manually sort and filter, and then prepare lists of identifications. This method was very time-consuming, requiring hours to days of manual work for each "shotgun" proteomics experiment.

The development of DTASelect by Dr. David Tabb at Scripps Institute, San Diego, CA (Tabb, 2002) solved the major problems of data sorting and filtering. This software was provided as freeware to non-profit institutes and was immediately obtained and tested by ORNL. This software can take any number of LC-MS/MS analyses and sort and filter peptide identifications to provide hmtl and text output files of identified proteins. These files can be re-filtered at any time after analysis. Examples of

DTASelect file outputs from "shotgun" experiments can be found at (http://compbio.ornl.gov/biofilm_amd/analysis/analysis_lcq/) and at (http://compbio.ornl.gov/rpal_proteome/analysis/sequest_tryptic/). The DTASelect algorithm can perform the required sorting and filtering processes for thousands of MS/MS spectra in seconds to minutes, compared with hours required with the original methods we used.

At the same time, Dr. David Tabb introduced the Contrast software (Tabb, 2002). This software had the capability of comparing outputs from DTASelect files from multiple "shotgun" proteomics experiments. This software formed the basis for all comparative studies discussed in Chapters 4 and 5. Example Contrast output files can also be found at the websites given above. The SEQUEST algorithm, along with the DTASelect and Contrast software packages, has become the central core of the proteomics pipeline discussed below.

While the DTASelect and Contrast software packages solved many of the problems encountered early in our proteomics efforts, they did not solve the major problems that existed as we moved to high-throughput characterization of large numbers of microbial proteome samples (Chapters 5-7). The first of the main problems encountered was the speed of analysis. At the time, all SEQUEST analyses were done on a single PC processor which could not handle the massive load of data generated by the mass spectrometers. Second, we had no efficient way to release datasets to our collaborators or to the scientific community as a whole after publication. The latter is an especially important point. Currently, proteomics suffers since, for the most part, large proteomics publications are only accompanied with a list of identified proteins but there has been little effort to establish open access web-based proteomic results which contain more information than just a simple list of proteins. This is absolutely necessary for the field of proteomics to mature and flourish, as pointed out in Carr et al. 2004 and Pedrioli et al. 2004.

Through a collaborative effort between experimental and computational researches at ORNL a high-throughput proteome informatics pipeline has been developed and fully implemented at ORNL. This pipeline automatically takes LC-MS/MS raw

output files, generates the necessary input text files for SEQUEST, processes the MS/MS spectra with SEQUEST on a multi-processor cluster, concatenates the output files and automatically runs DTASelect and Contrast by user-defined settings. The resultant DTASelect and Contrast files are then posted to a secure website for viewing by collaborators. Upon publication, these websites are made publicly available giving the scientific community open access to all the results files including directly linkable MS/MS spectra output files for every identified peptide. The ORNL - UC, Berkeley AMD (Acid Mine Drainage) Community Proteome Study (http://compbio.ornl.gov/biofilm_amd/) and the ORNL *Rhodopseudomonas Palustris* Proteome Study (http://compbio.ornl.gov/rpal_proteome/) are, to our knowledge, the first examples of completely open access proteome results websites with detailed result files and explanation files.

Biological Data Mining

The extraction of confident and clear biological information from proteomics datasets is a challenge for all laboratories involved in proteomics efforts. The first and most straightforward task is to extract the proteins confidently identified from a given proteome dataset. While this is generally straightforward, there is no set standard on how the identified proteins should be filtered. The need to filter peptide and protein identifications from SEQUEST or any other search algorithm output should not be overlooked. This is a key step in quality proteomic experiments and often is not done correctly by the community as a whole. For all studies presented here, we filtered SEQUEST individual MS/MS identifications at the following cross-correlation (Xcorr) values [Xcorrs of at least 1.8 (+1), 2.5 (+2), 3.5 (+3)]. We then filtered the protein identifications at 2 unique peptides per protein identification. We have found these filter levels to be conservative, generally giving less than 1-5% false positive rates depending on the data sample size and the database size. By extracting protein identifications based on confident filtering levels, the analytical chemists can be certain the protein lists that are provided to the biologists are as accurate as possible.

The next major challenge is comparing the proteins identified under two or more metabolic states and determining proteins exhibiting major changes in abundance. This is a challenge for quantitative data but even more so for semi-quantitative data. Currently, we accomplished this task by comparing replicate analyses of metabolic states with the Contrast software. The identified proteins are manually inspected for major differences in % sequence coverage and number of peptides identified. Table 2.4 illustrates an unknown protein showing a major difference between an aerobic state and anaerobic state in *R. palustris* (see Chapter 5). For this case, it was easy to determine that the protein was up-regulated in one state compared to the other but in many cases the results are not so obvious. Currently, we use a basic rule of a replicated difference of at least 30% sequence coverage and/or 4 or more unique peptides between two metabolic states to indicate a potential difference in expression (see detailed discussions in Chapters 4 and 5). This current process is very time-consuming, requiring manual analysis of over 1,000 protein identifications for any given proteome comparison.

Another potential avenue for making large-scale comparisons between proteome datasets is mapping the identifications onto metabolic pathway maps such as KEGG maps (Kyoto Encyclopedia of Genes and Genomes) (Figure 2.11). This methodology has some shortcomings since many proteins do not directly map onto metabolic pathways and we have not completely figured out how to incorporate indicators of abundance such as % sequence coverage and number of unique peptides into the metabolic maps. Examples of some of our first attempts to map large proteome datasets onto metabolic maps can be found at http://compbio.ornl.gov/rpal_proteome/analysis/keggmaps/html/map01100.html.

The extraction of biological information from proteomics datasets is clearly one of the great challenges facing proteomic endeavors in the future and will be an active area of research. While much progress was made on this in this dissertation, it is by no means a completed research effort. Another major challenge is developing models of the system and proposing hypotheses that can be tested based on proteomic results. This is clearly a challenge since many of the proteins identified are of unknown function. Chapters 4, 5, and 7 illustrate potential hypotheses and models developed from proteomic data but clearly this is also an area of much needed active research.

Table 2.4: Unknown protein identified as up-regulated in *R. palustris* study.

	Aerob1	Aerob2	Anaerob1	Anaerob2
RPA3501			91.3	89.1
K.TSVSLEEAFWNGMK.E +1			2.9888	3.4893
K.TSVSLEEAFWNGMK.E +2			2.9401	3.9379
K.TSVSLEEAFWNGMKEISSVR.D +2			4.371	4.4753
K.TSVSLEEAFWNGMKEISSVR.D +3			4.1719	4.9396
R.ALQAQQQAVADTK.T +1			2.8616	2.9495
R.ALQAQQQAVADTK.T +2			4.5562	4.4718
R.ALQAQQQAVADTKTESSLTAH.- +2			5.0691	5.2418
R.ALQAQQQAVADTKTESSLTAH.- +3				4.163
R.DMTLSELVGEIDSNR.Q +1			2.5702	2.8333
R.DMTLSELVGEIDSNR.Q +2			4.2027	4.0589
R.DMTLSELVGEIDSNR.Q +3				4.2024
R.DMTLSELVGEIDSNRQQGNLSSAIR.L +3			5.0664	5.1391
R.LFVLDYFR.S +1			2.233	2.045
R.LFVLDYFR.S +2			3.3294	3.2622
R.SIVVAGHK.T +1			1.8586	1.8177
R.SIVVAGHKTSVSLEEAFWNGMK.E +2				2.9167
R.SIVVAGHKTSVSLEEAFWNGMKEISSVR.D +3			6.0787	
R.SRALQAQQQAVADTK.T +2			3.5345	

Numbers in row with RPA3501 are total % coverage.
Numbers next to sequences are charge states.
Numbers in columns under growth state are Xcorr values.

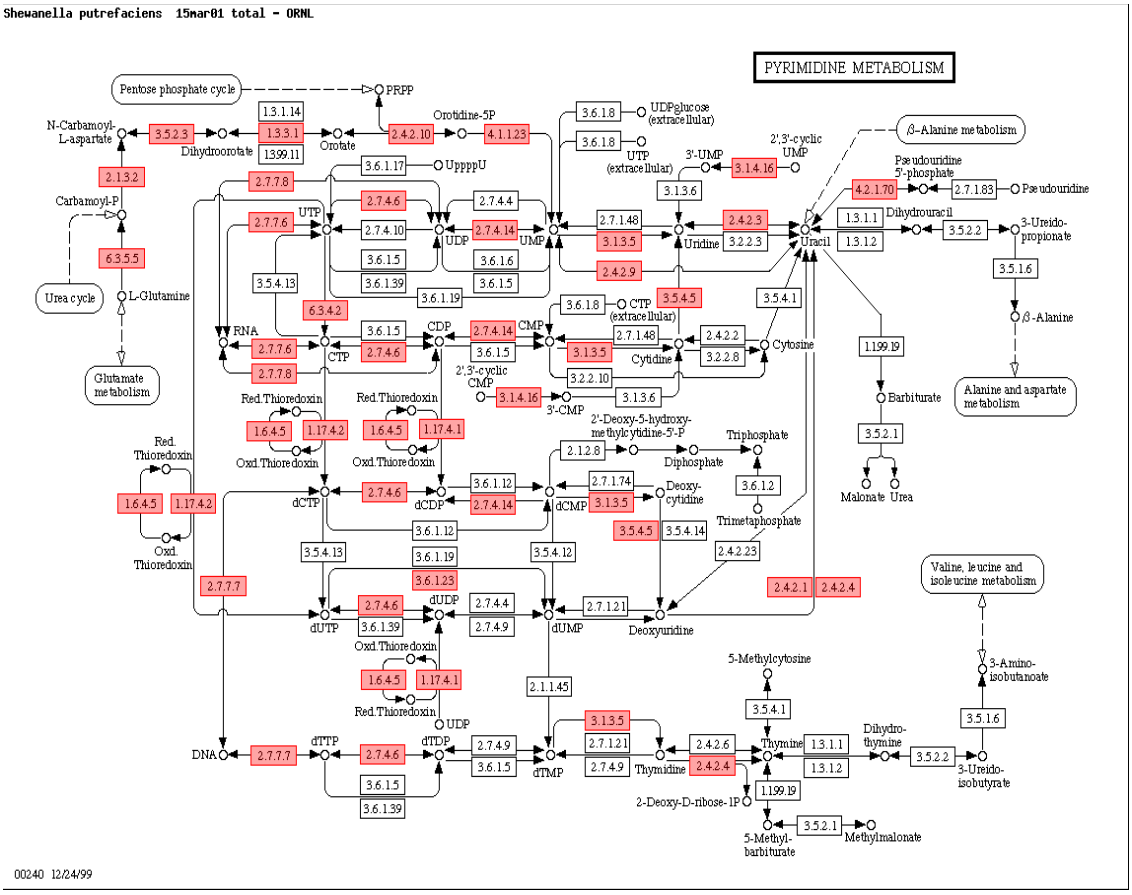


Figure 2.11: KEGG map of pyrimidine metabolism from *S. oneidensis* proteome. A KEGG map from the *S. oneidensis* MR-1 WT global proteome analysis (VerBerkmoes, 2002). Proteins highlighted in red were confidently identified.

Chapter 3

Application of the Integrated Top-Down Bottom-Up Methodology for the Characterization of Protein Complexes and Proteomes

All of the data presented below has been published as

Strader, M.B.; VerBerkmoes, N.C.; Tabb, D.L.; Connelly, H.M.; Barton, J.W.; Bruce, B.D.; Pelletier, D.A.; Davison, B.H.; Hettich, R.L.; Larimer, F.W.; and G.B. Hurst. Characterization of the 70S Ribosome from *Rhodopseudomonas palustris* using an Integrated "Top-Down" and "Bottom-Up" Mass Spectrometric Approach. *Journal of Proteome Research,* 2004; 3, 965-978.

VerBerkmoes, N.C.; Bundy, J.L.; Hauser, L.; Asano, K.G.; Razumovskaya, J.; Larimer, F.W.; Hettich, R.L.; and J.L. Stephenson Jr. Integrating "Top-Down" and "Bottom-Up" Mass Spectrometric Approaches for Proteomic Analysis of *Shewanella oneidensis*. *Journal of Proteome Research*, 2002; 1, 239-252.

All MS, sample preparation, experiments and data analysis on Rhodopseudomonas ribosomal complex were performed as a joint effort between Nathan C. VerBerkmoes, Brad Strader,and David Tabb, with assistance from Robert Hettich on top-down analysis.

All MS, sample preparation, experiments and data analysis on Shewanella proteome was performed by Nathan C. VerBerkmoes with assistance from Robert Hettich on top-down analysis.

Introduction

In the rapidly evolving field of proteomics, there is considerable interest in developing methods for large-scale, rapid, and robust analyses of proteins from complex biological samples. One of the major goals of these techniques is to obtain rapid identification of proteins as well as complete characterization of their intact molecular forms. Two major methods, or approaches, are currently employed for the analyses of complex protein mixtures. The most common method, often called bottom-up or "shotgun" proteomics, involves the digestion of a single protein, protein complexes, or proteomes with an enzymatic or chemical protease which creates small peptides of ~7-30 amino acids from the intact protein. These small peptides are very amenable to liquid chromatography separation and mass spectrometry analysis, usually through the method of tandem mass spectrometry, to obtain sequence information. The identification and

molecular form of the intact protein is literally built back up from the detection and identification of the resultant peptides. While this bottom-up proteomics approach is excellent for identifying a large number of proteins, it provides very limited molecular information about the intact proteins because rarely are peptides encompassing the entire protein sequence recovered from this methodology. While this may be a limitation, the bottom-up technique is the most widely applied to MS-based proteomics applications due to the straightforward nature of the methodologies.

An alternative strategy for proteome analysis, introduced by McLafferty et al. (Mortz, 1996; Kelleher, 1998), the top-down method, identifies proteins using accurate mass measurement and/or tandem mass spectrometry of intact proteins in order to generate sequence information. Since an intact mass is measured, this method may be advantageous for the detection of post-translational modifications, amino acid substitutions, and N-terminal processing. Such modifications to the intact protein may be overlooked in analyses by proteolysis-based (bottom-up) approaches, where only a fraction of the total theoretical peptide population of a given protein may be detected. Although means exist to include such modifications and potential amino acid changes in peptide-based searching algorithms, each possible modification introduced into a search increases the complexity and analysis time. Furthermore, without prior knowledge of the potential position of N-terminal processing (see Chapter 7), it is very difficult to identify cleavage positions on a large scale. The top-down approach also facilitates the detection of incorrectly predicted translational start sites. This approach is excellent for providing molecular level information for the intact proteins, but is limited in the numbers of proteins that may be detected from a given organism due to limited dynamic range, as well as the relative scarcity of bioinformatic tools to efficiently analyze this type of data.

We have developed a comprehensive method for protein characterization from complex mixtures that integrates features of both the top-down and the bottom-up approaches, capitalizing on the unique capabilities of each method. To our knowledge, we were the first group to develop detailed methodologies for the combination of these two techniques. This chapter describes the evolution of those methods and illustrates examples of the characterization of a protein complex and a whole proteome. The

analysis of single proteins to protein complexes to whole proteomes represents an increasing complexity which further challenges the analytical method. This chapter discusses advantages and disadvantages for the integrated methodology for the analysis of each of these, and discusses how the methodology should best be applied with its current level of technical development.

Our first attempt of the integrated top-down bottom up methodology was the characterization of an isolated single protein and a single protein in a complex mixture. We developed the technology on a series of protein variants (I68M, I68Q, Y69F, and Q67Y) from plasmid encoded R67 dihydrofolate reductase (DHFR) from *Escherichia coli*. The goal of these experiments was to develop rapid methodology for the characterization of recombinant over-expressed protein products as either isolated proteins or the over-expressed protein product in the initial cellular lysate. The analytical goal was to verify the position of the point mutation as well as verify the intact state of the over-expressed protein. The results from these initial experiments are not discussed in this chapter, but can be found in VerBerkmoes et al. 2002. These experiments represent our first attempts to develop and apply the integrated top-down bottom-up platform and resulted in a successful demonstration of the combined technology for rapid characterization of over-expressed recombinant proteins from either purified protein isolates or crude proteome mixtures.

The next step up in the level of complexity was the application of the methodology to the characterization of a protein complex. For these studies, we chose the 70S Ribosome from *Rhodopseudomonas palustris*. The ribosome has been a model protein complex for the development of MS-based proteomics techniques due to the ease of purification, the limited complexity and the presence of numerous post-translational modifications (Link, 1999). The ribosome is the universal macromolecular machine involved in translating the genetic code into proteins. Bacterial ribosomes are composed of a small subunit (30S) containing about 20 proteins and a single rRNA (16S), and a large subunit (50S) consisting of over 30 proteins and two rRNAs (23S and 5S). The ribosome from *Escherichia coli* is the most extensively characterized of the bacterial ribosomes. Ribosomes from bacterial species studied so far exhibit most, if not all, of the

homologues to ribosomal proteins found in *E. coli*. Furthermore, PTMs of ribosomal proteins from other bacteria tend to be similar to PTMs of *E. coli* proteins, with some variations in the corresponding modification positions. Thus, the purified ribosome from *Rhodopseudomonas palustris* made in excellent test case for the development of the integrated approach for characterizing protein complexes. For this study, the bottom-up approach was expanded to the use of 1D and 2D LC-MS/MS methodologies for the analysis of the enzymatically digested protein complex. This was necessary due to the increased complexity of the protein complex. The top-down methodology was moved from the ES-ion trap with ion-ion capabilities to the high resolution and high mass accuracy FT-ICR instrument. For these experiments, we performed LC-ES-FT-ICR for intact protein measurements. We have found this instrument to be superior in comparison with the ES-ion trap with ion-ion capabilities for the measurement of complex protein mixtures such as protein complexes and whole proteomes.

The next step up in the level of complexity was the attempted application of the methodology to the characterization of proteins directly from whole proteomes. For this study, we chose the *Shewanella oneidensis* MR-1 WT proteome. Concurrent analysis of the whole proteome by "shotgun" proteomics techniques provided in-depth knowledge of this proteome (VerBerkmoes, 2002). Here, we will limit discussion to the integrated top-down bottom-up characterization of the proteome. Detailed discussion on the entire proteome analysis by "shotgun" proteomics can be found in VerBerkmoes et al. 2002. For these studies, the proteome soluble fraction was separated at the intact protein level by strong anion exchange (SAX), the individual fractions were split in half with one half digested with trypsin and analyzed by 1D-LC-MS/MS and the other half analyzed by direct infusion ES-FT-ICR for the top-down application. We will illustrate the characterization of N-terminal processing with this example and point to advantages of the integrated approach for proteome analysis and the large challenges that still exist.

Materials and Methods

Chemicals and reagents

All salts, buffers, dithiothreitol (DTT), Bacterial Protease Inhibitor Cocktail, diethyl pyrocarbonate (DEPC), guanidine HCl, trifluoroacetic acid, glacial acetic acid, sucrose and RNase-free DNase I, were obtained from Sigma Chemical Co. (St. Louis, MO). In addition to using DEPC treated water to make buffers, RNase Away® (Molecular BioProducts, San Diego, CA) was also used to treat labware and bench-top surfaces to minimize RNase activity during the ribosome purification procedure. Sequencing-grade trypsin was purchased from Promega (Madison, WI). Formic acid was obtained from EM Science (Gibbstown, NJ). HPLC grade acetonitrile and water were used for all LC-MS analysis (Burdick & Jackson, Muskegon, MI). Ultrapure water used for sample buffers was obtained from a Milli-Q system (Millipore, Bedford, MA). BCA assay reagent and standards were obtained from Pierce Chemical Co. (Rockford, IL). Fused silica was purchased from Polymicro Technologies (Phoenix, AZ).

Methodologies for characterization of the ribosomal protein complex

Cell growth and preparation of 70S ribosomes

The wild-type strain, *Rhodopseudomonas palustris* CGA009 (a gift from Caroline Harwood, Dept. of Microbiology, University of Iowa), was grown either aerobically or anaerobically in a glass-walled fermentation vessel. Briefly, aerobic growth conditions, with air injected through the bottom of fermentation vessel, required media supplemented with 10 mM succinate (carbon source) without illumination (to eliminate photosynthesis). Anaerobic growth conditions required 10 mM succinate with the additional requirement of illumination and exclusion of air. All fermentations were run at 30^0C at pH 6.8. Cells were harvested at mid-log growth phase (O.D660 of ~0.8), and washed twice in ice-cold French Press buffer (100 mM ammonium chloride, 50 mM magnesium acetate, 20 mM Tris-HCl (pH 7.5), 1.0 mM DTT, 0.5 mM EDTA). After resuspending cells in the same buffer, a French Pressure cell (Thermo Spectronic, Madison, WI) was used to disrupt cells by applying 16,000 psi three times for 1 minute. DNase I was added to the resultant

suspension to degrade contaminant DNA for subsequent removal. Cellular debris was removed by centrifuging the lysate twice at 30,000 g in a SS-34 Sorval rotor for 30 minutes at 4^0C. The collected supernatant was then quick-frozen with liquid nitrogen and stored at -80^0C.

To separate 70S ribosomes initially from the remaining cellular components, the supernatant was layered at a 1:1 ratio (wt/wt) over a high salt sucrose cushion (20 mM Tris-HCl, pH 7.5, 50 mM magnesium acetate, 100 mM ammonium chloride, 1 mM DTT, 0.5 mM EDTA, 1.1 M sucrose) and centrifuged at 100,000 g in a Ti60 Beckman rotor for 16 hours at 4^0C. The ribosomal pellet was resuspended in a small volume (1-3 mL) of French Press buffer, aliquoted and stored at -80^0C for further use.

70S ribosomes were further purified and fractionated using sucrose density fractionation. Briefly, samples were layered on top of a 7%-30% linear sucrose gradient (10 mM Tris-HCl, pH 7.5, 6 mM magnesium acetate, 50 mM ammonium chloride, 1 mM DTT, 0.5 mM EDTA) and centrifuged at 85,000 g for 4 hours. After centrifugation, the gradients were fractionated and the absorbance at 260 nm was used to identify fractions containing ribosomes. Fractionated ribosomes were then pooled and recovered by centrifugation at 100,000 g for 16 hours.

Ribosomal protein extraction and the removal of contaminant rRNA was performed using the acid extraction method. The resuspended ribosomes were combined with 0.1 volume of 1 M magnesium chloride, then with 2 volumes of glacial acetic acid, and mixed by inversion for 2 hours at 4^0C. The insoluble fraction containing the contaminant rRNA was removed by centrifugation at 17,000 g for 30 minutes at 4^0C. After overnight dialysis in a 3,500 MWCO dialysis cassette (Slide-A-Lyzer, Pierce, Rockford, IL) against Ultrapure water, the protein samples were quantitated using the BCA assay.

LC-MS-MS for bottom-up proteomic analysis

All samples to be analyzed by the bottom-up approach were first digested with trypsin following the manufacturer's protocol and then desalted using C18 reverse-phase extraction (Sep-Pak, Waters, Milford, MA). Samples were then concentrated to ~0.1-1

μg/μl in a vacuum centrifuge (Savant Instruments, Holbrook, NY) and filtered with a 0.45 μm Ultrafree-MC filter (Millipore, Bedford, MA). Final peptide samples to be injected were in 100% H_2O with either 0.1% TFA (1D LC-MS-MS) or 0.1% formic acid (2D LC-MS-MS).

One-dimensional (1D) capillary LC-MS-MS experiments were performed with an UltiMate HPLC coupled to an LCQ-DECA or LCQ-DECA XP Plus quadrupole ion trap mass spectrometer (Thermo Finnigan, San Jose, CA), equipped with an electrospray source. Injections of typically 10-20 μg peptide digest were made using a Famos (LC Packings) autosampler with a 50 μl loop directly onto the column. The flow rate was 4 μl/min with a 160 min linear gradient from 100% solvent A (95% H_2O/ 5% ACN/ 0.5% formic acid) to 100% solvent B (30% H_2O/ 70% ACN/ 0.5% formic acid). The C18 column (300 μm i.d. x 25 cm, 300Å pore size, 5 μm particles; Vydac 218MS5.325 or Vydac 238EV5.325) was connected to the electrospray source with 100 μm i.d. fused silica tubing. Typical electrospray (ES) voltage was 4.5 kV and typical heated capillary temperature was 200-225^{o}C. The mass spectrometer was operated in the data-dependent MS-MS mode with dynamic exclusion enabled and a repeat count of 1. In this mode, four parent ions from each mass spectrum were chosen automatically for MS-MS analysis based on an ion's (1) abundance in the mass spectrum, and (2) absence from an "exclusion list" of parent ions that had, more times than the "repeat count" setting, been subjected to MS-MS analysis in the previous 1 minute time window. Data-dependent LC-MS-MS was performed over a parent ion m/z range of 400-2000. In some experiments, to increase dynamic range, separate injections were made while scanning several narrower parent ion ranges (m/z 400-1000, m/z 980-1500, and m/z 1480-2000) in addition to the full m/z range of 400-2000 (multiple mass range scanning, see Chapter 2 for details).

Two dimensional (2D) LC-MS-MS experiments were performed using a similar setup, with the following changes. Injections of 10 to 30 μg sample were made with the Famos autosampler onto a strong cation exchange column (LC Packings SCX, 500 μm i.d. x 15 mm), located on 10-port switching valve A of a Switchos system (LC Packings). The first dimension separation consisted of a series of step gradient elutions from the

SCX column affected by 9-12 subsequent injections, using the Famos autosampler, of ammonium acetate salt at concentrations of 25 mM, 50 mM, 100 mM, 200 mM, 400 mM, 600 mM, 800 mM, 1 M, and one to four injections of 2 M. Peptides eluting from the SCX column after each salt injection were captured on an LC Packings reverse-phase precolumn (300 μm i.d. x 5 mm, 300Å PepMap) on Switchos valve B. After washing salt from the precolumn, Valve B was switched to direct flow from the reverse-phase precolumn in back flush mode onto a nano-scale Vydac 218MS5.07515 C18 analytical column (75 μm i.d. x 15 cm, 300Å pore size, 5 μm particles). This second dimension separation employed a 150 min gradient, going from solvent A (95% H_2O/ 5% ACN/ 0.1% formic acid) to solvent B (30% H_2O/ 70% ACN/ 0.1% formic acid) at 200 nl/min to elute peptides into the mass spectrometer via a Thermo Finnigan nanospray source. The LCQ was run in the data-dependent mode with dynamic exclusion enabled and a repeat count of 2.

Protein identification from bottom-up data analysis

The entire published *R. palustris* database (Larimer, 2004) was used initially to analyze MS-MS spectra from bottom-up experiments using the SEQUEST algorithm (Thermo Finnigan). Initial searches were configured to include only tryptic peptides. Data representing the best 1D and 2D runs from the preliminary results were then re-analyzed using SEQUEST by searching against all predicted peptides, without specifying tryptic cleavages. For SEQUEST post-translational modification searches, a subset of the *R. palustris* sequence database was used. This database contained all ribosomal proteins and other proteins for which at least one peptide was observed in either the best 1D run or the best 2D run. For PTMs, we specified the following: The first search allowed mass shifts of 14 Da to detect methylation on lysine and arginine residues, and 16 Da to detect oxidations on methionine, cysteine, and tryptophan residues. The second search permitted mass shifts of 28 Da to detect dimethylations on lysines and arginines and 16 Da for methionine, cysteine and tryptophan residues. The third search permitted mass shifts of 42 Da to detect acetylations and trimethylations on lysine and arginine and 16 Da for methionine, cysteine, and tryptophan residues. The fourth search permitted 46 Da

to detect β methylthiolation on aspartic acid and 16 Da for methionine, cysteine, and tryptophan residues. Two more searches aimed at identifying N-terminal modifications were performed to identify methylations (14 Da) and acetylation/trimethylation (42 Da) at the N-termini of peptides. Note that the PTMs specified for the bottom-up vs. the top-down searches differ. Particular amino acid residues or termini can be specified for the bottom-up search, but not for the top-down search; furthermore, the tools for performing the searches differ in their natures and limitations.

The programs, DTASelect and Contrast (Tabb, 2002), were used to assemble, filter, and compare the identifications from SEQUEST searches on various experimental datasets. DTASelect's default SEQUEST score cutoffs were used; spectra from singly-charged peptides were required to exceed 1.8 in Xcorr, while Xcorr values for doubly- and triply-charged peptides were required to exceed 2.5 and 3.5, respectively. Contrast combines DTASelect results from several different bottom-up experiments to summarize numbers of peptides identified and other parameters, grouped by protein (see chapter 2 for detailed discussion of proteome informatics).

Electrospray FT-ICR for top-down proteomic analysis

High resolution mass spectra were acquired using an UltiMate HPLC (LC Packings/Dionex, Sunnyvale, CA) coupled to a 9.4 T HiRes electrospray Fourier transform ion cyclotron resonance mass spectrometer, ESI-FTICR MS (IonSpec, Lake Forest, CA). The HPLC flow rate was 4 μl/min with a 60 min linear gradient from 100% solvent A (95% H_2O/ 5% acetonitrile [CAN]/ 0.5% formic acid) to 100% solvent B (5% H_2O/ 95% ACN/ 0.5% formic acid). A C4 reverse-phase column (model 214MS5.325, 300 μm i.d. x 15 cm, 300Å pore size, 5 μm particles, Grace-Vydac, Hesperia, CA) was directly connected to the Analytica electrospray source with 100 μm i.d. fused silica capillary tubing. Ions were generated with a 3,700 V potential between a grounded needle and heated transfer capillary, accumulated in an external hexapole for 2 seconds, transferred into a high-vacuum region using a quadrupole lens system, and then detected in the mass analyzer. A broadband mass resolution of at least 50,000 (full width at half maximum) at m/z 1,000 was possible because ion detection was achieved in an ultra high

vacuum regime (~2 x 10-10 Torr). Standard proteins (ubiquitin, myoglobin) and peptides (leucine enkephalin, gramicidin S) were used for mass calibration. The high-resolution mass measurement enabled isotopic resolution of multiply-charged ions. The charge state of a multiply-charged ion could, therefore, be determined by directly measuring its isotopic spacing. Deconvoluted molecular mass spectra were generated with the IonSpec software. By calibrating on the calculated values of the most abundant isotopic peaks for the six different charge states (7+ to 12+) of the protein standard ubiquitin, the deconvoluted mass spectrum yielded a measured molecular mass that was within 0.025 Da (3 ppm) of the calculated value. The external calibration procedure enabled molecular measurement accuracy of ≤ 10 ppm for most proteins with molecular masses up to 40 kDa, although the mass errors were slightly larger for the more minor abundance proteins, for which the isotopic packets are somewhat distorted.

Because the mass resolution was at least 50,000 for the intact protein measurements, the molecular masses of these proteins could be measured with isotopic resolution. The measured most abundant isotopic mass (MAIM) of each molecular ion region was used as an approximation of the protein's isotopically-averaged molecular mass in order to query a database of all possible *R. palustris* proteins. This database query with the MAIM values was conducted with a reasonably large molecular mass tolerance window (+ 5 Da) to accommodate the fact that the abundances (but not the mass values) of the ions in the measured isotopic packet may vary somewhat from their calculated values. This is especially noticeable in the larger proteins, where the abundances of the isotopes around the average molecular mass are very similar. Even slight variations in the mass spectrometric measurements can result in peak abundance variations of a few percent, which can alter the most abundant isotope observed in these cases. This search usually revealed between 1 and 4 possible protein matches within the "crude" 5-Da window, with a close match to at least one protein in the database. Calculated masses for both intact proteins and proteins with N-terminal methionine truncation for all possible *R. palustris* proteins were searched in this initial screen. To refine a tentative protein match, the isotopically-resolved molecular mass region of the suspected protein was calculated, based on its sequence, and compared to the measured

data from the FT-ICR-MS experiment. Because these experiments were conducted under external calibration conditions, the mass of the most abundant isotopic peak for each matched protein from the database was required to be within 10 ppm (i.e., a few millidaltons) of the measured value for the more abundant signals, with somewhat lower mass accuracy (less than 30 ppm) permitted for more minor species. For the entire suite of 54 possible ribosomal proteins, an intact protein look-up table was extracted from the full *R. palustris* protein database; this intact protein table contained intact molecular masses, methionine-truncated molecular masses, and all possible combinations of methionine truncation with single acetylation and multiple methylations (up to 9). The experimental FT-ICR-MS data were used to query this look-up table for tentative PTM protein forms. All possible matches were compared against the results obtained from the bottom-up data.

Methodologies for characterization of the *Shewanella oneidensis* proteome by the combined top-down bottom-up technique

Cell growth and pre-fractionation

S. oneidensis MR-1 cells (4 L culture in LB Broth) were grown aerobically, harvested in mid-log growth phase (OD_{600}= 1.0) and washed twice with 50 mM Tris pH 7.5. Cells were resuspended in ice-cold lysis buffer (50 mM Tris, pH 7.5 with 1% Bacterial Protease Inhibitor Cocktail) and disrupted by sonication (Misonix, Farmingdale, NY) on ice with a microtip probe using 5 s bursts with a 5 s rest period for 5 min. Unbroken cells were pelleted by centrifugation at 5000 *g* for 30 minutes and discarded. The first fractionation was prepared by pelleting insoluble material at 20,000 g for 1 hour. The supernatant was collected and frozen at -80^{0}C until a later fractionation was performed ("crude lysate"). For these studies, only the crude lysate was used. "Shotgun" proteome characterization of the entire proteome can be found in VerBerkmoes et al. 2002. The crude fraction was analyzed by the BCA assay (Pierce, Rockford, IL) to determine protein concentrations.

Anion exchange fractionation

Samples of the crude lysate (2.0 mL) were fractionated on a Pharmacia (Piscataway, NJ) Source 15Q PE 4.6/100 quaternary ammonium strong anion exchange column attached to an Akta FPLC® (fast protein liquid chromatography) system using a linear gradient of 1 M NaCl in 20 mM Tris, pH 8.0 in 30 column volumes (approx. 51 mL) at a flow rate of 2.0 mL/min. Fractions (1 mL) were automatically collected for analyses by both the bottom-up and top-down approaches and stored at –80°C until needed. Fractions were split in half, with one half analyzed by the bottom up technique and the other half by the top-down technique.

LC-MS/MS for bottom-up proteomics

All fractions destined for bottom-up analysis were digested with trypsin following the manufacturer's protocol (Promega, Madison, WI). The samples were then de-salted with a C_{18} Sep-Pak (Waters, Milford, MA), dried to completion in a centrifugal evaporator (Savant Instruments, Holbrook, NY), resuspended in 5% TFA, and filtered through a 0.2 μm filter (Schleicher & Schuell, Keene, NH). Samples were resuspended in the appropriate amount of solvent for the number of necessary injections, depending upon the experiment.

All LC-MS/MS experiments were performed on an UltiMate HPLC (LC Packings, a division of Dionex, San Francisco, CA) coupled to an LCQ-DECA ion trap mass spectrometer (Thermo Finnigan, San Jose, CA) equipped with an electrospray source. The HPLC was operated in the capillary flow rate mode (4.0 μL/min), using a plug-in to the Xcalibur software provided by LC Packings. Two columns were used for 1D experiments, a VYDAC (Hesperia, CA) C_{18} column (218MS5.315, 300 μm i.d. x 15 cm, 300Å with 5 μm particles) and an LC Packings C_{18} column (300 μm i.d. x 15 cm, 100Å with 3 μm particles). The solvents used for chromatography were as follows: A: 95% H_2O/ 5% acetonitrile/ 0.5% formic acid, and B: 30% H_2O/ 70% acetonitrile/ 0.5% formic acid.

For all LC-MS data acquisition, the LCQ was operated in the data-dependent mode, where the top four peaks in every full MS scan were subjected to MS/MS analysis.

The settings for MS/MS were as follows: default charge state: 3; default isolation width: 3; normalized collision energy: 35%; activation q: 0.250; activation time: 30.00 ms. Five microscans were acquired for every full and MS/MS scan. The dynamic exclusion feature of the Xcalibur software was also enabled, with the following settings used: exclusion mass width: +/- 2.5 m/z; repeat count: 1; repeat duration: 0.5 min; exclusion duration: 1.00 min.

The 15 fractions obtained from the anion exchange separation were analyzed via a 1D-LC-ES-MS/MS with multiple mass range scanning experiment employing two mass range scans. For each experiment, approximately 200-800 μg of starting material was used, depending on the protein concentration of the fraction. After digestion and clean up, the samples were diluted to a total volume of 30 μl to allow for two injections over two mass ranges: 400-1000 m/z and 980-2000 m/z. The gradient for each of the injections was as follows: 0-10 min 100% A, 10-130 min 40% B, balance A, 130-145 min 50% B, balance A, 145-165 min 100% B. The entire half of the fraction was used for each digest. To aid in the top-down analysis and show reproducibility, other fractions from multiple anion exchange runs were analyzed (the whole fraction was not used and a single mass range was employed).

Protein identification from bottom-up data analysis

A *Shewanella oneidensis* MR-1 protein database was created from the preliminary genome sequence obtained from The Institute for Genomic Research website at http://www.tigr.org. For the bottom-up approach, all MS/MS data was searched off-line on a dual processor (1.7 GHz Pentium Xeon) workstation (Dell, Round Rock, TX) with the SEQUEST algorithm (Thermo Finnigan) against the annotated *S. oneidensis* protein database. The global SEQUEST settings used were as follows: threshold: 100,000; enzyme: trypsin; number of internal missed cleavage points: 4; Peptide mass tolerance: +/- 3.0 (average mass); fragment ion mass tolerance: +/-0.4 (monoisotopic mass); parent mass range 300-5000 daltons; filters 4 of 5: Xcorr 1.0; DelCN 0.1; sp 500; Rsp 5.0; 30% fragment ions. The output data from this search was first stored in Microsoft Excel for later use and then filtered further and sorted by gene locus number with DTASelect

software (Tabb, 2004). The default settings of DTASelect were used for all searches, which include a minimum cross-correlation (Xcorr) of 1.8 for +1 peptides, 2.5 for +2 peptides, and 3.5 for +3 peptides. We required a minimum of two unique peptides with the above qualifications for any given gene locus to be accepted as a positive hit.

ES-FT-ICR for top-down proteomics

A portion (about 500 μL) of each FPLC fraction was prepared for the top-down proteomics approach by dialysis against 4 L of H_2O for 10 hours in a 3500 MWCO dialysis microtube (Pierce, Rockford, IL). The procedure was repeated, with the second dialysis run for 4 hours. Samples were stored at -80^0C until analysis. Samples were prepared for MS by mixing 60 μL of sample with 40 μL of acetonitrile and 2 μL of acetic acid.

All mass spectra were acquired with an IonSpec (Irvine, CA) 9.4-Tesla HiRes electrospray Fourier transform ion cyclotron resonance mass spectrometer (ES-FT-ICR-MS), essentially as described above. The *S. oneidensis* MR-1 protein database was created from the preliminary genome sequence obtained from The Institute for Genomic Research website at http://www.tigr.org. From this data, an *S. oneidensis* intact protein database search tool was created to provide a means to search both intact average molecular mass and molecular mass minus N-terminal methionine. The most abundant molecular masses, as calculated from the ES-FT-ICR-MS data, were used to search this database. For confirmation of mass assignments in top-down experiments, the theoretical most abundant molecular mass for intact protein ions was calculated with IsoPro.

Results

Top-down and bottom-up characterization of the 70S ribosome

The 70S ribosome from *R. palustris* was characterized with the integrated top-down and bottom-up technique. Figure 3.1 illustrates the strategy for the top-down and bottom-up approach adopted in this study. Integration of results was achieved, as shown by the dotted-line arrows in Figure 3.1, by using protein identifications from analysis of

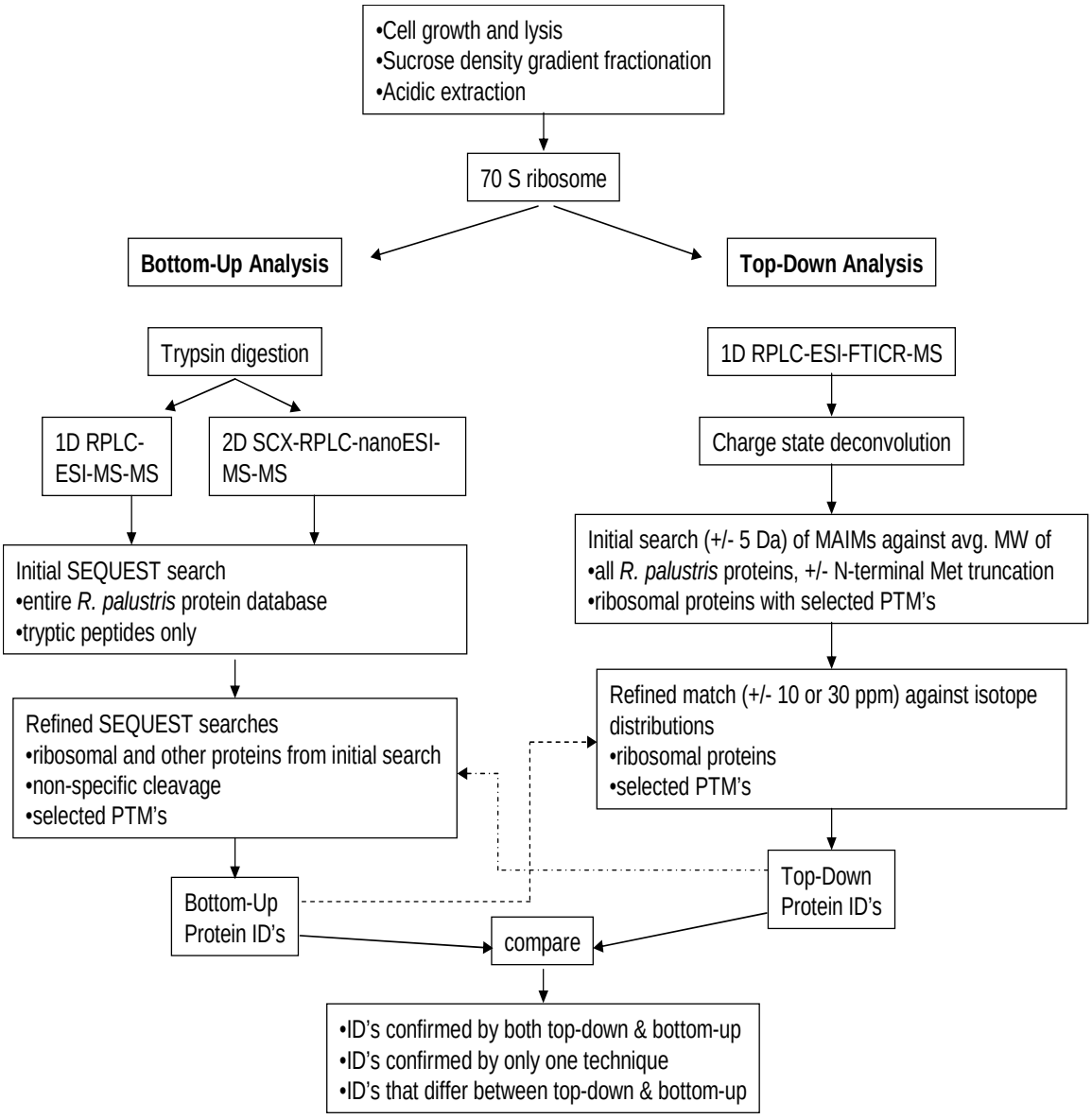


Figure 3.1: Strategy for top-down, bottom-up MS analysis of ribosomal proteins. Integration of results from the two approaches was achieved, as the dashed and dot-dash arrows show, by iteratively using the results from each approach to augment and expand the results of the other.
Figure provided by Dr. Greg Hurst and Dr. Brad Strader.

top-down data to refine analysis of bottom-up data, and vice versa, in an iterative manner to increase the number of characterizations of ribosomal proteins obtained. For example, identification of a methylated protein by the top-down approach could provide motivation to examine more closely the bottom-up results for the presence of a methylated peptide from that protein. The combined top-down bottom-up MS analysis identified a total of 53 of the predicted 54 ribosomal proteins. The data indicated the presence of 21 proteins for the small subunit and 33 for the large subunit (S20 and L26 are identical). No orthologue of *E. coli* S22 was identified for *R. palustris* ribosomes. We also identified isoforms for L7/L12 from the large subunit. The traditional nomenclature for ribosomal proteins was adopted from studies of *E. coli*, where L1-L36 represents ribosomal proteins of the large subunit and S1-S22 denote proteins from the small subunit. In this paper, each of the *R. palustris* ribosomal proteins (RRP) is named after the corresponding ribosomal protein in *E. coli*. The L7/L12 isoforms were therefore named RRP-L7/L12A and RRP-L7/L12B (discussed later).

HPLC separation strategies for bottom-up analysis

The mixture complexity of a tryptic digest from purified ribosomes is intermediate between that of a single protein digest and a digest from a whole proteome. Therefore, we compared several chromatographic strategies for the peptide separation in the bottom-up approach, including one dimensional (1D) reverse phase liquid chromatography (RPLC), and two dimensional (2D) separations employing both strong cation exchange (SCX) and RPLC (2-dimensional switching LC/LC-MS/MS; see Chapter 2 for details). The criterion for this comparison was maximum sequence coverage of ribosomal proteins, which would be necessary for a comprehensive examination of post-translational modifications.

After optimizing both separation and MS protocols, and examining sequence coverage obtained from initial SEQUEST searches, we selected the best 1D and 2D data sets for further SEQUEST analysis tailored for identifying PTMs. Table 3.1 compares the results obtained using the various separation strategies. A simple 1D RPLC separation required the least measurement time, but provided the smallest number of

Table 3.1: Summary of bottom-up analyses of ribosomal proteins[a].

Run #	LC Method	Mass Ranges[b]	Sample Amount	Spectra Produced[c]	Proteins Identified[d]	High-Scoring Spectra[e]	Identified Peptides	Average Sequence Coverage
1	1D	1	10 µg	4402	41	338	186	31%
2	1D	4	72 µg	10845	52	1071	604	51%
3[a]	1D	4	48 µg	12906	51	1198	610	57%
4[a]	2D	1	10 µg	46033	51	3737	821	60%
5	2D	1	30 µg	54712	50	5199	672	56%

a) Data from Runs 3 and 4 (shown in bold) subjected to more detailed SEQUEST analysis (see text for details).
b) Number of mass ranges for MS measurement (see Materials and Methods).
c) Total MS-MS spectra acquired.
d) Search considered all possible *R. palustris* proteins, peptides resulting from trypsin digestion, and no PTMs.
e) Spectra that met the default Xcorr cutoffs (see Materials and Methods).

protein and peptide identifications. The same 1D RPLC separation, repeated four times, each time targeting a different m/z range (multiple mass range scanning, see Chapter 2 for details), yielded a 3-fold larger number of peptide identifications. This four m/z range 1D experiment was performed twice, identifying 604 different peptides in the first run and 610 in the second; the second run was selected for more extensive SEQUEST analysis due to slightly higher average sequence coverage per ribosomal protein and overall number of peptides identified, despite the use of less sample. Both the four m/z range 1D and the 2D strategies resulted in similar numbers of identified ribosomal proteins and overall sequence coverage from the initial SEQUEST searches. Of the two 2D experiments, the run using 10 μg of sample produced a significantly larger number of peptide identifications than the 30 μg run, and so the former was chosen for further SEQUEST analysis. Although requiring the longest measurement time, the 2D method required less starting material than the multiple-mass-range 1D measurement, and produced the most confidently identified spectra, probably by decreasing the complexity of the peptide mixture introduced into the mass spectrometer at any particular time. Loading three times the sample amount on the 2D system resulted in a slight decrease in the mean sequence coverage of ribosomal proteins, while the sequence coverage on common contaminants was increased. This is a common observation of overloading 2D systems. This comparison suggests that 1D separations with multiple mass range scanning and 2-dimensional switching LC/LC-MS/MS are complementary in regards to quality of results obtained, amount of sample required, and time requirements. Indeed, if time and sample permits, the use of both techniques is advantageous since some peptides from some proteins will be more readily identified by one technique over the other.

