To the Graduate Council:

I am submitting herewith a dissertation written by Brian Keith Erickson entitled "Characterization of the Extracellular Proteome of a Natural Microbial Community with an Integrated Mass Spectrometric / Bioinformatic Approach." 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 Hettich , Major Professor

We have read this dissertation
and recommend its acceptance:

Alison Buchan

Loren Hauser

Cynthia Peterson

Brynn Voy

Accepted for the Council:

Carolyn R. Hodges
Vice Provost and Dean of the Graduate School

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

Characterization of the Extracellular Proteome of a Natural Microbial Community with an Integrated Mass Spectrometric / Bioinformatic Approach

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

Brian Keith Erickson

December 2010

Copyright © 2010 by Brian Keith Erickson

All rights reserved.

DEDICATION

I dedicate this dissertation to my family, especially my wife, Alison Russell Erickson. Their support, encouragement and love has shaped who I am today and without them I would not have had the opportunities that have preceded this dissertation.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Robert Hettich for his support and consistent, thoughtful, and embedded approach throughout the course of this work. I would like to acknowledge my committee members: Dr. Alison Buchan, Dr. Loren Hauser, Dr. Cynthia Peterson, and Dr. Brynn Voy for their time, excellent comments, suggestions, and thought provoking discussions. I would like to thank the Organic and Biological Mass Spectrometry group at Oak Ridge National Laboratory for their patience and guidance. I would also like to thank the staff, Dr. Cynthia Peterson and Dr. Albrecht von Arnim of the Genome Science and Technology program at the University of Tennessee for the professional enrichment and personal support. I would like to thank my wife, Alison, and close friends, (Dr.) Scott Younger and Dr. Rich Giannone, for their invaluable friendship and support. Finally, I would like to thank my parents, Jim and Jennifer; and Jeff and Gretchen for their love and support.

ABSTRACT

Proteomics comprises the identification and characterization of the complete suite of expressed proteins in a given cell, organism or community. The coupling of high performance liquid chromatography (LC) with high throughput mass spectrometry (MS) has provided the foundation for current proteomic progression. The transition from proteomic analysis of a single cultivated microbe to that of natural microbial assemblages has required significant advancement in technology and has provided greater biological understanding of microbial community diversity and function.

To enhance the capabilities of a mass spectrometric based proteomic analysis, an integrated approach combining bioinformatics with analytical preparations and experimental data collection was developed and applied. This has resulted in a deep characterization of the extracellular fraction of a community of microbes thriving in an acid mine drainage system. Among the notable features of this relatively low complexity community, they exist in a solution that is highly acidic (pH < 1) and hot (temperature > 40°C), with molar concentrations of metals. The extracellular fraction is of particular interest due to the potential to identify and characterize novel proteins that are critical for survival and interactions with the harsh environment.

The following analyses have resulted in the specific identification and characterization of novel extracellular proteins. In order to more accurately identify which proteins are present in the extracellular space, a combined computational prediction and experimental identification of the extracellular fraction was performed. Among the hundreds of proteins identified, a highly abundant novel cytochrome was targeted and ultimately characterized through high performance MS. In order to achieve deep proteomic coverage of the extracellular fraction, a metal affinity based protein enrichment utilizing seven different metals was developed and employed resulting in novel protein identifications. A combined top down and bottom up analysis resulted in the characterization of the intact molecular forms of extracellular proteins, including the identification of post-translational modifications. Finally, in order to determine the effectiveness of current MS methodologies, a software package was designed to characterize the > 100,000 mass spectra collected during an MS

experiment, revealing that specific optimizations in the LC, MS and protein sequence database have a significant impact on proteomic depth.

TABLE OF CONTENTS

Chapter 1: Introduction to Proteome Mass Spectrometry and Rationale for Characterization of an Extracellular Fraction of a Natural Microbial Community……...1

Chapter 2: An Integrated Experimental / Computational MS Based Platform for the Proteomic Characterization of a Natural Microbial Community……..16

Chapter 3: Computational Prediction and Experimental Validation of Signal Peptide Cleavages in the Extracellular Proteome of a Natural Microbial Community……..32

Chapter 4: An Integrated, Comparative Metal Affinity Enrichment and Proteomic Characterization of Novel Proteins from a Natural Microbial Community……58

Chapter 5: High Resolution Mass Spectrometry for the Characterization of a Novel, Growth Stage Dependent Cytochrome……...84

Chapter 6: The Design and Implementation of Software for Integrating Bottom up and Top Down MS Datasets for the Characterization of an Extracellular Fraction from a Natural Microbial Community……..……....96

Chapter 7: Development of a Spectral Assignment Approach to Evaluate Assigned versus Unassigned Tandem Mass Spectra in the Proteomic Analyses of Microbial Isolates and Communities…….....117

Chapter 8: Conclusions and Significance of the Characterization of the Extracellular Fraction in a Natural Microbial Community…….....143

References…….….....150

Vita…….....161

LIST OF TABLES

3.1: Summary of Computational Prediction and MS Identification of Signal Peptide Cleaved Proteins from 5 Distinct Extracellular AMD Samples……41

3.2: Distribution of Predicted Signal Peptide Cleaved Proteins for the Dominant Microbes in the AMD Microbial Community……42

3.3: Results of Edman N-terminal Sequencing of Select Secretome Proteins from the AMD Microbial Community……46

3.4: Highly Conserved and Replicated Signal Peptide Cleaved Proteins with Confirming N-terminus Spectra……49

3.5: NSAF Comparison of Select, Highly Differential Signal Peptide Cleaved Proteins……...51

3.6: Pfam Domain Analysis of Conserved and High Confidence Signal Peptide Cleaved Proteins…….....……53

4.1: Number of Proteins Labeled as Uniquely IMAC Bound or Unbound After Analysis of MS Data from Chromatographic Fractions……..…68

4.2: Average Amino Acid Abundance within Protein Groups that are Bound to Specific Metals, or Groups of Metals…….....71

4.3: Identified Flagellar Proteins and Their Pattern of IMAC Enrichment…...…….75

4.4: Pfam Identifications from Enriched and Novel Extracellular Proteins…...…..82

6.1: Subset of Ribosomal Proteins Identified by TD and BU MS…….105

6.2: Example TD Protein Identifications with Extremely Low PPM…….111

6.3: PTM – Methionine Cleavage..…….113

6.4: Results of TD and BU Analysis of the 29 Extracellular Fraction from the AMD Microbial Community……...….114

7.1: Number of Predicted Proteins in the Protein Sequence Databases ……..123

7.2: Results of Spectral Analysis on 4 Samples Including Microbial Isolates and Communities….……..126

LIST OF FIGURES

1.1: Integration of Genomic Sequence Information for Spectrum Identification….2

1.2: A View Inside the Mine in Redding, CA…….10

1.3: Illustration of the "-omics" Funnel …….……13

2.1: Illustration of the Interconnected Disciplines for Proteomic Analyses of Complex Microbial Communities…….17

2.2: Illustration of the Mine Tunnel System in Redding, CA……20

2.3: Illustration of a FTICR Mass Spectrometer…….……26

2.4: Illustration of an Ion Trapping Mass Spectrometer…….……..29

3.1: Representation of Signal Peptide and Cleavage……..…..35

3.2: Computation and Experimental Determination of Signal Peptide Cleavage.....37

3.3: Venn Diagram of Predicted and Measured Signal Peptides……44

3.4: Functional Distribution of Identified Signal Peptide Cleaved Proteins……48

3.5: NSAF Cluster Analysis of Signal Peptide Cleaved Proteins Identified in 28 Biofilm Samples……55

4.1: Schematic Representation of IMAC Enrichment……65

4.2: Number of Proteins Identified per Column……..……67

4.3: General Classification of Metal Binding……70

4.4: Heat Map of Proteins Found in Chromatographic Fractions……..76

4.5: Distribution of Functions Among Reductive Non-reductive and Ni/Co Metals……..78

5.1: MS Spectra of C-drift Cyt_{579}.……88

5.2: Deconvoluted IRMPD Fragmentation Spectrum of Cyt_{579} -16,119 Da Species……..90

5.3: Intact Mass Measurement of C-drift DS2 Exhibiting Two States of Additional N-terminal Truncation……91

5.4: MS Spectra of AB-Muck……93

5.5: MS Measurement of C75m……94

6.1: Flowchart of Integration of Accurate Intact Protein Mass (AIPM) and Bottom up Searching Algorithms…….....103

6.2: High Resolution Mass Spectrum of Ribosomal Subunits……..……107

6.3: Protein Identifications from the BU and TD MS Analysis of the Extracellular Fractions……..110

7.1: Distribution of Spectra Among All Samples.....130

7.2: Number of Spectra Assigned to a Peptide Among Microbial Isolates and Communities..……133

7.3: Classification of Spectra in a MS Experiment…….……..137

7.4: Distribution of +1 Charged, Unassigned MS2 Among Microbial Isolates and Communities……..…….……138

ABREVIATIONS AND SYMBOLS

ACN	Acetonitrilie
AIPM	Accurate Intact Protein Mass
AMD	Acid Mine Drainage
BU	Bottom Up
CID	Collision-Induced Dissociation
Cyt_{579}	Cytochrome 579
DS1 - 2	Developmental Stage 1 - 2
DTT	Dithiothreitol
ESI	Electrospray Ionization
ETD	Electron Transfer Dissociation
FDR	False Discovery Rate
FTICR	Fourier Transform Ion Cyclotron Resonance
FWHM	Full Width at Half Maximum
HMM	Hidden Markov Model
IMAC	Immobilized Metal Affinity Column
IRMPD	Infrared Multiphoton Dissociation
***Lepto*II/III**	*Leptospirillum* Group II/III
LC	Liquid Chromatography
MAIM	Most Abundant Isotopic Mass
MS	Mass Spectrometry
MS/MS	Tandem (fragment) Mass Spectrum
MS1	Full (survey) Spectrum
MS2	Tandem (fragment) Mass Spectrum
M/Z	Mass to Charge
MudPIT	Multidimensional Protein Identification Tool
NSAF	Normalized Spectral Abundance Factor
pI	Isoelectric Point
PPM	Parts per Million

PTM	Post-translational Modification
RP	Reverse Phase
SCX	Strong Cation Exchange
SRP	Signal Recognition Particle
TD	Top Down

Chapter 1

Introduction to Proteome Mass Spectrometry and Rationale for Characterization of an Extracellular Fraction of a Natural Microbial Community

---

1.1: Introduction

Understanding the molecular foundations that enable the fundamental mechanisms of life has been the backbone of dedicated biological research. “How does this process occur?”, “Why does this mechanism proceed?”, and “When does the process begin and end?” are examples of the general questions directed towards biological systems. Addressing these questions inevitably leads to the identification and characterization of the fundamental biomolecules, such as proteins, that are responsible for executing these functions. Towards this aim, the study of proteins has resulted in tremendous insight into the biological mechanisms that enable life. The research into the group(s) of proteins responsible for critical biological processes has required tremendous leaps in technology and data integration. A key experimental platform that has that has proven to be successful in identifying and characterizing proteins is biological mass spectrometry.

Technological advances have generally preceded biological discovery and illustrate their tightly coupled interplay. An excellent example of this coordination is the rapid explosion of genomic technologies and the resulting biological information which has provided immeasurable amounts of data. Initially, purified proteins were measured and characterized with the high resolution and accurate mass capabilities of mass spectrometers. It was later realized that as opposed to measuring single proteins, it was possible to measure the entire suite of proteins present. The availability of genomic information has provided the basis for MS based proteomic analyses. By utilizing the genome information and resulting predicted protein sequences it is possible, and routine, to identify thousands of proteins in a single MS experiment (**Figure 1.1**).[1]

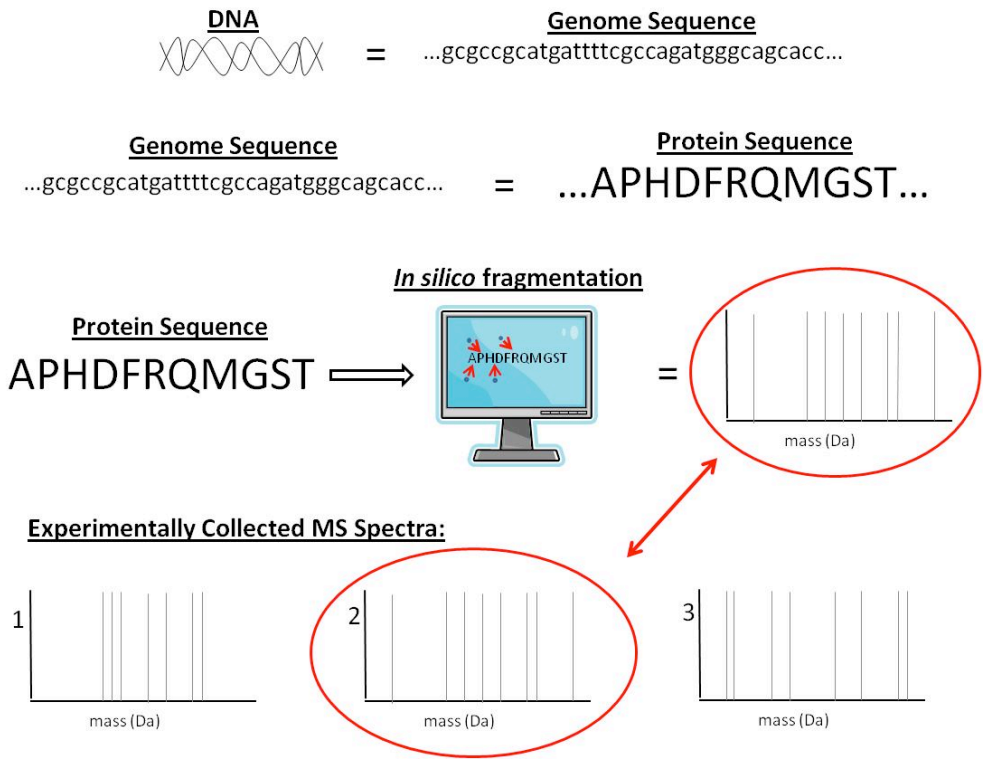

Figure 1.1: Integration of Genomic Sequence Information for Mass Spectrum Identification
The sequencing of DNA and resulting genome sequence is used to predict protein sequences. The protein sequences are then used to match a computational fragment spectrum to the experimentally derived fragment spectrum (circled in red).

Proteomics comprises the identification and characterization of the complete suite of proteins expressed by a particular sample.[2, 3] The sample can consist of a particular organism, cell type of an organism, or more recently, a community of organisms. The proteomic information obtained through this methodology provides an excellent view of the cellular functions and also provides temporal or spatial comparisons. In order to advance the field and application of MS based proteomics, the integration of various computational aspects were explored and applied in this dissertation. Additionally, the focus towards a particular subcellular fraction has resulted in a significantly deeper and more comprehensive view of the proteins present. The use of MS based proteomics is an excellent example of technology integration and adaptation, and has resulted in the identification and characterization of thousands of proteins, as well as tangible biological discoveries, including the characterization of growth state dependent protein export and a redox variable novel cytochrome, as discussed below for an acid mine drainage (AMD) microbial community.

Additionally, combinations of experimental and computational advancements have resulted in new approaches for deep proteomic characterization. Ultimately, this work succeeds in furthering the application of biological mass spectrometry by providing the next iteration of technology integration. Key discoveries discussed within provide anchors for further development as well as pointers towards protein targets of potentially critical value. For the sample sets described within, there exists hundreds of thousands of proteins that could be potential targets for specific biochemical analysis. Through this work, it is possible to effectively identify a functionally relevant and concise group of target proteins for future analysis.

1.2: Current State of Mass Spectrometric Proteomics

A principal factor in the transition to, and application of, MS based proteomics is the advancements and availability of genome sequencing Although biological mass spectrometry of isolate proteins or metabolites can proceed effectively without genomic information, the progression towards complex proteomic analysis generally requires the foundation of a genome sequence.[4] The advancements in genome sequencing are

mirrored by the increase in proteomic coverage. Concurrently, the continuing improvement in community genomics has paved the way for community proteomics analysis. One area of intense focus centers on microbial proteomics[5, 6]. Once previously reliant on lab cultured species, the direct measurement of not only one species but a community of species, sampled directly from their natural environment, is significantly more coomon. This has resulted in novel views of microbial existence and their impact of various systems, including the ocean, soil, and the human gastrointestinal tract.[7-9]

The evolution of MS based proteomics is based on the use of 2D gel electrophoresis for protein separation and identification, and the desire to overcome many of the inherent shortcomings of gel based proteomics.[10, 11] Many of the previous inefficiencies and challenges have been addressed or minimized with the application of MS based proteomics.[12] The foundation of mass spectrometry is the mass measurement of ions. Depending on the physical hardware of the mass spectrometer, a range of performance metrics are available (mass range, throughput, and mass resolution and accuracy). The parameters of the measurement are often inversely related, such that sacrifices in throughput are necessary for gains in mass resolution and accuracy or vice-versa.[13] The introduction of complicated samples, such as a microbial community containing millions of peptide ions, will present an over abundance of ions to the mass spectrometer. The sheer number of ions present will limit the range of ion detection and necessitates a separation of ions. The coupling of high performance liquid chromatography to the mass spectrometers results in a distributed elution of ions into the MS. An on-line separation, directly coupled to the ion source of the mass spectrometer, provides a robust and rapid method of separation. For complex samples, the application of multidimensional separations, combining two or more orthogonal separations is ideal. Development of an online platform has resulted in the wide use of 2D chromatographic online separations.[14] The ability to pre-load a specific quantity of peptides, which are in turn chromatographically separated and directly injected into the mass spectrometer for mass measurement, results in an extremely

rapid characterization of complex samples. Currently, the identification of thousands of proteins requires less than twenty-four hours of instrument time.

MS based proteomics has progressed rapidly in recent years. Technological advancements in sample preparation, liquid chromatography, and mass spectrometric instrumentation have allowed for previously unimagined high-throughput proteomic depth. Bottom up (BU) MS is the most common and widely used platform for MS based proteomics. A BU analysis involves the proteolytic digestion of proteins into representative peptides which are then chromatographically separated and detected in the mass spectrometer.[15, 16] The advancement of MS based proteomics is evident in the types of samples analyzed as well as in the results of the analysis. The first large scale proteomic analysis of *Saccharomyces cerevisiae* resulted in the identification of 1,484 proteins.[17] In 2002, Florens *et al* published a proteomic view of *Plasmodium falciparum*, the causative agent of malaria.[18] In this study, the depth of the proteomic coverage was increased with the identification of over 2,400 proteins. The general MS measurement and identification of isolate or single species is now fairly well established. With the publication of the AMD community genome in 2004, it was feasible to identify the community proteome of the dominant members within the AMD community.[19, 20] Over 2,000 proteins were initially identified and this number has now exceeded 10,000 protein identifications across multiple samples representing a diverse collection of growth states and locations.

A complementary approach to BU is the direct analysis of intact proteins without prior enzymatic digestion. The direct measurement of intact proteins, along with the elucidation of protein sequence through fragmentation, is termed top down (TD) MS.[21] The direct measurement of the native protein provides specific information on the state of the protein as it exists within the sample. This is significantly different from what is measured during a BU experiment, as the protein is inferred from the peptides but never directly measured. These techniques are considered to be complementary as opposed to competing, as each provides different metrics about the protein.[22] A highly useful application for TD MS is the identification of post-translational modifications (PTMs). The various modifications that are present on proteins are often key factors in cellular

signaling. The broad range of potential PTMs presents a challenge for unambiguous identification. For example, many post-translational modifications are highly labile. Through the course of a BU experiment, it is likely that a modification, such as a phosphorylation, will be inadvertently lost and not measured. For a TD experiment, the presence of the phosphate group can be directly measured providing evidence for the existence of that modification. Additionally, TD MS provides the ability to identify multiple forms of proteins that are differentially modified. Applications of TD MS can include the targeted analysis of isolated proteins or, more recently, proteomic analyses of complex samples. The direct measurement of large proteins or subunits of a protein complex presents a unique opportunity to characterize the molecular forms of proteins. Proteins exceeding 100,000 Da as well as membrane associated proteins and complexes can be isolated, measured and fragmented for specific sequence level information.[23, 24]

The application of MS to a wide range of samples has required technological gains in chromatography and MS instrumentation. The computational challenges associated with the proteomic analyses of increasingly complex samples were also initially addressed and have now spawned numerous available algorithms with unique attributes. Peptide MS analysis results in the collection of mass spectra, which are interrogated by computational algorithms that utilize the protein sequence database, determined from the genomic sequence. The protein sequence database is used to generate peptide fragments, *in silico*. These computationally generated, theoretical fragmentation spectra are compared and scored against the experimentally derived spectra. Matches between the spectra then provides reliable sequence information about the peptide which can then be mapped back to the parent protein, resulting in a protein identification. The SEQUEST algorithm is a staple in database peptide assignment.[25] More current approaches have increased throughput as well as provided means for BU PTM identification.[26, 27]

The application of MS based proteomics has provided a reproducible, high-throughput methodology for large-scale protein identification. The advancements in technology and data analysis are providing increased proteomic depth. What still

remains a fundamental challenge is the ability to extract relevant biological information from the datasets. This is especially true with the AMD community sample set. Within the protein sequence database, over 60% of the proteins have an unknown function. This, in turn, results in many of the proteomic identifications having an unknown function. One potential method to begin to characterize these proteins is through an integrated approach. Advancements in each stage of the proteomic characterization: sample preparation, separation, measurement, and bioinformatic processing, can provide a range of information which can used in order to begin to associate protein function.

1.3: Introduction to Microbial Proteomics

Estimations of prokaryotic abundance on Earth's surfaces exceed $1x10^{30}$ cells, with the number of estimated species ranging from 10^5 to 10^6.[28] The enormous population and astounding variety of species highlights the often unseen role that microbes play in maintaining a homeostatic cycle.[29] Microbial impact ranges from environmental remediation of heavy metals, beneficial stabilization of the human gut, to harmful acidification of acid mine drainage.[6, 9, 20, 30] Historically, microbial isolates cultured in the lab were the target model system for breakthroughs in genomic research. It has become clearer now that the genomic recombination, community interaction, and strain diversification witnessed in natural environments necessitates that analysis be directed towards natural microbial communities. The fact that ~80 – 99% of microorganisms cannot be successfully cultured within the lab dictates that sampling directly from the environment is necessary. Only recently has the technology become available to begin to understand these microbial species as they exist *in situ*.

Proteomic analysis of samples derived directly from their natural environment present numerous unique challenges. The physical collection of the sample requires, in many cases, a laborious effort, as is the case with the thriving microbial community living in a mine discussed within. Other recently studied communities, each with unique and consistent challenges, include the ocean, soil, and the human gastrointestinal tract.[7-9] In each case, the ability to collect a suitable amount of biomass must

considered. Additionally, the environment in which the microbial community exists in can present additional preparatory complications. A significant challenge lies in the ability to efficiently extract protein from the microbial cells that are present in a wide variety of matrices.

Microbial diversity presents both a lifelong avenue of scientific exploration and an often frustrating path towards biological inference. The limited availability of existing microbial protein characterization often restricts the rapid characterization of microbial proteins identified from uncultured species. Additionally, the relatively small number of sequenced microbial genomes provides a small foundation for attempting to infer protein function from natural samples. As mentioned previously, MS based proteomics relies heavily on the community genome sequences and the availability of matched genomes greatly enhances the ability to accurately identify proteins. Therefore, a primary challenge that must be addressed is the reliance of natural community samples on a suitable genomic sequence for MS based characterization.

A single MS analysis of a community sample is now capable of generating tens of gigabytes (GB) of data. This amount of data must be efficiently processed and organized in order for meaningful identifications and characterizations. The basis for this large amount of data lies in the complexity found within the natural samples. Compared to isolates which may express 2,000 – 3,000 proteins at a specific time, a community may contain hundreds of members each expressing thousands of proteins. In order to measure the large amount of proteins present the mass spectrometer must collect spectra for a significantly longer time resulting in a significant gain in data. Many of the informatics tools were not initially intended to handle datasets of the size and manner that are produced from microbial community proteomics. The advancement of computational tools is yet another challenge inherent with the analysis of microbial communities.

1.4: A Thriving Microbial Community in Acid Mine Drainage

A low-diversity community that populates acid mine drainage (AMD) biofilms has served as a model system for the development of community proteomics as well as for

for investigations into community development and structure (**Figure 1.2)**.[6, 31] Most of the AMD biofilms are dominated by *Leptospirillum* Group II (LeptoII), a Fe(II)-oxidizing, chemoautotrophic bacterium.[20, 32] The biofilms exhibit distinct developmental stages that vary in microbial community composition. Early developmental stages (DS1) are dominated by LeptoII; however, late developmental-stage biofilms (DS2) diversify, with increasing abundance of *Leptospirillum* Group III (LeptoIII), archaea, and eukaryote populations.[33] The microbial community exists in an extreme solution consisting of molar concentrations of Fe, sub-molar concentrations of Zn, Cu, As, a pH < 1, and temperatures exceeding 40°C.

The AMD community is an ideal sample set for the continued technological advancement of biological MS based proteomics and detailed characterization of natural microbial development and structure. The relatively low complexity of the AMD community presents a graduated challenge for characterizing natural community samples. The presence of only five dominant organisms has resulted in a highly refined community genome and resulting proteome. Furthermore, the limited number of abundant species has resulted in deep proteomic coverage of the most represented species. Advancements at all levels of the proteomics pipeline are possible and adjustments to the sample preparation, chromatography, and informatics have resulted in > 10,000 proteins being identified from this community. The biological characterization of the microbial community has resulted in several notable discoveries. For example, a dramatic shift in protein expression is observed depending on the developmental state. This is related to a shift in the dominant organism, which has been hypothesized to reflect the initial colonization by one species and then a subsequent shift towards a second species for continued growth. This also correlates well with the characterization of multiple metabolic pathways, several of which are spread among multiple microbial members, providing evidence for metabolic portioning. Finally, the presence of PTMs, including signal peptide cleavage, n-terminal methionine cleavage and additions of oxidation, methylation and acetylation provide clues for protein stability and signaling.

Figure 1.2: A View Inside the Mine in Redding, CA

The highly acidic stream is shown flowing through a section of the mine tunnel system. Image courtesy of Dr. Jillian Banfield

Acid mine drainage is a worldwide phenomenon and results in significant environmental contamination. The exposure of iron-sulfides (pyrite) common from mining results in the acidification of drainage water. The colonization of microbes results in a 10^6 increase in the acidification rate. The acidic AMD solution can potentially contaminate streams, municipal water and irrigation sources. Understanding the microbes that colonize this system will provide future treatments in order to slow, or ideally prevent the rapid acidification.

1.5: AMD Extracellular Fraction

Of particular interest are extracellular proteins that mediate interactions between the microorganisms and the environment. Due to the complexity of the samples from the AMD community, cellular fractionation has been employed in order to provide a more manageable set of proteins. An obvious fraction for characterization is the portion of proteins that reside and function outside of the cell. It is not unreasonable to expect that there exists in this fraction numerous novel proteins that are responsible for maintaining the tight coordination between the extreme environment and the microbes. A focused analysis provides several benefits with respect to both technology and biology. From a technical standpoint, the reduced complexity of the extracellular fraction provides opportunities to increase proteomic depth by enabling more specific enrichments of the fraction or adjustments to the chromatography. Additionally, applications of informatic techniques, including sequence based analysis, is more amenable to the smaller set of proteins identified in the extracellular fraction. The biological insight gained by studying the extracellular fraction can provide numerous clues to microbial survival in the extreme environment. At a basic level, a comprehensive identification of the proteins that reside outside the cell can illustrate mechanisms that enable microbial existence. The breadth and width of the identified proteins provides a view of how the microorganisms cope in the AMD solution. Expected functions of these proteins include those involved in transport of various solutes, including metals; enzymatic proteins responsible for protein turnover and defense, and cytochromes for metabolic processes and electron transport.

Finally, as the analysis of the AMD microbial community has progressed, it has become clear that in order to more fully understand how the organisms populate, thrive, and interact in the environment it is necessary to focus more attention on specific proteins or groups of proteins. **Figure 1.3** illustrates the principle in progressing towards more targeted analyses. As the genome and initial proteome have been described and updated, it follows that a progression towards a more specific cellular fraction is necessary, in an effort to identify key proteins within those fractions. The result of this approach is the generation of a select subset of specific protein targets, who, through targeted MS identification and computational analysis can be more fully characterized.

1.6: Application and Advancement of MS Proteomics

BU MS analyses result in the identification of thousands of proteins. TD analysis will typically result in fewer identification but is compensated by providing specific details of an intact protein. In this manner, BU is widely used as a tool for the comprehensive view of a proteome, while TD is more efficiently applied in a targeted manner. In either case, the identification of proteins is not the sole information point obtained. During a BU experiment, the mass spectrometer will target, isolate, and fragment thousands of peptides. Peptides that are more abundant in a sample will be targeted more often. This information is recorded during the experiment and is termed a spectral count. The greater the spectral count, the more abundant a particular peptide is relative to the other peptides in the sample. This does not necessarily provide absolute quantification of a peptide or protein, although inclusion of stable isotope based labels can provide this metric. The result of the MS analysis is the identification of a protein, relative quantification, and in many cases, information about PTMs. This dissertation attempts to expand the application of the generated data by integrating and creating novel software approaches. In order to more fully characterize the extracellular fraction, the existing MS proteomics pipeline was adapted and adjusted to provide increased protein identifications as well as specific characterizations of as many proteins as

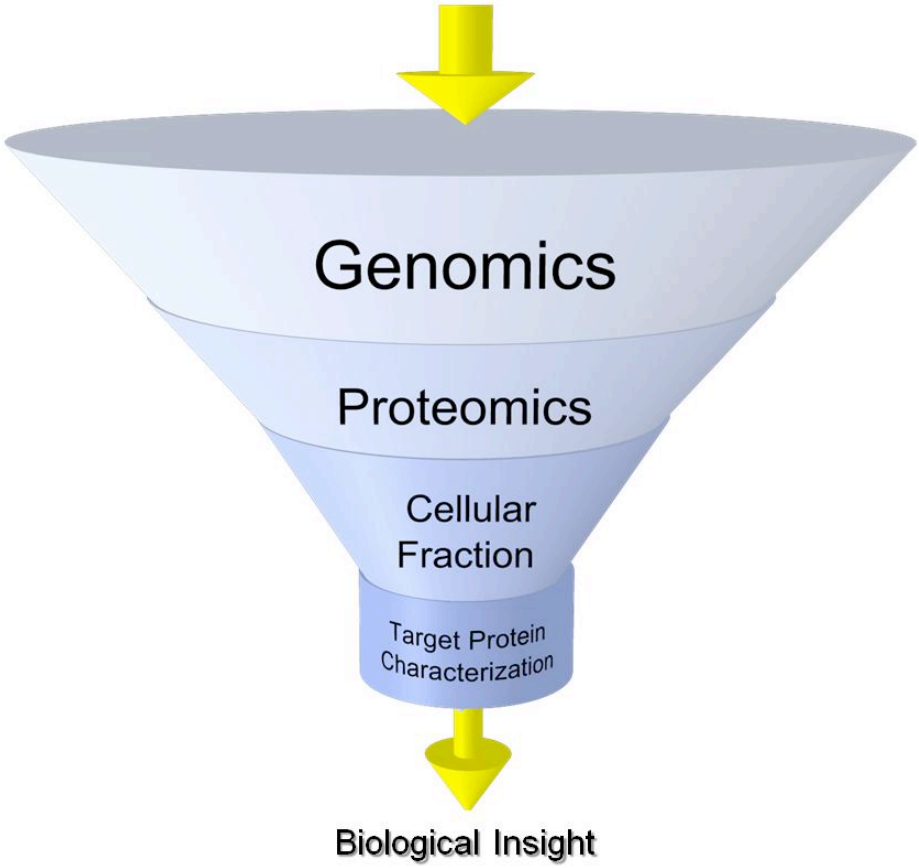

Figure 1.3: Illustration of the "-omics" funnel
The funnel describes the progression towards the analysis of subcellular fractions in an effort to more specifically identify and characterize critical proteins.

possible. This dissertation represents a progression of MS based proteomics and the foundation for future analysis of community samples.

1.7: Scope of the dissertation

This dissertation will describe the technological and informatic advancements that have been achieved during the analysis of the AMD natural microbial community. As described above, the focus on the extracellular fraction has resulted in not only a significantly more comprehensive identification of the protein members present in the extracellular fraction, but has also identified and characterized functionally critical proteins. Chapter 2 will provide a detailed methodological overview of MS based proteomics. The discussion points will provide a description of the complete sample flow from collection, preparation, and digestion, MS spectrum collection, and data analysis. Chapter 2 will also introduce a number of software tools developed for MS based proteomics. Chapter 3 highlights a targeted identification and characterization of proteins present in the extracellular space by a combined computational signal peptide prediction and experimental protein verification. In this chapter a confident identification of the proteins residing out the cell is assembled. This then lays the groundwork, by providing the initial protein identification, for the remaining chapters. Chapter 4 introduces and highlights the use of metal affinity columns for the enrichment and identification of novel extracellular proteins. In this chapter the application of seven different metal affinity columns will be discussed as well as an analysis of the greater than 100 novel proteins identified through this methodology. Chapter 5 highlights the targeted MS characterization of a cytochrome that exhibits a variable redox state correlating well with the growth state of the biofilm. Chapter 6 will introduce novel software that is designed to efficiently integrate top down and bottom up MS datasets. This software was then applied to an integrated MS analysis and resulted in the identification of greater than 300 intact proteins and numerous PTMs. Chapter 7 discusses the design, development, and application of software that characterizes the hundreds of thousands of spectra that are collected during a bottom up MS analysis. Through this characterization, it is possible to more accurately gauge the effectiveness

of the analysis as well as identify target spectra for more advanced interrogation. This chapter will also introduce the marked differences between the MS analysis of microbial isolates and communities in the ability to assign fragment spectra to peptides. This work presents a significant contribution to MS based proteomics analysis, with a focus towards a natural microbial community. Chapter 8 summarizes the technological integration and resulting biological inference as a framework for future studies, as well as additional insight about the colonization and existence of the microbial members in the AMD community.

Chapter 2

An Integrated Experimental / Computational MS Based Platform for the Proteomic Characterization of a Natural Microbial Community

Portions of included text are adapted from:

Melissa R. Thompson, Karuna Chourey, Jennifer M. Froelich, Brian K. Erickson, Nathan C. VerBerkmoes, Robert L. Hettich, **"**Experimental Approach for Deep Proteome Measurements from Small-Scale Microbial Biomass Samples**",** Analytical Chemistry, 2008**,** 80 (24), 9517-9525.

2.1: Introduction

The methodology for the proteomic characterization of a complex community contains aspects of analytical technologies, bioinformatics, and fundamental biology (**Figure 2.1**). It is the interconnected nature between the three key disciplines that have allowed MS based proteomics to become the principle platform for rapid and accurate proteomic characterization.

In order to efficiently characterize the thousands of proteins present in the extracellular fraction of the AMD microbial community, an optimized liquid chromatographic (LC) – tandem mass spectra (MS/MS) platform was utilized.[34] The online separation of the complicated microbial sample is necessary in order to allow the mass spectrometer sufficient time to accurately measure a mass or to target, isolate and fragment a specific ion. Additionally, the LC separation provides tremendous gains in dynamic range. The microbial sample that is ionized into the mass spectrometer contains a range of proteins or peptides that are present at highly variable concentrations. The chromatographic separation aides in the ability to measure even low abundance proteins or peptides by limiting the total ion population present in the mass spectrometer. Finally, coupling the liquid chromatography to the mass

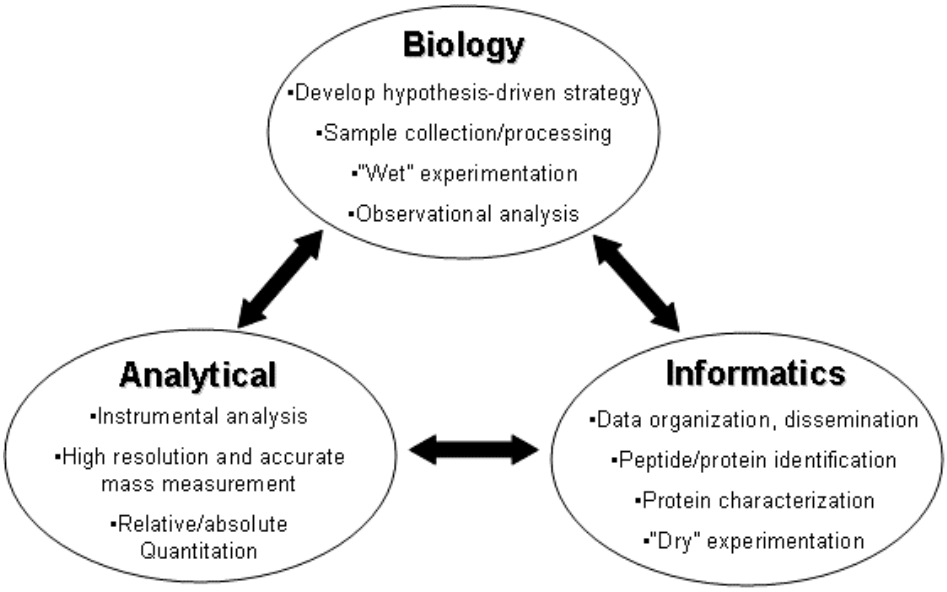

Figure 2.1: Illustration of the Interconnected Disciplines for Proteomic Analyses of Complex Microbial Communities

spectrometer directly results in a rapid separation with no additional offline preparation.

Although a variety of mass spectrometers exist, several are aptly suited for the characterization of complex samples. Depending the on the physical hardware of the mass spectrometer the performance metrics will vary (mass range, throughput, and mass accuracy and resolution). The performance metrics of the measurement are often in concert with each other, such that sacrifices in throughput are necessary for gains in mass resolution and accuracy or vice-versa. One primary consideration is the compatibility of the mass spectrometer with the ion source. The commonly used electrospray ionization (ESI) is considered a soft ionization (generally does not induce source fragmentation) technique and typically results in multiply charged gas-phase ions.[35] Electrospray is commonly the ionization technique of choice for proteomic analyses due to the ability to interface directly with online chromatography and the compatibility with organic molecules, such as proteins and peptides. The ability to generate multiply charged ions provides an increase in measureable mass range, as the mass spectrometer measures mass to charge (m/z), and an ideal ion for fragmentation. Two mass spectrometers which are preferred for ESI compatibility and useful operating figures of merit for protein and peptide mass measurement are the Fourier Transform Ion Cyclotron Mass Spectrometer (FTICR) and the Linear Trapping Quadrupole (LTQ).[36, 37] The FTICR and the LTQ differ in the performance metrics, with the FTICR providing extremely high mass accuracy and resolution, at the cost of throughput, and the LTQ providing tremendous gains in throughput, but sacrificing mass accuracy and resolution. For this reason, and others described below, the FTICR is the preferred instrument for intact protein analysis (TD) and the LTQ is well suited for peptide analysis (BU).

The final step in the proteomic methodology is the informatics. The purpose of this is to facilitate, depending on the sample, the assignment of MS spectra to either proteins or peptides. Fragmenting an ion in the mass spectrometer is a valuable technique for determining the sequence of a particular ion and most software packages rely on the fragment information for identification. The computational assignment of TD and BU spectra benefit from fragmentation, but the fragmentation itself is significantly

more amenable to BU analyses. The primary means of assignment for BU analyses relies on software that generates predicted fragmentation spectra, from the protein sequence database. The predicted spectra are then compared to experimental fragment spectra in order to identify matches and determine the probable peptide sequence. TD relies on the high mass accuracy and, in some cases, fragmentation of the intact protein. The use of specific software and particular parameters is intended to minimize the presence of false assignment, or the false discovery rate. Details of the informatics assignment are highlighted below.

2.2: Reagents / Solvents

Chemical reagents (i.e., guanidine HCl, acetic acid, dithiothreitol (DTT)) were acquired from Sigma Chemical Co. (St. Louis, MO) and were used as supplied without further purification. Modified sequencing grade trypsin (Promega, Madison, WI) was used for all protein digestions. TFE was purchased from Fluka (Buchs, Switzerland, Catalog No. 96924). HPLC-grade water and acetonitrile were obtained from Burdick & Jackson (Muskegon, MI), and 99% formic acid was purchased from EM Science (Darmstadt, Germany).

