Bioinformatic and Experimental Approaches for Deeper Metaproteomic Characterization of Complex Environmental Samples

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

Ramsunder Mahadevan Iyer
December 2017

Dedication

This dissertation is dedicated to the memory of

Manoj Kosare. He was a great friend and an awesome colleague with whom I shared a special bond in the short period of three years I worked with him.

Demise: 3rd November, 2014.

Acknowledgments

I would like to take this opportunity to thank all the people who helped, encouraged and supported me in this amazing doctoral journey. I would not have survived the rigors of graduate school without their unwavering support and belief that has resulted in me getting this hard-earned degree.

First of all, I would like to thank my adviser and mentor Dr. Robert L. Hettich who has been a constant source of inspiration and guidance in the five years I spent here at ORNL. I will always cherish the one-on-one conversations that were filled with honest feedbacks about my work, areas requiring improvement and loads of wisdom. I will never forget Bob asking me to close the door before the start of any meeting with him so that we do not overwhelm our neighbors with the sheer enthusiasm and excitement with which we discussed science. Also, I cannot forget the "Teacher Bob". His amazing ability to explain complicated concepts in a simple "hand-waving" manner is something I will miss immensely. Life in graduate school has been a mixed bag of satisfaction and disappointments and Bob always urged me to stay positive and helped me never lose my morale when the going got tough. I will always look back at the time spent here as a crucial phase for my self-development. I hope to apply the guidance and knowledge I received from Bob as I move ahead in my professional career.

I would like to thank my esteemed committee members Dr. Maria Cekanova, Dr. Gladys Alexandre, Dr. Margaret Staton and Dr. Dale Pelletier for taking the time out during committee meetings and providing me with valuable feedbacks that shaped up my dissertation. Their knowledge and critical feedbacks helped me to look at my research from different perspectives and broaden the scope of my projects.

I would like to thank Dr. Karuna Chourey for being a patient mentor and guide me on several projects where I had the opportunity to work with her. Her constant support enabled me to think critically about my projects and frame specific solutions that the collaborators required. I thank Dr. Ritin Sharma for helping me during my early days in the lab and teaching the experimental approaches for biological mass spectrometry and subsequent data analysis. I would like to thank Dr. Richard Giannone for his valuable inputs pertaining to the technical aspects of mass spectrometry and providing me with suggestions whenever required. I thank the other members in my group Suresh Poudel, Alfredo Blakeley-Ruiz, Ivan Villalobos Solis, Mallory Paige, Alexander Cope, David Reeves and Dr. Weili Xiong. I will miss our family of proteomics enthusiasts who are doing some really cool science and were always in for lively discussions during our weekly proteome meetings and with whom I had a blast during the annual ASMS retreats. A special thanks to my lab colleague and a good friend Chen Qian with whom I shared my office space. I am thankful to all the other members of organic and biological

mass spectrometry group especially Becky Maggard for her help with poster edits, printing and other administrative work, Keiji Asano and Dr. Greg Hurst for their help with instrument troubleshooting and their support in patiently resolving any other issues that I had during my stay here at ORNL.

I would also like to acknowledge the UT/ORNL graduate school of genome science & technology for giving me this wonderful opportunity and a conducive environment to pursue my doctoral studies. I thank Dr. Albrecht von Arnim for helping out with the smooth transition during my early days at UT and also giving me opportunities for inviting esteemed scientists from my field and hosting scientific talks which played a crucial role in my overall development. I am thankful to the GST staff especially Roger Gray, Terrie Yeatts and Anamika Missra for helping with my travel plans to scientific conferences and other administrative work during my stay at the graduate school.

Life in Knoxville would not have been so lively and magnificent without the amazing friends who were more like my extended family here. I thank all my friends especially Sagar Utturkar, Snehal Joshi, Snigdha Sewlikar, Arkadipta Bakshi, Preeti Chandrachud, Sudershana Nair, Aditya Barde, Ansul Lokdarshi and Chaithanya P Kalavagunta who were always there for me when I needed them. I will always cherish the time I spent with all these wonderful people and hope to share a bond that will last a lifetime. I am also thankful to my fiancée Archana Raghunath for her patience while I compiled my dissertation. She is quite a recent entry in my life but nonetheless a person with whom I wish to create some lasting memories.

I would also like to acknowledge two of my most amazing teachers Dr. Avinash Upadhyay and Dr. Jayasankar Neelakantan who inspired me to pursue a career in academics. I can never forget their passion for teaching which ignited the spark for science in students like me and make a career out of it.

A very special thanks to my parents who provided me with the best education and always motivated me to pursue my goals. Their dedication and hard work has resulted in me becoming the person that I am today. They instilled in me the most important values of honesty, obedience, responsibility and perseverance that I will always carry along with me as I move ahead in my personal and professional life. I would also like to acknowledge my brother Krishna Natarajan, and my nephew and niece Siddhith and Vihithi for their frequent visits to Knoxville and never making me feel away from home.

Finally, and most importantly, my utmost gratitude to my sister Lavanya Mahadevan for being my biggest support system. She has always been my ‘go to’ person during any crisis and always inspired me to give my best shot. Her guiding presence is undeniably the foundation upon which my entire graduate career has been built. Her relentless faith and constant motivation has been the most important factor that helped me realize my goals.

Abstract

The coupling of high performance multi-dimensional liquid chromatography and tandem mass spectrometry for characterization of microbial proteins from complex environmental samples has paved the way for a new era in scientific discovery. The field of metaproteomics, which is the study of protein suite of all the organisms in a biological system, has taken a tremendous leap with the introduction of high-throughput proteomics. However, with corresponding increase in sample complexity, novel challenges have been raised with respect to efficient peptide separation via chromatography and bioinformatic analysis of the resulting high throughput data. In this dissertation, various aspects of metaproteomic characterization, including experimental and computational approaches have been systematically evaluated. In this study, robust separation protocols employing strong cation exchange and reverse phase have been designed for efficient peptide separation thus offering excellent orthogonality and ease of automation. These findings will be useful to the proteomics community for obtaining deeper non-redundant peptide identifications which in turn will improve the overall depth of semi-quantitative proteomics.

Secondly, computational bottlenecks associated with screening the vast amount of raw mass spectra generated in these proteomic measurements have been addressed. Computational matching of tandem mass spectra via conventional database search strategies lead to modest peptide/protein identifications. This seriously restricts the amount of information retrieved from these complex samples which is mainly due to high complexity and heterogeneity of the sample containing hundreds of proteins shared between different microbial species often

having high level of homology. Hence, the challenges associated with metaproteomic data analysis has been addressed by utilizing multiple iterative search engines coupled with *de novo* sequencing algorithms for a comprehensive and in-depth characterization of complex environmental samples.

The work presented here will utilize various sample types ranging from isolates and mock microbial mixtures prepared in the laboratory to complex community samples extracted from industrial waste water, acid-mine drainage and methane seep sediments. In a broad perspective, this dissertation aims to provide tools for gaining deeper insights to proteome characterization in complex environmental ecosystems.

Table of Contents

Chapter 1 - Overview of Mass Spectrometry Based Proteomics and Bottlenecks Associated with Metaproteomic Characterization 1

1.1 Mass Spectrometry and its Applications in Biological Systems: 1

1.2 Mass Spectrometry based Proteomics in Systems Biology: 2

1.3 Metaproteomics- Utilizing the Power of High-Throughput Mass Spectrometry for Probing Complex Microbial Communities: 4

1.4 Mass Spectrometry as a Method of Choice for Characterization of Microbial Mixtures: 5

1.5 Experimental Challenges Associated with Community Proteomics: 6

1.6 Computational Challenges Associated with Community Proteomics: 8

1.6.1 Database choice - A Critical Factor Influencing Protein Identifications: 9

1.6.2 The Issue of Protein Identification and Inference: 10

1.7 De Novo Sequencing: 11

1.8 Scope of the Dissertation: 14

Chapter 2 - Experimental and Computational Approaches for Mass-spectrometry Based Community Proteomics 16

2.1 Generalized Workflow of MS based Proteomic Studies: 16

2.2 Digestion of Proteins into Peptides: 18

2.3 Multidimensional Chromatography: 20

2.4 Introducing Charged Particles to the Mass Spectrometer: 24

2.4.1 The Process of Electrospray Ionization: 24

2.5 Analysis of Peptide Ions Inside the Mass Spectrometer: 27

2.5.1 Mass analyzer: 27

2.5.2 Basic Steps in a Tandem Mass Spectrometry Experiment: 35

2.5.3 Data-Dependent Acquisition of MS/MS Spectra: 36

2.6 Database Matching of Tandem Mass Spectra: 37

2.6.1 Peptide Nomenclature: 37

2.6.2 Database Mapping of MS/MS Spectra: 39

2.6.3 Controlling False Discovery Rate: 42

2.6.4 Assembling Peptides to Proteins: 43

2.7 Quantitative Evaluation of MudPIT Data: 43

Chapter 3 - Optimization of Salt Pulse Step Elution Conditions for Improved Depth and Enhanced Coverage of Unique Peptides in Multi-dimensional LC-MS/MS Proteome Measurements 45

3.1 Separation by Two-Dimensional Chromatography to Enhance Unique Peptide Measurements: 45

3.2 Current Status and Limitations in Peptide Chromatography: 46

3.3 Materials and Methods: 50

3.3.1 Protein Extraction and Enzymatic Digestion: 50

3.3.2 Nano 2D LC-MS/MS Measurement: 51

3.3.3 Data Analysis: 52

3.3.4 De novo Sequencing: 53

3.4 Results and Discussion: 53

3.4.1 The Problem of Front Loading: 53

3.4.2 Proteome Metrics of Conventional 22 h vs. Modified 22 h Schemes: 55

3.4.3 Gravy Index of Peptides: 63

3.4.4 Percentage and Numerical Gains in Total Peptides between the Conventional 22 h and Modified 22 h Schemes: 65

3.4.5 Design of Shorter MudPIT Schemes for Improved Sample Throughput: 67

3.4.6 Distribution of Unique Peptides Across Modified 22 h and Modified 13 h Schemes: 71

3.5 Conclusions: 71

Chapter 4 - Evaluating the Impact of Multiple Search Engines and De Novo Sequencing Algorithms for Obtaining Deeper Proteome Coverage in Complex Environmental Samples 74

4.1 De Novo Sequencing as an Alternative Tool for Peptide Validation-Case Study Using a Thiocyanate Degrading Microbial Community Sample: 74