Protein sequence coverage and protein identifications from bottom-up analysis

All but two ribosomal proteins were observed in the bottom-up analyses that were chosen for SEQUEST analysis tailored for PTM identification (see Table 3.2). The 1D, four mass-range analysis failed to observe RRP-L34, and the 2D analysis did not identify RRP-L34 or RRP-L36. These two proteins both have a high percentage of basic residues. RRP-L34 has five lysines and twelve arginines in its sequence of 44 residues, for an

Table 3.2: Sequence coverage and peptide identifications for (bottom-up) 1D and 2D analysis.

Name	Sequence Coverage		# Peptide Identifications*	
	1D	2D	1D	2D
RRP-L1	66.0%	80.0%	25	43
RRP-L2	46.0%	58.0%	12	19
RRP-L3	79.0%	92.0%	23	45
RRP-L4	68.0%	78.0%	22	19
RRP-L5	54.0%	45.0%	10	14
RRP-L6	36.0%	50.0%	9	19
RRP-L7	51.0%	60.0%	16	18
RRP-L9	90.0%	56.0%	13	17
RRP-L10	91.0%	91.0%	27	45
RRP-L11	61.0%	69.0%	10	20
RRP-L13	81.0%	87.0%	17	24
RRP-L14	80.0%	84.0%	11	16
RRP-L15	78.0%	85.0%	20	24
RRP-L16	65.0%	77.0%	14	26
RRP-L17	66.0%	76.0%	13	22
RRP-L18	73.0%	73.0%	12	27
RRP-L19	78.0%	72.0%	18	24
RRP-L20	51.0%	57.0%	10	11
RRP-L21	58.0%	22.0%	5	4
RRP-L22	63.0%	57.0%	12	18
RRP-L23	44.0%	86.0%	5	13
RRP-L24	89.0%	89.0%	11	14
RRP-L25	74.0%	62.0%	20	30
RRP-L27	63.0%	87.0%	8	15
RRP-L28	77.0%	55.0%	13	13
RRP-L29	45.0%	32.0%	5	5
RRP-L30	91.0%	91.0%	6	6
RRP-L31	80.0%	100.0%	6	9
RRP-L32	58.0%	58%	2	2
RRP-L33	56.0%	66.0%	4	9
RRP-L34	0	0	0	0
RRP-L35	27.0%	49.0%	3	5
RRP-L36	22.0%	0	1	0
RRP-S1	37.0%	39.0%	12	16
RRP-S2	72.0%	87.0%	29	41
RRP-S3	57.0%	68.0%	16	26
RRP-S4	78.0%	85.0%	20	29
RRP-S5	72.0%	70.0%	22	25
RRP-S6	63.0%	69.0%	19	21
RRP-S7	72.0%	82.0%	22	28
RRP-S8	87.0%	71.0%	17	21
RRP-S9	84.0%	74.0%	18	27
RRP-S10	39.0%	66.0%	4	5
RRP-S11	86.0%	33.0%	10	13
RRP-S12	67.0%	63.0%	11	11
RRP-S13	33.0%	72.0%	6	21
RRP-S14	36.0%	50.0%	6	7
RRP-S15	87.0%	99.0%	14	17
RRP-S16	44.0%	56.0%	8	16
RRP-S17	83.0%	79.0%	12	10
RRP-S18	58.0%	61.0%	7	11
RRP-S19	98.0%	84.0%	12	19
RRP-S20	36.0%	23.0%	5	3
RRP-S21	23.0%	51.0%	3	9

* # of different peptide Ids including +1, +2 and +3 charges states for identical peptides.

average of 2.6 residues between trypsin cut sites; RRP-L36 averages 2.7 residues between trypsin cut sites. Because these sequences are so rich in trypsin cleavage sites, many of the resulting peptides fall below the lower m/z limit for isolation and fragmentation. Interestingly, we identified the intact mass of RRP-L36 but not RRP-L34 from the FT-ICR analysis (discussed in more detail later).

Top-down characterization

Intact proteins from three separate aerobically grown ribosome samples were examined by LC-FT-ICR-MS, and the resulting data were pooled. From this top-down analysis, we identified 42 intact *R. palustris* ribosomal proteins. The four largest ribosomal proteins (RRP-S2 at 36 kDa, RRP-S1 at 62.8 kDa, RRP-L2 at 31.6 kDa, and RRP-S3 at 26.3 kDa) were not observed. Even though the FT-ICR-MS has sufficient mass range to observe these species, prior experience with intact proteins suggests that larger species, such as these, are difficult to elute from the C4 reverse-phase column under the experimental conditions employed for the top-down liquid chromatography. It is likely that the increased hydrophobicity of these larger proteins results in irreversible binding on the reverse-phase column, making these proteins difficult, if not impossible, to elute from the column.

Figure 3.2 presents an example of data from the top-down approach. Figure 3.2A shows a total ion chromatogram of the purified ribosome sample from the reverse-phase separation, and Figure 3.2B is the deconvoluted mass spectrum corresponding to the chromatographic peak at 1152 seconds. At least ten different molecular species were observed in this spectrum, with molecular masses ranging from 7-11 kDa. For each observed species, the most abundant isotopic mass (MAIM) peak was used to query the entire *R. palustris* protein database for tentative protein identifications. This search was conducted by examining all intact and N-terminal methionine truncated proteins for possible matches. Note that this search did not consider all possible post-translational modifications of all the possible proteins, as the number of such possibilities would preclude searching in a meaningful fashion. Although more definitive information could be obtained by conducting tandem MS on the intact proteins, this is difficult on the

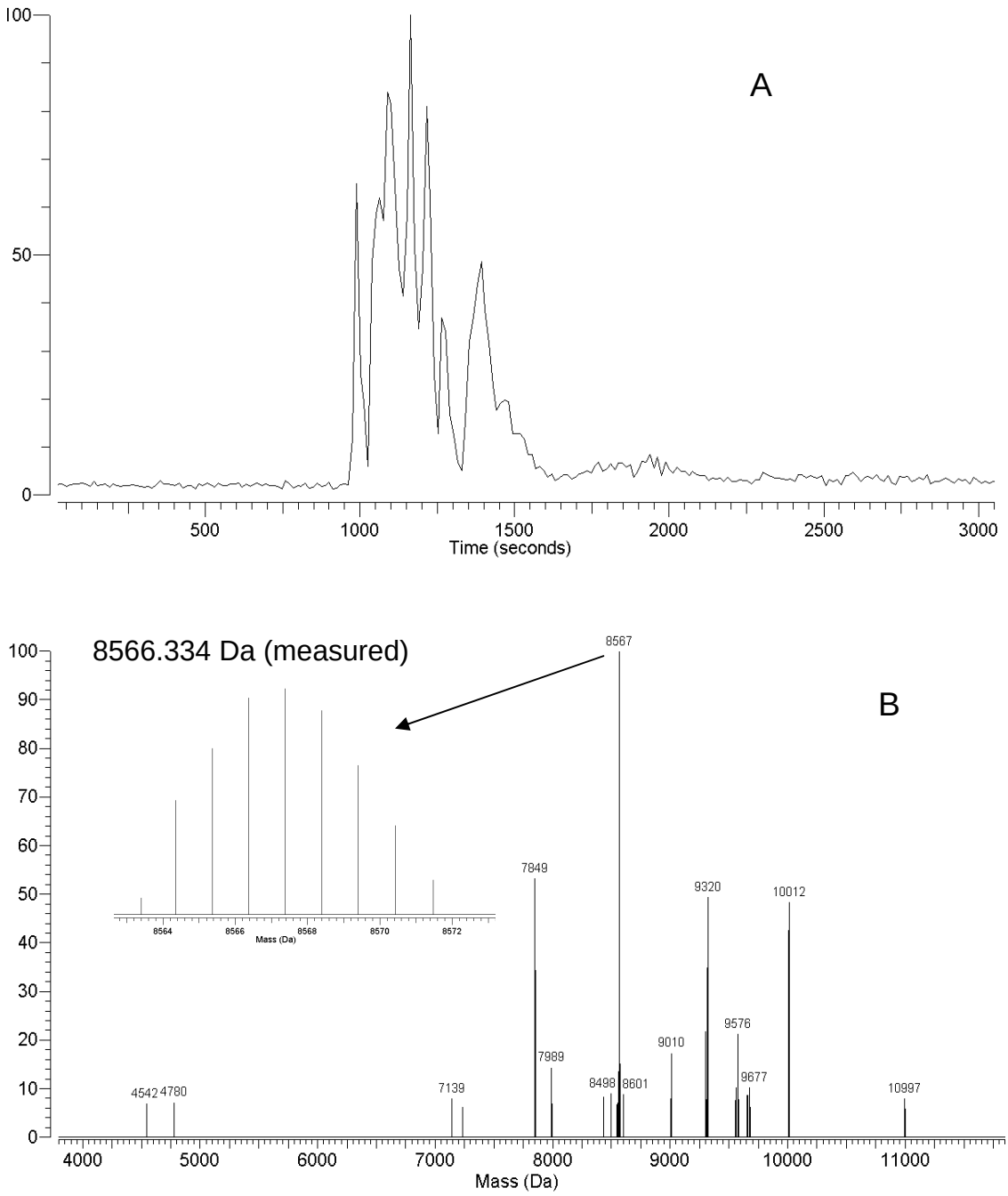


Figure 3.2: LC-ES-FT-ICR measurement of intact masses for top-down analysis. (A) Total ion chromatogram. (B) Deconvoluted mass spectrum corresponding to the chromatographic peak at 1152 seconds. Inset illustrates the isotopic resolution of the component at nominal mass 8,567 Da.

timescale of our chromatography. Our focus here was to compensate by correlating the top-down data with the bottom-up data for improved validation of tentative identifications. This approach, while probably less feasible for entire proteomes as discussed below, is well suited to simpler systems such as the purified ribosome complex. An isotopically-resolved pattern was then calculated from the elemental composition of each tentatively identified intact protein, and compared with the measured isotopic packet for final validation. The inset in Figure 3.2B illustrates the isotope pattern of the component at nominal mass 8,567 Da. The measured isotopic packet of this species was consistent with the calculated isotopic packet of intact RRP-L31; the measured isotopically resolved peak at 8,566.334 Da is within 2 parts per million of the calculated isotopically averaged value for this protein (8,566.315 Da). If this measured protein mass at 8566.334 Da is used to query the entire *R. palustris* proteome, the next closest match is a methionine-truncated hypothetical protein (gene RPA1934), which differs by 3 Da (360 ppm error) from the measured mass. In addition to the large mass error, RPA1934 is a hypothetical protein that was not measured in our bottom-up analysis. The next nearest ribosomal protein match to this measured value would be the methionine-truncated S18, which differs by 396 Da (45,000 ppm error) from the measured mass. This takes into account all possible ribosomal proteins, including intact, methionine truncated, or containing any variation of acetylation and/or methylation, to the extent specified in the experimental section. Thus, within the constraints of our search, only RRP-L31 was found to be consistent with the measured mass of the 8,567 Da species. Likewise, the component in Figure 3.2B at nominal mass 7849 Da had a MAIM of 7849.239 Da. This value is within 3 ppm of the calculated MAIM of 7849.213 Da for the methionine truncated RRP-L29. The RRP-L31 species is only present as the intact gene product, whereas the RRP-L29 is only present in the methionine truncated form. By searching for intact gene products as well as methionine truncated forms, five ribosomal proteins could be identified in this mass spectrum. The remaining 4-5 species observed in this mass spectrum could not be identified. These may represent altered forms of ribosomal proteins that are not readily identifiable, such as other truncation products, or could be due to contaminant proteins isolated with the ribosome.

Table 3.3 is the summary of intact protein identifications by the high-resolution FT-ICR-MS top-down technique. In total, 42 proteins were tentatively identified, with the majority (25) at better than 10 ppm mass accuracy, and only 3 differing by >30 ppm from the calculated value. Of these 42, ten correspond directly to the predicted gene products, 21 are processed by only methionine truncation, and the remaining 11 appear to be modified by further acetylation and/or methylation. Two proteins, RRP-L24 and RRP-S8, were found to be present in two different forms. The most highly modified species identified was RRP-L11, which is methionine-truncated, and contains multiple methylations and/or acetylations. About ten additional species were measured from the ribosome sample, but could not be identified. It is likely that these species correspond to the other ribosomal proteins, but are altered substantially (possibly by combinations of other PTMs, oxidation, and more extensive truncation) such that they are beyond the scope of our simple look-up table or they could be common contaminants identified in the bottom-up analysis as well.

As described above, RRP-L34 and RRP-L36 were either identified poorly or not at all by bottom-up analysis, most likely because of their high basic content. Both proteins should, however, form positive ions quite readily in the ES source and therefore be detected by FT-ICR analysis. While RRP-L36 could be matched to an isotopic packet from the FT-ICR analysis at 5063.952 Da, RRP-L34 was absent. Crystallographic structures of the 70S ribosome from *Thermus thermophilus* indicate that L34 is located at the base of the large subunit surface (Yusupov, 2001). Biochemical isolation of *R. palustris* 70S ribosomes may have resulted in stripping of this protein from the subunit surface.

Post-translational modifications of R. palustris ribosomal proteins

An important goal of this study was to search for PTMs of prokaryotic ribosomal proteins. Assignment of a particular PTM by only one proteomic technique is certainly possible, but PTM assignments can be strengthened by using the integrated approach, especially when results from the two approaches corroborate one another. The combined approach often allowed the identification of the modification positions and helped

Table 3.3: Ribosomal protein identification by top-down ESI-FT-ICR-MS.

Protein	Modification	Calc. Mass[a]	Meas. Mass[a]	Mass error (ppm)
L1	loss of Met	23877.832	23877.449	*16.0*
L3	plus Methyl	25622.463	25622.159	*11.9*
L5	plus 2 Methyl	21064.992	21064.576	*19.7*
L6	loss of Met	19272.408	19272.674	*-13.8*
L7/L12	loss of Met + 3 Methyl	12754.070	12754.089	*-1.5*
L9	none	21178.022	21178.268	*-11.6*
L10	loss of Met	19067.739	19067.617	*6.4*
L11	loss of Met+Acet+ 9 Methyl	15507.107	15507.246	*-9.0*
L14	none	13488.498	13488.645	*-10.9*
L15	none	16836.243	16836.259	*-1.0*
L17	plus 3 Methyl	15716.353	15716.056	*18.9*
L18	loss of Met	12904.93	12905.157	*-17.6*
L19	none	14296.764	14296.899	*-9.4*
L21	loss of Met	13358.081	13358.533	*-33.8*
L22	loss of Met	13826.007	13825.644	*26.2*
L23	none	10907.949	10908.021	*-6.6*
L24	loss of Met	10998.226	10998.231	*-0.5*
L24	loss of Met + Methyl	11012.241	11012.146	*8.6*
L29	loss of Met	7849.213	7849.239	*-3.3*
L30	loss of Met	7092.967	7092.988	*-3.0*
L31	none	8566.315	8566.334	*-2.2*
L32	loss of Met	6860.730	6860.636	*13.7*
L33	loss of Met + Methyl	6248.504	6248.450	*8.6*
L35	loss of Met	7415.278	7415.278	*0.0*
L36	none	5063.971	5063.952	*3.8*
S4	loss of Met + Methyl	23441.536	23441.690	*-6.6*
S5	loss of Met	20522.086	20522.411	*-15.8*
S7	loss of Met	17556.270	17556.629	*-20.4*
S8	loss of Met	14477.631	14477.683	*-3.6*
S8	loss of Met+Acet+4 Methyl	14575.704	14575.619	*5.8*
S10	none	11667.363	11667.404	*-3.5*
S11	loss of Met + Methyl	13760.215	13760.314	*-7.2*
S12	none	13874.799	13875.167	*-26.5*
S13	loss of Met	14313.985	14313.596	*27.2*
S14	loss of Met	11331.399	11331.900	*-44.2*
S15	loss of Met	10010.563	10010.562	*0.1*
S16	loss of Met	12017.595	12017.575	*1.7*
S17	loss of Met	9553.253	9553.316	*-6.6*
S18	plus 6 Methyl	9178.219	9177.834	*41.9*
S19	loss of Met	10087.371	10087.379	*-0.8*
S20	loss of Met	9577.324	9577.387	*-6.6*
S21	none	10062.669	10062.722	*-5.3*

a) MAIM (most abundant isotopic mass)

identify the presence of isoforms. For both analyses, we included in PTM searches the N-terminal modifications of methionine truncation, methylation, acetylation and β-methylthiolation. In addition, the search included β-methylthiolation of aspartic acids, single acetylations and mono-, di- and trimethylated lysines or arginines, all of which have been previously identified in ribosomal proteins from *E. coli* and eukaryotic-cell organellar ribosomes thought to have evolved from bacteria by endosymbiosis (Kowalak, 1996; Arnold, 1999; Yamaguchi, 2000). Phosphorylation, a common PTM in eukaryotic ribosomal proteins, has not been identified in prokaryotic ribosomal proteins, and was not included in the subset of modification searches.

N-terminal methionine truncations

The most common PTM identified by the integrated approach was truncation of the start methionine. We identified this modification in 32 *R. palustris* ribosomal proteins. The top-down technique identified an N-terminal truncation if the measured intact mass for a protein matched that obtained by subtracting the mass contributed by a methionine residue (131.0405 Da) from the mass calculated from the DNA-derived amino acid sequence. Twenty seven ribosomal proteins met this criterion. The bottom-up technique validated N-terminal truncations by identifying N-terminal peptides without a methionine; 16 truncated proteins could be identified in this way, while 10 were identified with the N-terminal methionine intact. For the proteins identified to have truncated N-termini, 10 were recognized by both techniques, 17 were identified by intact mass data alone and five were identified only by bottom-up data. Although a majority of these truncations were only identified by one technique, the combination of two MS approaches allowed a larger proportion of the N-termini to be surveyed than would have been possible with either one of these strategies, and provided increased confidence in the assignment for those proteins for which N-terminal methionine truncation was identified by both approaches.

Overall, the N-terminus truncation states of 45 proteins could be determined unambiguously. We were unable to determine whether seven ribosomal proteins contained an N-terminal methionine. For these species, it is possible that a longer

truncation than simple methionine loss could have occurred. Among these, RRP-L16, RRP-L20, RRP-L27, RRP-L34, and RRP-S2 yielded neither bottom-up data for the N-terminal peptide, nor top-down molecular mass information. For RRP-L5 and RRP-S12, on the other hand, conflicting information resulted from the two techniques. For these two proteins, the bottom-up data indicated a methionine truncation while the top-down data did not. While this result was a bit puzzling, one possible explanation for the conflicting data was that these two proteins existed both with and without methionine truncations. Since the relative abundances of this pair were unknown, as were their relative chromatographic behavior and mass spectrometric responses as intact proteins or as N-terminal peptides, there was no expectation that they necessarily would be observed to the same extent in the top-down vs. bottom-up experiments.

Methylation, acetylation, and β-methythiolation PTMs

In contrast to assigning N-terminal methionine truncations, identifying positions of acetylations, methylations, or β-methythiolation, is more complex because these modifications often result in isoforms; furthermore, acetylation and methylation can occur either on residue side chains or N-termini. Table 3.4 summarizes the PTM assignments for ribosomal proteins determined from the integrated approach. We assigned a particular PTM if at least one of the bottom-up data sets agreed with top-down data, or if bottom-up data from both the 1D and 2D separations were consistent. A few examples are given below. For more detailed descriptions, see Strader et al. 2004.

RRP-L3

A MAIM peak in the top-down data at 25,622.159 Da was consistent with singly methylated RRP-L3. Bottom-up analysis by 1D LC-MS-MS identified one peptide on which either K155 or K158 was methylated. In this spectrum, y-series ions containing K155, K158, K160 and K161 (y_{17}, y_{18}) exhibited a 14 Da shift corresponding to methylation. Other y-series ions (y_8, y_{10}, y_{11}, y_{12}) showed no m/z shift relative to an unmethylated peptide; these unshifted y-series ions eliminated K160 and K161 as locations for the methylation. The SEQUEST search identified no unmethylated peptides

Table 3.4: Post-translational modifications of *R. palustris* ribosomal proteins.

Protein	Modification	Residue(s)
RRP-L3	methylation	K155 or K158
RRP-L7/L12[a]	A: 2 methylations and 1 methylation	K69, K86
	B: trimethylation or acetylation	K86, K89
RRP-L11	Acetylation or trimethylation	K40
RRP-L30	methylation	N-terminus or K3[b]
RRP-L33	methylation	N-terminus or K3[b]
RRP-S12[c]	β-methylthiolation	D88

a) Present as two isoforms, A and B.
b) Insufficient data to distinguish between methylation at the N-terminus or at K3.
c) Present in both modified and unmodified forms.

covering this region of L3. While SEQUEST identified another tryptic peptide from L3 on which R170, R177 and R185 were all methylated, we assessed RRP-L3 to be singly methylated at either K155 or K158 because this scenario was supported by both the intact mass and the singly methylated peptide spectrum. L3 has previously been reported to be singly methylated in *E. coli* at Q150 (Arnold, 1999). While the amino acid that was modified differs between RRP-L3 and the *E. coli* homologue, the data suggested that this single methylation was conserved between these two species.

RRP-L7/L12

A molecular mass from the top-down data at 12,754.089 Da indicated that this protein was modified by methionine truncation, plus either multiple methylation (three methyl groups) or acetylation. At low resolution, the latter two modifications are isobaric (i.e., 42 Da) and would not be resolved. However, this relatively small protein was observed in high abundance in the FT-ICR mass spectra and could be measured with high resolution and high mass accuracy. Thus, the measured MAIM of 12,754.089 Da suggested that this protein is trimethylated (calculated MAIM of 12,754.070 Da; 1.5 ppm less than measured value) rather than acetylated (calculated MAIM of 12,754.035; 4.2 ppm less than measured value), as shown in Figure 3.3, although more extensive measurements would be required to definitively make this assignment solely from intact mass data. 1D LC-MS-MS analysis suggested that K69 is dimethylated and K86 is singly methylated, with both multiple overlapping peptides and spectra for different charge states of the same peptide in evidence for K86. Also identified from 1D data were acetylation or trimethylation at K86 and K89. The 2D LC-MS-MS analysis, on the other hand, identified K69 as dimethylated, and K6, K70, K86, and K100 as singly methylated (with K69 evidenced by two spectra representing multiple charge states of the same peptide, and K86 evidenced by multiple overlapping peptides), while other spectra indicated that K86 and K89 could be acetylated or trimethylated (on multiple overlapping peptides). These results are consistent with the existence of two isoforms of this protein; an increased abundance of one isoform over the other may explain why only one form is observed in the top-down data.

ADLQKIVDDLSSLTVLEAAELAKLLEEKWGVSAAAAVAVAAAPGAGGAAAPAEEKTEFTVVLASA
GD**K***KIEVIKEVRAITGLGL**K***EAKDLVEGAPKPLKEGVNKEEAEKVKAQLEKAGAKVELK

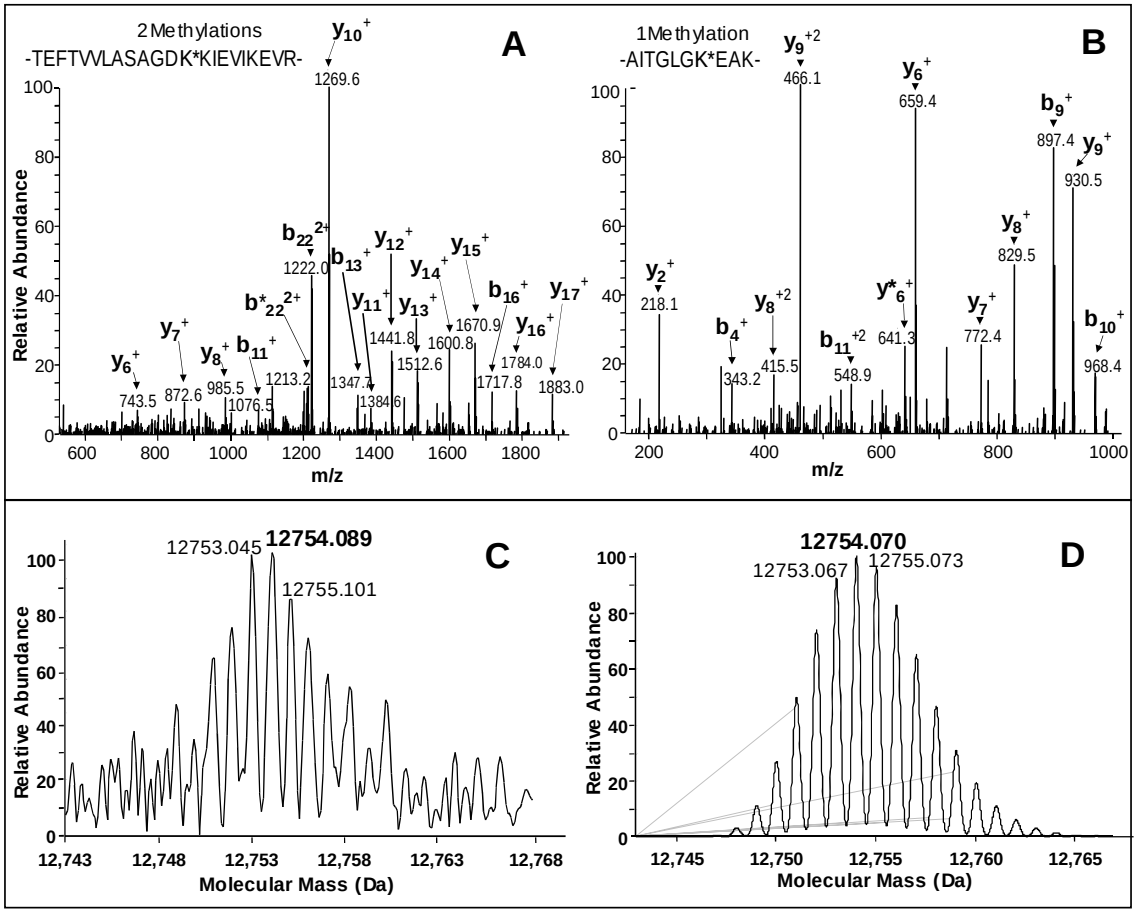


Figure 3.3: A comparison of top-down and bottom-up data for RRP-L7/L12. Fragmentation spectra from (A) peptide T57-R78 bearing 2 methylations at K69. (B) peptide A79-K89 bearing a single methylation at K86. (C) measured, and (D) calculated isotopic distributions for intact trimethylated RRP-L7. The sequence with potential methylation modifications is shown at the top of the figure.

The isoform of RRP-L7/L12 for which an intact mass was observed is best explained by two methylations at K69 and a single methylation at K86. Figure 3.3 shows MS-MS spectra representing peptides with di-methylated K69 and mono-methylated K86 residues (Figure 3.3 A, B) and the measured and calculated isotopic distributions determined for the intact mass (Figure 3.3 C, D). Both MS-MS spectra indicate that y or b ions containing the modified residue are shifted by the appropriate mass. For example, in Figure 3.3A, y_{10}-y_{17} and b_{13}, b_{16} and b_{22}^{2+}, containing modified K69, are shifted by 28 Da (di-methylation), while y_6-y_9, and b_{11}, which do not contain K69, appear at the same m/z values as for an unmodified peptide. For these bottom-up data, the spectra from Figure 3.3A and B give information about the modified residues but not the number of isoforms that exist for RRP-L7/L12 with these modifications. Without the top-down data, it would not be possible to definitively assign these two modified peptides to a single isoform.

The second isoform of RRP-L7/L12 we assigned is acetylated or trimethylated at residues K86 and K89. Evidence for this isoform is found in MS-MS spectra corresponding to peptides with modified K86 and K89 in both 1D and 2D separations. An MS-MS spectrum for this modified peptide shows b ions (b_{22}^{2+}, b_{11}, b_{12}, b_{14}, b_{16}) containing K86 and K89 shifted by 84 Da, indicating two acetylations or six methylations, while the y ions (y_5, y_6, y_9, y_{13}-y_{17}, y_{20}^{2+}) that do not contain K89 or K86 have no m/z shift. The L7/L12 protein in *E. coli* and other bacteria is known to exist in two isoforms: L7 is N-terminally truncated and acetylated, and L12 is N-terminally truncated and methylated at K81 (Arnold, 1999).

RRP-S12

A molecular mass of 13,875.167 Da corresponding to unmodified RRP-S12 was observed by the top-down approach. Bottom-up data from both 1D and 2D analyses, however, indicated the presence of both modified and unmodified RRP-S12. β-methylthiolation was observed in multiple spectra representing the same charge state. In an MS-MS spectrum assigned to peptide V86-R93, y_6, y_7 and b_3 ions containing modified D88 are shifted by 46 Da corresponding to β-methylthiolation, while b_2, y_4, and y_5 ions

not containing D88 are unshifted relative to an unmodified peptide. This novel PTM also occurs at D88 of the *E. coli* S12 ribosomal protein (Kowalak, 1996).

While some of our data suggest that other ribosomal proteins might possess PTMs, those reported in Table 3.4 include only cases for which supporting evidence from two or more different separation, or MS approaches, were found. For example, not included in the robust PTM assignments listed in Table 3.4 are several modified proteins identified from top-down data only. These include L5, L17, L24, S4, S8, S11, and S18. Similarly, although 2D LC-MS-MS provided bottom-up evidence for methylation at K6 and K100 of RRP-L7/L12, lack of evidence for these two modifications in the 1D LC-MS-MS data led to their exclusion from Table 3.4.

Top-down and bottom-up characterization of the crude lysate from *S. oneidensis*

Fractionation of the Intact Proteins by Anion-Exchange FPLC

For the characterization of the S. oneidensis proteome by the top-down and bottom-up approach we applied strong anion exchange fractionation to the intact proteins from the crude lysate. The anion exchange FPLC separation of the *S. oneidensis* crude lysate was judged to be reproducible, based on the identities of the proteins detected from each fraction in successive identical separations (samples run in triplicate - data not shown). SDS-PAGE analysis of the individual fractions revealed that each fraction typically contained a large number of proteins, although some fractions were enriched in a particular protein over another (data not shown). The UV trace of the anion exchange run, along with the number of proteins that were identified per fraction, is shown as Figure 3.4. Analysis of these fractions by 1D LC-MS/MS using two mass range scans identified 395 unique proteins, which is somewhat less than the analyses of the crude lysate by the shotgun approach (see VerBerkmoes, 2002, for discussion on "shotgun" analysis of the *S. oneidensis* proteome). Possible reasons for this observation may be sample losses during separation or the limited number of m/z ranges that were employed in MS analysis. However, data from this type of experiment is useful for correlation with top-down analyses of these same fractions. This method is optimal when targeted

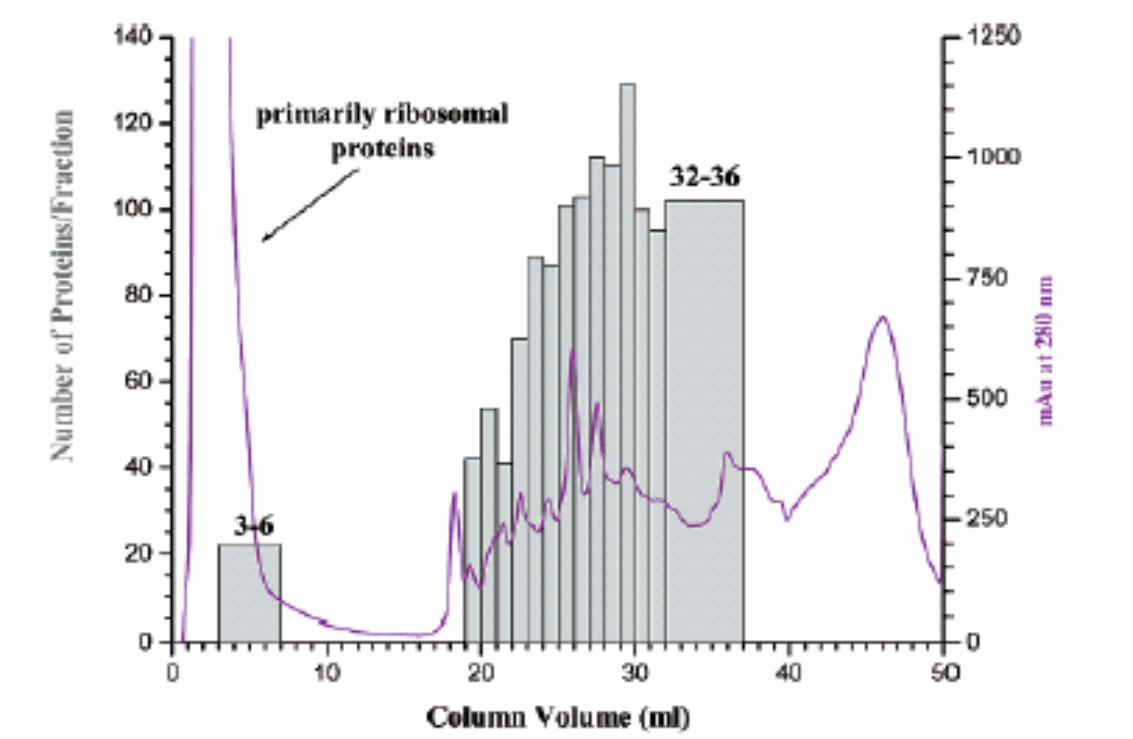


Figure 3.4: Anion exchange fractionation of *S. oneidensis* crude lysate.
The *S. oneidensis* crude lysate intact proteins were fractionated by strong anion exchange. Collected fractions were split in half, with one half used for top-down characterization and the other half for bottom-up characterization. The purple trace indicates the UV absorbance; the grey bars indicate the number of proteins detected by the bottom-up analysis. All fractions were collected 1 min after elution from column and detection by UV. The beginning fractions (10-20) were found to be devoid of proteins as well as the later fractions (38-50 mainly DNA) by previous studies and thus not analyzed in this study.

analysis of a particular protein is desired, since a given protein is isolated in a given fraction or fractions. The alternative of trying to compare large independent "shotgun" analysis with large independent top-down analysis, where fraction collection is not used, would be impractical.

Top-down MS Proteomic Analysis of S. oneidensis

The following protocol was employed to identify proteins by the top-down MS approach. ES-FT-ICR-MS was used to measure the masses of the proteins for all of the fractions examined by the bottom-up method. Although there was some variation between fractions, in general, between 5 and 20 proteins were observed in each fraction. Because the mass resolution was between 50,000 and 100,000, the molecular masses of these proteins could be measured with isotopic resolution (i.e., at the milli-Dalton level). The measured mass of the most abundant isotope of each molecular region was used to query a tabulated protein database for *S. oneidensis*. For several of the proteins, this search resulted in a tentative identification based on a molecular mass match with a protein in the database. Both intact proteins and proteins with N-terminal methionine truncation (the most common PTM for bacteria) were searched in this matter. To examine the protein match, the isotopically-resolved molecular mass region of the suspected protein was calculated based on its sequence (using IsoPro or similar isotopic modeling tool) and compared to that measured in the high resolution FT-ICR-MS experiment. Because these experiments were conducted with external calibration, the molecular masses of the matched proteins from the database were required to be within 15 ppm of the measured values for the more abundant proteins, with the minor proteins matched with somewhat lower mass accuracy (less than 25 ppm). For any tentative matches, the bottom-up MS data from the same fraction was examined to verify the identification of the same protein in that experiment. Although this process was conducted manually, efforts are underway to automate the correlation searching between the top-down and bottom-up MS data. For proteins that did not directly match the database query of intact protein molecular masses, the bottom-up data was examined for abundant candidate proteins that were within 3,000 Da of the measured protein.

Common protein modifications (such as acetylation or signal peptide cleavages) of the candidate proteins were evaluated in an attempt to match the measured molecular mass with the suspected proteins. Tandem mass spectrometry experiments were also conducted on many of the intact proteins. While additional structural information was observed, in most cases, the sequence coverage of the observed product ions was not extensive enough to provide a definitive identification. Rather, this method was used to confirm a suspected protein match, or to differentiate between two proteins with very similar molecular masses.

This general experimental approach is demonstrated in Figure 3.5a-b. Figure 3.5a illustrates the ES-FT-ICR-MS of fraction #23 from the FPLC separation. Note the complexity of the electrospray mass spectrum, with different charge states in close proximity. This is where the capability for high resolution mass measurements becomes essential, as it is possible to directly measure the isotopic spacing and thereby determine the charge state for a majority of isotopic envelopes. This feature helps resolve areas such as the m/z 922-926 region, as illustrated in the inset. Figure 3.5b is the deconvoluted molecular mass spectrum derived from Figure 3.5a. The electrospray mass spectrum simplifies into the dozen intact proteins detected in this fraction. Each of the "nominal" molecular masses shown in this figure consists of a well-separated isotopic envelope, as shown in the insert for the 30,213 Da species. For Figure 3.5b, four of the proteins were identified with the database search (as labeled on the figure), and were confirmed in the bottom-up data for this same fraction.

The proteins identified via top-down analyses of the intact anion-exchange fractions of the *S. oneidensis* lysate are shown in Table 3.5. All of these proteins were also detected in analyses of these fractions by the bottom-up approach. It is important to note that the accurate molecular mass measurements (even with an uncertainty of up to 25 ppm) of the ES-FT-ICR-MS approach greatly simplifies the database searching of intact proteins. For example, for any the proteins listed in Table 3.5, there were no more than three possible matches within 20 ppm of the input value. This greatly limits the possible matches, which are then verified with the bottom-up data. For about one-half of

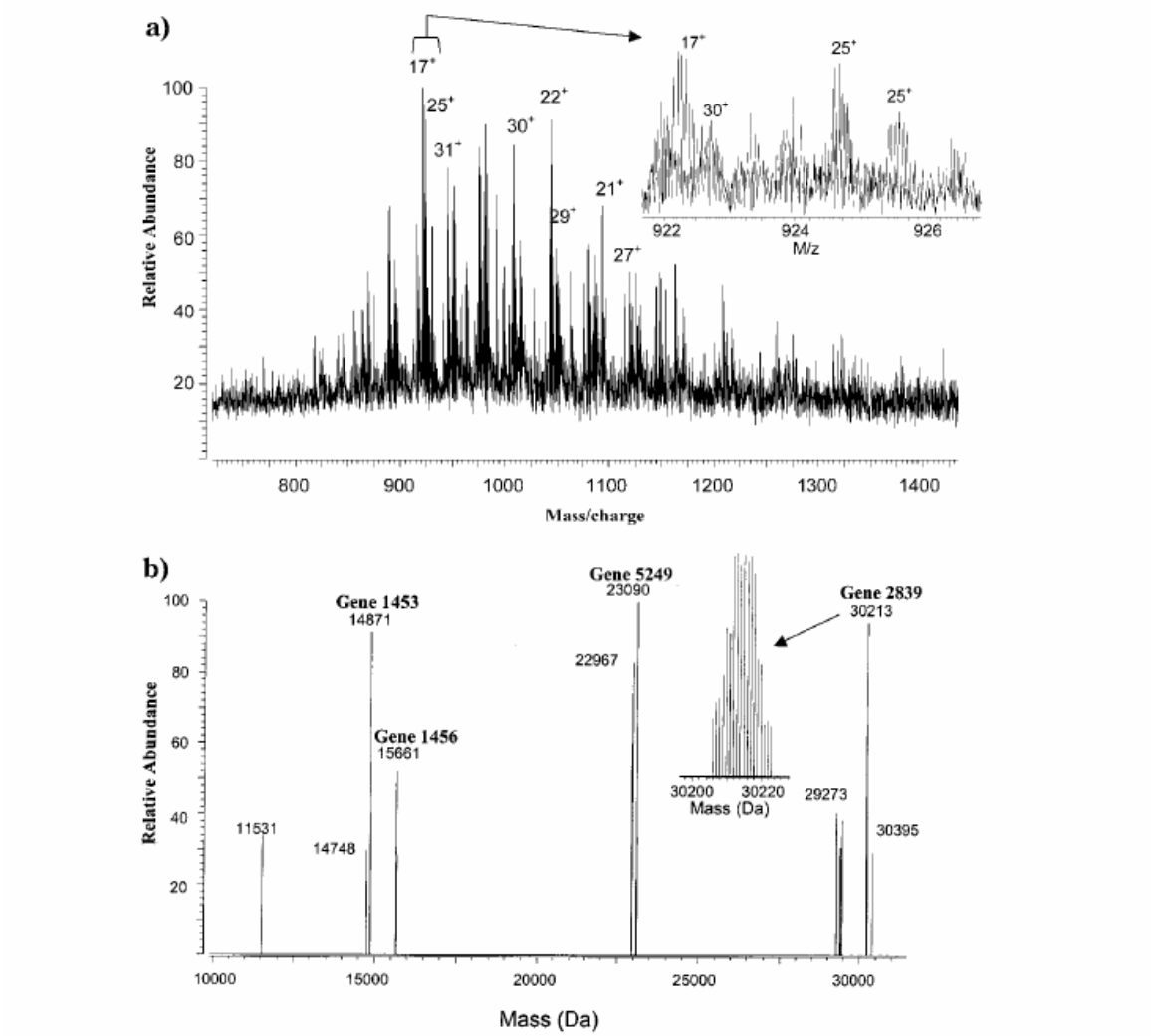


Figure 3.5: Mass spectra of FPLC fraction #23 (intact proteins) from *S. oneidensis*. a) ES-FTICR-MS revealing complex mixture of proteins (close up of m/z 925 region shown as an inset), b) deconvoluted molecular mass spectrum revealing the presence of about one dozen proteins detected in this sample. In each case, an isotopically-resolved cluster was measured (as shown in the inset). High resolution mass measurement was sufficient for identification of at least four proteins in this sample, as designated by the gene number labels.

Table 3.5: Proteins identified from *S. oneidensis* via the top-down approach.

Gene Number	Putative Protein Function	Modification	Calculated Mr (Da)	Measured Mr (Da)	Delta (ppm)
93	Acyl Carrier/protein dehydratase	None	18806.672	18806.196	25.3
516	Fumarate Reductase	Unknown	62446.931	61058.86	-
591	Glutathione synthetase	loss of Met	34830.028	34830.264	6.8
796	Thioredoxin	Start site change/loss of Met (ox)	11755.116	11755.306	16.2
1445	Ribosomal protein L31	None	15742.351	15742.523	10.9
1453	Ribosomal protein S6	loss of Met	14871.335	14871.251	5.6
1456	Ribosomal protein L9	None	15661.524	15661.567	2.7
1507	Conserved hypothetical	loss of Met	18220.595	18220.755	8.2
2697	Nucleoside diphosphate kinase	loss of Met	15350.788	15350.993	13.4
2839	Conserved hypothetical	loss of Met	30212.644	30212.891	8.2
3241	XTP pyrophosphatase	loss of Met	22310.208	22310.737	23.7
3552	Sigma 54 Modulation protein	None	11041.774	11041.83	5.1
3954	Putative periplasmic protein	loss of signal peptide	26444.645	26444.688	1.7
4128	Ribosomal protein L22	None	12070.668	12070.856	15.6
4133	Ribosomal protein L14	None	13455.401	13455.598	14.6
4137	Ribosomal protein S8	None	14038.516	14038.809	20.9
4445	GroES	None	10212.546	10212.655	10.7
4479	Malate dehydrogenase	None	32136.139	32136.534	12.3
4989	Elongation factor TS	Deamidation ($-NH_3$)†	30395.759	30396.062	10.0
5100	DNA Binding protein HU-beta	None	9445.081	9445.13	5.2
5152	Dihydrodipicolinate synthase	None	30943.998	30943.659	11.0
5249	Adenylate kinase	None	23093.001	23093.026	1.1

the proteins listed in Table 3.5, there were no other possible matches within 30 ppm relative to the identified protein. This implies that accurate molecular mass determination, at least for bacterial systems, is a useful tool for protein identification.

This list of identified proteins in Table 3.5 is much shorter than that obtained by the bottom-up approach, due in large part to the matrix ionization effects associated with direct ES-MS of these complex mixtures. The LC-ES-FT-ICR methodologies discussed in the above section on the ribosomal complex were developed after this study. The addition of RP chromatography after the SAX fractionation would have been a very useful additional technique. The combination of SAX and RP separations for intact proteins is currently under investigation. In all probability, most of the proteins in Table 3.5 represent species that were detected in relatively high abundance in a particular fraction, as judged by the amount of sequence coverage observed using the bottom-up approach. High-resolution molecular masses were determined for at least 70 discrete species from the FPLC fractions by FT-ICR-MS, but about 50 of these could not be matched readily to any protein located in our database. There are at least three possible reasons for the difficulty in easily determining the identities of these species: 1) they may represent modified versions of the expressed proteins (signal peptides, other PTM, etc.); 2) the protein database may have errors in the predicted gene start sites; or 3) they may be due to proteolytic breakdown products. To further support the protein identification procedure, MS/MS experiments were conducted in several cases to obtain fragment ions that could be used to verify (or in some cases determine) suspected protein identities. As the rules for protein fragmentation and the methods for fragmentation become better developed, this additional piece of structural information will aid greatly in identifying "unknown" species.