2.3: AMD Sample Collection and Extracellular Preparation

Biofilm samples were collected by our collaborators (Jill Banfield group, University of California at Berkeley) from various locations of the mine in Redding, CA.[20] Designations of the collection site are represented by the streams from which they originate (**Figure 2.2**). The identification of signal peptide cleaved proteins utilized five different samples: AB-End, AB-Front, AB-Muck (Friable), AB-Muck (DSII) and UBA. The metal enrichment and characterization of cytochrome 579 utilized the AB-Muck sample. Each of the samples represented a different location or biofilm growth state and each contained approximately $1x10^{10}$ cells. AB-End was an earlier growth state biofilm than AB-Front and AB-Muck, which were designated as Developmental Stage II (DSII). AB-Muck (Friable) exhibited a unique shift in the dominant microbial species,

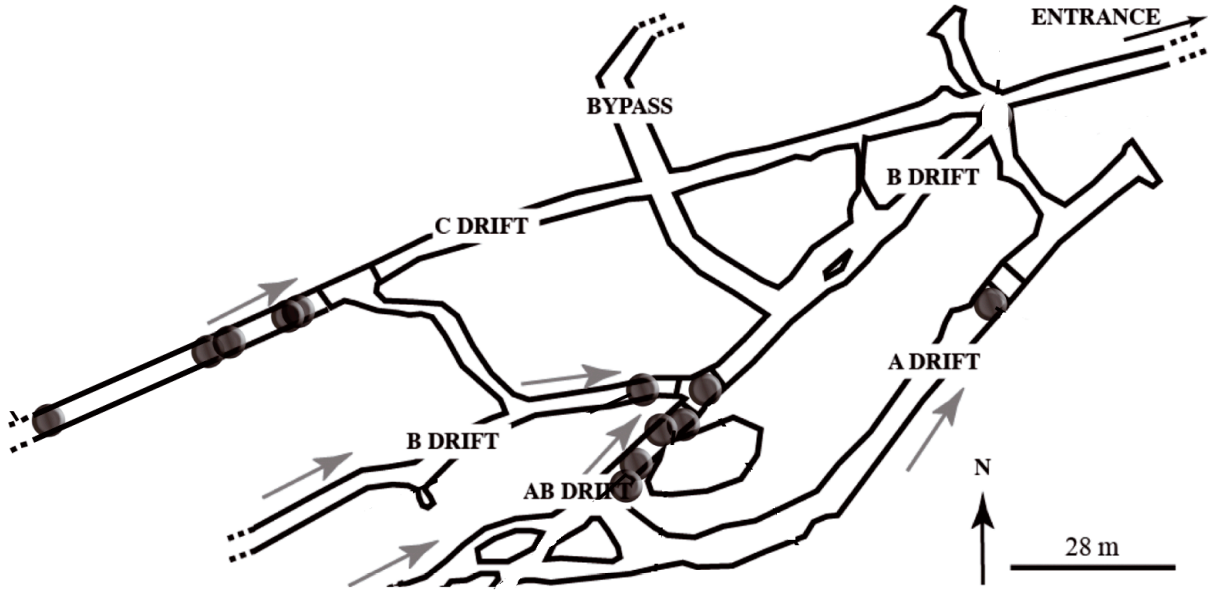

Figure 2.2: Illustration of the Mine Tunnel System in Redding, CA

The various tunnel streams are named (A drift, B drift, etc.). Dark circles represent areas of sampling for proteomic analyses.
Figure taken from Denef *et al.*[38]

while UBA exists only in the A-drift region of the mine. Each of the samples were immediately frozen in dry ice, and stored in a -80°C freezer until further processing. Collected samples were processed to produce an extracellular fraction while limiting inadvertent cellular lysis. The frozen samples were thawed and processed, as follows, at 4 ºC. Cells were suspended in 3 volumes of H_2SO_4 (pH 1.1), washed by rotation for 30 min, and recovered by centrifugation at 12,000g 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. Since the extracellular fraction was collected after treatment of the biofilm by cold osmotic shock, it is likely enriched in both periplasmic and secreted proteins. Proteins within the extracellular fraction were precipitated with ice-cold 10% trichloroacetic acid, and the pellet was rinsed with cold methanol and air-dried.

2.4: Protein Preparation for MS Measurement

For TD MS analysis, proteins were enriched and purified before direct measurement. BU MS experiments required additional preparation in order to enzymatically generate peptides for MS analysis.

For low complexity intact protein analysis (such as cytochrome 579 characterization), proteins were prepared for direct infusion into the mass spectrometer. Enriched protein samples were desalted with Zip-tip (C4, Millipore, Billerica, MA, USA) pipette tips and eluted with 100% acetonitrile (0.1% acetic acid, v/v).[39] The Zip-tip C4 pipette tip provides a reliable method for concentrating and desalting proteins prior to mass measurement. Following elution, the purified proteins can be diluted to a compatible concentration (500 nM – 100 uM) and measured by directly infusing the protein sample at a flow rate of 2.5 uL/min. For complex protein samples intended for intact mass measurement, the samples are loaded directly onto a chromatography column for desalting and on-line separation which is described below.

For BU measurement, the proteins are prepared for enzymatic digestion into peptides. The generation of peptides relies on the use of commercially available proteases, the most common being trypsin. Trypsin enzymatically cleaves the peptide

backbone c-terminal to lysine (K) and arginine (R). In order to prevent self-cleavage, the trypsin (Promega, Madison, WI) has been chemically modified by the manufacturer.[40] Proteins are denatured and reduced in order to eliminate any potential secondary or tertiary structure which may inhibit trypsin activity. This is accomplished by suspending the sample in 2 mL of 6 M guanidine-HCl, 10 mM dithiothreitol (DTT), at 60ºC for 1 hour prior to the introduction of trypsin. After denaturation and reduction of disulfide bonds, the sample is diluted 6-fold in 50 mM Tris-HCl/10 mM CaCl2 (pH 7.8) providing a suitable solution for tryptic activity. Sequencing-grade trypsin was added at ~1:100 (w/w), and digestions were performed with gentle rocking at 37ºC for 18 hours. This was followed by a second addition of trypsin at 1:100 and an additional 5-hour incubation. The samples were then treated with 20 mM DTT for 1 hour at 37ºC as a final reduction step, and immediately de-salted with Sep-Pak Plus C18 (Waters, Milford, MA).[41-43] All samples were concentrated and solvent exchanged into 0.1% formic acid in water by centrifugal evaporation to ~1 mg/mL starting material, filtered, aliquoted, and frozen at –80ºC until LC-MS/MS analysis

2.5: Liquid Chromatography

To efficiently measure the thousands of proteins or peptides present in a sample, the use of an on-line liquid chromatographic (LC) separation is employed.[44, 45] Depending on the complexity of the sample, this may include a single or double, ideally orthogonal, phase(s) of separation. The most common form of stationary phase for general separation of proteins and peptides is reverse-phase (RP). The reverse phase consists of silica bonded to variable length alkyl chains. The common versions of RP are C4, C8, C18 and their application is dependent on the sample set. For both intact protein and peptide mass measurement, the application of liquid chromatography results in significant gains in the ability to measure less abundant ions. The gains in dynamic range enable deep proteomic coverage among the thousands of expressed proteins. The LC occurs online, directly coupled with the mass spectrometer. This provides performance gains and reduces the potential for sample loss by eliminating additional offline sample handling. The composition of the solvent is adjusted over time

such that the concentration of the mobile phase is increased (gradient elution). Ideally, the gradient elution should provide increased resolution over an isocratic separation.

2.5.1: Intact Protein Separation

Intact protein chromatography typically utilizes shorter alkyl chain stationary phases (C4, C8), due to the increased size of intact proteins and correspondingly the greater available surface area, resulting in increased hydrophobic affinities for the stationary phase. On the other hand, BU MS analysis, where peptides are required to be separated, the increased hydrophobic interactions achieved with C18 provide more efficient separation. The non-polar surface of the stationary phase results in elution of polar molecules before non-polar molecules, thereby supporting the term reverse-phase. The elution buffer must contain a suitable organic solvent, such as methanol or acetonitrile (ACN). ACN is often utilized due to its volatile nature and subsequent compatibility with MS. The elution phase is often run in a flow, whereby the concentration of the organic (ACN) is increased over a period of time. The determination of the gradient is a function of the sample, instrument capabilities, and desired peak resolution. The high sensitivity of current MS instrumentation allows for small concentrations of proteins or peptides to be pre-loaded onto the chromatography column. Variations in ionization efficiency and the complexity of the sample will impact the loading concentration, but generally ~200 µg of sample is loaded. The stationary phase is loaded via a high pressure cell into a fused silica capillary connected to union containing a filter (0.5 µM, Upchurch Scientific, WA) acting as a frit. After the stationary phase is loaded, the desired sample can then be deposited onto the stationary phase. Utilizing PEEK ferrules and unions (Upchurch Scientific, WA), the fused silica is then coupled to a fused silica nanospray emitter (New Objective, MA).

2.5.2: Peptide Separation

For samples containing thousands of peptides, a two-dimensional separation is employed. Two parameters must be considered when utilizing a 2D separation: 1.) the sample throughput and 2.) the proteomic coverage. In practice, these metrics vary

inversely with each other and a balance must be sought that adequately assesses both metrics. A widely accepted platform for 2D separation that balances both the throughput and coverage is termed: Multidimensional Protein Identification Tool (MudPIT).[17] This methodology utilizes both RP and strong-cation exchange (SCX) to more completely separate complex peptide samples. For MudPIT separation, peptides are initially loaded and bound to the strong cation exchange material. During the on-line separation, increasingly concentrated steps of ammonium acetate (a volatile, MS compatible salt) are introduced across the SCX, resulting in the step elution of a subset of loaded peptides onto the next phase of separation, the RP. The peptides are then gradually eluted by the ACN gradient directly into the MS prior to the next salt pulse. For the BU analysis of the extracellular fractions, the mass spectrometer was coupled on-line with an Ultimate HPLC (LC Packings, a division of Dionex, San Francisco, CA). The system utilized a 2D nano-LC tandem mass spectrometry (2D-LC-MS/MS) setup. The flow rate from the pump was maintained at ~100 µL/min, which was then split pre-column to provide an approximate flow of ~200–300 nL/min at the nanospray tip. The split-phase columns were prepared in-house and consisted of SCX material (Luna SCX 5µ 100Å Phenomenex, Torrance, CA) and C18 RP material (Aqua C18 5µ 125Å Phenomenex). For all samples, ~200–500 µg of protein material was loaded off-line onto the back of the multi-phase column. The loaded RP-SCX column was then positioned on the instrument behind a ~15 cm C18 RP column (Aqua C18 5µ 125Å Phenomenex) also packed via a pressure cell into a Pico Frit tip (100 µm with 15 µm tip New Objective, Woburn, MA). All samples were analyzed via a 24-hour 12-step 2D analysis.

2.6: Mass Spectrometric Measurement

The selection of the appropriate MS instrument is dependent on the sample and desired measurement. Two platforms that are amenable to on-line liquid chromatographic separation are the Fourier transform ion cyclotron resonance (FTICR) and the linear trapping quadrupole (LTQ ion trap) mass spectrometers.[36, 37] Both instruments provide specific advantages and disadvantages for the measurement of

proteins and peptides respectively. In general, the primary differences between ion trap and FTICR mass spectrometers are the duty cycle (the time when the instrument is usefully operated), mass resolution, and mass accuracy. The ion trap mass spectrometers are ideally suited for complex peptides samples, due to their extremely fast scan time (~700 ms), moderate resolution (~3,000) and accuracy (unit mass) and ability to utilize collision-induced dissociation (CID) for ion fragmentation. FTICR-MS, on the other hand, requires more time to generate a spectrum (~1.5 secs) but is capable of significantly higher mass resolution (100,000) and mass accuracy (1×10^{-3} Da.).

2.6.1: Intact Protein MS Measurement

FTICR-MS measurement has more recently been applied to complex proteomic measurement of intact proteins.[46-49] FTICR provides unrivaled mass resolution and mass accuracy, both of which are critical for precise intact protein mass determination. The resolution and mass accuracy are a function of the high magnetic field (9.4T) and high vacuum ($1\text{-}3 \times 10^{-10}$ Torr) present within the instrument **(Figure 2.3)**. The high resolution of the instrument provides the ability to discern the complicated isotopic distribution of intact proteins, providing several benefits. Specific identification of individual peaks within the isotopic packet allows for the unambiguous determination of the charge state. This then results in the ability to correctly calculate the neutral mass. Additionally, intact protein mass spectra are generally highly complicated due to the high charge states of the ions. This results in mass spectra containing numerous charge states across the mass range, with numerous isotopic peaks for each charge state. The high resolution of the FTICR allows for the visualization of individual peaks as opposed to large waves of peaks. The high mass accuracy of the FTICR mass spectrometer is useful for identifying the chemical makeup of ions or assigning protein identifications based on mass alone. Utilizing the high mass accuracy provides yet another layer for confident protein identification.

Samples prepared for FTICR-MS analysis were diluted into 50/50/0.1 (V/V/V) H_2O/ACN/Acetic Acid and infused into the Micromass Z-Spray source attached to a Varian (Lake Forest, CA) 9.4-Tesla (Cryomagnetics Inc., Oak Ridge, TN) HiRes

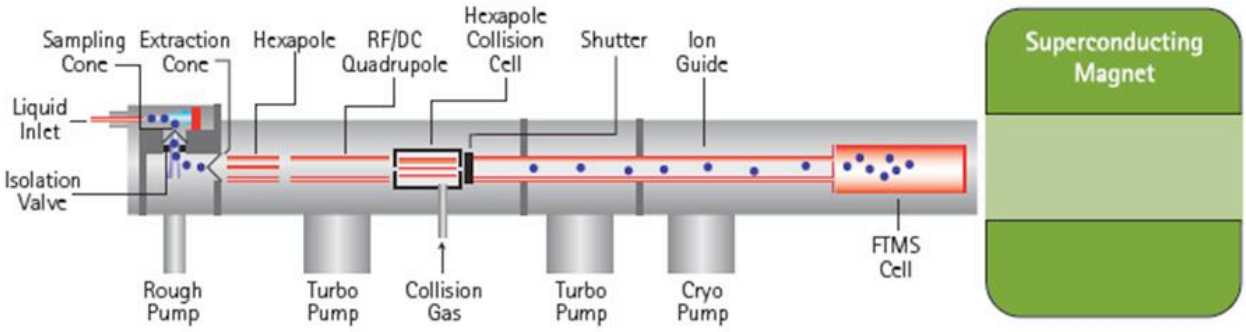

Figure 2.3: Illustration of a FTICR Mass Spectrometer

Following ionization and injection, the ions are transmitted through a series of ion guides and are eventually measured within the FTMS cell, which is located within the superconducting magnet.

Image: Varian Inc.

electrospray FTICR mass spectrometer. MS fragmentation was achieved through infrared multiphoton dissociation (IRMPD) with the 10 micron wavelength of a carbon dioxide laser. Parent charge states of protein were manually selected and isolated in the selection quadrapole prior to mass analysis in the FTMS analyzer cell. For internal mass calibration (better than 3 ppm achievable), a 350 nM spike of ubiquitin was introduced into the appropriate proteome sample. M/z values were manually extracted from spectra, deconvoluted, and plotted with Origin 8.

2.6.2: Peptide MS Measurement

Ion trapping mass spectrometers are aptly suited for the high throughput analysis of complex proteomic samples. The millisecond scan speed enables these types of mass spectrometers to rapidly select, isolate, and fragment ions, resulting in the sequence specific information necessary for a BU MS methodology. Their compatibility with CID based fragmentation results in predictable and consistent fragmentation of peptides.

The mass spectrometer itself is composed of three main components: an ion source, a mass analyzer and detector.[34] In general, the MS experiment proceeds as follows: a gas phase ion is introduced into the mass spectrometer which then passes through a series of ion guides to a mass analyzer for mass selection and then to a detector for ion detection and signal amplification. In general the ion trapping mass spectrometers are ideal for BU proteomic analyses. They provide moderate resolution, high sensitivity, rapid scan speed, the capability to perform multiple, consecutive fragmentations (MS^N), and compatibility with various fragmentation techniques (CID, ETD). The principle behind trapping an ion lies in producing a region within the ion trap where radial and axial stability is achieved. This occurs through the oscillation of the RF voltage. The ions can be trapped axially by application of a constant voltage to the end electrodes. Target ions can be isolated by sequentially scanning out (adjusting the RF) the ions that fall outside the desired m/z window. The ions remaining within the trap are ideally the target ions for fragmentation. Application of a excitation current

across the end caps excites the remaining ions thereby inducing thousands of physical collisions with the helium present in the trap.[50]

During the entire chromatographic process, the LTQ (Thermo Fisher Scientific, San Jose, CA) mass spectrometer was operated in a data-dependent MS/MS mode, as detailed below (**Figure 2.4**). The use of a data-dependent MS analysis provides real-time parent ion selection for fragmentation. In this mode, the software controlling the instrument fully automates the selection, isolation, and fragmentation of parent ions found in the MS1 spectra.[50, 51] Data-dependent parent ion selection is intensity based; therefore the peptides that are most abundant are typically selected. This allows for a sufficient ion population to be measured following fragmentation. 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 Fisher Scientific). The LTQ mass spectrometer was operated with the following parameters: nanospray voltage (2.4 kV), heated capillary temp 200ºC, full scan m/z range (400–1700). 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 both 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 3 minutes on the LTQ. The parameters for dynamic exclusion vary depending on the instrument platform and are adjusted based on empirical analysis of the parent ion selection. The use of dynamic exclusion prevents the repeated targeting, isolation and fragmentation of high abundant ions. Due to the high complexity of community microbial samples, it is likely that less abundant ions will never be targeted for fragmentation within the specified parameters (1 MS1 -> 5 MS2). By utilizing dynamic exclusion, abundant ions are excluded from additional fragmentation (after a repeat of one fragmentation), allowing the mass spectrometer to target lower abundance ions thereby increasing the depth of the proteomic identification.[52, 53]

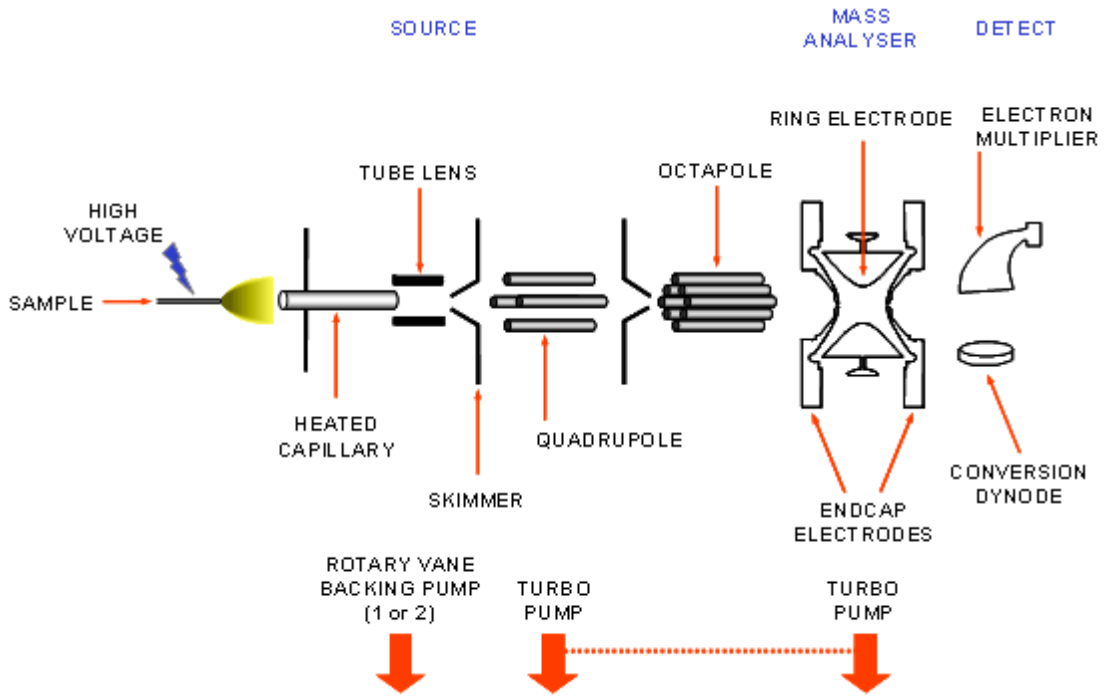

Figure 2.4: Illustration of an Ion Trapping Mass Spectrometer

Following ionization and injection, the ions are transmitted through a series of ion guides and are eventually contained within the ion trap. Within the ion trap, selection, isolation and fragmentation occur prior to ejection and detection by the electron multipliers.

Image: Thermo Scientific

2.7: Proteome Informatics

The assignment of experimentally collected spectra to either proteins or peptides is largely automated. This is necessary, as current methodologies are capable of collecting greater than 100,000 spectra in a single experiment. The method for identifying spectra varies depending if the sample consists of proteins or peptides. The assignment of spectra generated from the fragmentation of peptides relies on the use of protein sequence databases. In this manner, the protein sequence database serves as a template whereby the predicted sequences are used to simulate the experimental fragmentation. The predicted fragmentation is then matched to the experimentally collected fragmentation and scored, as detailed below. The assignment of spectra representing intact proteins is based on the high mass accuracy. Frequently this is not sufficient, as the proteins may be moderately or highly modified, thereby resulting in high mass error matches. The combination of a BU based protein identification and intact protein mass measurement provides sufficient confidence for a specific identification.

2.7.1 Database Searching

In order to assign spectra to peptides, the raw data must be extracted from each of the thousands of spectra. During the MS analysis, MS1 (survey) and MS2 (fragmentation) spectra are collected. The MS2 scans contain the results of fragmenting a parent ion with the inert gas, which is helium for the LTQ instruments. The resulting fragmentation pattern is the basis for the peptide identification. Extraction of the m/z values is followed by conversion to a deconvoluted neutral mass. The moderately low resolution by the ion trapping MS instruments requires that in many cases, +2 and +3 charge states must be considered when calculating the neutral mass of the fragments. Following charge state determination and deconvolution, each individual spectrum is represented by a peak list providing the neutral mass and intensity. Database matching software, such as SEQUEST, utilizes the supplied sequence database to perform an *in silico* digestion of the protein sequences. Due to the predictable nature of the enzymatic digestion and the CID based fragmentation, it is

possible to accurately predict the experimental fragmentation. A preliminary selection of possible candidate peptides is obtained by selecting peptides within +/- 3 Da. of the parent ion. This subset of peptides is then further culled by a preliminary score which compares the predicted versus experimental fragments. Two significant metrics of the computational and experimental matching are the XCorr value and DeltCN. The XCorr represents the final cross-correlation value of the spectra with the best match having the highest value. This represents how well the predicted spectrum matches the experimental spectrum. The DeltCN dictates how different the best hit is to the next. Larger values will, in general, provide greater confidence in the peptide assignment. Specific parameters for SEQUEST searches are described within the respective chapters.

Assignment of spectra representing intact proteins relies heavily on the accuracy of the mass measurement. Due to the inconsistent fragmentation observed with intact proteins, identification through fragmentation is often not possible. Thus, the integration of a BU identification with a high mass accuracy measurements is often required. A software platform developed for this approach is discussed in detail in chapter 6.

The methodology described above provides a robust and rapid platform for the complete proteomic characterization of complex microbial communities. Each stage of the sample preparation is intended to provide an optimal solution for MS measurement. This results in high quality spectra that are amenable to the computational algorithms highlighted above. The described platform has provided a solid experimental foundation for the targeted analyses highlighted below.

Chapter 3

Computational Prediction and Experimental Validation of Signal Peptide Cleavages in the Extracellular Proteome of a Natural Microbial Community

Portions of included text are adapted from:

Brian K. Erickson, Ryan S. Mueller, Nathan C. VerBerkmoes, Manesh Shah, Steven W. Singer, Michael P. Thelen, Jillian F. Banfield, Robert L. Hettich, "Computational Prediction and Experimental Validation of Signal Peptide Cleavages in the Extracellular Proteome of a Natural Microbial Community", Journal of Proteome Research, 2010, 9 (5), 2148-2159

Brian K. Erickson's contributions include computational prediction of datasets, experimental preparation of samples, experimental LC-MS/MS analysis, data parsing, and primary authorship.

3.1: Introduction

Determination of the cellular location(s) of proteins provides important contextual information, which can support proposed functions. Of particular interest are proteins that mediate interactions between a microorganism and its environment and operate under external conditions that may differ substantially from conditions in the cytoplasm. These secreted proteins are critical for nutrient transport, as well as organismal communication and survival (i.e., defense).

A primary, but not exclusive, method of protein transport to the extracellular region, periplasmic space, or outer membrane of gram-negative bacteria is through signal peptide mediated transport.[54] In this highly conserved process, trafficking of the protein is dependent on the presence of a specific sequence of amino acids, typically located within the first 50 amino acids of the N-terminus.[55, 56] Targeting generally

occurs through two pathways, one involves the signal recognition particle (SRP) and occurs co-translationally, and the other involves SecB, and occurs post-translationally. Following targeting, protein transport to the cytosolic membrane occurs through a complex of proteins known as the translocase, which includes membrane proteins as well as an ATPase.[57] The result of these activities is the directed transport of a protein and cleavage of the signal peptide. Additional models of protein secretion are utilized within gram-negative bacteria, but were not specifically probed within this study.

The ability to computationally predict signal peptides has advanced significantly in recent years.[58, 59] Current prediction algorithms utilize machine learning techniques, such as neural networks and hidden Markov models (HMM) to increase accuracy and precision.[60-62] These computational techniques identify patterns of amino acid composition in the N-terminal region of a protein in order to ascertain if a signal peptide is present. The pattern recognition has been optimized through the use of training sets and is specific for eukaryotes, gram-negative, and gram-positive bacteria. The training sets are composed of hundreds to thousands of experimentally verified signal peptide sequences. SignalP-3.0 is one widely used and accepted algorithm that effectively identifies the presence of a signal peptide along with the probable cleavage site. The current version of SignalP (3.0) has been improved over previous iterations by utilizing expanded training sets containing additional experimentally verified signal peptides, as well as including HMM resulting in increased prediction accuracy. For gram-negative bacteria, prediction of a signal peptide as well as identification of the cleavage site is proposed to be > 90% accurate.[63]

Experimental data provides important confirmations of signal peptide predictions. N-terminal sequencing techniques, such as Edman sequencing, can be used to verify the sequences of mature protein forms, but this is not an effective method for profiling the thousands of proteins present in microbial community proteomes. An alternative approach is to verify signal-cleaved peptides using shotgun multidimensional liquid chromatography tandem mass spectrometry (2D-LC-MS/MS). Since peptide assignments using shotgun proteomics depend on the presence of the exact predicted

peptide sequences in databases, signal-cleaved peptides and non-cleaved peptides can be readily distinguished (**Figure 3.1**)[64, 65]. Mass spectrometry is appropriate for signal peptide analysis in microbial community samples due to its unrivaled throughput, as well as its high dynamic range and mass accuracy.[2] Mass spectrometry also provides relative quantification of peptides and proteins, allowing for detection of trends in abundance patterns of exported proteins across samples.[66]

In this study, we evaluated the approach of integrating computational prediction of signal peptide-containing proteins with high-throughput mass spectrometry to validate signal peptide predictions for a diverse mixture of proteins from a natural microbial community. We focused on microbial biofilms with limited species richness from an acid mine drainage (AMD) system.[5, 6, 67] The biofilms grow in hot (40 ºC), pH ~ 1.0 solutions that contain near molar concentrations of metals (in particular Fe). Proteogenomic analyses, which combine proteomic measurements and metagenomic data, have been previously applied to these biofilms to catalogue and evaluate abundance patterns for thousands of proteins from the most abundant bacterial and archaeal populations.[19, 20, 38, 68, 69] As of yet, a specific identification and characterization of the secreted proteins present in the extremophilic AMD system has not been completed. The analysis presented in this study has broad implications for characterizing extremophilic microbial communities. High confidence identification of the secretomes will provide vital clues into microbial community interaction, function, and survival at the environmental and cellular interface. The combination of protein enrichment in the secretome and the presence of signal-cleaved peptides provide strong evidence for protein localization and clues to protein function. Characterization of the changes in the abundances of signal-cleaved proteins across microbial communities from biofilms growing in different geochemical environments and of different growth states permits a greater understanding of the roles of these proteins in situ. A subset of these secreted proteins may be critical for organismal survival in the highly acidic environment, and should provide unprecedented insight into the global acid mine drainage phenomenon. Finally, the methodologies presented within can be readily applied to a variety of microbial systems for specific prediction/characterization of their secreted proteins.

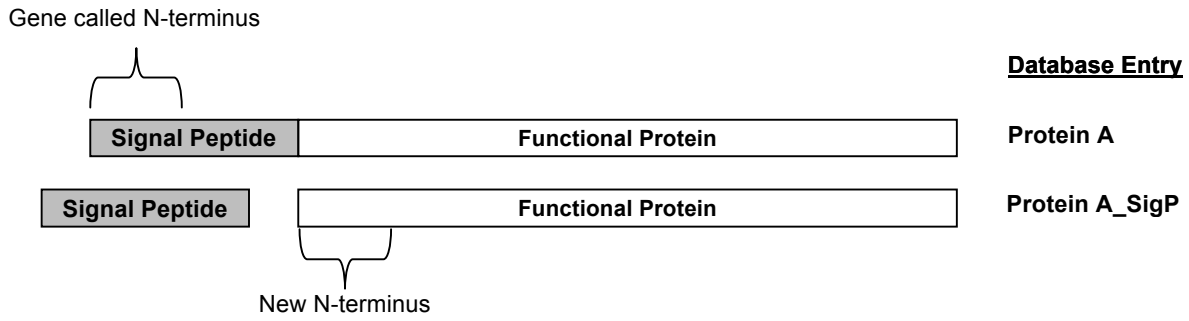

Figure 3.1: Representation of Signal Peptide and Cleavage
The original peptide of a protein was termed the "gene called N-terminus." Following removal of a signal peptide, a new N-terminus will be present on a protein. This is termed the "new N-terminus." The protein database contains both the pre-processed form of the protein, "Protein A" and the signal peptide cleaved sequence, "Protein A_SigP."

3.2: Materials and Methods

3.2.1: Signal Peptide Prediction

The experimental approach for this study consisted of parallel computational prediction and mass spectrometric identification of signal peptide cleaved proteins (**Figure 3.2**). SignalP-3.0[70] was downloaded from (http://www.cbs.dtu.dk/cgi-bin/nph-sw_request?signalp) and locally installed. Each of the 16,171 proteins present in the AMD database (Biofilm_5wayCG_UBA_08052007.fasta) was individually submitted to SignalP-3.0 for analysis using a Perl script which iteratively selects and copies each FASTA formatted protein sequence from the sequence database, along with accompanying header information, to a separate temporary text file. This text file was submitted to SignalP-3.0, which was executed with the following parameters: organism set to gram-negative bacteria, output short format, quiet analysis and protein sequence truncation after the first 50 amino acids. Results of SignalP-3.0 were exported to a temporary file, and identification of signal peptides was accomplished by parsing the results of the hidden Markov model analysis conducted by SignalP-3.0. Following the prediction of a signal peptide, the position of the cleavage site was noted in order to generate the processed protein sequences. Sequences of proteins predicted to have a signal peptide were truncated at the predicted cleavage site and their protein names were appended with “SigP” in the new protein sequence database. This database was labeled “Biofilm_5wayCG_UBA_08052007_SigP_Removed.fasta” and contains both the original gene-called protein sequence and if predicted to be present, a signal peptide cleaved protein sequence. The complete, SignalP-3.0 derived database can be found at: “http://compbio.ornl.gov/biofilm_amd_extracellular_proteome”

The lack of archaeal training data sets limits the effectiveness of SignalP-3.0 in predicting archaeal signal peptides. For this reason, the predictions and identifications of signal peptide cleaved proteins in this study were generally found for the abundant gram-negative bacterial populations in AMD biofilms.

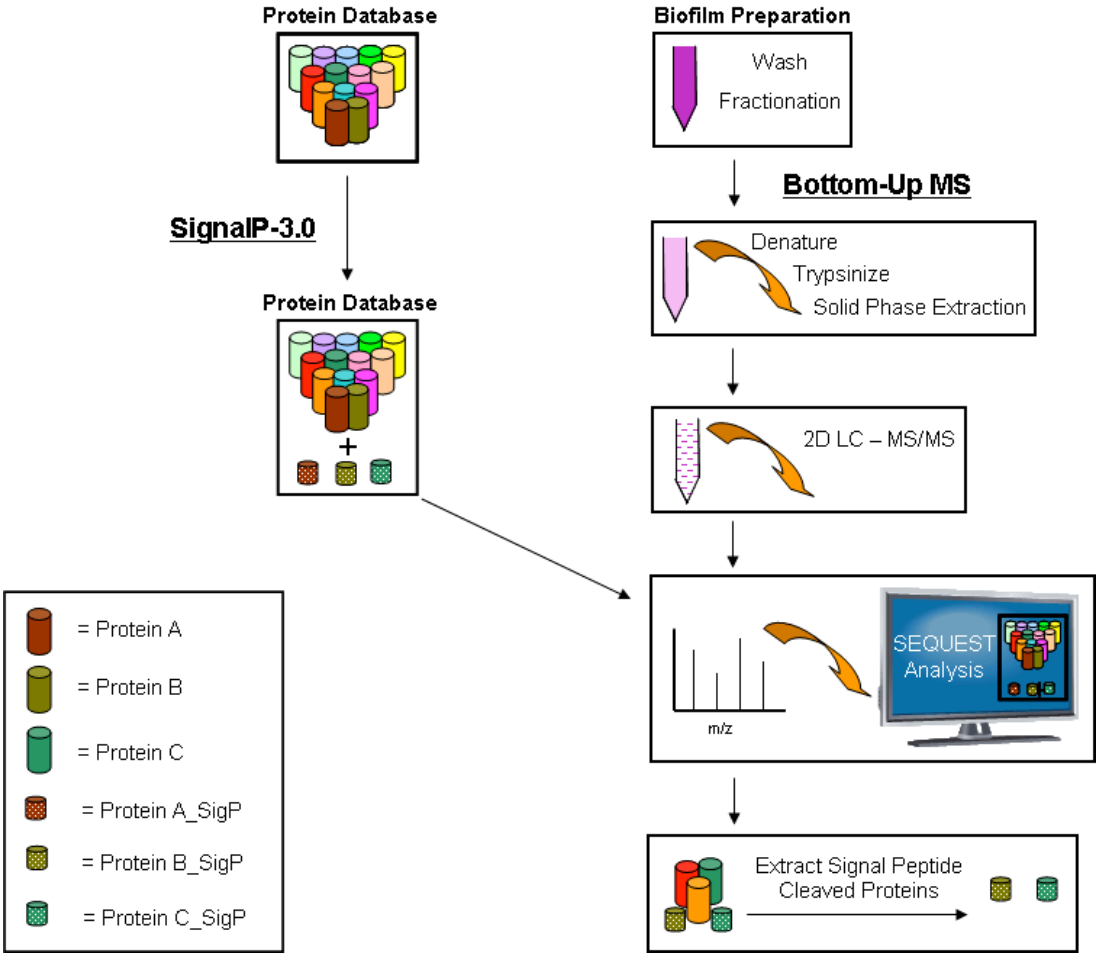

Figure 3.2: Computation and Experimental Determination of Signal Peptide Cleavage
The crude biofilm was fractionated and digested into peptides for LC-MS/MS. The acid mine drainage protein database was subjected to signal peptide cleavage prediction with SignalP-3.0. The proteins with predicted signal peptide cleavage were appended to the database with the signal peptide sequence removed and noted with a "SigP". The pre-processed protein sequence was retained in the database. Following MS/MS analysis the spectra were searched with SEQUEST utilizing the signal peptide appended database and parsed for signal peptide cleavage identifications.

3.2.2: Proteome Bioinformatics

The "Biofilm_5wayCG_UBA_08052007_SigP_Removed" database contains annotated proteins from the abundant microbial members of AMD biofilms.[69] The protein database also includes common contaminants (trypsin, keratin, etc.). Protein assignment of the MS/MS spectra was accomplished with the SEQUEST algorithm[25] and was run using the following parameters: enzyme type, trypsin; Parent Mass

Tolerance, 3.0; Fragment Ion Tolerance, 0.5; up to 4 missed cleavages allowed, and fully tryptic peptides only. Resulting output files were sorted and filtered using DTASelect with the following parameters: tryptic peptides only, deltCN value of at least 0.08, and XCorr values of at least 1.8 (+1), 2.5 (+2), 3.5 (+3) with a two peptide minimum. Cross-comparison among DTASelect outputs was accomplished with Contrast[71] and an in-house script that provides similar functions. Rapid filtering of the signal peptide cleaved proteins identified in the DTASelect output was accomplished using a Perl script which extracted protein identifications containing the "SigP" designation to an additional table. Accompanying information regarding protein sequence coverage, number of peptide identifications, and spectral counts were also recorded. In order to support the identification of a signal peptide cleavage, the DTASelect output was parsed for the predicted, pre-processed signal peptide. Identifications of a pre-processed signal peptide were noted along with accompanying spectral counts. A false positive rate of <1% was calculated based on forward-reverse database searching according to Elias *et al*.[72] All databases, peptide and protein results, MS/MS spectra and supplementary tables for all database searches are archived and made available as open access via the following link: "http://compbio.ornl.gov/biofilm_amd_extracellular_proteome". All MS ".raw" files or other extracted formats are available upon request.

Highly expressed signal-cleaved proteins with confirming new N-terminus spectra were submitted in batch form to Pfam for protein family and domain analysis.[73, 74] The parameters for the search included a merged global and local strategy and an E-value cutoff of 1.0 The resulting Pfam hits were further filtered with an E-value cutoff < 1×10^{-3}, exceeding the stringency outlined in Altschul *et al*. [75]

3.2.3: N-terminal Sequencing

Complementary experimental verification of signal peptide cleavage was accomplished on selected secretome proteins by Edman degradation. The extracellular fraction extracted from 50 mL of biofilm from the C-drift location collected in November, 2005 was fractionated by column chromatography on a SP-Sepharose FF column, as previously described.[76] After elution of Cyt_{579}, a 0-2 M NaCl gradient was applied at pH 5.0, (30 mL, 3 mL fractions). Greater than 95% of the proteins recovered from the NaCl gradient were present in the 1.4 M - 2.0 M NaCl fractions. These fractions were precipitated with 10% tricholoroacetic acid in an ice bath and redissolved in 10 µl of SDS-PAGE sample buffer. The final protein weight dissolved in the sample buffer was 20-40 µg. The samples were visualized by SDS-PAGE (15% polyacrylamide pre-cast gel, Bio-Rad), transferred to a polyvinylidene fluoride (PVDF) membrane, and the bands excised for sequencing. N-terminal sequencing of the visualized proteins was performed as previously described.[19]

3.2.4: Hierarchical Cluster Analysis of Signal Peptide Cleavage in 28 AMD Proteomes

The abundance patterns of computationally predicted and experimentally verified signal peptide cleaved proteins were examined across a distinct set of 28 biofilm samples collected over a period of 4 years from 6 different locations within the Richmond Mine, from a different study.[38] The abundances of individual proteins were calculated using normalized spectral abundance factors (NSAF), which are based on the spectral counts of peptides for a given protein.[18, 77] Resulting NSAF values were ASIN-transformed and used to cluster proteins and samples using Cluster version 3.0.[78] Clustering of mean-centered and scaled NSAF values was performed using an uncentered Pearson correlation metric and groups were defined using average linkage clustering. Heat maps were visualized with TreeView.[79] To determine whether correlations exist between protein abundances and developmental state of the biofilm, samples were labeled as either a high- or low-developmental stage based on the numbers of archaea detected within each community, as previously determined (Mueller

et al., submitted). Low developmental stage samples are highlighted in green and high developmental stage biofilms are highlighted in blue for each heatmap presented. Detection of differentially expressed signal peptide-containing proteins between developmental stages was achieved using the significance analysis of microarrays technique.[80] Two class unpaired Wilcoxan tests were performed using 500 permutations. Significant genes were assessed at a <10% false discovery rate (FDR).

3.3 Results

Shotgun proteomics via 2D-LC-MS/MS provides the critical cataloging of proteolytic peptides, thereby enabling the discovery and validation of signal peptide cleavage events. The complete AMD protein database was interrogated for signal peptide prediction along with concurrent LC-MS/MS measurements of community biofilm samples in order to ascertain: 1) whether the protein was expressed and detected, and 2) if so, did it reveal a new N-terminal sequence that would be representative of the processed, mature form of the protein?

3.3.1: SignalP-3.0 Prediction Results

The computational prediction of signal peptides resulted in 1,480 signal-cleaved proteins out of 16,171 proteins (9%) from the AMD database (**Table 3.1**). Approximately 18% of the proteins from gram-negative organisms were predicted to contain a signal peptide and more than half of the signal peptide predictions were from the dominant organisms, two strains of *Leptospirillum* group II (395 from the CG strain and 397 from the UBA strain). Leptospirillum group III contained 304 predicted signal peptide containing proteins (11.1% of its total annotated proteome), representing > 20% of the signal peptide database (**Table 3.2)**.

3.3.2: Experimental MS Results

Extracellular fractions from five distinct biofilms from different locations in the Richmond Mine were analyzed in triplicate by 2D LC-MS/MS. Overall, the MS analysis resulted in the identification of 3,388 total proteins. 531 proteins with predicted signal

Table 3.1: Summary of Computational Prediction and MS Identification of Signal Peptide Cleaved Proteins from 5 Distinct Extracellular AMD Samples

	# of IDs	% of Total DB
SignalP-3.0 Prediction	1,480	9%
Measured Protein IDs (All)	3,388	21%
Measured Protein IDs (SigP)	531	3%

Table 3.2: Distribution of Predicted Signal Peptide Cleaved Proteins for the Dominant Microbes in the AMD Microbial Community

	# of Proteins in SigP Database	% of SigP Database
***Leptospirillum* II**	792	53.5
***Leptospirillum* III**	304	20.5
G-plasma	66	4.5
***Ferroplasma* 1**	105	7.1
***Ferroplasma* 2**	136	9.2
Unassigned	77	5.2
Total	1,480	100.0

peptides were identified in at least one of the five sample sets (**Figure 3.3**). After removal of orthologous proteins, 377 non-redundant signal peptide cleaved proteins were identified. From these results, 115 non-redundant proteins were measured and identified as signal peptide cleaved proteins in *all samples and technical replicates*. 46 of the 531 proteins were determined to have signal peptide cleavages with high confidence on the basis of the *presence of a least one spectra corresponding to the new, processed N-terminus and MS identification in all samples and replicates*. 125 total proteins were identified in at least one sample with spectra matching to the new N-terminus generated by signal peptide cleavage. Although the identification of the new, signal peptide cleaved N-terminus provides strong support for the classification of that protein as signal peptide cleaved, the absence of a new N-terminal peptide identification *does not necessarily* indicate that the protein does not contain a signal peptide. For example, there are numerous proteins predicted to contain signal peptides for which no N-terminal peptides were experimentally identified with the current methods employed. The identification of a new N-terminus is dependent on the predicted signal peptide sequence and resulting MS peptide identification and would not be confused with simple tryptic cleavage to result in a new N-terminal identification. Computationally, the new N-terminus is designated in such a way that it is distinguishable from N-terminal tryptic peptides. Therefore, any new N-terminal identification is the result of a specific signal peptide computational prediction and corresponding experimental verification. Among the secretome results, several proteins not predicted to contain a signal peptide by SignalP-3.0 were still identified in the MS analysis. Examples include highly abundant proteins such as GroEL, numerous ribosomal subunits, and various transcription factors. Cell lysis or incomplete fractionation could account for these abundant proteins, which are frequently identified in proteomic analyses of the AMD microbial community.