4.2 Computational Bottlenecks Associated with Database Search Strategy for Complex Metaproteomes: 80

4.3 Combining Results from Multiple Search Engines: 81

4.4 De Novo Sequencing- Advantages and Limitations: 82

4.5 PepExplorer- A Pattern Recognition Tool for Mapping de novo Sequencing Results: 83

4.6 Materials and Methods: 84

4.6.1 Sample Types: 84

4.6.2 Protein Extraction and Enzymatic Digestion: 84

4.6.3 Nano 2D LC-MS/MS Measurement: 86

4.6.4 Data Analysis: 86

4.6.5 De Novo Sequencing: 87
4.6.6 De novo Results Parsing via PepExplorer: 87
4.7 The Computational Pipeline: 88
4.8 Results and Discussion: 90
4.8.1 Comparison of Proteomic Results of Three Different Samples Analyzed Against Database Search Algorithms: 90
4.8.2 Taxonomic Attribution of Peptides via Unipept: 90
4.8.3 Protein Gains after PepExplorer Data Integration for AMD Sample: 96
4.8.4 Functional Annotation of PepExplorer Inferred Protein List for the AMD Sample: 99
4.9 Conclusions: 104
Chapter 5 – Applications of Metaproteomics for Investigation of Microbial Dynamics in Ground Water Specimens 106
5.1 Introduction: 106
5.2 Materials and Methods: 108
5.2.1 Sample Types: 108
5.2.2 Protein Extraction and Proteomic Analysis: 109
5.2.3 Metaproteome Bioinformatics: 111
5.3 Results and Discussion: 111
5.3.1 Impact of Database Versions on Protein/Peptide Identifications: 111
5.3.2 Peptide Redundancy Between the Single and Concatenated Metagenomes: 116
5.3.3 Spectral Quality Assessment of Single and Concatenated Metagenomic Assemblies: 118
5.3.4 Expression of Genes Involved in SCN^- Degradation Confirmed by Proteomics: 122
5.4 Conclusions: 125
Chapter 6 - Applications of Metaproteomics to Investigate the Expression of Multi-heme Cytochromes in Methane Seep Sediments 126
6.1 Introduction: 126
6.2 Materials and Methods: 129
6.2.1 Sample Collection: 129
6.2.2 Cellular Lysis, Protein Extraction and Sample Preparation: 129
6.2.3 Nano 2D LC-MS/MS Measurement: 131
6.2.4 Bioinformatic Data Analysis: 132
6.3 Results and Discussion: 133
6.4 Conclusions: 142

Chapter 7 - Conclusions, Current Trends and Future Perspectives 143
7.1 Conclusions from this Dissertation Work: 143
7.2 Current Trends in Mass Spectrometry Based Proteomics: 146
7.3 Future Outlook, Perspectives and Demands: 149
References 152
Vita 177

List of Tables

Table 2.1: Performance metrics of MS instruments described in this dissertation. 34

Table 3.1: The gradient elution profiles for tested MudPIT schemes 56
Table 3.2: Ammonium acetate concentrations and time durations for the conventional 22 h, modified 22 h and modified 13 h schemes..... 57
Table 3.3: Overview of proteomic results from samples measured by the conventional 22 h and modified 22 h schemes for 6iso sample 58
Table 3.4: Overview of proteomic results from samples measured by the conventional 22 h and modified 22 h schemes for C-elegans sample 58

Table 4.1: Overview of proteomic results from samples measured by the conventional 22 h and modified 22 h schemes for the Thiocyanate community sample 75

Table 6.1: Multiheme cytochromes containing less than 10 heme motifs searched with a database of 2246 proteins derived from the sixteen genomes of SEEP-SRB1 and SEEP-SRB4 clades. 136
Table 6.2: Multi-heme cytochromes detected in each replicate run across all three sites (Eel River Basin, Santa Monica Basin and Hydrate Ridge). 137

List of Figures

Figure 1. 1: Peptide reconstruction via De Novo Sequencing 12

Figure 2.1: Schematic illustration of the MudPIT Workflow 17
Figure 2.2: Chromatographic Separation of Peptides 23
Figure 2.3: Schematics of Electrospray Ionization 26
Figure 2.4: Block Diagram of Linear Trapping Quadrupole 29
Figure 2.5: Block Diagram of LTQ Velos 30
Figure 2.6: Block Diagram of Orbitrap ELITE Hybrid Mass Spectrometer 32
Figure 2.7: Data dependent acquisition. 38
Figure 2.8: Nomenclature for naming the ions formed from peptide backbone cleavages 40

Figure 3.1: Distribution of unique peptides across each salt pulse (sp) for the Conventional 22 h (3.1A) and Modified 22 h (3.1B) schemes for 6iso sample. 54
Figure 3.2: Percentage of total proteins (3.2A) and peptides (3.2B) repeating across technical replicates for 6iso sample. 59
Figure 3.3: Scatter plot matrix and Pearson's correlation values of normalized spectral counts (nSpc) for conventional 22 h (3.3A) and modified 22 h (3.3B) schemes for the 6iso sample. 61
Figure 3.4: Proteome metric improvement in the modified 22 h scheme. 62
Figure 3.5: Hydrophilic (3.5A & 3.5C) and hydrophobic (3.5B & 3.5D) peptide elution profiles for conventional 22 h and modified 22 h schemes for C.elegans samples and 6iso samples. 64
Figure 3.6: Numerical gains in peptide counts between the conventional 22 h and modified 22 h schemes for 6iso (3.6A) and C.elegans (3.6B) samples. 66
Figure 3.7: Peptide and protein counts of 6iso tryptic peptides shown on a linear scale for conventional 22 h and modified 13 h schemes. 69
Figure 3.8: Distribution of Protein groups for 6iso sample 70
Figure 3.9: Distribution of unique peptides across each salt pulse (sp) for modified 13 h (3.9A) and modified 22 h (3.9B) schemes. 72

Figure 4.1: Distribution of concatenated peptides across individual salt pulses for technical triplicates of the thiocyanate community sample. 76
Figure 4.2: Distribution of PepNovo+ scores for collected mass spectra for the thiocyanate community sample 78
Figure 4.3: Schematic illustration of the designed computational pipeline used for analyzing the metaproteomic data. 89
Figure 4.4: Peptide (4.4A) and protein (4.4B) counts respectively of the three samples searched against either four search algorithms ("Combined") or one search algorithm (MSGF+). 91

Figure 4.5: Taxonomic attribution of 6iso (4.5A) and AMD (4.5B) metaproteomic data obtained via de novo sequencing and filtered via Unipept. 94
Figure 4.6: Taxonomic annotations common between combined database search algorithms and de novo sequencing for 6iso and AMD metaproteomic data. 95
Figure 4.7: Average number of proteins detected in the AMD sample using a single database search engine (MSGF+), combined database search engines and summation of proteins obtained using the combined database search engines and Pepexplorer mapping of de novo predicted peptides to the database. 97
Figure 4.8: Graphical user interface of the PepExplorer results browser 98
Figure 4.9: The graphical report of the protein sequence coverage showing the extension of area covered by the predicted peptides..... 100
Figure 4.10: Functional annotation of AMD metaproteomic data obtained using GhostKOALA.102
Figure 4.11: Taxonomic distribution of five most abundant phyla and subsequent genus based distribution of the most abundant phyla for the PepExplorer filtered protein list..... 103

Figure 5.1: Schematics of the sample types chosen for the thiocyanate study 110
Figure 5.2: Schematic illustration of the database versions used in the thiocyanate study..... 112
Figure 5.3: Peptide (5.3A) and protein (5.3B) counts respectively of each replicate when the thiocyanate data was searched with single and concatenated metagenomes. 114
Figure 5.4: Bipartite graph explaining the process of protein assembly in IDPicker..... 115
Figure 5.5: Venn diagram depicting the peptide redundancy i.e. the total number of shared peptides between the single and concatenated metagenomic databases..... 117
Figure 5.6: Stacked histograms representing the distributions of ScanRanker scores for collected mass spectra for control planktonic (5.6A), experimental biofilm (5.6B) and control biofilm (5.6C) respectively searched with single metagenomic assembly. 119
Figure 5.7: Stacked histograms representing the distributions of ScanRanker scores for collected mass spectra for control planktonic (5.7A), experimental biofilm (5.7B) and control biofilm (5.7C) respectively searched with concatenated metagenomic assembly. 120
Figure 5.8: Metaproteomics in SCN– reactor showing expression of genes involved in SCN– degradation and by product breakdown..... 123

Figure 6.1: Representative operon structure from organisms containing large multiheme cytochromes found in SEEP-SRB1..... 128
Figure 6.2: Metaproteomic Sample collection sites. 130
Figure 6.3 (A-F): Manual validation of acquired mass spectra 138

Chapter 1 - Overview of Mass Spectrometry Based Proteomics and Bottlenecks Associated with Metaproteomic Characterization

---

1.1 Mass Spectrometry and its Applications in Biological Systems:

Mass spectrometry mainly involves the measurement of the mass to charge ratio of ions present in a sample. It has become a widely used analytical tool in biological research for the characterization of biomolecules like sugars, proteins, lipids, oligonucleotides etc.

The first applications of mass spectrometry (MS) in biological research can be traced back to 1940's when heavy stable isotopes were used as tracers to understand carbon-di-oxide production in animal models [1]. Since then, rapid advancements in technology have diversified the range of MS in biological research. Currently, the applications of mass spectrometry in life-sciences encompass such diverse areas as screening infants for metabolic disorders [2], profiling changes in protein expression between cells grown in different growth conditions [3], determining the presence of minerals in food [4], characterizing pharmaceutical drugs and *in-vivo* metabolism [5] and studying functional gene expression in complex microbial ecosystems [6]. The growing popularity of MS in research can be exemplified from a search in PubMed (an online archive of life-science journals) for the phrase "Mass Spectrometry" that resulted in over 280,000 total hits, with over 46,900 articles published since 2015. The importance of MS in biological research was further highlighted in 2002 when John Fenn and Koichi Tanaka were awarded the 2002 Noble prize in chemistry for their work on the "*Development of soft ionization methods for mass spectrometric analysis of biological macromolecules"* [7, 8].

By definition, a mass spectrometer is an instrument that can ionize a sample and measure the mass to charge ratio of the resulting ions. However, the functional versatility of this mass to charge measurement has enabled this instrument to become a vital tool in a wide range of fields, including life sciences. This versatility is imparted because of the instrument's ability to give quantitative and qualitative information on the elemental, isotopic, and molecular composition of both organic and inorganic specimens. In addition, the samples that can be analyzed include gas, liquid, and solid states that have masses ranging from single atoms (several Da) to proteins over 300 KDa [9]. Post-translational modifications (PTM) of proteins, an important regulatory process for the cellular localization and eventual function, can also be quantified via MS [10]. The coupling of MS instrumentation with high throughput computing clusters and associated software have dramatically improved the breadth of information that can be obtained in these complex studies and thus helped MS to permeate into an extensive range of research domains. Lastly, the interface of liquid chromatography with mass spectrometry (LC-MS) has allowed separation and characterization to be performed simultaneously thus making it a leading analytical technique in biotherapeutics [11].

1.2 Mass Spectrometry based Proteomics in Systems Biology:

Systems biology is an inter-disciplinary field consisting of genomics, transcriptomics, proteomics and metabolomics that is aimed at understanding the biological systems at the global level [12]. Proteomics, which is the comprehensive study of the expression of the entire compliment of proteins in different organs, tissues or cell types, is a key component of the systems biology approach. Proteomics plays an essential role in our understanding of biology as a dynamic

system since the proteome is directly derived from the genome and in turn regulates both gene expression and cellular metabolism. The analytical versatility of the mass spectrometer has been extensively exploited in systems biology based proteomic studies involving complex correlations among molecular components.

There are two general strategies for studying Mass spectrometry (MS) based proteomics. The first is a discovery based approach in which, presence and absence of a certain protein is carried out by differential comparison of the entire proteome obtained from cellular extracts grown in different physiological conditions. This technique has demonstrated the capability to identify several thousand proteins and presents a global picture of the proteome under consideration [13]. However, specific functional characterizations like protein-protein interactions cannot be studied using this method. For this, a much more targeted approach is employed in which specific proteins along with their binding partners are extracted and studied systematically.