As mentioned previously, one major advantage of using the top-down approach is that it provides information on the intact molecular mass of the protein, which may be useful in the detection of post-translational modifications or N-terminal processing. As an example, the major component of anion exchange fraction #19 was found by the bottom-up approach to be a putative periplasmic protein from an ABC transport system. For this particular protein, the bottom-up MS technique was successful at identifying

tryptic peptides corresponding to 85% sequence coverage of the protein. However, when a portion of this FPLC fraction containing the intact proteins was analyzed via ES-FT-ICR-MS, no mass could be found that matched to the *in silico* average molecular mass of this predicted protein at 29,366.7 Da. Rather, the major component of the fraction was found to have a most abundant isotopic molecular mass of 26,444.688 Da. Inspection of the proteins in this FPLC fraction identified by the bottom-up method indicated that this putative periplasmic protein was the only species within 1,800 Da in molecular mass to this measured value. Since periplasmic proteins typically are known to have signal sequences, the tryptic peptides identified for this protein were examined in more detail. Even though 85% of the sequence was identified, there were no peptides from the first 33 amino acids of the N-terminal end, implying, but not proving, the removal of a signal peptide. Amino acid residues were successively removed from the N-terminus of the predicted peptide sequence in an attempt to match the predicted and measured masses. When the first 28 amino acid residues (MLNVKSHMKSLLGLVVAASMLTVLPAQS) were eliminated, the resulting protein was found to have a theoretical molecular mass of 26,444.645 Da (this is the most abundant isotope). This is illustrated in Figure 3.6a-b, in which the measured isotopic molecular region (Figure 3.6a) is compared to the calculated isotopic molecular region of the truncated protein (Figure 3.6b), based on removal of the 28-amino acid signal peptide from the protein discussed above. Note that the expected most abundant peaks differ by only 0.043 Da (1.6 ppm). Additionally, we performed MS/MS analysis on the intact protein using the FT-ICR-MS/MS. An abundant y_{106} ion (charge retained on C-terminus fragment) and a minor b_{140} ion (charge retained on N-terminus fragment) completely supported the proposed truncated version of the putative periplasmic protein. Analysis of this putative sequence by the "SignalP" signal sequence prediction tool (Nielsen, 1997) predicted that the initial 28 residues corresponded to the signal peptide for this particular sequence, in exact agreement with the MS data. This information, summarized in Figure 3.7, clearly illustrates the utility of using data from both bottom-up and top-down MS approaches.

The top-down analyses were also useful in confirming the preliminary protein annotation of the *S. oneidensis* genome. For example, thioredoxin, a major component of

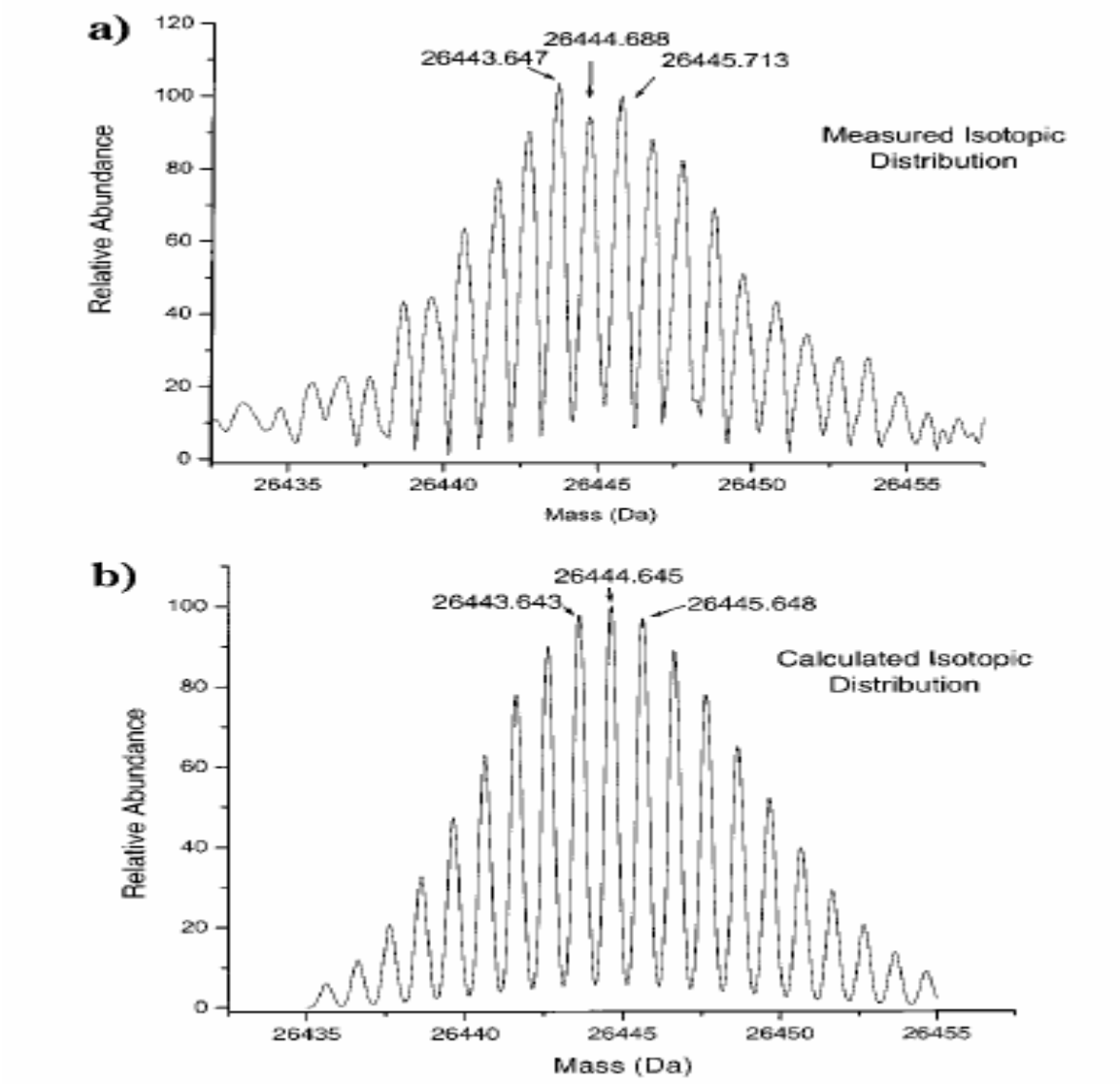


Figure 3.6: ES-FT-ICR spectra of putative periplasmic protein.
Comparison of the measured molecular isotopic distribution of the measured ~26,444.7 Da component (a) with the calculated molecular isotopic distribution of the truncated form of gene 3954 (b). The three most abundant isotopic peaks are labeled in each case. The difference in the expected most abundant isotopic peak between the calculated and measured values is 0.043 Da (1.6 ppm).

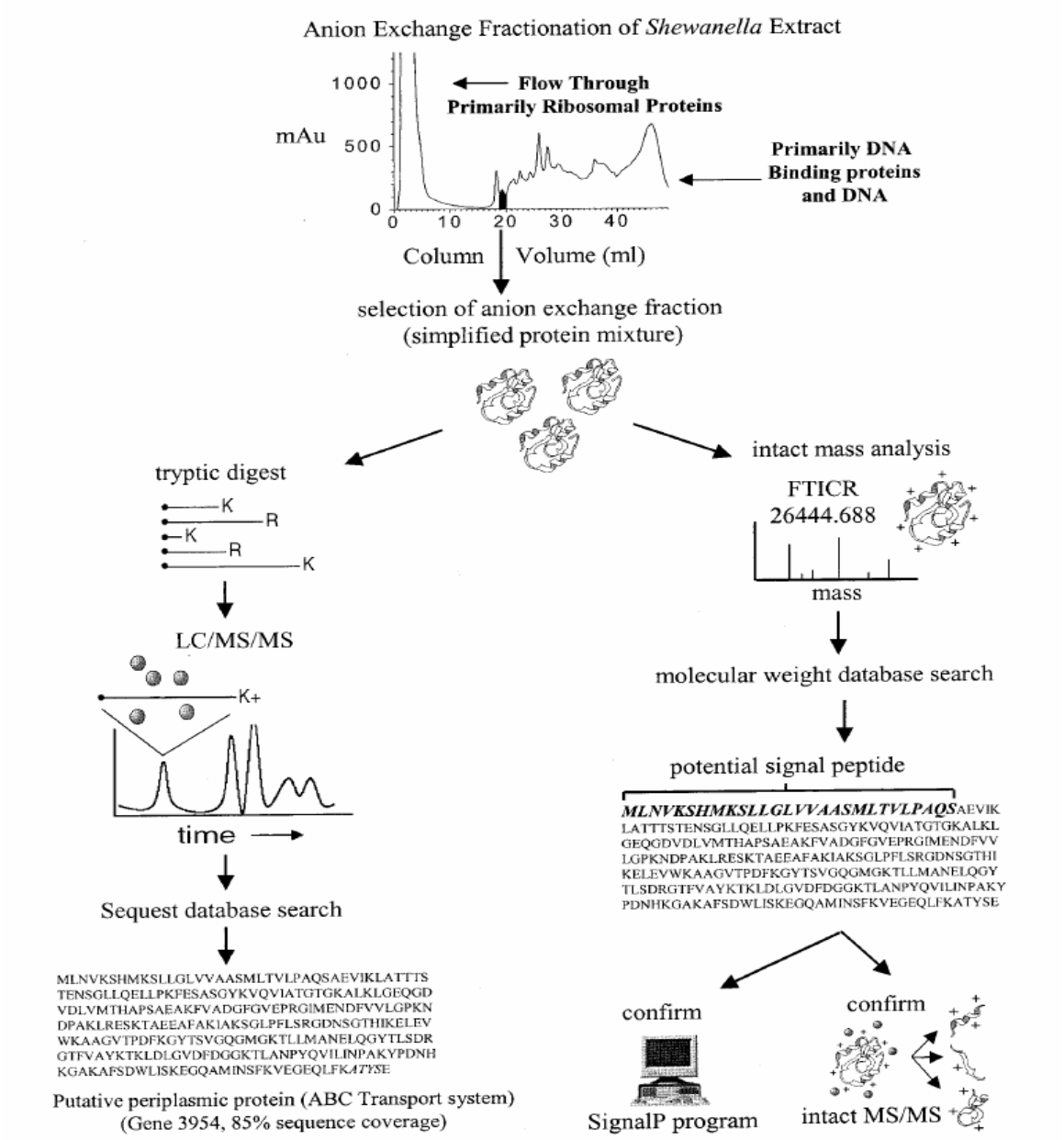


Figure 3.7: Combination of the bottom-up and top-down proteomic analysis. Extracts are first fractionated via anion exchange and are then split for proteolytic digestion and molecular weight determination. Analysis of the proteolytic digestion products is performed using 1D or 2D LC-MS/MS techniques followed by database searching. The undigested fractions undergo a dialysis step followed by analysis using ES-FT-ICR. By simplifying the fractions via anion exchange and using a 9 tesla magnet, increased dynamic range is obtained for intact mass analysis. In this example, a putative periplasmic protein (ABC transporter system) is identified using the bottom-up approach and its subsequent signal peptide confirmed via intact mass analysis, use of the SignalP program, and MS/MS of the intact protein.

anion exchange fraction #22, as determined by bottom-up analysis, was not observed at the predicted molecular mass 12789.9 Da in analysis of this fraction by FT-ICR-MS. A major species in this fraction occurring at 11755.306 Da was found to match the predicted molecular mass of oxidized thioredoxin within 0.190 Da when the start site for translation was shifted to the second Met of the theoretical protein sequence (in the final protein mass, this second Met residue was also eliminated). Clearly, when theoretical protein sequences based on genomic data are used for top-down analyses, the fidelity of the translational start site is a major issue, and future bioinformatics tools to analyze this type of data will need to take this factor into account.

The largest intact protein for *S. oneidensis* observed by the top-down MS method was measured at 61,058.86 Da. The most abundant protein (and the only one within 60 $\pm$ 10 kDa) in this particular FPLC fraction, as determined by the bottom-up method, was fumarate reductase, predicted to have an expressed molecular mass of 62,448.1 Da. Since the bottom-up MS technique revealed no other proteins close in mass to this value, it is suspected that the measured species corresponds to a truncated version of the fumarate reductase subunit. The presence of multiple heme groups on this protein complicates the assignment of an accurate mass value. Although the exact identity of this species has not yet been determined, this is one of the largest intact proteins determined to date in a bacterial proteomics approach by FT-ICR based methods. Furthermore, this is one of the key proteins thought to be involved in the metal ion reduction process.

Conclusion

We have developed and demonstrated an integrated top-down and bottom-up technology for the analysis of single proteins (VerBerkmoes, 2002), protein complexes (Strader, 2004), and whole proteomes (VerBerkmoes, 2002). This integrated technology has advantages that neither technology currently has on its own. Namely, the bottom-up technique is very good at identifying the majority of proteins in a sample, but not providing detailed information on the intact state of those proteins. The top-down technique is currently limited in the identification process, but can provide information on the intact state of the protein. We have demonstrated that with the combination of the

two techniques, the data from one set can help guide data mining from the other, and vice versa.

With the current state of technology, it is obvious that the integrated approach works well for simple mixtures such as purified proteins and protein complexes. The detailed integrated characterization of the 70S ribosome as a model system is one of the most complex protein machines in bacteria as far as shear number of proteins and potential modifications. The clear forward path for this methodology is to test its performance with a large number of protein complexes from microbes, such as proteins, in the oxidative phosphorylation chain, ATP synthase, the photosynthetic reaction center, the nitrogenase complex, and the RubisCO complex, to name just a few. To fully validate the usefulness of the methodology for characterizing protein complexes, it will be necessary to test the methodology on a large number of known and unknown protein complexes. The current enrichment of a large number of protein complexes through the ORNL Center for Molecular and Cellular Systems: A Research Program for Identification and Characterization of Protein Complexes (http://doegenomestolife.org/research/ornl.shtml) provides excellent starting materials for these types of studies. While great promise was shown with the 70S ribosome, this complex is very easy to purify and many of the proteins are ideal for MS analysis through either the top-down or bottom-up approach. Some of the protein complexes listed above, especially ones with a number of membrane embedded proteins, would be far more challenging.

The application of the top-down and bottom-up approach to the characterization of whole proteomes is a much greater challenge. It should be noted that our first experimental efforts on whole proteomes came when we were just developing the technology, but clearly, the results indicated much developmental work is needed. While it was possible to make discoveries, such as the N-terminal processing point of the periplasmic protein, the depth of information was limited. This is a major challenge for the integrated approach. The first major challenge is the effective separations of intact proteins in at least two dimensions. The first dimension separation must accommodate both soluble and insoluble proteins and provide high resolution. The strong anion

exchange method we employed is not effective for separating insoluble proteins and was very low resolution. A potential solution to this problem was introduced by Meng et al. 2002. In this study, they introduced the idea of dissolving the proteome in an acid labile surfactant and then separating the entire proteome in the first dimension with continuous elution electrophoresis. The acid labile surfactant is then removed by the addition of high acid and the proteins are separated in a second dimension by reverse-phase. As described in the characterization of the 70S ribosome, we have successfully integrated reverse-phase on-line separations with ES-FT-ICR, overcoming this major limitation.

The second major limitation for the integrated approach is the inability of FT-ICR instruments to apply data-dependent isolations and tandem mass spectrometry to intact proteins on liquid chromatography time scales. If this limitation could be overcome, more detailed structural information could be obtained from the top-down analysis, allowing for more confident protein identifications and PTM analysis, especially for the many proteins detected by the top-down technique but not confidently assigned. One potential instrumental advancement that could potentially overcome this problem is the recently developed integrated linear ion trap FT-ICR (Syka, 2004). This instrument has been shown to provide rapid data-dependent MS/MS of peptides with high resolution FT-ICR measurements, but has not been rigorously tested for the analysis of intact proteins.

The final major limitation for the integrated approach is the necessary proteome informatic tools for the analysis of top-down data as well as the integration of the top-down and bottom-up datasets. While many laboratories are currently working on improving software for bottom-up analysis, very few are working on software for top-down MS data and even less for integrating the two datasets. While this is currently an active area of research in our laboratories, much work is still needed to develop robust and user-friendly algorithms.

If all of the above limitations can be addressed in the next few years, the impact of this integrated top-down/bottom-up technology on the MS-based proteomics field could be immense. This type of technology is absolutely necessary as proteomic applications move beyond simply attempting to identify proteins to completely characterizing their intact states, including all post-translational modifications, correct

start and stop sites, identifying amino acid substitutions, and identifying N-terminal processing. While not important in prokaryotic systems, these techniques may also be applied to the very complex problem of accurately determining splice site variants in higher order eukaryotes.

Chapter 4

Shotgun Proteomics for the Characterization of the *Shewanella oneidensis* Fur Regulon

Some of the data presented below has been published as Xiu-Feng Wan, Nathan C. VerBerkmoes, Lee Ann McCue, Dawn Stanek, Heather Connelly, Loren J. Hauser, Liyou Wu, Xueduan Liu, Tingfen Yan, Adam Leaphart, Robert L. Hettich, Jizhong Zhou, and Dorothea K. Thompson. Defining the *Shewanella oneidensis* FUR Regulon: Integration of Genome-Wide Expression Analysis, Proteome Characterization, and Regulatory Motif Discovery. *Journal of Bacteriology* (2004), 186, 8385-8400. *All MS sample preparation, experiments and data analysis were performed by Nathan C. VerBerkmoes.*

*Complete datasets for the microarray and proteomic analyses and other supplementary material are available on the web site http://digbio.missouri.edu/~wanx/fur/fur.html

Introduction

Virtually all microbial systems require iron, which participates in many major cellular functions, including respiration, the trichloroacetic acid (TCA) cycle, enzyme catalysis, gene regulation, and DNA biosynthesis (for a review, see Andrews, 2003). Free Fe(II), however, can be detrimental because of its ability to catalyze Fenton reactions and the formation of highly reactive, damaging hydroxyl radicals (Touati, 2000). Consequently, the dynamics of intracellular iron concentrations must be precisely controlled and managed to prevent iron-induced toxicity due to excessive levels of free iron.

A diversity of prokaryotic organisms utilizes Fur (the ferric uptake regulator) to control iron homeostasis at the level of transcription. With the large-scale sequencing and annotation of numerous microbial genomes, it is apparent that Fur is widely distributed throughout the Bacterial domain. Fur homologs have been reported for a variety of bacteria, including *Escherichia coli* (Hantke, 1987), *Vibrio cholerae* (Litwin, 1992), *Vibrio anguillarum* (Tolmasky, 1994), *Neisseria species* (Berish, 1993; Thomas and Sparling, 1994, 1996), *Helicobacter pylori* (Bereswill, 2000), *Shewanella oneidensis* MR-1 (formerly *Shewanella putrefaciens* strain MR-1; Thompson, 2002), to cite just a few.

In *E. coli* and other bacteria, Fur is an iron-responsive, homodimeric

metalloprotein that complexes with Fe(II) to repress the transcription of genes/operons determining siderophore biosynthesis and transport in response to high intracellular Fe(II) concentrations (reviewed in Andrews, 2003). Fur accomplishes the repression of iron-scavenging systems and genes involved in other iron-related functions by binding to a specific sequence element, often referred to as the Fur box, in the target promoters of iron-regulated genes, thus effectively blocking transcription by the RNA polymerase holoenzyme (Bagg and Neilands, 1987; de Lorenzo, 1987). In response to iron limitation, Fur no longer binds to the operator site, and transcription from target promoters resumes (Figure 4.1).

Thompson et al. have described previously the partial transcriptome analysis of a *fur* insertion mutant of the metal ion-reducing bacterium *Shewanella oneidensis* MR-1 using DNA microarrays containing polymerase chain reaction (PCR)-generated amplicons corresponding to 691 predicted genes (Thompson, 2002). Since the publication of this study, sequence determination and closure of the *S. oneidensis* 5-Mbp genome was completed by The Institute for Genomic Research (TIGR) (Heidelberg, 2002), making it feasible to conduct a comprehensive microarray analysis of the dynamics of the MR-1 transcriptome in response to physiological perturbations or genetic mutations. *S. oneidensis*, a facultatively anaerobic γ-proteobacterium, possesses remarkably diverse respiratory capacities that have important implications with regard to the potential for bioremediation of metal contaminants in the environment. In addition to utilizing oxygen as a terminal electron acceptor during aerobic respiration, *S. oneidensis* can anaerobically respire various organic and inorganic substrates [i.e., fumarate, nitrate, chromium, thiosulfate, trimethylamine N-oxide (TMAO), Fe(III), and Mn(III)]. In contrast to other bacteria with well-characterized fur genes that utilize iron for assimilatory metabolism only, *S. oneidensis* uses iron for both the biosynthesis of cellular enzymes and macromolecules (assimilatory processes) and energy production (dissimilatory processes).

The advent of advanced analytical technologies, in particular DNA microarrays and high performance liquid chromatography-tandem mass spectrometry, and sophisticated computational methods have enabled a detailed, global characterization of

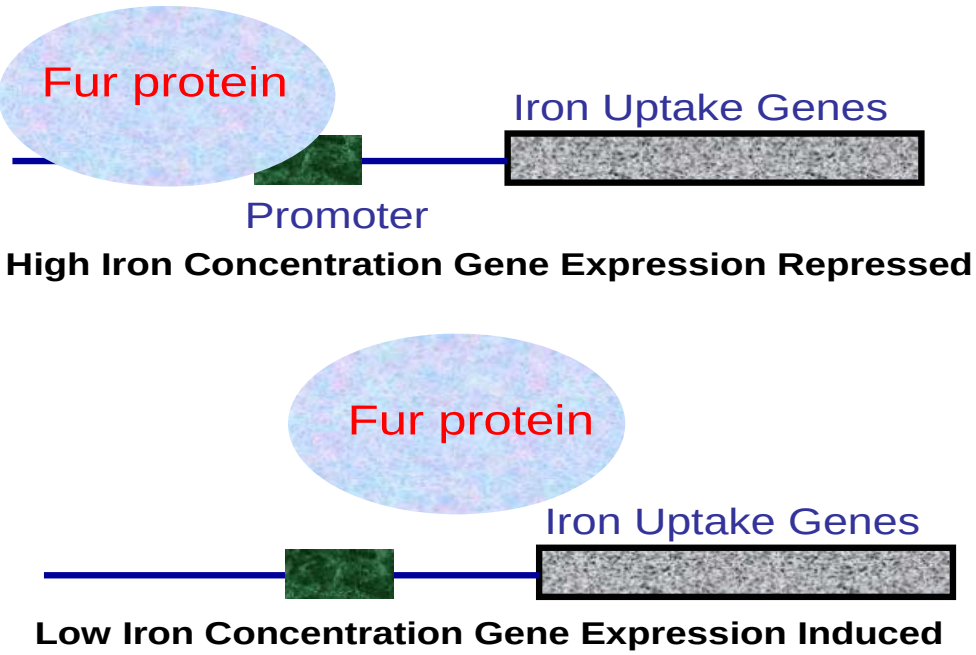


Figure 4.1: General operation of the ferric uptake regulator (fur) protein.
Top panel indicates the situation where iron concentrations are high and not limiting. The fur protein binds to promoters of iron uptake genes and represses their expression. Bottom panel indicates the situation where iron concentrations are low and limiting. The fur protein dissociates from the promoter and iron uptake genes are expressed.

microbial cellular processes that have previously been unattainable. Microarray-based genomic technology is a powerful tool for studying gene functions and regulatory networks at the transcript level. On the proteomic level, mass spectrometry has become an important tool for establishing the identity of proteins via peptide mass mapping or tandem mass spectrometry (see Chapters 1 and 2). More recently, a "gel-less" method of proteome characterization has been developed that combines liquid-based peptide separation (liquid chromatography) with high-resolution molecular mass (MS) measurements. This is one of the most promising approaches to overcome some of the limitations of two-dimensional polyacrylamide gel electrophoresis (2D-PAGE), and has become quite successful for the identification of large-scale microbial proteomes (Washburn, 2001; Lipton, 2002; VerBerkmoes, 2002; Peng, 2003; Corbin, 2003 and Chapters 1 and 2). Previous proteome studies of *S. oneidensis* with alternate electron acceptors such as fumarate, nitrate, or Fe(III) (Beliaev, 2003) and the *fur* mutant (Thompson, 2002) have failed to identify many of the Fe(III) transports and TonB-dependent receptors/transporters that were predicted to be up-regulated under these conditions. Furthermore, microarray analyses confirmed up-regulation of these genes at the transcript level. The previous studies were all carried out with 2D-PAGE-MS, which has known limitations in identifying membrane-bound proteins such as transporters and receptors. Thus, the major goal of the proteomics efforts discussed below is to determine if proteins predicted to be up-regulated by the microarray data could be confirmed through proteome analysis by LC-MS/MS (Figure 4.2).

A quantitative comparison of different microbial growth states is currently a serious challenge for MS-based proteomics efforts. This is mainly due to detection biases, which can arise due to differential protein extraction, matrix effects in the ionization processes, and biases in digestion and sample clean up. While great effort has been put forth into relative quantitation of proteins between different growth states using technologies such as isotope coding affinity tags (ICAT) (Gygi, 1999), metabolic labeling (Oda, 1999; Paša-Tolic, 1999), and ^{18}O water labeling (Yao, 2001) none of these techniques clearly solve the problem of relative quantitation between two samples. Specifically, the ICAT technology requires the labeling of cysteine residues, which are

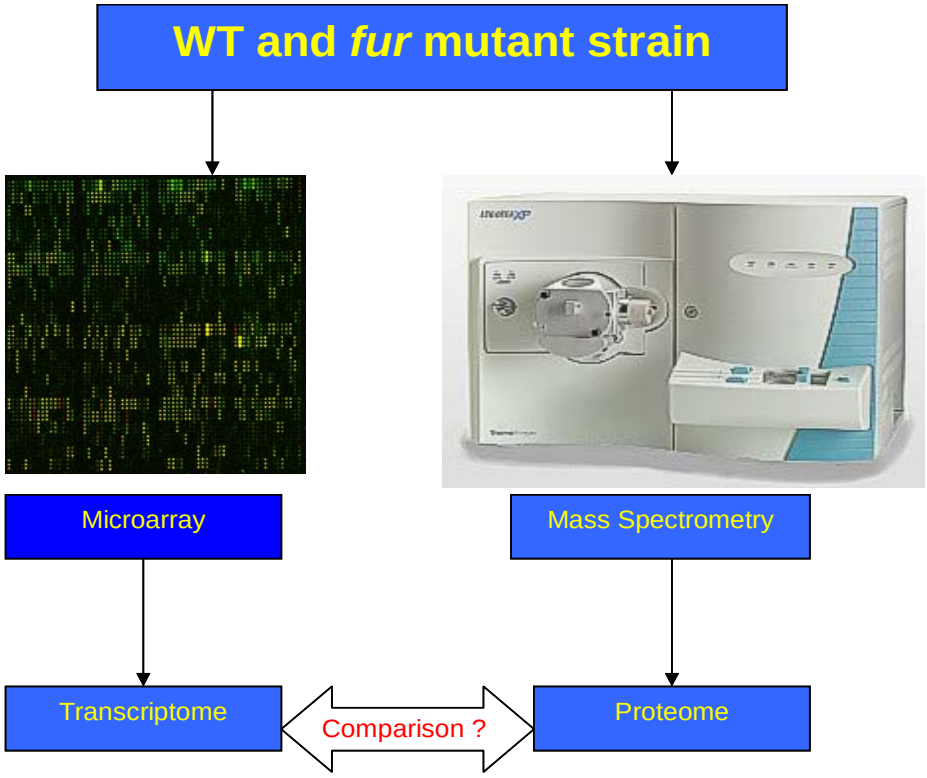


Figure 4.2: Comparison of microarray and proteome data.
Depicts general concept of comparing microarray and MS-based proteomics data. In this study the WT and *fur* mutant strain of *S. oneidensis* were prepared under aerobic conditions and analyzed by the two separate methodologies and compared.

not very prevalent in microbial systems when compared with eukaryotic systems. Indeed, ~60% of the *R. palustris* predicted proteome contains 2 or less cysteines per protein and 20% contains no cysteines at all (Chapter 1). Metabolic labeling with ^{15}N has shown the most promise in microbial systems (Lipton, 2002; Washburn, 2002), but requires strict control of nitrogen intake. Indeed, many microbial species cannot be cultivated under conditions where strict control of nitrogen intake is required. We have not currently been able to grow *S. oneidensis* under strict metabolic conditions allowing for nitrogen labeling. Labeling peptides during trypsin digestion with ^{18}O water is a potential alternative approach (Yao, 2001) but the expense involved in labeling the number of samples used in this study and the need for high-resolution mass spectrometers limits its use for large-scale proteome comparisons (these types of instruments that are capable of high mass resolution measurements as well as data-dependent MS/MS analysis were not available to us for this study). Below we will discuss a potential alternative to exact quantitation for determining proteins showing large-scale differences between the wild-type (WT) and *fur* mutant in aerobically grown *S. oneidensis* samples.

Materials and Methods

Chemicals and reagents

Unless otherwise stated, chemical reagents were obtained from Sigma Chemical Co. (St. Louis, MO). Modified sequencing grade trypsin from Promega (Madison, WI), was used for all protein digestion reactions. The water and acetonitrile used in all sample clean up and HPLC applications was HPLC grade from Burdick & Jackson (Muskegon, MI) and the 98% formic acid used was purchased from EM Science (Darmstadt, Germany).

Microarray analysis

All experimental details of the microarray analysis were conducted by the Thompson group and can be found in Wan et al. 2004. While the microarray work is

referred to in the comparison to the proteome data below, this was not the focus of this section of this dissertation and thus, is not included.

Preparation of whole cell lysates and protein extraction for proteome analysis

For HPLC-MS/MS analysis, *S. oneidensis* parental and FUR2 strains were grown in 1 L cultures (a total of 2 L/strain) under aerobic conditions, pelleted by centrifugation, washed twice in ice-cold 50 mM Tris pH 7.5, and stored at –80°C until analysis. For protein extraction, cell pellets were resuspended in ice-cold 50 mM Tris pH 7.5 and disrupted by sonication. Unbroken cells were pelleted by centrifugation (5,000 g × 15 min) and the suspension of membrane and soluble proteins was aliquoted into 1 ml tubes and frozen at –80°C until analyzed. For all LC-MS/MS analyses, aliquots of wild-type and FUR2 samples were quantitated for total protein amount using the BCA protein assay reagent (Pierce Biotechnology, Inc., Rockford, IL). Equal protein quantities of each sample were denatured with 6 M guanidine and 5 mM DTT at 60°C for 1 h and then diluted in 50 mM Tris (pH 7.5)/ 5 mM $CaCl_2$ to obtain a final guanidine concentration of 1 M. Sequencing grade trypsin (Promega, Madison, WI) was added at 1:100, and digestion reactions were run for 16 hours. Trypsin was added a second time at 1:100 and digestion was run for another 6 hours, followed by a final reduction step with 10 mM DTT for 1 h. Samples were immediately desalted with a C_{18} Sep-Pak (Waters, Milford, MA) and concentrated using a centrifugal evaporator (Savant Instruments, Holbrook, NY) to ~10 µg/µl and filtered to remove insoluble material. For equivalent LC-MS/MS analysis (see below), great care was taken to load equal quantities of wild-type and mutant samples onto the LC-MS/MS system.

LC-MS/MS and data analysis

Proteomes of WT and *fur* mutant were analyzed by three different shotgun LC-MS/MS techniques: 1D LC-MS/MS with 5 injections and 5 m/z ranges scanned, 2D LC-MS/MS with 1 injection and 1 m/z range scanned, and 2D LC-MS/MS with 2 injections and 2 m/z ranges scanned (as explained below). Although the use of multiple injections consumes more sample and more instrument time, this technique permits a more detailed

measurement of each portion of the mass spectrum, thereby enhancing the detection capabilities of the mass spectrometer due to relaxed dynamic range considerations (see chapter 2 for details). For example, the 1D LC-MS/MS with 5 injections and 5 m/z ranges scanned provides a significantly larger number of peptides detected than a 1D LC-MS/MS experiment involving 1 injection and 1 m/z range scanned (or in fact 5 injections and 1 m/z range scanned in each run).

One-dimensional LC-MS/MS experiments were performed with an UltiMate HPLC (LC Packings, a division of Dionex, San Francisco, CA) coupled to an LCQ-DECA ion trap mass spectrometer (Thermo Finnigan, San Jose, CA) equipped with an electrospray source. Injections were made with a Famos (LC Packings) autosampler onto a 50 µl loop. Peptides were injected onto a VYDAC 218MS5.325 (Grace-Vydac, Hesperia, CA) C18 column (300 µm i.d. × 25 cm, 300 Å with 5 µm particles) at a flow rate of 4 µl/min and separated over 240 minutes from 95% H_2O/ 5% ACN/ 0.5% formic acid to 30% H_2O/ 70% ACN/ 0.5% formic acid. Peptides were eluted directly into an electrospray source (Thermo Finnigan) with 100 µm i.d. fused silica. For all 1D LC/MS/MS data acquisition, the LCQ was operated in the data-dependent mode, where the top four peaks in every full MS scan were subjected to MS/MS analysis. Dynamic exclusion was enabled with a repeat count of 1 and exclusion duration of 1 minute. Five separate 50 µl injections of each sample were made and five segmented m/z ranges were scanned to increase total proteome coverage.

Two-dimensional LC-MS/MS experiments were performed on a 2D HPLC system (LC Packings) coupled to an LCQ-DECA ion trap MS equipped with a Finnigan nanospray source. Sample and salt (ammonium acetate) injections are made with the Famos autosampler onto a LC Packings SCX column (500 µm i.d. × 15 mm), which sits on Valve A of the Switchos system. Peptides that elute from the SCX column are captured on an LC Packings precolumn (300 µm i.d. × 5 mm, 300 Å PepMap) on valve B. After desalting on the precolumn, the precolumn flow was switched in-line with a nano resolving column VYDAC 218MS5.07515 C18 (75 µm i.d. × 15 cm, 300Å with 5 µm particles) connected directly to a nanospray source (Thermo Finnigan). After the injection and each subsequent salt step, a reverse-phase gradient was run for 160 minutes

to elute peptides into the mass spectrometer (same solvent system as above). The mass spectrometer was operated as described above except a dynamic exclusion repeat count of 2 was employed. For the 1 m/z range experiment, one 50 µl injection of each sample was made with 1 m/z range scanned (400-2000 m/z) and 11 salt steps ranging from 20 mM ammonium acetate to 2 M ammonium acetate. For the 2 m/z range experiment, a 50 µl injection was made followed by 8 salt bumps with the mass spectrometer scanning from 400-1000 m/z. A second 50 µl injection was then made followed by 8 salt bumps with the MS scanning from 990m/z-2000 m/z.

The resultant MS/MS spectra from each LC-MS/MS analysis were searched with SEQUEST (Thermo Finnigan) against all predicted proteins from the *S. oneidensis* TIGR annotation (Heidelberg, 2002) plus predicted proteins from the ORNL annotation. ORNL annotation methods use three different genome modeling programs: Glimmer (Delcher, 1999; Salzberg, 1998), Critica (Badger and Olsen, 1999), and Generation (http://compbio.ornl.gov/generation/). The results from all three algorithms are combined, followed by an automated resolving of overlapping genes, creating a final gene list. The database for proteome analysis was prepared by identifying those protein sequences identified by the ORNL team which were not included in the published TIGR protein translation files and appending these proteins to the TIGR database (M. Land, L. Hauser, and F.W. Larimer, personal communication). The raw output files were filtered and sorted with DTASelect (Tabb, 2002) (DelCN of at least 0.8 and Xcorrs of at least 1.8 [+1], 2.5 [+2], and 3.5 [+3]) and the three WT and FUR2 analyses were compared with Contrast (Tabb, 2002). A list was made of all proteins showing significant change of at least 30% sequence coverage and/or 4 or more unique peptides between the WT and *Fur* sample in all three replicate analyses.

Results

Transcriptome analysis

All experimental results of the microarray analysis were conducted by the Thompson group and can be found in Wan et al. 2004. While the microarray work is

referred to in the comparison to the proteome data below, this was not the focus of this section of this dissertation and thus is not included.

Proteome analysis of a *fur* deletion mutant

Whole-cell lysates of the *S. oneidensis* wild-type and FUR2 strains were digested with trypsin and analyzed by three similar "shotgun" LC-MS/MS methodologies: 1-dimensional LC-MS/MS with multiple mass range scanning (VerBerkmoes, 2002), 2-dimensional LC-MS/MS with a single m/z range, and 2-dimensional LC-MS/MS with two m/z ranges. The proteomes of the wild-type and FUR2 strains were compared by using qualitative analysis of percent (%) sequence coverage and the number of peptides identified in replicate "shotgun" LC-MS/MS analyses. To determine the level of variation in a single sample, the wild-type *S. oneidensis* proteome was analyzed in four separate experiments using 1D-LC-ES-MS/MS with multiple mass range scanning. It was determined that a change of sequence coverage of 30% and/or 4 or more unique peptides was an appropriate cut-off to determine if a protein had a significant change in concentration between two samples above general experimental variation (N. VerBerkmoes, unpublished data). The resultant MS/MS spectra were then searched with SEQUEST (Eng, 1994) and filtered with DTASelect (Tabb, 2002) using a conservative filtering technique (fully tryptic ends with cross-correlations [Xcorrs] of 1.8 for singly-charged ions [+1], 2.5 for doubly-charged ions [+2], and 3.5 for triply-charged ions [+3]). Conservative filters were selected for unambiguous protein identification; however, it is important to note that some true proteins are not identified in this scheme.

Table 4.1 presents the number of proteins identified from each experiment, the total amount of protein loaded onto the system, and the average % sequence coverage per protein. While the total number of proteins identified per analysis changed dramatically, the average % sequence coverage did not. We found this variation was due mainly to a large number of proteins identified with 1-2 peptides per protein; variations of this type were not considered in the final analysis. The three analyses for the wild-type and mutant were compared with the Contrast software, and the protein table was manually evaluated to determine proteins that showed reproducible changes above the stated cutoffs. The

Table 4.1: Total proteins identified by experiment.

Proteome Method[a]		Total protein loaded	No. of proteins identified[b]	Average sequence coverage (%)
FUR2	1D_5m/z	5.0 mg	490	19.78
	2D_1m/z	0.5 mg	765	17.95
	2D_2m/z	1.0 mg	807	18.39
WT	1D_5m/z	5.0 mg	555	21.90
	2D_1m/z	0.5 mg	611	18.13
	2D_2m/z	1.0 mg	673	18.20
Total		N/A	1104[c]	19.06[d]

[a]Three different methods were used to analyze the proteomes of the *fur* deletion mutant (FUR2) and wild-type *S. oneidensis* strains: (i) 5 m/z refers to a five-part 1D-LC-MS/MS experiment, which involved 5 injections with 4 segmented m/z ranges and 1 full m/z range; (ii) 1 m/z refers to a single 2D-LC-MS/MS experiment that involved one injection and 11 subsequent salt steps analyzed by MS; and (iii) 2 m/z refers to two 2D-LC-MS/MS experiments, which included two injections each with 8 subsequent salt steps analyzed by MS over two m/z ranges.
[b]At least 1 peptide per protein was detected with an Xcorrs of at least 1.8 (+1), 2.5 (+2), and 3.5 (+3).
[c]Sum of the number of non-redundant proteins identified for both WT and FUR2 samples.
[d]Average of sequence coverage per protein detected.

entire proteome dataset is provided in Supplementary Table S2 (http://digbio.missouri.edu/~wanx/fur/fur.html). We then compiled the list of proteins and compared it with the transcriptome data (Table 4.2). In general, the results showed that the microarray expression data correlated well with the proteomics data for genes showing large-scale differences at the level of transcription.

Correlation between transcriptome and proteome data

A total of 30 proteins from the 1,104 proteins comprising the proteome dataset passed the base criteria of a reproducible change of at least 30% sequence coverage and/or 4 unique peptides. Of those 30 proteins, the expression patterns for 13 protein species correlated very well with the gene expression data, while 12 proteins determined to have large-scale changes in abundance by LC-MS/MS analysis did not show any significant change at the mRNA level, and 2 proteins showed an inverse correlation between microarray data and proteome data. Two proteins (encoded by SO2304 and SO4422) identified by proteomic analysis as showing significant changes in abundance were originally annotated as pseudogenes and thus not included on the microarray slides.

Genes showing strong correlation between microarray and LC-MS/MS analysis

Of the 13 proteins showing good correlation between proteome data and microarray data, 11 were up-regulated and 2 were down-regulated in the *fur* deletion mutant relative to the parental strain. Of the 11 up-regulated proteins, 7 proteins (SO1482, SO1580, SO3669, SO3914, SO4516, SO4523, and SO4743) were annotated as transporters or receptors involved in siderophore-mediated iron uptake. This is significant because none of these proteins were identified in the previous study employing 2D-PAGE analysis (Thompson, 2002) and points to the advantage of LC-MS/MS techniques for detecting transport/binding proteins. Furthermore, this subset of proteins showed the largest fold changes in expression, with many being represented by 10 or more unique peptides in FUR2 and either 0 or 1 peptide in the wild-type. The genes encoding these 7 proteins also displayed large-scale differences (>5-fold induction) in expression, with the exception of two genes (SO1580 and SO3914), which

Table 4.2: Comparison of transcriptome and proteomics data.

ORF	Gene Product	Proteomics		Microarray Mean
		Fur	WT	
		Peptides	Peptides	(FUR2/WT)
UP-Regulated in *Fur*				
SO0314	ornithine decarboxylase, inducible	**20**	**8**	0.41
SO0401	alcohol dehydrogenase, zinc-containing	**14**	**1**	N/A
SO0442	phosphoribosylaminoimidazolecarboxamide formyltransferase/IMP cyclohydrolase	**9**	**1**	0.92
SO0453	peptidyl-prolyl cis-trans isomerase FkbP	**4**	**0**	1.17
SO0520	heavy metal efflux pump, CzcA family	**6**	**1**	0.77
SO0798	conserved hypothetical protein	**9**	**0**	26.7
SO0958	alkyl hydroperoxide reductase, C subunit	**10**	**1**	0.5
SO1190	conserved hypothetical protein	**13**	**3**	15.47
SO1482	TonB-dependent receptor, putative	**22**	**0**	26.81
SO1580	TonB-dependent heme receptor	**12**	**1**	2.48
SO2001	5'-nucleotidase (*ushA*)	**8**	**2**	1.11
SO3667	conserved hypothetical protein	**10**	**1**	59.66
SO3668	conserved hypothetical protein	**3**	**0**	37.5
SO3669	heme transport protein (*hugA*)	**30**	**7**	25.45
SO3914	TonB-dependent receptor, putative	**16**	**4**	2.83
SO4133	uridine phosphorylase (*udp*)	**6**	**3**	0.75
SO4516	ferric vibriobactin receptor (*viuA*)	**8**	**0**	8.21
SO4523	iron-regulated outer membrane virulence protein	**34**	**4**	5.61
SO4743	TonB-dependent receptor, putative	**19**	**1**	26.2
SO2304	alanine dehydrogenase	**9**	**1**	N/A
SO4422	TonB-dependent receptor, iron-siderophore	**11**	**2**	N/A
SO4422	Similar to ferric aerobactin receptor	**8**	**0**	N/A
SO1377	conserved hypothetical protein	**0**	**15**	1.47
Down-Regulated in *Fur*				
SO1778	decaheme cytochrome c (*omcB*)	**0**	**11**	0.19
SO1779	decaheme cytochrome c (*omcA*)	**0**	**6**	0.17
SO2363	cytochrome c oxidase, cbb3-type, subunit II	**2**	**8**	0.94
SO2907	TonB-dependent receptor domain protein	**4**	**28**	1.32
SO3733	hypothetical protein	**0**	**12**	0.94
SO4077	TonB-dependent receptor, putative	**0**	**8**	0.97
SOA0048	prolyl oligopeptidase family protein	**0**	**7**	1.24

exhibited only a 2- to 3-fold increase in mRNA abundance.

Three conserved hypothetical proteins (SO1190, SO3667, and SO3668) were highly up-regulated in the fur deletion mutant and found to have good correlation between the gene and protein expression data. SO3667 and SO3668 appear to be organized in an operon with SO3669 (heme transport protein, hugA), which was also identified as being up-regulated in FUR2 by both microarray and LC-MS/MS analysis. It should be noted that the protein species encoded by gene SO3668 was only detected with 3 peptides in the FUR2 strain compared with 0 in the wild-type and thus did not meet our criteria for acceptance for the proteome data. However, this protein was included in Table 4.2 because of its location in a probable operon with SO3367 and SO3669 and its apparent co-regulated expression (37.5-fold change) as identified by microarray analysis. Sequence analysis of gene product SO3668 revealed that the first half of the protein contains no trypsin cleavage sites, which possibly explains why fewer than 4 unique peptides were detected for this protein by the LC-MS/MS method.

Gene SO0798 was clearly identified as being up-regulated in the fur deletion mutant. A total of 9 peptides were detected in FUR2 compared with no peptides identified in the WT sample. Microarray analysis indicated the same up-regulation with an expression ratio (FUR2/WT) of 26.7. The public annotation (Heidelberg, 2002) for this gene is a conserved hypothetical but an alternative annotation performed at Oak Ridge National Laboratory (ORNL; see discussion below) identified this gene as having sequence similarity to a TonB-dependent receptor. The proteomic data provide preliminary evidence that this gene may indeed be a TonB-dependent receptor related to iron or other heavy metal uptake.

Two proteins (SO1778 and SO1779) showed substantial down-regulated expression in the fur mutant as identified by both DNA microarray and LC-MS/MS analysis. The genes omcA (SO1779) and omcB (SO1778) encode decaheme c-type cytochromes that are putative outer membrane lipoproteins and are likely involved in the complex multicomponent branched electron transport system of *S. oneidensis*. Previous studies suggest a role of OmcA and OmcB in the reduction of Mn(IV) by *S. oneidensis* (Myers and Myers, 2001).

Genes showing significant expression changes by proteomic analysis but not by microarray analysis

Of the 30 proteins identified as changing by proteomic analysis, 12 showed no substantial differences in mRNA levels by microarray hybridization. Six of these proteins were up-regulated in the FUR2 strain and 6 were down-regulated as determined by LC-MS/MS analysis. SO0520, annotated as a heavy metal efflux pump protein, was identified with 6 peptides in the fur mutant sample with only 1 peptide in the wild-type sample, but the sequence coverage in both cases was very low due to the large size of the protein. Both SO2907 and SO4077 encode putative TonB-dependent receptors and showed a significant decrease in protein abundance in FUR2 compared with the WT. For SO2907, 28 peptides were detected in the wild-type sample compared to only 4 in the fur mutant, whereas 8 peptides for SO4077 were detected in the WT compared with 0 in FUR2. In both cases, the microarray data did not correspond to the proteomics data. The other 8 proteins showed significant changes at the protein level and were of unknown function or had putative functions in energy metabolism, protein folding and stabilization, or the biosynthesis of purines, pyrimidines, nucleosides, or nucleotides. For example, SO1377 (conserved hypothetical protein) had 15 peptides detected in the wild-type sample with none detected in the FUR2 analysis. Furthermore, this protein was repeatedly detected in our baseline proteome analysis of *S. oneidensis* in a previous study (VerBerkmoes, 2002). Another hypothetical protein, SO3733, was detected repeatedly in the WT with a total of 12 unique peptides but was never detected in the FUR2 strain. This protein was also repeatedly detected in our previous study of the WT strain (VerBerkmoes, 2002). While the protein and mRNA levels do not seem to correlate for these two genes, their unknown function and the fact that they are not expressed at a detectable level in the protein samples for the fur mutant make them interesting cases for further study.