Clearly our experimental approach will be most successful for identifying soluble secreted proteins. We recognize that predicted signal-peptide proteins designed for membrane insertion would likely be under-represented in our datasets. We used the transmembrane predictor tool TMHMM[81] to interrogate the entire set of SignalP predicted proteins (1480), and find that about 30% of them contain one or more

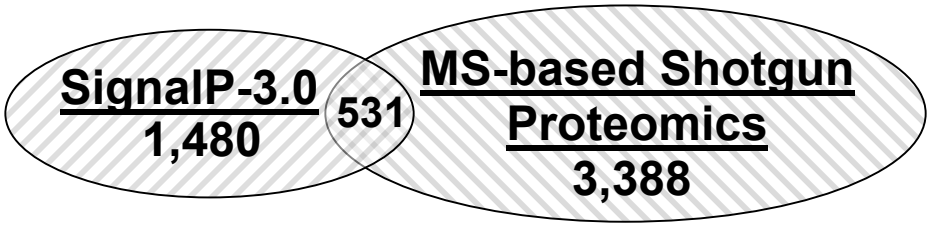

Figure 3.3: Venn Diagram of Predicted and Measured Signal Peptides
Computational analysis of the acid mine drainage protein database with SignalP-3.0 resulted in the prediction of 1,480 proteins with a signal peptide. Following MS/MS analysis a total of 3,382 proteins were confidently identified. Among these 531 proteins were predicted to contain a signal peptide cleavage and were ultimately identified through mass spectrometry.

transmembrane helices. As expected, our identified proteins showed no evidence of transmembrane regions. For the signal-peptide proteins *not identified* in this current study, we propose the following possible scenarios: 1) they were not expressed, 2) they are expressed at levels too low to detect, or 3) they are membrane proteins and thus escape detection by this method. While algorithms such as TMHMM can predict the last category, these cannot definitively define why other proteins went undetected.

3.3.3: Signal Peptide Prediction Disparities

In five cases, peptides predicted by SignalP-3.0 to be cleaved from the mature protein were identified in the uncleaved form by 2D-LC-MS/MS. These five proteins, represented by only 15 spectra are derived from *Leptospirillum* Group II (4) and Group III (1). The four from *Leptospirillum* Group II are conserved proteins of unknown function, whereas the *Leptospirillum* Group III protein has no known function. Of special note, we identified alternative forms of four of the *Leptospirillum* Group II proteins that had the predicted cleaved N-terminus. This could indicate that proteins are incompletely processed, that there were lysed cells with unprocessed protein in the extracellular fraction, or the identifications could be wrong (due to false positive spectral assignments). However, it is important to note that these five cases represent a minority of the signal cleaved proteins detected and verified.

3.3.4: Validation of N-terminal Protein Sequences

Edman degradation sequencing was used to confirm some of the N-termini of proteins in the extracellular fraction predicted by SignalP and identified by MS. The N-termini of seven *Leptospirillum* group II gene products were determined, and all of these correlated to the predicted N-termini determined through this study, except for the protein encoded by *Leptospirillum* group II UBA scaffold 8692 gene 12. This protein has an amino acid variation (measured-AG*T*PSEKLIQ, predicted-AG*D*PSEKLIQ), accounting for the discrepancy (**Table 3.3**). Two of the most abundant secreted proteins are encoded by *Leptospirillum* group II UBA scaffold 8524 gene 128 and *Leptospirillum* group II UBA scaffold 8524 gene 180. The predicted N-terminus was

Table 3.3: Results of Edman N-terminal Sequencing of Select Secretome Proteins From the AMD Microbial Community

Band Number	Gene Product	Observed N-terminal Sequence
1	UBA_LeptoII_8241_114	ASTTKGWVFR*
2	UBA_LeptoII_8524_128	SDVVGVVDVL*
3	UBA_LeptoII_8692_12	AGTPSEKLIQ[1]
4	UBA_LeptoII_8241_693	ASNITI*
5	UBA_LeptoII_8049_366	(A)DAYKTGH*
6	UBA_LeptoII_8524_180	DQAAPAAPA*
7	UBA_LeptoII_8524_180	AAKKKPAKKA
8	UBA_LeptoII_8524_180	GKAKPSMFV MGKAKKPSMF
9	UBA_LeptoII_8524_180	KKAAKKPMKK
10	UBA_LeptoII_8241_349	(EA)HMDHHRMMMR*

Highlighted sequences correspond with SignalP-3.0 prediction.
The "*" indicates mass spectrometric confirmation.
[1] UBA_8692_12 and its CG homolog have predicted sequence AGDPSEKLIQ.

observed for *Leptospirillum* II UBA scaffold 8524 gene 128, which is annotated to be a putative outer membrane protein (OmpH); however, a second form of the protein with the same N-terminus was present at higher molecular weight . This second form may represent a post-translational modification of the protein. For the protein product of *Leptospirillum* II UBA scaffold 8524 gene 180, five additional N-termini were identified in addition to the predicted N-terminus, which suggests that this protein is highly susceptible to protease cleavage.

3.4: DISCUSSION

This integrated computational/experimental study revealed a large complement of proteins that are actively transported beyond the cytosol in the dominant bacterial AMD community members. Given that the periplasms and outer membranes of cells are exposed to the very acidic, metal-rich environment, proteins localized there, including those involved in Fe^{2+} oxidation and electron transport,[82] must be adapted to these environmental challenges.

Figure 3.4 summarizes the functional grouping of signal peptide cleaved proteins. Several proteins identified as transported across the cytoplasmic membrane were annotated as efflux/protein transporters (8%), cytochromes (~6%), dehydrogenases, proteases, and reductases, as described in Goltsman *et al*.[68] This finding correlates well with an experimental investigation of the secretome of *Bacillus subtilis,* a gram-negative bacteria*,* where many proteases, dehydrogenases, and metal binding proteins were also highly abundant.[83] Over 58% of the identified signal peptide cleaved proteins are currently annotated as having an unknown function. Two novel cytochromes, Cytochrome 579 (Cyt_{579}) and Cytochrome 572 (Cyt_{572}) are highly abundant within the AMD biofilms. In particular, Cyt_{579}, thought to function as an electron transfer protein[76], was identified in all 15 MS experiments (270 spectra corresponding to the predicted new N-terminus of Cyt_{579} were identified).

The proteins with the highest confidence signal peptide cleavage are those that contain spectra matching to peptides representing the new N-terminus. **Table 3.4** lists 46 proteins for which signal peptide cleaved peptides were identified in all 15

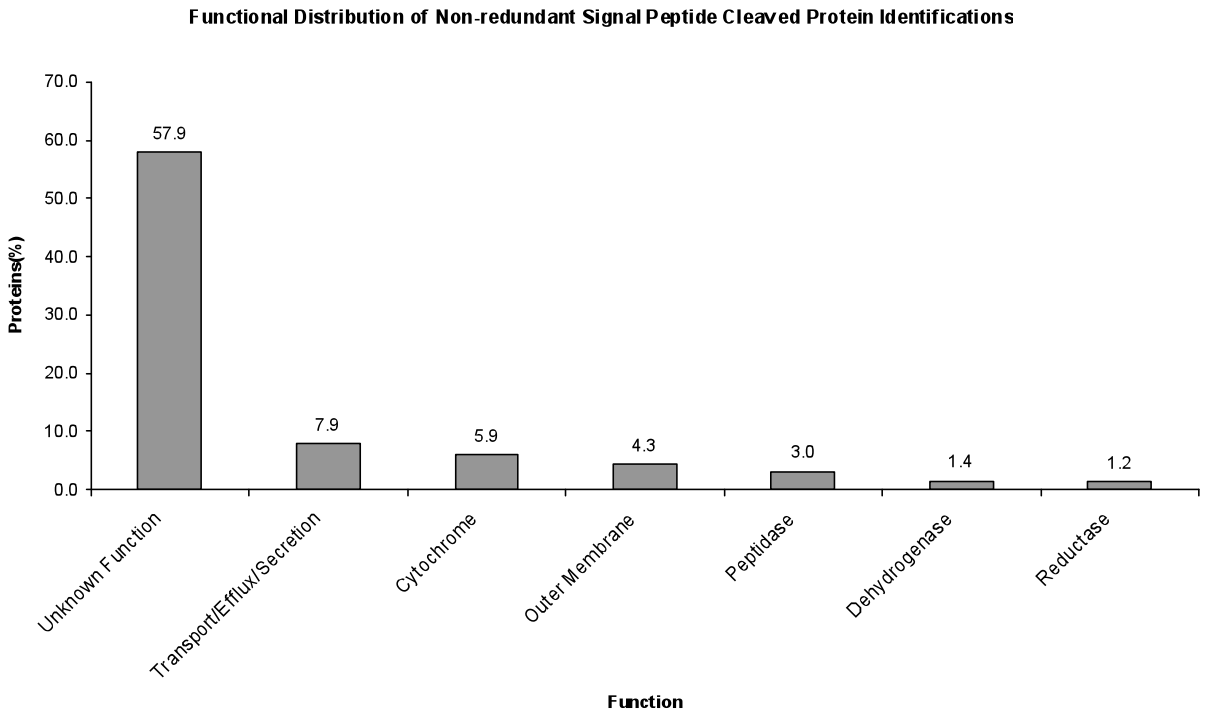

Figure 3.4: Results of Edman N-terminal Sequencing of Select Secretome Proteins From the AMD Microbial Community
Functional analysis of the 377 non-redundant proteins predicted to contain a signal peptide and identified through mass spectrometry. Proteins annotated as hypothetical or with an unknown function are widely present and comprise over 58% of the identified proteins. Expected extracellular proteins such as cytochromes, reductases and peptidases were also identified.

Table 3.4: Highly Conserved and Replicated Signal Peptide Cleaved Proteins with Confirming N-terminus Spectra

Name	N-terminal Spectra	Function
UBA LeptoII Scaffold 8049 GENE 83 SigP	531	Protein of unknown function
5wayCG LeptoII Contig 11390 GENE 17 SigP	531	Protein of unknown function
UBA LeptoII Scaffold 8241 GENE 693 SigP	529	Periplasmic phosphate binding protein
UBA LeptoII Scaffold 8524 GENE 180 SigP	490	Protein of unknown function
UBA LeptoII Scaffold 8062 GENE 372 SigP	290	Cytochrome 579 Variant 1
UBA LeptoII Scaffold 8062 GENE 147 SigP	290	Cytochrome 579 Variant 2
UBA LeptoII Scaffold 8135 GENE 9 SigP	277	Conserved protein of unknown function
5wayCG LeptoII Contig 11233 GENE 46 SigP	277	Conserved protein of unknown function
UBA LeptoII Scaffold 8241 GENE 153 SigP	236	Protein of unknown function
UBA LeptoII Scaffold 8524 GENE 128 SigP	198	Putative outer membrane protein (OmpH)
UBA LeptoII Scaffold 8241 GENE 297 SigP	100	Protein of unknown function
5wayCG LeptoII Contig 11184 GENE 47 SigP	93	Protein of unknown function
UBA LeptoII Scaffold 7931 GENE 111 SigP	65	Putative cytochrome
5wayCG LeptoII Contig 11238 GENE 99 SigP	65	Putative cytochrome
UBA LeptoII Scaffold 7931 GENE 73 SigP	54	Peptidyl-prolyl cis-trans isomerase (EC 5.2.1.8)
5wayCG LeptoII Contig 11238 GENE 58 SigP	54	Peptidyl-prolyl cis-trans isomerase (EC 5.2.1.8)
UBA LeptoII Scaffold 8241 GENE 348 SigP	35	Putative outer membrane protein
UBA LeptoII Scaffold 7931 GENE 365 SigP	35	Protein of unknown function
UBA LeptoII Scaffold 7931 GENE 101 SigP	33	Protein of unknown function
5wayCG LeptoII Contig 11238 GENE 88 SigP	33	Protein of unknown function
UBA LeptoII Scaffold 8062 GENE 53 SigP	32	Protein of unknown function
5wayCG LeptoII Contig 11216 GENE 10 SigP	32	Hypothetical protein
UBA LeptoII Scaffold 8135 GENE 71 SigP	26	Secretion protein HlyD
UBA LeptoII Scaffold 8062 GENE 151 SigP	25	Protein of unknown function
UBA LeptoIII Contig 7952 GENE 72 SigP	22	YceI family protein
UBA LeptoII Scaffold 8049 GENE 48 SigP	21	Protein of unknown function
UBA LeptoII Scaffold 8049 GENE 366 SigP	20	Conserved protein of unknown function
UBA LeptoII Scaffold 8062 GENE 173 SigP	20	Cytochrome C peroxidase (EC 1.11.1.5)
UBA LeptoII Scaffold 8062 GENE 32 SigP	17	Protein of unknown function
UBA LeptoII Scaffold 7931 GENE 352 SigP	12	Protein of unknown function
5wayCG LeptoII Contig 10608 GENE 3 SigP	11	Hypothetical protein
5wayCG LeptoII Contig 10961 GENE 20 SigP	10	Protein of unknown function
UBA LeptoIII Contig 9432 GENE 53 SigP	9	Hypothetical protein
UBA LeptoII Scaffold 8524 GENE 248 SigP	9	Conserved protein of unknown function
5wayCG LeptoII Contig 11111 GENE 93 SigP	5	Protein of unknown function
UBA LeptoII Scaffold 8241 GENE 81 SigP	5	Protein of unknown function
5wayCG LeptoII Contig 11277 GENE 93 SigP	5	Protein of unknown function
5wayCG LeptoII Contig 11391 GENE 14 SigP	4	Conserved protein of unknown function
UBA LeptoII Scaffold 8241 GENE 238 SigP	3	Protein of unknown function
UBA LeptoII Scaffold 8241 GENE 573 SigP	3	Protein of unknown function
UBA LeptoII Scaffold 8524 GENE 269 SigP	3	Protein of unknown function
UBA LeptoII Scaffold 8241 GENE 522 SigP	2	Putative peptidyl-prolyl cis-trans isomerase
UBA LeptoII Scaffold 8524 GENE 127 SigP	2	Putative bacterial surface antigen (D15)
UBA LeptoII Scaffold 8241 GENE 298 SigP	1	Putative outer membrane protein
UBA LeptoII Scaffold 8524 GENE 249 SigP	1	Putative OmpA family protein

extracellular samples (triplicate analysis of 5 AMD biofilms). These highest confidence cases include functionally distinct proteins from *Leptospirillum* group II (CG and UBA strains) and *Leptospirillum* group III. In addition to proteins of unknown function, cytochromes, isomerases, and outer membrane proteins were also identified. Several proteins of unknown function exhibited high spectral counts, suggesting that they are metabolically critical. For example, we identified 531 spectra for the cleaved N-terminus of the protein encoded by *Leptospirillum* group II UBA scaffold 8049 gene 83 and its ortholog, CG scaffold 11390 gene 17. Another example is the *Leptospirillum* group II protein encoded by UBA scaffold 8524 gene 180. This ~9.6 kDa signal peptide cleaved protein contains a C-terminal region with a high scoring peptidoglycan-binding domain. This domain has been previously implicated in metalloprotease functionality.[84] Edman sequencing has also identified additional N-terminal cleavages of this protein, suggesting alternate functions that may include signal transduction or peptidic defense. By identifying signal-cleaved proteins that are constitutively and highly expressed across all samples, this study has identified a conserved pool of target proteins that are strong candidates for further in-depth functional analyses.

Table 3.5 lists some secretome signal peptide cleaved proteins whose relative abundance *differs* according to sampling location or biofilm growth state, based on calculated NSAF values.[77] These may reflect responses to differences in the surrounding physiochemical environment as the result of changing growth state and sampling location. Subtle changes in the pH, temperature, or concentrations of heavy metals could induce changes in expression of specific proteins, such as cytochromes, solute transporters and co-factors, as well as dehydrogenases, thioredoxins, cytochromes, and quinones. As expected, many of the proteins that exhibit the largest changes in abundances are currently annotated with an unknown function. In some cases, the differences in expression are quite dramatic, as in a *Leptospirillum* group III protein from scaffold 9532 gene 30, which exhibits nearly a 100 fold increase in expression in the UBA location relative to the AB-Front or AB-End samples. However, it must be noted that the reduced expression of this protein could be partly accounted for by the lower abundance of this organism in the AB-drift biofilms. The protein exhibits

Table 3.5: NSAF Comparison of Select, Highly Differential Signal Peptide Cleaved Proteins

Gene ID	AB End	AB Front	UBA	AB Muck Friable	AB Muck GSII	Function
UBA_LeptoIII_Contig_9532_GENE_30_SigP	0.0078	0.0378	0.1049	0.0091	0.0201	SMP-30/Gluconolaconase/LRE domain protein
UBA_LeptoIII_Contig_9320_GENE_13_SigP	0.0215	0.0492	0.0632	0.0037	0.0203	Phosphate ABC transport
UBA_LeptoIII_Contig_7442_GENE_13_SigP	0.0156	0.0519	0.0284	0	0.0264	Hypothetical Protein
UBA_LeptoIII_Contig_9568_GENE_73_SigP	0.0156	0.0546	0.0365	0	0.0219	Outer membrane chaperone Skp (OmpH)
UBA_LeptoII_Scaffold_8049_GENE_220_SigP	0	0	0.0215	0	0	Hypothetical Protein
UBA_LeptoII_Scaffold_8524_GENE_128_SigP	0.1169	0.1048	0.0699	0.0889	0.1095	Putative outer membrane protein (OmpH)
5wayCG_LeptoII_Contig_10608_GENE_3_SigP	0.0468	0.086	0.0434	0.0619	0.0661	Hypothetical Protein
5wayCG_LeptoII_Contig_11216_GENE_10_SigP	0.0468	0.0879	0.0434	0.0619	0.0661	Hypothetical Protein
5wayCG_LeptoII_Contig_11276_GENE_204_SigP	0.0468	0.0852	0.0434	0.0619	0.0659	Hypothetical Protein
5wayCG_LeptoII_Contig_11391_GENE_1_SigP	0	0.0248	0	0	0.0114	Protein of Unknown Function
UBA_LeptoIII_Contig_9432_GENE_53_SigP	0.0325	0.0739	0.1156	0.0133	0.032	Hypothetical Protein
UBA_LeptoIII_Contig_7980_GENE_4_SigP	0.012	0.0581	0.0201	0.0101	0.0362	Hypothetical Protein
UBA_LeptoII_Scaffold_8241_GENE_349_SigP	0.0926	0.0697	0.0278	0.0965	0.0844	Protein of Unknown Function
UBA_LeptoIII_Contig_7442_GENE_12_SigP	0.0064	0.0349	0.0454	0	0	Cytochrome
UBA_LeptoII_Scaffold_8049_GENE_83_SigP	0.0832	0.0751	0.0833	0.0516	0.0694	Protein of Unknown Function
UBA_LeptoIII_Contig_9424_GENE_148_SigP	0.0172	0.0501	0.023	0.0174	0.0278	Hypothetical Protein
UBA_LeptoII_Scaffold_8241_GENE_298_SigP	0.0823	0.085	0.0946	0.075	0.0839	Putative outer membrane prote
UBA_LeptoII_Scaffold_8135_GENE_9_SigP	0.0752	0.0455	0.1111	0.0933	0.0667	Conserved Protein of Unknown Function
UBA_LeptoII_Scaffold_8241_GENE_153_SigP	0.074	0.0452	0.0278	0.0549	0.0634	Protein of Unknown Function
UBA_LeptoIII_Contig_9545_GENE_10_SigP	0	0.0434	0	0	0.0114	Hypothetical Protein
UBA_LeptoIII_Contig_9205_GENE_91_SigP	0.0136	0.0453	0.0655	0.0118	0.0263	Hypothetical Protein
UBA_LeptoII_Scaffold_7931_GENE_87_SigP	0.0678	0.0468	0.0523	0.0434	0.0461	Putative Cytochrome C
UBA_LeptoII_Scaffold_8241_GENE_348_SigP	0.0667	0.0295	0.0136	0.0362	0.035	Putative outer membrane protein
UBA_LeptoIII_Contig_9568_GENE_74_SigP	0.0023	0.0081	0	0	0	Surface antigen (D15)
UBA_LeptoII_Scaffold_8241_GENE_114_SigP	0.065	0.0514	0.0966	0.0526	0.0566	Putative glycosyl hydrolase
UBA_LeptoII_Scaffold_8241_GENE_238_SigP	0.0612	0.0491	0.0412	0.0355	0.0421	Protein of Unknown Function
UBA_LeptoII_Scaffold_8241_GENE_693_SigP	0.0607	0.0434	0.0742	0.0449	0.0483	Periplasmic phosphate binding protein
UBA_LeptoII_Scaffold_8062_GENE_53_SigP	0.0594	0.0912	0.0492	0.0673	0.0686	Protein of Unknown Function
UBA_LeptoII_Scaffold_8241_GENE_121_SigP	0.0573	0.0368	0.0304	0.0372	0.0406	Peptidase S
UBA_LeptoII_Scaffold_7931_GENE_365_SigP	0.0555	0.0374	0.0246	0.0355	0.0356	Protein of Unknown Function
UBA_LeptoII_Scaffold_8524_GENE_249_SigP	0.0542	0.08	0.0576	0.0654	0.0712	Putative OmpA family protein
UBA_LeptoII_Scaffold_7931_GENE_101_SigP	0.0518	0.0354	0.0754	0.0237	0.0238	Protein of Unknown Function
Unass_bact_scaff_903_GENE_5_SigP	0.0514	0.0155	0.0082	0.0167	0.0204	Hypothetical Protein
Unass_bact_scaff_1131_GENE_3_SigP	0.0497	0.04	0	0.0279	0.0428	Hypothetical Protein
5wayCG_LeptoII_Contig_11277_GENE_32_SigP	0.0472	0.0251	0.0309	0.0311	0.0263	Putative peptidase M16
UBA_LeptoII_Scaffold_8049_GENE_192_SigP	0.0461	0.033	0	0.0236	0.0343	Protein of Unknown Function

high BLAST sequence similarity (E-value > 9^{-59}) to numerous proteins containing a NHL repeat. This feature has been shown to confer catalytic activity in monooxygenases and serine/threonine kinases.[85] Additionally, several high scoring BLAST hits correspond to SMP-30/gluconolaconase/LRE domain-containing proteins. This annotation describes a region of sequence similarity observed in a variety of bacterial and archaeal enzymes. A putative ABC Transporter, 5wayCG *Leptospirillum* group III contig 9320 gene 13, was also inferred to show variation in abundance levels among samples. Finally, an annotated cytochrome encoded by UBA *Leptospirillum* III scaffold 7442 gene 12 is identified in relatively high abundance in the AB-End, UBA and AB-Front samples, but is not identified in the AB-Muck samples. This is in stark contrast to the previously mentioned Cyt_{579} which is ubiquitously identified in all samples. These results suggest that protein expression patterns reflect varying responses to local environmental conditions or biofilm age.

We conducted Pfam domain analysis on the 46 proteins identified with a signal peptide cleaved N-terminus. Nine proteins contain domains currently annotated in the Pfam database (**Table 3.6**), including cytochromes, outer membrane folds, catalytic sites from metabolic enzymes, and multiple Pfam domains. These domains correspond well with the predicted cellular extracytosolic location of the proteins. Additional domains include those involved in lipid binding, proteolytic digestion, and protein folding. Pyrrolo-quinoline quinone (PQQ) illustrates a common repeat that results in a characteristic beta-propeller structure found within proteins utilizing prosthetic quinones (integral members of electron transport chains).[86] Within our analysis, the PQQ repeat was identified in a *Leptospirillum* group II protein, encoded by scaffold 8241 gene 348, which is currently annotated as an outer membrane protein. The *Leptospirillum* group II protein, from scaffold 8062 gene 173 displays a high scoring (9.7×10^{-79}) Pfam identification to a cytochrome c peroxidase domain (CCP_MauG). CCP_MauG proteins have been found within the periplasmic space of gram-negative bacteria and are known to use two heme groups to reduce hydrogen peroxide without the formation of free radicals.[87] Another prevalent domain was the NHL tandem repeat (described above), which was identified multiple times within two proteins currently annotated as having

Table 3.6: Pfam Domain Analysis of Conserved and High Confidence Signal Peptide Cleaved Proteins

Name	Start #	End #	Pfam Acc #	E-value	Pfam ID
5wayCG LeptoII Contig 11233 GENE 46 SigP	86	115	PF08450.3	2.60E-06	SGL
5wayCG LeptoII Contig 11233 GENE 46 SigP	137	165	PF01436.12	5.90E-06	NHL
5wayCG LeptoII Contig 11233 GENE 46 SigP	25	52	PF01436.12	7.50E-05	NHL
5wayCG LeptoII Contig 11233 GENE 46 SigP	195	223	PF01436.12	8.40E-03	NHL
5wayCG LeptoII Contig 11238 GENE 58 SigP	13	177	PF00160.12	9.00E-60	Pro_isomerase
5wayCG LeptoII Contig 11238 GENE 99 SigP	150	233	PF00034.12	1.20E-03	Cytochrom_C
5wayCG LeptoII Contig 11391 GENE 14 SigP	13	178	PF04264.4	5.90E-54	YceI
UBA LeptoII Scaffold 7931 GENE 111 SigP	150	233	PF00034.12	1.20E-03	Cytochrom_C
UBA LeptoII Scaffold 7931 GENE 338 SigP	69	104	PF08238.3	9.60E-10	Sel1
UBA LeptoII Scaffold 7931 GENE 338 SigP	177	212	PF08238.3	2.90E-08	Sel1
UBA LeptoII Scaffold 7931 GENE 338 SigP	141	176	PF08238.3	9.30E-08	Sel1
UBA LeptoII Scaffold 7931 GENE 338 SigP	105	140	PF08238.3	6.80E-06	Sel1
UBA LeptoII Scaffold 7931 GENE 338 SigP	33	68	PF08238.3	1.80E-04	Sel1
UBA LeptoII Scaffold 7931 GENE 338 SigP	213	248	PF08238.3	2.00E-03	Sel1
UBA LeptoII Scaffold 7931 GENE 73 SigP	13	177	PF00160.12	9.00E-60	Pro_isomerase
UBA LeptoII Scaffold 8049 GENE 366 SigP	14	179	PF04264.4	3.90E-53	YceI
UBA LeptoII Scaffold 8062 GENE 173 SigP	1	171	PF03150.5	9.70E-79	CCP_MauG
UBA LeptoII Scaffold 8135 GENE 9 SigP	101	130	PF08450.3	8.20E-07	SGL
UBA LeptoII Scaffold 8135 GENE 9 SigP	152	180	PF01436.12	5.90E-06	NHL
UBA LeptoII Scaffold 8135 GENE 9 SigP	25	52	PF01436.12	6.40E-05	NHL
UBA LeptoII Scaffold 8135 GENE 9 SigP	210	238	PF01436.12	8.40E-03	NHL
UBA LeptoII Scaffold 8241 GENE 298 SigP	40	139	PF00691.11	1.30E-24	OmpA
UBA LeptoII Scaffold 8241 GENE 348 SigP	254	292	PF01011.12	7.40E-05	PQQ
UBA LeptoII Scaffold 8241 GENE 348 SigP	107	144	PF01011.12	1.20E-04	PQQ
UBA LeptoII Scaffold 8241 GENE 348 SigP	213	250	PF01011.12	1.40E-04	PQQ
UBA LeptoII Scaffold 8241 GENE 348 SigP	399	435	PF01011.12	1.30E-02	PQQ
UBA LeptoII Scaffold 8241 GENE 522 SigP	66	145	PF09312.2	3.90E-16	SurA_N
UBA LeptoII Scaffold 8241 GENE 522 SigP	163	256	PF00639.12	2.00E-08	Rotamase
UBA LeptoII Scaffold 8241 GENE 522 SigP	5	26	PF09312.2	1.40E-03	SurA_N
UBA LeptoII Scaffold 8241 GENE 693 SigP	1	156	PF01547.16	7.00E-04	SBP_bac_1
UBA LeptoII Scaffold 8524 GENE 127 SigP	432	748	PF01103.14	8.70E-36	Bac_surface_Ag
UBA LeptoII Scaffold 8524 GENE 127 SigP	252	330	PF07244.6	4.20E-22	Surf_Ag_VNR
UBA LeptoII Scaffold 8524 GENE 127 SigP	333	405	PF07244.6	5.20E-18	Surf_Ag_VNR
UBA LeptoII Scaffold 8524 GENE 127 SigP	8	79	PF07244.6	1.10E-14	Surf_Ag_VNR
UBA LeptoII Scaffold 8524 GENE 127 SigP	160	249	PF07244.6	7.10E-14	Surf_Ag_VNR
UBA LeptoII Scaffold 8524 GENE 127 SigP	80	157	PF07244.6	2.00E-12	Surf_Ag_VNR
UBA LeptoII Scaffold 8524 GENE 128 SigP	1	159	PF03938.5	6.80E-18	OmpH
UBA LeptoII Scaffold 8524 GENE 180 SigP	39	90	PF01471.9	5.30E-12	PG_binding_1
UBA LeptoII Scaffold 8524 GENE 249 SigP	129	224	PF00691.11	1.10E-42	OmpA
UBA LeptoIII Contig 7952 GENE 72 SigP	13	178	PF04264.4	1.00E-58	YceI

unknown functions (encoded by *Leptospirillum* group II CG contig 11233 gene 46 and *Leptospirillum* group II UBA scaffold 8135 gene 9).[85] A YceI-like domain was also found with high confidence (3.90×10^{-53}) in a protein of unknown function from *Leptospirillum* group II (encoded by scaffold 8049 gene 366). This domain is characterized by a beta-barrel motif and functions in lipid binding. A previous study of *E. coli* resulted in the identification YceI as one of three proteins currently annotated with an unknown function but showed a marked response to pH.[88] Domain prediction is not conclusive evidence for protein function, but it does provide valuable insight when coupled with the determination of extracellular location and signal peptide cleavage. The previous high-scoring domain identifications highlight the diversity of the extracellular fraction as well as the need for continued study.

The abundances of the signal peptide cleaved proteins identified in this study were examined using a more extensive and previously published dataset for 28 biofilm samples[38] to more comprehensively define changes in protein abundances across the AMD environment. Samples have been classified as low or high developmental stage biofilms based on their observed maturity (see experimental procedures).

Of the 377 non-redundant signal peptide cleaved proteins identified in this study, 174 were also found in the 28 biofilm proteomes. The previous study focused on the whole cellular proteome and thus the extracellular fractions of these samples were not implicitly retained and analyzed separately. Thus, this captures the composite total of all proteins identified, whether or not they are specifically exported to the extracellular region. The lower rate of identification of signal peptide cleaved proteins in this case is consistent with their inferred periplasmic or extracellular location. Clustering of the NSAF values for these proteins revealed distinct trends in the protein abundances with respect to developmental stage (**Figure 3.5A**). Each row in Figure 5 represents one of the 174 identified signal peptide cleaved proteins, with yellow indicating high expression (MS detection) and blue indicating low expression (MS detection). Based on the clustering of samples (across the x-axis), it is evident that the abundances of signal peptide cleaved proteins correlates significantly with biofilm growth state. Specifically, samples representing low developmental stage biofilms (green highlights) generally

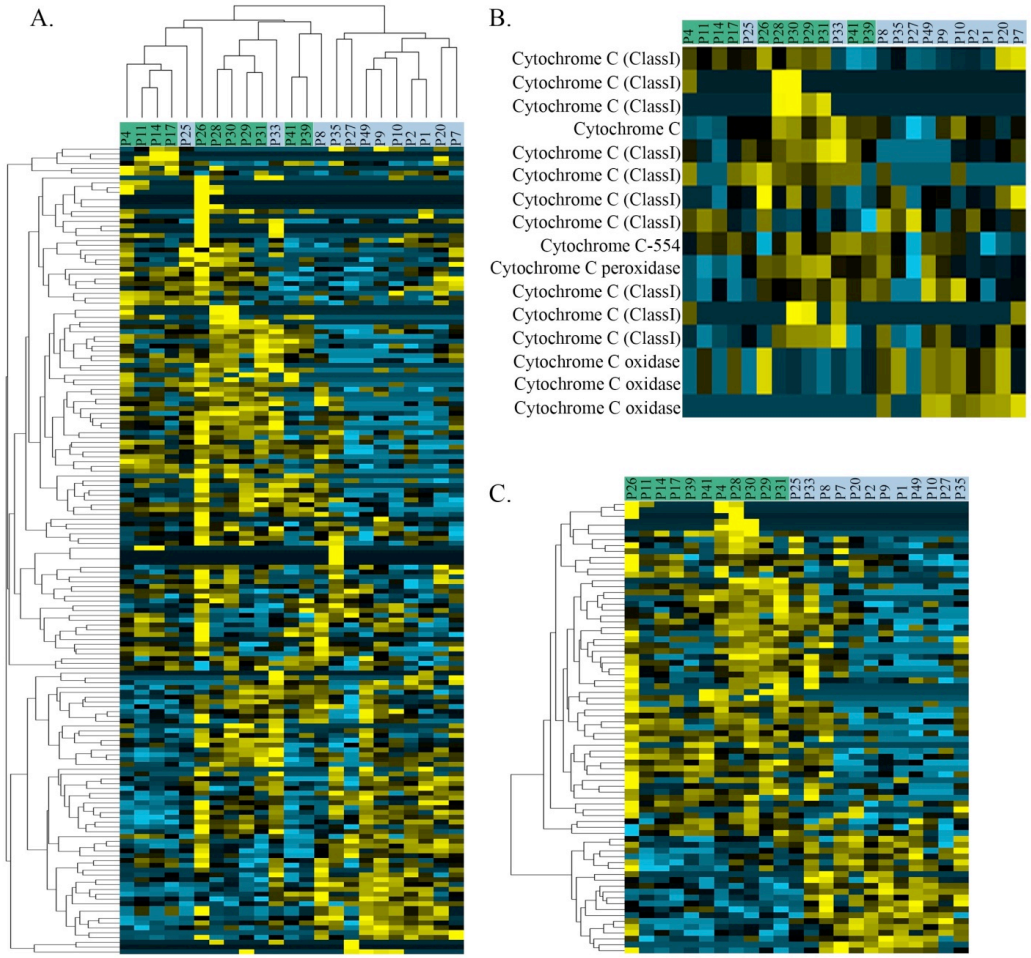

Figure 3.5: NSAF Cluster Analysis of Signal Peptide Cleaved Proteins Identified in 28 Biofilm Samples

Cluster analysis of NSAF values from proteins identified within the 28 biofilm samples. 5A displays the low developmental stage (green) and high developmental stage (blue) of the 28 biofilm samples and the clustering based on proteomic expression. Rows represent individual signal peptide cleaved proteins with yellow indicating increased expression and blue indicating low expression. Figure 5B highlights the growth state expression dynamics of AMD cytochromes. Figure 5C represents a subset of the proteins identified in figure 5A that exhibit dramatic expression changes as a function of growth state. Among these, numerous cytochromes are present in the early growth states, while several chemotaxis proteins are present in late growth stages.

cluster tightly, but separately from a cluster of samples representing high developmental stage biofilms (blue highlights). When the clustering of proteins based on their abundances across samples is examined (down the y-axis), it is noted that there is a subset of predicted signal peptide cleaved proteins that exhibit high abundances in low developmental stage biofilms, but low or no detectable expression in high developmental stage biofilms. Similarly, another subset displays no or low expression in low developmental stages and increased expression in high biofilm developmental stages.

An interesting result of these analyses was the NSAF-based abundance trends of numerous signal peptide cleaved cytochromes (**Figure 3.5B**). In general, we detect early expression of class I cytochromes, whereas cytochrome oxidases appear to be abundant later in development. These results most likely denote shifts in metabolism, which occur as biofilms age. These results are consistent with the increased abundance of Cyt_{579} and *c*-type cytochromes in early development stage biofilms.[89]

Other significant differences in the abundances of proteins between the two developmental stages were also defined (**Figure 3.5C**), with many currently annotated with no known function. As identified in the analysis of the five biofilms, we noted that the low developmental stage displays numerous cytochromes that are not identified in high developmental stages. Conversely, it was found that two chemotaxis sensory proteins were in greater abundance in high developmental stages. An increase in chemotaxis protein expression may result from the depletion of nutrients that may occur as biofilms age and more organisms colonize the environment. Proteomic adaptation, through dynamic expression of signal peptide cleaved proteins, may assist these microbes in identifying regions of the biofilm where nutrients are not limiting.

Finally, the potential proteomic adaptation of secreted proteins to the highly acidic AMD environment was probed by utilizing the pool of predicted and identified signal peptide cleaved proteins as a representation of the extracellular fraction. Protein adaptation to acidic environments has been examined in two previous studies that have compared the calculated isoelectric points (pI) of proteins from organisms that tolerate and/or grow within highly acidic conditions and those from more mesophilic

organisms.[90, 91] The results of these two studies are in conflict, with one finding significant differences and the other not. One reason for this conflict may be that both studies include the complete genomes of each organism to calculate median pIs. Including all *predicted proteins* in these analyses, even those that are not exposed to highly acidic extracellular environments, can introduce unintended biases. In an earlier study, we examined the predicted pIs for proteins of the extracellular fraction of the AB-End biofilm and determined that the distribution exhibited a bi-modal appearance with the largest proportion of proteins falling between 9 - 11.[19] However, this study did not explicitly resolve proteins with signal peptides. In this current work, the median pI of proteins predicted to have signal peptides from *Leptospirillum* Group II was compared to the remaining pool of proteins from this organism and a significant difference was observed (median pI of SigP proteins = 9.10, median pI of remaining proteins = 6.95; t-test; *p*-value < 1×10^{-6}). This distribution of pIs closely followed the previous study, with a significant proportion falling between 9 – 9.9. In this study, the protein sampling size was over ten times larger than the previous study, providing increased confidence in the pI determination. A potential caveat of this methodology is the inability to include secondary or tertiary protein structure. For example, it has previously been found that a maltose-binding protein from a thermoacidophilic bacteria has a calculated pI of 6.5 and a measured pI of 10, and this discrepancy is due to the large number of basic residues constituting the solvent exposed face of the protein.[92] Therefore, future studies examining protein adaptation to various environments will need to account for perceived differences in amino acid sequence and pI within the context of protein localization and the three-dimensional structure of a given protein. Given that the identified signal peptide cleaved proteins are known to be functional outside of the cytosol, they would serve as excellent candidates for detailed biochemical analysis of protein adaptation to extreme environmental conditions of the AMD environment.

3.5: CONCLUSIONS

We have integrated computational prediction with experimental verification as a methodology for validating, characterizing, and comparing signal peptide cleavage from

an acidophilic microbial consortium. The ability to validate computational prediction of signal peptide cleavage by mass spectrometry at the peptide level has enabled refinement of the secretome. Analysis of the AMD protein database resulted in the prediction of over one thousand potential signal peptide cleaved proteins. Without experimental verification, the validity and confidence of the assignments is uncertain. By combining the prediction with high throughput LC-MS/MS techniques, we were able to confidently identify hundreds of signal peptide cleaved proteins. No marked differences in signal peptide cleaved protein identification were observed relative to the distribution of species in the biofilm, as expected. What is evident though is the degree of conservation and divergence of exported signal peptide cleaved proteins from varying sampling locations. This supports the inference that the proteome is dynamic, depending on local environmental conditions or biofilm age. These results are also supported by examining the expression patterns of the proteins identified in this study within a larger sample set (28 samples) representing four years of sample collection. Here, distinct sets of signal peptide cleaved proteins were associated with both low and high developmental stage biofilms. This study also highlights the predominance of proteins that are annotated as either hypothetical or with an unknown function in the expressed proteomes, since the majority of identified signal peptide cleavage proteins fall within these two categories. By combining the results of Pfam analysis with the newly obtained information of potential cellular location and signal peptide cleavage, it is possible to at least partially decipher the role of some of the putative unknown proteins. The integrated prediction and identification of proteins that are specifically targeted to extra-cytosolic locations and the characterization of their expression patterns in this study has identified numerous proteins that are essential for many key functions within the AMD system.

Chapter 4

An Integrated, Comparative Metal Affinity Enrichment and Proteomic Characterization of Novel Proteins from a Natural Microbial Community

Portions of included text are adapted from:

Brian K. Erickson, Korin E. Wheeler, Ryan Mueller, Steven W. Singer, Nathan C. VerBerkmoes , Mona Hwang, Jillian F. Banfield , Michael P. Thelen, and Robert L. Hettich, "Evaluation of a Varied Metal Affinity Enrichment Strategy for Expanding the Dynamic Range of Extracellular Proteome Characterization for a Natural Microbial Community" Manuscript in preparation.

Brian K. Erickson's contributions include experimental preparation of samples, experimental LC-MS/MS analysis, data parsing, and primary authorship.

4.1: Introduction

Essential to a complete understanding of the organization and functions of microorganisms within natural environments is the ability to characterize the spectrum of expressed proteins that reveal detailed metabolic activity information. Identification of low-abundance proteins and other proteins that are difficult to detect remains a formidable challenge due to dominant, high-abundance proteins in the complex proteome. In particular, highly abundant proteins interfere with the measurement of more minor proteins by causing problems in both chromatographic separations and mass spectrometric measurement. For example, minor protein biomarkers and sensor proteins in particular are generally difficult to identify within the complexity of the proteome.