MS based proteomics can be experimentally carried out using two different strategies. 1) The ‘‘bottom-up’’ strategy involves digesting the crude protein extract into peptides using proteolytic enzymes such as trypsin. This is followed by separation of peptides using liquid chromatography coupled to tandem mass spectrometry. 2) ‘‘Top-down’’ strategy involves MS analysis of intact proteins that have not been cleaved by proteolytic enzymes. Here, intact proteins are subjected to direct fragmentation and subsequent mass analysis. All the experiments described in this dissertation were carried out using the bottom-up approach [14].

Some of the current studies where MS based proteomics has contributed to development of systems biology include 1) Nutritional sciences; where systems analysis of normal and nutrient-perturbed networks with respect to a specific diet has enabled the modeling of cellular signaling responses [15], 2) Drug Discovery; where systems biology has resulted in the identification of protein network components and characterization of post-translational modifications in the drug discovery process [16], 3) Neurological Diseases; where integration of data from genetic and proteomic experiments has identified pathways involved in the pathogenesis of neurological diseases [17].

1.3 Metaproteomics- Utilizing the Power of High-Throughput Mass Spectrometry for Probing Complex Microbial Communities:

The coupling of high performance multi-dimensional liquid chromatography and tandem mass spectrometry for deep proteome characterization of microbial proteins from complex samples (metaproteomics) has heralded a new era of scientific discovery for these crucial unculturable systems. Metaproteomic investigations have highlighted some interesting aspects of functional gene expression within microbial consortia, particularly those that contain limited microbial membership diversity. However, applications of proteomic investigations for complex microbial assemblages such as groundwater, seawater soils etc. still present challenges that need to be addressed. Nonetheless, metaproteomics has played a major role in enhancing our understanding of the microbial world and link microbial community composition to function.

Earlier work on metaproteomics was heavily focused on lower complexity microbial systems such as those found in acid mine drainage (AMD) [18-20], sludge water bioreactor [21] and

microbial communities in gnotobiotic mice [22]. These models provided an initial understanding that propelled the design of customized proteomic workflows for more complex systems. Some of the key areas where environmental metaproteomics has been used to gain molecular insights include 1) Bioremediation; where microbial activities resulted in remediating toxic metal contamination sites like soil [23]. 2) Carbon cycling as a means to regulate carbon flow in the ecosystem [24]. 3) Bioenergy that deals with the characterization of microbial species involved in the conversion of cellulosic material to biofuels for bioethanol, biodiesel and biohydrogen production [25] and 4) Understanding how microbial species impact human health in various body sites including oral, gastrointestinal and genital tracts [26].

Inspecting metaproteomic datasets from these studies have revealed crucial information about microbial community structure, function, and dynamics. These observations have enhanced our understanding on how micro-organisms co-ordinate, co-operate and compete for nutrient resources and how they share the metabolic 'work' amongst themselves to ensure community survival and provides defense from external environmental sources. At the higher level, this information is useful in understanding host-microbial interactions such as bacterial/plant or bacterial/animal interfaces.

1.4 Mass Spectrometry as a Method of Choice for Characterization of Microbial Mixtures:

Mass spectrometry has served as a standard analytical platform for characterization of constituent molecules (peptides, proteins and other metabolites) in microbial mixtures. Recently, MS-based approaches have been used as alternatives, or in some cases completely replaced conventional methodologies for microbial characterization in research as well as

clinical microbiology laboratories. A series of commercially available systems are now routinely employed in hospital, clinical, and research labs worldwide [27]. Some of the highly effective MS systems for microbial characterization include MALDI-based tandem instruments—trap-TOF [28], TOF-TOF [29] FTMS [30] as well as single particle mass spectrometry (SPMS) analyzers [31]. Applications of electrospray ionization coupled with tandem MS for rapid characterization of microorganisms (including their proteomes and metabolomes) have been gradually making their way in 1) Identifying bacteria with unsequenced genomes [32], 2) In bacterial metaproteomics [33] and 3) Proteogenomics [34, 35].

1.5 Experimental Challenges Associated with Community Proteomics:

One of the major bottlenecks associated with microbial metaproteomics is the process of protein extraction from complex environmental samples. These samples often contain mixtures of various organic and inorganic materials that includes humic acid, lignin, chemical chelation, cell exudation and various degradation products. Different protein extraction methods have been designed depending on the features of the environmental samples that includes soils, sediments, ground and sea water samples, acid mine drainage, biofilms, marine organic particles and symbionts. However, due to the heterogeneous species composition, wide dynamic range in protein abundance levels, and the proteins binding to the membrane or soil matrix that are difficult to extract, there is no standard and efficient protocol for extracting proteins from environmental samples. Thus, design of new extraction protocols that can lead to enhanced quantification of proteins from these complex matrices is an area of active research in metaproteomic studies.

Another major caveat in the experimental workflow is the separation of peptide mixtures. Proteolytic digestion of proteins extracted from these community samples generally results in complex peptide mixtures. Sample fractionation is a crucial step, as it simplifies complexity prior to a mass spectrometric measurement. 2-dimensional gel electrophoresis is an efficient protocol for protein separation as it can effectively resolve protein isoforms and modification states. However, this process is hampered by its limited dynamic range and low-throughput, and thus has gradually been replaced with liquid chromatography (LC). In general, multi-dimensional separations consisting of strong cation exchange (SCX) with reverse phase (RP) are used to separate complex peptide mixture in an automated, high-throughput manner. Here two different peptide properties i.e. charge and hydrophobicity are exploited to obtain good orthogonality. This dramatically improves the resolution and better sampling of low abundant species that results in the identification of thousands of proteins from these samples. However, disadvantages of LC include its generally higher operational cost and less favorable concentration sensitivity.

Besides this, some of the other experimental challenges encountered during metaproteomic analyses include 1) Further optimization of proteome fractionation procedures for efficient extraction of soluble, membrane and extracellular fractions, 2) Quantification methods that do not require *in situ* metabolic labeling and 3) Use of mass spectrometers with fast scan speeds and high mass accuracies for resolving complex proteomes within optimum time frames.

1.6 Computational Challenges Associated with Community Proteomics:

A multidimensional liquid chromatography/ tandem mass spectrometry measurement of a community sample generates molecular masses for thousands of tryptic peptides over a course of an extended chromatographic elution profile, which can last for 22-24 hours. Naturally, this process generates tens to thousands of fragmentation spectra, most of which correspond to a specific peptide sequence. Thus, manual inspection of these myriads of peptide sequences is impossible and requires complex bioinformatic tools for data interpretation.

Over the past few years, the development of metaproteomics has been driven by the ability to sequence microbial genomes via high throughput sequencing [36]. Thus, the quality of a metaproteomic data is related to the metagenome from which it was derived. High-throughput DNA sequencing platforms like 454, Illumina and Ion Torrent have had dramatic impacts on metaproteome measurements. Best bioinformatic practices in metaproteomic data analysis involves constructing a predicted metagenomic database of the sample under consideration and computationally converting this metagenome to a metaproteome database and then searching the fragmentation spectra against this database to infer peptides and proteins. With the widespread availability and reduced cost of sequencing platforms, metagenome sequencing of samples has become a standard routine.

Despite these developments, the field suffers from the exceptional complexity and heterogeneity of the sample that hamper data evaluation. Protein hits in metaproteome measurements are derived from several hundreds of different species, thus making taxonomic binning a tremendous task. Another challenging issue routinely encountered is the functional

annotation of proteins, as community proteomics is chiefly interested in specific functions performed by the dominant member in the ecosystem [37]. Other intrinsic complexities present in the biological sample, like homologous proteins/domains, horizontal gene transfer and strain level variations, require additional proteomic considerations that can maximize protein identification without compromising the false discovery rate (FDR).

These computational bottlenecks have partially been addressed by developing softwares that are dedicated for the analysis of metaproteomic datasets [38, 39]. For exhaustive peptide and protein identifications, these softwares feature multiple database search algorithms which lead to enhanced identification reliability. Also, the use of concatenated metagenomes by combining several samples from different time points can boost metaproteome coverage. This issue will be addressed in detail in chapter 5. One solution in the measurement space employs the use of high mass accuracy, high resolution mass spectrometers such as FTICR or Orbitraps that have very high discriminatory power (≤ 5 parts-per-million mass accuracy) to resolve peptides of similar mass to charge ratios. This high resolving power permits high fidelity assignments during database search at an extraordinarily low FDR (false discovery rate) levels (≤ 0.1 %) [40].

1.6.1 Database choice - A Critical Factor Influencing Protein Identifications:

In spite of the development of dedicated tools (like multiple iterative search engines discussed above) for metaproteomic characterization, the choice of protein sequence database remains a major factor influencing the annotation of tandem mass spectra. Protein identification chiefly depends on matching experimental mass spectra obtained from the sample under study with the *in-silico* generated theoretical spectra obtained from a protein sequence database.

Environmental samples derived from soil, groundwater etc. may contain thousands of different microbial species, and the selection of a well-suited database is rather an arduous task. Secondly, these large databases result in an exponential increase in search space, which poses several FDR-related issues which impact the peptide to spectrum matches [41, 42]. Despite outstanding efforts undertaken in addressing metagenome research, the above-mentioned factors pertaining to database selection is still a major limiting factor influencing in-depth peptide/protein identifications. Cross-species identification is often a successful approach due to the high sequence similarity among orthologous genes from closely related species [43], but a single amino acid change can seriously affect peptide-to-spectrum matching, resulting in erroneous protein identification. Error-tolerant DB searches, as well as DB-independent strategies like *de novo* (peptide) sequencing (discussed below), have been proposed as alternative solutions to classical DB search [42, 44-46]. However, an optimized and standardized mass spectrometry (MS) data analysis pipeline that can cater to all the bottlenecks concerning metaproteomic characterization is not yet available.

1.6.2 The Issue of Protein Identification and Inference:

Apart from database choice, the validation of peptides/proteins identified is major issue to tackle in shotgun metaproteomic experiments [47]. The problem arises because of the fact that a given peptide identification can be potentially matched to several proteins due to sequence redundancy. In the metaproteomic landscape, this issue becomes even more complicated as the peptide sequences could either originate from different proteins within a single organism (intra-species variation) or more likely from homologous proteins originating in different

species (inter-species variation). This recurrence of a protein or in certain cases a part of a protein sequence is mainly due to sequence conservation of functional domains between closely related organisms. As a result, the protein assembly software will attribute these redundant sequences to several species thus hampering the process of unique identification and correct assignment of species. Hence, additional efforts are required while assigning protein identifications to avoid false positives [48].

1.7 *De Novo* Sequencing:

Although database searching is the current gold standard for analyzing proteomics data, it suffers from a major drawback in that it relies heavily on a user provided database. This is acceptable in the case of microbial isolates where the database complexity is relatively low. However, in case of complex environmental samples, the metagenome may be unavailable or incomplete. This is because some of the low abundant microbial species in the sample may be uncultivable and hence the database lacks the complete genomic sequences. These naturally underrepresented species that contribute to the metaproteome may go undetected due to the absence of their corresponding sequences in the database.