Genes showing inverse correlation between proteome data and microarray analysis

Of the 30 genes identified by LC-MS/MS as showing significant change in protein abundance, only two had inverse correlation with the microarray data. Proteins

encoded by SO0314 (ornithine decarboxylase) and SO0958 (alkyl hydroperoxide reductase, C subunit) were both found to be significantly up-regulated in the proteome of FUR2. Both, though, were found to be down-regulated in the fur mutant sample by microarray analysis with FUR2/WT ratio values of 0.41 (SO0314) and 0.50 (SO0958).

Verified expression of proteins from pseudogenes

A protein sequence database based on the *S. oneidensis* MR-1 genome was constructed from two sources for our proteomic studies. This database included 4,778 protein sequences from the TIGR annotation (http://www.tigr.org/; Heidelberg, 2002), plus an additional 355 protein sequences provided by the Genome Analysis and System Modeling Group in the Life Sciences Division at ORNL. These 355 additional sequences consisted largely of the protein translations of ORFs annotated as pseudogenes by TIGR; putative or partial translations of pseudogenes are not typically included in protein databases released by genome projects.

Proteome analysis of the *S. oneidensis* FUR2 and WT strains revealed that 9 ORFs (see Supplementary Table S3 online) previously annotated as pseudogenes were expressed and detected with high confidence (i.e., at least two unique peptides identified per protein). However, many more proteins were detected with single-peptide identifications. Of particular interest were the following two genes that were found to be highly expressed in the *fur* mutant but not in the wild-type strain: SO2304, annotated as an alanine dehydrogenase with an authentic point mutation; and SO4422, annotated as a siderophore receptor gene with one or more frameshifts (Table 4.2). Interestingly, the ORNL annotation predicted two ORFs in place of SO4422: a TonB-dependent receptor/iron siderophore receptor and a ferric aerobactin receptor. The protein products for these genes were easily identified in FUR2 (8 or more peptides) but were either detected at very low levels or not at all in the wild-type strain (2 or fewer peptides). Transcript expression data were not generated for these genes because pseudogenes were not included in the fabrication of *S. oneidensis* whole-genome microarrays.

The remaining 7 proteins identified showed no significant difference in sequence coverage or the number of peptides detected between the FUR2 and WT strains. These

genes are still of interest due to the fact that the proteomics data verify their expression despite the apparent frameshifts and/or point mutations identified in the genome sequence that resulted in their annotation as pseudogenes. These 7 proteins are encoded by SO0590, annotated as a phosphatidylserine decarboxylase (ORNL gene 700, identified with a total of 41.2% sequence coverage and 5 unique peptides); SO3130, a glutamate tRNA synthetase catalytic subunit (ORNL genes 1951 and 1952, 16.9% coverage with 2 peptides, and 24% coverage with 2 peptides, respectively); SO2937, a putative ribosomal large subunit pseudouridine synthase (ORNL gene 2085, 9.6% coverage and 2 peptides); SO2756, a probable peroxidase (ORNL gene 2202, 81.4% coverage and 12 peptides); SO3798, a hypothetical protein (ORNL gene 3461, 18.7% coverage and 2 peptides); SO1211, peptide chain release factor 3 (ORNL gene 4738, 12% coverage and 2 peptides); and SO1900, a putative acyl-CoA synthetase (ORNL gene 5161, 11.8% coverage and 3 peptides). In light of these findings, we examined these genes in more detail. Of primary interest is SO4422, initially identified as a pseudogene by TIGR, and as two proteins in the ORNL annotation due to a frameshift in the middle of the gene. Proteome analysis clearly revealed the expression of the intact gene product, since numerous peptides were detected from before and after the frameshift. The question this poses for SO4422, as well as others in this list, is how these genes are being expressed when sequence data indicate that they are likely pseudogenes with premature stops or frameshifts. SO1900 and SO3789 seem to be intact, despite their original annotation as pseudogenes; while for the remaining genes listed above the most likely explanation may be sequencing errors in the final compiled genome DNA sequence or differences between the sequenced strain and individual laboratory strains. These new findings emphasize the value of using proteomic analyses to verify genome annotations and suggest that it may be useful to provide, in some form, potential protein translations of pseudogenes for proteome analyses. While proteome methodologies do not generally give a definitive reason for a difference, they are very good at identifying potential differences between a published genome and the actual genome of the strain the individual laboratory is using.

Conclusions

With the complete sequencing of numerous microbial genomes, the next challenge is to verify the annotated functions and to determine the biological roles of functionally undefined genes by using integrative experimental approaches. In this study, we used targeted mutagenesis, genome-wide expression profiling, and MS-based proteome analysis to characterize the *S. oneidensis* Fur regulon. This present study builds upon and substantially expands a previous investigation of an *S. oneidensis* fur homolog (Thompson, 2002) by providing a comprehensive description of the Fur regulon. In addition, proteomic verification of the recent genome-wide prediction of proteins in *S. oneidensis* was provided for a large number of genes and led to the identification of protein products of previously annotated pseudogenes.

In addition to genomic structural analysis and transcriptomic viewing through array technologies, proteomic analyses constitute an important component of functional genomic studies because they enable the most essential level of gene expression to be visualized. Recently, *S. oneidensis* has been the focus of a number of proteome studies of different magnitudes that employed various technologies (Beliaev, 2002b; Devresse et al, 2001; Giometti, 2003; Mohan, 2003; Thompson, 2002; Vanrobaeys, 2003; VerBerkmoes, 2002). In one of our previous studies, a fur insertional mutant (FUR1) of *S. oneidensis* was compared with the wild-type strain using two-dimensional polyacrylamide gel electrophoresis (2D-PAGE), followed by micro-liquid chromatography-electrospray ionization tandem mass spectrometry (micro-LC-ESI-MS/MS) (Thompson, 2002). While this analysis revealed 11 major protein species exhibiting significant changes in abundance between the wild-type and fur mutant, only two of these proteins were from the expected class of transport/binding proteins, and many of the proteins showing large-scale differences in expression at the mRNA level by DNA microarray analysis were not identified. This may be due to the fact that typical 2D-PAGE methods have difficulty in capturing small proteins, proteins with widely ranging isoelectric points, and a large proportion of membrane-associated proteins. To extend the protein identification to a more comprehensive level, we analyzed whole proteomes of aerobically grown wild-type and fur deletion strains by gel-less qualitative "shotgun" MS proteomics. This method is

very useful in rapidly determining large-scale protein differences between two samples but cannot identify exact fold changes and is insufficient at detecting small changes in protein abundance. Variations of this qualitative methodology to determine changes in protein expression without isotopic labels have recently been proposed (Chelius, 2002; Gao, 2003). While these newer methodologies may show the potential of giving more exact quantification, this was not the major point of our study. Rather, we were interested in identifying proteins (mainly transporters, receptors and binding proteins) predicted to be up-regulated in the *fur* deletion mutant by microarray analysis but yet to be clearly identified by proteome analysis. Thus, we employed a very simple method to screen for proteins showing large differences in protein amounts between the two samples by percentage sequence coverage and number of peptides found across a triplicate measurement using three similar LC-MS/MS techniques.

While it would have been preferable to analyze the two *S. oneidensis* strains with a more quantitative technique such as isotope-coded affinity tags (ICAT) (Gygi, 1999) or ^{15}N metabolic labeling (Oda, 1999; Pasa-Tolic, 1999) this was not possible due to the limitations of these technologies. For example, ICAT technology involves labeling cysteine residues in proteins with either a heavy or light label, which then allows for accurate quantification of the peptides/proteins by "shotgun" LC-MS/MS. While this methodology is very useful for proteomic analysis of eukaryotic systems, it becomes less global for prokaryotic systems, where the average number of cysteine residues per protein is approximately half that found in eukaryotic systems. In *S. oneidensis*, the average number of cysteine residues per protein is 3.1 compared with 6.2 in the yeast *Saccharomyces cerevisiae*. Furthermore, 19% of the *S. oneidensis* predicted proteome contains 0 cysteine residues, 18% have 1 cysteine residue, and 16% have 2 cysteine residues, making it impossible to obtain multiple (three or more) quantitative tryptic peptide measurements of ~50% of the predicted proteome. While metabolic labeling via ^{15}N can be very effective for quantitation in microbial species since every peptide will be labeled, in this study, the microarray experiments were conducted with cells grown in LB media. To coordinate with the microarray data, the proteome studies were conducted with the same samples, which precluded the use of a defined growth media.

Replicate whole-proteome analysis of aerobically grown WT and FUR2 strains resulted in the identification of a total of 1,104 proteins from both. Proteins showing dramatic changes in abundance (4 or more unique peptides and/or 30% sequence coverage) were highlighted for comparison with the microarray data (Table 4.2). Of those induced genes likely to be members of the Fur regulon based on microarray and motif identification data, proteins encoded by SO0798 (conserved hypothetical protein), SO1190 (conserved hypothetical protein encoded by the SO1188-89-90 operon), SO1482 (putative TonB-dependent receptor), SO1580 (TonB-dependent heme receptor), SO3669 (heme transport protein, HugA, part of the SO3669-68-67 operon), SO4516 (ferric vibriobactin receptor, ViuA), SO4523 (iron-regulated outer membrane virulence protein, IrgA), and SO4743 were identified by LC-MS/MS and showed significantly greater abundance levels in FUR2. Expression of two conserved hypothetical proteins, encoded by genes SO3667 and SO3668, was also up-regulated in the FUR2 strain as demonstrated by LC-MS/MS analysis. Genes SO3667 and SO3668 are arranged together in a probable operon with hugA (SO3669) and their deduced protein products are predicted to be soluble based on the complete absence of transmembrane helices. In this case as well, the *in vivo* abundance levels of these conserved hypothetical proteins were consistent with the observed increases in mRNA expression for the corresponding genes.

Other proteins detected in the large-scale proteome analysis had abundance levels that inversely correlated with the transcriptome data (Table 5). An ornithine decarboxylase (SO0314), a heavy metal efflux pump (SO0520), alkyl hydroperoxidase reductase (SO0958), and uridine phosphorylase (SO4133), for example, showed increased abundance in FUR2, while their corresponding transcript levels were slightly down-regulated as identified by microarray hybridization. While this lack of one-to-one correlation between the proteomics and microarray data may be somewhat surprising, it is important to keep in mind that protein stability, modifications, and/or turnover may alter the protein abundances expected from the microarray data (Pratt, 2002; Eymann, 2002; Corbin, 2003). In addition, it is not clear why the expression of these proteins would be affected by a fur deletion, although differences in their expression may be related more to the cell's attempt to cope with high intracellular iron levels in the absence

of a functional Fur regulator. The inconsistencies between the transcriptome and proteome datasets emphasize the importance of investigating gene expression from the perspectives of both transcription and translation to account for the different levels of regulation possible in prokaryotes. A significant finding to emerge from the proteomic analysis was the identification of protein products encoded by genes originally annotated in the *S. oneidensis* MR-1 genome published by Heidelberg et al. 2002, as having authentic point mutations or frame shifts, (i.e., as pseudogenes). The direct identification of these genes at the protein level confirms their existence.

A noteworthy observation was the significant repression of OmcA (SO1779) and OmcB (SO1778) in the *fur* deletion mutant by DNA microarray analysis and proteome analysis. Both genes encode a decaheme cytochrome c. Cytochromes OmcA and OmcB contain c-type hemes and are localized to the outer membrane of *S. oneidensis* (Myers and Myers, 1992). A study by Myers and Myers 2001, indicated a role for OmcA and OmcB in the anaerobic reduction of MnO_2 by *S. oneidensis*. Recently, evidence that these outer membrane cytochromes are cell-surface exposed suggests that OmcA and OmcB may directly contact extracellular metal oxides at the cell surface (Myers and Myers, 2003). In cells harboring a fur deletion mutation, expression of omcA was repressed 5.9-fold and 4.8-fold under aerobic and anaerobic respiratory conditions, respectively (Wan, 2004). Similarly, OmcB, which resides immediately downstream of OmcA on the genome, exhibited a 5.3-fold and 4.0-fold reduction in mRNA abundance under the two growth conditions tested. The decrease in omcA and omcB expression at the transcript level positively correlated with the apparent absence of the OmcA and OmcB proteins in FUR2 under aerobic growth conditions. Other than cytochrome c oxidase (SO2363), these were the only cytochromes showing significant changes in protein abundance based on LC-MS/MS analysis. Very little is known about the expression and regulation of OmcA and OmcB, with the exception that OmcA showed increased expression in *S. oneidensis* cells exposed to a shift from aerobic growth to anaerobic growth in the presence of fumarate, Fe(III), or nitrate (Beliaev, 2002b).

In conclusion, we have used a combination of genomic expression profiling and high-resolution proteomic analysis using LC-MS/MS to define the Fur regulon in the

dissimilatory metal-reducing bacterium *S. oneidensis*. Fur functions primarily as a negative regulator of siderophore/receptor-mediated iron transport in *S. oneidensis*, although a deletion mutation in fur had pleiotropic effects on gene expression. The proteome analysis clearly indicated the up-regulation of numerous transporters and receptors with potential function in heavy metal and siderophore uptake under the *fur* deletion mutant. The up-regulation of these proteins corresponded well with microarray data. Four conserved hypotheticals were clearly up-regulated in the *fur* deletion mutant and their up-regulation was confirmed at the transcript level. The proteins are high quality candidates for future functional studies such as gene knockouts, interaction analysis, and metal transport uptake assays. Finally, our studies implicate Fur as an activator of omcA and possibly omcB transcription and translation, although further experiments are needed to confirm this hypothesis and to determine the exact function of these proteins.

Chapter 5

Determination of the Baseline Proteome of the Versatile Microbe *Rhodopseudomonas Palustris* Under its Major Metabolic States

All of the data presented below is in preparation for submission Nathan C. VerBerkmoes, Manesh Shah, Patricia Lankford, Dale Pelletier, Michael B. Strader, David L. Tabb, William. H. McDonald, John. W. Barton, Gregory B. Hurst, Loren Hauser, Brian H. Davison, J.T. Beatty, Caroline S. Harwood, Robert F. Tabita, Robert L. Hettich, and Frank W. Larimer. Determination of the Baseline Proteome of the Versatile Microbe *Rhodopseudomonas Palustris* under its Major Metabolic States. *Journal of Bacteriology* (2005). *All MS sample preparation, experiments and data analysis were performed by Nathan C. VerBerkmoes*

*Complete datasets for the proteomic analyses and other supplementary material are available on the web site http://compbio.ornl.gov/rpal_proteome/

Introduction

Rhodopseudomonas palustris is a purple nonsulfur anoxygenic phototrophic bacterium in the α-proteobacteria family. It is found widely distributed in the environment preferring soil and water samples. *R. palustris* is of great interest due to its high metabolic diversity and ability to degrade simple aromatic hydrocarbons (lignin monomers). It has exceptional metabolic versatility in its modes of energy generation and carbon metabolism. Specifically, this microbe is capable of four major metabolic modes: photoheterotrophic (energy from light and carbon from organic carbon sources) and photoautotrophic growth (energy from light and carbon from carbon dioxide), as well as chemoheterotrophic (carbon and energy from organic compounds) and chemoautotrophic growth (energy from inorganic compounds and carbon from carbon dioxide). It can degrade complex aromatic hydrocarbons and chlorinated pollutants. Furthermore, *R. palustris* is capable of producing hydrogen gas as a by-product of nitrogen fixation making it a potential biofuel producer. *R. palustris* also has the potential to act as a greenhouse gas sink by converting CO_2 into cells. Since most of these metabolic states can easily be produced in laboratory settings, it makes *R. palustris* a model system for the study of diverse metabolic modes and their control. The genome

of this microbe had recently been completed and annotated revealing 4,836 potential protein encoding genes in a 5.459 Mb genome (Larimer, 2004). The genome sequencing effort has paved the way for detailed system biology studies such as transcriptomics, proteomics and protein-protein interactions studies.

Our goal is to use multi-level systems biology studies such as the proteomics technologies of proteome profiling and protein-protein interaction studies (Buchanan, 2002) as well global gene knockouts and transcriptome profiling to obtain a greater understanding of the diverse metabolic states of this microbe and the proteins important to the individual metabolic states. For the initial foundation of this project, we have characterized the baseline proteome of *R. palustris* wild-type strain grown under six conditions including photoheterotrophic, chemoheterotrophic, nitrogen fixation, photoautotrophic, stationary phase, as well as with benzoate as an alternate carbon source. The basic methodology for baseline proteome analysis involves fractionating cells grown under each condition by centrifugation techniques into four major fractions (crude, membrane, pellet and cleared), followed by digestion with trypsin and analysis by liquid chromatography coupled with electrospray (ES)-tandem mass spectrometry (MS/MS) ("shotgun" proteomics Washburn, 2001; VerBerkmoes, 2002; Lipton, 2002; Peng, 2002; Corbin, 2003). This proteome study resulted in the overall identification of 1,664 proteins with conservative filtering constraints. This is the first proteome analysis of *R. palustris* to date providing a deep characterization into the diverse metabolic function of this microbe.

Qualitative analyses of these growth conditions have revealed 311 proteins exhibiting large-scale differences between conditions, many of these being hypothetical or conserved hypothetical proteins showing strong correlations with different metabolic modes. In total, 325 pure hypothetical and conserved hypothetical proteins were identified representing 20% of the identified proteins. A completely novel operon of five proteins was found to be expressed only under the anaerobic states with no evidence of expression under aerobic states. Proteins known to be associated with given growth states such as nitrogen fixation, photoautotrophic and growth on benzoate were up-regulated under those states illustrating the effectiveness of the methodology. The results

of this study are being integrated with other large-scale system studies of *R. palustris*. Notably, the identified proteins, fractionation information and their corresponding growth states have been cataloged in a database. From this database, we have extracted 965 high quality targets for tandem affinity purification for the analysis of protein-protein interactions (Buchanan, 2002). This study illustrates how baseline proteome analysis can be directly coupled with other systems biology studies. By providing all of the datasets as open access, clearly defined and for direct download, we hope these datasets can be used by the scientific community as a whole and provide an example on how large-scale proteome datasets can easily be shared with others in the community.

Materials and Methods

Chemicals and reagents

All salts, DTT, trifluoroacetic acid (TFA), and guanidine used in this work were obtained from Sigma Chemical Co. (St. Louis, MO). Protein concentrations were determined with BCA reagents from Pierce Chemical Co. (Rockford, IL). Modified sequencing grade trypsin, from Promega (Madison, WI), was used for all protein digestion reactions. The water and acetonitrile used in all sample clean up and HPLC applications was HPLC grade from Burdick & Jackson (Muskegon, MI) and the 98% formic acid used in these applications was purchased from EM Science (an affiliate of MERCK KgaA, Darmstadt, Germany).

Cell growth and production of protein fractions

R. palustris strain CGA010, a hydrogen-utilizing derivative of the sequenced strain (obtained from C.S. Harwood) was grown under 6 major conditions for this study: chemoheterotrophic-aerobically in the dark with succinate, photoheterotrophic-anaerobically in the light with succinate or benzoate, photoautotrophic-anaerobically in the light with sodium bicarbonate and H_2, photoheterotrophic stationary phase with succinate, and photoheterotrophic N_2 fixing. A LhaA (light harvesting apparatus assembly protein) mutant (Young, 1998) was grown as a secondary control for

chemoheterotrophic growth. All cultures were grown anaerobically in the light or aerobically in the dark with shaking in defined mineral medium at 30°C to mid-log phase (except the stationary sample) (Kim, 1991). All anaerobic cultures were illuminated with 40 or 60 W incandescent light bulbs. Carbon sources were added to final concentration of 10 mM succinate (all metabolic states except benzoate and photoautotrophic), 3 mM benzoate (benzoate growth) or 10 mM sodium bicarbonate (photoautotrophic and benzoate growth). Growth was monitored spectrophotometrically at 660 nm. For the photoheterotrophic N_2 fixing cultures, ammonium sulfate was replaced by sodium sulfate in the culture medium and N_2 gas was supplied in the head space. For the photoautotrophic growth, H_2 was supplied in the head space. The photoheterotrophic stationary phase culture was grown exactly as the photoheterotrophic log phase state except the culture was allowed to proceed into stationary phase ($OD_{660\ nm}$ >2.0). Cell extracts were prepared as follows: cells were harvested by centrifugation, washed twice with ice-cold wash buffer (50 mM Tris buffer [pH 7.5] with 10 mM EDTA) and resuspended in ice-cold wash buffer. Cells were then lysed with sonication and unbroken cells were removed with low-speed centrifugation (5,000 g x 15 min). Four proteome fractions were created from the cellular extract by ultracentrifugation (100,000 g for 1 hour creates membrane and crude fractions and then 100,000 g for 18 hours creates pellet and cleared fractions). All four proteome fractions were quantified with BCA analysis, aliquoted and frozen at -80^0C until digestion.

Digestion of proteome fractions

All proteome fractions from all growth states were processed in exactly the same way. Briefly, 5 mg of each proteome fraction was diluted in 6 M guanidine and 5 mM DTT then heated at 60°C for 1 hour. The guanidine and DTT were diluted 6-fold with 50 mM Tris/10 mM $CaCl_2$ (pH 7.8) and sequencing grade trypsin was added at 1:100 (wt/wt). The digestions were run with gentle shaking at 37°C for 18 hours followed by a second addition of trypsin at 1:100 and additional 5-hour incubation. The samples were then treated with 10 mM DTT for 1 hour at 60 °C as a final reduction step. Samples were immediately de-salted with Sep-Pak Plus C_{18} solid phase extraction (Waters, Milford,

MA). All samples were concentrated and solvent exchanged into 0.1% TFA in water by centrifugal evaporation to ~10 μg/μL starting material, filtered, aliquoted and frozen at -80°C until LC-MS/MS analysis.

LC-MS/MS analysis

The four proteome fractions from all growth states discussed above were analyzed in duplicate via one-dimensional LC-MS/MS experiments performed with an Ultimate HPLC (LC Packings, a division of Dionex, San Francisco, CA) coupled to an LCQ-DECA or LCQ-DECA^{XPplus} quadrupole ion trap mass spectrometer (Thermo Finnigan, San Jose, CA). Automated 50 μL injections were made with a Famos autosampler (LC Packings) onto the HPLC column. Flow rate was set at 4 μL/min with a 240-min gradient for each LC-MS/MS run. A VYDAC 218MS5.325 (Grace-Vydac, Hesperia, CA) C_{18} column (300 μm i.d. x 25 cm, 300Å with 5 μm particles) was used for all separations. The column was directly connected to the Finnigan electrospray source with 100 μm i.d. fused silica. For each new growth state, a new HPLC column was used to prevent cross-contamination. For all LC-MS/MS data acquisition, the LCQ was operated in the data dependent mode with dynamic exclusion enabled (repeat count 1), where the top four peaks in every full MS scan were subjected to MS/MS analysis. For all experiments, the mass spectrometer was operated with the following parameters: ES voltage 4.5 kV, heated capillary 200°C, 5 microscans averaged for full scans and MS/MS scans, 5 m/z isolation widths for MS/MS isolations and 35% collision energy for collision-induced dissociation. To increase dynamic range in the 1D-LC-MS/MS analysis, separate injections were made with a total of 8 overlapping segmented m/z ranges scanned (referred to as gas phase fractionation or multiple mass range scanning).

Data analysis

The resultant ~450 LC-MS/MS runs were all processed as follows. The MS/MS spectra from all files were first searched with SEQUEST (Thermo Finnigan) with two modes: tryptic only (only fully tryptic peptides were considered) and non-tryptic (fully tryptic, partially tryptic and fully-non tryptic peptides were considered). The MS/MS

spectra from all files were then searched with DBDigger (Tabb, 2005) with only the fully tryptic option. For all database searches, an *R. palustris* proteome database was used, which contained 4,833 proteins and 36 common contaminants (http://compbio.ornl.gov/rpal_proteome/databases). All resultant output files from SEQUEST and DBDigger were organized by growth state and run number (all fractions from a single proteome analysis were combined) and then filtered by DTASelect (Tabb, 2002) at the 1-peptide, 2-peptides and 3-peptides level with the following parameters: SEQUEST, delCN of at least 0.08 and cross-correlation scores (Xcorrs) of at least 1.8 (+1), 2.5 (+2) and 3.5 (+3); DBDigger, delCN of at least 0.08 and MASPIC scores of at least 25 (+1), 30 (+2) and 45 (+3). The filtered DTASelect files from proteome replicates were compared with Contrast (Tabb, 2002) to ensure quality reproducibility (at least 70% similar protein identifications at the 2-peptide filter level between replicates were required). Contrast was then used to create pairwise comparisons of growth states as well as a global comparison of all growth conditions (http://compbio.ornl.gov/rpal_proteome/analysis). For the evaluation of protein fractionation and the creation of tandem affinity purification targets, the fractions from individual proteome analysis were not concatenated but rather analyzed individually by DTASelect and then compared with Contrast (this was only done with the SEQUEST fully tryptic dataset).

Results

The goal of this study was to obtain baseline proteome analysis of *R. palustris* under major metabolic states. This is possible since *R. palustris* can easily be grown under all metabolic states listed above except chemoautotrophic growth. Furthermore, variations of these major metabolic modes, such as nitrogen fixation, can also be readily accomplished. For this study, 7 major growth states were investigated. These can be broken down into two major categories: aerobic growth in the dark and anaerobic growth in the light (Figure 5.1). For aerobic growth, two states were evaluated: wild-type chemoheterotrophic, in which succinate was used as a carbon and energy source and the samples were fully aerated in the dark. A LhaA mutant (light harvesting assembly

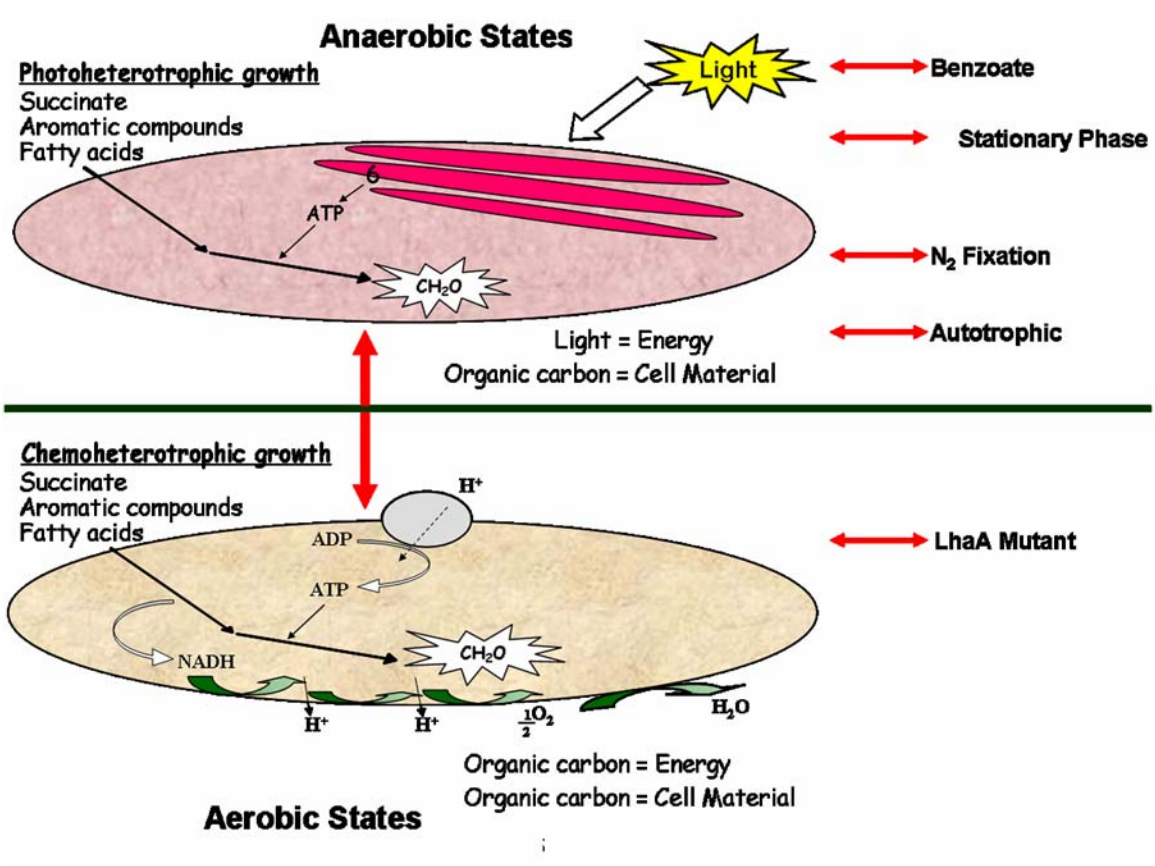


Figure 5.1: Major metabolic states interrogated in this study.
Top states are all anaerobic grown in the light without oxygen. Bottom states are all aerobic grown in the dark with oxygen present.
Figure adopted from Larimer, 2004 and modified.

mutant) was grown under the same conditions as a secondary control for aerobic growth (the mutation is thought to only effect the assembly of the light harvesting complex and should not have a major effect on aerobic growth in the dark). For anaerobic growth, five states were characterized; all were wild-type strain and all were kept anoxygenic with light as the energy source. The core state was photoheterotrophic where succinate was provided as a carbon source and ammonium sulfate as the nitrogen source, photoautotrophic had CO_2 and solid sodium bicarbonate substituted as the carbon source, photoheterotrophic nitrogen fixation had N_2 substituted as the nitrogen source, photoheterotrophic benzoate had benzoate substituted as the carbon source and photoheterotrophic stationary phase was grown into stationary phase while all other states were harvested in mid-log phase.

All growth states were prepared with the exact same protocol. Cells were harvested, washed and lysed by sonication. Four proteome fractions were created by ultracentrifugation, digested with trypsin and then analyzed in duplicate by an automated 1D-LC-ES-MS/MS technique which employed multiple mass range scanning. Multiple mass range scanning is a simple technique to increase dynamic range for proteome measurements, which involves injecting the same sample repeatedly with overlapping narrow mass ranges scanned by the mass spectrometer (VerBerkmoes, 2002). We have found this technique to be very reproducible and simple to implement for a large-scale study of many samples. Its main disadvantage is the large amount of sample needed since multiple injections must be made; this was not a concern for this study since plenty of protein could be obtained from a single 2 L culture of *R. palustris.* A total of ~450 LC-ES-MS/MS runs were created from this study (7 growth states, 4 proteome fractions, 8 runs per fraction, 2 duplicates of each growth state), all were searched individually with SEQUEST (Eng, 1994) and DBDigger (Tabb, 2005), filtered with DTASelect and compared with Contrast (Tabb, 2002). All resultant DTASelect files and Contrast files used in this study, as well as the protein database, can be downloaded from the *Rhodopseudomonas Palustris* Proteome Study Website (http://compbio.ornl.gov/rpal_proteome/). This site also contains directly linkable

spectra for all identified peptides, a step towards open access proteome results (Carr, 2004; Pedrioli 2004).

The entire dataset was processed three separate ways: SEQUEST fully tryptic, SEQUEST non-tryptic, and DBDigger fully tryptic. DBDigger is a new search algorithm developed at Oak Ridge National Laboratory (ORNL), which provides better accuracy and sensitivity than SEQUEST. The results from each of these searches filtered at 1-peptide, 2-peptides and 3-peptides are shown in Table 5.1 (the numbers in these tables also include some common contaminants, thus numbers listed below have these removed and are smaller). Clearly, non-specific searches identify many more proteins than fully tryptic SEQUEST searches do but with a greater level of false positives. There is currently controversy in the proteomics field over the use of non-specific searches for data generated from trypsin digestions, thus we will confine our discussions to the fully tryptic datasets, but we provide the non-specific dataset for comparison. DBDigger has just recently been published and has not been tested by the community as a whole and this is the first large-scale use of the algorithm for MS/MS data analysis. From the fully tryptic dataset of proteins identified by at least two peptides, 1,691 proteins were identified by DBDigger and 1,664 proteins were identified by SEQUEST. The combined list from these two algorithms resulted in the overall identification of 1,805 proteins; 1,549 are shared between the lists, with 140 proteins identified only by DBDigger and 116 identified only by SEQUEST. Table S1 contains all proteins identified by each algorithm as well as corresponding sequence coverage for each protein. Figure 5.2 shows the scatter plot of % sequence coverage for each protein and for each algorithm plotted against each other. This scatter plot clearly shows that each algorithm is providing very similar results with few outliers. Proteins found on the x-axis or y-axis were only identified by one of the two algorithms. Those with very high sequence coverage found by only one of the programs are the most significant outliers and should be investigated in more detail in a future study. Since DBDigger is a new search algorithm, which has not been fully tested, we will focus our biological analysis below on those proteins confidently identified from the SEQUEST fully tryptic dataset. The dataset is provided as the first major comparison between the two algorithms.

Table 5.1: Total identified proteis by search algorithm and filtering level.

Filtering Level	SEQUEST fully tryptic[a]	SEQUEST non-tryptic[a]	DBDigger fully tryptic[b]
1-peptide	2752	4482	2785
2-peptides	1670	2888	1698
3-peptides	1317	1833	1338

a) Filters for SEQUEST Xcorrs of at least 1.8 (+1), 2.5 (+2) and 3.5 (+3)
b) Filters for DBDigger MASPIC scores of at least 25 (+1), 30 (+2) and 45 (+3)

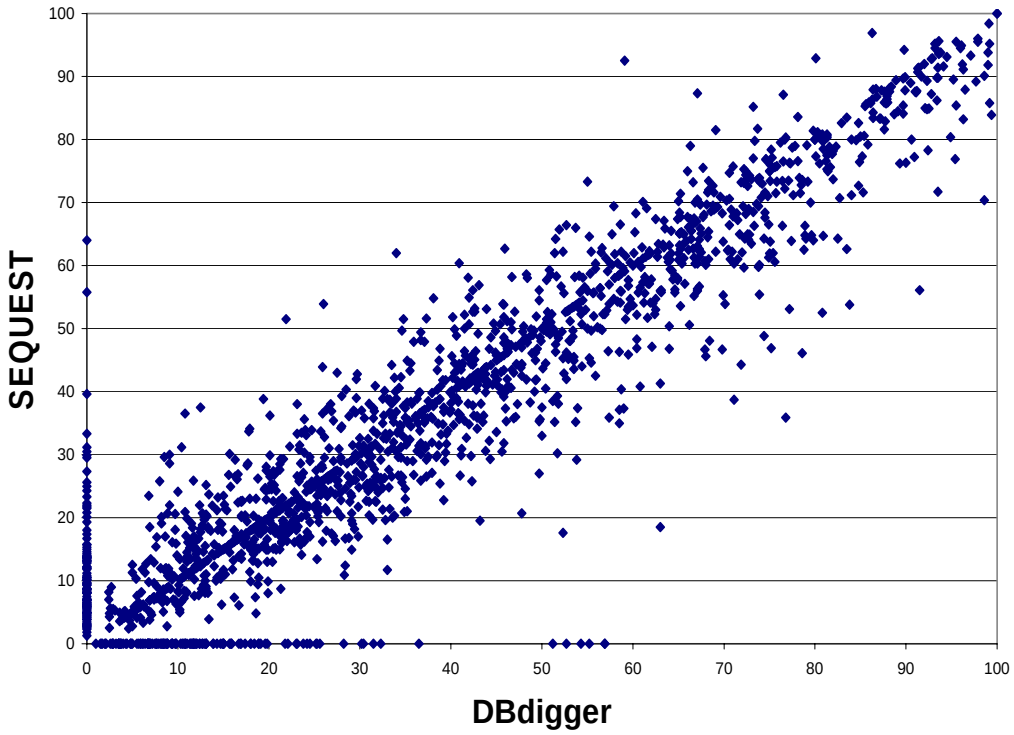


Figure 5.2: Scatter plot of SEQUEST vs. DBDigger % sequence coverage per protein.
Illustrates the correspondence of SEQUEST % sequence coverage (y-axis) to DBDigger % sequence coverage (x-axis) for every protein identified by one or both algorithms at the 2-peptide level with filters shown in Table 5.1.

Major features of the proteome

The identified protein totals from fully tryptic SEQUEST searches of each growth state analysis filtered at 1-peptide, 2-peptides and 3-peptides are shown in Table 5.2 Statistical analysis of a single growth state by the peptide/protein prophet software (Keller, 2002; Nesvizhskii, 2003) indicated a 5% false positive rate at the 1-peptide level, a 1% false positive rate at the 2-peptides level and virtually no false positives at the 3-peptides level (personal communication A. Nesvizhskii). We felt that 1% false positive was acceptable so we accepted all proteins identified with at least two fully tryptic peptides for further analysis. After the removal of common contaminants (trypsin, keratin, etc.) from the final protein list, a total of 1,664 proteins were identified. The entire list of identified proteins with predicted functions, % sequence coverage, functional categories, pIs and molecular weights (MW) can be found in Supplemental Table 2. The identified proteins with sequence coverage from individual growth states can be found in Supplemental Table 3. To determine if this methodology had any major biases against certain protein forms, we compared the identified proteins' pI and MW ranges with those predicted from the entire genome (Figure 5.3). We found no major biases against the MW or pI of the proteins identified in this study. The protein with the lowest pI detected was RPA0060 (pI 3.82, MW 14.4 kDa), a conserved unknown protein, which was detected with ~50% sequence coverage from every growth state in this study. The protein with the highest pI detected was RPA4197 (pI 12.37, MW 5065.6 kDa), ribosomal protein L36, which was detected with an average of 29% sequence coverage in three of the seven growth states. Ribosomal protein L36 was also the smallest protein detected as well. The largest protein detected was RPA3958 (pI 4.87, MW 215.1 kDa), a conserved unknown protein, which was detected with an average of ~11% sequence coverage from every growth state in this study.

The functional categories for the identified proteins are shown in Table 5.3 (these functional categories are based on the ORNL annotation scheme for all bacteria (http://genome.ornl.gov/microbial/). We also mapped all of the identified proteins onto KEGG pathways (Kyoto Encyclopedia of Genes and Genomes) and indicated which

Table 5.2: Identified proteins by growth state and filtering level.

Growth Condition:	3-Peptide	2-Peptide	1-Peptide
LhaA mutant Chemoheterotrophic Run 1	714	931	1287
LhaA mutant Chemoheterotrophic Run 2	722	930	1322
WT Chemoheterotrophic Run 1	721	941	1350
WT Chemoheterotrophic Run 2	631	809	1235
WT Photoheterotrophic Run 1	692	884	1263
WT Photoheterotrophic Run 2	724	891	1251
WT Photoheterotrophic Stationary Phase Run 1	725	921	1369
WT Photoheterotrophic Stationary Phase Run 2	746	930	1405
WT Photoautotrophic Run 1	764	961	1455
WT Photoautotrophic Run 2	695	900	1391
WT Photoheterotrophic Nitrogen Fixation Run1	768	995	1441
WT Photoheterotrophic Nitrogen Fixation Run2	807	1018	1508
WT Benzoate Photoheterotrophic Run1	624	814	1273
WT Benzoate Photoheterotrophic Run2	686	884	1285
Total Proteins Identified	1320	1670	2752

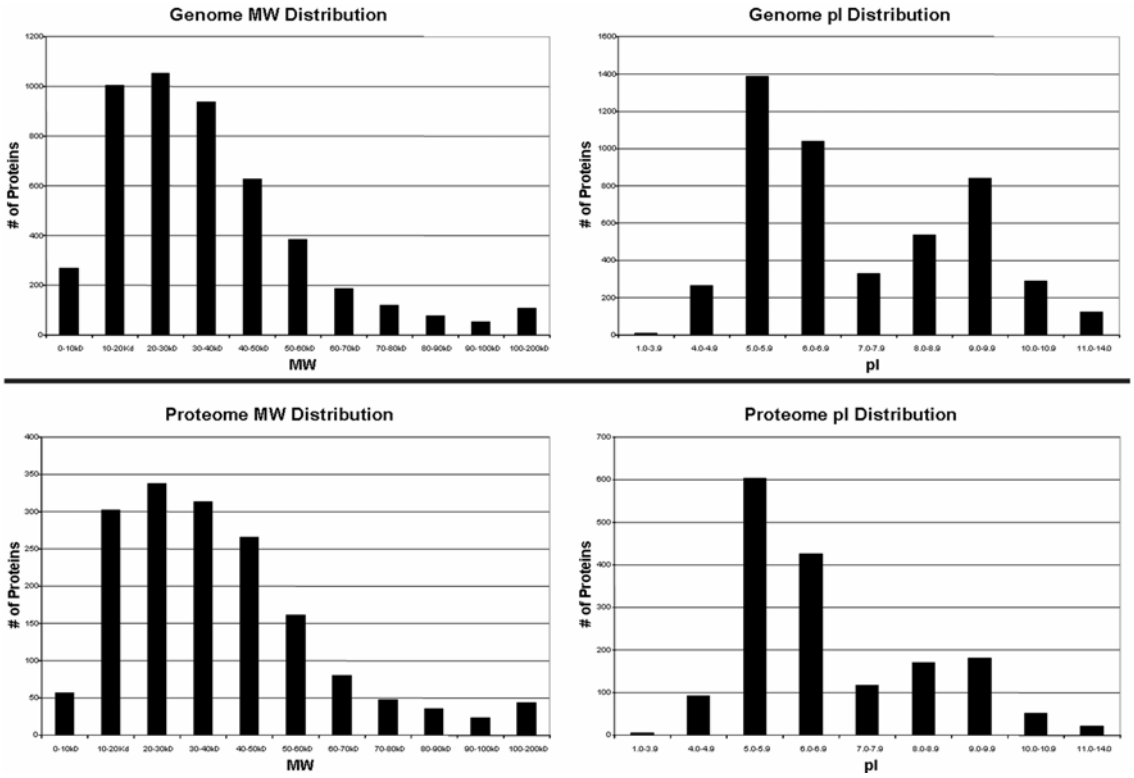


Figure 5.3: MW and pI comparisons for genome and proteome.
Comparison between predicted MW and pI distribution from all proteins predicted from the genome (top) with those identified from the proteome with at least two unique peptides (bottom).

Table 5.3: Functional categories.

Category	Proteins	Genome Prediction	% Identified
Unknowns and Unclassified	325	1407	23.10
Replication and Repair	22	126	17.46
Energy Metabolism	142	306	46.41
Carbon and Carbohydrate Metabolism	67	107	62.62
Lipid Metabolism	98	158	62.03
Transcription	59	283	20.85
Translation	122	168	72.62
Cellular Processes	207	524	39.69
Amino Acid Metabolism	104	181	57.46
General Function Prediction	157	420	37.38
Metabolism of Cofactors and Vitamins	80	150	53.33
Transport	168	699	24.03
Signal Transduction	73	231	31.60
Purine and Pyrimidine Metabolism	40	56	71.43
Total	1664	4816	34.57

metabolic states they were detected under (note this was only done for those proteins which were found in pathways predicted by KEGG, the entire pathways can be found at http://compbio.ornl.gov/rpal_proteome/analysis/). The proteome was dominated with hypothetical and conserved hypothetical proteins, 325 in total being confidently identified. This represents 23% of the predicted hypothetical proteins from the genome. In our classification scheme, proteins' names are changed from hypothetical and conserved hypothetical to unknown and conserved unknown when they are confidently identified with at least two unique peptides. A total of 107 unknown proteins and 218 conserved unknown proteins were identified. The proteomes' second most dominant category was proteins involved in general cellular processes such as chaperones, proteases, flagella proteins, stress proteins and some general enzymatic proteins. A total of 207 proteins were identified in this category representing 40% of those predicted from the genome. The *R. palustris* genome contains two separate copies of GroEL (RPA 1140 and RPA 2164) and GroES (RPA1141 and RPA2165). Proteomics clearly identified each copy of each subunit with a number of unique peptides. Both copies were expressed under all growth states interrogated in this study. The proteomes' third most dominant category was the transport proteins. A total of 168 proteins were detected from this category representing 24% of those predicted from the genome. The genome sequencing and annotation effort predicted widespread use of transport systems in this microbe with 325 complete transport systems representing almost 15% of the predicted genome (Larimer, 2004). This is larger than most microbes where transport is generally predicted at 5-6%. Approximately 10% of the detected proteome were proteins involved in transport. Many of the detected transport proteins were the ATP-binding cassette (ABC) systems. In many cases only part of the entire operon was detected, generally the periplasmic binding proteins were detected at the highest sequence coverage, while the embedded membrane transporter and related ATPases were detected with lower sequence coverage or not at all. This may account for part of the difference between the larger number of predicted transport proteins and the actual number detected. Eleven tonB-dependent receptors/iron transporters were detected suggesting iron acquisition is

important for *R. palustris* as also indicated by the genome annotation. A number of these were found to be up-regulated under aerobic growth as discussed below.

The categories which were identified with the highest percentage of those proteins from the genome included translation (72%), purine/pyrimidine metabolism (71%), carbon and carbohydrate metabolism (62%), lipid metabolism (62%), and amino acid metabolism (57%). This is to be expected since many of the proteins in these categories are necessary under all metabolic modes. In a previous study of the purified 70S ribosome from *R. palustris*, we clearly identified 53 of the 54 predicted ribosomal proteins (Strader, 2004). From this study, we clearly identified 50 of the 54 predicted ribosomal proteins directly from the proteome. The missed ribosomal proteins are all small and highly rich in lysine residues. Digestion of such proteins results in small peptides, which are not readily amenable to LC-MS/MS. A total of 18 of the potential 20 tRNA synthetases were confidently identified. Most of the ribosomal proteins and tRNA synthetases were found under every growth state characterized (Table S3). While only 40 proteins were detected from purine and pyrimidine metabolism, this represented most of the predicted proteins from this group. As was the case with many of the translation proteins, many of the purine/pyrimidine metabolism proteins were found under most of the metabolic states. Carbon and carbohydrate metabolism was dominated by glycolysis/gluconeogensis and TCA proteins. The entire pathway for each was detected with most proteins detected under every metabolic state. A total of 98 of the predicted 158 proteins involved in lipid metabolism were detected.

The proteome categories of replication and repair, energy metabolism, transcription, general function prediction, metabolism of co-factors and vitamins, and signal transduction were all identified but many with less than 50% of that predicted from the genome. Replication and repair was detected with the smallest percentage and smallest numbers of proteins with only 22 proteins and 17% of the predicted proteome. This may be due to the low abundance of these proteins and their rapid turnover; they may only be used during DNA replication and repair and then quickly degraded. Proteins involved in transcription were also not detected in large quantities, with only 59 total proteins detected representing 21% of those predicted from the genome. The proteins

involved in signal transduction are also generally considered to be of lower abundance. A total of 73 proteins representing 31% of those predicted from the genome were confidently detected. Many of these were methyl-accepting chemotaxis proteins, two-component response regulators, and two-component sensor histidine kinases. The detection of these proteins clearly indicates the increased dynamic range of LC-MS/MS techniques over 2D-PAGE-MS techniques where such low abundance proteins are rarely detected.