Affinity chromatography has been established as an effective method to increase depth of proteomic identification.[93] Recent approaches include the use of combinatorial libraries [94, 95] to enrich for low abundance proteins in model systems. Such approaches are useful to indiscriminately enrich for low-abundance species, or other challenging proteins. More targeted studies have included heparin chromatography for enrichment of brain signaling proteins [96] and the use of lectin affinity chromatography to capture the glycoproteome [97]. However, all of these approaches suffer related problems in robustness and selectivity/sensitivity of enrichment.

Metals provide a unique ability to bind a variety of ligands, including proteins, with relative specificity. Recent work has demonstrated metal-affinity chromatography[98] and mass spectrometry to selectively enrich proteins[99-101] , including post-translational modifications such as phosphorylation[25]. Thus far, immobilized metal affinity chromatography (IMAC) studies have utilized primarily a single metal affinity column in proteomic enrichment approaches[102, 103]. Enrichments utilizing one metal present several limitations, including ambiguous specificity and the pervasive masking of low abundance proteins by ubiquitous proteins. IMAC specificity of biomolecules for metals is roughly dictated by hard and soft lewis acidity of the metal. Hard acids, such as Fe^{3+} or Mg^{2+}, preferentially bind to ligands with an oxygen (hard bases), i.e. phosphates or aspartic/glutamic acid; soft acids, such as Cu^{+} and Hg^{2+}, preferentially bind to thiols, i.e. cysteine. However, the strength of metal binding is generally dictated by the Irving-Williams series, which states that biomolecules will nonspecifically bind to metals higher in the series (i.e., greater negative hydration energy). IMAC enables the enrichment of proteins with metal-affinity, but has several limitations. IMAC is not specific for physiologically active metal binding, nor does it work well for strongly bound metal–protein complexes[101]. The use of IMAC columns provides efficient and reproducible enrichment of selective proteins, thus highlighting the potential application of this approach for deepening proteomic coverage.

The coupling of affinity-based protein purification with high-performance liquid chromatography and high-throughput mass spectrometry (MS) brings together powerful tools to increase the dynamic range of proteomic identifications. This combination is

aptly suited for extensive protein purification, identification, and characterization. MS has become the gold-standard tool for comprehensive and high throughput proteomic characterizations, providing an unparalleled depth of proteomic coverage. The ability of MS to provide high confidence protein identification, relative quantification, and details of protein modification make this technique well suited for the identification and characterization of a broad spectrum of proteins within the proteome.[2, 3] The high dynamic range and broad characterization achievable through MS based proteomics is unmatched by previous methods, including 2D SDS-PAGE.

Here, we report a method of selective IMAC enrichment and MS identification of proteins across a library of biologically active metals: copper, cobalt, manganese, magnesium, nickel, zinc, and iron. This broad spectrum of metals provides enrichment across an array of ligands and enables identification of a wide range of proteins. In addition, enrichment within specific metal columns provides insights that can be harnessed for selective purification of a particular protein or groups of proteins for further studies. Metals were chosen as affinity tags in this study to enhance the depth of proteome coverage within a natural, well-studied acid mine drainage (AMD) microbial community from the Richmond Mine at Iron Mountain (Redding, CA). The AMD community is low diversity and well characterized, serving as an ideal system for investigations into biogeochemical interactions[19, 20, 68]. Of particular interest are extracellular proteins that mediate interactions between the microorganisms and the environment, which includes low pH and metal rich conditions. These iron oxidizing bacteria maintain metal homeostasis within molar concentrations of iron and millimolar concentrations of copper, zinc and arsenic.[104] Extracellular proteins are critical for nutrient transport, organismal communication, and defense mechanisms. This study provides as yet another puzzle-piece in the emerging picture of the AMD microbial community system, complementing details already known about extracellular proteins[105-107], community membership[19], growth state dependence[108], and strain variation[69, 109] in this model environmental microbial consortium. The coupling of IMAC enrichment and MS analysis serves to deepen the dynamic measurement to identify

lower abundance proteins, including proteins of unknown function, and may open a new route for targeted studies of metabolic functions.

4.2: Methodology

Immobilized metal affinity chromatography (IMAC) was used to enrich for proteins with metal binding ligands. A 5ml HiTrap Chelating HP column (GE Healthcare) was equilibrated with 0.1M metal in buffer A (20mM, pH5 MES), followed by loading of the metal salt in the same buffer. Metals tested included $ZnSO_4$, $CuSO_4$, $Fe_2(SO_4)_3$, $CoSO_4$, $MgSO_4$, $NiSO_4$, and $MnSO_4$. Following metal loading, the column was rinsed with 50ml binding buffer to assure that only bound metal was left on the column. A total of 0.5mg extracellular protein was then added to the metal loaded column and rinsed with 50ml binding buffer. Eluent with protein from the buffer rinse were labeled as the 'unbound' column fraction. The column was then rinsed with buffer B, 20mM MES with 0.05M EDTA and 0.5M NaCl at pH 5. The eluent after washing with buffer B was labeled the 'bound' column fraction. Each column run was repeated, for a total of three technical replicates.

Binding of proteins to the IMAC column was monitored by absorbance at 280nm, 1D SDS PAGE, and Bradford protein assays. After initial analysis, the protein column fractions were precipitated with 100µl 100% trichloroacetic acid solution for every 900µl of protein solution. The sample was then incubated at 4°C for 1 hr. Samples were centrifuged (10min at 25000*g*) and supernatant was removed. The protein pellet was washed with 4°C methanol and air-dried. Samples were stored at -80°C until trypsin digestion.

The "Biofilm_AMD_CoreDB_04232008.fasta" database contains annotated proteins from the abundant microbial members of AMD biofilms. The protein database also includes common contaminants (trypsin, keratin, etc.). Protein assignment of the MS/MS spectra was accomplished with the SEQUEST algorithm[25] and was executed with the following parameters: enzyme type, trypsin; Parent Mass Tolerance, 3.0; Fragment Ion Tolerance, 0.5; up to 4 missed cleavages allowed, and fully tryptic peptides only. Resulting output files were sorted and filtered using DTASelect[71] with the

following parameters: tryptic peptides only, deltaCN value of at least 0.08, and Xcorr values of at least 1.8 (+1), 2.5 (+2), 3.5 (+3) with a two peptide minimum. Cross-comparison among DTASelect output was accomplished with Contrast and an in-house script that provides similar functions.

The proteins were assigned designations according to their enrichment in the IMAC columns and their resulting normalized spectral abundance factor (NSAF)[77] values from the mass-spectrometry runs. For each metal, a protein was assigned a characterization of: “IMAC bound”, “unbound”, or “false” in accordance to criteria based upon their detection in each technical replicate and abundance within the bound and unbound fractions. The protein was classified as “IMAC bound” if both of the following conditions were satisfied in both replicates: 1). The abundance in the bound fraction was non-zero (indicating detection); 2). The abundance in the unbound fraction was less than the abundance in the bound fraction, *or* the protein was undetected in the unbound fraction. The protein was classified as “unbound” if both of the following conditions were satisfied with both replicates: 1). the abundance in the unbound fraction was non-zero (indicating detection); 2). The abundance in the bound fraction was less than the abundance in the unbound fraction, *or* the protein was undetected in the bound fraction. All other proteins were labeled “false”, indicating the thresholds were not met for IMAC bound or unbound designation.

The results were clustered using Cluster version 3.0 (http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm#ctv) and visualized as heat maps with TreeView software. Proteins of unknown function exhibiting enrichment on a specific IMAC column were batch submitted to Pfam for domain and motif analysis[110]. Pfam was executed with the following parameters: merged global and local strategy and an E-value cutoff of 1.0. The resulting identifications were then further filtered at an E-value $< 1 \times 10^{-3}$.

4.3: Results

Bottom up LC-MS/MS measurements of metal affinity enriched extracellular protein samples from the AMD community resulted in an expansive dataset of proteins (**Figure 4.1**). The dataset was interrogated for reproducibility, identification of groups of proteins with specific versus universal IMAC enrichment, and IMAC enrichment of abundant amino acids.

4.3.1: IMAC Column Binding

Each of the seven metals loaded onto the IMAC columns were divalent, with the exception of Fe(III). The ferric column was utilized because ferrous iron is insoluble at pH 5.0. All divalent metals yielded a significantly higher concentration of proteins in the unbound column fraction when compared with the bound fraction, as analyzed by absorbance at 280 nm and 1D-SDS-PAGE. Notably, within the divalent metals, cobalt had the highest abundance of bound proteins and magnesium had the lowest.

In contrast, the bound fraction from the ferric column had a much higher concentration of protein. SDS PAGE analysis shows selective binding of an abundant ~16 kDa protein, the dominant soluble *Lepto II* protein cytochrome 579. This protein appears primarily in the unbound fraction of each of the other IMAC columns.

A control column was run without metal loaded to verify that chromatographic results were indicative of metal affinity and not affinity for the column resin. Not surprisingly, the results from the control column deviated significantly from the metal loaded IMAC columns, indicating few proteins bound to the IMAC column beads alone. The bound protein fraction contained only a small amount of residual protein, as indicated by silver stained SDS PAGE with only a faint band at 20kDa. MS results indicate this band is due to the dominant soluble *Lepto II* protein cytochrome 579.

4.3.2: 2D-LC-MS/MS Measurements

In total, 485 non-redundant proteins were identified across the thirty-two IMAC preparations with subsequent mass spectrometric measurements. This value includes replicate determinations of the bound and unbound fractions from each of the seven

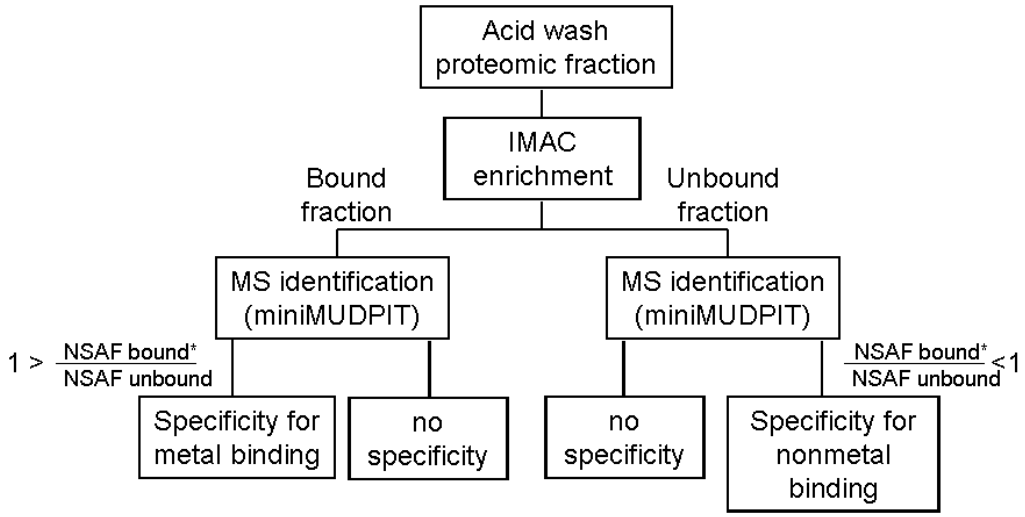

Figure 4.1: Schematic Representation of Experimental Design

Each step was repeated twice to provide technical replicates. As indicated by the asterisk, a protein was labeled with specific binding only if the appropriate ratio was found in both replicates.

metals, as well as the control column. The total number of identified proteins correlates well with the expected number of proteins in the extracellular fraction identified in a previous, non-enriched analysis (531 extracellular proteins).[105] A pooled average of 370 non-redundant proteins was identified in each column. The identified proteins were often found across multiple IMAC columns, with an overlap ranging from 13% to 93% among the columns. Reproducibility among replicates was very good, at an average of 9% variability, with several replicates exhibiting a deviation as low as 0.5 or 1%. One exception must be noted for the cobalt-IMAC, which had a higher standard deviation of 27% for the unbound and 48% for the bound column fractions. It appears that a remarkably high number of proteins were identified in one of the two mini-MUDPIT measurements (472 identified proteins compared to an average of 354 in other samples), potentially inflating the relative error for the Co-IMAC. In every column except Fe^{3+}, there were a total of two to five times as many proteins identified within the IMAC bound fraction as compared to the unbound fraction (**Figure 4.2**).

The least number of proteins (35) identified was within the bound fraction of the control column. Correspondingly, the highest number of proteins identified was within the unbound fraction (438) of the same column. This disparity indicates that few proteins have affinity for the column resin (without loaded metal).

4.3.3: Classification of Specificity of IMAC Enrichment

Selection criteria based upon each protein's binding profile and normalized spectral abundance factors (NSAF) from MS experiments were used to assign metal specificity, if any, to each protein.[77] Each protein was designated as "bound", "unbound", or "false" for each metal. A definitive bound or unbound designation indicates that the protein was detected in both replicates and had a consistently higher NSAF value in the respective column fraction. Criteria were designed to reduce false positives. Proteins that were either poorly detected or were not detected in abundance in either the bound or unbound fraction of a column were labeled 'false'.

Table 4.1 summarizes the results of this classification of enriched proteins as uniquely bound or uniquely unbound to each IMAC column. The number of proteins

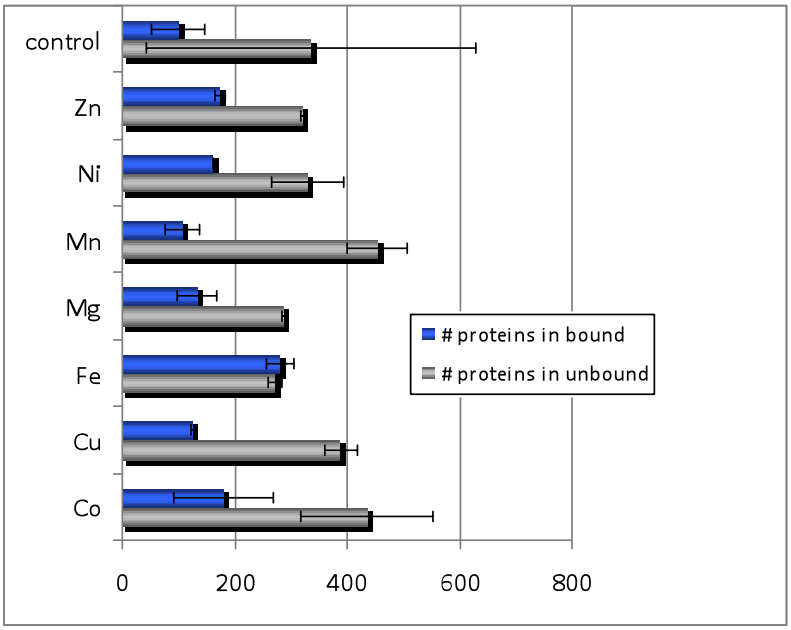

Figure 4.2: Number of Proteins Identified per Column

Number of proteins identified by MS-proteomics within the unbound (grey) and bound (blue) chromatographic fractions. Error bars indicate standard deviation from average.

Table 4.1: Number of Proteins Labeled as Uniquely IMAC Bound or Unbound After Analysis of MS Data from Chromatographic Fractions

	Fe^{3+}	Co^{2+}	Mn^{2+}	Ni^{2+}	Cu^{2+}	Mg^{2+}	Zn^{2+}	control
Unbound	65	67	49	59	57	45	88	0
Bound	72	210	274	177	229	158	165	55
Total	137	277	323	236	286	203	253	55

designated as bound for each metal follows the following trend: Mn(II)> Cu(II) > Co(II) > Ni(II) > Zn(II) > Fe(III). Somewhat surprisingly, this trend is inconsistent with Hard Soft Acid Base theory, used to predict metal binding preferences and the Irving-Williams series for divalent metals (Mn(II) < Co(II) < Ni(II) < Cu(II) > Zn(II)) generally used to predict the strength of metal binding for divalent metals. Additionally, the trend does not overlay with the metal's abundance in the AMD environment.

Of the 485 non-redundant protein identifications, each was classified within three general categories: bound to a specific IMAC column (specific IMAC enrichment), bound to multiple IMAC columns (universal IMAC enrichment), or non-bound to any metals. The majority of identified proteins (295, 61%) can be classified as universally IMAC enriched proteins; 137 (28%) classified as specifically IMAC enriched; and the remaining 53 proteins (11%) classified as preferentially non-bound for any of the IMAC columns (**Figure 4.3**).

4.3.4: Amino Acid Specificity of IMAC Enrichment

Clearly, IMAC enrichment of native proteins is limited to selection of proteins with exposed ligands for metal binding. Identification of proteins enriched by IMAC is not necessarily indicative of functional metal binding, nor strength of metal binding beyond a minimum threshold. It has been previously shown that IMAC purification can be influenced by biases in the residue content of proteins (particularly histidine). In an effort to determine if any such biases exist in the proteins identified through the IMAC-MS/MS analysis, the residue content of the 485 non-redundant proteins were evaluated (**Table 4.2**). Among histidine, cysteine, and methionine, no bias was observed. However, among the proteins that were identified as binding to Zn, Cu, and Mg, a slight increase in histidine presence was observed, while the proteins binding to Co, Mn, and Ni resulted in a slight decrease in histidine content. Further analysis of the residue content of identified proteins by highlighting physiochemical properties both before and after sequence alignment resulted in similar results, displaying no discernable bias among amino acid content.

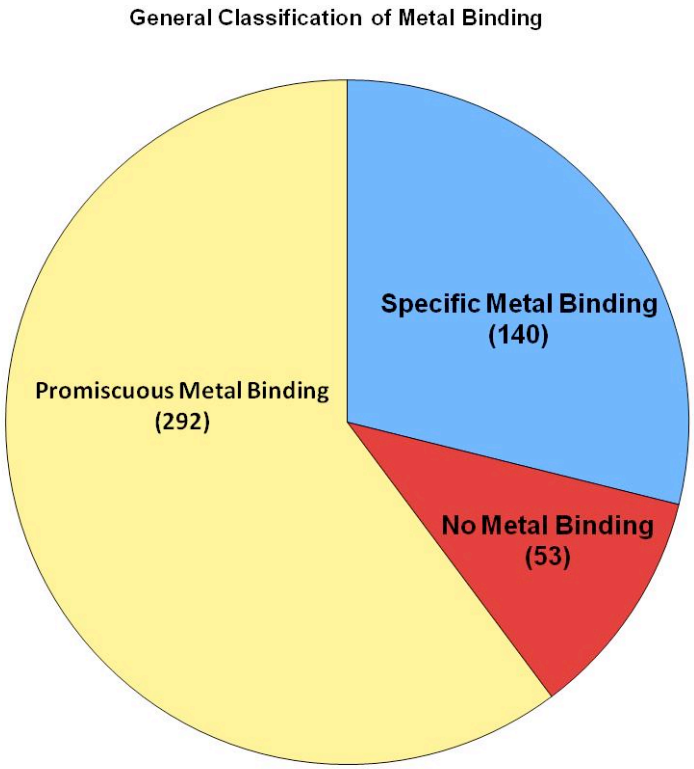

Figure 4.3: General Classification of Metal Binding

The majority of identified proteins exhibited binding affinity for multiple metals

Table 4.2: Average Amino Acid Abundance within Protein Groups that are Bound to Specific Metals, or Groups of Metals

(averages)	total aa	His	Cys	Met	acidic	basic	hydrophobic	hydrophilic
all metals	228.51	1.90	0.91	2.46	17.28	29.48	44.57	55.43
all	220.50	1.63	1.09	2.39	17.98	28.63	42.65	57.35
divalents	232.92	2.15	0.70	2.82	16.92	32.23	44.25	55.75
ZnCuMg	206.42	2.45	1.24	2.10	12.96	31.12	43.37	56.63
CoMnNi	186.80	1.22	0.75	3.35	18.40	26.62	41.73	58.27
Cu	244.50	1.91	0.38	2.55	18.16	31.47	42.53	57.47
Ni	339.60	1.74	0.43	3.04	25.60	30.04	44.14	55.86
Co	232.50	1.25	0.51	2.97	19.91	28.87	42.60	57.40
Zn	234.64	2.22	1.14	2.86	19.49	33.58	42.40	57.60
Mn	263.69	2.32	1.08	3.37	22.87	33.84	42.86	57.14

4.3.5: PFam Analysis of Specifically Enriched Proteins of Unknown Function

Among the identified IMAC enriched proteins, it is apparent that a significant subset of the proteins has an unknown function. A portion of the proteins with an unknown function were identified as binding to a specific metal (specific IMAC enrichment). These proteins contain no reasonable sequence homology to any currently characterized proteins. Forty five proteins of unknown function were chosen for more detailed analysis because they exhibited selective binding for one specific metal. Functional domains and motifs were computationally predicted by batch submission of these forty five unique protein sequences to Pfam[110]. Fifty-six unique Pfam identifications were predicted, representing multiple domains and motifs, with some proteins containing multiple, high scoring hits.

4.4: Discussion

Among the 485 extracellular proteins identified here, 116 had not been identified in previous MS measurements of the AMD community proteome [19, 105, 108]. Among multiple replicated MS analyses of the extracellular fraction,[105] it is estimated that approximately 500 – 600 proteins can be reproducibly measured. Thus, the 116 newly identified proteins represent a 23% increase in proteomic measurement depth of the extracellular fraction. The ability to bind and enrich proteins allows for deeper coverage of the proteome and, additionally, verification of numerous proteins that were previously only predicted to exist. These newly identified proteins, formerly annotated as hypothetical, are likely of relatively low abundance due to their novel identification following IMAC enrichment. The 116 proteins originate from each of the major microbes within the community and comprise a variety of functional annotations. The ability to identify and characterize a significant number of additional proteins in the extracellular space highlights the benefits of an expansive enrichment and MS detection strategy.

The newly identified proteins have a wide variety of predicted functions within the extracellular space. As reported previously, over 57% of the extracellular proteome consists of proteins of unknown function[105]. It is not surprising then, that a significant portion (35%) of the newly identified proteins reported here have an unknown function.

Additionally, a subset of the novel proteins were previously annotated as hypothetical, but can now be re-annotated as 'proteins of unknown function' due to their definitive measurement. The remaining proteins have expected extracellular functional annotations, including secretion/efflux/transport, cytochromes, dehydrogenases/reductases, and kinases. Among the newly identified proteins, neither physicochemical (molecular weight or pI) biases nor functional bias appear within the subgroup. The IMAC enrichment pattern is similar to that for the entire extracellular proteome identified. Indeed, the binding pattern for the newly identified proteins clearly shows that the vast majority bound to the cobalt and manganese loaded IMAC columns. In fact, in combination with the copper column, which has a very different protein enrichment pattern, this cobalt and manganese IMAC columns enriched for all but 13 of the 116 (89%) newly identified proteins.

4.4.1: Analysis of Universal and Specific IMAC Enrichment

It is expected that soluble extracellular proteins identified here have adapted to exposure to the acid mine drainage, including adaption to the heavy metal rich environment. Insights into metal ligation chemistries or patterns may provide clues into ligation of the AMD metals. With the IMAC enrichment procedure, universally IMAC enriched proteins dominated the dataset at 61%, but unfortunately provide few clues for characterizing the metal chemistries of the AMD community. With many mechanisms of metal binding, it is difficult to derive the potential mechanisms of metal binding to these promiscuous proteins. For example, the trend in the number of proteins identified within the IMAC bound fraction of each of the metal columns (**Table 4.1**) does not follow the Irving-Williams series of the relative stabilities of complexes formed by metal ions, a trend of ionic radius, binding preferences as dictated by hard-soft acid base theory, nor does it follow with the abundance of metals within the AMD environment. Rather, as expected, it is simply indicative of significant metal binding based on the molecular exterior of that protein. Thus, the number of proteins that bind to each metal is likely dictated by a combination of the ratio of available hard and soft ligand sites within proteins and the radius of available metal binding sites. Similarly, the promiscuous

binding of proteins to multiple metals could be explained through several mechanisms, including numerous metal binding sites or high metal affinity residues within one protein; one metal site that binds only hard or soft metals; or potential denaturation of the protein to expose non-native surface residues.

Nearly all proteins with ribosomal structural roles were identified as universally IMAC enriched proteins. The identification of ribosomal subunits in the extracellular fraction is not uncommon and can be attributed to their pervasive abundance from unavoidable cellular lysis. The identification of ribosomal proteins as universally IMAC enriched is not surprising. Although there is no evidence of the involvement of metal ions in peptide bond formation, metals are believed to play an important structural role in rRNA folding and stabilization of the compact tertiary rRNA structures providing a basis for the observed metal affinity[111]. Magnesium is the only divalent metal known to be abundant in the ribosome, along with a lesser abundance of zinc; however, the bounty of phosphates in the ribosome may lead to nonspecific interactions with cationic metals.

Conversely, the remaining 190 (39%) IMAC specific enriched proteins were preferentially found in the bound and unbound fractions of a column. The specific enrichment of these proteins provides enhanced opportunities for further characterization. For example, eleven flagella proteins, the majority of flagellar proteins in the database (11/18, 61%), demonstrate a clear specificity towards non-metal binding, or show universal enrichment within the unbound fraction of the IMAC column (**Table 4.3**). It is interesting that flagellar proteins, designed to be exposed to the metal rich environment, have no metal affinity.

The IMAC enrichment data for each protein were clustered to reveal trends within the pattern of proteomic IMAC enrichment amongst the library of seven metals and enrichment for each individual protein (**Figure 4.4**). After clustering, the eight IMAC columns were divided into four distinct groups: 1.) the control with no metal; 2.) iron, the only trivalent metal; 3.) Co^{2+}, Mn^{2+}, Ni^{2+} ; and 4.) Cu^{2+}, Mg^{2+}, Zn^{2+}.

As an outlier from the binding trends of the other six metals, the cluster of enrichment results from the ferrous IMAC column showed that only 7 of the identified

Table 4.3: Identified Flagellar Proteins and Their Pattern of IMAC Enrichment
Proteins identified preferentially within the bound fraction of an IMAC column are indicated with a blue box and 1; proteins preferentially identified within the unbound fraction are shown with a yellow box and -1; proteins either unidentified within a column or those with no bound/unbound preference are shown in a grey box with a 0.

Cell Motility proteins	Co	Cu	Fe	Mg	Mn	Ni	Zn	control
Epl_15865_87_COG1681 Archaeal flagellins	0	0	0	1	1	0	-1	0
LII_11111_14_Putative flagellin	-1	-1	-1	-1	-1	-1	-1	0
LII_11111_17_Flagellar hook-associated protein (FlgL)	0	0	1	1	0	-1	0	0
LII_11111_21_Putative flagellin	-1	-1	-1	-1	-1	-1	-1	0
LII_11111_26_Flagellar basal body rod protein	-1	0	0	-1	-1	-1	-1	0
LII_11277_262_Probable flagellar hook protein FlgE	-1	-1	-1	-1	-1	-1	-1	0
LII_11277_263_Probable flagellar hook capping protein FlgD	-1	-1	-1	0	-1	-1	-1	0
LII_8241_208_Probable flagellar hook protein (FlgE)	-1	-1	-1	-1	-1	-1	-1	0
LII_8241_209_Probable flagellar hook capping protein (FlgD)	-1	0	-1	0	-1	-1	-1	0
LII_8241_641_Putative flagellar basal body rod protein	-1	0	0	-1	-1	-1	-1	0
LII_8241_645_Putative flagellin	-1	-1	-1	-1	-1	-1	-1	0
LII_8241_649_Flagellar hook-associated protein (FlgL)	0	0	0	1	0	-1	0	0
LII_8241_652_Putative flagellin	-1	-1	-1	-1	-1	-1	-1	0
LIII_8063_25_flagellin domain protein	-1	-1	-1	-1	-1	0	-1	0
LIII_8063_31_flagellin domain protein	0	-1	0	-1	-1	-1	-1	0
LIII_8063_36_flagellar basal-body rod protein FlgG	-1	1	0	0	0	0	-1	0
LIII_9612_10_flagellin domain protein	-1	-1	-1	-1	-1	-1	-1	0
Unasn_10454_1_COG1749 Flagellar hook protein FlgE	1	0	0	1	0	0	0	0
Unasn_4203_2_COG1344 Flagellin and related hook associated proteins	-1	-1	-1	-1	-1	0	-1	1

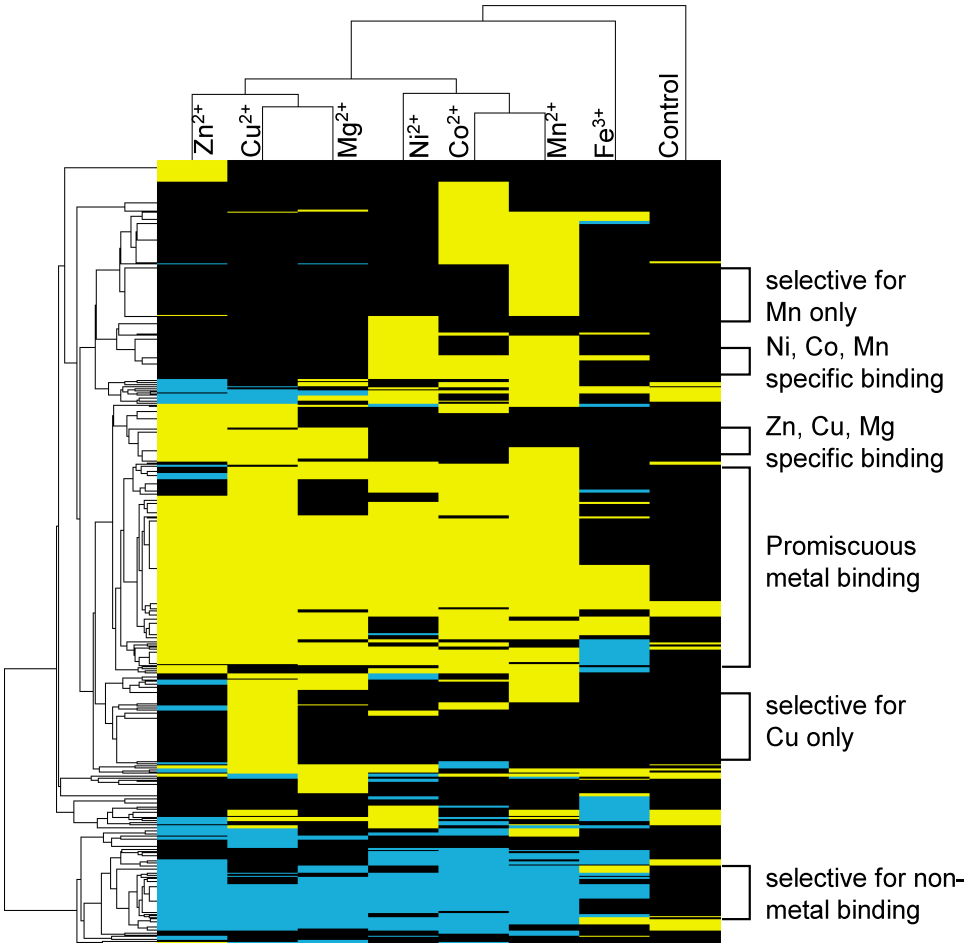

Figure 4.4: Heat Map of Proteins Found in Chromatographic Fractions
Metals are clustered from the left: Fe^{3+}, Co^{2+}, Mn^{2+}, Ni^{2+}, Cu^{2+}, Mg^{2+}, Zn^{2+}, Control. Proteins identified as metal-bound are shown in yellow, metal-unbound in blue, and unenriched in black. Metal binding specificity falls into six major clusters, identified to the right.

proteins had preferential binding for Fe^{3+}. Although iron is of great physiological importance to the AMD microbial community, the abundance of environmental iron is divalent. Immobilized Fe(III) is known to weakly interact with the carboxylic and phenolic groups and strongly bind phosphate groups, with chelation within a four-membered ring complex[112]. The majority of proteins selectively enriched by Fe(III) were proteins of unknown function; however, one is a putative type I cytochrome. Although interesting, many cytochromes have been previously shown to nonspecifically bind to iron through exposed protein surface interactions[98].

4.4.2: Functional Insights into Enriched Proteins

Of the proteins clustered that bind a specific metal or group of metals, each cluster has a wide range of cellular functions (**Figure 4.5**). The largest groups of specifically IMAC enriched proteins were involved in translation, ribosomal structure, and biogenesis, or post-translational modification, protein turnover, and chaperones. Again, a majority (54%) of enriched proteins identified in this study have no known function.

More minor trends were observed within other functional categories. Cell wall/ membrane/ envelope biogenesis proteins (COG function M), carbohydrate transport and metabolism (COG function G) and nucleotide transport and metabolism proteins (COG function F) were nearly universally enriched by IMAC columns loaded with divalent metals, but were not enrichment by ferric iron. The post-translational modification, protein turnover, chaperones (COG function O) was primarily enriched by manganese. Secondary metabolites biosynthesis, transport and catabolism proteins (COG function Q) were only enriched by cobalt and manganese. Finally, proteins involved in signal transduction mechanisms (COG function T) were largely enriched by manganese, along with nickel and cobalt.

A protein currently annotated as a TonB protein (*Leptospirillum* group II scaffold 11277_gene 177) stood out as a universally divalent-IMAC enriched protein, with specificity for each metal except Fe(III). TonB was not identified in either control

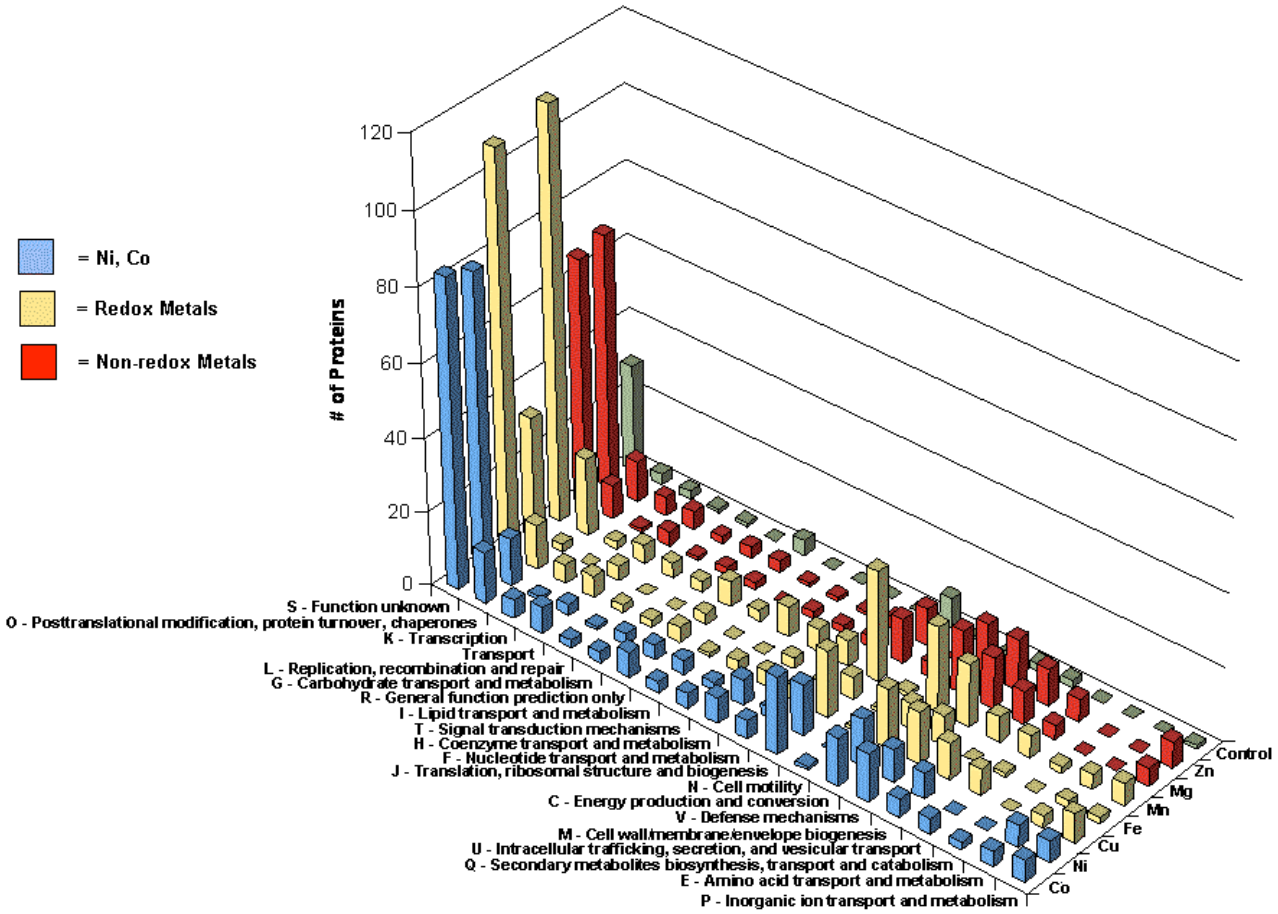

Figure 4.5: Distribution of Functions Among Reductive, Non-reductive and Ni/Co Metals

column's bound replicates, illustrating its marked specificity for enrichment in IMAC columns loaded with divalent metals. TonB is a moderately characterized protein that is a component of the TonB-dependent transport (TBDT) mechanism. This complex has previously been shown to be critical for the uptake of low abundance iron from the environment in gram-negative bacteria.[113] BLAST analysis[114] of this protein against the non-redundant database revealed a C-terminal region of the protein to show suitable sequence similarity (E-score < $1x10^{-8}$) to other known TonB proteins, while the remaining 60% of the sequence did not show significant similarity to known TonB proteins. Variation in TonB sequences is not uncommon, but in this instance the moderate amount of sequence conservation may imply that the identified *Lepto*II TonB protein is a highly diverged form, a probable adaption to the iron rich environment of the AMD. Additionally, recent characterizations have shown that TBDT also requires a nickel co-factor and that potential substrates of TBDT may include a number of metals including Cu, Co, and Ni .[115] Based upon the pattern of divalent IMAC column enrichment shown here, one may speculate that this critical metal transport protein has been ecologically tuned in *LeptoII* to increase affinity for essential, but less abundant metals in the AMD environment.

An alternative grouping of proteins in accordance with their affinity for reductive and non-reductive IMAC enrichment resulted in insights into the enriched proteins in the AMD microbial community (**Figure 4.5**). Although proteins of unknown function account for the majority of functional classifications in both categories, several categories were particularly highly represented among proteins exhibiting affinity for the reductive metals, such as: PTM, protein turnover, chaperones; translation, ribosomal structure, biogenesis; energy production; and defense mechanisms. Coenzyme transport and metabolism were also among the most striking functional categories to exhibit protein specificity.

Grouping IMAC enriched proteins by oxidation-reduction activity enables possible insights into the role of redox activity in the proteome of the AMD community. As outlined by Mounicou et. al.[101], copper, iron, and manganese have biological functions that are predominantly redox active, while magnesium and zinc serve non-redox roles.

Cobalt and nickel are not considered in this analysis because of their ability to serve in either category. We find that, overall, IMAC columns loaded with reductive metals (Cu^{2+}, Fe^{3+}, Mn^{2+}) enriched for more proteins than IMAC columns loaded with non-reductive metals (Mg^{2+}, Zn^{2+}). As mentioned earlier, the Fe^{3+} IMAC column remained unique in its relatively low number of specifically IMAC enriched proteins. In general, among the three reductive metals, the number of proteins exhibiting specificity remained relatively similar. Additionally, proteins exhibiting specific enrichment on non-reductive metal loaded IMAC columns also remained relatively constant, but of a smaller quantity than the proteins specifically enriched by reductive metal loaded IMAC columns. A slight bias in population is observed between the groupings (ie., redox vs non-redox) but within the groupings the abundances of the representative functions appears consistent. This property could allow for a targeted enrichment of a specific functional grouping by utilizing either a redox or non-redox active metal.

4.4.3: Insights into Enriched Proteins of Unknown Function

One theme across studies of the AMD community proteome is the large number of proteins of unknown function[19, 68, 105, 116, 117]. Based on their abundance and pervasive expression, these functionally unknown proteins are predicted to be critical for microbial survival[68, 116, 117]. The potential for conserved IMAC enrichment among these proteins highlights the effectiveness of the IMAC and MS analysis in identifying key target proteins from the thousands present in the community. The combination of highly specific, experimental IMAC enrichment events and Pfam analysis of these proteins of unknown function provides additional evidence for the role of these proteins within the microbial community.

Forty-five proteins that exhibited specific metal binding and had an unknown function were submitted to Pfam for domain and motif prediction. Among the domains represented within these proteins, several show interesting IMAC enrichment and metal binding properties. A 139 residue portion of a protein enriched by the Co column (located on contig 12989, gene 6 from an undetermined species within the AMD community) resulted in a high scoring ($7x10^{-32}$) match to the COXG Pfam ID (**Table**

4.4). The COXG domain has been identified in hundreds of chemolithoautotrophic microbes and is integral member of the hydrolysis of CO.[118] The structure of the COXG domain is related to the protein family ArsR.[119] This family contains over six hundred Pfam sequences and is frequently found in microbial proteins functioning as metallosensitve transcriptional repressors.[120] Another enriched protein from the Ni column resulted in a predicted high scoring ($3.2x10^{-57}$) domain with the Pfam ID thermopsin (unassigned protein from contig 436, gene 3). The thermopsin domain represents a family of acid and temperature stable proteases. Previous assays determined thermopsin was optimally active between 25 - 78ºC and pH 2, similar to conditions found with the AMD system.[121] The proteins identified to contain the thermopsin domain are generally large, oligomeric proteins that are known to contain metal co-factors.[122] The high scoring domains identified in these two highlighted proteins support the notion that these unknown proteins may function as critical degradative enzymes. The representative COG category (protein turnover, O) was among the highest represented categories of all the identified IMAC enriched proteins. Finally, among the high scoring Pfam identifications, eight of the proteins are novel identifications. The COXG containing protein described above had not been identified in previous analyses of the extracellular fraction. This protein is an excellent example of the methodology and provides a specific target for future experiments considering the novel protein identification, specificity to the Co-IMAC column and high scoring COXG domain (implying potential as a metallosensitive transcriptional repressor). The ability to identify specific targets that may play critical metal related roles from a suite of tens of thousands of proteins further illustrates the applicability of the described method and necessity for continued advancement towards a complete proteomic characterization.