De novo sequencing offers an extended approach to common database searching strategy and can be used as an alternative tool for peptide identification. Here, instead of matching the spectrum to a database-derived theoretical spectrum, the *de novo* algorithm examines the sequence of a spectra directly from the tandem mass spectrum [49]. Thus, it side-steps the problem of unrecovered sequences by not relying on the protein sequence database at all. **Figure 1.1** shows a representative example of peptide reconstruction directly via tandem mass

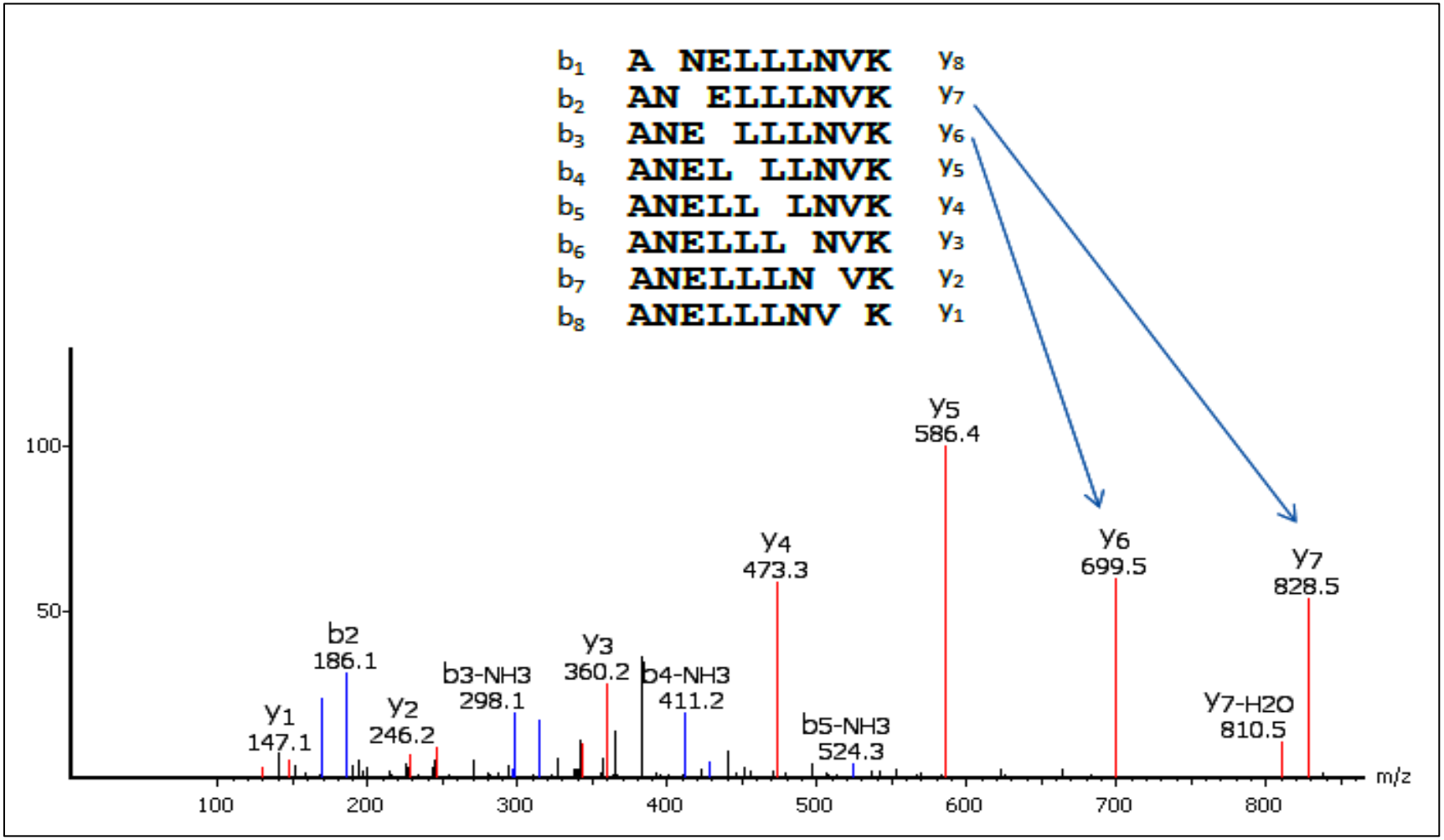


Figure 1. 1: Peptide reconstruction via De Novo Sequencing

spectra. The main idea of *de novo* sequencing is to use the mass difference between two fragment ions to calculate the mass of an amino acid residue on the peptide backbone. The mass can usually uniquely determine the residue. For example, the mass difference between the y7 and y6 ions in **Figure 1.1** is equal to 129, which is the mass of glutamic acid (E) residue. Similarly, the next adjacent residue between y6 and y5 can be determined as Leucine (L) by the mass difference. Such a process can be continued until all the residues are determined. Thus, if one can identify either the *y-ion* or *b*-ion series (the peptide nomenclature of *b and y* type ions is discussed in chapter 2) in the spectrum, the peptide sequence can be determined.

However, the spectrum obtained from the mass spectrometry instrument does not tell the ion types of the peaks, which requires a computer algorithm to figure out the process of *de novo* sequencing. Modern computer architectures now allow *de novo* sequencing to be used as a powerful and cost-effective validation method for peptide identification. However, it should be noted that distinguishing amino acids or their combinations by direct interrogation of fragmentation spectra require high mass accuracy instrument platforms. For example, the masses of the Lysine and Glutamine residues differ by just 0.037 Da and thus requires a mass spectrometer that can afford a mass accuracy within a few ppm range that can distinguish closely related species. Hence, *de novo* sequencing should be used as a complimentary approach only when mass spectrometers possessing high mass accuracies (like Orbitraps discussed in chapter 2) are used for proteomic measurements. Secondly, the essential step of mapping these *de novo* peptides back to a protein sequence is not performed by the *de novo* algorithm. Also, using *de novo* derived data for taxonomic and functional interpretation of

metaproteomics data has not been explored so far. These aspects will be discussed in broader detail in chapter 4.

1.8 Scope of the Dissertation:

This dissertation will primarily focus on two key factors for enhancing metaproteome research: 1) Optimization of chromatographic separations in order to obtain deeper non-redundant peptide identifications and 2) Optimized informatics pipeline consisting of multiple iterative search engines coupled with *de novo* sequencing algorithms for a comprehensive and in-depth characterization of complex environmental samples. As demonstration of these two major aspects, this dissertation will also describe the deployment of these experimental and computational strategies to evaluate various natural environmental microbial communities.

Chapter 2 will provide a detailed explanation of biological mass spectrometry with a special focus on discovery proteomics. The foundations of multidimensional chromatography and tandem mass spectrometry, ionization modes, data-dependent acquisition, peptide/protein identification from MS/MS spectra will be discussed in broader detail.

Chapter 3 will describe the optimization and refinement of complex peptide separation profiles on samples of varying complexities. Here, the concept of “peptide front-loading” will be introduced, which impacts proteome depth and modifications in experimental protocols will be proposed to circumvent this issue. The findings from this chapter will be useful to the proteomics community for obtaining deeper non-redundant peptide identifications which in turn will improve the overall depth of semi-quantitative proteomics.

Chapter 4 will address the computational bottlenecks associated with community proteomics. The use of multiple search engines coupled with the integration of orthogonal information form *de novo* sequencing algorithms will be discussed here. Samples of varying complexities will be used to evaluate the range and effectiveness of our approach.

Chapter 5 will present a case study where time-series genome resolved metagenomics and endpoint metaproteomics were used to reveal changes in the microbial community of a newly inoculated Thiocyanate (SCN^-) bioreactor. This work will highlight the applicability of metaproteomics to gain a mechanistic understanding of contaminant degradation by a microbial community and provide critical biological insights into the expression patterns of microbial functionality in the community.

Chapter 6 will present another case study where metaproteomics was used to assess the expression of microbial multi-heme cytochromes *in situ* at three different sites along the west coast of North-America that contain methane seeps. This study highlights the applicability of metaproteomics to reveal the expression of proteins involved in extracellular electron transfer from ANME archaea to their bacterial partners thus mediating the process of reverse methanogenesis in oceanic sediments.

Finally, chapter 7 will summarize the major accomplishments of this dissertation research and provide a framework for future experiments to obtain deeper proteome coverage in diverse sample types at an unprecedented level of molecular detail. It will conclude by outlining the current state of mass spectrometry based proteomics and highlight major areas that need further research for comprehensive protein identification.

Chapter 2 - Experimental and Computational Approaches for Mass-spectrometry Based Community Proteomics

---

2.1 Generalized Workflow of MS based Proteomic Studies:

In this chapter, the overall approach of proteomic measurements of biological samples by liquid chromatography and tandem mass spectrometry will be discussed in broader detail.

Multidimensional liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) is a standard analytical platform for the identification and quantification of peptides and proteins in complex biological samples. MudPIT (Multidimensional protein identification technology) is a widely used workflow for the separation of proteolytic peptides [50]. The main goal of MudPIT mass spectrometry is to maximize the identification of proteins with varied abundances. The intrinsic capability of MS allows only a limited number of peptide identifications, with the identifications biased towards the more abundant species in the sample. MudPIT measurements usually employ longer run times (~20 hrs.) that results in better separation of complex samples. This in turn enhances the overall peptide identifications [51]. Prior to scanning by the mass spectrometer, the peptide mixtures are resolved by 2-dimentional chromatography. The main goal of peptide separation by chromatography is to maximize unique peptide identifications and in turn improve the overall protein identifications. **Figure 2.1** outlines the generalized schematics of MudPIT workflow. Here, the proteins are first digested to peptides. These peptides are then loaded onto a biphasic column (usually strong cation exchange [SCX] and reverse phase[RP]). Different fractions of peptides are then eluted off this column by multidimensional chromatography and they subsequently enter the mass