A total of 142 proteins, which is nearly 50% of the proteins, predicted to be involved in energy metabolism were confidently detected. These include many of the proteins involved in photosynthesis and the oxidative phosphorylation chain. As with the ABC transporters, it was found that very often only part of the entire known protein complexes in these pathways were confidently identified. For example, ATP synthase is encoded by two operons, the complex consists of proteins RPA0175-RPA0179 and RPA0843-RPA0846 (see KEGG map at: http://compbio.ornl.gov/rpal_proteome/analysis/keggmaps/html/map00193.html). For this case, all of the proteins encoded by the first operon are predicted to be connected to the membrane but not directly embedded and were detected with high sequence coverage under every metabolic state. The second operon encodes two proteins, which are associated with the membrane (RPA0843 ATP synthase B chain and RPA0844 ATP synthase, subunit B') but not directly embedded. These proteins were also readily detected under all metabolic states. The final two proteins of this operon, RPA0845 (ATP synthase subunit C transmembrane protein) and RPA0846 (ATP synthase subunit A) are both small membrane embedded proteins. Neither RPA0845 nor RPA0846 were detected in any of the metabolic states even though they are expressed at high abundance with the rest of the complex. This illustrates a common problem in “shotgun” proteomics applications in that small proteins embedded in the membrane of membrane associated complexes are very difficult to routinely detect. The routine detection of such proteins is a current area of research and will be addressed in a future manuscript. This same problem was also encountered for proteins involved in photosynthesis, the cytochrome-c oxidase complex and the NADH-ubiquinone dehydrogenase complex. The NADH-

ubiquinone dehydrogenase complex is another example of gene duplication in *R. palustris*. Two operons (RPA2937-RPA2952 and RPA4252-RPA4264) are predicted to encode proteins for this complex. A total of 6 proteins were identified from the first operon and 4 proteins from the second operon indicating both operons are indeed expressed. These proteins were observed across all metabolic states indicating expression under all metabolic states.

Metabolic state comparisons

One of the goals of this study was to identify the major differences at the protein level between the major metabolic states of *R. palustris*. A quantitative comparison of many different growth states is currently a serious challenge for MS-based proteomics efforts. While great effort has been put forth into relative quantitation of proteins between different growth states using technologies such as isotope coding affinity tags (ICAT) (Gygi, 1999), metabolic labeling (Oda, 1999; Pasa-Tolic, 1999) and ^{18}O water labeling (Yao, 2001) none of these techniques have been clearly shown to be effective for large-scale studies of many growth states in microbial systems. Specifically, the ICAT technology requires the labeling of cysteine residues, which are not very prevalent in microbial systems when compared with eukaryotic systems. Indeed, ~60% of the *R. palustris* predicted proteome contains 2 or less cysteines per protein and 20% contains no cysteines at all. Metabolic labeling with ^{15}N has shown the most promise in microbial systems (Lipton, 2002; Washburn, 2002) but requires strict control of nitrogen intake, a serious concern in bacterial species that can fix nitrogen. Indeed, many microbial species cannot be cultivated under conditions where strict control of nitrogen intake is required. Labeling peptides during trypsin digestion with ^{18}O water is a potential alternative approach but the expense involved in labeling the number of samples used in this study and the need for high-resolution mass spectrometers limits its use for large-scale proteome comparisons. A more practical problem with quantitative comparisons of many growth states is that the necessary informatic tools are still under development. The Contrast program was used in this study to compare the many different growth states and

replicates in this study. No such global comparison program for quantitative proteomics datasets has been developed at this point.

It has been shown that semi-quantitative comparisons of proteome datasets based on the % sequence coverage, # of identified peptides, and the repeat count for a protein (how many MS/MS sequencing events are acquired per protein) are all indicators of protein abundance (Liu, 2004). In a previous study of *Shewanella oneidensis*, we used the % sequence coverage and number of unique peptides per protein identified in triplicate analysis of a control and a *fur* mutant to compare relative protein abundances (Wan, 2004). We found this technique to be very indicative of proteins showing major differences between the two growth states and the results compared favorably with microarray data from the same samples. Below we used the same technique of comparing the % sequence coverage and the number of unique peptides per protein between different metabolic states to identify those proteins showing large-scale differences between compared states. We used the general rule that a protein must have a replicated difference of at least four peptides and/or 30% sequence coverage between the two states being compared to be called a major difference (Wan, 2004). It is very important in such comparisons to process and analyze samples under the same conditions to achieve the best reproducibility possible. Table 5.4 illustrates the reproducibility of proteins detected between duplicate runs for each metabolic state. At the two peptides filtering level, between 70-80% reproducibility was achieved for each metabolic state. In our experience, that is the best that can be achieved with current "shotgun" proteomics technologies. Most of the proteins which did not replicate between runs were identified with less than three to four peptides. None of these were considered in the comparison of the metabolic states. Proteins identified as showing major differences were then compared across all growth states to determine if trends in expression could be determined. In total we found 311 proteins exhibiting major differences between growth states. The entire list of proteins sorted by compared metabolic states can be found in supplemental table S4. It should be noted that this technique is only useful in determining proteins exhibiting large-scale differences in expression between growth states and generating hypothesis about these proteins, which can be tested in future

Table 5.4: Reproducibility of identified proteins by growth state.

Growth Condition	1 peptide	2 peptide
LhaA mutant Chemoheterotrophic	71.7%	75.7%
WT Chemoheterotrophic	68.1%	70.8%
WT Photoheterotrophic	70.5%	77.7%
WT Photoheterotrophic Stationary	70.1%	79.1%
WT Photoheterotrophic Nitrogen Fixation	69.8%	75.7%
WT Photoautotrophic	70.3%	77.3%
WT Photoheterotrophic Benzoate	67.7%	73.7%

studies. Exact quantification of the protein differences cannot be established with this technique.

Chemoheterotrophic WT vs. chemoheterotrophic LhaA mutant

The chemoheterotrophic LhaA mutant (Young, 1998) was analyzed as a secondary control for aerobic growth. It was used in the global comparison of aerobic and anaerobic states (Figure 5.1) since this mutation is not believed to have any major effect on aerobic growth in the dark. For this to be effective, very few differences should be seen between the chemoheterotrophic wild-type and chemoheterotrophic LhaA mutant. This was indeed the case, as only 6 proteins were found to have significant difference between the WT and LhaA mutant (Table S4, 1st tab). Thus we concluded that this mutant could be used as a secondary control for aerobic growth and directly compared with the anaerobic states as discussed below.

Chemoheterotrophic vs. photoheterotrophic

The chemoheterotrophic vs. the photoheterotrophic states are the base states, which all other states are compared against in this study, as shown in Figure 5.1. In theory, these two states should be very different with chemoheterotrophic obtaining energy from carbon compounds (succinate) and the photoheterotrophic obtaining energy from light. In reality, many of the genes involved in photosynthesis were found under every state studied whether the samples were grown in the dark or light. For example, gene RPA1548, which encodes for the H subunit of the photosynthetic reaction center, was found under every metabolic state, though with slightly lower sequence coverage under the aerobic states, which were grown in the dark. It seems that no matter what the condition, *R. palustris* attempts to gain energy through photosynthesis by expressing proteins involved in photosynthesis. Indeed, the hallmark phenotype of photosynthesis, the red coloring of the cell membranes was observed for every metabolic state, though the red coloring was much more pronounced under photo states. Nonetheless, many differences were found at the protein level between these two states. In total, 31 proteins were found to be clearly up-regulated under the chemoheterotrophic state and 56 proteins

were found to be up-regulated under the photoheterotrophic state (Table S4, 2nd tab). A total of 8 unknown and conserved unknown proteins were found to be up-regulated under the chemoheterotrophic state. The unknown proteins RPA2269, RPA2471, RPA3930 all showed strong correlation with the aerobic states with very little expression under any of the anaerobic states. The conserved unknown protein RPA4179 was found with ~90% sequence coverage under both the WT and LhaA mutant under aerobic conditions and was not found under any anaerobic states except nitrogen fixation and benzoate growth where it was also found with high sequence coverage (>50%). Additionally, proteins involved in iron uptake and utilization such as RPA1876 (putative TonB-dependent iron siderophore receptor), RPA2120 (putative hemin binding protein), RPA3480 (fiu putative outer membrane receptor for iron transport), and RPA4152 (fbpA periplasmic iron binding protein FbpA precursor) were all up-regulated under the aerobic state.

A total of 17 unknown and conserved unknown proteins were found to be up-regulated under the photoheterotrophic state. The unknown proteins RPA1494, RPA1495, RPA1620, RPA2333, RPA2334, RPA2335, RPA2336, RPA2338, and RPA3786 all showed strong correlation with the anaerobic states with very little expression under any of the aerobic states. The unknown protein RPA3011 was only found with high sequence coverage under photoheterotrophic growth; with all other anaerobic states it was found with less than 10% coverage or not at all. The operon of unknown proteins from RPA2333-2338 is an especially interesting case; this entire operon, except RPA2337, was found to show strong expression under anaerobic states but no expression in the aerobic states (Table 5.5). The lack of detection of RPA2337 cannot be explained; the protein has no predicted transmembrane domains and predicted cleavage by trypsin indicates 5-6 peptides that should easily be detected. None of the proteins in this operon have been found to have strong similarity to any genes in sequenced microbial genomes to date except RPA2333. RPA2333 does have strong similarity to a putative cation transport ATPase but does not have the predicted transmembrane domains generally associated with such a transport ATPase. This coupled with the fact that the rest of the proteins in the operon do not show any similarity to any other know protein indicates we have no strong evidence for the function of

Table 5.5: Unknown operon identified in the anaerobic metabolic states.

Locus	lhaa	Aerob	Anaerob	Stat	Auto	N2	Benz	Functional Assignement
RPA2333	0	0	18	19	4	24	13	unknown protein
RPA2334	0	0	45	52	12	42	42	unknown protein
RPA2335	0	0	14	20	0	14	14	unknown protein
RPA2336	0	0	79	74	74	74	69	unknown protein
RPA2337	0	0	0	0	0	0	0	unknown protein
RPA2338	0	0	28	42	14	37	33	unknown protein

*Numbers in table report the percentage of residues in protein sequences that were identified by at least two peptides passing the filtering criteria, numbers are averages of two runs.
#Cells highlighted in grey are anaerobic states.

RPA2333. *Thus we have detected a completely novel, expressed operon with potential function under anaerobic growth*. The entire operon was mainly detected in the membrane fraction with some detection in crude and pellet fractions (see discussion below), suggesting a potential membrane embedded protein complex or large protein complex. The lack of predicted transmembrane domains (RPA2335 has 2, RPA2338 has 2 and the rest have 0) suggests the possibility of a large protein complex which is fractionating with the membrane fraction. Further evidence came from a separate study of the purification of the 70S ribosome (Strader, 2004). In screening of the fractions from the first sucrose gradient enrichment of the 70S ribosome from the photoheterotrophic state, we detected the entire operon in one fraction (except RPA2337). Upon further purification of the ribosome, the operon was not detected. Most proteins found in the fractions of the sucrose gradient were indeed proteins involved in large multimeric protein complexes. While this evidence is anecdotal, this operon is clearly a good target for future functional studies such as gene knockouts and protein interaction studies through tagging protocols or biochemical enrichment.

The conserved unknown proteins RPA0932 and RPA0934 along with the putative protease RPA0933 also make up an operon that shows expression only under the anaerobic states. RPA0933 can be speculated to be involved in processing of protein(s) necessary for anaerobic growth, which are not required under aerobic growth. The functions of the two adjacent proteins from the operon are clearly unknown. The conserved unknown proteins RPA1659, RPA3501, RPA4127 all showed strong correlation will the anaerobic states with very little expression under any of the aerobic states. The conserved unknown protein RPA3501 is an interesting case, showing 80-90% sequence coverage with a high repeat count and a high number of peptides under all anaerobic states (suggesting high expression), yet it was undetected in the aerobic states. Thus, all indicators of abundance point to this protein being highly expressed under anaerobic states yet we have no clue to its function. Again, this gene is a good candidate for future functional studies. Some other highlights of the photoheterotrophic state include the expression of a universal stress protein (RPA1260). This protein was found to be expressed only under the anaerobic states with no expression under the aerobic

states indicating a potential function in the stress response during anaerobic growth. The expression of the cbb operon (RPA4641-RPA4645) is up-regulated in the photoheterotrophic state and this trend was also observed across all anaerobic states with less expression found in the aerobic states.

Photoheterotrophic vs. nitrogen fixation

The photoheterotrophic state vs. the photoheterotrophic nitrogen fixation state was an ideal test for the effectiveness of this methodology since many of the proteins associated with nitrogen fixation are known and should primarily be expressed under the nitrogen fixation condition. This was indeed found to be the case. In total, 12 proteins were found to be clearly up-regulated under the photoheterotrophic state and 40 proteins were found to be up-regulated under the photoheterotrophic nitrogen fixation state (Table S4, 3rd tab). Most of the proteins thought to be involved in nitrogen fixation were found to be clearly up-regulated under the nitrogen fixation condition and not detected to any great extent under any of the other conditions. These include RPA0274, a nitrogen regulatory protein; RPA2593 and RPA2595, nitrogen assimilation regulatory proteins; RPA4209, glutamine synthetases; and the entire nif regulon RPA4602-4632 (RPA4633 is also part of this regulon but was barely detected). The nif regulon contains the functional protein complex for nitrogen fixation (RPA4618-4620) which was clearly up-regulated. The major nitrogen fixation proteins, as well as some other proteins showing expression only under nitrogen fixation conditions, are compared with expression levels for all other metabolic states in Table 5.6. *The clear identification of these proteins under the nitrogen fixation states and not under other states indicates the effectiveness of this qualitative technique for comparing large numbers of metabolic states in microbial systems without exact quantitation*. As indicated in Table 5.6, a number of proteins not directly predicted to be involved in nitrogen fixation were also identified only under nitrogen fixation conditions. A few examples include RPA0761, a possible oligopeptide transporter and RPA3669, a putative urea short-chain amide/branched-chain amino acid uptake ABC transporter periplasmic solute-binding protein precursor. The expression of these two proteins could be involved with the bacteria attempting to gain nitrogen by the

Table 5.6: Some proteins identified only under nitrogen fixation.

Locus	lhaa1	Aerob	Anaerob	Stat	Auto	N2	Benz	Functional Assignement
RPA0274	0	0	0	0	0	90	15	GlnK, nitrogen regulatory protein P-II
RPA0761	0	0	0	0	0	25	0	possible oligopeptide ABC transporter,
RPA1206	0	0	0	0	0	20	0	aldehyde dehydrogenase
RPA1927	0	0	0	0	0	54	0	unknown protein
RPA1928	0	0	0	0	0	67	0	ferredoxin-like protein [2Fe-2S]
RPA2156	0	0	0	0	0	41	0	unknown protein
RPA2593	0	3	0	0	0	14	2	nitrogen assimilation regulatory protein ntrC
RPA3669	0	0	0	0	0	74	0	amino acid uptake ABC transporter
RPA4209	0	0	0	0	0	43	0	glutamine synthetase II
RPA4602	0	0	0	0	0	42	0	ferredoxin like protein, fixX
RPA4603	0	0	0	0	0	38	0	nitrogen fixation protein,fixC
RPA4604	0	0	0	0	0	34	0	electron transfer flavoprotein alpha chain
RPA4605	0	0	0	0	0	40	0	electron transfer flavoprotein beta chain fixA
RPA4608	0	0	0	0	0	9	0	nitrogenase cofactor synthesis protein nifS
RPA4610	0	0	0	0	0	16	0	Protein of unknown function, HesB/YadR/YfhF
RPA4612	0	0	0	0	0	17	0	ferredoxin 2[4Fe-4S] III, fdxB
RPA4613	0	0	0	0	0	59	0	DUF683
RPA4614	0	0	0	0	0	41	0	DUF269
RPA4615	0	0	0	0	0	76	0	nitrogenase molybdenum-iron protein nifX
RPA4618	0	0	0	0	0	69	0	nitrogenase molybdenum-iron protein beta chain,
RPA4619	0	0	0	0	0	72	0	nitrogenase molybdenum-iron protein alpha chain,
RPA4620	0	0	0	0	0	63	0	nitrogenase iron protein, nifH
RPA4623	0	0	0	0	0	87	0	fixU, nifT
RPA4631	0	0	0	0	0	55	0	ferredoxin 2[4Fe-4S], fdxN
RPA4632	0	0	0	0	0	16	0	NIFA, NIF-SPECIFIC REGULATORY protein
RPA4714	0	0	0	0	0	42	0	unknown protein

*Numbers in table report the percentage of residues in protein sequences that were identified by at least two peptides passing the filtering criteria, numbers are averages of two runs.

uptake of amino acids or peptides from the media (while amino acids/peptides were not directly added to the media, the cells may have been scavenging them from broken-down dead cells and broken-down excreted proteins). Three unknown proteins (RPA1422, RPA2126, and RPA4717) were also only detected under nitrogen fixation, suggesting a potential unknown role in the nitrogen fixation process. These are good future targets for functional studies such as gene knockouts.

Photoheterotrophic vs. photoautotrophic

The photoheterotrophic state was compared with the photoautotrophic state to determine proteins important to the carbon fixation process. In total, 18 proteins were found to be clearly up-regulated under the photoheterotrophic state and 37 proteins were found to be up-regulated under the photoautotrophic state (Table S4, 4th tab). Three unknown proteins and 1 conserved unknown protein were found to be down-regulated in the photoautotrophic state. The glutamate synthase complex, as well as the protochlorophyllide reductase complex, was found to be down-regulated. As expected, the ribulose-bisphosphate carboxylase large chain (RPA1559) and small chain (RPA1560) (RubisCO form I) were clearly up-regulated under autotrophic growth. The only growth states that RubisCO form I was detected with high sequence coverage was autotrophic and benzoate growth (see discussion on benzoate growth below). This is expected since this is the key enzyme involved in carbon fixation. The RubisCO form II protein (RPA4641) was detected under all growth states. Interestingly, RPA1561, a cbbX protein homolog part of the RubisCO form I operon, was clearly up-regulated under autotrophic growth. It was also detected under benzoate growth indicating a clear co-expression with RubisCO form I. A total of 10 conserved unknown proteins and 3 unknown proteins were also indicated as up-regulated under autotrophic growth. A number of these unknown proteins such as RPA1114, RPA1243, RPA1244, RPA2786, RPA3309, RPA3568, and RPA4704 showed a strong expression pattern under autotrophic, stationary and benzoate growth. In those three metabolic states, carbon is a limiting factor in growth suggesting a potential function of carbon uptake or scavenging for the unknown proteins.

Photoheterotrophic log vs. photoheterotrophic stationary

The photoheterotrophic log phase was compared to stationary phase to determine proteins induced by the stress response of growth late into stationary phase. In total, 13 proteins were found to be down-regulated under the photoheterotrophic stationary state and 25 proteins were found to be up-regulated under the photoheterotrophic stationary phase (Table S4, 5th tab). Of the 13 proteins detected as down-regulated in stationary phase, five were also found down-regulated in the photoautotrophic state. These include RPA1542 and 1545 components of the protochlorophyllide reductase complex, RPA1975 a periplasmic mannitol binding protein, RPA2977 a ribonucleotide reductase, and two unknown proteins RPA2297 and RPA3011. While previous work in our laboratories in the *Escherichia coli* K12 strain comparing mid-log and stationary phase clearly indicated many proteins involved in protein turnover, folding and stress response up-regulated in the stationary phase (VerBerkmoes unpublished data), this was not found to be the case for *R. palustris*. Not a single protein annotated as involved in stress response, protein turnover or protein folding was detected up-regulated in the stationary phase. This may be due to a lack of knowledge of the function of proteins involved in this stress response process in *R. palustris.* Indeed, 9 conserved unknown proteins and 2 unknown proteins were identified as up-regulated in the stationary phase samples. Seven of these, RPA1114, RPA1243, RPA1244, RPA2786, RPA3309, RPA3568, and RPA4704, were also detected in the autotrophic phase as up-regulated. Thus, these proteins could be involved in attempting to assimilate carbon or the general stress response.

Photoheterotrophic succinate vs. Photoheterotrophic benzoate

The photoheterotrophic state with succinate as a carbon source was compared with the photoheterotrophic state with benzoate as a carbon source to determine proteins important to benzoate degradation. In total, 29 proteins were found to be clearly up-regulated under the photoheterotrophic succinate state and 44 proteins were found to be up-regulated under the photoheterotrophic benzoate state (Table S4, 6th tab). Only one unknown protein and 1 conserved unknown protein were found to be up-regulated in the photoheterotrophic succinate state. The photoheterotrophic succinate state had a greater

expression of many ABC transport systems periplasmic proteins such as RPA0580, a putative branched chain amino acid transporter, RPA3093, possible urea/short-chain binding protein, as well as RPA3725, RPA4019, RPA4029, RPA4648. The entire carbon monoxide dehydrogenase operon (RPA4666-4668) was found to be up-regulated under the photoheterotrophic succinate state. As expected, the photoheterotrophic benzoate state had many proteins expected to be involved in benzoate degradation up-regulated. The entire benzoate degradation regulon from RPA0650-RPA0662 was clearly up-regulated and not detected to any significant level under any other metabolic states. This is another clear example of the methodology detecting changes in expression in expected proteins (see highlighted proteins in Table S4, 6th tab). The only protein from this regulon not detected was the transcription factor RPA0663. A 4-hydroxybenzoyl-CoA reductase complex (RPA0670, RPA0671, *note RPA0672 was detected with only two peptides and thus not listed in table*) and the 4-hydroxybenzoyl-CoA ligase (RPA0669) were only confidently detected under benzoate growth. One protein of the associated transporter was detected under all conditions and the regulator was not detected at all. The RubisCO operon, including the cbbX protein, was confidently detected with high sequence coverage as was the case in photoautotrophic growth. In this case, the microbe is fixing carbon as a means of balancing its redox potential from the oxidation of benzoate. While not large-scale differences, some proteins of the pim operon (RPA3713, RPA3714, RPA3715, and RPA3717) were detected at elevated levels in comparison with the photoheterotrophic succinate state. These proteins are annotated as being involved in lipid metabolism but actually function at the bottom of the benzoate degradation pathway. Two unknown proteins (RPA1422 and RPA2786) were clearly up-regulated. RPA1422 was also found to be up-regulated under the nitrogen fixation state. RPA2786 was also found to be up-regulated under the photoautotrophic state and the stationary phase state suggesting possible unknown related function between the two states. Three conserved unknown proteins (RPA1777, RPA3309, and RPA4137) were up-regulated. RPA1777 was only confidently detected under benzoate growth except for a single two peptide identification under the LhaA mutant aerobic state. RPA3309 showed the similar pattern of detection as RPA2786 with high sequence coverage under the photoautotrophic

state and the stationary state. RPA4137 was detected under all anaerobic states but was not detected at all under the aerobic states; it had the highest sequence coverage under autotrophic and benzoate growth.

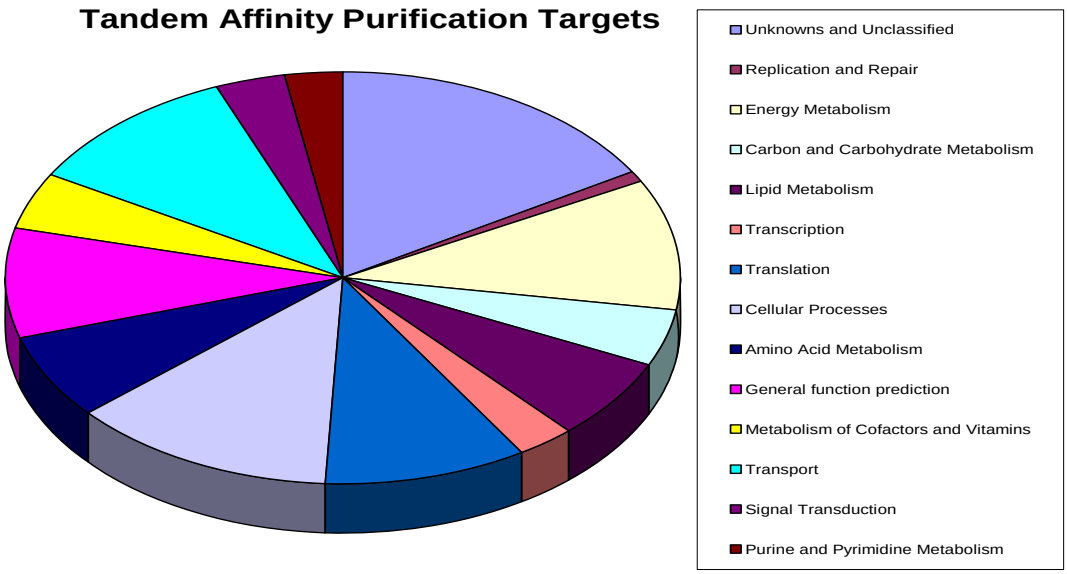
Tandem affinity targets

One goal of this project was to determine the abundant proteins expressed under the major metabolic states and provide targets for large-scale tandem affinity purification (TAP) of protein complexes (Puig, 2001; Buchanan, 2002; Ho, 2002; Gavin, 2002) thus integrating baseline proteome analysis with an ongoing large-scale analysis of protein complexes (Buchanan, 2002). The reason this is necessary is that the correct determination of high quality targets and the metabolic state to express such targets cannot be directly inferred from the genome. While a protein may be predicted to be involved in a complex from the genome annotation or association in an operon with other known proteins, its expression cannot be verified from the genome information. Furthermore, as illustrated above many proteins are only expressed at a high level under certain metabolic states. Thus, the protein target should be grown under the correct metabolic state to provide the most optimal conditions for affinity purification of the target and potential interacting proteins. Previous large-scale studies of protein complexes and protein-protein interactions have not taken this type of information into account (Ho, 2002; Gavin, 2002). We have determined through this study and other previous studies that many known proteins involved in protein complexes will pellet during centrifugation process. In the case of this study, we applied two high speed centrifugation steps. The first was for 1 hour at 100,000 g to create an enriched membrane fraction. The second was for 18 hours at 100,000 g to create a pellet and cleared fraction. We have found most known protein complexes to pellet during the first or second spin and the cleared fraction is nearly devoid of known protein complexes. We then extracted the entire dataset for those proteins with at least four unique peptides found in either the crude membrane, or pellet fraction in any given metabolic state. Table S5 contains the entire list of proteins, functional categories and a pie chart (2^{nd} tab) with the distribution of each. In total, 965 proteins were identified as potential targets for

affinity purification. The metabolic states and fraction each protein was found in can be accessed on the *R. palustris* proteome website (http://compbio.ornl.gov/rpal_proteome/analysis/) allowing for rapid look-up of targets in choosing the correct metabolic state for target growth. The distribution of TAP targets shows the exact same distribution as the identified proteome (Figure 5.4). Many known protein complexes are in the list including, ribosomal proteins, ATP synthase complex, succinyl-CoA synthetases, cytochrome C oxidase, pyruvate dehydrogenase, nitrogenase proteins, RubisCO, GroEL/GroES, and RNA polymerase subunits, as just a few examples. In addition, 107 conserved unknown and 50 unknown proteins were identified as high quality targets. These make excellent candidates in attempting to identify potential functions for unknown proteins. If they can be isolated with known protein complexes, their functions may be inferred.

Conclusions

In this study we have characterized the *R. palustris* proteome under 6 major metabolic states. We have confidently identified 1,664 proteins representing 34% of the predicted proteome. Over 300 proteins were identified as exhibiting large-scale differences between metabolic states many of these being conserved unknown or unknown proteins. The proteome analysis of a large number of metabolic states is clearly necessary to begin to understand how microbes change their proteome to adapt to the resources present. Since exact quantitation of a large number of metabolic states is not currently straightforward with "shotgun" proteomics, other methods of semi- quantitative approaches must be applied. This study presents a first step towards that goal and illustrates how this can be accomplished with semi-quantitative indicators of abundance such as percent sequence coverage and number of unique proteins. The clear up-regulation of proteins known to be involved in nitrogen fixation, autotrophic carbon assimilation and benzoate degradation illustrates the effectiveness of this technique. The conserved unknown and unknown proteins that were identified as up-regulated under given metabolic states make excellent targets for future studies since they may have essential functions under those states.



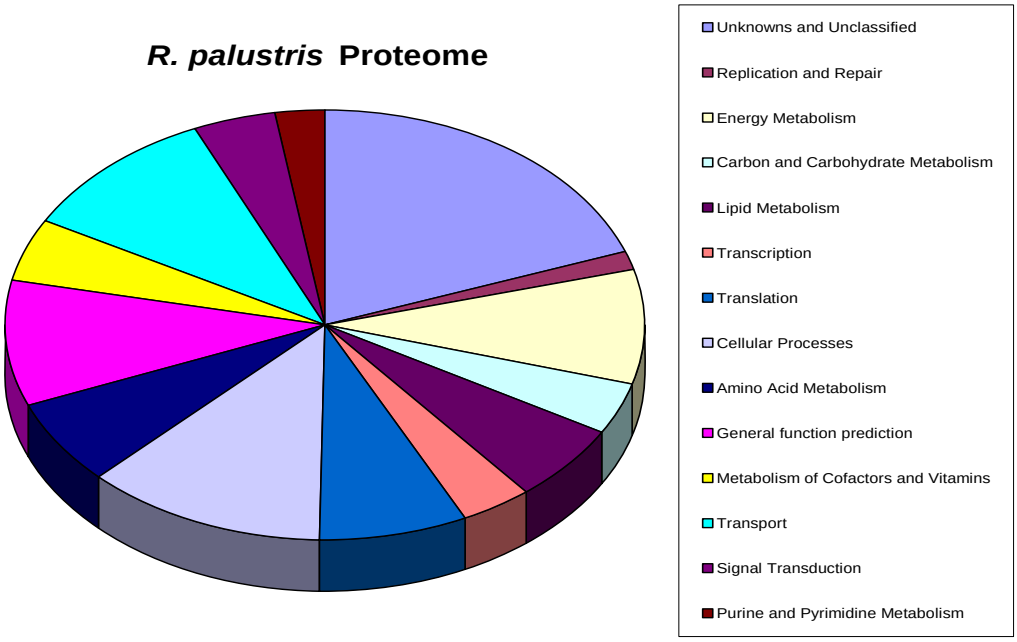


Figure 5.4: Functional categories of TAP targets and proteome.
Comparison between functional categories of tandem affinity purification targets and observed proteome.

We have demonstrated why proteome analysis of many metabolic states should be undertaken before large-scale analysis of protein complexes from a microbe. By creating a database of potential affinity targets and their expression patterns, logical choices for target expression can be made. It is absolutely necessary to express a target protein under the same conditions as the potential interacting partners are expressed, to have the optimal chance for success. If a target protein is expressed under a condition where the potential interacting proteins are not expressed then the potential for false negative results are guaranteed. This information cannot always be directly inferred from the genome annotation. In some cases, the predicted function of a target protein and thus the metabolic state it should be expressed under can be inferred from the genome annotation. However, this will not always be the case. This is clearly true for conserved hypothetical and hypothetical proteins where no information on potential function or metabolic state expression exists.

We are currently investigating the potential of proteomics to study microbial species directly from natural samples (Chapters 6 and 7). *R. palustris* is a perfect target for such studies since its genome has been sequenced, it is ubiquitous, and it can change its metabolic state to optimally survive with whatever source of energy and biological building blocks are present. By characterizing the proteome under the major metabolic states, we have observed the potential proteins which may be expressed in the community settings. More importantly, we have created a databank of MS/MS spectra from observed peptides, which can act much as synthetic peptides, for the verification of detection of proteins from *R. palustris* in the environment.

In this study, we have compared two search algorithms for processing MS/MS spectra. SEQUEST is considered by many to be the gold standard search algorithm for "shotgun" proteomics applications. DBDigger was recently developed at ORNL and provides more flexibility and speed than SEQUEST with the same accuracy and precision (Tabb, 2005). This was the first large-scale comparison of the two algorithms on a bacteria proteome dataset. The results clearly indicated that both algorithms gave very similar results adding confidence to the overall identifications as well as advancing the usefulness of the new search engine. The lack of open access to the results of large-scale

proteome datasets generated from "shotgun" proteomics is clearly a hindrance to future progress in this field (Carr, 2004; Pedrioli 2004). For this study, we created a completely open access website (ORNL *Rhodopseudomonas Palustris* Proteome Study Website) for the repository of all the results from the different search methods and search engines for direct download. We have made all identified MS/MS spectra directly accessible and have provided clear descriptions of all the data files. For the future directions of this endeavor, we hope to create a fully interactive website where researchers can access the datasets, re-search the datasets, re-filter the datasets and compare the datasets as they choose. While we were not able to accomplish this lofty goal for this study, we believe the steps we have taken are further than any other study to this point and bring us much closer to truly open access proteome results.

Chapter 6

Evaluations of "Shotgun" Proteomics for Characterizing the Complex Metaproteomes of Microbial Communities

Introduction

The previous chapters have demonstrated the application of "shotgun" proteomics to the characterization of microbial isolates grown in batch cultures. Microbial physiology, biochemistry, evolution and ecology have been studied in great detail due to the importance of microbes in industry, human health, agriculture and the environment. The greatest effort to date in molecular level studies of microbes has mostly focused on cultured species grown to large densities in laboratory settings. While this methodology of study is very useful in understanding the fundamental physiology and biochemistry of microbes, it does not truly study the microbes under their natural settings and many microbes cannot be isolated and cultured for these types of studies. Microbes in nature always exist in communities with other microbes, under a constant struggle for nutrients and space, where their existence is very often dependent upon metabolites from other microbes in the community for survival. By definition, microbial communities are composed of mixed cultures of interacting populations of microbes and their environments.

While these communities have been studied in the past by many different methodologies, they have not been studied in detail by system level molecular biology techniques such as genomics, transcriptomics, proteomics or metabolomics. While these techniques could offer system level information on the molecular level of the physiology and biochemistry of the community, they have been slow to develop for communities due to the complexity of natural microbial communities and the limits of the current technologies. Researchers have begun to explore genomic DNA-based analysis techniques for these microbial communities (Smalla and Sobecky, 2002; Tyson, 2004; Venter, 2004). This area of work is often characterized as "whole community genomics." Early work in this field has shown that the biggest technical challenges will be found in communities, which are very diverse. Indeed, deep sequencing of microbial communities

will probably be limited to simpler communities dominated by a few species in the near future. Recent research and discussion have suggested that microarrays may be used to study microbial communities at the transcriptome level, but have also highlighted the difficulty involved with diverse communities and the advantages of working in simpler communities (Wu, 2001; Zhou, 2003). Genome sequencing forms the core database for any systems level analysis of a living species or a community, but it only provides the blueprint for the metabolic possibilities a species or community possesses. Currently, microarray analysis provides the most detailed analysis of transcript levels, which relate to protein expression levels, but the measurement is based on the intermediate between DNA and protein. Proteins and proteomes are the actual functioning units of cells and communities of cells.

Over the last decade, the emerging field of proteomics has now evolved to the point of being able to provide detailed, diverse and precise information for proteomes for sequenced organisms from batch isolates (see Chapters 4 and 5). While not as fully-developed as the genomics and transcriptomics technologies, MS-based proteomics offers a diversity of techniques for the direct measurement, at a molecular level, of the essential and functional components of any living system, proteins. With the advent of community genome sequencing projects, it may now be possible to make direct proteome measurements on simple and stable microbial communities that have been sequenced at the whole genome level.

Our first efforts in this field were not for deep proteome coverage but for simple microbial identification. In VerBerkmoes et al. 2004, we examined the potential of "shotgun" proteomics to analyze a target species in a background of 3 microbial species and 1 plant species. We tested the capability of a common commercial MS-based "shotgun" proteomics platform for the detection of the target species (*E. coli*) at four different concentrations and four different time points of analysis. We also tested the effect of database size on positive identification of the four microbes used in this study by testing a small (13-species) database and a large (261-species) database. The results clearly indicated that this technology could easily identify the target species at 20% in the background mixture at 60, 120, 180 or 240 minutes analysis time with the small database.

The results also indicated that the target species could easily be identified at 20% or 6% but could not be identified at 0.6% or 0.06% in either a 240-minute analysis or a 30-hour analysis with the small database. Recent efforts, illustrated below, suggested that not only can identifications be made at less than 1% target species, but small coverage of the proteome could be achieved as well. The effects of the large database were severe on the target species where detection of unique peptides above background at any concentration used in this study was impossible, though the three other microbes used in this study were clearly identified above background when analyzed with the large database. These results were confirmed with the below test for *E. coli* while, again, the other species were not effected as much. The reason for this will be discussed. This initial study pointed to the potential application of this technology for microbial detection but highlighted many areas of needed research before the technology would be useful in real world samples.

The primary goal of this chapter is to demonstrate initial testing and developments we have done with simple mixtures of microbial species as simulants for microbial communities to obtain deep proteome characterization rather than just identification. These types of tests are absolutely necessary to determine the current state of proteomics technologies and to develop new technologies. It is more logical to work with known mixtures of microbes simulating microbial communities than to try and develop these technologies on undefined precious samples from real communities. Current proteomic techniques have not been shown to probe 'deep' into community systems to infer biochemical and physiological function; however, proteomics are predicted to be applied into this area (Sauer, 2003; Macarthur and Jacques, 2003; Casado, 2004).

The proteome analysis of any microbial community would be difficult with any current technology. The primary theoretical and practical concerns are:

1) The level of sequence information available on the community. Current proteomic methodologies are heavily reliant on existing genome sequence information. If a large number of microbes present in the community are unsequenced or not related to sequenced organism, current technology will be

very limited in its application. The best situation is if the actual environmental sample has been sequenced to the fullest extent possible.

2) The level of diversity and dynamic range associated with the species of interest in the community. Preliminary work in our laboratories (discussed below) has suggested practical limitations in studying a microbe, which is less than 1% in the community. The total number of species and their relative concentrations will be very important in the ability of current technologies to make any useful measurements of community proteomes.

3) The quantity of biomass available for study. While MS-based proteomic measurements continue to become more and more sensitive, practical applications of sample preparation and LC-MS/MS require at least 100 ug-1 mg of crude proteome starting material for effective analysis.

4) The level of interrelatedness and/or diversity at the base pair level amongst members of the same species in the community and between species in the community. Since MS-based proteomics relies on molecular level sequence dependent measurements on either enzymatically prepared peptides or whole proteins, the diversity at the amino acid level for individual peptides or proteins from the same species or between species will be very important.

When all of these facts are considered, it is clear that, with current technology, the direct measurement of complicated environmental samples such as a typical soil or water sample, which may contain tens of thousands of species at different concentrations with very little sequence information, will be difficult, if not impossible. A possible starting place for community proteomics is environmental samples, which may contain a much simpler microbial community of 10-100 species, but which is dominated by 5-10 microbes. It should be theoretically possible to extend current MS-based proteomic methodologies to make biologically relevant measurement from such microbial

communities. Natural communities that might be a promising area of initial research are extremophile niches. These communities are found in hostile environments such as very high salt, high or low pH, and very high or low temperatures. The communities are generally thought to contain a simpler mixture of species with a few predominant species due to the harsh living conditions. One such system, which theoretically fulfills all the requirements for an initial community proteome study (community must have sequence information available, be fairly simple in total number of species, and have plenty of biomass available), is the recently sequenced community populating the Acid Mine Drainage (AMD) at the Richmond Mine in Iron Mountain, CA (Tyson , 2004). The proteome characterization of this community is discussed in Chapter 7. This chapter focuses on methodology testing and development with an artificial mixture of microbial isolates made up of *Shewanella oneidensis* MR-1, *Escherichia coli* K-12, *Rhodopseudomonas palustris* CGA010, and *Saccharomyces cerevisiae*. We tested the effects of concentration and database size on the depth of coverage that could be obtained with current "shotgun" proteomics techniques. Hopefully, the methodologies used here and the lessons learned can be applied in the characterization of real microbial communities.

Materials and Methods

Chemicals and reagents

All salts, DTT, trifluoroacetic acid (TFA), and guanidine used in this work were obtained from Sigma Chemical Co. (St. Louis, MO). Protein concentrations were determined with BCA reagents from Pierce Chemical Co. (Rockford, IL). Modified sequencing grade trypsin, from Promega (Madison, WI), was used for all protein digestion reactions. The water and acetonitrile used in all sample clean up and HPLC applications was HPLC grade from Burdick & Jackson (Muskegon, MI) and the 98% formic acid used in these applications was purchased from EM Science (an affiliate of MERCK KgaA, Darmstadt, Germany).

Cell growth and production of protein fractions

The four microbes used in this study (*E. coli* K-12, *R. palustris* CGA010, *S. cerevisiae*, and *S. oneidensis* MR-1) were all grown individually to mid-log phase and mixed after cell harvesting at appropriate concentrations based on wet-cell weight. Table 6.1 illustrates the different mixture concentrations used in this study. *R. palustris* was designated as the target species. Cell extracts were prepared as follows: cells were harvested by centrifugation, washed twice with ice-cold wash buffer (50 mM Tris buffer [pH 7.5] with 10 mM EDTA) and resuspended in ice-cold wash buffer. Cells were then mixed at the indicated concentrations in Table 6.1 by wet-cell weight. Cells were then lysed with sonication and unbroken cells were removed with low-speed centrifugation (5,000 g x 15 min). Two proteome fractions were created from the cellular extract by ultracentrifugation (100,000 g for 1 hour). The soluble fraction is referred to as crude soluble; the pellet is referred to as the membrane fraction. The membrane fraction was washed once with wash buffer and pelleted again by ultracentrifugation. The pellet was then resuspended in wash buffer with the aid of gentle sonication. Both proteome fractions were quantified with BCA analysis, aliquoted and frozen at -80 ^{0}C until digestion.

Digestion of proteome fractions

Both proteome fractions from all concentrations were processed in exactly the same manner. Briefly, 5 mg of each proteome fraction was diluted in 6 M guanidine and 10 mM DTT then heated at 60°C for 1 hour. The guanidine and DTT were diluted 6-fold with 50 mM Tris/10 mM $CaCl_2$ (pH 7.8) and sequencing grade trypsin was added at 1:100 (wt/wt). The digestions were run with gentle shaking at 37°C for 18 hours, followed by a second addition of trypsin at 1:100 and additional 5-hour incubation. The samples were then treated with 20 mM DTT for 1 hour at 37°C as a final reduction step. Samples were immediately desalted with Sep-Pak Plus C_{18} solid phase extraction (Waters, Milford, MA). All samples were concentrated and solvent exchanged into 0.1% FA in water by centrifugal evaporation to ~10 μg/μL starting material, filtered, aliquoted and frozen at -80°C until LC-MS/MS analysis.

Table 6.1: Concentrations of 4 test microbes in artificial mixtures.

Organism	Mix 1	Mix 2	Mix 3	Mix 4	Mix 5
E. coli	25%	32%	33%	33%	33.3%
S. cerevisiae	25%	32%	33%	33%	33.3%
S. oneidensis	25%	32%	33%	33%	33.3%
R. palustris	25%	5%	1%	0.2%	0%

LC-MS/MS analysis

All samples were analyzed by 2-dimensional MudPIT LC/LC-MS/MS (described in Chapter 2). An UltiMate HPLC (LC Packings, a division of Dionex, San Francisco, CA) was used for the separation process. The pump provided a flow rate of ~100 µL/min, which was split pre-column to provide an approximate flow of ~200-300 nL/min at the nanospray tip. The split-phase columns were constructed as follows: the back column was packed with approximately 3.5 cm of strong cation exchange material (Luna SCX 5 µm 100A Phenomenex, Torrance, CA) into a 100 µm fused silica via a pressure cell followed by 3.5 cm of C-18 reverse phase (RP) material (Aqua C-18 5 µm 200A Phenomenex). The sample was then loaded off-line onto the dual phase column. For all samples, ~500 µg of protein was loaded onto the back dual phase column. The loaded RP-SCX column was then positioned on the instrument behind a ~15 cm C_{18} RP column (Jupiter C18 5 µm 300A Phenomenex) also packed via pressure cell into Pico Frit tip (100 µm with 15 µm tip New Objective, Woburn, MA). The entire column system was positioned into the nanospray source (Thermo Finnigan) on either the LCQ or LTQ mass spectrometers.

All samples were analyzed via a 24-hr 12-step MudPIT analysis consisting of increasing concentration (0-500 mM) salt pulses of ammonium acetate followed by 2-hour reverse phase gradients from 100% aqueous solvent (95% H_2O/ 5% ACN/ 0.1% formic acid) to 50% organic solvent (30% H_2O/ 70% ACN/ 0.1% formic acid). During the entire chromatographic process, the LCQ mass spectrometer operated in a data-dependent MS/MS mode detailed below. The chromatographic methods and HPLC columns were virtually identical for all analyses. The LC-MS system was fully automated and under direct control of the Xcalibur software system (Thermo Finnigan).

The LCQ MS was operated with the following parameters: nanospray voltage (2.4 kV), heated capillary temp 200^{0}C, full scan m/z range (400-1700). The data-dependent MS/MS mode was set up with the following parameters: 4 MS/MS spectra for every full scan, 5 microscans averaged for full scans and MS/MS scans, 5 m/z isolation widths for MS/MS isolations and 35% collision energy for collision-induced dissociation. To

prevent repetitive analysis of the same abundant peptides, dynamic exclusion was enabled with a repeat count of 1 and an exclusion duration of 1 min.

Data analysis

The resultant LC-MS/MS files were all processed as follows. The MS/MS spectra from all files were searched with SEQUEST (Thermo Finnigan) against three databases DB4, DB13 and DB large. DB4 contained the four species contained in the mixtures, *E. coli* K-12, *R. palustris* CGA010, *S. cerevisiae*, and *S. oneidensis* MR-1, with a total of 20,595 proteins. DB13 contained the four species contained in the mixture plus 8 other bacteria and 1 plant (*Arabidopsis thaliana*), with a total of 84,606 proteins. DB 261 contained the four species in the mixture plus most sequenced microbes to date and both sequenced plant species (*Arabidopsis thaliana* and *Oryza sativa*). The different *E. coli* species that have been sequenced were removed from the database. This database had a total of 261 species and 1,011,612 protein entries. The alternative databases were used to test the ability to detect the species contained in the database against the background of all other sequenced species. All resultant output files from SEQUEST were filtered by DTASelect (Tabb, 2002) with the following parameters: SEQUEST, delCN of at least 0.08 and cross-correlation scores (Xcorrs) of at least 1.8 (+1), 2.5 (+2) and 3.5 (+3). The filtered DTASelect files were then extracted with in-house developed Perl scripts to obtain the numbers of unique and non-unique peptides presented below.

Results

The initial experiments for this study have focused on testing current 2D-LC-MS/MS methodologies to analyze an artificial 4-microbe mixture with *E. coli, R. palustris, S. cerevisiae* and *S. oneidensis*. Two aspects of community proteomics were analyzed in this initial study: the first was to determine at what level functionally meaningful results could be obtained from a target species *(R. palustris*) whose concentration was decreased over a range of concentrations from 25% to 0.2%. All experiments were conducted with current 2D-LC-MS/MS technologies. The second was to test the effects of database size on the identification of unique peptides from the four

microbial species in the sample. Though peptides may be identified above a certain filtering level, in a large database of many proteins (and especially if that database contains many related species with many conserved proteins), there will exist a large number of replicate tryptic peptides with the exact same sequence. If a specific peptide is found in multiple species, then it cannot be used as a unique peptide for identification purposes. This issue was handled with the DTASelect algorithm since it labels all unique peptide identifications with an asterisk; non-unique peptides are listed under all possible originating proteins and do not contain an asterisk. Unique peptides are defined as those peptides from a given protein that are unique to a specific species in a given database. Non-unique peptides refer to those peptides that were not unique across a given database, meaning the same exact peptide is found in multiple proteins. Non-unique cannot be used for confident assignment of protein identification. We used in-house designed Perl scripts to extract the number of uniquely identified peptides and proteins from any given analysis and database search.