4.5: Conclusions

The combination of a library of IMAC columns and bottom up LC-MS/MS analyses has resulted in identification of hundreds of metals and their pattern of enrichment across seven biologically active metals. Direct detection provided characterization of IMAC enrichment for 485 proteins, including 116 newly identified

Table 4.4: Protein Identifications from Enriched and Novel Extracellular Proteins

A subset list of high scoring Pfam domain / motif analysis of IMAC enriched proteins that exhibited specific binding to one IMAC column, but currently do not have a known function.

Protein Name	IMAC Column	Alignment Start	Alignment End	E-value	Pfam Acc. #	Pfam ID
5wayCG_Leptoll_Cont_10961_GENE_10	Mn	92	453	1.20E-104	PB004231	Pfam-B_4231
*UBA_Leptolll_Cont_9205_GENE_68	Cu	55	453	7.20E-94	PB004231	Pfam-B_4231
*UBA_Leptolll_Cont_9545_GENE_21	Cu	56	454	8.80E-94	PB004231	Pfam-B_4231
5wayCG_Unassgn_Cont_436_GENE_3	Ni	1	255	3.20E-57	PF05317.4	Thermopsin
*AMD_Gpl_Cont_12267_GENE_89	Mn	4	119	6.50E-43	PF01981.9	PTH2
*UNL_Unassgn_Cont_12989_GENE_6	Co	5	144	7.10E-32	PF06240.6	COXG
UBA_Leptoll_Scaffold_8241_GENE_12	Cu	1	64	1.70E-30	PF04384.6	DUF528
*5wayCG_Leptoll_Cont_11195_GENE_17	Cu	6	94	6.50E-29	PF01910.10	DUF77
UBA_Leptoll_Scaffold_8241_GENE_689	Cu	6	94	6.50E-29	PF01910.10	DUF77
*5wayCG_Leptoll_Cont_11238_GENE_107	Cu	90	155	2.80E-12	PF08308.4	PEGA
*UBA_Leptoll_Scaffold_7931_GENE_119a	Cu	90	155	2.80E-12	PF08308.4	PEGA
*UNL_Unassgn_Cont_3050_GENE_1	Cu	2	65	1.80E-06	PF04264.6	Ycel
5wayCG_Leptoll_Cont_10961_GENE_49	Cu	64	93	0.0002	PF08238.5	Sel1
UNL_Apl_Cont_17472_GENE_13	Ni	17	131	0.0027	PF07992.7	Pyr_redox_2

*Indicates newly identified protein as a result of IMAC enrichment

proteins that expand the depth of extracellular proteomic coverage by greater than 20%. The enriched proteins represented the abundant microbial members from the mine and are predicted to perform a variety of cellular functions. IMAC enrichment did not show any discernable bias, including the amino acid composition of the proteins. Among the identified proteins, three classifications of IMAC enrichment were identified: universal IMAC enrichment, specific IMAC enrichment, and no metal binding. The majority of proteins were enriched by the Co, Cu and Mn columns and many of the newly identified proteins were enriched here based upon binding to these columns. The prevalence of unknown proteins was expected and apparent. In order to identify potential targets for future studies, predicted domains and motifs were identified and resulted in numerous high scoring regions indicating potential functions. This methodology highlights the use of integrated technologies and has provided a specific target list of proteins, extracted from a metaproteome of tens of thousands, that my play critical roles in metal related activities that is necessary for the survival of the extremophilic microbes that thrive in the acid mine.

Chapter 5

High Resolution Mass Spectrometry for the Characterization of a Novel, Growth Stage Dependent Cytochrome

Portions of included text are adapted from:

Steven W Singer, Brian K Erickson, Nathan C VerBerkmoes, Mona Hwang, Manesh B Shah, Robert L Hettich, Jillian F Banfield, Michael P Thelen. "Posttranslational Modification and Sequence Variation of Redox-active Proteins Correlate with Biofilm Life Cycle in Natural Microbial Communities.", ISME Journal, 2010, May, Epub ahead of print.

Brian K. Erickson's contributions include sample preparation, experimental FT-ICR data collection, data analysis, and authorship.

5.1: Introduction

Proteomic measurements of an early developmental stage biofilm identified two atypical cytochromes expressed at high levels. These were initially identified as *LeptoII* proteins of unknown function, and later were determined to be the membrane Cytochrome 572 (Cyt_{572}) and periplasmic Cytochrome 579 (Cyt_{579}).[19] Both cytochromes were purified from a mixed developmental-stage biofilm and characterized biochemically.[76, 116] Cyt_{572}, localized to the outer membrane, is a 57-kDa multimeric protein that oxidizes Fe(II) at low pH, and thus is likely the Fe(II) oxidase for LeptoII in the biofilm.[116] Inspection of metagenomic data sets showed six sequences corresponding to *LeptoII* variants of Cyt572 in the Ultra Back A (UBA) and 5way Community Genomics (CG) databases (5way CG and UBA databases refer to environmental genomic databases obtained from DNA isolated from biofilms collected at distinct sites in the Richmond Mine. 5way CG was sampled at the 5-way

convergence of streams in June 2004, and UBA was sampled in the upper A drift in June 2005). Genomic sequences were assembled and these were deposited in the databases of these names, indicating that multiple variants of the cytochrome may be expressed in the same biofilm sample.[20, 69] Cyt_{579} was characterized as a 16-kDa monomeric protein localized to the periplasm of *LeptoII*.[76] Cyt_{579} was isolated as a mixture of polypeptides with different N-terminal cleavage sites. Redox reactions with Fe(II) demonstrated pH-dependent Fe(II) oxidation that was inconsistent with its assignment as the Fe(II) oxidase for *LeptoII* and suggested an electron transfer function. Our working model positions Cyt572 as the Fe(II) oxidase on the outer membrane of *LeptoII* cells, which oxidizes Fe(II) to Fe(III) and transfers electrons to Cyt_{579}. This scheme is analogous to the proposed role of an outer membrane bound c-type cytochrome, Cyc2, and rusticyanin, a periplasmic Cu protein, in *Acidithiobacillus ferrooxidans*, an acidophilic Fe(II)-oxidizing bacterium found in AMD environments.[123] Recent analysis of multiple AMD biofilm proteomes from different developmental stages has shown that the community switches from rapid Fe(II)-dependent autotrophic growth in early developmental stages to partitioning of fixed carbon to heterotrophs in late developmental stages.[124] The effects of aging on biofilm morphology, microbial community composition, and protein expression led us to speculate that electron-transfer proteins, critical for Fe(II) oxidation, could change physically over the biofilm life cycle, either from specific posttranslational modifications or from genetic variation as a result of mutation or recombination. Characterization of cytochromes by high resolution intact protein mass spectrometry from biofilms representing early and late developmental stages (DS1 and DS2, respectively) showed that both posttranslational modifications and expressed sequence variants are correlated with biofilm development.

5.2: Methodology

Both Cyt_{579} and Cyt_{572} were purified from the biofilms as described previously[116, 117] and stored at 4°C. No change in the redox properties of samples of either protein was observed after 6 months at 4°C, and minimal degradation was observed by SDS-

PAGE. In all cases, visible spectra for both oxidized and reduced samples were identical to our previously published spectra for Cyt_{579} and Cyt572. Enrichment of c-type cytochromes was achieved by fractionation of the extracellular fraction of the C75m sample from November 2006. The 95% $(NH_4)_2SO_4$ precipitate of the acid wash fraction was dialyzed for 16h against 20mM H_2SO_4/100mM $(NH_4)_2SO_4$ pH 2.2. The dialysate was loaded onto a SP-Sepharose (GE Healthcare, Piscataway, NJ, USA) fast flow column and washed with this pH 2.2 buffer. Cyt_{579} was eluted by a step increase to pH 5.0 in 100mM NaOAc, and the remaining proteins were eluted with a 0–2M NaCl gradient in the same buffer. Characteristic visible absorption spectra for c-type cytochromes were observed throughout the 1.2–2.0M fractions, with 1.4 and 1.5M fractions containing the highest concentrations of the c-type cytochrome as measured by visible absorbance of the a-band at 552nm for samples reduced with sodium ascorbate. The individual heme bands were visualized by separation of proteins on 15% SDS-PAGE by the method of Francis and Becker (1984). The stained bands were excised from the gel and digested with trypsin.[125]

5.2.1: Cyt_{579} Intact Mass Measurement

Purified samples of Cyt_{579} were further prepared for characterization of the intact proteins by high resolution top-down MS. Cyt_{579}-enriched samples were desalted with Zip-Tip (C4, Millipore, Billerica, MA, USA) pipette tips and eluted with 100% ACN (0.1% acetic acid, v/v). The protein fraction was then diluted into 50/50/0.1 (v/v/v) H_2O/ACN/acetic acid and infused into the Micromass Z-Spray source attached to a Varian (Lake Forest, CA, USA) 9.4-T (Cryomagnetics Inc., Oak Ridge, TN, USA) HiRes electrospray FT-ICR (Fourier transform ion cyclotron resonance mass spectrometer or an electrospray source coupled to the LTQ-Orbitrap-XL (Thermo Fisher Scientific, San Jose, CA, USA). MS fragmentation was achieved through collisionally activated dissociation, electron-transfer dissociation or infrared multiphoton dissociation. Parent charge states of Cyt_{579} were manually selected, isolated and fragmented (collisionally activated dissociation or electron-transfer dissociation) in the ion trap before high-resolution mass measurement in the Orbitrap. For infrared multiphoton dissociation on

FT-ICR, parent charge states of Cyt_{579} were manually selected and isolated in the selection quadrapole before mass analysis in the FTMS analyzer cell. A 350-nM spike of ubiquitin was introduced into the C-drift DS1 and AB-Muck DS1 samples for internal mass calibration. M/z values were manually extracted from spectra, deconvoluted and plotted with Origin 8 (OriginLab, Northhampton, MA, USA).

5.3: Results

5.3.1: Intact Protein Characterization of Cyt_{579}

High-resolution MS measurement of the purified protein from the C-drift DS1 sample revealed two molecular species by distinct isotopic distribution packets. Identification of the most abundant isotopic masses (MAIM) from each distribution corresponded to molecular species of 16,131.541 and 16,119.562 Da. (**Figure 5.1a**; external calibration). The measured mass of 16,131.541 Da corresponds to gene UBA_8062_372_S98A (Cyt_{579}) with a predicted signal peptide cleaved at the N-terminus resulting in the final sequence: (N-AELDILKP…). This observation was confirmed by PCR amplification and sequence analysis of the gene encoding for Cyt_{579} from DNA recovered from the C-drift DS1 sample. Twenty-nine clones were obtained after transformation of the PCR amplicon, all of which had an identical sequence to UBA_8062_372_S98A. The mature isoform of Cyt_{579} corresponds to the same sequence but lacking the seven C-terminal amino acids (…GNLKPE) and ending in (…FLNTAAK); this was the dominant variant expressed in the previously characterized Cyt_{579} preparations from C-drift biofilm samples [126]. The second most abundant distribution, 16,119.562 Da, was inferred to be a modified form of Cyt_{579}. In order to obtain the most accurate mass measurement possible, an internal calibration utilizing ubiquitin as a standard was used and resulted in a measured mass of 16,119.549 Da (**Figure 5.1b**). Collisional dissociation of this molecular ion resulted in 16 fragment ions, all corresponding to the sequence of Cyt_{579}, including abundant fragment ions corresponding to a sequence tag of MFWVVA, which is unique to Cyt_{579}. This exact sequence tag was also confirmed by

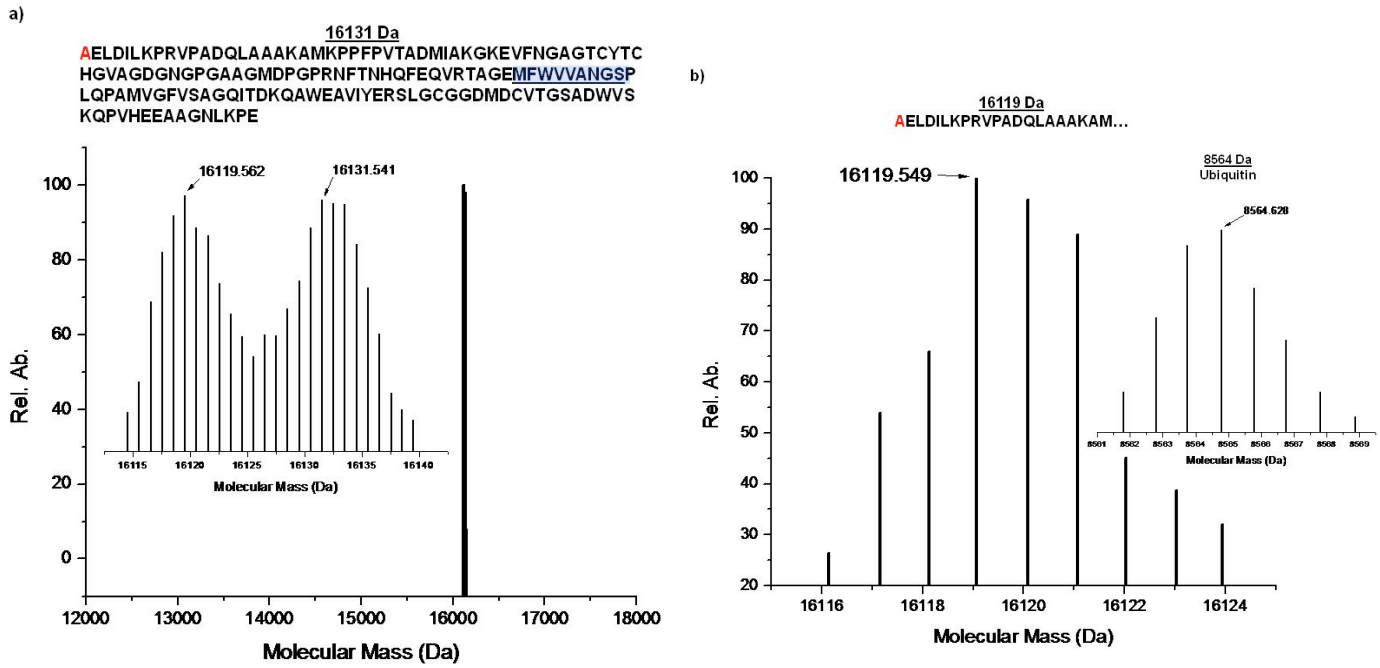

Figure 5.1: MS Spectra of C-drift Cyt_{579}

N-terminal residues are highlighted in red. a) The full length sequence of Cyt_{579} corresponding to mass 16,131 Da. Intact mass measurement of C-drift DS1 resulted in the identification of two abundant mass distributions, including the truncated product at mass 16,119 Da as described in the text. Residues highlighted in blue indicate the unique sequence tag. b) Partial sequence displaying truncated n-terminus and accurate mass measurement of C-drift DS1 following mass calibration utilizing an internally spiked ubiquitin standard (inset).

ETD measurements. Additionally, IRMPD fragmentation resulted in an expanded sequence tag of FWVVANGS (**Figure 5.2**), confirming the CAD and ETD results. The mass errors of the predicted versus measured fragment ions corresponding to the sequence tag were each less than 10 ppm, providing significant confidence in the identification of the 16,119 Da species as a modified form of Cyt_{579}. Previous bottom-up peptide measurements of this modified protein verified the sequence of the expected Cyt_{579} protein, plus the presence of an oxidation of residue Met-21. While the sum total of the fragment ions and peptide mass spectra uniquely support the assignment of the 16,119 Da species as a Cyt_{579}, neither set provided complete sequence coverage and thus it was impossible to unambiguously determine the modified form of this truncated version. However, based on all the information, the most likely assignment of the 16,119 Da species is modification of Cyt_{579} by oxidation (likely at Met-21), accompanied by loss of CO from the intact protein (CO loss from intact proteins is not unusual). We searched extensively for the location of the CO loss, but were unable to pin it down in the fragmentation or peptide data. However, the calculated mass for a Cyt_{579} protein with these specific modifications is 16,119.530 Da, in excellent agreement with the measured value (less than a 2 ppm mass error). For C-drift DS2 sample, measured masses correspond to additional N-terminal truncations, resulting in masses of 14,574 and 14,319 Da (N-AELDILKPRV… and N-AKAMKPPFPV…, respectively; **Figure 5.3**). These masses were previously observed in mixed developmental stage C-drift samples and also are derived from UBA_8062_372_S98A, with an identical cleavage at the C-terminus as the Cyt_{579} isoform characterized in C-drift DS1.

5.3.2: Cyt_{579} Isolated from AB-Muck Developmental Stages

To establish the generality of the developmental stage-dependent alterations of Cyt_{579}, the protein was purified from the AB-Muck DS1 samples and DS2 samples. As with the C-drift samples, substantially more Cyt_{579} was extracted from the early development stage sample. MS analysis of intact protein from the AB-Muck DS1 sample revealed a molecular ion identical to that identified from the C-drift DS1 samples

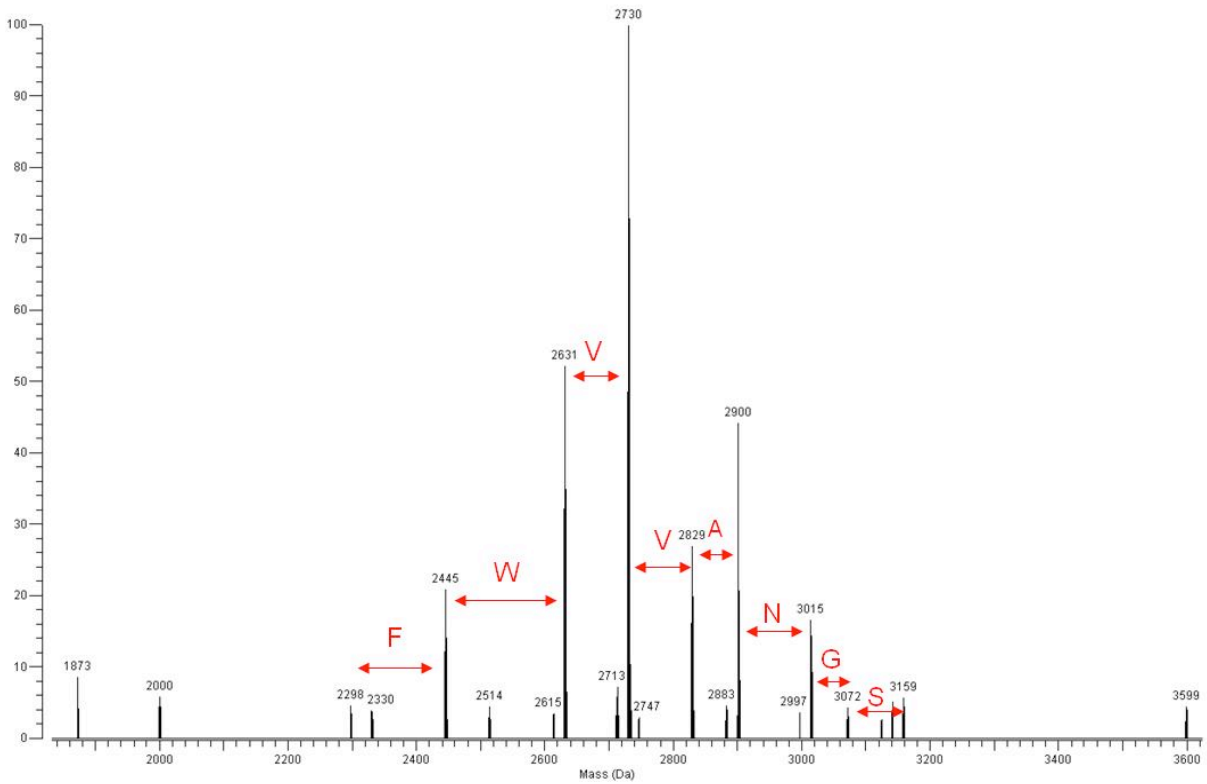

Figure 5.2: Deconvoluted IRMPD Fragmentation Spectrum of Cyt_{579} -16,119 Da Species

The Cyt_{579} unique sequence tag FWVVANGS is displayed in red.

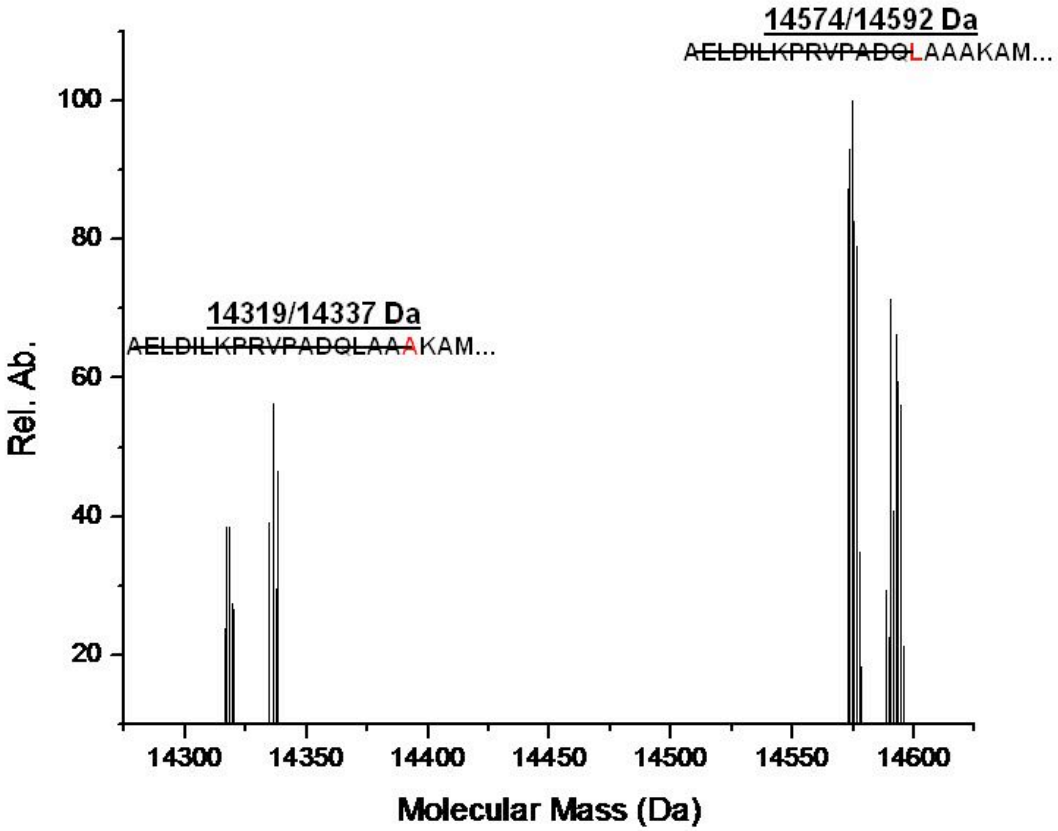

Figure 5.3: Intact Mass Measurement of C-drift DS2 Exhibiting Two States of Additional N-terminal Truncation
The doublets may indicate (-H_2O) loss and were confirmed to be Cyt_{579} through MS fragmentation. Strikethrough highlights the cleaved sequence.

(16,119.540 Da), indicating the N-terminus and Cyt_{579} variant were the same for both (**Figure 5.4a**). Internal mass calibration of the ABM DS1 sample resulted in a 0.022 Da difference between the ABM DS1 and C-drift DS1 Cyt_{579}. The AB-Muck DS2 sample had molecular ions corresponding to the LAAA N-terminus previously observed for C-drift DS2 sample, but lacked the ion distribution corresponding to the AKA N-terminus (Figure 2b). Redox experiments in the presence of 30 mM Fe(II) were very similar to the results obtained for the C-drift samples, and redox titrations at pH 4.3 indicated that the DS1 Cyt_{579} had a midpoint potential of 600 mV whereas the DS 2 preparation had a midpoint potential of 450 mV.

5.3.3: Cyt_{579} from C75m Site

Although the biofilms collected from the C75m site remained fairly constant in thickness across sampling times, the pH was lower (0.70) for the June than November 2006 sample, at which time the pH was 1.0. In addition, archaea were significantly more abundant than in the biofilm growing at this site in June than in November. Cyt_{579} was extracted from both C75m biofilms, which was more typical in morphology and microbial community composition to the early development stage biofilms from C-drift and AB-Muck sites described above. The yield of Cyt_{579} from the November C75m sample was similar to the yields obtained from the other early growth stage biofilms; however, the yield from the June sample was dramatically lower (75 times less Cyt_{579} was extracted from the June biofilm). This result is consistent with an ultra-structural study of the same biofilm [127] in which immunohistochemical detection of Cyt_{579} demonstrated that Cyt_{579} expression was localized only at the biofilm-water interface, and that a majority of the *LeptoII* cells did not express Cyt_{579}. Intact protein MS characterization of Cyt_{579} indicated that the proteins purified from the June and November samples had identical masses of 15,690 Da, which is the mass of the 8062_372 S98A Cyt_{579} with an ILKP N-terminus, identical to the isoform observed in the previous study on Cyt_{579} from the biofilms (**Figure 5.5**). Although insufficient Cyt_{579} was recovered from the June C75m sample for redox analysis, a midpoint potential of 590 mV was determined for Cyt_{579} from the November C75m sample.

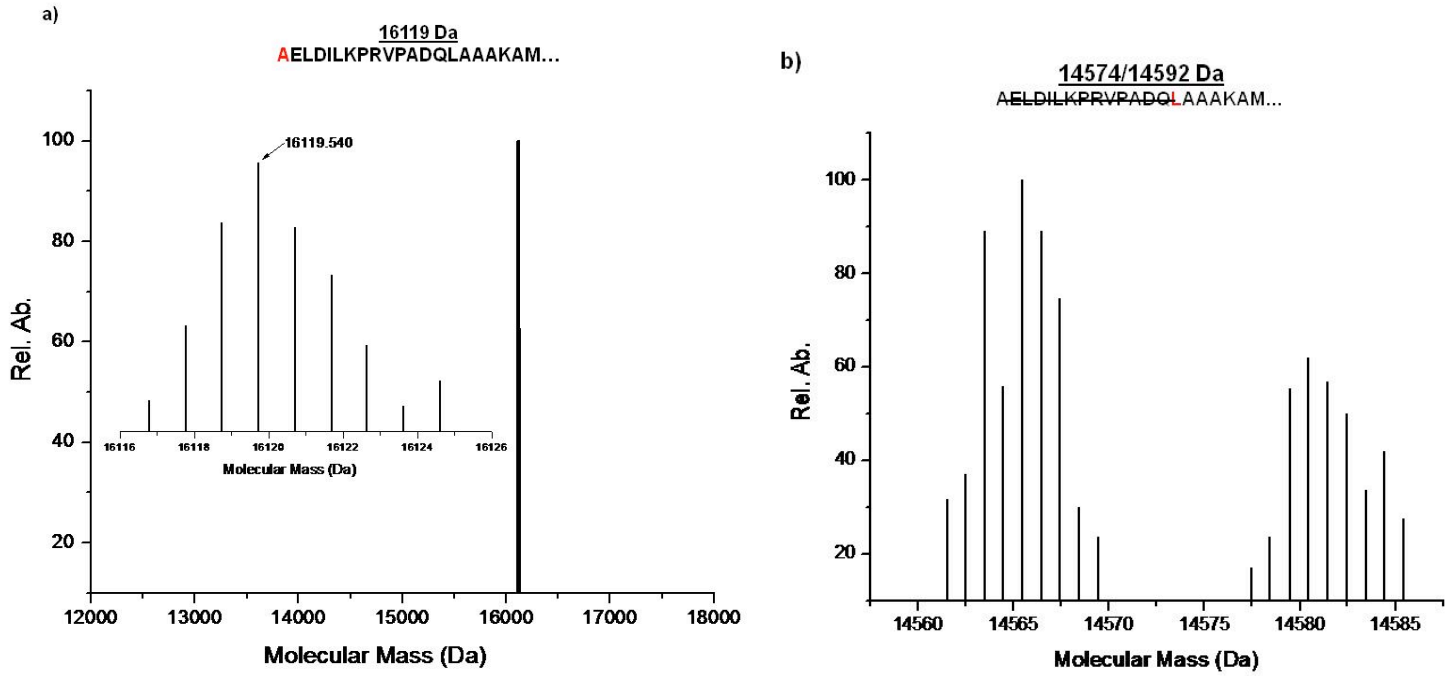

Figure 5.4: MS Spectra of AB-Muck.
N-terminal residues are highlighted in red. a) MS measurement of ABM DS1 exhibiting the same N-terminal truncation as C-drift DS1. b) Intact MS measurement of ABM DS2 resulting in conformation of the LAA N-terminal truncation.

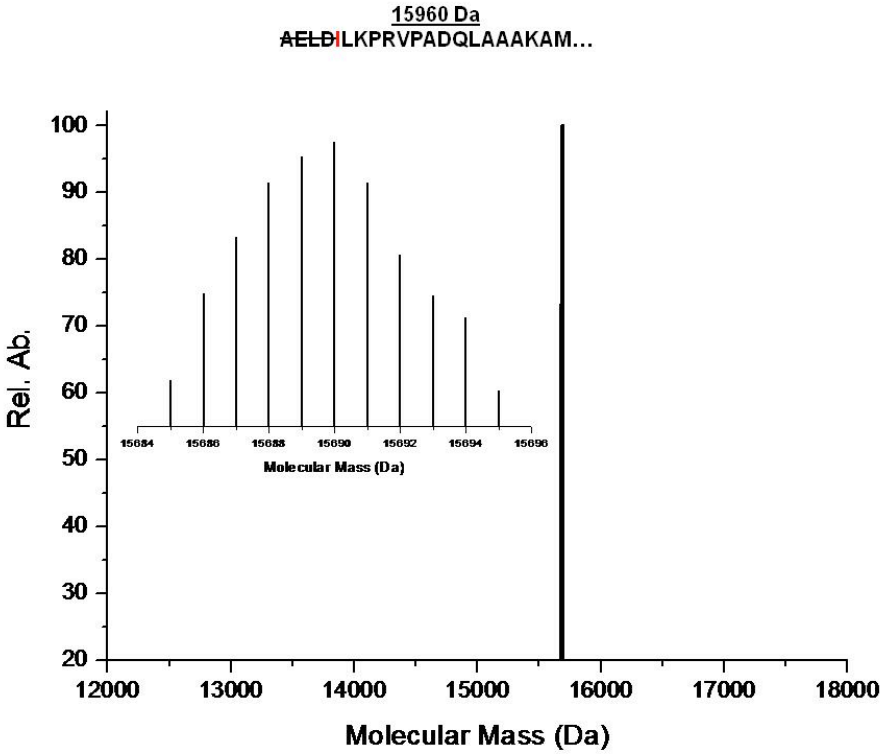

Figure 5.5: MS Measurement of C75m
Cyt_{579} resulted in an accurate mass of 15,960 Da corresponding to a ILKP N-terminus. The N-terminal residue is highlighted in red.

5.4: Conclusions

Previous studies with single species biofilms have described changes in protein type and abundance as these biofilms mature.[128, 129] Our work is unique in that it shows posttranslational modification, sequence variation, and truncation all play an important role in individual proteins in the life cycle of a natural, multispecies biofilm. This study shows that combining both high-throughput proteomics measurements and targeted biochemical studies can identify highly expressed proteins in natural microbial communities that may be sensitive to changes in the environment or species composition. These observations are critical to link biochemical pathways to the functioning of natural microbial communities.

Chapter 6

The Design and Implementation of Software for Integrating Bottom up and Top Down MS Datasets for the Characterization of an Extracellular Fraction from a Natural Microbial Community

Portions of included text are adapted from:

Vilmos Kertesz, Heather M. Connelly, Brian K. Erickson, Robert L. Hettich, "PTMSearchPlus: Software Tool for Automated Protein Identification and Post-Translational Modification Characterization by Integrating Accurate Intact Protein Mass and Bottom-Up Mass Spectrometric Data Searches", Analytical Chemistry, 2009, 81 (20), 8387-8395

Brian K. Erickson's contributions include computational design and implementation, experimental preparation of samples, experimental LC-MS/MS analysis, data analysis, and authorship.

6.1: Introduction

Various mass spectrometric approaches are available for characterizing complex protein mixtures by either interrogating the intact proteins (using accurate intact protein mass (AIPM) or top-down approaches) or their constitutive proteolytic peptides (termed "bottom up" (BU)).[49] While the BU approach is more well-developed and widely represented, each of these methods features a unique set of strengths and weaknesses. Clearly, the comprehensive characterization of complex proteomes will require further development in each method.

Top-down mass spectrometry for intact protein characterization was first introduced with electrospray-Fourier transform ion cyclotron resonance mass spectrometry (ESI-FTICR-MS).[130-132] The dynamic range, sensitivity, and mass accuracy achievable by high performance FTICR-MS affords not only high resolution

protein identification in most cases, but also detailed information about the molecular state of intact proteins. This high resolution measurement can reveal protein details that include post-translational modifications (PTMs), truncations, mutations, signal peptides, and isoforms due to the ability to accurately measure covalent modifications that alter the molecular mass.[16, 133] While intact protein measurement methodologies provide a powerful analytical approach, there are some remaining challenges for this approach. For example, on-line chromatography of intact proteins is often difficult due to the wide range of protein sizes and hydrophobicities, intact proteins often do not yield extensive fragmentation information, and the resulting data is often difficult to analyze and to interpret due to limited bioinformatics tools.[130]

The more common peptide or BU mass spectrometric approach to identify proteins and their modifications involves enzymatic digestion of proteins with a protease such as trypsin, Glu-C, or cyanogen bromide to generate a peptide mixture. This peptide mixture is then analyzed by MS/MS methods to generate peptide fragmentation spectra that are compared to theoretical spectra of possible peptide candidates from a database using different searching algorithms.[134] This “shotgun” proteomics approach is able to efficiently provide a comprehensive list of proteins present even in a large multi-protein complex. However, vital information about the molecular nature of the protein may be missed if the peptides containing particular modifications or variations escape detection. Furthermore, identifying peptides that come from a complex protein mixture may not provide information to distinguish between isoforms of the same protein.

One obvious solution to a more comprehensive characterization of complex protein mixtures would involve an integrated intact protein and proteolytic peptide characterization approach, which would exploit the unique strengths of each method. In such an integrated approach, the information from the comprehensive list of proteins identified by their intact molecular mass can be compared against information from the comprehensive list of proteolytic peptides corresponding to the same protein, thus revealing detailed information about modified protein isoforms. The correlation between the two methods can provide detailed PTM location and identity, and may be more

generically applicable than fragmentation information from the intact proteins. It is important to realize that while accurate molecular masses can be measured for most intact proteins (provided they are within the accessible molecular range of the mass spectrometer employed), the quality of the tandem mass spectra from intact proteins varies greatly and in some cases is not sufficient to provide much detailed information. We were one of the first groups to demonstrate integrated intact protein and proteolytic peptide measurement approach for the characterization of the *Shewanella oneidensis* proteome[135], and have extended this for the 70S ribosomal complex from *Rhodopseudomonas palustris*.[136] For these studies, most of the integrated datasets were interrogated manually. Integrated intact protein and proteolytic peptide approaches have seen increased development in the last several years, focusing on both experimental[137-141] and computational aspects, but range greatly in their ability to handle high resolution datasets and how the scoring is conducted.

While there are a variety of software searching tools for BU data analysis (i.e., SEQUEST[25], Mascot[142], X!Tandem[143]), there are relatively few tools for top-down and AIPM analyses. The current software standard for top-down work is ProSightPTM (*commercially available from ThermoFisher Scientific Corporation as ProSightPC*), which combines a number of search engines and a browser environment into a web application that allows the user to analyze AIPM and corresponding protein fragmentation data.[144] This program uses the masses of intact proteins and the tandem mass spectrometry information (i.e. product ion masses) of the same proteins to provide protein and PTM identifications. This software relies on the use of top-down dissociation methods that are often not as comprehensive for complex mixtures as BU methods employing an enzymatic digestion. Frequently employed methods to generate intact protein fragments include collision-induced dissociation (CID), infrared multiphoton dissociation (IRMPD), electron capture dissociation (ECD)[145], or electron transfer dissociation (ETD). PROCLAME is another top-down software analysis tool that uses intact protein mass measurements to determine sets of putative protein cleavage and modification events to account for the measured protein masses observed.[146] PROCLAME provides a reasonable prediction algorithm, but is unable to

incorporate tandem mass spectrometry (MS/MS) data within the process. More recently, BIG Mascot was introduced and operates in a similar approach as ProsightPTM, utilizing intact protein mass and corresponding product ion masses generated from intact protein dissociation.[147] While there is some progress in the demonstration of computational software to integrate AIPM and BU datasets, as listed above, this active field is still very much under development.

In this report, we describe a new software algorithm, *PTMSearchPlus,* which provides a comprehensive search method to enable the integration of AIPM identification with the BU generated peptide data to faster and more confidently identify proteins and their associated PTMs. The software can perform independent AIPM or BU searches, as well as integrate both approaches. By combining these two search capabilities, the results from the AIPM search can be used to limit the number of the proteins that are used to generate the peptide database for the BU search ("AIPM predicted" search) and in return, the results of the BU search can be used as confirmation for the proteins with associated PTMs found in the AIPM search. The limitation of the database used in the BU search based on the results of the AIPM search may reduce the search time dramatically, allowing the user to search for more PTMs on proteins and peptides within a reasonable time frame. The power of this integrated search method is demonstrated using data from analysis of a protein standard mixture and a complex *Escherichia coli* ribosomal protein mixture, and finally extended to an extracellular AMD sample. In addition to the integration approach, we also present a novel way to reduce the number of peptide candidates in a BU search when multiple PTMs are probed. The method allows the user to limit the number of possible PTMs on a peptide based on chemical considerations that may result in a significant decrease in the number of peptide candidates. Dramatic increases in search throughput with this method are demonstrated using data from a complex *Escherichia coli* protein mixture database.

6.2: Materials and Methods

6.2.1: LC-FTICR-MS for AIPM Mass Spectrometry

All capillary HPLC-FTICR-MS experiments were conducted with an Eksigent NanoLC-2D HPLC interfaced directly to a Micromass Z-Spray source on a Varian (Lake Forest, CA) 9.4-Tesla (Cryomagnetics Inc., Oak Ridge, TN) HiRes electrospray Fourier transform ion cyclotron resonance mass spectrometer.[148] A C4 reverse-phase column (Phenomenex Jupiter, 300Å with 5μm particles) was packed via a pressure cell in-house and was employed for all intact protein separations.

The ribosomal purification eluent or extracellular fraction consisting of 5-20 μg of total protein was injected onto the column and eluted at 2.5 μl/min into the electrospray ion source of the FTICR-MS. The gradient was run from 90% solvent A (95/5/0.1 (v/v/v) H_2O/ACN/formic acid) to 100% solvent B (95/5/0.1 (v/v/v) ACN/H_2O/formic acid) over a 60-min linear gradient. Calibration of the mass spectrometer was accomplished externally using a ubiquitin solution resulting in a mass accuracy of ±3-10 ppm and resolution of 50,000-160,000 (FWHM).

6.2.2: 1D LC-MS-MS for BU mass spectrometry.

For all peptide samples, one-dimensional (1D) LC-MS-MS experiments were performed with a Famos/Switchos/Ultimate HPLC System (Dionex, Sunnyvale, CA) coupled to an LTQ quadrupole ion trap mass spectrometer (Thermo Finnigan, San Jose, CA) equipped with a nanospray source, as previously described.[42] A 160-minute linear gradient from 100% solvent A (95% H_2O/5% ACN/0.1% formic acid) to 100% solvent B (30% H_2O/ 70% ACN/0.1% formic acid) was employed. For all 1D LC-MS-MS data acquisition, the LTQ was operated in the data dependent mode with dynamic exclusion enabled (repeat count 2), where the five most abundant peaks in every MS scan were subjected to MS-MS analysis. Data dependent LC-MS-MS was performed over a parent *m/z* range of 400-2000.

6.2.3: Software

PTMSearchPlus was developed using Delphi 3 computer language (Borland Software Corp., Scotts Valley, CA) under Microsoft© Windows XP Home Edition (Microsoft Corp., Redmond, WA) operating system and can be run in any 32-bit Windows environment with at least 256 MB RAM. Currently, the program is freely available upon request to any government or educational institute.

6.3 Results

PTMSearchPlus currently supports the following search options:

- a standalone AIPM search.
- a standalone BU search using the MyriMatch[26] scoring algorithm.
- an integrated AIPM and MyriMatch-based BU search.

These search options are discussed briefly below.

6.3.1: Standalone AIPM Search

Deconvoluted isotopic peak envelopes from FTICR-MS measurements were matched against calculated isotopic peak envelopes of modified and non-modified proteins from a database, which contains FASTA formatted protein sequences. A match was judged on the mass difference of the most abundant peaks of the experimental and calculated isotopic envelopes. In general, a maximum difference of 50 mDa was declared as a match in the searches.

6.3.2: Standalone BU Search

In this mode, the software used the MyriMatch[26] scoring algorithm to compare modified and non-modified peptides of a given protein database against BU mass spectra information stored in MS2 files.[149] Peptides with scores above a certain limit were assigned as a match and used in calculation of protein coverage's.

6.3.3: Integrated AIPM and BU Search

The hollow arrows in **Figure 6.1** illustrate the most straightforward approach to integrate AIPM and BU searching algorithms in general. In this case, AIPM and BU data were searched independently using the specified full PTM database, and the results were then compared and combined. This approach was considered to be a "complete" search, as all proteins (and their possible PTMs) were checked against the two different -AIPM and BU- datasets.