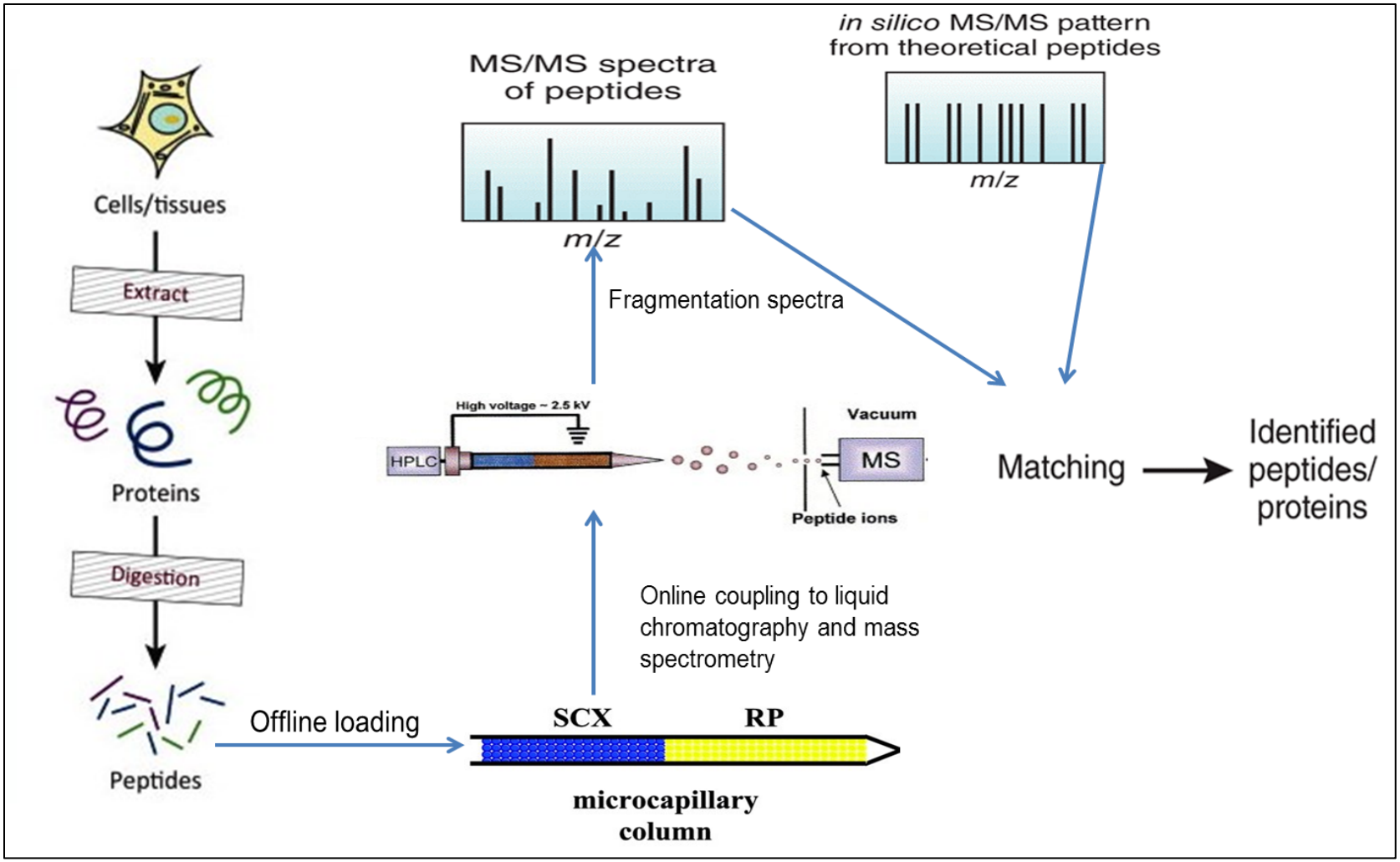


Figure 2.1: Schematic illustration of the MudPIT Workflow

spectrometer. Here the peptides of unique mass are first isolated and then fragmented. The mass/charge values of the resulting fragment ions are then measured to generate the MS/MS spectrum. The experimental spectra provide an empirically generated fingerprint of the peptide. Following this, the sequenced genome of the sample under consideration is subjected to *in-silico* digestion to generate theoretical spectra. Finally, the experimental and theoretical spectra are compared to identify peptides and proteins [52, 53]. A ~20 hr. MudPIT run typically generates hundreds to thousands of MS/MS spectra from which thousands of proteins can be sampled. Quantitative information on relative protein abundances can be estimated in a number of ways and will be discussed in section 2.7.

2.2 Digestion of Proteins into Peptides:

In order to obtain the peptide mixture for LC-MS/MS analysis, the biological sample is typically processed as follows:

- The environmental sample is subjected to cell lysis. For this, the sample is boiled in lysis buffer containing SDS which breaks open the cells (SDS dissolved in 100 mM Tris-HCl buffer pH 8.0). SDS is a denaturing anionic surfactant which works by disrupting the hydrophobic interactions within the cell membrane. This damages the cellular integrity resulting in release of the proteins into the solution [54].
- The resulting crude protein extract is subjected to TCA (Trichloroacetic acid) precipitation, which separates the proteins from the remaining cell debris. TCA primarily acts by dehydration of the water shells around the protein [55]. In addition, the anionic TCA may trigger partial protein unfolding through disruption of the electrostatic

interactions which determine the native tertiary structure of the protein. As a result, the usually well-hidden hydrophobic interior of the protein becomes exposed to the solvent [55].

- Protein denaturation is then achieved using urea. Direct interaction of urea with charged amino acid residues weakens the intermolecular bonds, thus weakening the overall secondary structure of protein. Once, the gradual unfolding of protein occurs, urea can then easily access the hydrophobic inner core of the protein thus speeding up the denaturation process [56]. In the next step, disulfide bonds in the protein are subsequently reduced with DTT (dithiothreitol) and the resulting thiol groups are alkylated using iodoacetamide to prevent bond reformation.
- In the next step, the protein sample is digested to generate peptides that is subsequently sequenced by mass spectrometry. Protein digestion can be achieved both chemically and enzymatically. Cyanogen Bromide is a widely used chemical for protein digestion [57]. However, the main method of choice for protein digestion is enzymatic digestion by trypsin. It is a widely used endoprotease which acts by cleaving at the carboxyl side of the arginine and lysine residues unless these amino acids are followed by proline [58]. Tryptic peptides are between 10 to 20 amino acids long depending on the frequency of lysine and arginine residues, this range in molecular weight of peptides is ideal for measurements with the majority of mass analyzers. Also, the low cost of trypsin makes it an ideal candidate for mass spectrometry based proteomics. This process of proteolytic digestion of peptides for subsequent LC-MS/MS analysis is called bottom up proteomics.

- The samples are then transferred to a 10 kDa molecular cut off filter. An acid-salt solution of NaCl and formic acid is then added on top of the filter. The salt helps in efficient removal of peptides from the filter and helps in reducing the surface adsorption of peptides while the acid aids in protonation of peptides. This filter is spun at 4500 g, which causes the elution of tryptic peptides while all the remaining cellular debris are retained by the filter.
- Finally, the standard peptide amount to be loaded for each MS measurement is determined by Bicinchoninic Acid Assay (BCA) assay kit [59]. This assay is based on the principal of Cu^{2+} reduction to Cu^{1+} by peptides in the alkaline solution. The Cu^{1+} ions are then detected by bicinchoninic acid. Initially, a standard curve is plotted by measuring the optical density of bovine serum albumin (BSA) prepared via serial dilution. Following this, the optical density of sample peptide solution is measured. Then using the standard curve as the reference, the peptide concentration in the sample is determined.

2.3 Multidimensional Chromatography:

The major advantage of MudPIT analysis is that it allows for the quantification of more low abundant proteins from a complex sample. This is achieved by chromatographic separation of a digested sample over long periods of time so that the low abundant species are also given sampling opportunities.

- Initially, the peptide sample is loaded under high pressure on a fused silica microcapillary back-column (150 µm inner diameter), containing two stationary phases in tandem, a ~ 4.5 cm strong cation exchange (SCX, 5 µm particle size, 100 Å pore size) resin followed by

~4.5 cm reverse phase (RP, 5 µm particle size, 125 Å pore size) resin. The SCX stationary phase consists of aliphatic sulfonic acid groups that are negatively charged in aqueous solution that results in the binding of strong basic analytes. The reverse phase resin consists of octadecyl carbon chain (C18)-bonded silica. Due to its non-polar and hydrophobic nature, the molecules in the polar mobile phase are adsorbed to the hydrophobic stationary phase of the reverse phase resin. An online two-dimensional liquid chromatography setup employing SCX and RP is a widely-used method for peptide fractionation because of its ability to exploit two distinct peptide chemical properties (charge and hydrophobicity), thus offering reasonable orthogonality.

- The presence of RP upstream of SCX allows the sample to be loaded directly after digestion. At this point, the peptides are irreversibly bound to the RP resin.
- In the next step, the back-column is washed offline for 15 mins with solvent A (95% HPLC grade water, 5% acetonitrile, 0.1% formic acid) and then five ramps of solvent A and solvent B (70% acetonitrile, 30% HPLC grade water, 0.1% formic acid) for 30 mins (Solvent A and Solvent B are the mobile phases). This initial washing step removes the residual urea or salt that may present in the sample. This step also translocates the bound peptides from the upstream RP onto the SCX resin.
- The biphasic back column with peptides bound to SCX is now mounted between a high-power liquid chromatography (HPLC) system and an in house pulled nanospray emitter (100 µm i.d) packed with 12-15 cm of reverse phase resin. Thus, the peptides can be resolved chromatographically and then eluted directly into the mass spectrometer. This procedure is called online MudPIT.

- To resolve the peptides, a series of eleven ~2 hour chromatographic steps are employed (total runtime is 22 hours). Here, the peptides are translocated from the SCX to the downstream RP using a volatile salt (usually ammonium acetate) and then eluted into the mass spectrometer with a solvent A/ solvent B gradient. The salt concentration is increased gradually in each step so that different sub-population of peptides are resolved chromatographically and detected by the mass spectrometer. **Figure 2.2** provides the schematics of peptide separation by chromatography used in all 2D-LC-MS/MS experiments.
- A typical salt concentration scheme widely used in the proteomics community generally starts with 25 mM of ammonium acetate and increases by 25-50 mM increments, topping out at 500 mM in the last pulse [50, 60-63]. Thus, the analysis consists of eleven strong cation exchange steps followed by a ~ 2-hour reverse phase gradient, constituting an approximate runtime of 22 hours. However, reducing the initial concentration of salt to smaller incremental windows provided a better resolution in chromatography resulting in enhanced proteome coverage. This aspect will be discussed in broader detail in chapter 3.

  Although, the above presented method of sample analysis has a 20-fold increase in run-time, it also increases sampling time, reduces noise and improves the detection sensitivity, especially for low abundant peptides.

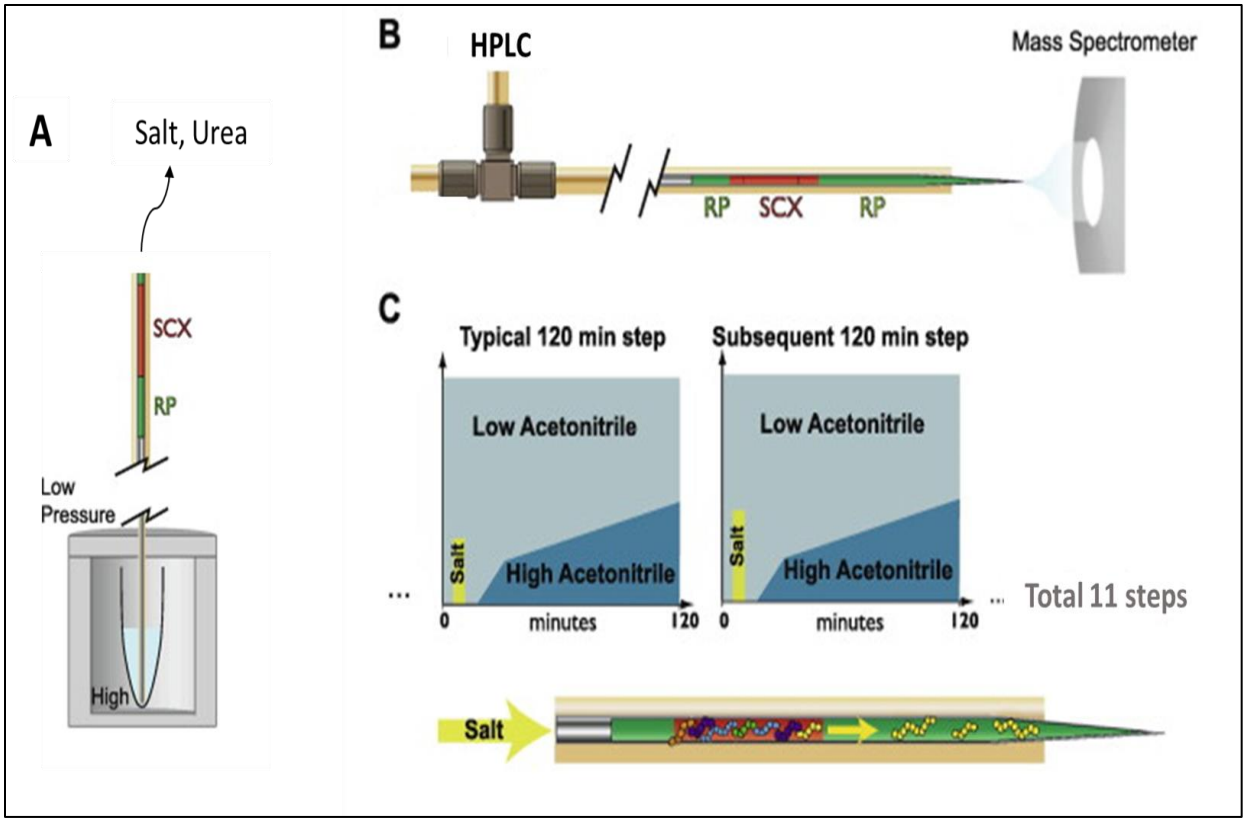


Figure 2.2: Chromatographic Separation of Peptides

(A) *Sample loading; peptides from digested protein sample are pressure loaded under onto a 150 µm inside diameter microcapillary tube packed with SCX and RP. The column is washed offline that removes the residual urea or salt that may present in the sample. This step also helps in desalting of sample during loading and translocates the bound peptides from the RP onto the SCX resin.* ***(B)*** *Peptides bound to the SCX resin are resolved chromatographically to RP and then eluted directly to the mass spectrometer.* ***(C)*** *Gradual elution of sample; peptides are eluted gradually over a 22 hr period to achieve near total identifications. Each ~120 min chromatography step uses a series of increasing salt concentrations to transfer the peptides from SCX to the RP resin. The lower diagram demonstrates the peptides (yellow) moving from SCX (red) to the RP (green) resin. After each salt step, the acetonitrile gradient is used to elute the peptides from RP to the mass-spectrometer for analysis and detection. Adapted from fig. 3 Banks et al (*10.1016/j.pep.2012.09.007*)*