The first experimental test was to test the ability of 2D LC-MS/MS on an ES-quadrupole ion trap to detect peptides and proteins from *R. palustris* at decreasing concentrations from 25% to 0% in the mixture of the other three microbes which were kept at the same concentration. Figure 6.1 illustrates these results against the DB4 database. At the 25% concentration, all bacterial species were readily identified with at least 2,000 unique peptides. *S. oneidensis* was the highest with nearly 3,500 peptides, while *E. coli* was found at nearly 3,200 peptides, and *R. palustris* at nearly 2,000 peptides. The differences in the number of unique peptides are most likely due to differential lyses and slight variations in mixing concentrations (this cannot be exactly determined since it is just an estimation on wet cell pellet weight). The massive difference in the number of *S. cerevisiae* is not due to small errors in mixing, but rather due to differential lysing. It is known that *S. cerevisiae* does not lyse as well by sonication as bacterial samples but we wanted to test the results with our current proteomics pipeline (Chapter 2). We are currently investigating alternative lyses procedures for mixtures to attempt to develop optimal lyses processes for the multitude of cell types that might be encountered in the natural environment.

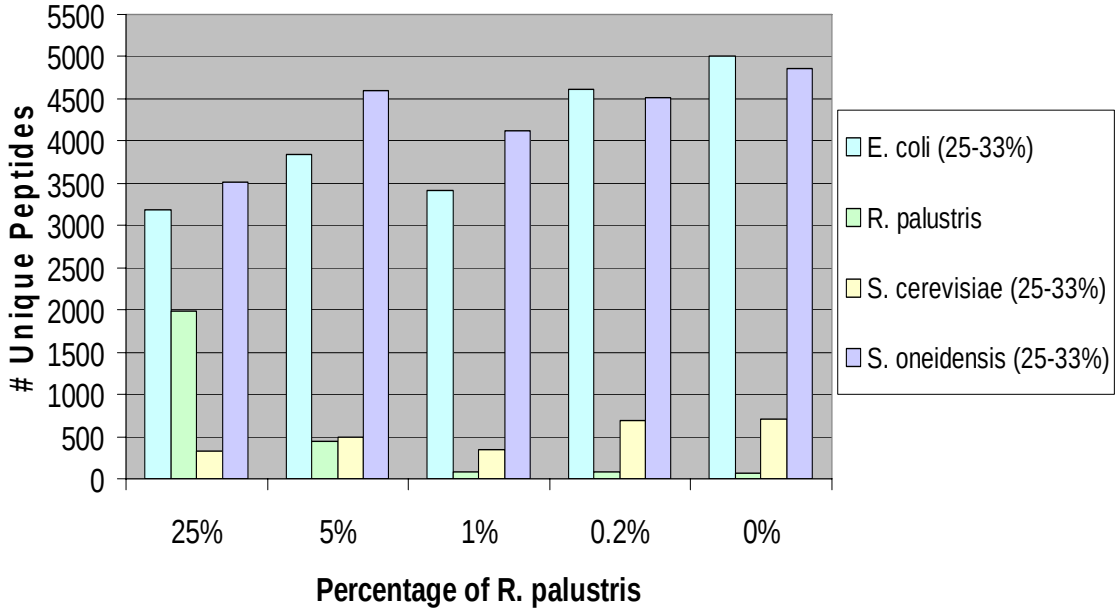


Figure 6.1: Number of unique peptides identified with the DB4 database.
Illustrates the number of unique peptide identifications against the DB4 database. At the 25% concentration, all bacterial species were readily identified with at least 2,000 unique peptides. As the *R. palustris* target is diluted, its number of unique peptides identified rapidly drops and peptides from the other microbes are detected at a higher level.

As expected, the number of detected peptides from *R. palustris* dropped dramatically from the 25% concentration to the lower concentrations with ~500 peptides detected at the 5% level and less than one hundred peptides at the 1% and the 0.2%. The number of detected peptides at the 1% and 0.2% could not be distinguished from the level of false positive hits to *R. palustris* at the 0% level, suggesting no useful results can be obtained at less than 1% (this was not actually the case, as described below). We then compared the identified proteins' functional categories from *R. palustris* to determine if we were just detecting abundant ribosomal proteins at the lower percentages or if we were actually still getting a representative overview of the proteome. Figure 6.2 (top left, 1,695 total proteins) compares the functional categories of the proteins identified from the global characterization of the *R. palustris* proteome (see chapter 5) with the functional categories of the identified proteins at 25% (610 total proteins), 5% (227 total proteins) and 1% (70 total proteins). Interestingly, while the numbers of proteins dropped significantly, the functional categories were not significantly affected, even at the 1% level. This suggested that, while deep proteome measurements were not being made at the lower percentage, the methodology was still able to take a representative snapshot of the proteome.

The next question addressed was to determine if any of the peptides/proteins identified at the 1% and 0.2% were actually real peptide and protein identifications or simply false positives, as with the 0%. This was done by manual validation of the identifications. Table 6.2 illustrates the number of peptides identified from an abundant protein from *R. palustris*, the H subunit of the photosynthetic reaction center complex (PuhA). This protein was found to be detected under every growth state in *R. palustris* global proteome characterization (Chapter 5) with high sequence coverage, high number numbers of peptides and a high repeat count suggesting high abundance (Liu, 2004). Thus, it was a good candidate to determine if identifications were made at low percentages since we had sequencing spectra from virtually all potential detected peptides from this protein, which can act like synthetic peptides for MS/MS verification (note the same peptide sequence always gave the exact same MS/MS spectra under similar CAD energy regimes). Table 6.2 indicates that a unique peptide was identified from this

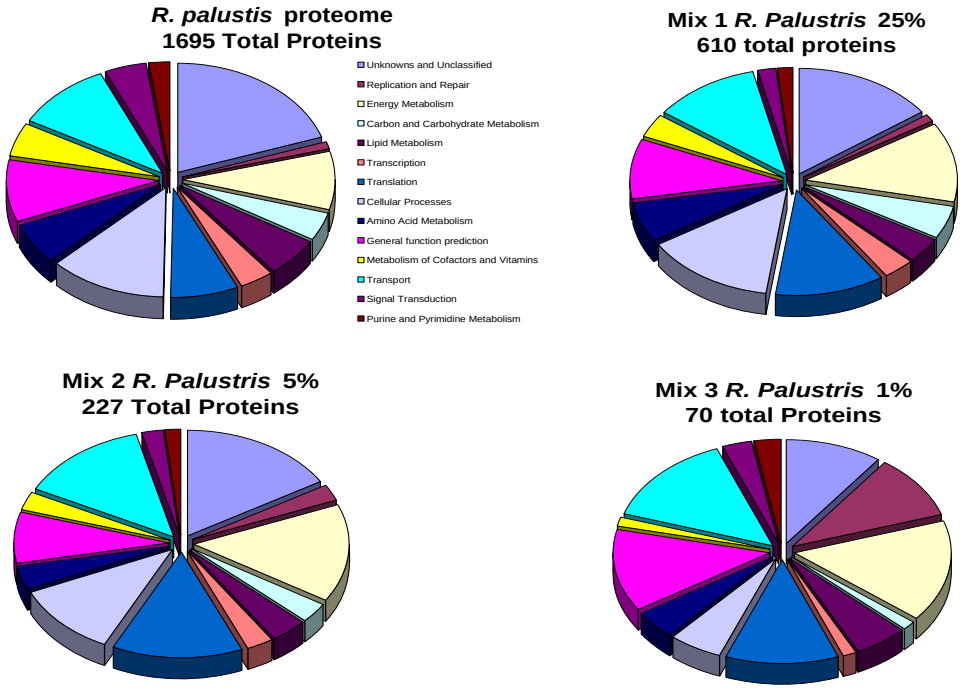


Figure 6.2: Functional categories of *R. palustris* proteins at different concentrations. Compares the functional categories of the proteins identified from the global characterization of the *R. palustris* proteome (1,695 total proteins, top left) with the functional categories of the identified proteins at 25% (610 total proteins, top right), 5% (227 total proteins, bottom left) and 1% (70 total proteins, bottom right).

Table 6.2: Identified peptides from PuhA at each concentration of *R. palustris.*

Mixture	% Coverage	Protein		
Mix 1	**44.70%**	**puhA H subunit of photosynthetic reaction center complex**		
rpal:RPA1548	Charge State	Xcorr	DelCN	Peptide
*	3	4.1787	0.5624	K.TVPSTSNDRPNVALTPAAPWPGAPFVPTGNPFADGVGPGSYAQR.A
*	3	5.8592	0.4355	R.ADVPELGLDNLPIIVPLR.A
*	2	5.6884	0.5109	R.ADVPELGLDNLPIIVPLR.A
*	2	3.788	0.4484	R.ADVPELGLDNLPIIVPLRAAK.G
*	1	2.7334	0.2724	R.YLEVEVAK.S
*	2	4.0343	0.5105	R.VLLPVPFALINDPFGK.V
*	3	5.8444	0.5249	R.VLLPVPFALINDPFGK.V
*	2	3.2296	0.6166	R.VLLPVPFALINDPFGKVSVDAIR.G
*	2	4.5225	0.5718	K.VSVDAIRGDQFAGVPTTSKGDQVSK.L
*	3	3.8473	0.3707	K.VSVDAIRGDQFAGVPTTSKGDQVSK.L
Mix 2	**42.70%**	**puhA H subunit of photosynthetic reaction center complex**		
rpal:RPA1548	Charge State	Xcorr	DelCN	Peptide
*	2	2.5409	0.3636	K.IGVPAPPDPK.T
*	3	3.5846	0.4472	K.TVPSTSNDRPNVALTPAAPWPGAPFVPTGNPFADGVGPGSYAQR.A
*	3	5.8143	0.4663	R.ADVPELGLDNLPIIVPLR.A
*	2	5.6008	0.4809	R.ADVPELGLDNLPIIVPLR.A
*	2	4.1629	0.369	R.ADVPELGLDNLPIIVPLRAAK.G
*	2	2.9884	0.485	R.VLLPVPFALINDPFGK.V
*	2	3.7545	0.4465	R.GDQFAGVPTTSKGDQVSK.L
Mix 3	**17.30%**	**puhA H subunit of photosynthetic reaction center complex**		
rpal:RPA1548	Charge State	Xcorr	DelCN	Peptide
*	3	5.1517	0.583	K.TVPSTSNDRPNVALTPAAPWPGAPFVPTGNPFADGVGPGSYAQR.A
Mix 4	**7.10%**	**puhA H subunit of photosynthetic reaction center complex**		
rpal:RPA1548	Charge State	Xcorr	DelCN	Peptide
*	2	4.4881	0.4467	R.ADVPELGLDNLPIIVPLR.A
Mix 5	**N/A**			

protein at the 1% and the 0.2% levels, with good cross-correlation scores. Furthermore, both peptides were identified at the 25%, 5%, and in the global proteome analysis, allowing for direct comparisons of the resultant MS/MS spectra.

Figure 6.3 illustrates the MS/MS spectra for the identified peptide at 25% and 1% and Figure 6.4 illustrates the MS/MS spectra for the identified peptide at the 25% and the 0.2%. Clearly, these MS/MS spectra are almost identical, indicating true identifications can be made at the 1% and 0.2% level. Furthermore, the identification of this protein also tells us that *R. palustris* is actively producing a protein involved in a metabolic function that is photosynthesis. This illustrates why it is useful, if possible, to have global characterizations of a microbial proteome isolate before attempting to characterize that microbial proteome in the environment. This is because the MS/MS spectra obtained from the isolates can be used as virtual synthetic peptide MS/MS spectra to verify low level identifications that might be made by only one or two peptides from environmental samples.

The next parameter investigated was the effects of database size. The reason for this was that research planned for the future will include attempting to characterize *R. palustris* in unknown mixtures of other microbes such as soil and water samples. Could a microbe, or a mixture of microbes, be characterized against a database of all other sequenced microbial species? Would there be enough unique peptides identified to confidently tell the microbe of interest apart from an unknown background of other species? Could useful measurements be made in large mixtures of microbes against large databases of potential microbes? While the simulations we employed did not lead to the final answers, they did make initial steps towards solving these problems and questions. To test this, we compared the levels of unique peptide identifications against a smaller database of just 13 species (DB13) and against a very larger database of 261 species (DB 261). The same proteome analysis shown in Figure 6.1 was searched against the DB13 database and the results are shown in Figure 6.5. The small database search did not have major effects on the results. The same basic trends of identified unique peptides were seen with ~250-400 peptides uniquely matching peptides from the other databases. It should be noted this was a conglomerate of all peptides matching to the other databases.

MS/MS at 25% and 1% *R. palustris*

TVPSTSNDRPNVALTPAAPWPGAPFVPTGNPFADGVGPGSYAQR

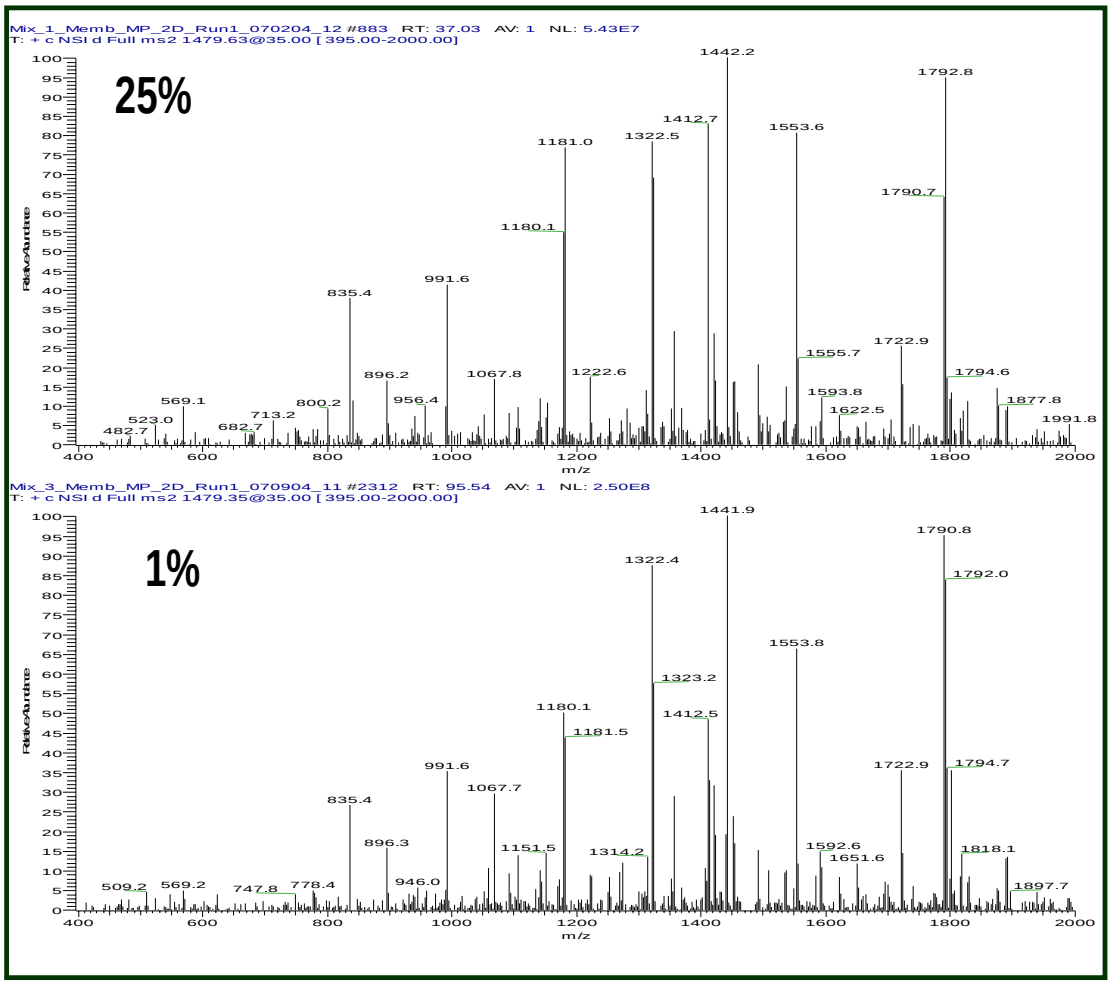


Figure 6.3: Diagnostic peptide identified from PuhA at 1% *R. palustris*.
Compares the MS/MS spectra of diagnostic peptide from PuhA detected at 25% and 1% *R. palustris* concentration.

MS/MS at 25% and 0.2% *R. palustris*

ADVPELGLDNLPIIVPLR

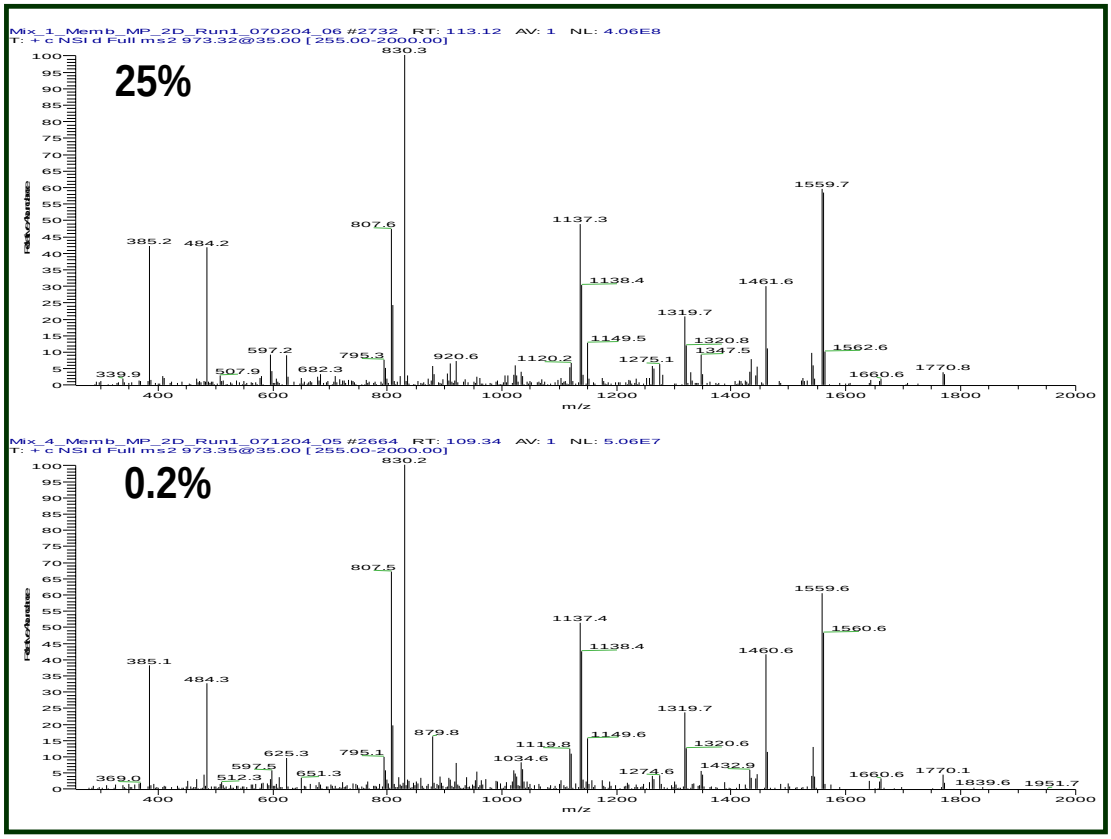


Figure 6.4: Diagnostic peptide identified from PuhA at 0.2% *R. palustris*. Compares the MS/MS spectra of diagnostic peptide from PuhA detected at 25% and 0.2% *R. palustris* concentration.

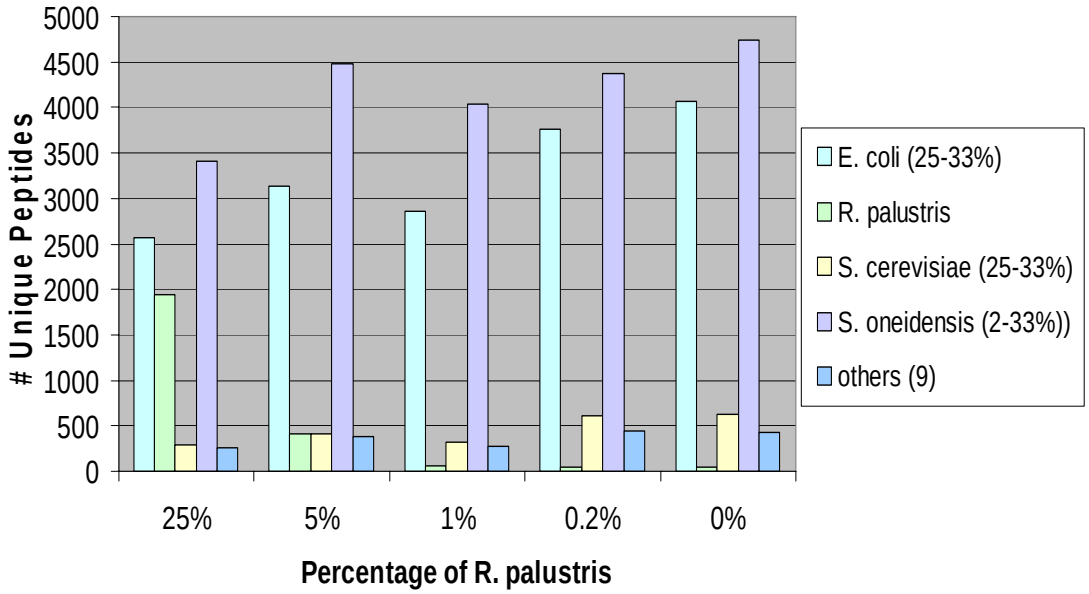


Figure 6.5: Number of unique peptides identified with the DB13 database. Illustrates the number of unique peptide identifications against the DB13 database. At the 25% concentration, all bacterial species were readily identified with at least 2,000 unique peptides. The same trends are seen as in Figure 6.1, with slightly lower peptide values showing that the small database does not have drastic effects on the # of unique peptide identifications.

It should be noted that this was a conglomerate of all peptides matching to the other 9 species in the database, with no more than 30-50 peptides from any given species. Clearly, the identifications of *R. palustris* at 1.0% and 0.2% were not above the average number of false positive hits to background species.

The dataset from just the Mixture 1 (25% all species) was then searched against the large database (DB 261). The results are shown in Figure 6.6 for the DB4 database search, the DB13 database search and the DB261 database search. Clearly, a major difference can now be seen. The number of unique peptides from *R. palustris*, *S. oneidensis* and *S. cerevisiae* were not affected in a major way. But the number of unique peptides from *E. coli* went from ~3,100 peptides with DB4 to ~2,550 peptides with DB13 to less than two hundred peptides with DB261. This was most likely due to the fact that a large number of closely related microbes to *E. coli* had been sequenced in comparison with the other three microbes in this study. These results illustrated two major points: 1) in the planned future experiments, it might be feasible to attempt to analyze *R. palustris* in a mixture of many other species against a very large database, assuming the concentration of *R. palustris* is relatively high; and 2) the ability to do this would highly depend on the species type and the database used. Without taking into account the uniqueness of species against the database one would be searching, it would be possible to conclude that the species was not present, even though it was present at high concentrations as was the case in this example with *E. coli*. Clearly, the best option would be to obtain whole genome sequencing of the environmental sample of interest if at all possible, as shown in Chapter 7.

Conclusions

This chapter demonstrated initial testing of proteome analysis of simple microbial mixtures. Initial studies have focused on evaluating the standard 24-hour 2D-LC-MS/MS experiment for the analysis of 4-component artificial microbial mixtures. *R. palustris* was designated as the target species, and its concentration in the samples were set to 20%, 5%, 1%, 0.2% and 0%, while all other species were kept constant. Significant *R. palustris* proteome measurements could be made at 25% and 5%, with 1% borderline. A

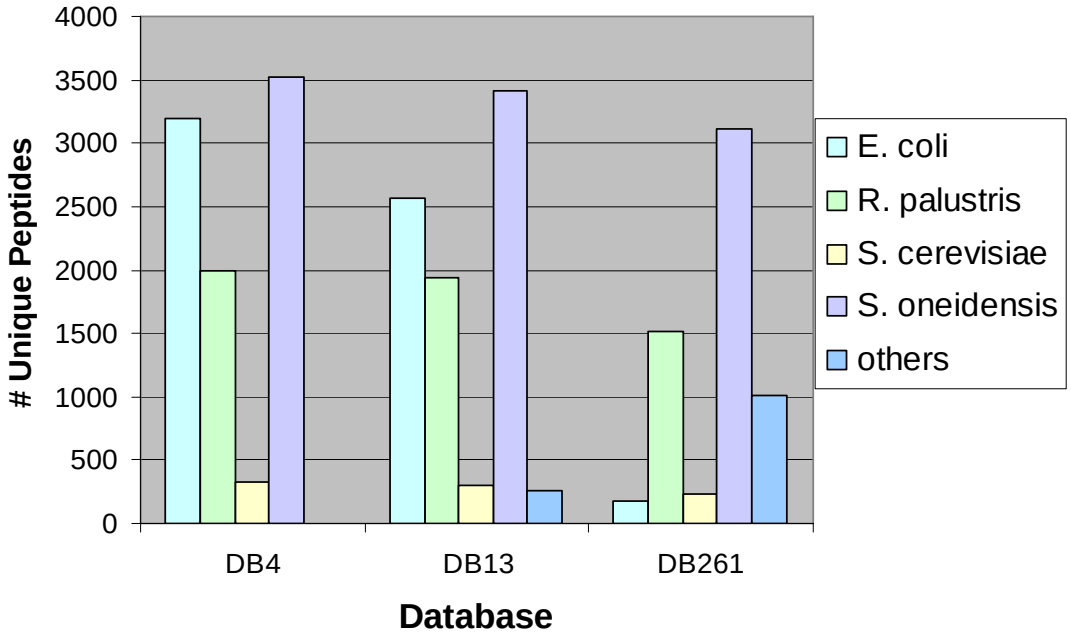


Figure 6.6: # of unique peptides identified from Mix 1 against all three databases. Illustrates the number of unique peptide identifications against all three databases with Mix 1 (25% of each microbe). While the # of peptide identifications was not severely affected by the increase in database size for *R. palustris*, *S. oneidensis* and *S. cerevisiae*, the increase database size dramatically affected *E. coli*.

few peptides could be detected confidently at the 0.2% level. Analysis of the expressed functional categories demonstrated that even at the 1% level an un-biased snapshot of the proteome was taken albeit at a very shallow level.

These initial tests illustrated that significant research development would be necessary to apply community proteomics to complex microbial mixtures where many components may be present at less than 5%. One potential improvement would be the implementations of the linear ion trap recently introduced (Schwartz, 2002). This instrument provides much better scan speed, dynamic range and sensitivity, as discussed in Chapter 2 and illustrated in Chapter 7. While this may help to identify more proteins at the 25% and 5%, it most likely would not allow for the needed increase in dynamic range necessary for detailed analysis of proteomes at less than 5%. A potential alternative that is currently being explored is the development of three-dimensional separation platform for community proteomes. The difficulty in creating such a platform would be the robust coupling of all three dimensions of separation. Another difficulty would be that peptides have a limited number of chemical properties to exploit in the separation process.

Current research has explored the use of RP-SCX-RP separations of peptides. While not truly orthogonal, this was straightforward in the coupling processes working around the first problem, but had shown serious limitations with the second problem. A second methodology being explored is the initial separation of the intact proteins by SDS-PAGE, followed by cutting the entire gel into district bands and in-gel digestions of the bands. The peptides were then eluted and analyzed by SCX-RP-ES-MS/MS. This showed great potential of working around the second problem by first separating the intact proteins, but we had found the coupling process to the second dimension, as well as the large in-gel digestions, to be experimentally difficult. The development of an effective 3D separation for complex microbial mixtures is still an area of active research.

While the use of additional separation space may increase the dynamic range, the best potential would be to increase the dynamic range of the mass spectrometers employed. While the quadrupole ion trap and the linear ion trap have very good dynamic range in the MS/MS mode, the dynamic range in full scan, where peptides are picked for

MS/MS, is not so good. The clear solution to this would be the use of the new linear ion trap FT-ICR which has recently become available (Syka, 2004). The proposed platform would include the coupling of a high resolution three-dimensional separation with the high dynamic range LTQ-FT-ICR (Figure 6.7).

We also explored the effects of database size on unique peptide identifications. While smaller databases would not have a dramatic effect on the number of identified peptides, larger databases would diminish the identification of unique peptides for any given species, but species such as *E. coli*, with many closely-related species in the database, are dramatically affected. The correct choice of database components will be of primary importance when working in unknown microbial communities. Clearly, the best situation would be to have the community of interest whole genome sequence.

To our knowledge, this was the first example of testing current "shotgun" proteomics techniques to characterize mixtures of microbial species. Quality results were obtained for the bacteria of a simple mixture of microbes, where all components were at equal concentrations. But the proteome coverage severely diminished as the target's species concentration was decreased. Currently, we are investigating alternatives in separation methodologies, MS methodologies, and cell lyses techniques. We are also investigating the potential for differential, or comparative, proteomics in mixtures by mixing *R. palustris* grown under different metabolic states into the 25% mixture with the other microbes. The goal here is to determine if the metabolic state of *R. palustris* can be accurately determined in a mixture of microbial species. Initial results from these studies are very promising. While significant progress has been made through this work and the work on a real microbial community presented in Chapter 7, clearly much research and development is needed to develop a robust and reproducible platform for the characterization of a diversity of microbial communities.

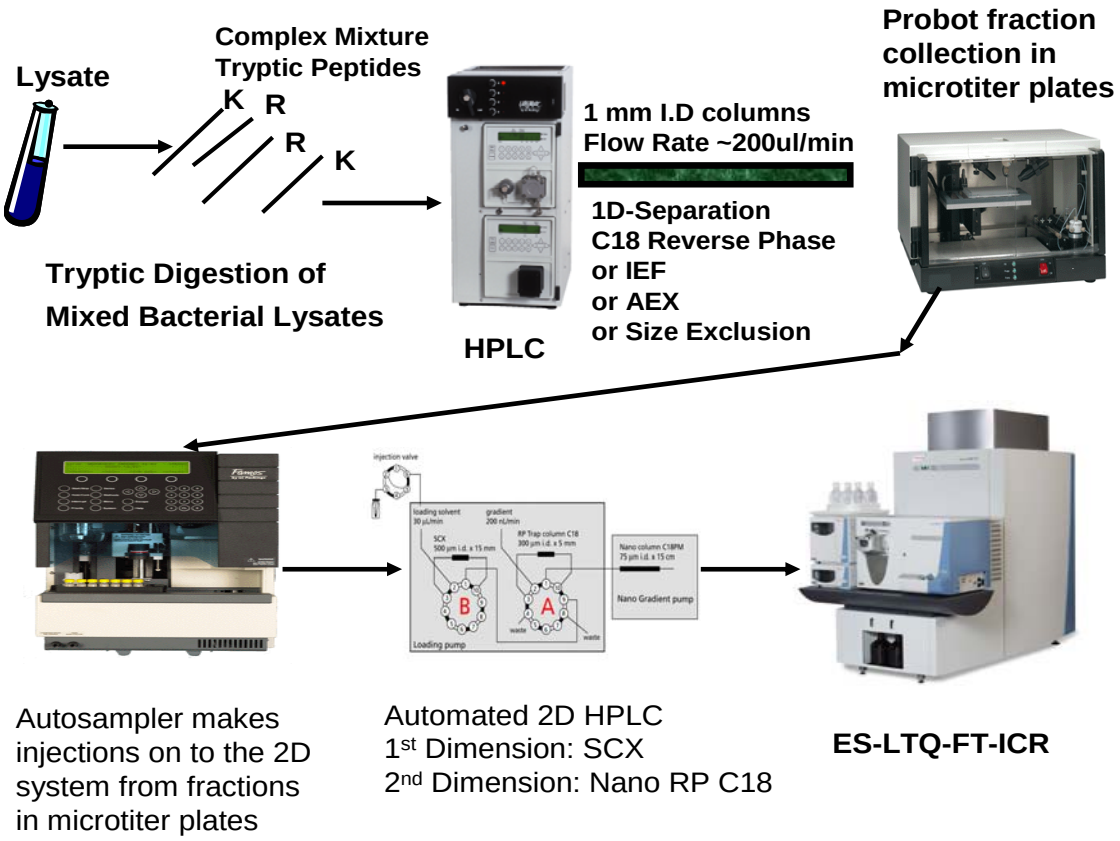


Figure 6.7: A potential solution for achieving high dynamic range measurements of microbial proteome mixtures.
The coupling of a high resolving three-dimensional separation with the high dynamic range potential of the LTQ-FT-MS instrument may allow for deep proteome characterization of microbial species less than 5% in a mixture, with more abundant microbial species.

Chapter 7

Mass Spectrometry-Based Proteome Analysis of the Acid Mine Drainage Community

Some of the data presented below has been submitted as Rachna J. Ram, Nathan C. VerBerkmoes, Michael P. Thelen, Gene W. Tyson, Brett J. Baker, Robert C. Blake II, Manesh Shah, Robert L. Hettich, and Jillian F. Banfield. Community proteomics of a natural microbial biofilm. *Science* (2004), in review. ***All MS sample preparation, experiments and data analysis were performed by Nathan C. VerBerkmoes.***

Complete datasets for proteomic analyses and other supplementary materials are available on the web site http://compbio.ornl.gov/biofilm_amd/

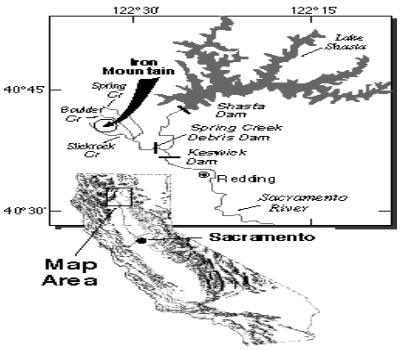
Introduction

Microbial communities play key roles in the Earth's biogeochemical cycles. Our knowledge of the structure and function of these communities is limited because analyses of microbial physiology and genetics have been largely confined to studies of organisms from the few lineages for which cultivation conditions have been determined. An additional limitation of pure culture-based studies is that potentially critical community and environmental interactions are not sampled. Recent acquisition of genomic data directly from natural samples has begun to reveal the genetic potential of communities (Tyson, 2004) and environments (Venter, 2004). Typically, more than 40% of the recovered genes in any genome are hypothetical, meaning that their predicted products share no significant sequence similarity to characterized proteins.

Acid Mine Drainage (AMD) has become a serious environmental problem since the advent of modern mining technologies. During deep mining projects, large quantities of the sulfide mineral in most rocks (pyrite FeS_2) become exposed to weathering and erosion by air and water. Mining increases the surface area of the sulfide ores exposed to air and water, thus increasing the acid generation. In mining areas where the rocks have very low buffering capacity, the buildup of acid and heavy metals in the water can become so great that the streams are actually toxic acidic heavy metal solutions called AMD. The generation of AMD sites can theoretically occur by just geochemical events, but it is hypothesized that the unique microbial communities found to populate these harsh environments may actually be greatly increasing the generation of AMD (reviewed

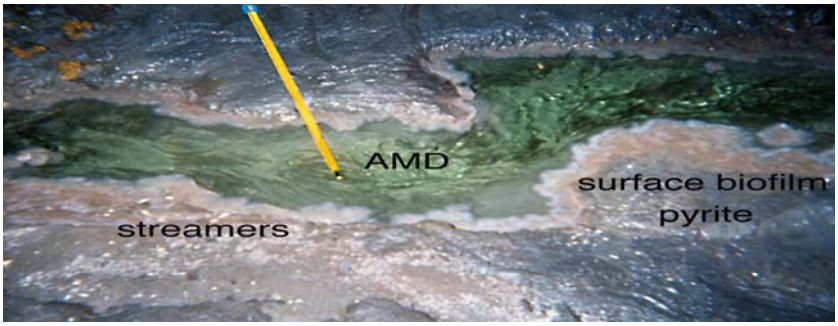
by Baker and Banfield, 2003). Thus, a greater understanding of the basic biology of the microbial communities populating such sites may allow for methods to slow the AMD process down or to apply such process in industrial applications. The microbial communities populating the AMD site at the Iron Mountain Mine in California is one of the most widely studied AMD communities to date (Schrenk, 1998; Edwards, 2000; Bond, 2000; Baker and Banfield, 2003). The location and pictures of general AMD biofilms can be found in Figure 7.1. While all the communities studied to date live in either very acidic conditions (0-1 pH) or mildly acidic conditions (1-3 pH), they all seem to be dominated by a few primary archaea or bacteria. The surface biofilm growing directly on the AMD stream (Figure 7.1) has been studied in great detail and several samples were collected from various places in the Richmond mine over the last few years. One of these samples was used for first attempt at community genome sequencing (Tyson, 2004) and location of its collection is near the entrance of the mine as illustrated in Figure 7.2. The random shotgun sequencing of the acidophilic biofilm allowed for near complete reconstruction of the genomes from two dominant organisms designated *Leptospirillum* group II and group III and *Ferroplasma* type II and type III and partial recovery of three other minor genomes. The remaining samples, while not exactly the same as the sample that was sequenced, provided excellent starting material for initial studies into methodology development for integrated genomic and proteome studies of microbial communities. This study is the first combination of cultivation-independent genomic and proteomic analyses to validate predicted genes, determine relative abundance and cellular localization of expressed proteins, and provide clues to protein function. The approach also enabled identification of the major investments of cellular resources and the physiological challenges faced by a self-sustaining, chemoautotrophic microbial community. The results from these studies, when coupled with the large amount of ecological and genomic information already available, will provide for even greater insight into the biology of the AMD community.

The Acid Mine Drainage Site at Iron Mountain Mine, CA



The Location

The AMD streams



The AMD streams enhanced

Figure 7.1: The acid mine drainage community of Iron Mountain, California.
Top left panel shows the location of the mine in California. Top right panel shows an AMD biofilm on top of an AMD stream. The bottom panel shows a close up look at an AMD stream.
Figure courtesy of Dr. Jillian Banfield.

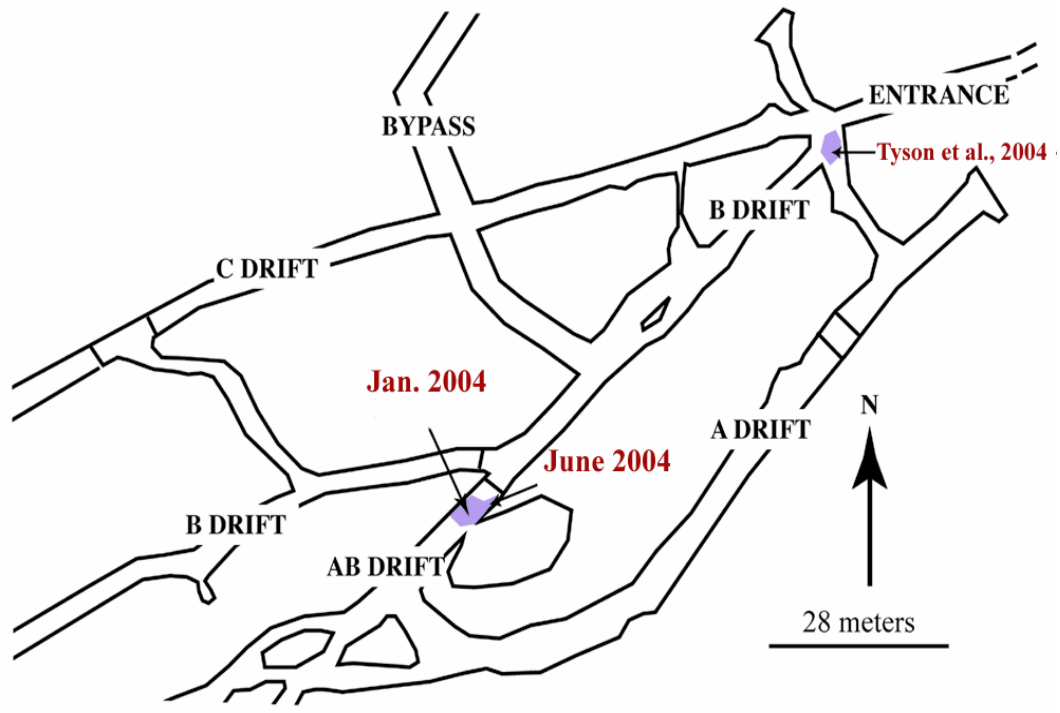


Figure 7.2: Map of biofilm sampling sites within Iron Mountain Mine, near Redding, California.
The original collection for genome sequencing was near the entrance to the mine noted as Tyson et al, 2004. The collection area for this study was near the AB drift, and was collected in 2004.
Figure courtesy of Dr. Jillian Banfield and Dr. Rachna Ram.

Materials and Methods

Chemicals and reagents

Unless otherwise stated, chemical reagents were obtained from Sigma Chemical Co. (St. Louis, MO). Modified sequencing grade trypsin, from Promega (Madison, WI), was used for all protein digestion reactions. The water and acetonitrile used in all sample clean up and HPLC applications was HPLC grade from Burdick & Jackson (Muskegon, MI) and the 98% formic acid used in these applications was purchased from EM Science (Darmstadt, Germany).

Collection of biofilm samples

Several hundred grams of biofilm suspension were collected from the Richmond mine, near the downstream confluence of drifts A and B (Figure 7.2 January and June 2004). A portion was fixed on-site (see below), while the rest was transported within the mine at ambient temperature and put on dry ice within an hour of collection, transported back to the laboratory, and then stored at -80°C. For this study, the samples collected in January, 2004, were the only ones used.

Whole-cell rRNA fluorescent *in-situ* hybridization (FISH) analysis

Samples were twice washed with 10 mM phosphate buffered saline (NaH_2PO_4, anhydrous, 1.9 mM phosphate, and 150 mM NaCl, adjusted to pH 1.2 with H_2SO_4), fixed overnight with 3 volumes of 4% paraformaldehyde to 1 volume of sample, and stored at -20°C within 8 hrs of collection. Hybridizations were performed on fixed samples, with incubation at 46°C and washing at 48°C for 15 min. Probe ARC915 was used for archaea, probe LF655 for all *Leptospirillum* groups, LF1253 for *Leptospirillum* group III, and EUB338 were used to visualize all bacteria members of the community.

Preparation of biofilm protein fractions for mass spectrometry

To prepare protein fractions, a sample of biofilm was thawed and cells from about 8 ml were processed at 4°C. Cells were suspended in 3 volumes H_2SO_4 (pH 1.1), washed by rotation for 30 min, and recovered by centrifugation at 12,000 x g for 20 min. This

wash was repeated once by resuspending the cell pellet in the same volume of sulfuric acid solution, and the two reddish-yellow supernatants were combined to form the extracellular fraction. Cells were resuspended in 20 ml 0.1 M sodium acetate (pH 5.0), placed on ice, and lysed by sonication using a micro-probe at high power with 30 sec pulses for 10 min. The suspension was centrifuged at 5,000 x g for 20 min, and the pellet containing cells and debris was re-extracted in the same manner. The combined supernatants constituted the cellular fraction. Centrifugation of the cellular fraction at 100,000 x g for 1 hr yielded a clear, yellowish supernatant enriched in soluble, cytoplasmic proteins. The reddish, translucent pellet resulting from ultracentrifugation was washed once by resuspension in sodium acetate buffer, the ultracentrifugation step was repeated, and the membrane pellet was resuspended in 1 ml of water. Fractions enriched for extracellular, whole cellular, and soluble proteins were precipitated with ice-cold 10% trichloroacetic acid, and the pellets were rinsed with cold methanol and air-dried. Membrane fractions were frozen on dry ice and later processed without precipitation.

For comparison of membrane treatments, cells from 8 ml biofilm were washed as described above, resuspended and split into two tubes. Cells recovered by centrifugation were then resuspended into either 12 ml of H_2SO_4 (pH 1.1) (for membrane sample "M1"), or 20 mM Tris-SO_4 (pH 8.0) (for the membrane sample "M2"). These were placed on ice and lysed by sonication as described above. After centrifugation, supernatants were diluted with either 40 ml of H_2SO_4 (pH 1.1) ("M1") or 0.1 M sodium carbonate (pH 11.0) ("M2") and mixed by rotation for 30 min at 4°C. These were centrifuged at 6000 x g to remove precipitates, and the supernatants were centrifuged at 100,000 x g for 1 hr. The membrane pellets were washed once by resuspension in the appropriate buffer, the ultracentrifugation step repeated, and each membrane pellet resuspended in 1 ml of the same buffer**.**

Each of the fractions described above were denatured and reduced in 6 M guanidine-HCl, 10 mM DTT, at 60^0C for 1 hr. Samples were diluted 6-fold in 50 mM Tris-HCl (pH 7.8) with 10 mM $CaCl_2$. Sequencing grade trypsin was added at ~1:100 (w/w) and digestions were run with gentle shaking at 37^0C for 18 hrs. This was followed

by a second addition of trypsin at 1:100 and additional 5-hr incubation. The samples were then treated with 20 mM DTT for 1 hr at 37^0C as a final reduction step, and immediately de-salted using Sep-Pak Plus C_{18} (Waters, Milford, MA). All samples were concentrated and solvent exchanged into 0.1% formic acid in water by centrifugal evaporation to ~10 μg/μL starting material, filtered, aliquoted and frozen at -80^0C until LC-MS/MS analysis.

Mass spectrometry

All samples were analyzed by 2-dimensional MudPIT LC/LC-MS/MS (described in Chapter 2). For the LCQ dataset, an Ultimate HPLC (LC Packings, a division of Dionex, San Francisco, CA) was used; for the LTQ dataset, a Surveyor HPLC (Thermo Finnigan, San Jose, CA) was used. Each pump provided a flow rate of ~100 μL/min, which was split pre-column to provide an approximate flow of ~200-300 nL/min at the nanospray tip. The split-phase columns were constructed as follows: the back column was packed with approximately 3.5 cm of strong cation exchange material (Luna SCX 5 μm 100A Phenomenex, Torrance, CA) into a 100 μm fused silica via a pressure cell followed by 3.5 cm of C-18 reverse phase (RP) material (Aqua C-18 5 μm 200A Phenomenex). The sample was then loaded off-line onto the dual phase column. For all samples, ~200-500 μg of protein was loaded onto the back dual phase column. The loaded RP-SCX column was then positioned on the instrument behind a ~15 cm C_{18} RP column (Jupiter C18 5 μm 300A Phenomenex) also packed via pressure cell into Pico Frit tip (100 μm with 15 μm tip New Objective, Woburn, MA). The entire column system was positioned into the nanospray source (Thermo Finnigan) on either the LCQ or LTQ mass spectrometers.