The filled arrow in **Figure 6.1** represents a different approach that was also implemented in *PTMSearchPlus* to limit search space for the BU search. The search space limitation was based on the proteins and PTMs found in the AIPM search. Using this approach, an AIPM search was conducted first, followed by assigning the union of possible PTMs found for a particular protein. For example, if *protein 1* was found in three different forms in the AIPM search, e.g. once with two methylations, once with a phosphorylation, and once with a β-methylthiolation, then the union of these PTMs was assigned to *protein 1.* This individually assigned PTM (two methylations + a phosphorylation + a β-methylthiolation in this example) represent the maximum PTM search space that was used to create PTM peptides of the given protein (*protein 1* in the example) in the BU search. For proteins not found in the AIPM search, peptides were generated without any PTM from the intact (non-modified) sequence of a given protein and tested in the BU search.

The advantage of this method over the "complete" search was the significant decrease in the number of theoretical peptide candidate sequences generated during the BU search by taking advantage of the "AIPM predicted" BU search. In this approach, peptide sequences for the BU search were generated based on the results of the AIPM search. Obviously, such a method requires good quality separation and identification of intact proteins. Otherwise, a valid, modified protein that was not identified in the AIPM, but truly existed in the sample, would not be represented in the BU search.

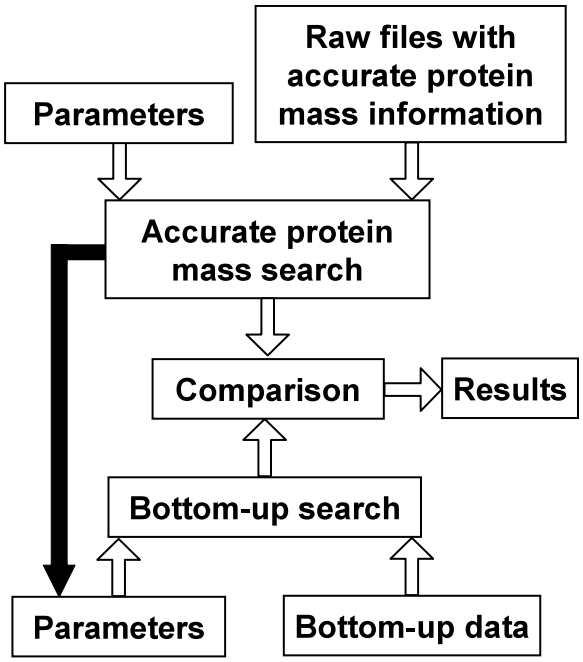

Figure 6.1. **Flowchart of Integration of Accurate Intact Protein Mass (AIPM) and Bottom up Searching Algorithms**
The filled arrow indicates the "AIPM predicted" BU search approach.

6.3.4: Decreasing the Number of Peptide Candidates by Restricting the Maximum Number of PTMs on a Single Peptide

To the best of our knowledge, the current commercially available BU search engines do not have the ability to limit the total number of different PTMs on a single peptide to a reasonable level that could be considered acceptable from a chemical viewpoint. However, a dramatic decrease in the number of peptide candidates and a noticeable search speed increase can be achieved when applying a limitation on the total number of PTMs on a single peptide, as described in the two scenarios below.

6.3.5: Evaluation of *PTMSearchPlus* for *Escherichia coli* Ribosomal Proteins

A full protein database of *Escherichia coli (K12)* provided a base for *PTMSearchPlus* to evaluate its effectiveness with a more complex sample. A purified ribosomal protein mixture was divided into two parts followed by their independent AIPM and BU analyses. A combined AIPM and BU search was performed on the data obtained. The search was accomplished using "complete" and "AIPM predicted" BU searches as well. The PTMs included in the AIPM search were mono-, di- and trimethylation on arginine and lysine, methionine truncation at the N-terminus, and disulfide formation between cysteine residues. Within the BU search the specified PTMs were mono-, di- and trimethylation on arginine and lysine, and methionine truncation at the N-terminus. (Note, that acetylation was not specified explicitly as a PTM, but must be considered when trimethylation with the same approximately 42 Da mass shift, was found.)

From this integrated AIPM-BU search, we identified 52 out of the total 54 possible ribosomal proteins, many of which were not modified or only exhibited methionine truncation. **Table 6.1** summarizes the PTM-containing ribosomal proteins and peptides confidently identified by an AIPM and/or a BU search. The four PTM proteins identified (L7/L12, L11, S5 and S11) all had PTMs that exactly matched with the PTM of the corresponding peptide found using an "AIPM predicted" BU search. This data demonstrates the unique advantage of coupling AIPM and the BU datasets, in which higher confidence is achieved by the related but independent measurements.

Table 6.1: Subset of Ribosomal Proteins Identified by TD and BU MS
Escherichia coli ribosomal proteins and peptides confidently identified with PTMs by accurate intact protein mass (AIPM) and bottom-up (BU) searches using *PTMSearchPlus*.

Protein	AIPM PTM	Δp	BU PTM peptides	BU score
L7/L12	(M loss) + TriMet/Ace	0.2	***(M loss)SIT(K+TriMet/Ace)DQIIEAVAAMSVMDVVELISAMEEK***	***85.61***
L11	TriMet/Ace	1.8	LQVAAGMANPSPPVGPALGQQGVNIMEFC(K+TriMet/Ace)AFNAK	43.43
S4	N/A	N/A	*C(K+Met)IEQAPGQHGAR*	33.71
S5	(M loss) + TriMet/Ace	18.5	(M loss)AHIE(K+TriMet/Ace)QAGELQEK	32.33
S11	(M loss) + Met	27.2	(M loss)A(K+Met)APIRAR	28.19

Namely, the AIPM data of these four proteins confirms that all of the PTM peptides were found in the BU search, i.e. no peptides with a PTM was missed. This confirmation is not available without coupling the approaches together. On the other hand, the BU search determines the location of the PTM that is difficult to ascertain by the AIPM search.

As an example, **Figure 6.2** presents corresponding identifications from AIPM and BU searches of the same protein. Figures 4a and 4b show calculated and measured isotopic distributions, respectively, of *50S ribosomal protein L7/L12* with methionine loss and trimethylation/acetylation found by the AIPM search. The mass difference between the most abundant peaks of the calculated and measured isotopic distributions was 0.2 ppm. The location of the trimethylation/acetylation identified by the AIPM search was determined by BU data. The spectrum in Figure 4c is assigned to a peptide of *50S ribosomal protein L7/L12* with a sequence of SIT(K+trimethylation/acetylation)DQIIEAVAAMSVMDVVELISAMEEK. This peptide contains a trimethylation on K5 and is also result of a methionine truncation of the original protein.

A "complete" BU search was also performed to check the validity of the "AIPM predicted" analysis with a complex sample. In the “complete” search, a peptide of protein S4 with a methylation was found, which had not been identified previously in the "AIPM predicted" BU search (see **Table 6.1**). The reason for missing the methylated peptide by the BU search was likely due to the lack of finding the modified S4 protein by the AIPM search (the unmodified S4 also was not detected). This resulted in a peptide database containing only the non-PTM peptides of S4 during the BU search. As S4 is a 23.5 kDa protein, the reason for not identifying it in the AIPM search is most likely that it was not eluted off the C4 reverse phase column used in the AIPM analysis. Manual inspection revealed very few peaks above 20 kDa identified by the AIPM analysis. At present, the integrated AIPM-BU search discussed above does not provide capability to track the PTM peptides of a protein that are not found by the AIPM method (i.e. if it didn’t elute

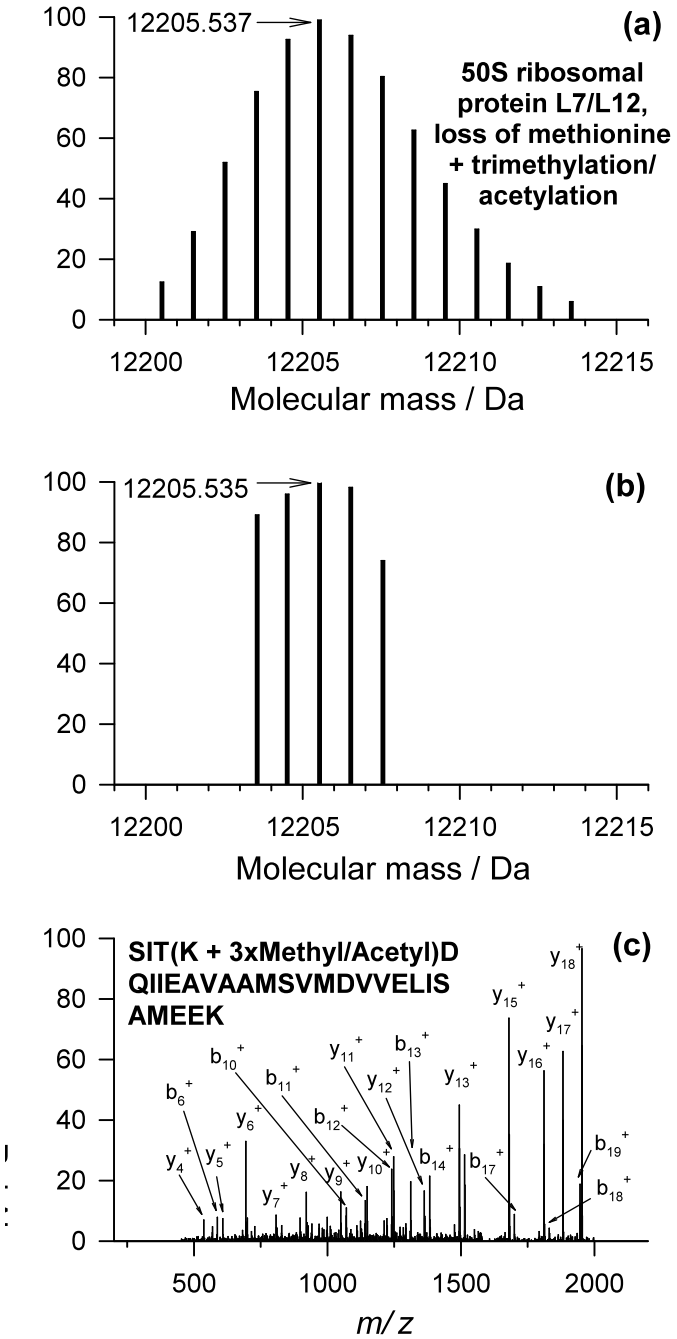

Figure 6.2: High Resolution Mass Spectrum of Ribosomal Subunits
(a) Calculated and (b) measured isotopic distributions for *50S ribosomal protein L7/L12* with methionine loss and monomethylation exhibiting 0.2 ppm mass difference between their most abundant peaks. (c) MS/MS spectrum of peptide SIT(K+3xMethyl/Acetyl)DQIIEAVAAMSVMDVVELISAMEEK of the same protein.

from the column or was not detected in the intact form for any reason). However, the "AIPM predicted" BU search does look for non-PTM peptides of proteins not found in the AIPM search. While the current version of *PTMSearchPlus* does not include this feature, confident identification of peptides of these proteins in the BU search could be used as a trigger to search for PTMs on peptides of a protein not identified in the AIPM search. Furthermore, based on the experimental data, a cutoff mass for the "AIPM predicted" BU search could be specified in the software to target this problem from another angle. The cutoff mass would define the mass that a protein has to exceed in order to generate its peptides using a "complete" BU search. This modification would decrease the chance to miss a PTM peptide even if the protein is not eluted from the separation column during the AIPM analysis, while keeping the speed advantage of the "AIPM predicted" BU search for proteins below the cutoff mass.

6.4: Bottom Up and Top Down Characterization of the Extracellular Fraction of the AMD Microbial Community

In order to assess the functionality of *PTMSearchPlus* with a significantly more complex sample set, an AMD extracellular fraction was submitted for discrete BU and TD analyses, as well as an integrated BU/TD analysis. In order to reduce the protein complexity, the extracellular fraction was subjected to cation exchange fractionation prior to MS measurement. The resulting twenty-nine fractions where then divided in half for BU and TD analysis. The intact protein MS measurement of the AMD extracellular fraction serves several purposes. First, the reduced complexity produced by off-line fractionation results in samples that are amenable to both BU and TD methodologies. This results in an excellent sample for methodological improvement and testing. Secondly, the direct measurement of the intact proteins form the extracellular fraction has not been previously obtained. The range of identifications, presence of PTMs or cleavages can be deduced from the TD measurement. Integrating the BU peptide dataset with the TD molecular form assignment provides increased support for a particular identification.

6.4.1: Bottom Up Analysis of 29 AMD Extracellular Fractions

Following the BU analysis of the 29 fractions, a total of 744 non-redundant protein identifications were made (**Figure 6.3**). Although more proteins were identified in this analysis than the previous characterization of the extracellular fraction (531), this number is consistent within the range of identifications from multiple extracellular analyses. Among the proteins identified numerous known cytoplasmic proteins (ie. ribosomal proteins) were also identified indicating unintended cell lysis occurred during sample processing. This may account for the increase in proteins identified. The range of proteins identified is remarkably similar to other analyses with numerous proteins with an unknown function present as well as protease, transporters and cytochromes. Each of the major microbial species are represented through the protein identifications including several archaeal species.

6.4.2: Top Down Analysis of 29 AMD Extracellular Fractions

The remaining portion of the 29 extracellular fractions was processed for LC-FTICR-MS. 387 non-redundant proteins were identified at < 5 parts per million (ppm) mass error (**Figure 6.3)**. Allowed modifications included cleavage of the n-terminal methionine (-131.04), disulfide bond (-2.016), methylation (+14.016) and oxidation (+15.995). Additionally, over 250 replicated, abundant masses were identified that were not identified from the sequence database. These are likely legitimate proteins due to their abundance, isotopic distribution, and intensity but for many possible reasons were unable to be assigned to a predicted protein. The most likely reasons include the presence of additional PTMs that were not included in the search or amino acid variations resulting in significant mass shifts. **Table 6.2** provides a subset of the list of the TD identifications. The proteins listed have a mass < 20 ppm with 29 of 38 having a mass error < 2 ppm. Among the proteins identified, a significant number currently have an unknown function but proteins performing expected functions (ie, protein degradation and transport) are also present. The remarkable mass accuracy achievable through FTICR-MS provides significant confidence in the identification of these proteins.

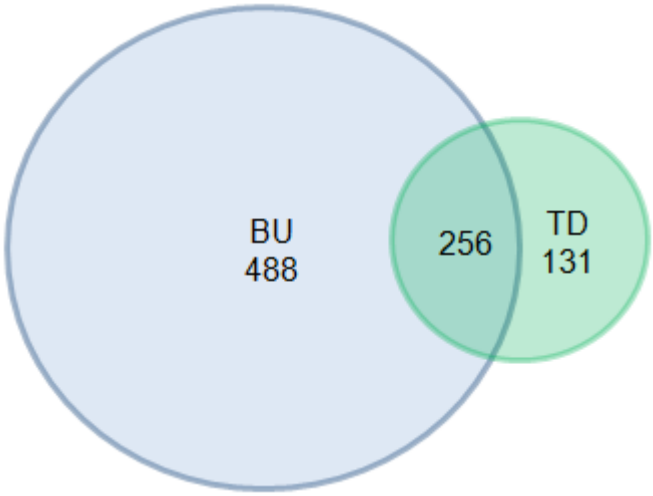

Figure 6.3: Protein Identifications from the BU and TD MS Analysis of the Extracellular Fractions

Table 6.2: Example TD Protein Identifications with Extremely Low PPM

Protein name	Protein description	Calculated Mass (Da)	Observed Mass (Da)	PPM
UBA_LeptoII_Scaffold_8135_GENE_46	Uncharacterized conserved protein	10064.021	10064.021	0.0
fer2_scaff_553_GENE_4	Hypothetical protein	12148.017	12148.017	0.0
UBA_LeptoII_Scaffold_8241_GENE_655	Hypothetical protein	13216.951	13216.951	0.0
fer2_scaff_10_GENE_60	Cysteinyl-tRNA synthetase	15409.446	15409.446	0.0
UBA_LeptoIII_397_37	E-Glycine cleavage system H protein (lipoate-binding)	14042.828	14042.827	0.1
gpl_scaff_305_GENE_7	Dehydrogenase (short-chain alcohol dehydrogenases)	26433.285	26433.289	0.1
5way_CG_LeptoII_scaff_8_GENE_1	Hypothetical protein	9696.871	9696.873	0.2
UBA_LeptoII_Scaffold_8524_GENE_180	Hypothetical protein	9696.871	9696.873	0.2
UBA_LeptoII_Scaffold_8049_GENE_275	Hypothetical protein	12052.827	12052.825	0.2
gpl_scaff_126_GENE_6	5'-3' exonuclease (including N-terminal domain of PolI)	14041.928	14041.925	0.2
5way_CG_LeptoII_scaff_76_GENE_21	Ferredoxin	12844.499	12844.502	0.3
UBA_LeptoII_Scaffold_8062_GENE_94	Undefined product	13174.130	13174.134	0.3
5way_CG_LeptoII_scaff_303_GENE_2	Adenylylsulfate reductase	12843.874	12843.878	0.3
Unass_arch_scaff_733_GENE_3	Hypothetical protein	12054.088	12054.093	0.4
5way_CG_LeptoII_scaff_374_GENE_1	Hypothetical protein	8973.303	8973.307	0.5
UBA_LeptoII_Scaffold_8135_GENE_15	Hypothetical protein	24829.289	24829.304	0.6
fer1_scaff_249_GENE_1	Amino acid transporters	21167.384	21167.371	0.6
fer2_scaff_307_GENE_5	Hypothetical protein	11640.258	11640.249	0.7
fer2_scaff_118_GENE_2	Hypothetical protein	12053.138	12053.128	0.8
UBA_LeptoIII_335_7	Hypothetical protein	14041.414	14041.427	0.9
gpl_scaff_269_GENE_6	RNA-binding protein Rrp4	24828.712	24828.736	1.0
UBA_LeptoIII_350_18	Transposase	26432.928	26432.960	1.2
UBA_LeptoIII_177_1	Hypothetical protein	20006.233	20006.206	1.3
Unass_arch_scaff_505_GENE_7	Hypothetical protein	19372.482	19372.508	1.4
fer1_scaff_176_GENE_1	Precorrin isomerase	19144.992	19144.966	1.4
UBA_LeptoII_Scaffold_8135_GENE_107	Predicted site-specific integrase-resolvase	7224.451	7224.440	1.5
5way_CG_LeptoII_scaff_59_GENE_15	Cytidine/deoxycytidylate deaminase	18082.600	18082.628	1.6
fer1_scaff_578_GENE_5	Ribosomal protein	17240.285	17240.256	1.7
UBA_LeptoIII_274_2	Transposase	16454.881	16454.909	1.7
5way_CG_LeptoII_scaff_75_GENE_9	Transposase	6513.394	6513.408	2.1
UBA_LeptoII_Scaffold_8062_GENE_180	Mg-dependent Dnase	28793.801	28793.865	2.2
gpl_scaff_460_GENE_3	Transketolase	16436.039	16436.096	3.5
UBA_LeptoII_Scaffold_8049_GENE_295	Transposase	16876.303	16876.237	3.9
UBA_LeptoII_Scaffold_8241_GENE_177	Transposase	19049.772	19049.848	4.0
UBA_LeptoIII_377_25	Ubiquinone oxidoreductase subunit 3 (chain A)	14325.071	14325.140	4.8
UBA_LeptoII_Scaffold_8027_GENE_48	Biopolymer transport protein	16688.035	16688.234	11.9
gpl_scaff_609_GENE_5	Permease	29567.654	29567.201	15.3
5way_CG_LeptoII_scaff_3_GENE_96	Multiple antibiotic resistance	21400.754	21401.148	18.4

174 proteins were identified with cleavage of the n-terminal methionine (MET). **Table 6.3** highlights a subset of these proteins that exhibited an extremely low ppm mass error. The cleavage of the n-terminal MET is not uncommon and observed in both the TD and BU datasets. It has been estimated that ~80% of all proteins in any given proteome will display cleavage of the n-terminal MET.[150] In this analysis only 44% (at < 5 ppm) of the proteins were identified in with n-terminal MET cleavage. The most common residues following the n-terminal cleavage are: Ala (A), Cys(C), Gly (G), Pro (P), or Ser (S). 16 of the 39 (~41%) proteins exhibiting n-terminal cleavage had a second residue matching the commonly observed set. The range of n-terminal cleavage and the second residue after cleavage do not explicitly follow the findings from the *E. coli* analysis. This is not surprising as many of the proteins identified are unique to the AMD community as evident in the extreme number of proteins with an unknown function. Identification of the methionine peptidase may provide additional details regarding the range and specificity of n-terminal MET cleavage.

6.4.2: Integration of TD and BU Datasets

Figure 6.3 illustrates that among both the TD and BU dataset, 256 proteins were identified in both. Among the proteins identified through each method, the sequence coverage ranges from a low of 6% to full peptide coverage of the protein. The TD identifications remain at < 5 ppm mass error. The identification of representative peptides provides substantial confidence in the assignment of the protein by high mass accuracy. The 131 proteins that were not identified in the BU analysis can be attributed to several factors. First, the BU analysis requires two representative peptides per protein for identification. If one peptide of a particular protein was identified but no additional peptides were also identified, that particular protein would not be included. Therefore, it is possible that a subset of the proteins not identified by the BU analysis fall in to this category. Secondly, if the BU proteins were predicted to contain PTMs, based on the TD identification, and the peptides, for many reasons, did not contain the specified modifications, the protein would not be identified though the BU analysis. **Table 6.4** displays example proteins that were identified in both the TD and BU

Table 6.3: PTM – Methionine Cleavage

Protein name	Protein description	2nd AA after Met	Calculated Mass (Da)	Observed Mass(Da)	PPM
UBA_LeptoII_Scaffold_8241_GENE_554	Hypothetical protein	S	24828.8325	24828.83238	0.0
fer1_isolate.689	Conserved hypothetical protein	E	7146.2828	7146.282764	0.0
UBA_LeptoIII_376_15	Hypothetical protein	Y	8539.6496	8539.649481	0.0
5way_CG_LeptoII_scaff_86_GENE_19	Hypothetical protein	S	22031.1217	22031.12136	0.0
gpl_scaff_196_GENE_16	Hypothetical protein	N	13314.1218	13314.12151	0.0
fer1_isolate.151c	Inorganic pyrophosphatase	K	19049.8468	19049.84796	0.1
5way_CG_LeptoII_scaff_150_GENE_15a	Ribosomal protein L34	S	5100.0811	5100.080697	0.1
UBA_LeptoII_Scaffold_8241_GENE_488a	Ribosomal protein L34	S	5100.0811	5100.080697	0.1
UBA_LeptoII_Scaffold_8049_GENE_307	Undefined product	G	15410.3741	15410.37197	0.1
5way_CG_LeptoII_scaff_22_GENE_15	Hypothetical protein	N	12862.5013	12862.50404	0.2
UBA_LeptoII_Scaffold_8062_GENE_162	Lactoylglutathione lyase	K	16510.482	16510.48566	0.2
fer2_scaff_89_GENE_13	Hypothetical protein	A	15409.7983	15409.79488	0.2
UBA_LeptoII_Scaffold_8049_GENE_252a	Hypothetical protein	G	9695.7711	9695.773381	0.2
UBA_LeptoII_Scaffold_8241_GENE_320	Ribosomal protein L27	A	10070.3452	10070.34841	0.3
gpl_scaff_133_GENE_15	Hypothetical protein	D	13028.9276	13028.92287	0.4
fer2_scaff_90_GENE_11	Hypothetical protein	Y	8489.8804	8489.877225	0.4
UBA_LeptoII_Scaffold_8692_GENE_102	Ribosomal protein S17	S	10583.4391	10583.44312	0.4
fer1_isolate.1510c	Hypothetical protein	K	10575.3343	10575.33838	0.4
UBA_LeptoII_Scaffold_8524_GENE_197	Undefined product	T	14110.0122	14110.00514	0.5
5way_CG_LeptoII_scaff_140_GENE_1	Cold shock protein	A	7227.9763	7227.980857	0.6
fer1_scaff_143_GENE_3	Hypothetical protein	R	8837.519	8837.513065	0.7
fer1_scaff_832_GENE_3	Superfamily II DNA and RNA helicases	D	22140.9493	22140.96454	0.7
Unass_bact_scaff_458_GENE_4	Hypothetical protein	K	16625.1411	16625.15336	0.7
fer2_scaff_56_GENE_27	Hypothetical protein	E	13216.1597	13216.14987	0.7
gpl_scaff_688_GENE_1	Hypothetical protein	A	8612.0668	8612.059662	0.8
UBA_LeptoIII_398_37	Hypothetical protein	I	6207.1567	6207.151501	0.8
UBA_LeptoIII_398_108	Hypothetical protein	T	5883.655	5883.650009	0.8
UBA_LeptoIII_371_28	Hypothetical protein	A	6514.4992	6514.493088	0.9
5way_CG_LeptoII_scaff_6_GENE_22	Hypothetical protein	S	13026.8824	13026.89475	0.9
UBA_LeptoII_Scaffold_8524_GENE_126	Undefined product	S	13026.8824	13026.89475	0.9
UBA_LeptoIII_345_6	Hypothetical protein	I	18938.27	18938.25164	1.0
fer1_isolate.1799c	Hypothetical protein	Y	10582.9321	10582.92182	1.0
UBA_LeptoIII_383_34	Hypothetical protein	P	13027.7055	13027.6922	1.0
Unass_arch_scaff_730_GENE_6	Ribosomal protein L1	R	16436.1151	16436.0956	1.2
fer2_scaff_93_GENE_9	Hypothetical protein	Y	14096.8348	14096.81408	1.5
UBA_LeptoIII_364_15	Hypothetical protein	S	7859.9772	7859.965395	1.5
5way_CG_LeptoII_scaff_27_GENE_50	Hypothetical protein	R	10577.1177	10577.13413	1.6
UBA_LeptoII_Scaffold_7931_GENE_426	Hypothetical protein	R	10577.1177	10577.13413	1.6
fer1_scaff_128_GENE_3	Hypothetical protein	D	9693.1206	9693.105504	1.6

a)

Gene ID	Sequence Coverag	Annotation
5way_CG_Leptoll_scaff_222_GENE_9	100	thioredoxin
5way_CG_Leptoll_scaff_46_GENE_4_SigP	100	hypothetical protein
UBA_Leptoll_Scaffold_7931_GENE_101_SigP	100	Hypothetical protein
UBA_Leptoll_Scaffold_8027_GENE_1	100	Thiol-disulfide isomerase and thioredoxins
UBA_Leptoll_Scaffold_8062_GENE_53_SigP	100	Hypothetical protein
UBA_Leptoll_Scaffold_8062_GENE_63	100	Hypothetical protein
UBA_Leptoll_Scaffold_8135_GENE_107	100	integrase-resolvase
UBA_Leptoll_Scaffold_8524_GENE_247	100	undefined product
UBA_Leptoll_Scaffold_8692_GENE_94	100	Hypothetical protein
5way_CG_Leptoll_scaff_137_GENE_1a	98	Ribosomal protein L33
UBA_Leptoll_Scaffold_8524_GENE_248_SigP	98	undefined product
UBA_Leptoll_Scaffold_8241_GENE_298_SigP	96.6	Outer membrane protein
Unass_bact_scaff_1107_GENE_1	96.5	Translation initiation factor IF3
UBA_Leptoll_Scaffold_8062_GENE_372	87	Cytochrome 579
UBA_Leptoll_Scaffold_8062_GENE_372_SNP1	87	Cytochrome 579

b)

Gene ID	M/Z Error	PTM	Function
UBA_Leptoll_Scaffold_8027_GENE_1	0.014571	1xDEM-1xDIS-	Thiol-disulfide isomerase and thioredoxins
UBA_Leptoll_Scaffold_8135_GENE_107	0.10324	N/A	Predicted site-specific integrase-resolvase
UBA_Leptoll_Scaffold_8062_GENE_372	0.137384	1xDIS-	Cytochrome 579

Table 6.4: Results of TD and BU Analysis of the 29 Extracellular Fraction from the AMD Microbial Community

a) Example protein identifications following bottom up analysis. Two proteins that exhibited 100% sequence coverage are highlighted. b) Results of TD analysis displaying complementary intact protein identification including any identified PTMs. Highlighted proteins were also identified in the BU analysis.

analysis. The highlighted rows are examples of proteins that were identified in the BU analysis with high sequence coverage and through the TD analysis with high mass accuracy and a low mass error. Both the high sequence coverage and the low mass error support the unambiguous identification of the proteins. Additionally, the highly abundant and previously characterized Cyt_{579} was identified in both the BU and TD analyses and is highlighted in **Table 6.4**.

6.5: Conclusions

PTMSearchPlus provides a novel computational approach for the integration of accurate intact protein mass (AIPM) and bottom-up (BU) searches to both confidently identify intact proteins and to characterize their PTMs. The required input data are a FASTA protein database, a selection of possible PTMs, the types and ranges of which can be specified, and both intact protein and proteolytic peptide mass spectra data collected from the same protein mixture. After a search is conducted, the software outputs a list of intact and PTM proteins matching the AIPM data with their respective peptides found by the BU search. This list also includes protein and peptide sequence coverage information, scores, etc. Furthermore, manual evaluation including visual inspection of annotated AIPM and BU mass spectra to evaluate, modify (e.g. remove obvious false positives, low quality spectra etc.) and (automatic) refiltering of the results is also possible in the software. Improvement in BU search speed when limiting the total number of possible PTMs on a peptide or performing an “AIPM predicted” search was also evaluated. All of these features of *PTMSearchPlus* were demonstrated using a protein standard mixture or a complex protein mixture from *Escherichia coli*. Also demonstrated was a unique advantage of coupling AIPM and the BU datasets mutually beneficial for both approaches: AIPM data can confirm that no PTM peptides were missed in a BU search, while the BU search determines the location of the PTM, which is not readily determined through an AIPM search alone. The “AIPM predicted” search resulted in the first analysis of intact proteins from the AMD microbial community. The initial results provided high confidence identifications of PTM proteins. A future analysis of the AMD TD, BU dataset, with additional combinations of PTMs will be performed.

Currently, development of a new scoring algorithm for the AIPM search is under way in which the score is based on mass and intensity differences of the peaks in the theoretical and measured isotopic envelopes. Future work also includes evaluation of using a cutoff mass for the "AIPM predicted" BU search. Furthermore, assessment of triggering a "complete" BU search of a protein when it is not identified by the AIPM search but confident identification of corresponding peptides by the BU search is available, will be accomplished.

Chapter 7

Development of a Spectral Assignment Approach to Evaluate Assigned versus Unassigned Tandem Mass Spectra in the Proteomic Analyses of Microbial Isolates and Communities

Portions of included text are adapted from:

Brian K. Erickson, Alison R. Erickson, Brian D. Dill, Nathan C. VerBerkmoes, Jillian F. Banfield, Robert L. Hettich, "Evaluation of Quality Matched Versus Quality Unmatched Tandem Mass Spectra in the Proteome Characterizations of Microbial Isolates and Communities", Manuscript in preparation.

Brian K. Erickson's contributions include software design and implementation, data analysis, and primary authorship.

7.1: Introduction

Mass spectrometric (MS) based proteomics analyses are capable of identifying thousands of proteins from a wide range of samples. The transition from analyzing microbial isolates towards complex, natural microbial samples has uncovered evidence of metabolic portioning, dynamic protein expression, and growth state dependent protein export.[68, 105] Although mass spectrometric based proteomics is remarkable in its ability to rapidly characterize thousands of proteins from a complex microbial community, the overall level of proteomic depth remains low when compared to the total suite of expressed proteins.[2, 12] Historically, the success of a given MS proteomic experiment was based on the number of proteins identified. When comparing the ~1400 proteins identified in a microbial isolate to the thousands identified in a community, it would appear that the overall methodology is performing well, as

evidenced by the increased protein identifications.[19, 42] A quantitative spectral analysis evaluation of identified vs. unidentified peptides, as described within this text, was developed in order to more fully describe the effectiveness of the current MS methodologies on complex microbial communities. This spectral analysis has revealed that although the number of identified *proteins* is increasing, the number of *peptide* identifications, used to piece together the presence of protein, has decreased dramatically. When comparing an isolate to a community sample, upwards of ~70% fewer tandem mass spectra are being assigned to peptides. It is obvious that determining the cause for this reduced number of spectral assignments is critical for achieving the desired proteomic depth of these complex communities.

Current MS methodologies of complex, microbial samples utilize a variety of sample preparation methods, but all depend highly on the enzymatic digestion of proteins, liquid chromatographic separation and MS measurement of peptides. The mass spectrometer operates in a data-dependent mode whereby a survey (MS1) scan results in peptide targets that are collisionally fragmented resulting in a tandem mass spectrum (MS2).[34] The MS2 spectra are computationally assigned to peptides that are present in a sequence database consisting of all predicted proteins. The computational assignment is performed by several notable programs including SEQUEST, Mascot or X!Tandem[25, 142, 143] and is based on pattern matching between predicted fragmentation of the computational sequence and the experimentally derived sequence. The assigned MS2 spectra are then filtered and assembled to represent predicted proteins in the database.[71] This represents the total suite of identified peptides and proteins from a given MS experiment.

During a standard MS analysis of a microbial community, the mass spectrometer will collect ~100,000 spectra. These spectra will represent a combination of survey (MS1) and tandem (MS2) spectra. A subset of the MS2 spectra will be assigned a predicted peptide and a separate subset will remain unassigned. Of the MS2 that are unassigned to a peptide, a portion will be of low quality (i.e., insufficient fragments and/or low intensity), whereas the remaining will be of high quality, but for a variety of potential reasons were not assigned to a peptide. By quantifying and categorizing the

total spectra collected during a MS experiment, it is possible to gain a more defined image of the experimental performance. Specifically, the number of assigned MS2 and the number of quality unassigned MS2 provide valuable metrics relating to the sample preparation, chromatography, mass spectrometer operating parameters and computational peptide assignment. A decrease in the number of assigned MS2 spectra, as observed when comparing the isolate and the microbial community, illustrates the ineffectiveness of the current methodology. Although the number of proteins identified increases between an isolate and community, the significant decrease in assigned MS2 spectra dictates that inefficiencies are present along the experimental MS path. Furthermore, the presence of quality unassigned spectra represents a significant subset of untapped proteomic information. Each of these unassigned spectra may correspond to a previously unidentified peptide from a previously unidentified protein. Determining the potential causes for the presence of these quality unassigned spectra and addressing means to assign a peptide identification is necessary to achieve deeper proteomic depth. Finally, these spectra are excellent datasets for submission to the ever improving suite of *de novo* spectral assignment algorithms, post-translational identification tools or algorithms for sequence tag discovery.[151]

To highlight the marked differences in spectral assignment between microbial isolates and communities, protein samples from *E. coli, R. palustris*, a low complexity microbial community living in acid mine drainage (AMD), and collected groundwater from a soil remediation site were compared. *E. coli*, and similarly *R. palustris*, represents a baseline for MS proteomic analysis, as the sample preparation is rather routine and uncomplicated and the protein databases are well characterized and curated.[152, 153] On the other hand, samples collected from the natural communities are present in extreme environmental conditions and contain complicated matrices. Furthermore, the protein databases differ significantly in their makeup and curation when compared to *E. coli*. The AMD microbial community has utilized significant resources in order to generate a suitable metagenome representing the abundant species within the community.[20] Although the species within the community show very

limited homology to previously studied organisms, the direct metagenome provides a very solid foundation for proteomic analysis.[19] The level of sequence accuracy compared to that of *E. coli* is, of course, substantially lower as the AMD community represents multiple species with several closely related strains. At the other end of the spectrum, the groundwater microbial community does not have a direct, matched metagenome and instead is dependent upon protein sequences from microbes that are assumed to closely resemble, at best estimate, those that are present in groundwater. MS based proteomics is being challenged with a variety of samples that the datasets utilized throughout this study represent. Through the use of spectral analysis, it is possible to gain a significantly deeper understanding of the success and failure of MS based proteomic measurements. It is then feasible to specifically optimize relevant procedures and parameters in order to increase proteomic depth. Ultimately, these gains in proteomic coverage will hopefully serve to further unlock the biological processes that allow these communities to thrive. The comparative spectral analysis between microbial isolates and communities described within presents the first targeted look at the experimental implications of community MS analysis and serves to provide a process to improve the experimental results.

7.2: Materials and Methods

7.2.1: Protein Sample Preparation and Tryptic Digestion

Escherichia coli K-12 and *Rhodopseudomonas palustris* lysates were used as representative low complexity, bacterial isolates for all experiments. Samples collected from the Acid Mine Drainage (AMD) in Redding, CA and groundwater in Rifle, CO were used to represent higher complexity, natural microbial community samples.

Approximately 3 mg of cells were processed via a single tube cell lysis method[43] and suspended in 6M guanidine/10mM DTT in order to lyse cells and denature proteins. The guanidine concentration was diluted to 1M with 50 mM Tris buffer/10mM $CaCl_2$ and sequencing grade trypsin (Promega, Madison, WI) was added to digest proteins to peptides. The complex peptide solution was desalted via C18 solid phase extraction,

concentrated and filtered (0.45um filter). For each 2D-LC-MS/MS analyses below, ~1/5 of the total sample was used.

A sample was collected from the AB End location of the Richmond mine (Iron Mountain, Redding, CA) and flash frozen on site. Whole cell fractions were obtained by acid lysis and extraction as previously described.[19]

Three filtered groundwater samples (named: Cristobal, Hanna and Borris) were collected from a water well (#4) located in Rifle, Colorado. All community samples were thawed and microbial cells were extracted from the bulk material. Approximately 3 mg of total protein was processed via single tube cell lysis, as above, and denatured in 6M guanidine/10mM DTT to lyse cells and denature proteins. The guanidine concentration was diluted to 1M with 50 mM Tris buffer/10mM $CaCl_2$ and sequencing grade trypsin (Promega, Madison, WI) was added to digest proteins to peptides. The complex peptide solution was desalted via C18 solid phase extraction, concentrated and filtered (0.45um filter). For each 2D-LC-MS/MS analyses below, ~1/4- 1/5 of the total sample was used. Each of the isolates and community trypsin-digested samples used in this study were analyzed by 22hr, 12-step MudPIT.

7.2.2: MS Analysis

Each of the trypsin-digested isolates, *E. coli* and *R. palustris,* were individually loaded onto a split-phase column (RP-SCX-RP) and analyzed via 2D-LC-MS/MS connected to a linear ion trap, LTQ (ThermoFisher Scientific) coupled to a nanoflow high performance liquid chromatography system (HPLC, Dionex U3000) using a nanospray ionization source (Proxeon). The LTQ settings were set to acquire a full MS scan (from 400 to 1700 m/z) followed by five data-dependent MS/MS, 2 microscans for both full and MS/MS scans, centroid data for all scans and 2 microscans averaged for each spectra, dynamic exclusion set at 1.

The trypsin-digested AMD biofilm community sample (sample_B_run2) was loaded (~150 µg) onto a split-phase column (RP-SCX-RP) and analyzed via 2D-LC-MS/MS with on a high performance LTQ-Orbitrap (ThermoFisher Scientific) coupled to a nanoflow high performance liquid chromatography system (HPLC, Dionex U300) using

a nanospray ionization source (Proxeon) as described previous.[19, 69] The Orbitrap settings were as follows: 30K resolution on full scans in Orbitrap, all data-dependent MS/MS in LTQ (top five), 2 microscans for both full and MS/MS scans, centroid data for all scans and 2 microscans averaged for each spectra, dynamic exclusion set at 1. The four filtered groundwater samples were analyzed individually via 1D-LC-(SCX-RP)-MS/MS on a LTQ-Orbitrap (ThermoFischer Scientific) with the same settings as described above.

7.2.4: Data Processing

All MS/MS datasets were searched with the SEQUEST algorithm and filtered with DTASelect/Contrast at the peptide level with a minimum Xcorr of 1.8 (+1), 2.5 (+2), 3.5 (+3) and a minimum deltCN of 0.08.[25, 71] Only proteins identified with two fully tryptic peptides from the 22 hr runs were considered for further biological study. Tandem MS/MS spectra were searched against the following databases: All tandem MS/MS collected from *Escherichia coli* K-12 (single experiment) were searched against a database containing proteins predicted to be encoded by its' genome, MS2 phage, and common contaminants (36 proteins). All MS/MS collected from *Rhodopseudomonas palustris* (single experiment) were searched against a database containing proteins predicted to be encoded by its' genome and contaminants. All tandem MS/MS collected from the AMD biofilm sample was searched against two different databases, (i) DB1: biofilm_db1 (Tyson et al. 2004; 12,148 proteins) and (ii) DB2: Biofilm_5wayCG_UBA_06162006 (Lo et al. 2007; 16,170 proteins) which contains additions from supplementary genomic sampling at the UBA location. All MS/MS collected for each groundwater sample were searched separately against the same database, rifle_geobacter7_01092008 (26,272 proteins) with the same parameters described above. Detailed information regarding each database can be found in **Table 7.1**.

Table 7.1: Number of predicted proteins in the protein sequence databases

Sample	Database	# Predicted proteins	Avg protein MW (Da)	Avg protein length (aa)	Description
Isolates	E. coli K-12	4,391	34,651.12	324.10	E. coli genome
	R. palustris	4,877	35,201.45	340.62	R. palustris genome
Groundwater	Rifle_geobacter7	26,272	37,171.06	344.16	7 Geobacter genomes (2008)
AMD	DB1	12,148	28,254.84	256.99	CG_allORFS+fer1_missing (2004)
	DB2	16,170	29,791.08	272.60	5wayCG+UBA (2006)

7.2.5: Spectral Analysis

Automation of the spectral analysis was accomplished through a collection of in-house developed Perl scripts which automated the following procedures. "mzXML" files representing all spectra collected during the MS analysis were parsed and divided into MS1 and MS2 spectra. Concurrently, the SEQUEST results were parsed in order to identify the subset of spectra assigned to a peptide. The remaining unassigned MS2 files were then sub-divided into poor quality and high quality unassigned MS2 spectra. The criteria for high quality spectra classification was based on intensity and fragment distribution as follows. The charge state of the precursor peptide was required to be greater than 1. This conservative metric increases the likelihood that the targeted ion is in fact a peptide and a not a small molecule or lipid. The absolute intensity, based on empirical analysis of quality spectra, was required to be above 2500 counts. Finally, the MS2 spectrum was required to contain at least three fragment peaks that were within 20% of the base peak intensity. Any unassigned MS2 spectra not adhering to these criteria were classified as poor unassigned MS2 spectra. The classification for each spectra was formatted for display in Excel.