2.4 Introducing Charged Particles to the Mass Spectrometer:

Matrix-assisted Laser Desorption/Ionization (MALDI) [64] and Electrospray ionization (ESI) [8] are the two major ionization modes for introducing charged analytes from the liquid or solid phase into a gas phase. MALDI uses a laser beam to dislodge analytes and create gas phase ions, while ESI uses a high voltage potential to desolvate the ions containing liquid droplets to the gas phase ions. The MALDI approach mostly generates ions of +1 charge state while ESI gives rise to multiply charged ions. MALDI is more amenable for top-down proteomics involving intact protein measurements while ESI can be easily coupled with on-line HPLC and tandem mass-spectrometry and is more suited for bottom-up proteomics approach. All the experiments outlined in this dissertation were carried out in the ESI mode.

2.4.1 The Process of Electrospray Ionization:

Electrospray Ionization (ESI) is an ionization technique best suited for the study of biological molecules. It is a comparatively a gentler ionization technique as it yields little fragmentation, and it can be coupled with tandem mass spectrometry to obtain structural information of peptides and proteins. In this method, the peptide sample is passed as a fine spray under a high electric field. This electric field forms a cone (called the Taylor Cone [65]) which is enriched in positive charges. The unstable end of the cone breaks up to yield a fine spray of positively charged droplets. These droplets are constantly losing solvent in flight due to evaporation and therefore shrink in volume. Finally, before reaching the so called Rayleigh limit [66], where charge balances the surface tension, these droplets undergo coulomb explosion [67] where

these droplets break to form smaller droplets which undergo desolvation to yield gas phase ions that enter the mass spectrometer for analysis **(Figure 2.3)**.

In the ESI mode, the tryptic peptides are maintained in an acidic environment (by addition of formic acid) to form multiply charged peptides. This multiple charging yields an effective mass range to the mass spectrometer and analysis of compounds with MW 100,000 Da is also possible. This is because multiple charging results in reduction in mass to charge ratio and allows us to measure a peptide whose MW exceeds the possible mass range that can be measured by the instrument. This also enables us to perform multiple measurements of molecular mass of the compound. The flow rates in electrospray can be varied from 1 nL/min-1 µl/min (nanospray) to 1-100 µl/min (microspray) to 100 µl/min-1mL/min (macrospray). A high voltage of 3-5 kV is applied to the edge of the capillary. Generally, the peptides to be analyzed are maintained in a mixture of water and organic solvents like acetonitrile as it helps in better drying of droplets to generate gas phase ions. The analytes suitable for ESI include charged inorganic anions and cations, organic acids and bases. Even polar neutral and non-polar neutral species are quite amenable to ESI.

One of the biggest advantages of ESI is that it can be coupled with online liquid chromatography techniques. Also, the sample can be maintained in a liquid state and hence volatility and thermal stability of the molecule can be maintained to generate gas phase ions. It can also be coupled with tandem MS to study amino acid sequence of proteins and study non-covalent interactions like post-translational modifications. However, ESI may result in photon transfer reactions which may result in species with higher proton affinity to retain charge. This may result in subsequent loss of valuable information as peptides that lose charge may not be

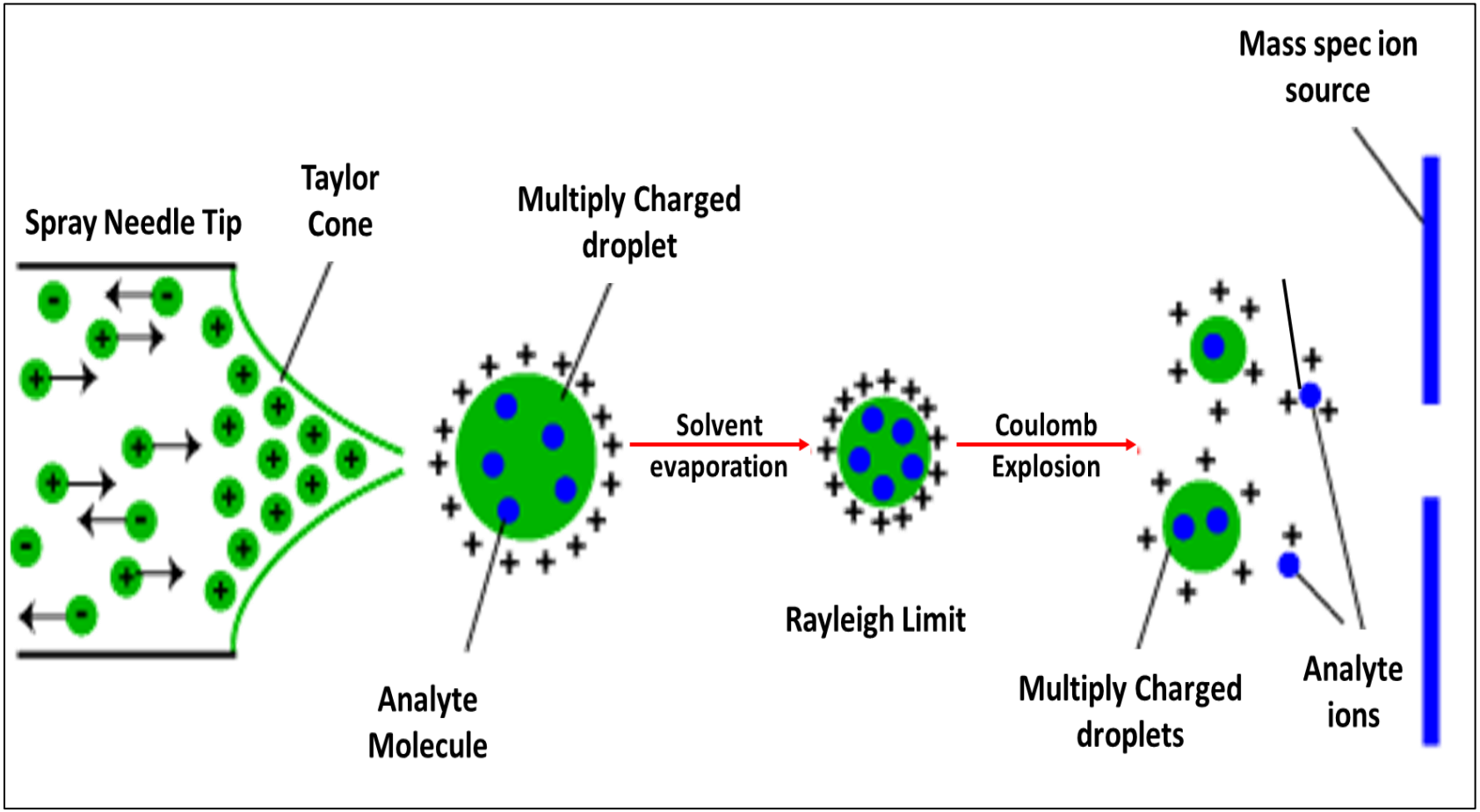


Figure 2.3: Schematics of Electrospray Ionization

measured by ESI. Multiple charging though useful to determine higher MW peptides may also prove to be a burden as the instrument needs a high mass resolving power to differentiate closely related species.

The proteomics experiments described in this dissertation were carried out in the nanospray ionization mode. The reduction in flow rate achieved by nanospray leads to enhanced reduction in droplet size which increases the sensitivity of peptide measurements.

2.5 Analysis of Peptide Ions Inside the Mass Spectrometer:

2.5.1 Mass analyzer:

Various methodologies exist for the analysis of peptides by generating their MS/MS spectra. The description below outlines the principles involved in mass analysis of the Thermo Scientific Linear Trapping quadrupole (LTQ) mass spectrometer. Before mass analysis can begin, the charged particles passing through entrance of the mass spectrometer (at a pressure of 760 Torr) are transported through a series of chambers that are sequentially at lower pressures (~20 µTorr). Ions are then stored for a brief period in a mass analyzer and are then sequentially ejected from the mass analyzer according to their mass/charge values and finally detected by dynode/multiplier type detection system.

The LTQ mass analyzer consists of a series of four ~6 cm long hyperbolic rods. These rods are cut into three sub-sections that are insulated from one another **(figure 2.4)**. This creates a central quadrupole in which ions are trapped between two end quadrupoles. A DC voltage is applied to each section with the central section at -14 V and two end sections at +20 V thus

creating a potential difference where ions can be trapped efficiently. The DC voltages applied to end sections are adjusted to -12V so that positively charged ions can enter the trap. Once the ions move to the central section of mass analyzer, the DC voltage at the end sections is increased, thus trapping a sub-population of positively charged peptide ions **(figure 2.4).**

Additional AC voltages are then applied across the rods of the central section so that ions can be held in stable trajectories. By manipulating these voltages, ions of a given mass/charge (m/z) can be destabilized so as to be ejected through a pair of slits in two of the central section rods. This selective ejection and subsequent detection gives rise to 'full spectrum MS' which provides information about the relative abundance of peptide species with different m/z values [68].

The next generation of instruments in the field of discovery proteomics was the LTQ Velos, which is a dual cell linear ion trapping instrument **(figure 2.5)**. Compared to LTQ, the LTQ Velos features a novel ion transmission pipeline that significantly enhances transfer efficiency. It also consists of a dual trap, one for ion trapping at high pressure, which improves trapping efficiency by 90%. The second trap that is coupled to ion detection system is operated at low pressure that significantly improves measurement resolution [69].