All samples were analyzed via a 24-hr 12-step MudPIT analysis consisting of increasing concentration (0-500 mM) salt pulses of ammonium acetate followed by 2-hour reverse phase gradients from 100% aqueous solvent (95% H_2O/ 5% ACN/ 0.1% formic acid) to 50% organic solvent (30% H2O/ 70% ACN/ 0.1% formic acid). During the entire chromatographic processes, the LCQ or LTQ mass spectrometers operated in a data-dependent MS/MS mode detailed below. The chromatographic methods and HPLC

columns were virtually identical for all analyses. The LC-MS system was fully automated and under direct control of the Xcalibur software system (Thermo Finnigan).

The LCQ and LTQ mass spectrometers were both operated with the following parameters: nanospray voltage (2.4 kV), heated capillary temp 200^0C, full scan m/z range (400-1700). The LCQ data-dependent MS/MS mode was set up with the following parameters: 4 MS/MS spectra for every full scan, 5 microscans averaged for full scans and MS/MS scans, 5 m/z isolation widths for MS/MS isolations and 35% collision energy for collision-induced dissociation. The LTQ data-dependent MS/MS mode was set up with the following parameters: 5 MS/MS spectra for every full scan, 2 microscans averaged for full scans and MS/MS scans, 3 m/z isolation widths for MS/MS isolations and 35% collision energy for collision-induced dissociation. To prevent repetitive analysis of the same abundant peptides, dynamic exclusion was enabled with a repeat count of 1 and an exclusion duration of 1 min on the LCQ and 3 min on the LTQ. All samples (5 biofilm fractions) were analyzed in triplicate on the LCQ with a single m/z range. In addition, the $NaCO_3$–treated membranes (M1) and the whole cell fraction were analyzed with a single 3 m/z range experiment which consisted of 3 individual 24-hr MudPIT analyses with three segmented m/z ranges (400-900, 850-1300, 1200-1700) on the LCQ. All samples (5 biofilm fractions) were analyzed in triplicate on the LTQ with a single m/z range except the crude soluble fraction, which was analyzed a single time due to lack of available sample.

Proteome informatics

Four protein databases were used for this study and each can be found at (http://compbio.ornl.gov/biofilm_amd/databases/). The first database (Biofilm_db1) contained all predicted proteins based on the community genomic analysis (Tyson, 2004) plus the *Ferroplasma acidarmanus* fer1 isolate genome. This database contained 12,148 entries and was the primary database used in all data analyses of the LCQ and LTQ datasets. A subset of LCQ and LTQ data was searched against two alternative databases. Biofilm_db2 was created by adding to Biofilm_db1 all proteins ≥100 amino acids encoded in regions of hypothetical genes after six-frame translation and consideration of

alternative start and stop sites, resulting in 15,646 protein entries. Biofilm_db3 was created by adding *Shewanella oneidensis* MR-1, *Rhodopseudomonas palustris* CGA009, *Escherichia coli* K-12, and *Saccharomyces cerevisiae* public protein databases to Biofilm_db1. The purpose of this database search was to determine the level of false positive identifications using microbial species known not to be in the sample. The results from these alternative database searches can be found in the analysis webpage (http://compbio.ornl.gov/biofilm_amd/analysis/). A final database (Biofilm_db1_snps_1) was created from the PCR information discussed below. This database was exactly the same as Biofilm_db1 except for two amino acid changes and removal of the N-terminus of the protein LeptoII_scaff_14_GENE_20. The new name for this protein is LeptoII_scaff_14_GENE_20_SNP1 in this database. This new database was used for searching the extracellular proteome LTQ data in order to test the hypothesis of two amino acid changes and N-terminal processing of the protein.

For all database searches, MS/MS spectra RAW files were first converted to mzXML format using ReAdW software program developed at the Institute for Systems Biology, Seattle, WA (http://www.systemsbiology.org), and available from SourceForge repository at (http://sashimi.sourceforge.net). The individual spectra for each RAW file were extracted from the mzXML file into corresponding DTA files (required as input to SEQUEST), using another software program from ISB, mzXML2Other. The spectra were then searched using SEQUEST, with the following parameters: enzyme type trypsin; Parent Mass Tolerance, 3.0; Fragment Ion Tolerance, 0.5; up to 4 missed cleavages allowed. Only fully tryptic peptide candidates were searched, non-specific cleavage was ignored due to potential false positive rates in large databases searched with the SEQUEST algorithm. The output data files were then filtered and sorted with the DTASelect algorithm (Tabb, 2002) using the following parameters: fully tryptic peptides only, with delCN of at least 0.08 and cross-correlation scores (Xcorrs) of at least 1.8 (+1), 2.5 (+2) and 3.5 (+3). DTASelect files from all proteome fractions analyzed by the LCQ and LTQ can be found in the analysis page (http://compbio.ornl.gov/biofilm_amd/analysis/) under the corresponding dataset. All DTASelect files were filtered at 1 peptide and 2 peptides per protein and are available for

download in a text format or a viewable html version where every identified spectrum can be viewed by clicking on the spectral number (first column, labeled by filename). The DTASelect results from all proteome fractions were then compared with the Contrast program (Tabb, 2002). The analysis page contains global contrast files (all the proteome fractions) filtered at 1 peptide, 2 peptides and 3 peptides per protein. The global analysis file filtered at 2 peptides from the LCQ dataset and the LTQ dataset were combined to give the final identified protein list. The analysis page also contains inter-fraction contrast files (compares multiple runs on same sample with same instrument platform) filtered at 1 peptide and 2 peptides, as well as pair wise comparisons (compare replicates runs of different proteome fractions) at 1 peptide and 2 peptides.

Results

We used mass spectrometry (MS)-based "shotgun" proteomics to characterize the protein complement of a relatively low complexity, natural microbial biofilm. Proteins could be identified because they were extracted from a sample similar to one for which genomic sequence is available (Tyson, 2004). The biofilm samples used in this study and prior work were collected from the underground regions of the Richmond Mine at Iron Mountain, near Redding, California (USA). These pink biofilms grew on the surface of very acidic (pH ~0.8) sulfuric acid-rich, hot (~42°C), metal-contaminated solutions. The previously characterized biofilm was collected from a location known as the '5-way' near the entrance to the mine (Figure 7.2, Tyson, 2004). In contrast, the samples used in this study were collected from ~42 m deeper into the mine near the confluence of the AB and B tunnels in January, 2004 (Figure 7.2). The biofilm formed a continuous, paper-thin film on the surface of a pool of slowly flowing acid mine drainage (Figure 7.3, top left and top right). This biofilm was much thinner than the biofilm present at the same location six months later (Figure 7.3, bottom right), indicating that it comprised an actively growing community. FISH analysis demonstrated that *Leptospirillum* group II dominated the sample, but it also contained *Leptospirillum* group III, *Sulfobacillus,* and archaea related to *Ferroplasma acidarmanus* (Figure 7.4). This is very similar in structure and composition to the community sequenced previously (Tyson, 2004).

Figure 7.3: Biofilm sample collection.
Top left: photograph of the biofilm from the AB end location (Figure 2) during collection in January 2004. The biofilm occurs as a continuous sheet over the surface of the AMD pool; lines are wrinkles that form due to movement of the solution. Top right: close-up photograph during sample collection showing that the biofilm is very thin and apparently homogeneous. Bottom right: image of the biofilm in the same location six months later. Figure courtesy of Dr. Jillian Banfield and Dr. Rachna Ram.

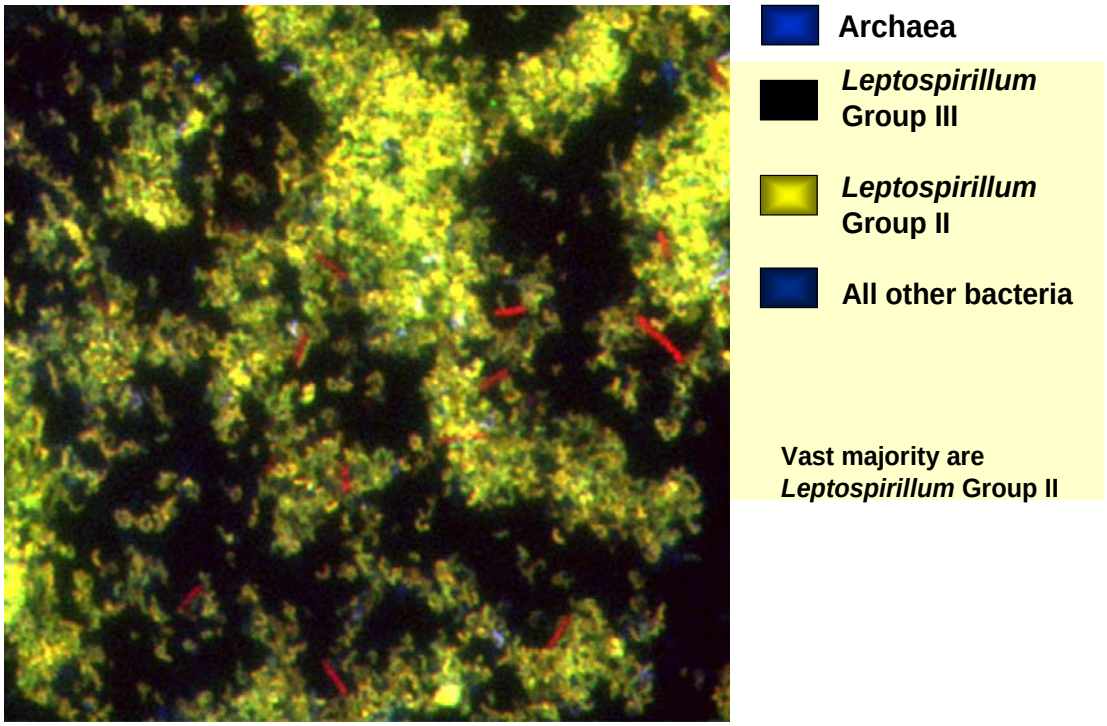


Figure 7.4: AMD biofilm composition from FISH analysis.
Fluorescence *in-situ* hybridization analysis of the biofilm collected from AB end in January 2004. In this image, *Leptospirillum* group II is yellow, *Leptospirillum* group III is white, *Sulfobacillus* sp. are red and archaea are blue.
Figure courtesy of Dr. Jillian Banfield and Dr. Rachna Ram.

For typical proteomic analyses, ~8 ml of biofilm was fractionated by washing, sonication, and centrifugation to yield extracellular proteins and samples enriched in proteins from whole cells, membranes (two different preparations), and cytoplasm (Figure 7.5). We combined two proteomic datasets (LCQ and LTQ) that were generated by triplicate analyses of the samples listed above. The LCQ dataset was generated on a 3-dimensional quadrupole ion trap mass spectrometer (LCQ-DECA XP plus, Thermo Finnigan, San Jose, CA). The LTQ dataset was generated on a 2-dimensional linear ion trap mass spectrometer (LTQ Thermo Finnigan). Both analyses used an identical "shotgun" proteomics approach via a two-dimensional (2D) nano-LC MS/MS system with a split-phase column as described in Chapter 2.

Genomic databases and protein detection

Proteins could be identified because comprehensive genomic data was available (Tyson, 2004). The environmentally-derived *Leptospirillum* group II and *Ferroplasma* type II composite genomes are near complete and *Leptospirillum* group III and Gplasma genomes are partially reconstructed. A partial environmentally-derived *Ferroplasma* type I genome is available, in addition to a complete genome of the closely related *Ferroplasma acidarmanus* isolate. In general, proteins could be assigned to organisms because the genes that encode them are on scaffolds that have been assigned to different organism types. From the genomic dataset, we created a database of 12,148 proteins (Biofilm_db1) that was used to identify MS/MS spectra.

All MS/MS spectra from the LCQ and LTQ datasets were searched with the SEQUEST algorithm (Eng, 1994) and filtered with DTASelect (Tabb, 2002) at the peptide level. Results of all replicate runs were compared with the Contrast program (Tabb, 2002) and evaluated based on matching of one peptide, two or more peptides, or three or more peptides per protein. All DTASelect files and Contrast files used in this study, as well as databases and resulting identifications, can be downloaded from the AMD Proteome Website Analysis Page (http://compbio.ornl.gov/biofilm_amd/). This website also contains directly linkable spectra for all identified peptides which are also downloadable, a step towards open access proteome results (Pedrioli, 2004; Carr 2004).

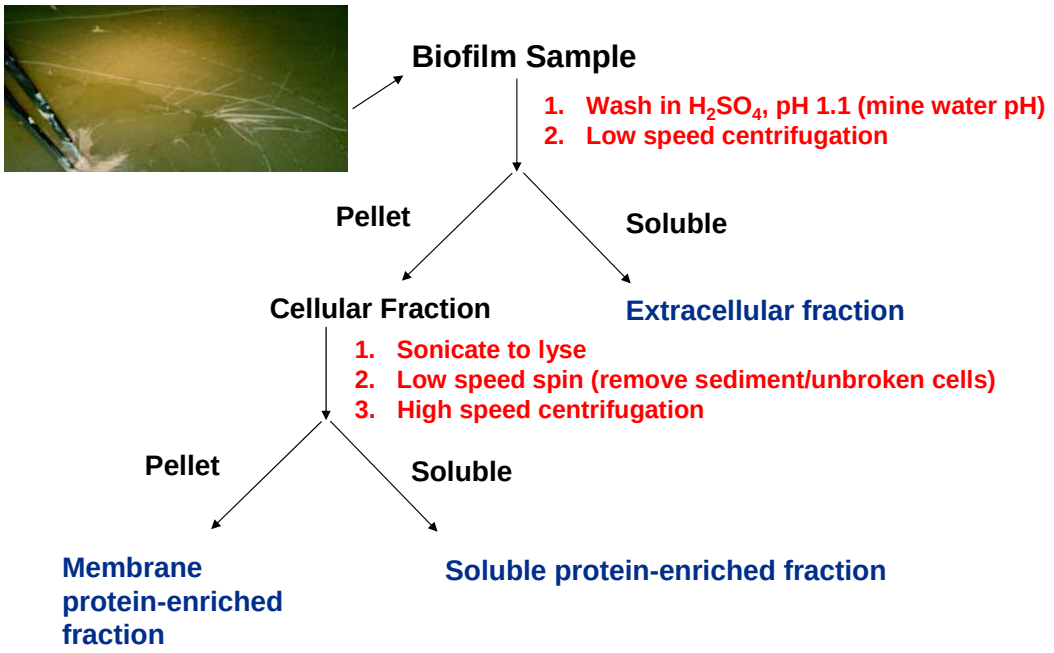


Figure 7.5: Fractionation of the AMD biofilm prior to MS-based proteomics. The biofilm sample (top left) is processed into four major fractions: an extracellular fraction, a whole cellular fraction, a soluble enriched protein fraction and a membrane enriched protein fraction. All fractions were analyzed in triplicate by LC/LC-MS/MS on LCQ and LTQ ion trap mass spectrometers.

One or more peptides from the combined LCQ and LTQ datasets were assigned to ~5,994 proteins (Table 7.1). This corresponds to ~49% of all proteins encoded by the genomes of the five dominant organisms. The entire list of identified proteins can be found at http://compbio.ornl.gov/biofilm_amd/supplemental (Table S1.pdf). Because of the anticipated false positive rate for identifications based on matching of only one peptide, we required matching of two or more peptides to a protein for confident detection (2,146 proteins were identified with this criterion). After removal of proteins duplicated in the genomic dataset (i.e., produced by the same gene but on different scaffolds separated due to strain heterogeneity), we detected 2,036 different proteins in the biofilm. All of the analyses below rely exclusively on proteins detected at the two peptide level from the combined dataset. The distributions of isoelectric points and molecular weights of these proteins were similar to those of all proteins predicted from genome data (Figure 7.6), indicating that there was no strong sampling or detection bias with this technique.

We detected 1,387 proteins from *Leptospirillum* group II, representing 48% of the estimated 2,877 genes in this organism (Table S2.pdf at http://compbio.ornl.gov/biofilm_amd/supplemental/). The dominance of *Leptospirillum* group II proteins in the biofilm was anticipated, based on the abundance of this organism type in the community (Figure 7.4). The fraction of proteins detected from *Leptospirillum* group II exceeds those of some prior proteomic studies of other microbial organisms. In part, the large number of detected proteins could reflect the presence of cells in many different growth stages, as well as microniches within the biofilm. The extensive sampling enabled analyses of the proteome of *Leptospirillum* group II in its natural context that are comparable to those achieved previously for pure cultures of other organisms. From the other species in the biofilm, 268 *Leptospirillum* group III, 84 *Ferroplasma* type I, 99 *Ferroplasma* type II, and 120 Gplasma proteins were detected. In addition, 34 proteins were detected on unassigned archaeal scaffolds and 39 on unassigned bacterial scaffolds. The low level of detection of other species in this community is not surprising in light of the dominance of *Leptospirillum* group II.

We performed a separate analysis to test the likelihood of matching unique

Table 7.1: Number of proteins detected at different filtering levels, derived from redundant protein counts from global Contrast files of the entire LCQ and LTQ datasets.

Filtering Level	LCQ dataset	LTQ dataset	Combined datasets
Liberal filters*	3127	5534	5994
Conservative filters**	1160	2077	2146
Ultra-conservative filters***	837	1419	1435

*Liberal filters requiring at least 1 peptide per gene;
**Conservative filters requiring at least 2 peptides per gene;
***Ultra-conservative filters requiring at least 3 peptides per gene.
Xcorrs of at least 1.8 (+1), 2.5 (+2) 3.5 (+3) were used in all cases.

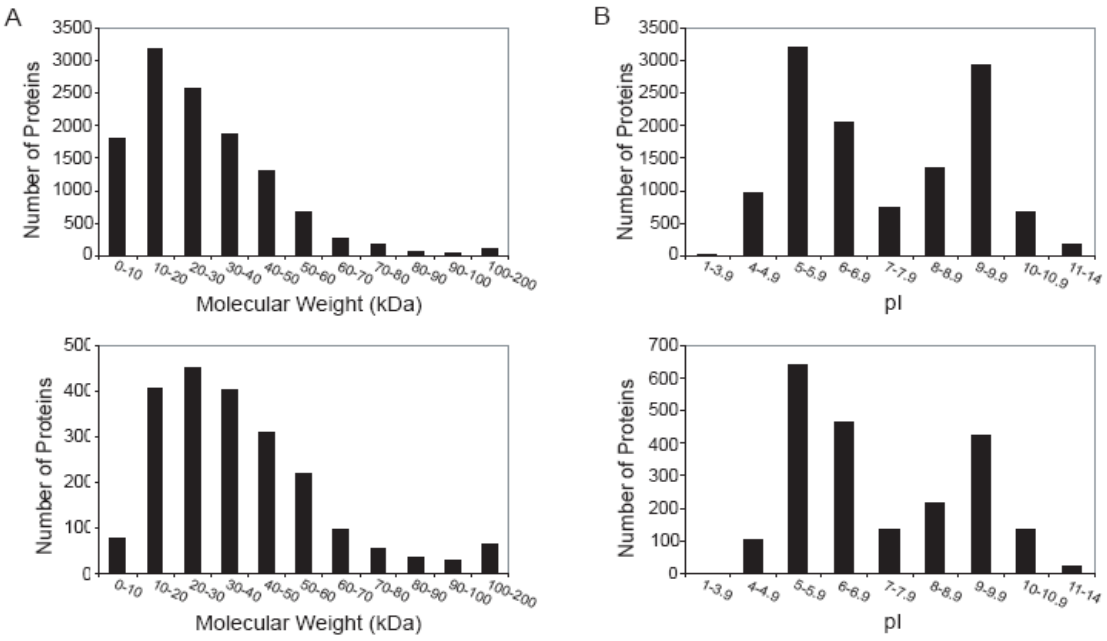


Figure 7.6: Genome and proteome MW and pI distributions.
Molecular weight (A) and isoelectric point (pI) distributions (B) of all proteins predicted from genomic analysis compared to those confidently detected by proteomic analysis. This analysis is based on predicted proteins from the biofilm_db1. The top panels are results for all proteins predicted in the genomic dataset and the bottom panels are the results for all proteins detected by proteomic analysis.

peptides from proteins not present in the samples ("false positives"). For this analysis, we supplemented the database with protein sequences derived from the genomes of *Escherichia coli*, *Shewanella oneidensis*, *Rhodopseudomonas palustris*, and *Saccharomyces cerevisiae*. This new database (Biofilm_db3) contained 31,900 proteins, causing a decrease in the number of unique peptides derived from mine organisms. Nonetheless, of the 6,605 unique peptides identified by LCQ analysis of the five biofilm fractions using this new database, 12% were false positives. Analysis of the 5,397 unique peptides that were derived from proteins for which at least two peptides were recovered revealed a false positive rate of only 2.8%. A similar analysis of the data from the LTQ analysis revealed a false positive rate of 16% at the one unique peptide level and 6.2% after considering only unique peptides from proteins matched by at least two peptides. This indicates that the likelihood of spuriously matching peptides in the MS/MS data is very low, especially after filtering at the two peptide level, and taking into consideration that some of the false positives could have derived from mine organisms for which we have incomplete or no genomic information. The increase incidence of false positives in the linear ion trap dataset was expected due to the faster scan speed of this instrument (~five times faster than the 3-dimensional ion traps). This illustrates the need to reconsider common proteome informatics practices as MS instrumentation evolves over time.

It is expected that concentrations of proteins from very low abundance members will be too low to be detected by community proteomics analysis with current technology (~1%, see chapter 6). Furthermore, we are unable to identify proteins from organisms such as *Sulfobacillus*, for which there is no genome sequence available. It is likely that the dominant strains present in this biofilm slightly differ from those in the previously characterized biofilm (Tyson, 2004). In most cases, a single amino acid substitution will prevent peptide detection. However, if amino acid-level divergence between strain populations in the biofilm is comparable to that observed previously (average <2%), matching of at least two peptides to most proteins should be possible. Numerous high quality spectra for unidentified peptides are available for further analysis once additional

genomic data becomes available. All MS raw files from this study are archived at ORNL and available for future analysis.

Abundant biofilm proteins

The most abundant proteins are likely to be critical to the biofilm community. Determining relative or absolute abundance of individual proteins is a recognized challenge in MS-based proteomics. Detection biases can arise due to differential protein extraction, matrix effects in the ionization processes, and biases in the digestion and sample clean up processes. Nonetheless, sequence coverage, the number of unique peptide hits (number of unique spectra detected per protein), and MS/MS spectral counts (the number of times a peptide is detected from a protein) are all indicators for protein abundance (Liu, 2004). We assume that potential detection biases will not be so large as to obscure general trends, or inferences about the relative abundance of broad functional categories.

We used peptide count, spectral count, and percent coverage to infer the 10 most abundant proteins in each fraction (full list can be found at http://compbio.ornl.gov/biofilm_amd/supplemental/, Peptide & MS/MS Spectra Counts, Sequence Coverage). All three measures gave generally similar results. Overall, the biofilm is dominated by hypothetical proteins. We defined hypothetical proteins as those lacking a significant BLAST match ($<e^{-10}$) to a protein with a functional assignment. By this definition, 42% of the predicted proteins encoded by the genomes of community members were hypothetical. Since many functionally annotated proteins have not been biochemically characterized, this approach is likely to underestimate the number of truly novel proteins. Predicted proteins with no significant similarity to any known protein are referred to as "unique". Those with similarity to predicted proteins but no close similarity to characterized proteins are described as "conserved". Unique and conserved novel proteins represented 15% and 2% of the abundant proteins, respectively.

The biofilm was also dominated by ribosomal proteins (13%), chaperones (11%), thioredoxins (9%), and proteins involved in radical defense (8%). Thioredoxins are involved in redox reactions in which proteins containing disulfide bonds are refolded.

There were at least four different and abundant disulfide isomerases detected in the extracellular fraction and ten in the entire proteomic dataset. This indicates that protein stability in pH <1 solutions is achieved in part by refolding carried out by abundant pH tolerant enzymes. Peroxiredoxin and some other highly detected proteins (i.e., rubrerythrin, catalase) are involved in defense against oxidative stress, suggesting that byproducts of aerobic respiration are an important challenge in the AMD environment.

Based on percent sequence coverage, the complement of proteins enriched in the extracellular fraction (>10% coverage and more than twice as abundant in the extracellular fraction as any other fraction) is dominated by unique novel proteins (64%), and contains only ~1% conserved novel proteins. The presumably metal and acid tolerant unique proteins are future research targets, as they likely play key roles in adaptation. We detected all predicted proteins for 9 putative operons composed only of hypothetical genes. For example, one operon encodes five *Leptospirillum*-specific proteins and another encodes three *Leptospirillum* group II-specific proteins, all of which were detected in the membrane and extracellular fractions.

Functional analysis of an abundant hypothetical protein

Of the proteins enriched in the extracellular fraction, the one with the highest sequence coverage is encoded by a hypothetical gene from *Leptospirillum* group II. 67% of the protein sequence from the community genomics dataset could be reconstructed from multiple overlapping peptides. No peptides were recovered from three discrete regions of the protein (Figure 7.7). The first is a 23 amino acid region predicted to be a signal peptide. Sequences for the gene determined by PCR amplification differed from that in the community genome dataset by ~3% at the nucleotide level and resulted in substitutions of one glutamate for aspartate at position 76 and one serine for glycine at position 139.

Using DNA sequence for the variant, peptide sequences were re-analyzed with a database containing the corrected amino acids. The protein was fully recovered after this modification, except for the predicted signal peptide (Figure 7.7). This example illustrates a more general approach, in which regions of proteins not detected by MS due

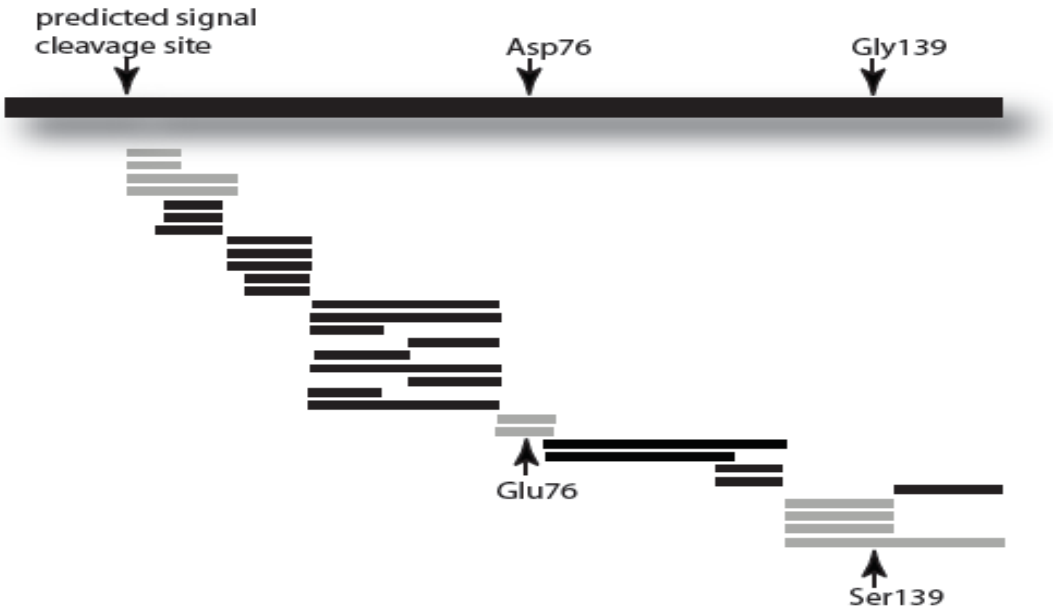


Figure 7.7: Recovery of peptides spanning the entire sequence of a natural variant of cytochrome 579.
The predicted sequence for cytochrome 579, based on the community genomic data (Tyson, 2004), is represented by a large black bar. Below, smaller black bars represent tryptic peptides identified through proteomic analysis that correspond to regions of cytochrome 579. Grey bars represent regions of the mature protein recovered by mass spectrometry after consideration of cleavage of an N-terminal signal peptide and two amino acid differences due to strain variation.

to post translational modification or strain variation can be resolved by PCR (either the full gene or just across variable regions). Thus, genomic data for a sequenced strain can enable proteome analysis and discovery of abundant protein variants in environmental samples for which no sequence data are available.

The abundant protein in the extracellular fraction is only weakly similar (e^{-6} by BLASTp) to previously studied c-type cytochromes and Fe/Pb permeases. This, and the presence of a heme-binding consensus sequence, suggested a role in electron transport. We verified that the predicted N-terminal sequence of this protein almost exactly matches that of an abundant heme-staining protein identified by SDS-PAGE analysis of the extracellular fraction. Interestingly, the peptide sequence differs in that it contains leucine in place of isoleucine encoded by the codon ATC, suggesting codon usage that differs from *E. coli*. However, the proteomic analysis is blind to this substitution because isoleucine and leucine share the same mass.

Abundant iron-oxidizing cytochromes with absorption peaks around 579 nm had previously been detected in *Leptospirillum ferrooxidans* (*Leptospirillum* group I) and *Leptospirillum ferriphilum* (very closely related to *Leptospirillum* group II) isolates (Hart, 1991, Blake, 1993). Amino acid sequences for these cytochromes had not been reported. Spectroscopic analysis of the purified *L. ferriphilum* cytochrome (cyt579) showed that the heme group in this protein is unusual and not of the typical *a*, *b* or *c*-type. At pH 2, this protein has an unusually high reduction potential (615 mV), appropriate for catalyzing iron oxidation. The first 40 amino acids of the iron-oxidizing cytochrome purified from the periplasm of *L. ferriphilum* match amino acid positions 24 to 64 of the putative cytochrome in *Leptospirillum* group II from the natural biofilm. This confirms that the mature protein lacks a 23 amino acid leader sequence. Thus, we reconstructed 100% of the mature protein via MS-based proteomics. The cleavage of a leader sequence from the N-terminus indicates that the mature protein is exported across the cytoplasmic membrane. The partitioning of cyt579 into the extracellular fraction of *Leptospirillum* group II corroborates its localization in the acid-exposed periplasm. Based on its distribution, its abundance, and the ability of its *L. ferriphilum* homolog to oxidize iron, we conclude that cyt579 may be central to iron oxidation by *Leptospirillum* group II.

We detected eight other membrane and periplasmic c-type cytochromes, and components of NADH dehydrogenase, succinate dehydrogenase, and the cytochrome bc complex. In addition, three hypothetical proteins with heme-binding motifs were detected. Using this information, we developed a working model for the iron oxidation pathway in *Leptospirillum* group II (Figure 7.8) in which cyt579 is the first step. Elucidating the roles of the other c-type cytochromes in this electron transport chain is an important objective for further study, as iron oxidation is central to energy generation in the AMD ecosystem. As the supply of ferric iron is the rate-limiting step for pyrite oxidation, the metabolic activity of iron-oxidizing microorganisms largely determines the rate of AMD formation. *Leptospirillum* group II dominates most biofilms from the Richmond Mine and is frequently detected at other mining sites and bioleaching plants. *Thus, cyt579 is potentially the key enzyme that connects the biology and geochemistry of metal-rich acidic environments.*

Functional categories of detected proteins in *Leptospirillum* group II

From the combined 2 peptide proteome dataset, 69% of the detected *Leptospirillum* group II ORFs encode proteins that could be assigned a function (required BLAST match of $<e^{-10}$ to proteins of known function). We assigned all detected *Leptospirillum* group II proteins to functional categories based on clusters of orthologous genes (COG) (Tatusov, 1997) in order to evaluate the degree of expression of hypothetical proteins and to estimate the relative investments by this organism in different metabolic activities. The most commonly detected proteins were unique and conserved hypotheticals (Figure 7.9). Proteins involved in amino acid metabolism, translation, and energy production and conversion were the next most commonly detected, followed by cell envelope biogenesis, coenzyme metabolism, and protein folding and modification. Within the category of coenzyme metabolism, proteins involved in cobalamin and heme biosynthesis were abundant. Heme is essential for manufacturing cytochromes such as cyt579. A high demand for cyt579 is consistent with the relatively low energy yield associated iron oxidation.

Heme is also required for construction of abundant proteins such as catalase and

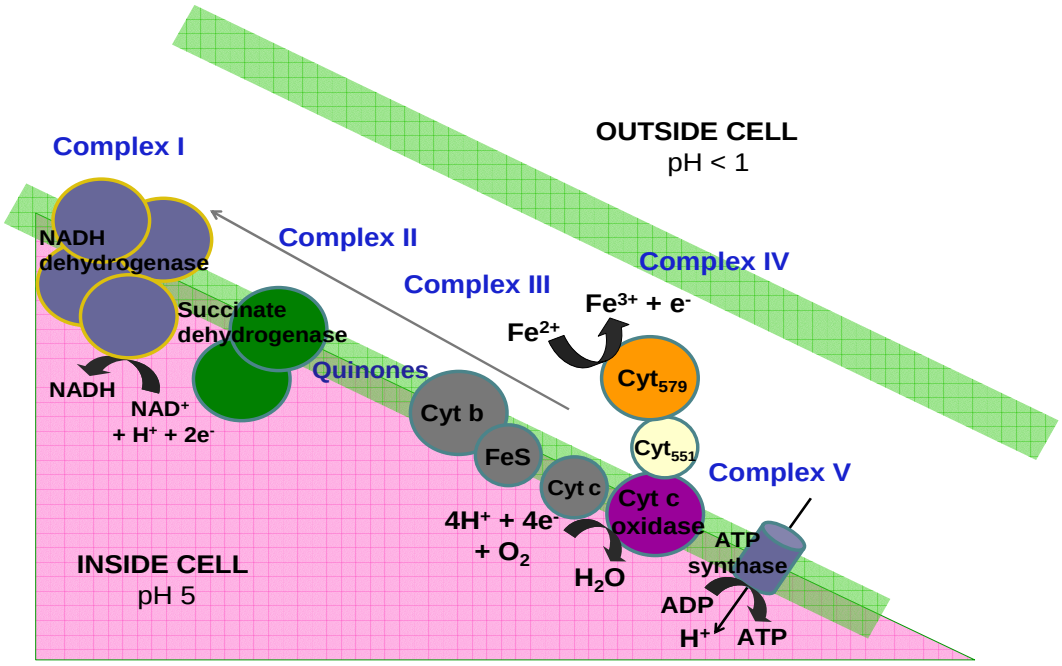


Figure 7.8: Potential mechanism for Fe^{2+} oxidation in *Leptospirillum* group II. Based on detected proteins, we inferred a possible arrangement of electron transport components associated with the cell wall of *Leptospirillum* group II. Quinones are inferred based on detection of UbiC, UbiB, and UbiE proteins. Iron oxidation by soluble, acid-stable cyt579 localized in the periplasm is coupled to oxygen reduction and the generation of ATP. Some electrochemical energy is siphoned towards producing NADH for use in biosynthetic pathways (arrows pointing to the left). Expressed proteins related to c-type cytochromes (possibly including cyt 551, 553, and four other distinct types) could be involved in the production of NADH and/or serve as intermediates in the transfer of electrons from cytochrome 579 to the terminal oxidase complex.
Figure courtesy of Dr. Jillian Banfield and Dr. Rachna Ram.

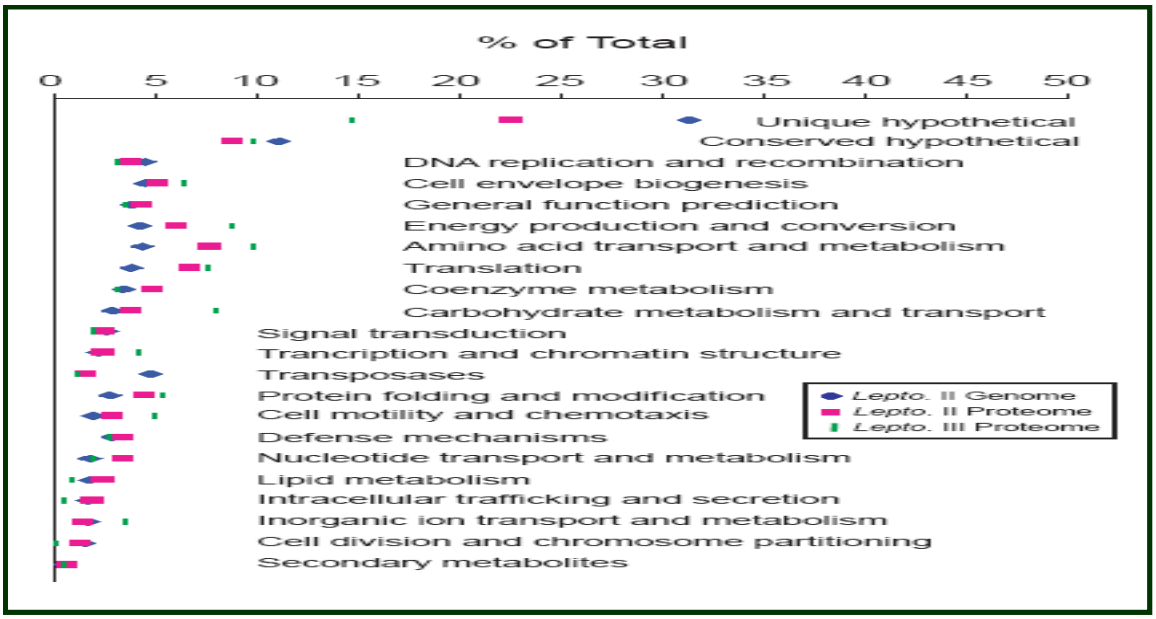


Figure 7.9: Functional categories of AMD proteome.
Functional categories of *Leptospirillum* group II proteins predicted from the genome dataset, and *Leptospirillum* groups II and III detected in the proteome. Depicted are the percent of total proteins in each functional category, based on the corresponding genome and proteome datasets.

peroxidase, important for peroxide and radical detoxification. Similarly, the abundance of enzymes involved in protein refolding may reflect the challenge associated with maintaining protein integrity in the hot, acid environment. The abundant disulfide isomerases may also construct and maintain the conformation of the abundant acid-exposed heme-based proteins localized in the periplasm. Proteins from COG families involved in secondary metabolite biosynthesis/transport/catabolism, cell division/chromosome partitioning and inorganic ion transport/metabolism contained the smallest numbers of detected proteins. In part, this may reflect our inability to assign these metabolic roles to novel environment- and lineage-specific proteins.

Despite the predominance of hypothetical proteins, it is notable that only 38% of the predicted "conserved hypothetical" and 35% of the predicted "unique hypothetical" proteins were detected. In contrast, we detected 86% of proteins involved in amino acid metabolism and 86% of those involved in translation. This suggests that many of the predicted hypothetical proteins are either non-functional, low abundance or are expressed under conditions different to those at the time of sampling. Relative to the genome, the proteome is enriched in proteins required for amino acid metabolism, translation, nucleotide metabolism, protein refolding and modification. In all other categories (except transposases), representation in the genome was similar to that in the proteome (Figure 7.9). This suggests that the diversity of genes for a particular function is a relatively good predictor of diversity in the proteins associated with that function in the proteome. The deviations may be clues to the demands placed on the organism at the time of sampling.

Conclusions

This description of the protein complement of a natural biofilm utilized relatively comprehensive genome sequence data from a closely related microbial community. Although the samples were different from those previously characterized by genomics, *it was possible to confidently detect 2,036 proteins*. This represents the first large-scale proteome characterization of a natural microbial community. The large number of detected unique and conserved novel proteins underscores the importance of proteins of

unknown function in the community. Novel proteins that were detected and abundant may be targeted for purification from biofilm samples and subsequent functional analysis. While the combined genomics and proteomics approach can validate the existence of such genes the methodology does not verify function of the gene products. Rather, it provides the information for directed purification and characterization of potential important unknown proteins from the community. The validation and subsequent characterization of identified proteins is the next step in understanding the function of the fascinating communities populating acid mines.

Even slight variations in protein sequence could be detected by the mass spectrometry methods used here, making possible future ecological studies in which dominant strain variants are followed in their natural environment over time. As the sensitivity and dynamic range of mass spectrometry methods improves, analysis of smaller samples will enable studies with higher spatial resolution and make it possible to differentiate peptides from closely related coexisting strains. Mass spectrometry-based *de novo* sequencing approaches on the horizon (Standing, 2003) and the potential for MS^3 analysis (Olsen, 2004) should reduce the requirement for exact gene sequence data, broadening the applicability of genome sequence information. The proteomic data presented here provides insights into the major challenges faced by life in an extreme environment. A similar combination of cultivation-independent genomic and proteomic methods could be extended to other communities containing uncultivated organisms of environmental, medical, or industrial importance. The main challenge for future studies will be the complexity and dynamic range of the communities. Clearly this was a very simple community and a good starting point for developing the necessary tools for characterizing the complex proteome of microbial communities. But much work is needed to improve upon the methodologies discussed here to tackle the complexity likely to be found in other natural microbial communities.

Chapter 8
Conclusion

In the studies presented here, we have attempted to build a high performance MS-based proteomics platform for the analysis of microbial proteomes from isolates to communities. We first discussed the methodologies of microbial growth, proteome extractions, and sample preparation and then moved into discussions of the evolution of LC-MS/MS methodologies used in these studies. Emphasis was placed on the development of the proteome informatics pipeline and the challenges faced in the extraction of relevant biological information from large proteome datasets. We then illustrated how these technologies can be applied to the analysis of microbial proteomes. Clearly, although great progress has been made through these studies, much work still needs to be done in technology development to bring proteomics to the level of whole genome sequencing and transcriptome analysis. Some potential avenues of research are discussed below.

The combination of the top-down and bottom-up MS methodologies for the characterization of individual proteins, protein complexes and whole proteomes was first conceived in our laboratories. While many proteomics groups were focusing on either top-down or bottom-up techniques, very few have tried to integrate the two technologies. Our initial efforts have demonstrated great promise for this integrated technology for individual proteins and protein complexes to obtain a detailed level of information not possible by either technique alone. This includes the determination of the position and number of post-translational modifications on the final intact protein product, as well as the determination of the number and position of amino acid changes (mutations) within intact proteins for most potential substations (Ile-Leu can't be resolved). The exact position of N-terminal cleavage positions of processed proteins was also demonstrated. The analysis of individual protein and protein complexes is somewhat straightforward with the combined technology with certain limitations, but the analysis of whole proteomes is currently beyond the analytical capability. This is primarily due to limitations in the top-down technology and limitations in software to integrate the

datasets. Top-down technology in its current form has difficulties with the complex mixtures found in whole proteome analysis. Potential 2D separations of intact proteins may overcome this limitation. Also, some protein types, such as proteins larger than 40-50 kDa, are not directly amenable to top-down analysis in complex mixtures. The primary technological advances needed for this combined technology include methods for data-dependent MS/MS on intact proteins on liquid chromatography time scales and improved MS methods for characterizing large proteins, that is >50 kDa proteins from complex mixtures. Another primary limitation of top-down analysis is the bioinformatic tools for querying protein databases. The isotopic packets of intact proteins and the MS/MS spectra of intact proteins are both much more complicated than those derived from peptide measurements. Improvements in proteome informatics for top-down data and informatics tools for combining top-down and bottom-up datasets to search for PTMs, amino acid substitutions, and N-terminal truncations are all necessary.

We introduced the first attempt to use semi-quantitative proteomics data to compare a WT and mutant strain of *Shewanella oneidensis*. The mutant was a gene knockout to the global regulator of iron uptake, the ferric uptake regulator. Microarray analysis of the transcriptome of the WT vs. the *fur* mutant indicated many proteins thought to be involved in heavy metal transport and utilization up-regulated in the *fur* knockout. This was expected since *fur* is an active repressor of the expression of these proteins under conditions where iron is not limiting. Previous proteome studies of the *fur* knockout by 2D-PAGE-MS had not identified many of the proteins predicted to be over-expressed by the microarray analysis. We employed a pure liquid-based LC-MS/MS approach using two varieties of "shotgun" LC-MS/MS techniques for a semi-quantitative comparison between the WT and *fur* mutant strain. We clearly identified many proteins thought to be involved in heavy metal transport and utilization as well as conserved unknown and unknown proteins up-regulated in the *fur* mutant strain. These results had very good correlation with the microarray data on the same samples, with only two proteins showing an inverse correlation.

While many membrane proteins, including transporters and receptors, were identified in the *S. oneidensis* study, the identification of such proteins, especially with

high sequence coverage, is still a major challenge in proteomics. As discussed in Chapter 5, many small hydrophobic proteins involved in multi-protein membrane complexes are very difficult to routinely detect, thus making comparative proteomics of such proteins a difficult process. While not highlighted to a great extent in this dissertation, much work has been done on developing alternative methods for digesting membrane proteins from microbial systems in our laboratories. Alternative digestions methods, such as dual CNBr/trypsin digestions, digestions with Proteinase K, and digestions with trypsin in the presence of high organic (80% Acetonitrile or 100% methanol), were all evaluated. None of these techniques proved to be reproducible or effective at routinely identifying small embedded hydrophobic membrane proteins. Of the three above mentioned techniques, we found only the CNBr/trypsin method to be effective, while the other two methods were actually found to be less effective than the normal trypsin digest. The CNBr/trypsin method has some major shortcomings, including the toxicity of CNBr, the difficulty and time-consuming sample preparation process and the difficulty with processing the data generated from these digests with SEQUEST. Clearly, this area of research needs much attention but a straightforward solution is not evident. Even if it is possible to digest the protein, it becomes very difficult to keep the resulting hydrophobic peptides in solution through sample preparation and liquid chromatography. Furthermore, their lack of basic residues, which are required for pre-forming protonated peptide ions in solution prior to electrospray ionization, creates another major problem.

We extended the semi-quantitative approach to the first large-scale proteome characterization of *Rhodopseudomonas palustris* under its major metabolic states. All 6 growth states, along with one mutant, were characterized in duplicate by "shotgun" proteomics. This study, along with Lipton et al. 2002, are the only two studies to date to use "shotgun" proteomics to characterize and compare a large number of metabolic states from a microbial system. This study demonstrated how important such large scale multi-metabolic state proteome studies are and demonstrated how the data, if made publicly available, can be integrated with other system level studies such as global gene knockouts and large scale analysis of protein complexes. Theses types of studies are absolutely essential to obtain a greater understanding of the function of microbial systems at a global

level and many more such studies should be pursued at the proteome level with other microbial species. We were clearly able to show the differential expression of many known, conserved unknown and unknown proteins across the metabolic states, including the detection of a completely novel operon expressed only under the anaerobic growth states. These types of experiments become most useful when they can be done under many different growth states with many replicates to build statistical confidence in the results, as well as textual understanding of protein expression patterns. Most current proteomics efforts in microbial systems involve the characterization of a proteome only under a single metabolic state or the comparison of two states or a mutant and a wild-type. While such studies are informative, all initial proteomics efforts for isolated microbes should first begin with a baseline analysis of that microbial proteome under as many metabolic states as possible with as many replicates as possible. Our landmark work with *R. palustris* has generated a proteome database than can be directly referenced by all in the community who are currently, and in future studies, carrying out more detailed studies with this bacterium.