7.3: RESULTS

Detailed classification and characterization of spectra from proteomic analyses provides enhanced feedback of MS performance and ultimately provides greater proteomic depth. The number of assigned and unassigned MS2 spectra serves as an excellent metric of MS run to run comparison, especially for determining the performance impacts between isolate and natural community samples. Spectral assignment was performed on 3 proteomic samples sets (*E. coli / R. palustris*, soil groundwater, and AMD biofilm).

7.3.1: Spectral Assignment

Following MS analysis, the raw spectra was converted to the mzXMLformat for further processing.[154] Spectral assignment is operating system independent, but will perform optimally depending on the physical computer hardware. On a 2.13 GHz dual-

core processor with 2 GB of system memory, the spectral assignment of 137,556 spectra required ~40 minutes of computation time. The processing time can be reduced with higher performance hardware.

7.3.2: Control Sample

Initially, a control MS experiment containing no loaded peptides was submitted for spectral analysis. This control analysis was intended to result in no quality MS2 spectra and served to illustrate the ability of the algorithm to discern between quality and poor MS2 spectra. Utilizing the same experimental, SEQUEST (*E. coli* and *R. palustris* protein database) and DTASelect parameters, spectral analysis of 1,158 total spectra resulted in the identification of 1,265 MS2 spectra (**Table 7.2A**). As expected, all of the 1,265 MS2 spectra were not assigned to any peptides and were classified as poor unassigned. These results demonstrated the ability of the spectral analysis algorithm used in this study to identify poor unassigned MS2 spectra.

7.3.3: Bacteria Isolates

The proteomic analysis of *E. coli* resulted in the identification of 1,193 proteins from a total of 137,556 spectra (25,206 MS1 spectra, 112,350 MS2 spectra) (**Table 7.2B**). The 1,193 protein identifications were generated from 41,448 spectra to peptide assignments by SEQUEST. This results in ~36% of collected MS2 spectra being assigned to a peptide. Among the remaining 79,902 unassigned MS2 spectra: 12,830 were classified as quality unassigned and 58,072 were classified as poor unassigned. An additional standard, *R. palustris* was analyzed and submitted for spectral analysis. The MS experiment resulted in the identification of 1,410 proteins from a total of 112,600 spectra (20,093 MS1 spectra and 92,507 MS2 spectra) (**Table 2B**). 22,460 spectrum to peptide assignments were made by SEQUEST, representing ~24% of the MS2 spectra collected during the experiment. The unassigned MS2 spectra were composed of 15,664 quality and 54,383 poor quality MS2 spectra. What is notable is the decrease in the percent of MS2 assigned to a peptide, 36% (*E. coli*) and 24% (*R. palustris*). This decrease in the number of assigned MS2 could be the result of

Table 7.2: Results of Spectral Analysis on 4 samples Including Microbial Isolates and Communities

						Unassigned MS2 Spectra	
	Sample	# of Proteins Identified	# of MS1 Spectra	# of MS2 Spectra	Assigned MS2	Poor	Quality
[A]Control	No protein	0	253	1265	0	1265	0
[B]Isolates	*E. coli*	1193	25206	112350	41448	58072	12830
	R. pal	1410	20093	92507	22460	54383	15664
[C]Groundwater	Cristobal	1671	21109	105545	11904	67741	25900
	Hanna	760	18635	93175	6748	85476	957
	Borris	61	17462	87310	2648	83862	800
[D]AMD	DB1	1057	28747	115127	20895	81530	12702
	DB2	1819	28747	115127	27107	77053	10967

numerous factors, including incorrect sequence information and variations in the sequence of post-translational modifications (PTM). The 15,664 quality unassigned MS2 spectra measured in the *R. palustris* sample lend support that the experimental preparation and measurement proceeded properly and that the decrease in MS2 assignments is the result of sequence polymorphisms or PTMs.

7.3.3: Groundwater Microbial Community

Three different community groundwater samples resulted in the identification of 1,671 (Cristobal), 760 (Hanna) and 61 (Borris) proteins. The total number of spectra collected for Cristobal, Hanna, and Borris was 126,654 / 111,810 / 104,772 respectively (**Table 7.2C**). The variance in total spectra collected is a result of the automated parent ion selection of the instrument and reflects differences in sample quality and peptide availability for fragmentation. The Cristobal sample resulted in the highest number of MS2 spectra to peptide assignments, 11,904 (~11% of MS2 spectra). Hanna resulted in 6,742 (~7% of MS2 spectra) and Borris with 2,648 (~3% of MS2 spectra) spectra to peptide assignments. Numbers of unassigned MS2 spectra correlate well with the previously reported MS2 assignment results and follow expected trends based on the number of proteins identified and reflect the reduction of viable MS2 spectra in each respective sample. That is, as the number of assigned MS2 decrease, the number of peptide identifications decreases. Additionally, the number of quality unassigned MS2 decreases while the number of poor unassigned increases. It is evident that the Cristobal sample performed relatively well and resulted in significant amounts of peptides that were amenable for identification. The Hanna and Borris samples, on the other hand, resulted in very few assignments or quality unassigned. This suggests errors in the sample collection or preparation and will be further discussed later in this study. The determination and categorization of the unassigned spectra provides significant insight into the quality of the sample, efficiency of the sample preparation, and instrument performance.

7.3.4: AMD Biofilm Microbial Community

The soluble fraction collected from the AMD biofilm community was searched by SEQUEST with two protein sequence databases. Database 1 consisted of predicted protein sequences directly generated from the genome sequence of the microbes present in the mine. Database 2 was a refinement of database 1 and was expected to more accurately reflect the proteomes expressed by the microbes (see materials and methods). Database 1 resulted in the identification of 1,057 proteins while database 2 identified 1,819 proteins (**Table 7.2D**). This marked increase in protein identification highlights the benefits of consistent protein sequence database refinement. As the protein sequence database more accurately resembles the expressed proteome, the deeper the achievable proteomic coverage. As before, each of the results were submitted for spectral analysis in order to determine what effect the database refinement had on the numbers of assigned MS2 spectra. The MS analysis consisted of 143,874 spectra (28,747 MS1 spectra and 115,127 MS2 spectra). Database 1 resulted in 20,895 assigned MS2 spectra and database 2 resulted in 27,107 assigned MS2 spectra, an increase of 6,212 assigned MS2 spectra. As expected the number of poor unassigned spectra decreased from 81,530 to 77,053 and the number of quality unassigned also decreased from 12,702 to 10,967. This indicates that an additional 1,735 quality unassigned MS2 spectra were assigned to a peptide solely from the refinement of the protein sequence database. This also indicates that SEQUEST was able to identify an additional 4,477 spectra that were previously classified as poor unassigned MS2. The refined database enabled SEQUEST to match spectra to peptides with sufficient scoring thresholds that were previously unobtainable.

The proteomic results of microbial isolates and communities share numerous similarities and differences. Based on the use of a consistent set of MS parameters, it could be expected that the distribution of spectra (number of MS1 and MS2) should be relatively similar among all samples. Following spectral analysis it is obvious that distribution of spectra is highly dependent on the sample. A common MS methodology utilizes a parent scan (MS1) to provide the specified ion targets for fragmentation (data dependent MS/MS). In the experiments performed for this study, the parameters were

set such that each MS1 was followed by five MS2 scans (assuming a suitable parent ion is available). In general, it was observed that among each sample grouping (standards, groundwater and AMD), the number of collected MS1 spectra was relatively similar. Between groupings, the number of scans varies slightly, with the AMD sample resulting in the highest number of collected MS1 spectra (**Figure 7.1**). Given that the experimental parameters and collection time are the same for all samples, the differences in collected MS1 scans relates to the number of suitable parent ion selected for fragmentation. A decrease in the number of MS1 scans indicates that more parent ions were selected for fragmentation, generating a higher number of MS2 scans. This is critical as this directly relates to the number of potential peptides being assigned and ultimately the number of proteins identified. For the AMD samples, the increase in the number of MS1 scans could indicate that there is an insufficient peptide load or that the chromatography is not performing efficiently, resulting in redundant peptide elution. Additionally, it would be expected, if sufficient parent peptides are available, that the ratio of MS1:MS2 should be 1:5 for each sample. Based on spectral analysis it was observed that for the groundwater community sample the ratio of MS1:MS2 was in fact 1:5. It could be hypothesized that due to the saturation of MS2 spectra, the ratio of MS1:MS2 could be increased to 1:6, for example, in order to target and fragment additional peptides. On the other hand, the isolates (1:4.5 MS1:MS2) and AMD (1:4 MS1:MS2) samples fall under the expected 1:5 ratio and would suggest adjustments in the methodology or sample load and/or chromatography. Quantifying the number of MS1 and MS2 spectra provides tangible feedback regarding critical MS operational parameters. Adjustments based on this analysis would result in increased numbers of MS2 spectra which could then lead to increased proteomic depth.

7.4: Discussion

The spectral analysis described above attempts to classify and characterize the spectra collected during a MS experiment. Utilizing this characterization provides additional significant feedback regarding the effectiveness of the experiment as it relates to sample preparation, experimental parameters and database to peptide

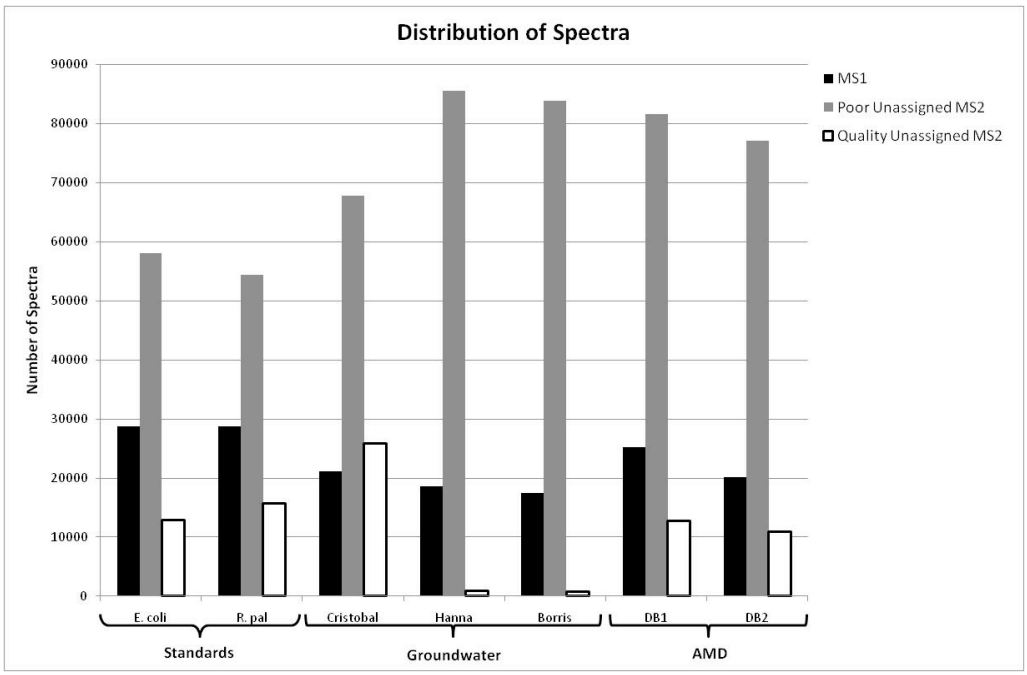

Figure 7.1: Distribution of Spectra Among All Samples

assignment. To illustrate the marked differences in microbial isolate and community proteomic samples, several MS datasets were analyzed via spectral analysis.

7.4.1: Protein database dependence

When a genome or metagenome of interest does not exist or is inaccessible for database searching, a significant, negative impact on the ability to assign MS/MS and generate statistically relevant peptide-protein identifications is observed. If one or more species has not been sequenced within a community sample or a protein has acquired a mutation (e.g., deletion, single amino acid polymorphism or translocation), the sequence database searching methodology will simply not identify these proteins. Although the amount of genomic sequencing data is increasing, the rate at which specific species of interest are sequenced to completion is not keeping pace. Additionally, if a genome has been sequenced, but was sequenced poorly or has insufficient sequence coverage prohibiting high-quality assembly, there is an increased chance to have multiple truncated open reading frames (ORFs) and proteins (not full length) which present which can hamper maximum MS/MS assignment. Thus, the quality and existence of single microbial genomes and environmental metagenomic sequences will impact the spectrum-peptide assignment and resulting proteome identifications. On the other hand, if a genome or metagenome is not available for the corresponding proteome, a comprehensive collection of published reference genomes (based on similarity to proteins of other species) will be concatenated into a single database. Although the addition of hundreds-to-thousands of reference genomes would be necessary to provide a wide array of sequence diversity and proteome coverage, the increased database size requires additional computational resources, increases spurious matches and false positives while decreasing the reliability and specificity of spectrum-peptide predictions.

In this study, we have chosen two representative microbial isolates, *R. palustris* and *E. coli*. Both isolates have finished genomes that are published and well-characterized. The AMD environmental community has also been sequenced as described by Tyson *et al* for DB1 and Lo et al. for DB2.[20, 69] DB1 contains genomic

sequences (*Leptospirillum* group II, *Leptospirillum* III, archaeal species, and other Bacteria and Eukarya) collected from the 5-way site (often referred to as the 5-way CG genomic dataset). Although DB1 is a complete composite metagenome, additional metagenomic sequences have since been acquired from another location in the mine, the UBA site, referred to as DB2 in this study. DB2 provides additional sequence diversity and sequence variants unlike DB1. To note, the average protein length and molecular weight are very similar and do not create any bias (**Table 7.1**). In this study, DB1 (older) and DB2 (updated) serve to represent the impact of the "quality" of a sequence database on MS/MS assignment. The groundwater environmental samples (Cristobal, Hanna, and Borris), on the other hand, do not have sequenced metagenomes. Therefore, based on previous literature, 7 isolate *Geobacter* genomes (*G. bemidjiensis*, *G.* M21, *G. sp.* FRC-32, *G. lovleyi* SZ, *G. metallireducens* GS-15, *G. uraniumreducens* RF4, *G. sulfurreducens*) were selected and concatenated into one database to represent the expected community metagenome for all three proteomes. These samples will serve to represent the impact of not having a matched genome or metagenome and sequence database and its affect on MS/MS assignment.

As demonstrated in **Table 7.2** and **Figure 7.2**, the quality of the sequence database has a significant impact on the number of peptide-protein identifications and % of MS/MS assigned. With the AMD proteome, the total number of assigned MS2 has increased from ~20,000 (DB1) to 27,000 spectrum to peptide assignments with DB2 providing deeper proteome coverage. The number of identified proteins has also increased with database quality from ~1,057 proteins with DB1 to ~1,819 proteins with the more representative DB2. Furthermore, if the well characterized isolates' database results are compared to either the groundwater community (reference genomes) or the AMD community (matched metagenome); the isolates' database is capable of assigning a higher proportion of MS2s compared to either community samples, especially the groundwater samples. Additionally, access to a sample derived database (DB1 or DB2) versus an estimated best-fit reference database (rifle_geobacter7) for community samples is significantly more truthful and effective

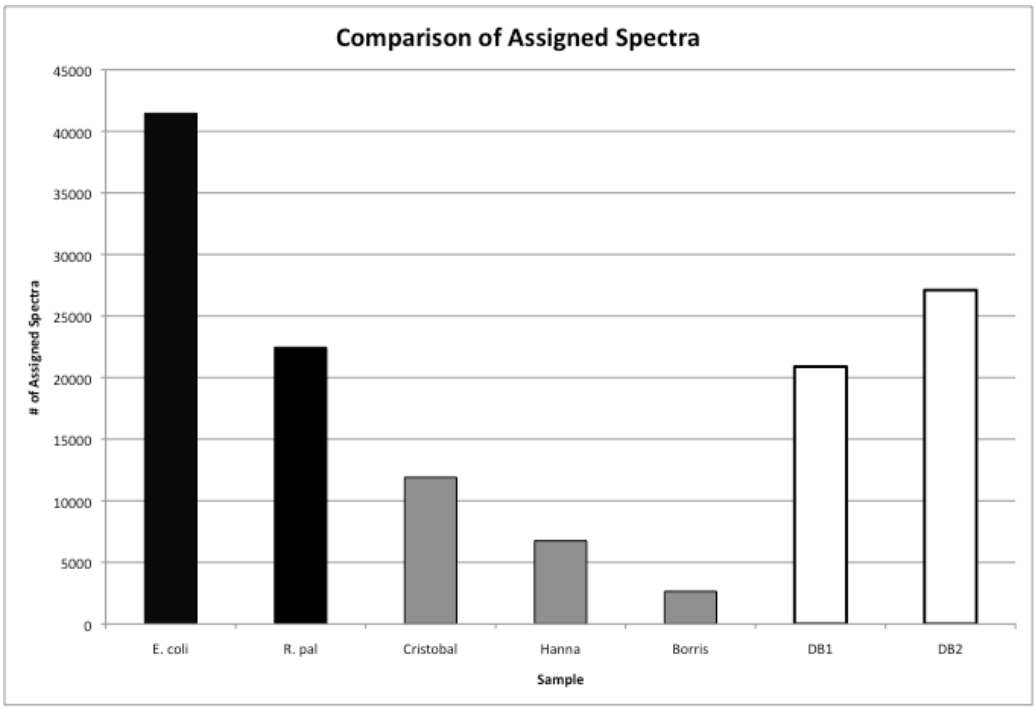

Figure 7.2: Number of Spectra Assigned to a Peptide Among Microbial Isolates and Communities

based on its ability to assign more MS2 spectra. This is especially significant considering that the groundwater database contains approximately twice the number of proteins as the AMD - DB1, highlighting that the total size of the database will not necessarily result in deep proteomic coverage. These results emphasize the need and value for creation and usage of the most representative and complete protein sequence database for the MS database searching methodology. The classification and assessment of collected and assigned MS-based spectra can be used as an additional tool to evaluate the successfulness of a database and its capability to maximize spectrum-peptide assignment.

7.4.2: Assigned MS2 Spectra

A critical, but relatively hidden metric of MS based proteomics is the determination of the number of MS2 spectra assigned to a peptide. As the protein identifications are directly tied to the assignment of peptides, it is useful to quantify the number of collected spectra that are contributing to the protein identifications. Currently, the number of peptides identified is routinely provided by most protein database searching algorithms. This peptide count should not be compared to the assigned spectrum count, as these algorithms remove redundant spectra assigned to the same peptide, unless explicitly specified not to. **Figure 7.2** illustrates the marked differences in assigned MS2 spectra observed among the samples. The *E. coli* sample resulted in the highest number (41,448) and the groundwater sample – Borris (2,648), the lowest number of assigned MS2 spectra. Among the isolates, the *R. palustris* sample resulted in approximately half (22,460) as many assigned MS2 spectra when compared to the *E. coli* sample. The *E. coli* sample serves as an excellent baseline, with ~37% of all MS2 spectra being assigned to a peptide due to its routine sample preparation and well curated protein sequence database. The significant decrease in assigned MS2 spectra between the *E. coli and R. palustris* sample was unexpected and could be the result of several factors. A likely scenario is that the protein sequence database for *R. palustris* is not as refined as the *E. coli* database, resulting in fewer spectrum-to-peptide matches. It is believed that the *R. palustris* sample was of sufficient quality, due to the large number of quality

unassigned MS2 spectra (discussed later). The groundwater samples provided an interesting and correlative look at the effects of both poor sample quality and the lack of a matching metagenome and highlights the differences in results between isolates and communities. The most successful groundwater analysis resulted in the assignment of 11,904 spectra, nearly 75% fewer spectra assigned to a peptide when compared to *E. coli.* The second and third groundwater samples were only able to assign 6,748 and 2,648 MS2 spectra to peptides. This is also fully reflected in the number of proteins identified: 1671, 760, and 61 respectively. Even though a relatively low number of MS2 assignments were made, the identification of 1,671 and 760 proteins indicates that even modest gains in the number of assigned MS2 spectra could result in significant gains in proteomic depth. The dramatic spread in assigned MS2 spectra, even with consistent experimental parameters, highlights differences that can be attributed to numerous causes. The most likely factor relates to the quality of the sample. For *E. coli*, the sample is a relatively pure isolate, suspended in MS compatible buffers and solvents. This is in stark contrast to the community samples that contain numerous species as well as extraneous organic compounds and harsh solution conditions. Additionally, the lack of a dedicated and matching protein sequence database significantly hinders the ability of SEQUEST to efficiently assign MS2 spectra to peptides. This is more accurately reflected in the number of quality unassigned MS2 spectra. Finally, the AMD sample illustrates the gains that are possible upon further refinement of the protein sequence database. DB1 resulted in 20,895 assigned MS2 whereas DB2 resulted in an additional 6,212 assigned spectra. The AMD sample is an excellent example of an experimentally optimized natural community sample with the number of assigned MS2 spectra comparable to the *R. palustris* isolate. The number of assigned AMD MS2 spectra is significantly increased over the groundwater samples and again highlights the necessity for a suitable protein sequence database.

7.4.3: Unassigned MS2 Spectra

Quantifying the number of MS2 spectra assigned to a peptide is direct representation of the sample preparation, experimental performance and ability to accurately assign

spectra to peptides in the protein sequence database. In a related but distinct manner, the number of unassigned MS2 spectra is an additional, useful diagnostic, and further highlights the differences between the analysis of isolates and communities. The unassigned MS2 spectra can be further sub-classified as poor quality or high quality (**Figure 7.3**). In general the number of unassigned MS2 correlates with the sample quality and protein sequence database, much like the assigned MS2. The unassigned MS2 spectra differ when they are further classified as either quality or poor. By including this metric, additional clues about the sample, operating conditions, and sequence database become apparent.

A significant amount of unassigned MS2 spectra were classified as poor. As detailed in the materials and methods, these spectra either did not contain sufficient fragment ions, sufficient intensity or carried a single charge (+1). The isolate samples (*E. coli* and *R. palustris*) were determined to serve as the baseline for comparison against the natural community samples. In *E. coli*, among the 112,350 total MS2 spectra collected, over 58,000 were classified as poor (51%). A similar proportion was observed for the *R. palustris* sample, 54,383 (~59%). Although not fully unexpected, it is obvious that much of the MS instrument time results in spectra that is largely unusable. This statistic increases in the natural community samples with a range of 67,741 – 85,476 poor unassigned MS2 spectra (**Figure 7.4**). There are several reasons for the increased numbers of poor unassigned MS2 spectra in the natural community samples, including an abundance of low intensity parent ions, sample contamination, or chromatography complexities. Among the criteria utilized to determine if a spectrum is quality or poor, the requirement for a charge state greater than one was chosen in order to produce a set of quality spectra that are more likely to be peptides. The proteolytic digestion by trypsin will, in most cases, produce peptides that contain a c-terminal 'K/R'. In addition to the c-terminus, the n-terminus is also a likely charge retaining location. Additionally, singly charged peptides will likely produce a smaller range of fragment ions due to the presence of uncharged fragments. The limiting number of fragments then results in MS2 spectra that contain insufficient data for a peptide assignment. Thus, it is more likely that desirable peptides will contain charges greater than one. It is possible

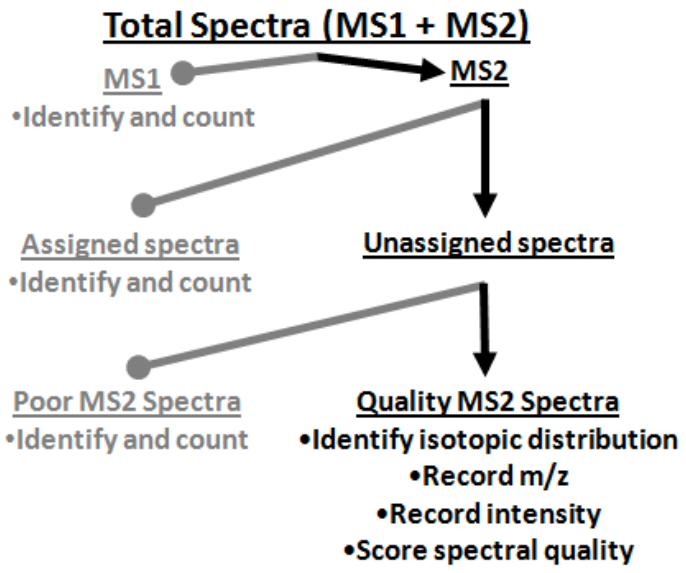

Figure 7.3: Classification of spectra in a MS experiment

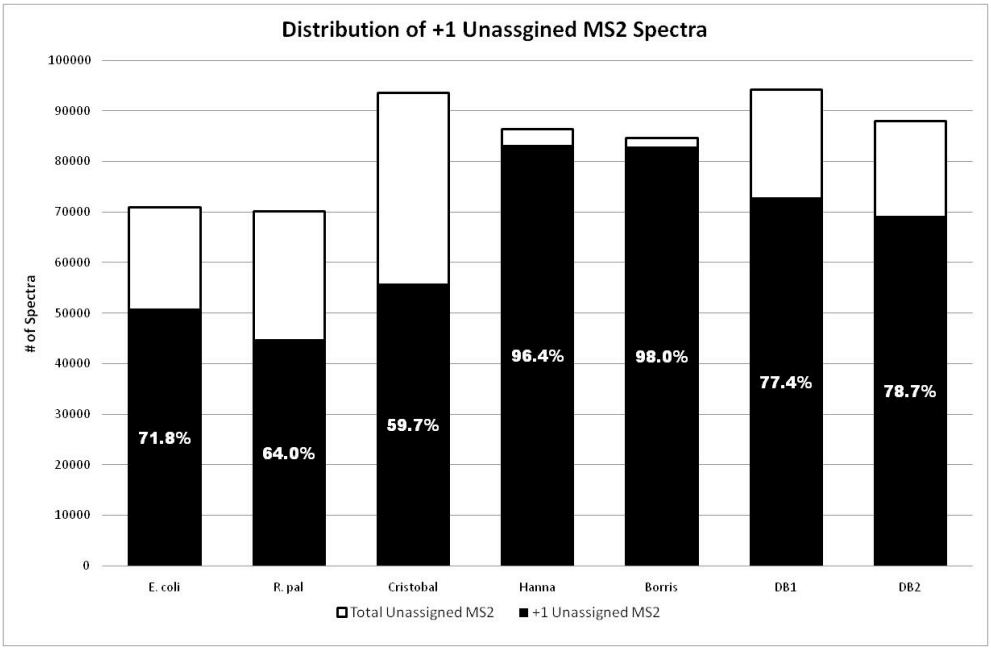

Figure 7.4: Distribution of +1 Charged, Unassigned MS2 Among Microbial Isolates and Communities

Percentages within each column represent the proportion of unassigned +1 MS2 for each sample.

that legitimate peptides contain a +1 charge, but the more conservative criteria was chosen in order to identify those spectra that are more likely peptides. **Figure 7.4** illustrates the number and proportion of unassigned MS2 spectra that carried a single charge. The isolate samples resulted in the fewest number of singly charged ions while two of the groundwater samples contained the largest number. Additionally, the isolate samples displayed among the lowest proportion of singly charged ions relative to all unassigned MS2 spectra (71% & 64% +1 MS2 of all unassigned MS2). Interestingly, the groundwater – Cristobal sample had a relatively low (when compared to the other natural community samples) number of singly charged ions: 55,924. This more closely resembled the distribution found in the isolate samples. Furthermore, with greater than 40% of the total unassigned MS2 spectra containing multiply charged ions, it is further apparent that a significant amount of valuable data is present and unused. The increase in not only the amount of unassigned MS2 spectra but the presence of greater numbers of singly charged ions in the natural samples may indicate preparatory or experimental optimization is necessary. Any reduction in the number of unassigned spectra, or more specifically, the number of singly charge ions will likely result in great gains in peptide assignment and ultimately protein identification.

The number of quality MS2 spectra varied from a high of 25,900 (Groundwater – Cristobal) to a low of 800 (Groundwater – Borris). The isolates and AMD sample resulted in ~11,000 – 16,000 quality unassigned MS2 spectra. These spectra represent quality data which, for a number of potential reasons, were not successfully assigned to a peptide. An obvious cause for the lack of a peptide assignment is a disconnect between the protein sequence database and the experimentally measured peptide. This disconnect could occur for several reasons including: incorrect (sequence polymorphism) or missing sequence in the protein sequence database, post-translation modification or chimeric MS2 spectra containing multiple fragmenting parents resulting convoluted spectra. The groundwater – Cristobal sample had over 25,000 quality unassigned spectra. This significant number again indicates that within the sample, numerous, multiply charged parents were isolated and fragmented and resulted in daughter spectra containing a range of fragment ions of sufficient intensity. It could be

concluded that the sample collection and preparation, as well as the measurement, proceeded successfully. It would then follow that the lack of peptide assignment could be likely based on the non-specific protein sequence database. On the other hand, the two remaining groundwater samples resulted in only 957 and 800 quality unassigned MS2 spectra. This significant loss in quality unassigned spectra (along with the low number of peptide assigned MS2 spectra) indicates that either the sample collection was poor, or the preparation failed. Based on repeated preparations, it could be concluded that the sample collection was not correct, resulting in insufficient peptide concentration. Finally, the AMD sample displayed a decrease in quality unassigned from DB1 to DB2, 12,702 to 10,067. This illustrates that as the protein sequence database was refined, previously unassigned quality spectra were successfully identified. Between the isolate and community samples, it appears the availability for a comprehensive protein sequence database is paramount if the sample collection proceeds as intended. The most significant factor affecting the number of quality unassigned was the availability of a suitable protein sequence database as evident in the groundwater – Cristobal sample, with 25,900 quality unassigned MS2 spectra, greater the twice that of *E. coli*, the designated baseline.

The high percentage of quality unassigned MS2s in community samples should not be left uncharacterized, as they could contribute significantly to or unravel parts of a complex proteome. Therefore, in conjunction with database search engines, *de novo* algorithms could be applied for the high quality unassigned spectra. This is especially useful for instances where a matched or sample derived genome or metagenome (ie, groundwater samples) is not available. One possible route is the submission of the entire dataset for *de novo* analysis, which requires significant amounts of computer time. This could include MS2 spectra of poor quality as well as the redundant assignment of spectra that had been previously identified by SEQUEST. An alternative approach utilizes only the unassigned quality MS2 spectra identified in this study. For comparison, the 25,900 unassigned quality MS2 spectra from the groundwater – Cristobal site were submitted to PepNovo for *de novo* interpretation.[151] Submission of all MS2 spectra (105,545) to PepNovo requires ~35 hours of computer time. By

reducing the set of submitted spectra to only the quality unassigned MS2 spectra, the *de novo* analysis only requires 8.5 hours, a 75% reduction in computer time. This significant reduction in compute time is a clear advantage of submitting only a subset of the collected MS2 spectra. Furthermore, this results in a culled results output containing only proposed sequences of high quality MS2 that had not been previously assigned to peptide. Analysis of the PepNovo output indicates that 99.8% (25,849 / 25,900) of the high quality spectra were *de novo* sequenced resulting in the prediction of over 15,000 novel peptides. For comparison, PepNovo analysis of all MS2 spectra from the second salt pulse, resulted in a decrease of assignable MS2, ~94% (9265 / 9772). This notable decrease in *de novo* sequenced MS2 can be attributed to the presence of poor quality MS2 spectra.

7.5: Conclusion

We have applied quantitative spectral analysis to microbial isolates and community MS proteomic results. By categorizing and quantifying the distribution of full scan and tandem mass spectra, a refined image of the experimental performance is obtained. This has resulted in a novel perspective of the effectiveness of the sample collection, preparation, experimental parameters, and database assignment, and will ultimately allow for specific optimizations in order to achieve greater proteomic depth. Key contrasts between isolate and community proteome samples are apparent when the quantity of assigned and unassigned spectra is compared. The *E. coli* isolate samples resulted in the highest number of assigned MS2 spectra while a natural groundwater microbial sample had the lowest. Multiple factors impact the ability to routinely identify spectra from the natural communities including sample collection and to a larger extent the accessibility of a match metagenome. This also largely influences the number of unassigned spectra. Remarkably, the natural samples resulted in a comparable number of quality unassigned MS2 spectra, with one particular groundwater sample exhibiting nearly double the number of spectra as the isolates, indicating that the sample collection and preparation were performing well. Identifying and targeting these unassigned quality spectra enabled the rapid application of de novo peptide

assignment and resulted in an additional 15,000 high scoring peptide identifications. The spectral analysis has highlighted several areas that must be optimized in order to gain a desired completeness of proteomic depth in natural community samples.

Chapter 8

Conclusions and Significance of the Characterization of the Extracellular Fraction in a Natural Microbial Community

---

8.1: Conclusions

The research presented within this dissertation represents a comprehensive characterization of the extracellular proteome from an extremophilic microbial community. The integration of *experimental* mass spectrometric and *computational* bioinformatic approaches has resulted in the design and application of novel methodologies resulting in the identification and characterization of novel proteins. Additionally, the creation of two software platforms has enabled a rapid integration of top down and bottom up data sets, and a new metric for determining the effectiveness of bottom up analysis of complex proteomes. The use of the developed methodologies and new software tools, as well as the novel protein identifications; provide the groundwork for the ever increasing characterization of the AMD microbial community in particular, and other natural microbial communities in general.

In order to more comprehensively identify and ultimately characterize community proteomes, it is evident that further integration and focused analyses are necessary. The application of computational signal peptide prediction with experimental MS identification illustrates the potential to utilize historically genomic tools for proteomic analysis. The reliance and availability of genomic information appears to be a rather untapped resource for proteomic characterization. The application of genomic tools and optimization of the genome will result in vastly more proteomic information. A number of genomic tools, including operon prediction, gene structure analysis and additional gene function prediction will result in a significantly improved genome. This directly translates into increased proteomic identifications with more descriptive functional annotations. Recent work has resulted in genome curation through MS peptide identification.[38] Although effective for generating a more representative genome, this

does not necessarily result in proteomic gains, as this builds upon existing protein identifications. Current efforts to improve the metagenome of microbial communities have shown that alterations in the sequencing, assembly, and annotation provide significant gains in proteomic analyses.

Deep proteomic coverage is necessary in order to fully understand and characterize the unique pathways, interactions, and partitioning that occurs in microbial communities. Direct analysis of subcellular fractions has increased the depth of proteomic coverage. However, the issues of dynamic range will continue to hamper the identification of low abundance proteins. Further separation techniques or combinations of techniques will be necessary in order to achieve the deep proteomic coverage necessary. The use of metal affinity column chromatography presents one possible route of increasing the dynamic range of MS based proteomics. This methodology has established that a combination of cellular fractionation and an affinity based enrichment can result in novel identifications. Alternative methods or platforms, including gel based fractionation also show great promise for the depth of proteomic coverage.

Although MS based proteomics is highly applicable to the rapid identification of large protein samples, the use of more specialized mass spectrometric instrumentation for targeted analyses is often overlooked. The characterization of a novel cytochrome from the AMD microbial community illustrates the potential for more targeted applications of high resolution instrumentation (FTICR). Initial experimental analyses utilizing Edman degradation provided a preliminary view of the now characterized n-terminal truncation; however, this approach was not able to fully characterize the truncation, nor was it able to provide the specific sequence tag buried within the middle of the protein. Both of these key data points were provided through the FTICR-MS analysis. The specific targets generated through MS analyses should be considered for further characterization with the use of advanced MS instrumentation. Details of the intact protein, including truncations and PTMs are readily measured. This can provide yet another layer of evidence for determining the functions of the numerous unknown proteins in the AMD the community. Furthermore, top down identification and characterization remains an under-utilized tool in community proteomics. Determination

of the depth and breadth of protein modification has been largely a targeted focus. The availability of instrumentation that allows for the rapid measurement of complex intact protein samples should be better integrated into the proteomics pipeline. A sizeable amount of optimization is still required for the direct analysis of complex intact protein samples but considering the advancement of robust fractionation techniques it is not unreasonable to include top down analyses in proteomic characterizations.

8.2: Optimization of MS based proteomics through spectral analysis

The use of spectral analysis has shed light on several metrics of proteomic measurements that were previously unnoticed. The ability to chart and categorize the spectra collected during a MS experiment provides real-time feedback on the effectiveness of the experiment. Quantifying the different types of collected spectra (MS1, assigned / unassigned MS2) reveals specific data points about the effectiveness of the sample preparation, chromatographic separation, instrumental setup and database assignment. In an effort to increase the proteomic depth, a significant amount of focus has been directed towards the optimization of sample preparation and advancements in MS instrumentation. What the spectral analysis has revealed is that current methodologies, although not without faults, are generating a significant portion of high quality spectra. Subtle optimizations in the chromatography, based on the results of the spectral analysis, could provide reduced instrument time, allowing for increased experimental replication and gains in statistical confidence. This also indicates that additional development should be directed towards increasing the portion of quality spectra that are assigned to a peptide.

Spectral analysis can be used to specifically track where in the chromatographic elution the majority of quality spectra are present. The use of a constant graduated elution generates several regions in the chromatogram where a significant portion of the spectra are poor quality and generate no useable data. Further optimizing the chromatography in order to prolong the region where high quality spectra are typically eluted and minimizing the region where low quality spectra are eluted could result in significant gains in the amount of spectra assigned to a peptide. This, of course,

directly relates to achieving a complete proteomic identification. Additionally, the chromatographic optimization could significantly reduce portions of the collection time, providing the potential for replicated MS analyses.

A final, and significant metric, is that a large portion of collected, quality spectra remain unassigned to peptides. The unassigned spectra essentially represent useful data that is unused. It is obvious, from this point alone, that specific focus should be given to the assignment of these unassigned spectra. Optimization of the sample preparation, chromatography or instrument parameters will likely result in gains in proteins identified, but a subset of the collected spectra, even with the mentioned optimizations will remain unassigned. This presents a significant and formidable challenge for MS based proteomics as there are a number of potential reasons for a high quality spectrum to remain unassigned. The use of *de novo* assignment is one potential route for assigning peptide sequences to these unassigned spectra. Currently, the significant amount of processor time required for *de novo* assignment of the hundreds of thousands of spectra limits its usefulness. By only identifying and then submitting the high quality unassigned spectra to *de novo* search algorithms, significant time is saved and the resulting peptide identifications are complementary to those peptides identified from the traditional database search. Targeting the thousands of unassigned spectra will result in tremendous gains in proteomic coverage, without the need for additional experimental optimization.

8.3: Generation of specific proteins for targeted analysis

Ultimately the fundamental goal of MS based proteomics is the generation of biological inferences. That is, can MS based proteomics provide tangible biological results? One of the most overlooked aspects is that within the thousands of identified proteins from a given sample, a number of specific targets are identified that are ideal for future analysis. This subtle, but critical point is often missed among the impressive amount of data generated through the analyses. The ability to hone in on several proteins out of hundreds of thousands truly illustrates the benefits of these methodologies. An excellent example of this is the identification, description, and

ultimately characterization of a novel cytochrome believed to be a key member in electron transport in the AMD system. The initial identification of a highly abundant protein led to the realization that multiple n-terminal sequences existed. Targeted analysis through high performance mass spectrometry led to an unambiguous identification and resulted in the correlation between n-terminal truncation and biofilm growth state. An additional example lies in the enrichment of proteins through metal affinity columns which resulted in the identification of over one hundred previously unidentified proteins. The low abundance of these proteins, along with the presence of high scoring domains and motifs presents these proteins as excellent targets for specific biochemical analysis. The identification of these proteins through the MS methodologies provides evidence that these, once predicted gene products, are in fact expressed and are viable targets to pursue. By targeting a subset of the proteins identified by MS, who are noted for a particular sequence or functional characteristic, it is possible to rapidly and efficiently describe these proteins with the goal of describing the biological pathways from which they represent.

8.4: Future directions

A progression towards additional integration of advanced bioinformatic principles with the analytical measurements will result in significant gains in proteomic characterization. Currently, the computational aspect of MS based proteomics is witnessing a surge of interest, with more and more focus directed towards developing more rapid algorithms that are capable of handling the immense datasets. A serious challenge exists though in the ability to efficiently integrate the new informatics approaches with the multitude of experimental data generated. As the data becomes more and more proprietary, due to data file size constraints, the informatics tools must be adept at handling and interpreting these formats. Currently, this step is often a challenge for data analysis, whereby raw data is required to be converted to plain text for analysis. Additionally, as the sample sets become increasingly complicated, the datasets are too becoming too large to efficiently interpret. Specific focus must be applied to create tools that are able to present data efficiently to the informatics

algorithms for analysis and then to efficiently present the results to the researcher for interpretation. One potential area of integration could be through the use of computational databases that utilize a user friendly front-end for interaction. This would result in an efficient manner in which to store both the experimentally derived data and the results of the data following analysis. This would also create a means to rapidly interrogate the results of not only one MS analysis but a catalog of archived data.

From an experimental point of view, the lack of top down integration is a fundamental are of mass spectrometry that must be included in future proteomic analyses. As the ability to reduce the complexity of community microbial samples increases, the applicability of TD analyses follows. The direct analysis of the intact proteins provides specific hooks towards biological function that is unseen with the more adapted bottom up methodologies. The current and next generation of MS instruments are capable of measuring the range of masses present in an sample containing intact proteins with both high accuracy and resolution. Although the chromatographic separation of intact proteins will remain an inherent obstacle, the optimization of new separation techniques, such as Gel-Eluted Liquid Fraction Entrapment Electrophoresis (GELFrEE), will result in complex reduced samples that are amenable to top down analyses.[155] Through the integration of top down and bottom data sets, the suite, range and frequency of post-translational modifications can be efficiently determined.