Further enhancement in mass accuracy can be achieved with mass spectrometers that have sophisticated detection systems such as the Orbitraps [70]. Orbitraps can be considered as a modified form of quadrupole ion trap, although orbitraps use a static electrostatic field as compared to quadrupole ions traps that uses a dynamic electrostatic field typically oscillating at ~1MH_z. The orbitrap's axially symmetrical electrodes create a combined 'quadro-logarithmic' electrostatic potential. In an analogy, orbitraps mass analyzers share the principle with yet

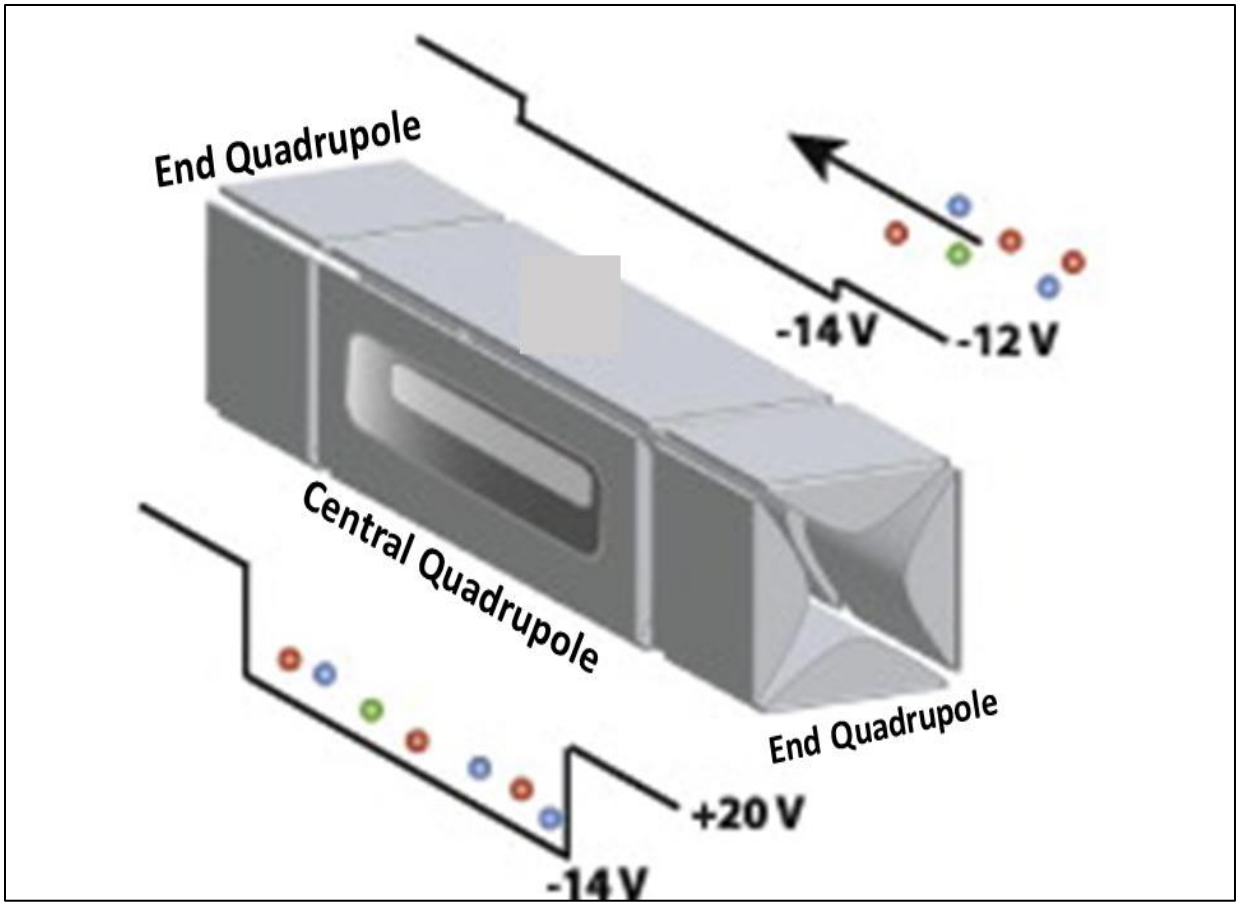


Figure 2.4: Block Diagram of Linear Trapping Quadrupole

Adapted from figure 5 Banks et. Al (10.1016/j.pep.2012.09.007)

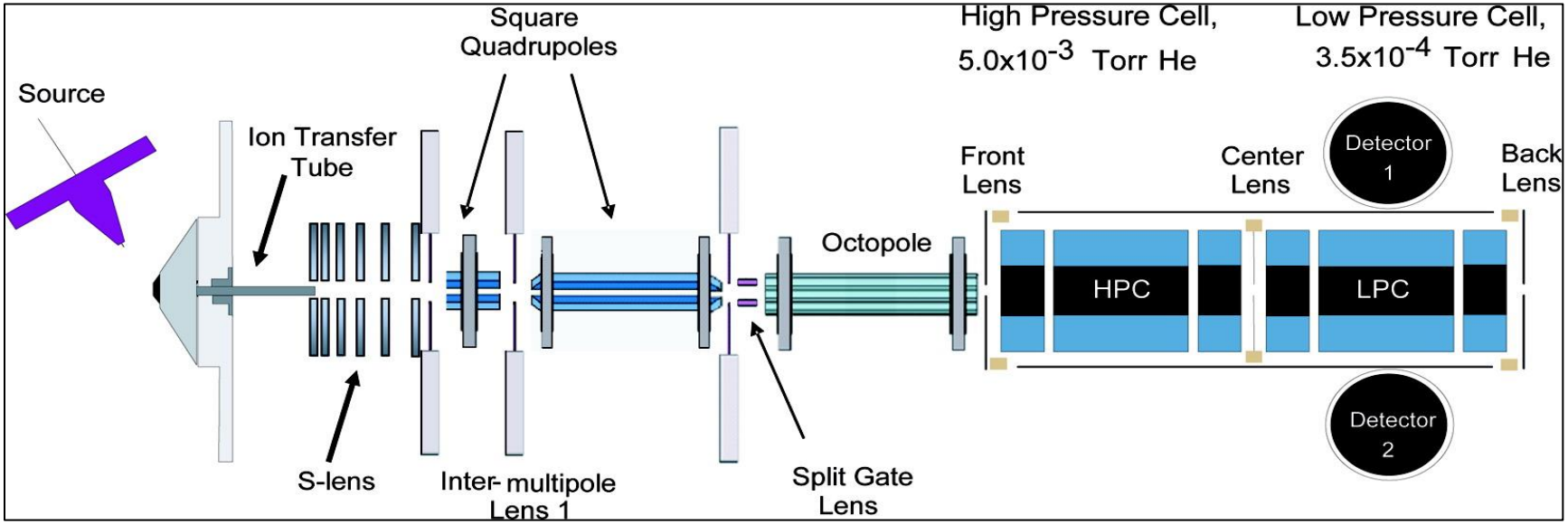


Figure 2.5: Block Diagram of LTQ Velos

Image source: Tonya Pekar Second; Justin D. Blethrow; Jae C. Schwartz; Gennifer E. Merrihew; Michael J. MacCoss; Danielle L. Swaney; Jason D. Russell; Joshua J. Coon; Vlad Zabrouskov; Anal. Chem. 2009, 81, 7757-7765.DOI: 10.1021/ac901278y)

another type of mass analyzer, the Fourier transform ion cyclotron resonance (FT-ICR). In FT-ICR, ion detection is performed by a broadband image current detection, followed by a fast Fourier transform (FFT) algorithm [71] that converts the recorded time domain signal into a m/z spectrum [72] . Here, an axial frequency is used which is independent of the energy and the special spread of the ions. High efficiency in mass resolution and mass accuracy is achieved in orbitraps because of this energy independence. Briefly, the orbitrap is a system of axially symmetrical mass analyzer that consists of a spindle shaped inner electrode and a barrel shaped outer electrode. Migration of ions takes place from linear ion trap into a curved gas-filled ion trap (c-trap). Ion trajectories are considerably slowed down by a nitrogen gas bath system which diverts the ions orthogonally into the space between outer and inner electrodes. Ion oscillation inside the analyzer is achieved around the central electrode due to the application of electrostatic field **(Figure 2.6)**.

All the mass spectrometry measurements described in this dissertation were performed by one of the three instruments described above. The increase in sample complexity has fueled the demand for instruments having higher sensitivity coupled with mass accuracy and speed. Some of the new generation mass spectrometers include the Q-Exactive hybrid quadrupole-Orbitrap mass spectrometer [73] and Orbitrap fusion [74] that deliver unprecedented molecular analysis of complex samples.

Some of the crucial figures of merits that are used to differentiate various mass analyzers are described below.

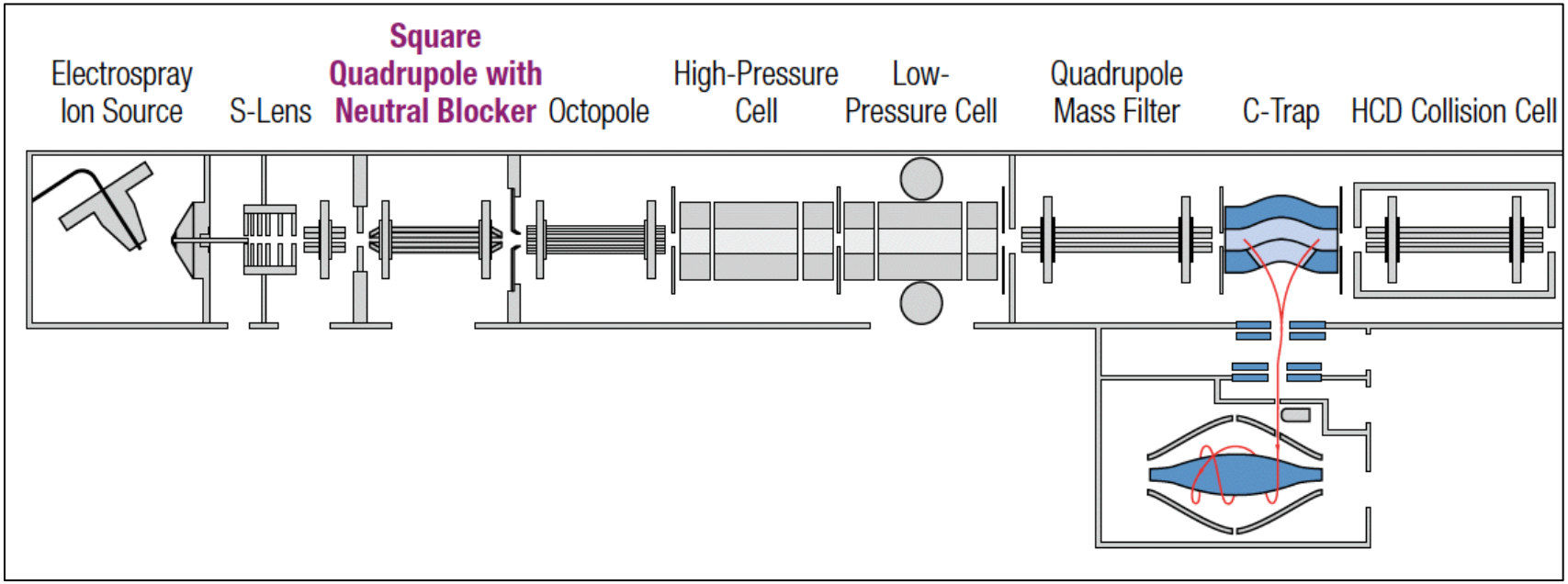


Figure 2.6: Block Diagram of Orbitrap ELITE Hybrid Mass Spectrometer

Image source: (http://planetorbitrap.com/orbitrap-elite#tab:features)

- *Mass Resolving Power:* Mass resolving power or the resolution in mass spectrometry refers to the ability of the instrument to differentiate two peaks having very small difference in mass to charge ratios. Resolution refers to the sharpness of the peak. The sharper the peak, the better is the resolution.