As an example, we decided not to pursue such large-scale baseline studies of the *S. oneidensis* proteome because such experiments were thought to be underway at other laboratories. This data is still not available, which has impeded our progress on more detailed proteome studies that we are currently undertaking on *S. oneidensis* chromium utilization. The needed future experimental project in *S. oneidensis* is a baseline analysis of the *S. oneidensis* proteome under anaerobic and aerobic conditions, with a large number of alternative electron acceptors. This dataset then must be made available in an open access format to the community as a whole as a reference point for future experiments.

The study genomes, transcriptomes and proteomes, of microbial isolates, is the obvious first step in attempting to understand microbial processes at a systems level. But this is only the first step. Eventually, these studies must be conducted with the natural communities in which these microbes live and have to cope with their natural surroundings. The reason for this is three-fold:

1) Many microbes cannot be cultured from their natural habitats. Indeed, it has been estimated that less than 1% of all microbes present on earth can be cultured in the laboratory.

2) Microbes in the natural environment are continually interacting with one another and these interactions, whether for providing each other with essential nutrients for survival or for the transmission of genetic information and thus genetic potential, are essential. The precise nature of these intricate interactions cannot be reproduced in a laboratory; thus they cannot fully be understood by the analysis of isolates or mixtures of isolates grown in the laboratory.

3) The metabolic responses of microbes to their natural environments are the key to understanding their survival mechanisms but the complexity of a natural environment can never really be fully simulated in the laboratory.

For these reasons, it is necessary to develop systems biology techniques to study microbial species directly from the natural environment. The three main reasons to study microbial systems from the environment discussed above are the main reasons these studies will be so difficult. The communities of interest can have hundreds to thousands of individual species, the community can be continually changing, and the natural environment can be continually changing. To date, two studies of employing whole genome sequencing of microbial communities have been accomplished (Tyson, 2004; Venter, 2004). To date, no major measurement of transcript levels from an environmental sample has been accomplished. Even with these daunting challenges, we decided to develop and evaluate methodologies to measure mixtures of microbial proteomes and, hopefully, natural community proteomes.

We demonstrated our first attempts at measuring proteomes of artificial microbial mixtures. Very simple mixtures of four microbial species (*S. oneidensis, R. palustris, E. coli,* and *S. cerevisiae*) were made with altering concentrations of *R. palustris*. These types of tests are absolutely necessary to determine the current state of proteomics technologies and to develop new technologies. It is more logical to work with known mixtures of microbes simulating microbial communities than to try and develop these technologies on undefined precious samples from real communities. We tested the

effects of concentration to determine how well we could measure a proteome of a minor species in the presence of more abundant species. While species identification and protein identifications could be made with high confidence at 1% and as low as 0.2%, the level of proteome coverage was insufficient at less than 25% target species.

This is clearly a limitation to the application of current proteomics to the analysis of microbial communities, since many of the species in a natural community will be present at less than 1%. The major challenge here as with any proteomics experiment is dynamic range. The potential methodology necessary to obtain the necessary dynamic range is the development of 3-dimensional separation techniques (see discussion in Chapter 6). While still under development, the addition of an extra dimension might provide some of the needed dynamic range. It will also be helpful to move all tests from the quadrupole ion trap to the linear ion trap. While this will have a dramatic increase in the number of proteins identified, we are not optimistic that this change alone will allow for deep proteome coverage of species less than 1%. Another potential avenue is rapid enrichment of certain microbial species from the environmental sample by cell sorting techniques. Potential collaborations with instrument manufacturers on this methodology could be critical. It will be important for any such experiments to be done rapidly so as not to introduce major changes to the microbial proteome.

From the mass spectrometry perspective, the most logical solution is the implementation of FT-ICR mass spectrometers. While these instruments were not capable of rapid data-dependent MS/MS during the course of this work, such an instrument, which combines a linear ion trap with an FT-ICR (Syka, 2004) has recently become available. The FT-ICR provides the best dynamic range of any mass spectrometer available in full scan mode, with ~100-1,000 times better performance in dynamic range than ion trap mass spectrometers. This is absolutely essential for obtaining deep coverage of low-abundance proteins or, in this case, low-abundance proteomes. This can be accomplished since the FT-MS allows for the detection of low level peptides in the full scan mode and the linear ion trap can then easily isolate and fragment those ions (the ion trap has excellent dynamic range in MS/MS mode but is somewhat more limited in full scan MS mode). The combined LTQ-FT-ICR provides a

potential solution to obtain quality proteome coverage of microbial species less than 1% in a mixture.

While methodologies for characterizing the proteome of complex microbial mixtures were still under development, we were provided the opportunity to test our techniques on a natural microbial community. The microbial communities populating the Acid Mine Drainage (AMD) streams of Iron Mountain, California, were a unique and perfect opportunity. The self-sustaining, chemoautotrophic microbial communities found in the AMD streams survive in the harshest of environments with very acidic conditions (pH <1.0), high metal concentrations, high temperatures, and a lack of fixed nutrients. The fact that these conditions are so hostile to life made these microbial communities perfect model systems for a combined genomic/proteomic approach. This is because very few microbial species are capable of surviving in the environment, so the mixtures obtained contain predominantly 4-5 dominant species. Previous characterization of a microbial biofilm from this community by whole genome sequencing (Tyson, 2004) paved the way for proteome initiative. We applied "shotgun" proteomics to a sample similar to the community characterized by whole genome sequencing. Although the samples were different from those previously characterized by genomics, it was possible to confidently detect 2,036 proteins. A large number of detected unique and conserved novel proteins underscored the importance of proteins of unknown function in the community. To our knowledge, this was the first large-scale characterization of natural microbial proteomes.

While our first characterization of a microbial community could be considered a great success, we also learned many technical areas that need to be addressed with future research in order to better characterize the AMD communities as well as other communities. As discussed above, the first biggest concern is extending the dynamic range. For the community we characterized, one microbe was dominant with 4 others less dominant. Very good coverage was obtained for the dominant microbe (~48% of the predicted proteome) but the other bacteria and archae were all only covered at 10% or less. FISH analysis clearly indicated the presences of the other bacteria and archae but all were at a very low percentage in comparison with the dominant bacterium.

The second major challenge is strain variation from the original sequenced genomes. It will rarely be the case that it will be possible to do the proteome characterization on the exact same sample as the genome characterization. It will be necessary to characterize the proteomes of many different microbial communities within an environmental niche to obtain spatial and temporal resolution of the communities. Proteome technologies allow for the analysis of many samples, but the time and monetary restraints of whole genome sequencing limits its application to a few representative samples. The problem with this is that genomes of communities will gradually change at the base pair level over time and space (strain variation). Since MS-based proteomics relies on exact correspondence between the predicted peptides and those for which mass spectra are collected, a single amino acid substitution is likely to prevent assignment of a peptide to a protein. Thus, the divergence of protein sequence away from the original sequence from the whole genome sequencing project can have grave consequences in the ability of MS techniques to identify given proteins.

We illustrated this divergence for two amino acids in an abundant protein in the biofilm sample and illustrated how a combined approach of PCR and MS analysis could verify the amino acid substitution. While this method worked for an individual protein, it is not practical on a whole proteome level. Clearly, higher throughput methods for characterizing strain variations must be developed. The first challenge is to determine which peptides have been modified; the second involves determining what the correct sequence is for the strain variant peptide. Research is needed for the combination of two potential methods.

A potential solution is the combination of MS^3 of fragment ions and high mass accuracy parent peptide measurements with a *de-novo* sequencing algorithm (reviewed by Standing, 2003) to detect and verify peptides and proteins amino acid changes from the community. The high mass accuracy measurement of parent peptide ions is necessary to validate predicted sequences from the *de-novo* sequencing algorithm. The high mass accuracy measurement could be accomplished with the LTQ-FT-MS instrument. This instrument allows for accurate measurement of peptides within 3 ppm from LC-MS analyses. This level of mass accuracy allows for good discrimination between potential

peptide sequences predicted from a *de-novo* sequencing attempt. If the mass of the predicted peptide is not within ~5 ppm of the measured mass, then the candidate sequence can be rejected and others can be considered. The MS^3 experiment offers another confirmation of predicted peptide sequences, as recently described in Olsen 2004. The concept introduced in this paper is that all major fragment ions from an MS^3 experiment should readily be assigned to the partial sequence of the original parent peptide. If they cannot be assigned, then the original predicted sequence may be incorrect. If all major ions can be assigned, then the predicted sequence is most likely correct. While MS^3 experiments are theoretically possible on ion traps and FT-MS instruments, they found no real application in "shotgun" proteomics because older versions of these instruments did not trap enough ions in the MS/MS experiment to make MS^3 a viable option in real-time LC-MS experiments. The new linear ion trap (Schwartz, 2002) has much greater ion storage capacity, making MS^3 experiments in real-time LC-MS a possibility. The Olsen 2004 work used the LTQ-FT-MS instrument for such an experiment, by conducting all the MS^3 experiments in the linear trap of the instrument. We also immediately verified this was possible on our own linear ion trap for peptides from the AMD sample (Figure 8.1). A potential research avenue would be to optimize this methodology for the analyses of large numbers of peptides from protein standard mixtures and then a representative AMD community proteome sample. The basic experimental plan for an MS^3 experiment follows three major steps: 1) parent peptide mass measurement [will be high mass accuracy on LTQ-FT-MS]; 2) parent mass isolation and fragmentation; and 3) isolation and fragmentation of the top three fragment ions from the MS/MS spectra resulting in three MS^3 experiments. The high mass accuracy of the parent ion, as well as the three MS^3 experiments, can each be used to independently verify the peptide sequence. The combination of high mass accuracy peptide measurements, MS^3, and the development of a robust *de-novo* sequencing algorithm that takes into account all experimental information is a potential avenue for determining strain level variants at a high-throughput level.

The final major technological advances needed for the characterization of natural microbial communities is in the area of proteome extraction and sample preparation.

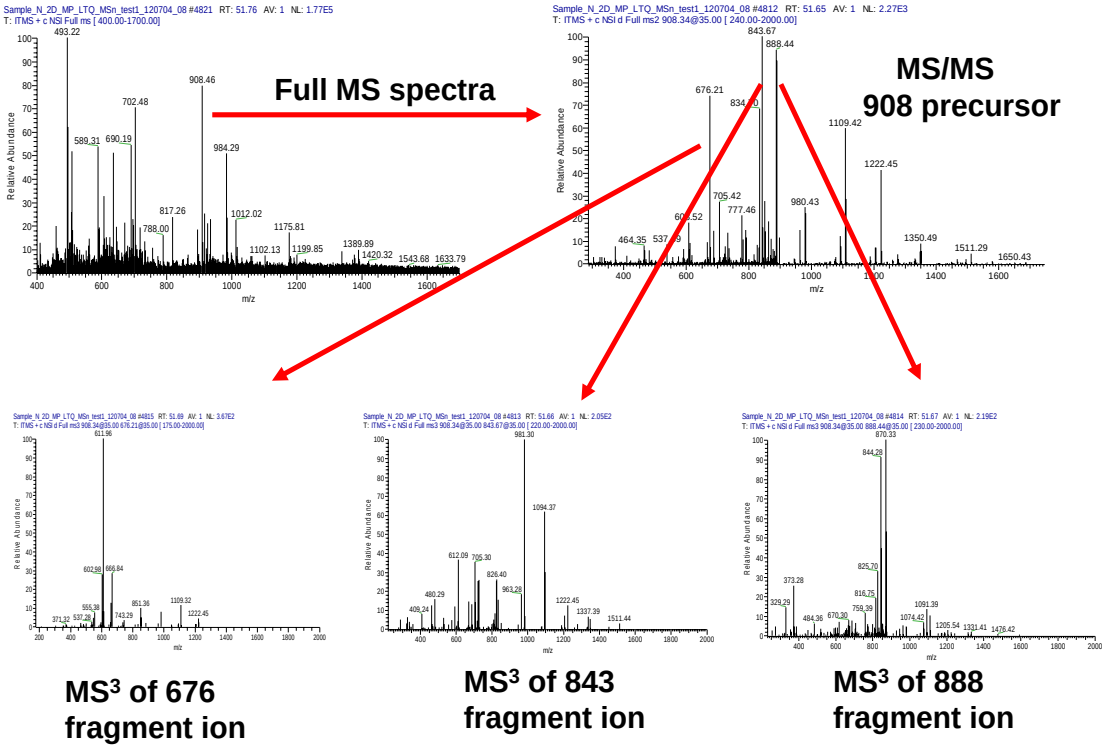


Figure 8.1: An MS^3 experiment on an LIT on a peptide from the AMD community. An MS^3 experiment on an unidentified peptide from the AMD community. The MS/MS spectra (top right) was of high quality yet no identification was made, suggesting the possibility of a strain variant. The steps in the process include a full MS scan (top left), followed by an MS/MS scan of abundant parent ions from the full scan. The three top abundant ions from each MS/MS scan are then subjected to further fragmentation in the MS^3 scans.

Probably, the most notable is the amount of starting material needed. The final goal should always be to reach techniques to measure proteomes directly from single cells, but that will probably not be accomplished for years to come. For the studies of isolates presented in the dissertation, the amount of starting cell mass is not as much of an issue since large quantities can always be cultured. For community samples, this is not always the case. While we were able to obtain large quantities of the AMD biofilm, one area of research we would like to explore is spatial diversity within a single biofilm. The goal of these studies will be to obtain cm^3 samples, extract the proteomes, analyze and compare. Our current sample preparation techniques are not optimized for these small types of proteome samples. Much research is needed to develop automated sample preparation techniques for such small starting materials.

While many technical challenges were overcome in the course of this dissertation, many more were realized. The field of proteomics is an exciting area with many opportunities for technological advances. In the next few years the rapid, reproducible and routine analysis of entire proteome from isolates can be expected. This will allow for global views of microbial physiology not attainable before. It will be necessary for MS-based proteomics to reach the level of whole genome sequencing where hundreds of microbial isolate proteomes under many different metabolic conditions are characterized every year. One of the greatest shortcomings of the field right now is the lack of dissemination of the dataset to the scientific community. This is mainly due to the competitive nature of the field right now. As discussed at great length in this dissertation this must be overcome and proteome datasets must be openly shared for the field to continue to mature.

The technology has already allowed for the characterization of hundreds of conserved unknown and unknown proteins at a rapid pace. One of the clearest challenges is the integration of the field of proteomics with rapid structural analysis, functional assays and genetic methods to develop rapid integrated methods to determine not only the identity of conserved unknown and unknown proteins but also their function. For

proteomics to become truly useful, at gaining insight into the function of the many unknown proteins present in any microbial species, this must be accomplished

The final major challenge is the application of this technology to natural communities. While small steps were taken in the course of this dissertation, much work is still needed. The complexity of natural communities is truly daunting but unless initial steps are taking to attack this important issue no progress will be made. Hopefully, the work presented in this dissertation brings us one step closer to the ultimate goals of comprehensive, reproducible and rapid characterizations of microbial proteomes from isolates to communities.

LIST OF REFERENCES

LIST OF REFERENCES

Andrews, S.C.; Robinson, A.K.; Quiñones R. (2003) Bacterial iron homeostasis. *FEMS Microbiol. Rev*., **27**, 215-237.

Arnold, R.J; Reilly, J.P. (1999) Observation of *Escherichia coli* ribosomal proteins and their posttranslational modifications by mass spectrometry. *Anal. Biochem.*, **269**, 105-112.

Baker, B.J.; Banfield, J.F. (2003) Microbial communities in acid mine drainage. *FEMS Microbiology Ecology,* **44**, 139-152.

Bagg, A.; Neilands, J.B. (1987) Ferric uptake regulation protein acts as a repressor, employing iron (II) as a cofactor to bind the operator of an iron transport operon in *Escherichia coli*. *Biochemistry*, **26**, 5471-5477.

Beliaev, A.S.; Saffarini, D.A. (1998) *Shewanella putrefaciens* mtrB encodes an outer membrane protein required for Fe(III) and Mn(IV) reduction. *J. Bacteriol*., **180**, 6292-6297.

Beliaev, A.S. et al. (2002a) Microarray transcription profiling of a *Shewanella oneidensis* etrA mutant. *J. Bacteriol*., **184**, 4612-4616.

Beliaev, A.S. et al. (2002b) Gene and protein expression profiles of *Shewanella oneidensis* during anaerobic growth with different electron acceptors. *OMICS*, **6**, 39-60.

Bennett, K.L.; Matthiesen, T.; Roepstorff, P. (2000) Probing protein surface topology by chemical surface labeling, cross linking, and mass spectrometry. *Methods Mol. Biol.*, **146**, 113-131.

Berish, S.A.; Subbarao, S.; Chen, C.Y.; Trees, D.L.; Morse, S.A. (1993) Identification and cloning of a fur homolog from Neisseria gonorrhoeae. *Infect. Immun*., **61**, 4599-4606.

Bereswill, S.; Greiner, S.; Van Vliet, A.H; Waidner, B.; Fassbinder, F.; Schiltz, E.; Kusters, J.G.; Kist, M. (2000) Regulation of ferritin-mediated cytoplasmic iron storage by the ferric uptake regulator homolog (Fur) of Helicobacter pylori. *J. Bacteriol*., **182**, 5948-5953.

Biemann, K. (1986) Mass spectrometric methods for protein sequencing. *Anal. Chem.*, **58**, 1288A-1300A.

Biemann, K. (1988) Contributions of Mass Spectrometry to Peptide and Protein Structure. *Biomed. Environ. Mass Spectrom*., **16**, 99-111.

Blake, R.C.; Shute, E.A.; Greenwood, M.M.; Spencer, G.H.; Ingledew, W.J. (1993) Enzymes of aerobic respiration on iron. *FEMS Microbiol. Rev.*, **11**, 9-18.

Blattner, F.R et al. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453-1474.

Bond, P.L.; Smriga, S.P.; Banfield, J.F. (2000) Phylogeny of Micro-organisms Populating a Thick, Subaerial, Predominantly Lithotrophic Biofilm at an Extreme Acid Mine Drainage Site. *Applied and Environmental Microbiology*, **66**, 3842-3849.

Buchanan, M.V. et al. (2002) Genomes to Life "Center for Molecular and Cellular Systems": a research program for identification and characterization of protein complexes. *OMICS*, **6**, 287-303.

Bumann, D.; Meyer, T.F.; Jungblut, P.R. (2001) Proteome analysis of the common human pathogen Helicobacter pylori. *Proteomics*, **1**, 473-479.

Bult, C.J. et al. (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science*, **273**, 1058-1073.

Carr, S.; Aebersold, R.; Baldwin, M.; Burlingame, A.; Clauser, K.; Nesvizhskii, A. (2004) The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data. *Mol. Cell Proteomics*., **3**, 531-533.

Casado, B. (2004) Proteomics for nasal secretion analysis. *Curr. Allergy Asthma Rep*., **4**, 224-229.

Chelius, D.; Bondarenko, P.V. (2002) Quantitative Profiling of Proteins in Complex Mixtures Using Liquid Chromatography and Mass Spectrometry. *J. Proteome Res.*, **1**, 317-323.

Corbin, R.W.; Paliy, O.; Yang, F.; Shabanowitz, J.; Platt, M.; Lyons, C.E.; Root K.; McAuliffe, J.; Jordan, M.I.; Kustu, S.; Soupene, E.; Hunt, D.F. (2003) Toward a protein profile of *Escherichia coli*: comparison to its transcription profile. *Proc. Natl. Acad. Sci*., **100**, 9232-9237.

Davis, M.T.; Spahr, C.S.; McGinley, M.D.; Robinson, J.H.; Bures, E.J.; Beierle, J.; Mort, J.; Yu, W.; Luethy, R.; Patterson, S.D. (2001) Towards defining the urinary proteome using liquid chromatography-tandem mass spectrometry II. Limitations of complex mixture analyses. *Proteomics*, **1**, 108-117.

Devreese, B.; Vanrobaeys, F.; Van Beeumen, J. (2001) Automated nanoflow liquid chromatography/tandem mass spectrometric identification of proteins from *Shewanella putrefaciens* separated by two-dimensional polyacrylamide gel electrophoresis. *Rapid Commun. Mass Spectrom.*, **15**, 50-56.

Dharmasiri, K.; Smith, D.L. (1996) Mass spectrometric determination of isotopic exchange rates of amide hydrogens located on the surfaces of proteins. *Anal. Chem.*, **68**, 2340-2344.

Dujon, B. (1996) The yeast genome project: what did we learn? *Tends Genet.*, **12**, 263-270.

Edman, P. (1950) Methods for determination of amino acid sequence in peptides. *Acta Chem. Scand.*, **4**, 283-293.

Edman, P.; Begg, G. (1967) A protein sequenator. *Eur. Jour. Biochem.*, **1**, 80-91.

Edwards, K.J.; Bond, P.L.; Gihring, T.M.; Banfield, J.F. (2000) An Archaeal Iron-Oxidizing Extreme Acidophile Important in Acid Mine Drainage. *Science*, **287**, 1796-1799.

Eng, J.K.; McCormack, A.L.; Yates, J.R. 3rd (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Mass Spectrom.*, **5**, 976-989.

Eymann, C.; Homuth, G.; Scharf, C.; Hecker, M. (2002) *Bacillus subtilis* functional genomics: Global characterization of the stringent response by proteome and transcriptome analysis. *J. Bacteriol.*, **184**, 2500-2520.

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

Fleischmann, R.D. et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae*. *Science,* **269**, 496-512.

Fraser, C.M. et al. (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science*, **270**, 397-403.

Fulda, S.; Huang, F.; Nilsson, F.; Hagemann, M.; Norling, B. (2000) Proteomics of *Synechocystis* sp. strain PCC 6803. Identification of periplasmic proteins in cells grown at low and high salt concentrations. *Eur. J. Biochem.*, **267**, 5900-5907.

Gao, J.; Opiteck, G.J.; Friedrichs, M.; Dongre, A.R.; Hefta, S.A. (2003) Changes in the protein expression of yeast as a function of carbon source. *J. Proteome Res.*, **2**, 643-649.

Gatlin, C.L.; Kleemann, G.R.; Hays, L.G.; Link, A.J.; Yates, J.R. 3rd (1998) Protein identification at the low femtomole level from silver-stained gels using a new fritless electrospray interface for liquid chromatography-microspray and nanospray mass spectrometry. *Anal. Biochem.*, **263**, 93-101.

Gavin, A.C. et al. (2002) Functional Organization of the Yeast Proteome by Systematic Analysis of Protein Complexes. *Nature*, **415**, 141-147.

Giometti, C.S. et al. (2003) Analysis of the *Shewanella oneidensis* proteome by two-dimensional gel electrophoresis under nondenaturing conditions. *Proteomics*, **3**, 777-785.

Griffin, T.J.; Gygi, S.P.; Ideker, T.; Rist, B.; Eng, J.; Hood, L.; Aebersold, R. (2002) Complementary profiling of gene expression at the transcriptome and proteome levels in *Saccharomyces cerevisiae*. *Mol. Cell Proteomics*, **1**, 323-333.

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

Grunenfelder, B.; Rummel, G.; Vohradsky, J.; Roder, D.; Langen, H.; Jenal, U. (2001) Proteomic analysis of the bacterial cell cycle. *Proc. Natl. Acad. Sci. USA*, **98**, 4681-4686.

Han, D.K.; Eng, J.; Zhou, H.; Aebersold, R. (2001) Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nat. Biotechnol.*, **19**, 946-951.

Hantke, K. (2001) Iron and metal regulation in bacteria. *Curr. Opin. Microbiol.*, **4**, 172-177.

Hart, A.; Murrell, J.C.; Poole, R.K.; Norris, P.R. (1991) An acid-stable cytochrome in iron-oxidizing *Leptospirillum* ferrooxidans. *FEMS Microbiol. Lett.*, **81**, 89-94.

Heidelberg, J.F. et al. (2002) Genome sequence of the dissimilatory metal ion-reducing bacterium *Shewanella oneidensis*. *Nat. Biotechnol.*, **20**, 1118-1123.

Ho, Y. et al. (2002) Systematic Identification of Protein Complexes in *Saccharomyces cerevisia*e by Mass Spectrometry. *Nature*, **415**, 180-183.

Hernychova, L.; Stulik, J.; Halada, P.; Macela, A.; Kroca, M.; Johansson, T.; Malina, M. (2001) Construction of a Francisells tularensis two-dimensional electrophoresis protein database. *Proteomics*, **1**, 508-515.

Hess, D.; Covey, T.C.; Winz, R.; Brownsey, R.W.; Aebersold, R. (1993) Analytical and micropreparative peptide mapping by high performance liquid chromatography/electrospray mass spectrometry of proteins purified by gel electrophoresis. *Protein Sci.*, **2**, 1342-1351.

Hillenkamp, F.; Karas, M.; Beavis, R.C.; Chait, B.T. (1991) Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. *Anal. Chem.*, **63**, 1193A-1203A.

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

Hunt, D.F.; Yates, J.R. 3rd; Shabanowitz, J.; Winston, S.; Hauer, C.R. (1986) Protein sequencing by tandem mass spectrometry. *Proc. Natl. Acad. Sci. USA*, **83**, 6233-6237.

Ideker, T.; Thorsson, V.; Ranish, J.A.; Christmas, R.; Buhler, J.; Eng, J.K.; Bumgarner, R.; Goodlett, D.R.; Aebersold, R.; Hood, L. (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, **292**, 929-934.

Jensen, O.N.; Larsen, M.R.; Roepstorff, P. (1998) Mass spectrometric identification and microcharacterization of proteins from electrophoretic gels: strategies and applications. *Proteins*, **2**, 74-89.

Jungblut, P.; Thiede, B. (1997) Protein identification from 2-DE gels by MALDI mass spectrometry. *Mass Spectrom. Rev.*, **16**, 145-162.

Ideker, T.; Galitski, T.; Hood L. (2001) A new approach to decoding life: systems biology. *Annu. Rev. Genomics Hum. Genet*., **2**, 343-372.

Keller, A.; Nesvizhskii, A.I.; Kolker, E.; Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem*., **74**, 5383-5392.

Kelleher, N.L.; Taylor, S.V.; Grannis, D.; Kinsland, C.; Chiu, H.J.; Begley, T.P.; McLafferty, F.W. (1998) Efficient sequence analysis of the six gene products (7-74 kDa) from the *Escherichia coli* thiamin biosynthetic operon by tandem high-resolution mass spectrometry. *Protein Sci*., **7**, 1796-1801.

Koller, A.; Washburn, M.P.; Lange, B.M.; Andon, N.L.; Deciu, C.; Haynes, P.A.; Hays, L.; Schieltz, D.; Ulaszek, R.; Wei, J.; Wolter, D.; Yates, J.R. 3rd (2002) Proteomic survey of metabolic pathways in rice. *PNAS*, **99**, 11969-11974.

Kim, M.K.; Harwood, C.S. (1991) Regulation of benzoate-CoA ligase in *Rhodopseudomonas palustris*. *FEMS Microbiol. Letts.*, **83**, 199-204.

Klose, J. (1975) Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals. *Humangenetik*, **26,** 231-243.

Kowalak, J.A.; Walsh K.A. (1996) Beta-methylthio-aspartic acid: identification of a novel posttranslational modification in ribosomal protein S12 from Escherichia coli. *Protein Science*, **5**, 1625-1632.

Larimer, F. W., et al. (2004) Complete genome sequence of the metabolically versatile photosynthetic bacterium *Rhodopseudomonas palustris.* *Nature Biotech*. **22**, 55-60.

Lipton, M.S.; Pasa-Tolic, L.; Anderson, G.A.; Anderson, D.J.; Auberry, D.L.; Battista, J.R.; Daly, M.J.; Fredrickson, J.; Hixson, K.K.; Kostandarithes, H.; Masselon, C.; Markillie, L.M.; Moore, R.J.; Romine, M.F.; Shen, Y.; Stritmatter, E.; Tolic, N.; Udseth, H.R.; Venkateswaran, A.; Wong, K.; Zhao, R.; Smith, R.D. (2002) Global analysis of *Deinococcus* radiodurans proteome by using accurate mass tags. *PNAS*, **99**, 11049-11054.

Langen, H.; Takacs, B.; Evers, S.; Berndt, P.; Lahm, H.W.; Wipf, B.; Gray, C.; Fountoulakis, M. (2000) Two-dimensional map of the proteome of Haemophilus influenzae. *Electrophoresis*, **21**, 411-429.

Liu, H.; Lin. D.; Yates, J.R. (2002) Multidimensional Separations for Protein/Peptide Analysis in the Post-Genomic Era. *Biotech*., **32**, 898-902.

Liu, H.; Sadygov, R.G.; Yates, J.R. 3rd (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.*, **76**, 4193-4201.

Link, A.J.; Eng, J.; Schieltz, D.M.; Carmack, E.; Mize, G.J; Morris, D.R.; Garvik, B.M.; Yates, J.R. 3rd (1999) Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.*, **17**, 676-682.

Litwin, C.M.; Boyko, S.A.; Calderwood, S.B. (1992) Cloning, sequencing, and transcriptional regulation of the *Vibrio cholerae* fur gene. *J. Bacteriol*., **174,** 1897-1903.

de Lorenzo, V.; Wee, S.; Herrero, M.; Neilands, J.B. (1987) Operator sequences of the aerobactin operon of plasmid ColV-K30 binding the ferric uptake regulation (fur) repressor. *J. Bacteriol*., **169**, 2624-2630.

de Lorenzo, V.; Giovannini, F.; Herrero, M.; Neilands, J.B. (1988) Metal ion regulation of gene expression. Fur repressor-operator interaction at the promoter region of the aerobactin system of pColV-K30. *J. Mol. Biol*., **203**, 875-884.

Macarthur, D.J.; Jacques, N.A. (2003) Proteome analysis of oral pathogens. *J. Dent. Res*., **82**, 870-876.

Mann, M.; Hendrickson, R.C.; Pandey, A. (2001) Analysis of proteins and proteomes by mass spectrometry. *Annu. Rev. Biochem.*, **70**, 437-473.

Martin, S.E.; Shabanowitz, J.; Hunt, D.F.; Marto, J.A. (2000) Subfemtomole MS and MS/MS peptide sequence analysis using nano-HPLC micro-ESI Fourier transform ion cyclotron resonance mass spectrometry. *Anal. Chem.*, **72**, 4266-4274.

Mawuenyega, K.G.; Kaji, H.; Yamauchi, Y.; Shinkawa, T.; Saito, H.; Taoka, M.; Takahashi, N.; Isobe, T. (2002) Large-scale identification of *Caenorhabditis* elegans proteins by multidimensional liquid chromatography – tandem mass spectrometry. *J. Proteome Res.*, in press for **Vol. 2**, Jan. 2003.

McCormack, A.L.; Schieltz, D.M.; Goode, B.; Yang, S.; Barnes, G.; Drubin, D.; Yates, J.R. 3rd (1997) Direct analysis and identification of proteins in mixtures by LC-MS/MS and database searching at the low-femtomole level. *Anal. Chem.*, **69**, 767-776.

McLuckey, S.A.; Stephenson, J.L. Jr. (1998) Ion/ion chemistry of high-mass multiply-charged ions. *Mass Spectrom. Rev.*, **17**, 369-407.

Meng F.; Cargile B.J.; Patrie S.M.; Johnson, J.R.; McLoughlin, S.M.; Kelleher, N.L. (2002) Processing complex mixtures of intact proteins for direct analysis by mass spectrometry. *Anal. Chem*., **74**, 2923-2929.

Mohan, D.; Pasa-Tolic, L.; Masselon, C.D.; Tolic, N.; Bogdanov, B.; Hixson, K.K.; Smith, R.D.; Lee, C.S. (2003) Integration of electrokinetic-based multidimensional separation/concentration platform with electrospray ionization-Fourier transform ion cyclotron resonance-mass spectrometry for proteome analysis of *Shewanella oneidensis*. *Anal. Chem*., **75**, 4432-4440.

Molloy, M.P.; Phadke, N.D.; Maddock, J.R.; Andrews, P.C. (2001) Two-dimensional electrophoresis and peptide mass fingerprinting of bacterial out membrane proteins. *Electrophoresis*, **22**, 1686-1696.

Mortz, E.; Vorm, O.; Mann, M.; Roepstorff, P. (1994) Identification of proteins in polyacrylamide gels by mass spectrometric peptide mapping combined with database search. *Biol. Mass Spectrom.*, **23**, 249-261.

Mortz, E.; O'Connor, P.B.; Roepstorff, P.; Kelleher, N.L.; Wood, T.D.; McLafferty, F.W.; Mann, M. (1996) Sequence tag identification of intact proteins by matching tandem mass spectral data against sequence data bases. *Proc. Natl. Acad. Sci. USA*, **93**, 8264-8267.

Myers, C.R.; Myers, J.M. (1992) Localization of cytochromes to the outer membrane of anaerobically grown *Shewanella putrefaciens* MR-1. *J. Bacteriol*., **174**, 3429-3438.

Myers, J.M.; Myers, C.R. (2001) Role for outer membrane cytochromes OmcA and OmcB of *Shewanella putrefaciens* MR-1 in reduction of manganese dioxide. *Appl. Environ. Microbiol.*, **67**, 260-269.

Nakanishi, T.; Okamoto, N.; Tanaka, K.; Shimizu, A. (1994) Laser-Desorption Time-Of-Flight Mass-Spectrometric Analysis Of Transferrin Precipitated With Antiserum - A Unique Simple Method To Identify Molecular-Weight Variants. *Biological Mass Spectrometry*,**23,** 230-233.

Neidhardt, F.C. (1987) *Escherichia coli* and *Salmonella typhimurium* Cellular and Molecular Biology. **Volume 2**, *American Society for Microbiology*.

Nielsen, H.; Engelbrecht, J.; Brunak, S.; von Heijne, G. (1997) A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Int. J. Neural. Syst*., **8**, 581-599.

Nesvizhskii, A.I.; Keller, A; Kolker, E; Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.*, **75**, 4646-4658.

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

O' Farrell, P.H. (1974) High Resolution Two-Dimensional Electrophoresis of Proteins. *Journal of Biol. Chem.*, **250**, 4007-4021.

Olsen, J.V.; Mann, M. (2004) Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. *Proc. Natl. Acad. Sci. USA*, **101**, 13417-13422.

Paša-Tolić, L.; Jensen, P.K.; Anderson, G.A.; Lipton, M.S.; Peden, K.K.; Martinović, S.; Tolić, N.; Bruce, J.E.; Smith, R.D. (1999) High-throughput proteome-wide precision measurements of protein expression using mass spectrometry. *J. Am. Chem. Soc.*, **121**, 7949-7950.

Pedrioli P.G. et al. (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotech.*, **22**, 1459-1466.

Peng, J.; Gygi, S.P. (2001) Proteomics: The Move to Mixtures. *J. Mass Spec*.,**36**, 1083-1091.

Peng, J.; Elias, J.E.; Thoreen, C.C.; Licklider, L.J.; Gygi, S.P. (2002) Evaluation of multi-dimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: The yeast proteome. *J. Proteome Res.*, **2,** 43-50.

Perkins, D.N.; Pappin, D.J.; Creasy, D.M.; Cottrell, J.S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551-3567.

Pratt, J.M.; Petty, J.; Riba-Garcia, I.; Robertson, D.H.; Gaskell, S.J.; Oliver, S.G.; Beynon, R.J. (2002) Dynamics of protein turnover, a missing dimension in proteomics. *Molecular & Cellular Proteomics*, **1**, 579-591.

Puig, O.; Caspary, F.; Rigaut, G.; Rutz, B.; Bouveret, E.; Bragado-Nilsson, E.; Wilm, M.; Seraphin, B. (2001) The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods*, **24**, 218-229.

Roepstorff, P.; Fohlman, J. (1984) Proposal for a Common Nomenclature for Sequence Ions in Mass-Spectra of Peptides. *Biomed. Mass Spec.*, **11**, 601.

Reynolds, K.J.; Yao, X.; Fenselau, C. (2002) Proteolytic ^{18}O labeling for comparative proteomics: evaluation of endoprotease Glu-C as the catalytic agent. *J. Proteome Res.*, **1**, 27-33.

Sauer, K. (2003) The genomics and proteomics of biofilm formation. *Genome Bio*l., **4**, 219.

Schrenk, M.O.; Edwards, K.J.; Goodman, R.M.; Hamers, R.J.; Banfield, J.F. (1998) Distribution of *Thiobacillus ferrooxidans* and *Leptospirillum ferrooxidans*: Implications for Generation of Acid Mine drainage. *Science*, **279**, 1519-1522.

Schwartz, J.C.; Senko, M.W.; Syka, J.E. (2002) A two-dimensional quadrupole ion trap mass spectrometer. *J. Am. Soc. Mass Spectrom.,* **13**, 659-669.

Sharp, J.S; Becker, J.M.; Hettich, R.L. (2003) Protein surface mapping by chemical oxidation: structural analysis by mass spectrometry. *Anal. Biochem*., **312,** 216-225.

Sharp, J.S; Becker, J.M.; Hettich, R.L. (2004) Analysis of protein solvent accessible surfaces by photochemical oxidation and mass spectrometry. *Anal. Chem*., **76**, 672-683.

Shen, Y.; Zhao, R.; Belov, M.E.; Conrads, T.P.; Anderson, G.A.; Tang, K.; Paša-Tolić, L.; Veenstra, T.D.; Lipton, M.S.; Udseth, H.R.; Smith, R.D. (2001) Packed capillary reversed-phase liquid chromatography with high-performance electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry for proteomics. *Anal. Chem.*, **73**, 1766-1775.

Shevchenko, A.; Jensen, O.N.; Podtelejnikov, A.V.; Sagliocco, F.; Wilm, M.; Vorm, O.; Mortensen, P.; Shevchenko, A.; Boucherie, H.; Mann M. (1996) Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. *Proc. Natl. Acad. Sci. USA*, **93**, 14440-14445.

Smalla, K.; Sobecky, P.A. (2002) The prevalence and diversity of mobile genetic elements in bacterial communities of different environmental habitats: insights gained from different methodological approaches. *FEMS Microbiol. Ecol*., **42**, 165-175.

Spahr, C.S.; Davis, M.T.; McGinley, M.D.; Robinson, J.H.; Bures, E.J.; Beierle, J.; Mort, J.; Courchesne, P.L.; Chen, K.; Wahl, R.C.; Yu, W.; Luethy, R.; Patterson, S.D. (2001) Towards defining the urinary proteome using liquid chromatography-tandem mass spectrometry. I. Profiling an unfractionate tryptic digest. *Proteomics*, **1**, 93-107.

Standing, K.G. (2003) Peptide and protein *de novo* sequencing by mass spectrometry. *Curr. Opin. Struct. Biol*., **13**, 595-601.

Strader, M.B. et al. (2004) Characterization of the 70S Ribosome from *Rhodopseudomonas palustris* using an integrated "top-down" and "bottom-up" mass spectrometric approach. *J. Proteome Res*., **3**, 965-978.

Stafford, G. (2002) Ion Trap Mass Spectrometry: A personnel perspective. *J. Am. Soc. Mass. Spectrom.*, **13**, 589-596.

Syka, J.E.; Marto, J.A.; Bai, D.L.; Horning, S.; Senko, M.W.; Schwartz, J.C.; Ueberheide, B.; Garcia, B.; Busby, S.; Muratore, T.; Shabanowitz, J.; Hunt, D.F. (2004) Novel linear quadrupole ion trap/FT mass spectrometer: performance characterization and use in the comparative analysis of histone H3 post-translational modifications. *J Proteome Res.*, **3**, 621-626.

Tabb, D.L.; Hayes-McDonald, W.; Yates, J.R. (2002) DTASelect and Contrast: Tools for Assembling and Comparing Protein Identifications from Shotgun Proteomics. *J. Proteome Res.*, **1**, 21-26.

Tabb, D.L.; Narasimhan, C.; Strader, M.B; Hettich. R.L. (2005) DBDigger: reorganized proteomic database identification improves flexibility and speed. *Anal. Chem*. Web Release Date: 05-Mar-2005.

Tatusov, R.L.; Koonin, E.V.; Lipman, D.J. (1997) A genomic perspective on protein families. *Science*, **271**, 631-637.

Tolmasky, M.E.; Wertheimer, A.M.; Actis, L.A.; Crosa, J.H. (1994) Characterization of the *Vibrio anguillarum* fur gene: role in regulation of expression of the FatA outer membrane protein and catechols. *J. Bacteriol*., **176**, 213-220.

Thompson, D.K. et al. (2002) Transcriptional and proteomic analysis of a ferric uptake regulator (Fur) mutant of *Shewanella oneidensis*: Possible involvement of Fur in energy metabolism, transcriptional regulation, and oxidative stress. *Appl. Environ. Microbiol*., **68**, 881-892.

Thomas, C.E.; Sparling, P.F. (1994) Identification and cloning of a fur homolog from Neisseria meningitis. *Mol. Microbiol*., **11**, 725-737.

Touati, D. (2000) Iron and oxidative stress in bacteria. *Arch. Biochem. Biophys*., **373,** 1-6.

Tyson, G.W. et al. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, **428**, 37-43.

Vanrobaeys, F.; Devreese, B.; Lecocq, E.; Rychlewski, L.; De Smet, L.; Van Beeumen, J. (2003) Proteomics of the dissimilatory iron-reducing bacterium *Shewanella oneidensis* MR-1, using a matrix-assisted laser desorption/ionization-tandem-time of flight mass spectrometer. *Proteomics*, **3**, 2249-2257.

Venter, J.C. et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66-74.

VerBerkmoes, N.C.; Strader, M.B.; Smiley, R.D.; Howell, E.E.; Hurst, G.B.; Hettich, R.L.; Stephenson, J.L. Jr. (2002) Intact protein analysis of Site Directed Mutagenesis Overexpression Products: Plasmid Encoded R67 Dihydrofolate Reductase. *Analytical Biochemistry*, **305**, 68-81.

VerBerkmoes, N.C.; Bundy, J.L.; Hauser, L.; Asano, K.G.; Razumovskaya, J.; Larimer, F.; Hettich, R.L.; Stephenson, J.L. Jr. (2002) Integrating "Top-Down" and "Bottom-Up" mass spectrometric approaches for proteomic analysis of *Shewanella oneidensis*. *J. Proteome Res.*, **1**, 239-252.

VerBerkmoes, N.C.; Hervey, W.J.; Shah, M.; Land, M.; Hauser, L.; Larimer, F.W.; Van Berkel, G.J.; Goeringer, D.E. (2005) Evaluation of "shotgun" proteomics for identification of biological threat agents in complex environmental matrixes: experimental simulations. *Anal. Chem.*, **77**, 923-932.

Wagner, M.A.; Eschenbrenner, M.; Horn, T.A.; Kraycer, J.A.; Mujen, C.V.; Hagius, S.; Elzer, P.; DelVecchio, V.G. (2002) Global analysis of the *Brucella melitensis* proteome: Identification of proteins expressed in laboratory-grown culture. *Proteomics*, **2**, 1047-1060.

Wan, X.; VerBerkmoes, N.C.; McCue, L.A.; Stanek, D.; Connelly, H.M.; Hauser, L.; Wu, L.; Liu, X.; Yan, T.; Leaphart, A.; Hettich, R.L.; Zhou, J.; Thompson D.K. (2004) Transcriptomic and Proteomic Characterization of the Fur Modulon in the Metal-Reducing Bacterium *Shewanella oneidensis*. *Journal of Bacteriology*, **186**, 8385-8400.

Washburn, M.P.; Wolters, D.; Yates, J.R. 3rd (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.*, **19**, 242-247.

Washburn, M.P.; Ulaszek, R.; Deciu, C.; Schieltz, D.M.; Yates, J.R. 3rd (2002) Analysis of quantitative proteomic data generated via multidimensional protein identification technology. *Anal. Chem.*, **74**, 1650-1657.

Wasinger, V.C.; Polack, J.D., Humphery-Smith, I. (2000) The proteome of *Mycoplasma genitalium*. Chaps-soluble component. *Eur. J. Biochem.*, **267**, 1571-1582.

Wiener, M.C.; Sachs, J.R.; Deyanova, E.G.; Yates, N.A. (2004) Differential mass spectrometry: a label-free method for finding differences in complex peptide and protein mixtures. *Anal. Chem.*, **76**, 6085-6096.

Wilm, M.; Shevchenko, A.; Houthaeve, T.; Breit, S.; Schweigerer, L.; Fotsis, T.; Mann, M. (1996) Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. *Nature*, **379**, 466-469.

Wolters, D.A.; Washburn, M.P.; Yates, J.R. 3rd (2001) An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.*, **73**, 5683-5690.

Yao, X.; Freas, A.; Ramirez, J.; Demirev, P.A.; Fenselau, C. (2001) Proteolytic ^{18}O labeling for comparative proteomics: model studies with two serotypes of *adenovirus*. *Anal. Chem.*, **73**, 2836-2842.

Wu, L.; Thompson, D.K.; Li, G.; Hurt, R.A.; Tiedje, J.M.; Zhou, J. (Dec 2001) Development and evaluation of functional gene arrays for detection of selected genes in the environment. *Appl. Environ. Microbiol.*, **67**, 5780-5790.

Yamaguchi, K.; Subramanian, A.R. (2000) The plastid ribosomal proteins. Identification of all the proteins in the 50S subunit of an organelle ribosome (chloroplast). *J. Biol. Chem.*, **275**, 28466-28482.

Young, C.S.; Beatty, J.T. (1998) Topological model of the *Rhodobacter capsulatus* light-harvesting complex i assembly protein LhaA (previously known as ORF1696). *J. Bacteriol.*, **180**, 4742-4745.

Young, M.M.; Tang, N.; Hempel, J.C.; Oshiro, C.M.; Taylor, E.W.; Kuntz, I.D.; Gibson, B.W.; Dollinger, G. (2000) High-throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. *Proc. Natl. Acad. Sci. USA*, **97**, 5802-5806.

Yusupov, M.M.; Yusupova, G.Z.; Baucom, A.; Lieberman, K.; Earnest, T.N.; Cate, J.H.D.; Noller, H.F. (2001) Crystal structure of the ribosome at 5.5 A resolution. *Science*, **292**, 883-896.

Zhou, J. (Jun 2003) Microarrays for bacterial detection and microbial community analysis. *Curr. Opin. Microbiol*., **6,** 288-294.

VITA

Nathan Christopher VerBerkmoes was born in Traverse City, MI on April 3rd, 1973. He was raised by his grandparents, Henry and Vickie VerBerkmoes, in Lake Linden, Michigan, where he graduated from Lake Linden-Hubbell High School in 1991. He first enrolled at Michigan State University and then transferred to the University of Tennessee where he graduated with a B.S. cum laude, Biochemistry and Cellular and Molecular Biology, in 2001.

He enrolled in the University of Tennessee-Oak Ridge National Laboratory Graduate School of Genome Science and Technology in 2001 to pursue his doctorate in Life Sciences. He graduated with a Ph.D. in 2005. He accepted a post-doctoral position at Oak Ridge National Laboratory in the Organic and Biological Mass Spectrometry Group.