8.5 Perspective

The research presented here provides a comprehensive proteomic characterization of an extracellular fraction from a microbial community. The challenges inherent with complex proteomic analyses have resulted in the coupling of high performance mass spectrometry along with the design, integration, and application of bioinformatic algorithms. The advancements in technological integration provide a suitable pathway to begin to target more complex microbial communities. The proteomic results presented within provide a solid foundation for targeted biochemical analyses. For example, identified proteins exhibiting dramatic expression changes that correlate with the biofilm developmental state or the TonB like protein exhibiting no

marked affinity for Fe but displaying limited sequence homology to known TonB proteins are excellent examples of target proteins from the hundreds of thousands that exist within the community. This work represents significant progress towards the ultimate goal of a complete identification and eventual characterization of the AMD microbial system.

Over the last five years microbial proteomics has shown remarkable progress. The transition from reporting summaries of protein identifications of cultivable, single isolate microorganisms, to global analyses of increasingly complex natural microbial communities has not halted the advancement and application of biological mass spectrometry. Over the next five years it is not unrealistic to imagine that complex community analyses will be considered as routine as isolate proteomics is currently considered. Intense research devoted to several of the avenues described above will raise the standards for MS based proteomics yet again. It is obvious that current proteomic discoveries do not present biological dead ends, but instead create even more paths for targeted research. One potential front that may pave the way for the next leap forward for proteomics is the advancements in metagenome sequencing. Next-generation sequencing technologies are promising greater accuracy with increased sequence coverage, reduced costs, and, higher throughput. A more accurate metagenome, from deeper sequence coverage and improved assembly, translates into increased spectrum to peptide assignments. With additional metagenomic sequence coverage of highly diverse microbial communities, from decreased costs and increased throughput, a greater pool of information will be available for sequence-based contrasts and comparisons. Assuming these metrics can be met; metaproteomics will stand to gain tremendously. As the ultimate goal is a gain in biological insight, continued focus on mining biological inferences from integrated metagenomic-metaproteomic datasets will solidify mass spectrometry as the foundation for cutting edge environmental proteomics.

REFERENCES

1. Domon, B.; Aebersold, R., Mass spectrometry and protein analysis. *Science* **2006,** 312, (5771), 212-7.
2. Cravatt, B. F.; Simon, G. M.; Yates, J. R., 3rd, The biological impact of mass-spectrometry-based proteomics. *Nature* **2007,** 450, (7172), 991-1000.
3. Aebersold, R.; Mann, M., Mass spectrometry-based proteomics. *Nature* **2003,** 422, (6928), 198-207.
4. Pandey, A.; Mann, M., Proteomics to study genes and genomes. *Nature* **2000,** 405, (6788), 837-46.
5. Wilmes, P.; Bond, P. L., Metaproteomics: studying functional gene expression in microbial ecosystems. *Trends Microbiol* **2006,** 14, (2), 92-7.
6. VerBerkmoes, N. C.; Denef, V. J.; Hettich, R. L.; Banfield, J. F., Systems biology: Functional analysis of natural microbial consortia using community proteomics. *Nat Rev Microbiol* **2009,** 7, (3), 196-205.
7. DeLong, E. F., Microbial community genomics in the ocean. *Nat Rev Microbiol* **2005,** 3, (6), 459-69.
8. Kent, A. D.; Triplett, E. W., Microbial communities and their interactions in soil and rhizosphere ecosystems. *Annu Rev Microbiol* **2002,** 56, 211-36.
9. Verberkmoes, N. C.; Russell, A. L.; Shah, M.; Godzik, A.; Rosenquist, M.; Halfvarson, J.; Lefsrud, M. G.; Apajalahti, J.; Tysk, C.; Hettich, R. L.; Jansson, J. K., Shotgun metaproteomics of the human distal gut microbiota. *ISME J* **2009,** 3, (2), 179-89.
10. O'Farrell, P. H., High resolution two-dimensional electrophoresis of proteins. *J Biol Chem* **1975,** 250, (10), 4007-21.
11. Gorg, A.; Obermaier, C.; Boguth, G.; Harder, A.; Scheibe, B.; Wildgruber, R.; Weiss, W., The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis* **2000,** 21, (6), 1037-53.
12. Peng, J.; Gygi, S. P., Proteomics: the move to mixtures. *J Mass Spectrom* **2001,** 36, (10), 1083-91.
13. Balogh, M. P., Debating resolution and mass accuracy. *Lc Gc North America* **2004,** 22, (2), 118-+.
14. Yates, J. R.; Ruse, C. I.; Nakorchevsky, A., Proteomics by mass spectrometry: approaches, advances, and applications. *Annu Rev Biomed Eng* **2009,** 11, 49-79.
15. Liu, H.; Lin, D.; Yates, J. R., 3rd, Multidimensional separations for protein/peptide analysis in the post-genomic era. *Biotechniques* **2002,** 32, (4), 898, 900, 902 passim.
16. VerBerkmoes, N. C.; Connelly, H. M.; Pan, C.; Hettich, R. L., Mass spectrometric approaches for characterizing bacterial proteomes. *Expert Rev Proteomics* **2004,** 1, (4), 433-47.
17. Washburn, M. P.; Wolters, D.; Yates, J. R., 3rd, Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* **2001,** 19, (3), 242-7.
18. Florens, L.; Carozza, M. J.; Swanson, S. K.; Fournier, M.; Coleman, M. K.; Workman, J. L.; Washburn, M. P., Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors. *Methods* **2006,** 40, (4), 303-11.
19. Ram, R. J.; Verberkmoes, N. C.; Thelen, M. P.; Tyson, G. W.; Baker, B. J.; Blake, R. C., 2nd; Shah, M.; Hettich, R. L.; Banfield, J. F., Community proteomics of a natural microbial biofilm. *Science* **2005,** 308, (5730), 1915-20.

20. Tyson, G. W.; Chapman, J.; Hugenholtz, P.; Allen, E. E.; Ram, R. J.; Richardson, P. M.; Solovyev, V. V.; Rubin, E. M.; Rokhsar, D. S.; Banfield, J. F., Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **2004,** 428, (6978), 37-43.
21. Siuti, N.; Kelleher, N. L., Decoding protein modifications using top-down mass spectrometry. *Nat Methods* **2007,** 4, (10), 817-21.
22. Kelleher, N. L.; Lin, H. Y.; Valaskovic, G. A.; Aaserud, D. J.; Fridriksson, E. K.; McLafferty, F. W., Top down versus bottom up protein characterization by tandem high-resolution mass spectrometry. *Journal of the American Chemical Society* **1999,** 121, (4), 806-812.
23. Ryan, C. M.; Souda, P.; Bassilian, S.; Ujwal, R.; Zhang, J.; Abramson, J.; Ping, P.; Durazo, A.; Bowie, J. U.; Hasan, S. S.; Baniulis, D.; Cramer, W. A.; Faull, K. F.; Whitelegge, J. P., Post-translational modifications of integral membrane proteins resolved by top-down Fourier transform mass spectrometry with collisionally activated dissociation. *Mol Cell Proteomics* **2010,** 9, (5), 791-803.
24. Thangaraj, B.; Ryan, C. M.; Souda, P.; Krause, K.; Faull, K. F.; Weber, A. P.; Fromme, P.; Whitelegge, J. P., Data-directed top-down Fourier-transform mass spectrometry of a large integral membrane protein complex: Photosystem II from Galdieria sulphuraria. *Proteomics* **2010**.
25. Yates, J. R., 3rd; Eng, J. K.; McCormack, A. L.; Schieltz, D., Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal Chem* **1995,** 67, (8), 1426-36.
26. Tabb, D. L.; Fernando, C. G.; Chambers, M. C., MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J Proteome Res* **2007,** 6, (2), 654-61.
27. Payne, S. H.; Yau, M.; Smolka, M. B.; Tanner, S.; Zhou, H.; Bafna, V., Phosphorylation-specific MS/MS scoring for rapid and accurate phosphoproteome analysis. *J Proteome Res* **2008,** 7, (8), 3373-81.
28. Whitman, W. B.; Coleman, D. C.; Wiebe, W. J., Prokaryotes: the unseen majority. *Proc Natl Acad Sci U S A* **1998,** 95, (12), 6578-83.
29. Pace, N. R., A molecular view of microbial diversity and the biosphere. *Science* **1997,** 276, (5313), 734-40.
30. Wood, T. K., Molecular approaches in bioremediation. *Curr Opin Biotechnol* **2008,** 19, (6), 572-8.
31. Wilmes, P.; Simmons, S. L.; Denef, V. J.; Banfield, J. F., The dynamic genetic repertoire of microbial communities. *FEMS Microbiol Rev* **2009,** 33, (1), 109-32.
32. Schrenk, M. O.; Edwards, K. J.; Goodman, R. M.; Hamers, R. J.; Banfield, J. F., Distribution of thiobacillus ferrooxidans and leptospirillum ferrooxidans: implications for generation of acid mine drainage. *Science* **1998,** 279, (5356), 1519-22.
33. Wilmes, P.; Remis, J. P.; Hwang, M.; Auer, M.; Thelen, M. P.; Banfield, J. F., Natural acidophilic biofilm communities reflect distinct organismal and functional organization. *ISME J* **2009,** 3, (2), 266-70.
34. Yates, J. R., 3rd, Mass spectral analysis in proteomics. *Annu Rev Biophys Biomol Struct* **2004,** 33, 297-316.

35. Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M., Electrospray ionization for mass spectrometry of large biomolecules. *Science* **1989,** 246, (4926), 64-71.
36. Marshall, A. G.; Hendrickson, C. L.; Jackson, G. S., Fourier transform ion cyclotron resonance mass spectrometry: a primer. *Mass Spectrom Rev* **1998,** 17, (1), 1-35.
37. Douglas, D. J.; Frank, A. J.; Mao, D., Linear ion traps in mass spectrometry. *Mass Spectrom Rev* **2005,** 24, (1), 1-29.
38. Denef, V. J.; VerBerkmoes, N. C.; Shah, M. B.; Abraham, P.; Lefsrud, M.; Hettich, R. L.; Banfield, J. F., Proteomics-inferred genome typing (PIGT) demonstrates inter-population recombination as a strategy for environmental adaptation. *Environmental Microbiology* **2009,** 11, (2), 313-325.
39. Pluskal, M. G., Microscale sample preparation. *Nat Biotechnol* **2000,** 18, (1), 104-5.
40. Strader, M. B.; Tabb, D. L.; Hervey, W. J.; Pan, C.; Hurst, G. B., Efficient and specific trypsin digestion of microgram to nanogram quantities of proteins in organic-aqueous solvent systems. *Anal Chem* **2006,** 78, (1), 125-34.
41. Chourey, K.; Thompson, M. R.; Morrell-Falvey, J.; Verberkmoes, N. C.; Brown, S. D.; Shah, M.; Zhou, J.; Doktycz, M.; Hettich, R. L.; Thompson, D. K., Global molecular and morphological effects of 24-hour chromium(VI) exposure on Shewanella oneidensis MR-1. *Appl Environ Microbiol* **2006,** 72, (9), 6331-44.
42. VerBerkmoes, N. C.; Shah, M. B.; Lankford, P. K.; Pelletier, D. A.; Strader, M. B.; Tabb, D. L.; McDonald, W. H.; Barton, J. W.; Hurst, G. B.; Hauser, L.; Davison, B. H.; Beatty, J. T.; Harwood, C. S.; Tabita, F. R.; Hettich, R. L.; Larimer, F. W., Determination and comparison of the baseline proteomes of the versatile microbe Rhodopseudomonas palustris under its major metabolic states. *J Proteome Res* **2006,** 5, (2), 287-98.
43. Thompson, M. R.; Chourey, K.; Froelich, J. M.; Erickson, B. K.; VerBerkmoes, N. C.; Hettich, R. L., Experimental approach for deep proteome measurements from small-scale microbial biomass samples. *Anal Chem* **2008,** 80, (24), 9517-25.
44. Sandra, K.; Moshir, M.; D'Hondt, F.; Verleysen, K.; Kas, K.; Sandra, P., Highly efficient peptide separations in proteomics Part 1. Unidimensional high performance liquid chromatography. *J Chromatogr B Analyt Technol Biomed Life Sci* **2008,** 866, (1-2), 48-63.
45. Sandra, K.; Moshir, M.; D'Hondt, F.; Tuytten, R.; Verleysen, K.; Kas, K.; Francois, I.; Sandra, P., Highly efficient peptide separations in proteomics. Part 2: bi- and multidimensional liquid-based separation techniques. *J Chromatogr B Analyt Technol Biomed Life Sci* **2009,** 877, (11-12), 1019-39.
46. Shen, Y.; Tolic, N.; Zhao, R.; Pasa-Tolic, L.; Li, L.; Berger, S. J.; Harkewicz, R.; Anderson, G. A.; Belov, M. E.; Smith, R. D., High-throughput proteomics using high-efficiency multiple-capillary liquid chromatography with on-line high-performance ESI FTICR mass spectrometry. *Anal Chem* **2001,** 73, (13), 3011-21.
47. Johnson, J. R.; Meng, F.; Forbes, A. J.; Cargile, B. J.; Kelleher, N. L., Fourier-transform mass spectrometry for automated fragmentation and identification of 5-20 kDa proteins in mixtures. *Electrophoresis* **2002,** 23, (18), 3217-23.
48. Durbin, K. R.; Tran, J. C.; Zamdborg, L.; Sweet, S. M.; Catherman, A. D.; Lee, J. E.; Li, M.; Kellie, J. F.; Kelleher, N. L., Intact mass detection, interpretation, and visualization to automate Top-Down proteomics on a large scale. *Proteomics* **2010**.

49. Bogdanov, B.; Smith, R. D., Proteomics by FTICR mass spectrometry: top down and bottom up. *Mass Spectrom Rev* **2005,** 24, (2), 168-200.
50. Schwartz, J. C.; Senko, M. W.; Syka, J. E., A two-dimensional quadrupole ion trap mass spectrometer. *J Am Soc Mass Spectrom* **2002,** 13, (6), 659-69.
51. Liebler, D. C., Shotgun mass spec goes independent. *Nat Methods* **2004,** 1, (1), 16-7.
52. 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., Towards defining the urinary proteome using liquid chromatography-tandem mass spectrometry. II. Limitations of complex mixture analyses. *Proteomics* **2001,** 1, (1), 108-17.
53. Zhang, Y.; Wen, Z.; Washburn, M. P.; Florens, L., Effect of dynamic exclusion duration on spectral count based quantitative proteomics. *Anal Chem* **2009,** 81, (15), 6317-26.
54. Rapoport, T. A., Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes. *Nature* **2007,** 450, (7170), 663-9.
55. Gierasch, L. M., Signal sequences. *Biochemistry* **1989,** 28, (3), 923-30.
56. von Heijne, G., The signal peptide. *J Membr Biol* **1990,** 115, (3), 195-201.
57. Driessen, A. J.; Manting, E. H.; van der Does, C., The structural basis of protein targeting and translocation in bacteria. *Nat Struct Biol* **2001,** 8, (6), 492-8.
58. Nair, R.; Rost, B., Sequence conserved for subcellular localization. *Protein Sci* **2002,** 11, (12), 2836-47.
59. McGeoch, D. J., On the predictive recognition of signal peptide sequences. *Virus Res* **1985,** 3, (3), 271-86.
60. Reinhardt, A.; Hubbard, T., Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res* **1998,** 26, (9), 2230-6.
61. Zhang, Z.; Wood, W. I., A profile hidden Markov model for signal peptides generated by HMMER. *Bioinformatics* **2003,** 19, (2), 307-8.
62. Nielsen, H.; Krogh, A., Prediction of signal peptides and signal anchors by a hidden Markov model. *Proc Int Conf Intell Syst Mol Biol* **1998,** 6, 122-30.
63. Bendtsen, J. D.; Nielsen, H.; von Heijne, G.; Brunak, S., Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* **2004,** 340, (4), 783-95.
64. Yates, J. R., 3rd; Eng, J. K.; McCormack, A. L., Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Anal Chem* **1995,** 67, (18), 3202-10.
65. Sadygov, R. G.; Cociorva, D.; Yates, J. R., 3rd, Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat Methods* **2004,** 1, (3), 195-202.
66. Bantscheff, M.; Schirle, M.; Sweetman, G.; Rick, J.; Kuster, B., Quantitative mass spectrometry in proteomics: a critical review. *Anal Bioanal Chem* **2007,** 389, (4), 1017-31.
67. Allen, E. E.; Banfield, J. F., Community genomics in microbial ecology and evolution. *Nat Rev Microbiol* **2005,** 3, (6), 489-98.
68. Goltsman, D. S.; Denef, V. J.; Singer, S. W.; Verberkmoes, N. C.; Lefsrud, M.; Mueller, R.; Dick, G. J.; Sun, C.; Wheeler, K.; Zemla, A.; Baker, B. J.; Hauser, L.; Land, M.; Shah, M. B.; Thelen, M. P.; Hettich, R. L.; Banfield, J. F., Community genomic and proteomic analysis of chemoautotrophic, iron-oxidizing "Leptospirillum rubarum" (Group II) and Leptospirillum ferrodiazotrophum (Group III) in acid mine drainage biofilms. *Appl Environ Microbiol* **2009**.

69. Lo, I.; Denef, V. J.; Verberkmoes, N. C.; Shah, M. B.; Goltsman, D.; DiBartolo, G.; Tyson, G. W.; Allen, E. E.; Ram, R. J.; Detter, J. C.; Richardson, P.; Thelen, M. P.; Hettich, R. L.; Banfield, J. F., Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. *Nature* **2007,** 446, (7135), 537-41.
70. Nielsen, H.; Engelbrecht, J.; Brunak, S.; von Heijne, G., Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng* **1997,** 10, (1), 1-6.
71. Tabb, D. L.; McDonald, W. H.; Yates, J. R., 3rd, DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J Proteome Res* **2002,** 1, (1), 21-6.
72. Elias, J. E.; Gygi, S. P., Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* **2007,** 4, (3), 207-14.
73. Bateman, A.; Coin, L.; Durbin, R.; Finn, R. D.; Hollich, V.; Griffiths-Jones, S.; Khanna, A.; Marshall, M.; Moxon, S.; Sonnhammer, E. L.; Studholme, D. J.; Yeats, C.; Eddy, S. R., The Pfam protein families database. *Nucleic Acids Res* **2004,** 32, (Database issue), D138-41.
74. Finn, R. D.; Mistry, J.; Schuster-Bockler, B.; Griffiths-Jones, S.; Hollich, V.; Lassmann, T.; Moxon, S.; Marshall, M.; Khanna, A.; Durbin, R.; Eddy, S. R.; Sonnhammer, E. L.; Bateman, A., Pfam: clans, web tools and services. *Nucleic Acids Res* **2006,** 34, (Database issue), D247-51.
75. Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J., Basic local alignment search tool. *J Mol Biol* **1990,** 215, (3), 403-10.
76. Singer, S. W.; Chan, C. S.; Zemla, A.; Verberkmoes, N. C.; Hwang, M.; Hettich, R. L.; Banfield, J. F.; Thelen, M. P., Characterization of Cytochrome 579, an unusual cytochrome isolated from an iron oxidizing microbial community. *Appl Environ Microbiol* **2008**.
77. Zybailov, B.; Mosley, A. L.; Sardiu, M. E.; Coleman, M. K.; Florens, L.; Washburn, M. P., Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. *J Proteome Res* **2006,** 5, (9), 2339-47.
78. de Hoon, M. J.; Imoto, S.; Nolan, J.; Miyano, S., Open source clustering software. *Bioinformatics* **2004,** 20, (9), 1453-4.
79. Saldanha, A. J., Java Treeview--extensible visualization of microarray data. *Bioinformatics* **2004,** 20, (17), 3246-8.
80. Tusher, V. G.; Tibshirani, R.; Chu, G., Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* **2001,** 98, (9), 5116-21.
81. Krogh, A.; Larsson, B.; von Heijne, G.; Sonnhammer, E. L., Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **2001,** 305, (3), 567-80.
82. Druschel, G. K.; Baker, B. J.; Gihring, T. M.; Banfield, J. F., Acid mine drainage biogeochemistry at Iron Mountain, California. *Geochemical Transactions* **2004,** 5, (2), 13-32.
83. Tjalsma, H.; Antelmann, H.; Jongbloed, J. D.; Braun, P. G.; Darmon, E.; Dorenbos, R.; Dubois, J. Y.; Westers, H.; Zanen, G.; Quax, W. J.; Kuipers, O. P.; Bron, S.; Hecker, M.; van Dijl, J. M., Proteomics of protein secretion by Bacillus subtilis: separating the "secrets" of the secretome. *Microbiol Mol Biol Rev* **2004,** 68, (2), 207-33.
84. Seiki, M., Membrane-type matrix metalloproteinases. *Apmis* **1999,** 107, (1), 137-43.
85. Husten, E. J.; Eipper, B. A., The membrane-bound bifunctional peptidylglycine alpha-amidating monooxygenase protein. Exploration of its domain structure through limited proteolysis. *J Biol Chem* **1991,** 266, (26), 17004-10.

86. Xia, Z.; Dai, W.; Zhang, Y.; White, S. A.; Boyd, G. D.; Mathews, F. S., Determination of the gene sequence and the three-dimensional structure at 2.4 angstroms resolution of methanol dehydrogenase from Methylophilus W3A1. *J Mol Biol* **1996,** 259, (3), 480-501.
87. Fulop, V.; Ridout, C. J.; Greenwood, C.; Hajdu, J., Crystal structure of the di-haem cytochrome c peroxidase from Pseudomonas aeruginosa. *Structure* **1995,** 3, (11), 1225-33.
88. Stancik, L. M.; Stancik, D. M.; Schmidt, B.; Barnhart, D. M.; Yoncheva, Y. N.; Slonczewski, J. L., pH-dependent expression of periplasmic proteins and amino acid catabolism in Escherichia coli. *J Bacteriol* **2002,** 184, (15), 4246-58.
89. Singer, S. W.; Erickson, B. K.; Verberkmoes, N. C.; Hwang, M.; Shah, M. B.; Hettich, R. L.; Banfield, J. F.; Thelen, M. P., Posttranslational modification and sequence variation of redox-active proteins correlate with biofilm life cycle in natural microbial communities. *ISME J* **2010**.
90. Hou, S. B.; Makarova, K. S.; Saw, J. H. W.; Senin, P.; Ly, B. V.; Zhou, Z. M.; Ren, Y.; Wang, J. M.; Galperin, M. Y.; Omelchenko, M. V.; Wolf, Y. I.; Yutin, N.; Koonin, E. V.; Stott, M. B.; Mountain, B. W.; Crowe, M. A.; Smirnova, A. V.; Dunfield, P. F.; Feng, L.; Wang, L.; Alam, M., Complete genome sequence of the extremely acidophilic methanotroph isolate V4, Methylacidiphilum infernorum, a representative of the bacterial phylum Verrucomicrobia. *Biology Direct* **2008,** 3, -.
91. Futterer, O.; Angelov, A.; Liesegang, H.; Gottschalk, G.; Schleper, C.; Schepers, B.; Dock, C.; Antranikian, G.; Liebl, W., Genome sequence of Picrophilus torridus and its implications for life around pH 0. *Proceedings of the National Academy of Sciences of the United States of America* **2004,** 101, (24), 9091-9096.
92. Schafer, K.; Magnusson, U.; Scheffel, F.; Schiefner, A.; Sandgren, M. O. J.; Diederichs, K.; Welte, W.; Hulsmann, A.; Schneider, E.; Mowbray, S. L., X-ray structures of the maltose-maltodextrin-binding protein of the thermoacidophilic bacterium Alicyclobacillus acidocaldarius provide insight into acid stability of proteins. *Journal of Molecular Biology* **2004,** 335, (1), 261-274.
93. Urh, M.; Simpson, D.; Zhao, K.; Richard, R. B. a. M. P. D., Chapter 26 Affinity Chromatography: General Methods. In *Methods in Enzymology*, Academic Press: 2009; Vol. Volume 463, pp 417-438.
94. Boschetti, E.; Righetti, P. G., The ProteoMiner in the proteomic arena: A non-depleting tool for discovering low-abundance species. *Journal of Proteomics* **2008,** 71, (3), 255-264.
95. Vanitha, T.; Shanhua, L.; Liliana, G.; Julia, L.; Lee, L.; David, H.; Egisto, B., Reduction of the concentration difference of proteins in biological liquids using a library of combinatorial ligands. *ELECTROPHORESIS* **2005,** 26, (18), 3561-3571.
96. Krapfenbauer, K.; Fountoulakis, M., Improved Enrichment and Proteomic Analysis of Brain Proteins with Signaling Function by Heparin Chromatography. In *Neuroproteomics*, 2009; pp 165-180.
97. McDonald, C. A.; Yang, J. Y.; Marathe, V.; Yen, T.-Y.; Macher, B. A., Combining Results from Lectin Affinity Chromatography and Glycocapture Approaches Substantially Improves the Coverage of the Glycoproteome. *Molecular & Cellular Proteomics* **2009,** 8, (2), 287-301.
98. Arnold, F. H., Metal-affinity separations: a new dimension in protein processing. *Biotechnology (N Y)* **1991,** 9, (2), 151-6.

99. Juan, L.-B.; José Luis, G.-A., Environmental proteomics and metallomics. *PROTEOMICS* **2006,** 6, (S1), S51-S62.
100. Kulkarni, P. P.; She, Y. M.; Smith, S. D.; Roberts, E. A.; Sarkar, B., Proteomics of metal transport and metal-associated diseases. *Chemistry* **2006,** 12, (9), 2410-22.
101. Mounicou, S.; Szpunar, J.; Lobinski, R., Metallomics: the concept and methodology. *Chemical Society Reviews* **2009,** 38, (4), 1119-1138.
102. She, Y. M.; Narindrasorasak, S.; Yang, S.; Spitale, N.; Roberts, E. A.; Sarkar, B., Identification of metal-binding proteins in human hepatoma lines by immobilized metal affinity chromatography and mass spectrometry. *Mol Cell Proteomics* **2003,** 2, (12), 1306-18.
103. Smith, S. D.; She, Y. M.; Roberts, E. A.; Sarkar, B., Using immobilized metal affinity chromatography, two-dimensional electrophoresis and mass spectrometry to identify hepatocellular proteins with copper-binding ability. *J Proteome Res* **2004,** 3, (4), 834-40.
104. Druschel, G.; Baker, B.; Gihring, T.; Banfield, J., Acid mine drainage biogeochemistry at Iron Mountain, California. *Geochemical Transactions* **2004,** 5, (2), 13.
105. Erickson, B. K.; Mueller, R. S.; VerBerkmoes, N. C.; Shah, M.; Singer, S. W.; Thelen, M. P.; Banfield, J. F.; Hettich, R. L., Computational Prediction and Experimental Validation of Signal Peptide Cleavages in the Extracellular Proteome of a Natural Microbial Community. *Journal of Proteome Research* 9, (5), 2148-2159.
106. Singer, S. W.; Erickson, B. K.; Verberkmoes, N. C.; Hwang, M.; Shah, M. B.; Hettich, R. L.; Banfield, J. F.; Thelen, M. P., Posttranslational modification and sequence variation of redox-active proteins correlate with biofilm life cycle in natural microbial communities. *Isme J.*
107. Jiao, Y.; Cody, G. D.; Harding, A. K.; Wilmes, P.; Schrenk, M.; Wheeler, K. E.; Banfield, J. F.; Thelen, M. P., Characterization of extracellular polymeric substances from acidophilic microbial biofilms. *Appl Environ Microbiol* 76, (9), 2916-22.
108. Denef, V. J.; Kalnejais, L. H.; Mueller, R. S.; Wilmes, P.; Baker, B. J.; Thomas, B. C.; VerBerkmoes, N. C.; Hettich, R. L.; Banfield, J. F., Proteogenomic basis for ecological divergence of closely related bacteria in natural acidophilic microbial communities. *Proc Natl Acad Sci U S A* 107, (6), 2383-90.
109. Wilmes, P.; Andersson, A. F.; Lefsrud, M. G.; Wexler, M.; Shah, M.; Zhang, B.; Hettich, R. L.; Bond, P. L.; VerBerkmoes, N. C.; Banfield, J. F., Community proteogenomics highlights microbial strain-variant protein expression within activated sludge performing enhanced biological phosphorus removal. *Isme J* **2008,** 2, (8), 853-64.
110. Finn, R. D.; Mistry, J.; Tate, J.; Coggill, P.; Heger, A.; Pollington, J. E.; Gavin, O. L.; Gunasekaran, P.; Ceric, G.; Forslund, K.; Holm, L.; Sonnhammer, E. L. L.; Eddy, S. R.; Bateman, A., The Pfam protein families database. *Nucl. Acids Res.* 38, (suppl_1), D211-222.
111. Kaczanowska, M.; Ryden-Aulin, M., Ribosome Biogenesis and the Translation Process in Escherichia coli. *Microbiol. Mol. Biol. Rev.* **2007,** 71, (3), 477-494.
112. Li, S.; Dass, C., Iron(III)-Immobilized Metal Ion Affinity Chromatography and Mass Spectrometry for the Purification and Characterization of Synthetic Phosphopeptides. *Analytical Biochemistry* **1999,** 270, (1), 9-14.
113. Wandersman, C. c.; Delepelaire, P., BACTERIAL IRON SOURCES: From Siderophores to Hemophores. *Annual Review of Microbiology* **2004,** 58, (1), 611-647.
114. Johnson, M.; Zaretskaya, I.; Raytselis, Y.; Merezhuk, Y.; McGinnis, S.; Madden, T. L., NCBI BLAST: a better web interface. *Nucl. Acids Res.* **2008,** 36, (suppl_2), W5-9.

115. Schauer, K.; Rodionov, D. A.; de Reuse, H., New substrates for TonB-dependent transport: do we only see the [`]tip of the iceberg'? *Trends in Biochemical Sciences* **2008,** 33, (7), 330-338.
116. Jeans, C.; Singer, S. W.; Chan, C. S.; Verberkmoes, N. C.; Shah, M.; Hettich, R. L.; Banfield, J. F.; Thelen, M. P., Cytochrome 572 is a conspicuous membrane protein with iron oxidation activity purified directly from a natural acidophilic microbial community. *ISME J* **2008,** 2, (5), 542-50.
117. Singer, S. W.; Chan, C. S.; Zemla, A.; VerBerkmoes, N. C.; Hwang, M.; Hettich, R. L.; Banfield, J. F.; Thelen, M. P., Characterization of cytochrome 579, an unusual cytochrome isolated from an iron-oxidizing microbial community. *Appl Environ Microbiol* **2008,** 74, (14), 4454-62.
118. Santiago, B.; Schubel, U.; Egelseer, C.; Meyer, O., Sequence analysis, characterization and CO-specific transcription of the cox gene cluster on the megaplasmid pHCG3 of Oligotropha carboxidovorans. *Gene* **1999,** 236, (1), 115-24.
119. Radauer, C.; Lackner, P.; Breiteneder, H., The Bet v 1 fold: an ancient, versatile scaffold for binding of large, hydrophobic ligands. *BMC Evol Biol* **2008,** 8, 286.
120. Busenlehner, L. S.; Pennella, M. A.; Giedroc, D. P., The SmtB/ArsR family of metalloregulatory transcriptional repressors: Structural insights into prokaryotic metal resistance. *FEMS Microbiol Rev* **2003,** 27, (2-3), 131-43.
121. Fusek, M.; Lin, X. L.; Tang, J., Enzymic properties of thermopsin. *J Biol Chem* **1990,** 265, (3), 1496-501.
122. Edmondson, S. P.; Qiu, L.; Shriver, J. W., Solution structure of the DNA-binding protein Sac7d from the hyperthermophile Sulfolobus acidocaldarius. *Biochemistry* **1995,** 34, (41), 13289-304.
123. Castelle, C.; Guiral, M.; Malarte, G.; Ledgham, F.; Leroy, G.; Brugna, M.; Giudici-Orticoni, M. T., A new iron-oxidizing/O2-reducing supercomplex spanning both inner and outer membranes, isolated from the extreme acidophile Acidithiobacillus ferrooxidans. *J Biol Chem* **2008,** 283, (38), 25803-11.
124. Denef, V. J.; Mueller, R. S.; Banfield, J. F., AMD biofilms: using model communities to study microbial evolution and ecological complexity in nature. *ISME J* **2010,** 4, (5), 599-610.
125. Haveman, S. A.; Holmes, D. E.; Ding, Y. H.; Ward, J. E.; Didonato, R. J., Jr.; Lovley, D. R., c-Type cytochromes in Pelobacter carbinolicus. *Appl Environ Microbiol* **2006,** 72, (11), 6980-5.
126. Singer, S. W.; Chan, C. S.; Zemla, A.; VerBerkmoes, N. C.; Hwang, M.; Hettich, R. L.; Banfield, J. F.; Thelen, M. P., Characterization of cytochrome 579, an unusual cytochrome isolated from an iron-oxidizing microbial community. *Applied and Environmental Microbiology* **2008,** 74, (14), 4454-4462.
127. Wilmes, P.; Remis, J. P.; Hwang, M.; Auer, M.; Thelen, M. P.; Banfield, J. F., Natural acidophilic biofilm communities reflect distinct organismal and functional organization. *ISME Journal* **2009,** 3, (2), 266-270.
128. Southey-Pillig, C. J.; Davies, D. G.; Sauer, K., Characterization of temporal protein production in Pseudomonas aeruginosa biofilms. *J Bacteriol* **2005,** 187, (23), 8114-26.
129. Teal, T. K.; Lies, D. P.; Wold, B. J.; Newman, D. K., Spatiometabolic stratification of Shewanella oneidensis biofilms. *Appl Environ Microbiol* **2006,** 72, (11), 7324-30.

130. Kelleher, N. L.; Taylor, S. V.; Grannis, D.; Kinsland, C.; Chiu, H. J.; Begley, T. P.; McLafferty, F. W., 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* **1998,** 7, (8), 1796-801.
131. Little, D. P.; Speir, J. P.; Senko, M. W.; O'Connor, P. B.; McLafferty, F. W., Infrared multiphoton dissociation of large multiply charged ions for biomolecule sequencing. *Anal Chem* **1994,** 66, (18), 2809-15.
132. Mortz, E.; O'Connor, P. B.; Roepstorff, P.; Kelleher, N. L.; Wood, T. D.; McLafferty, F. W.; Mann, M., Sequence tag identification of intact proteins by matching tanden mass spectral data against sequence data bases. *Proc Natl Acad Sci U S A* **1996,** 93, (16), 8264-7.
133. Larsen, M. R.; Roepstorff, P., Mass spectrometric identification of proteins and characterization of their post-translational modifications in proteome analysis. *Fresenius J Anal Chem* **2000,** 366, (6-7), 677-90.
134. Hunt, D. F.; Yates, J. R., 3rd; Shabanowitz, J.; Winston, S.; Hauer, C. R., Protein sequencing by tandem mass spectrometry. *Proc Natl Acad Sci U S A* **1986,** 83, (17), 6233-7.
135. VerBerkmoes, N. C.; Bundy, J. L.; Hauser, L.; Asano, K. G.; Razumovskaya, J.; Larimer, F.; Hettich, R. L.; Stephenson, J. L., Jr., Integrating 'top-down" and "bottom-up" mass spectrometric approaches for proteomic analysis of Shewanella oneidensis. *J Proteome Res* **2002,** 1, (3), 239-52.
136. 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.; Hurst, G. B., Characterization of the 70S Ribosome from Rhodopseudomonas palustris using an integrated "top-down" and "bottom-up" mass spectrometric approach. *J Proteome Res* **2004,** 3, (5), 965-78.
137. Borchers, C. H.; Thapar, R.; Petrotchenko, E. V.; Torres, M. P.; Speir, J. P.; Easterling, M.; Dominski, Z.; Marzluff, W. F., Combined top-down and bottom-up proteomics identifies a phosphorylation site in stem-loop-binding proteins that contributes to high-affinity RNA binding. *Proc Natl Acad Sci U S A* **2006,** 103, (9), 3094-9.
138. Johnson, K. A.; Paisley-Flango, K.; Tangarone, B. S.; Porter, T. J.; Rouse, J. C., Cation exchange-HPLC and mass spectrometry reveal C-terminal amidation of an IgG1 heavy chain. *Anal Biochem* **2007,** 360, (1), 75-83.
139. Whitelegge, J.; Halgand, F.; Souda, P.; Zabrouskov, V., Top-down mass spectrometry of integral membrane proteins. *Expert Rev Proteomics* **2006,** 3, (6), 585-96.
140. Wu, S.; Lourette, N. M.; Tolic, N.; Zhao, R.; Robinson, E. W.; Tolmachev, A. V.; Smith, R. D.; Pasa-Tolic, L., An integrated top-down and bottom-up strategy for broadly characterizing protein isoforms and modifications. *J Proteome Res* **2009,** 8, (3), 1347-57.
141. Wu, S. L.; Kim, J.; Hancock, W. S.; Karger, B., Extended Range Proteomic Analysis (ERPA): a new and sensitive LC-MS platform for high sequence coverage of complex proteins with extensive post-translational modifications-comprehensive analysis of beta-casein and epidermal growth factor receptor (EGFR). *J Proteome Res* **2005,** 4, (4), 1155-70.
142. Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S., Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999,** 20, (18), 3551-67.
143. Craig, R.; Beavis, R. C., TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004,** 20, (9), 1466-7.

144. LeDuc, R. D.; Taylor, G. K.; Kim, Y. B.; Januszyk, T. E.; Bynum, L. H.; Sola, J. V.; Garavelli, J. S.; Kelleher, N. L., ProSight PTM: an integrated environment for protein identification and characterization by top-down mass spectrometry. *Nucleic Acids Res* **2004,** 32, (Web Server issue), W340-5.
145. McLafferty, F. W.; Horn, D. M.; Breuker, K.; Ge, Y.; Lewis, M. A.; Cerda, B.; Zubarev, R. A.; Carpenter, B. K., Electron capture dissociation of gaseous multiply charged ions by Fourier-transform ion cyclotron resonance. *J Am Soc Mass Spectrom* **2001,** 12, (3), 245-9.
146. Holmes, M. R.; Giddings, M. C., Prediction of posttranslational modifications using intact-protein mass spectrometric data. *Anal Chem* **2004,** 76, (2), 276-82.
147. Karabacak, N. M.; Li, L.; Tiwari, A.; Hayward, L. J.; Hong, P.; Easterling, M. L.; Agar, J. N., Sensitive and specific identification of wild type and variant proteins from 8 to 669 kDa using top-down mass spectrometry. *Mol Cell Proteomics* **2009,** 8, (4), 846-56.
148. Connelly, H. M.; Pelletier, D. A.; Lu, T. Y.; Lankford, P. K.; Hettich, R. L., Characterization of pII family (GlnK1, GlnK2, and GlnB) protein uridylylation in response to nitrogen availability for Rhodopseudomonas palustris. *Anal Biochem* **2006,** 357, (1), 93-104.
149. McDonald, W. H.; Tabb, D. L.; Sadygov, R. G.; MacCoss, M. J.; Venable, J.; Graumann, J.; Johnson, J. R.; Cociorva, D.; Yates, J. R., 3rd, MS1, MS2, and SQT-three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications. *Rapid Commun Mass Spectrom* **2004,** 18, (18), 2162-8.
150. Frottin, F.; Martinez, A.; Peynot, P.; Mitra, S.; Holz, R. C.; Giglione, C.; Meinnel, T., The proteomics of N-terminal methionine cleavage. *Mol Cell Proteomics* **2006,** 5, (12), 2336-49.
151. Pitzer, E.; Masselot, A.; Colinge, J., Assessing peptide de novo sequencing algorithms performance on large and diverse data sets. *Proteomics* **2007,** 7, (17), 3051-4.
152. Blattner, F. R.; Plunkett, G., 3rd; Bloch, C. A.; Perna, N. T.; Burland, V.; Riley, M.; Collado-Vides, J.; Glasner, J. D.; Rode, C. K.; Mayhew, G. F.; Gregor, J.; Davis, N. W.; Kirkpatrick, H. A.; Goeden, M. A.; Rose, D. J.; Mau, B.; Shao, Y., The complete genome sequence of Escherichia coli K-12. *Science* **1997,** 277, (5331), 1453-62.
153. Larimer, F. W.; Chain, P.; Hauser, L.; Lamerdin, J.; Malfatti, S.; Do, L.; Land, M. L.; Pelletier, D. A.; Beatty, J. T.; Lang, A. S.; Tabita, F. R.; Gibson, J. L.; Hanson, T. E.; Bobst, C.; Torres, J. L.; Peres, C.; Harrison, F. H.; Gibson, J.; Harwood, C. S., Complete genome sequence of the metabolically versatile photosynthetic bacterium Rhodopseudomonas palustris. *Nat Biotechnol* **2004,** 22, (1), 55-61.
154. Pedrioli, P. G.; Eng, J. K.; Hubley, R.; Vogelzang, M.; Deutsch, E. W.; Raught, B.; Pratt, B.; Nilsson, E.; Angeletti, R. H.; Apweiler, R.; Cheung, K.; Costello, C. E.; Hermjakob, H.; Huang, S.; Julian, R. K.; Kapp, E.; McComb, M. E.; Oliver, S. G.; Omenn, G.; Paton, N. W.; Simpson, R.; Smith, R.; Taylor, C. F.; Zhu, W.; Aebersold, R., A common open representation of mass spectrometry data and its application to proteomics research. *Nat Biotechnol* **2004,** 22, (11), 1459-66.
155. Tran, J. C.; Doucette, A. A., Gel-eluted liquid fraction entrapment electrophoresis: an electrophoretic method for broad molecular weight range proteome separation. *Anal Chem* **2008,** 80, (5), 1568-73.

Vita

Brian Erickson earned his Bachelor of Science degree in Bioinformatics from Baylor University in 2003. He earned his Master of Science degree in Biotechnology at the University of Texas at San Antonio. He lives in Knoxville with his wife, Alison, bassets, Dexter and Jezebel and their amazingly vocal parrot, Chester.