  Resolving power can be illustrated by the following equation:

  Rm = m/ Δm resolution

  Where Δm is the value for minimum peak separation which enables us to distinguish two ionic species. Hence, Δm is the width of the peak measured at 50% fraction of the peak height. It is called "Full width and Half maximum". Resolving power has no units

- *Mass Accuracy:* Mass accuracy refers to the difference between true mass of an ionic species and the measured mass. Here true mass refers to the calculated mass obtained by adding up the masses of each atom in the ion and the measured mass refers to the experimentally obtained value. This difference has to be as small as possible for us to obtain high mass accuracy. It is measured by the following equation

  $\Delta m_{accuracy} = m_{true} - m_{measured}$

  Mass accuracy is measured in parts per million (ppm).

  $ppm = 10^6 [\Delta m_{accuracy} / m_{measured}]$.

  Both resolving power and mass accuracy are important figures of merit in mass spectrometry. Modern day mass spectrometers can differentiate between peaks having very minute differences in mass to charge which allows us to separate closely related species with a high degree of mass accuracy with respect to the true mass.

- *Dynamic Range:* It is the measure of the detection range of the mass spectrometer i.e. the ratio of largest to smallest detectable signal.
- *Mass Range:* It is the range of mass/charge that is amenable to detection by a given mass analyzer
- *Scanning Speed:* Scanning speed refers to the time taken by a given mass analyzer to measure m/z values over a given mass range.

MudPIT measurements consist of complex samples containing tens of thousands of peptides. Hence, it is imperative to employ a mass spectrometer having a high mass resolving power and high scanning speed that can differentiate closely related peptide species as well as resolve a complex proteome within a duration of 22-24 hours that constitutes a typical MudPIT setup.

Table 2.1: Performance metrics of MS instruments described in this dissertation.

Figure of Merit	LTQ-XL	LTQ-Orbitrap XL	LTQ-Orbitrap Elite
Mass resolving power	0.05 FWHM 1000-2000	7,500 - > 100,000 at m/z 400	15,000 - > 240,000 at m/z 400
Mass Accuracy	0.1 Da	< 3 ppm* with external mass calibration < 1 ppm* with internal mass calibration	< 3 ppm* RMS with external mass calibration < 1 ppm* RMS with internal mass calibration
Dynamic Range		>4,000 within a single scan guaranteeing specified mass accuracy	>5,000 within a single scan guaranteeing specified mass accuracy
Mass Range	m/z 15-200 m/z 50 - 2000 m/z 200 - 4000	m/z 50 - 2000 m/z 200 - 4000	m/z 50 - 2000 m/z 200 - 4000

$$*ppm = \frac{\text{Observed Mass} - \text{Theoretical Mass}}{\text{Theoretical Mass}} * 10^6$$

FWHM = Full Width Half Maximum

2.5.2 Basic Steps in a Tandem Mass Spectrometry Experiment:

Once the peptide ions enter the mass spectrometer, they are processed as follows:

- **Obtain a mass spectrum:** In this step, the ions that enter inside the mass spectrometer are used to generate a full mass spectrum and their m/z values are determined to locate the ion of interest
- **Isolate ions with m/z for the component of interest (Precursor ions):** In this method, other ions with m/z values are eliminated and the precursor ion is selected. Generally, the ion having the most abundant m/z value is first selected and then we move down sequentially. We can limit the threshold of ions so that only those ions within a pre-defined range of m/z value are selected for subsequent analysis.
- **Energetically activate or react the isolated (precursor) ion:** In this step the precursor ion is activated by causing a shift in its mass or charge. This is done by colliding the ion with electrons, photons or neutral molecules resulting in its fragmentation.
- **Mass analyze the product ions:** In this step the fragmented product ions are mass analyzed to obtain detailed structural information of the fragmented ion. In case of peptides, mass analysis of product ions can result in determining the amino acid sequence composition of the peptide.

Ion activation refers to increasing the internal energy of the ion resulting in its dissociation. The three most common ion activation methods are as follows:

i. **Collision-Induced dissociation (CID) or Collision-Activated dissociation (CAD):** Here the ions are collided with a neutral atom or molecule. It's important that the mass of the neutral molecule be as big as possible. This results in better fragmentation of the target ion which is then mass analyzed [75]. Hence neutral gases like argon or nitrogen are used. CID is the most common ion activation method used in proteomic studies.

ii. **Activation by Photon absorption: Photodissociation or Infrared multiphoton dissociation (IRMPD):** In this method, the fragmentation of the target ion is achieved by absorption of photons resulting in the target ion getting excited into a more energetic vibrational state. A threshold of this excitation results in cleavage of bonds causing fragmentation [76].

iii. **Activation by electrons: Electron induced dissociation (EID):** In this method, fragmentation of positively charged target ions are achieved by transferring low energy electrons to them. This method is useful for higher charge state ions and is more useful to study PTM's as EID only attacks the peptide backbone leaving the phosphorylation sites intact [77].

2.5.3 Data-Dependent Acquisition of MS/MS Spectra:

Data dependent acquisition (DDA) is used to improve the sampling of low abundant peptides. Typically, the initial full scan MS identifies the m/z values of precursor ions, and MS/MS scans then systematically isolates, fragments and characterizes these ions in order of decreasing

abundance. Now, in order to avoid repeated characterization of the same species, the ions are added to the 'dynamic exclusion (DE) list'. The time duration of exclusion list is user defined (for example, 30 seconds for an ion that has been sampled twice). This way, the ion is barred from repeated selection for a chosen time which allows low-abundant peptides to be sampled [78]. The choice of DE time is a function of chromatographic peak width and had been shown to improve the unique peptide counts in proteomic studies. This process is explained in broader detail in **Figure 2.7.**

2.6 Database Matching of Tandem Mass Spectra:

Each MudPIT run generates hundreds to thousands of MS/MS spectra. Each of these MS/MS spectra is a potential peptide fragment ion resulting from the digestion of a protein. In order to identify peptides, the experimental MS/MS spectra has to be matched with a predicted pattern of the fragment ion so that a list of proteins in the original sample can be identified [79]. Each experimental spectrum provides a characteristic fingerprint of the peptide which is correlated with the peptides derived from *in-silico* digest of the protein database by the same endoproteinases that was used during sample preparation (like trypsin). For this, various computational algorithms are put into use [52, 80-82].

2.6.1 Peptide Nomenclature:

The matching of the experimental spectra with the *in-silico* generated spectra makes use of the peptide dissociation pattern obtained from CID (Collisional induced dissociation). CID preferentially breaks the bond along the backbone of the amino acid chain. A wide variety of product ions are generated because of the bond cleavages along the polypeptide backbone or

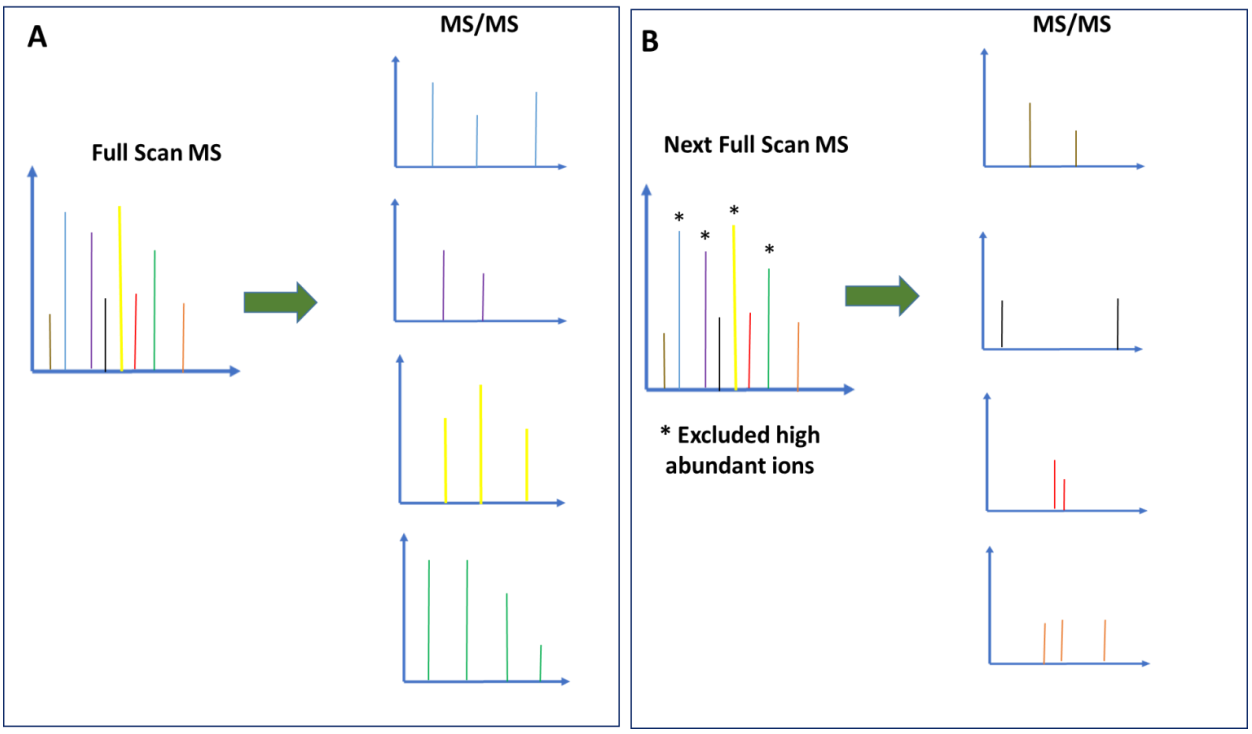


Figure 2.7: Data dependent acquisition.

A) *In the initial full scan, the mass spectrometer screens for the most abundant type of ions that are eluting (represented by blue, purple, yellow and green lines). These ions are then subjected to fragmentation.* ***B)*** *Now, to scout for low abundant peptides, the mass spectrometer puts these ions in the dynamic exclusion list such that these ions (represented by *) are excluded from MS/MS analysis for a chosen time. This in turn allows the analysis of ions represented by red, black, gray and orange lines which are in low abundance and would be selected for fragmentation.*

via partial or complete cleavage of amino acid side chains. **Figure 2.8** depicts the nomenclature used for naming ions formed from cleavages of the peptide backbone. The vertical lines correspond to bond cleavage and arrows signify the product ion. The product ions referred to by letters *a, b, c* correspond to the charge remaining with the N-terminal portion of the peptide ion while *x, y, z* denote product ions in which the charge is retained on the C-terminal portion of the ion [83].

Peptide sequences are, by convention, listed from the N- to the C-terminus, therefore the choice of the *a, b, c* nomenclature is mnemonic as these are the first three letters of the alphabet. Amongst these different types of cleavages, those bond cleavages which occur along the peptide backbone are the most important for sequence determination, particularly the *b* and *y* ions which result from cleavage of peptide bonds. The m/z values of these *b* and *y* fragment ions can be calculated and MS/MS spectrum can be predicted for every possible peptide. Finally, matches between experimental and predicted fragmentation patterns are made and cross-correlation scores are calculated to identify how good a match is. Ambiguous matches are then filtered out before final assembly of peptides to proteins.

2.6.2 Database Mapping of MS/MS Spectra:

The next step in the proteomics experiment is to identify peptides measured by the mass spectrometer and assemble them into proteins. To accomplish this, the tandem mass spectra are computationally matched against the predicted proteome of the species under study. The protein database is a fasta formatted database constructed from the genome of the organism under consideration. This database consists of all the predicted proteins irrespective of its

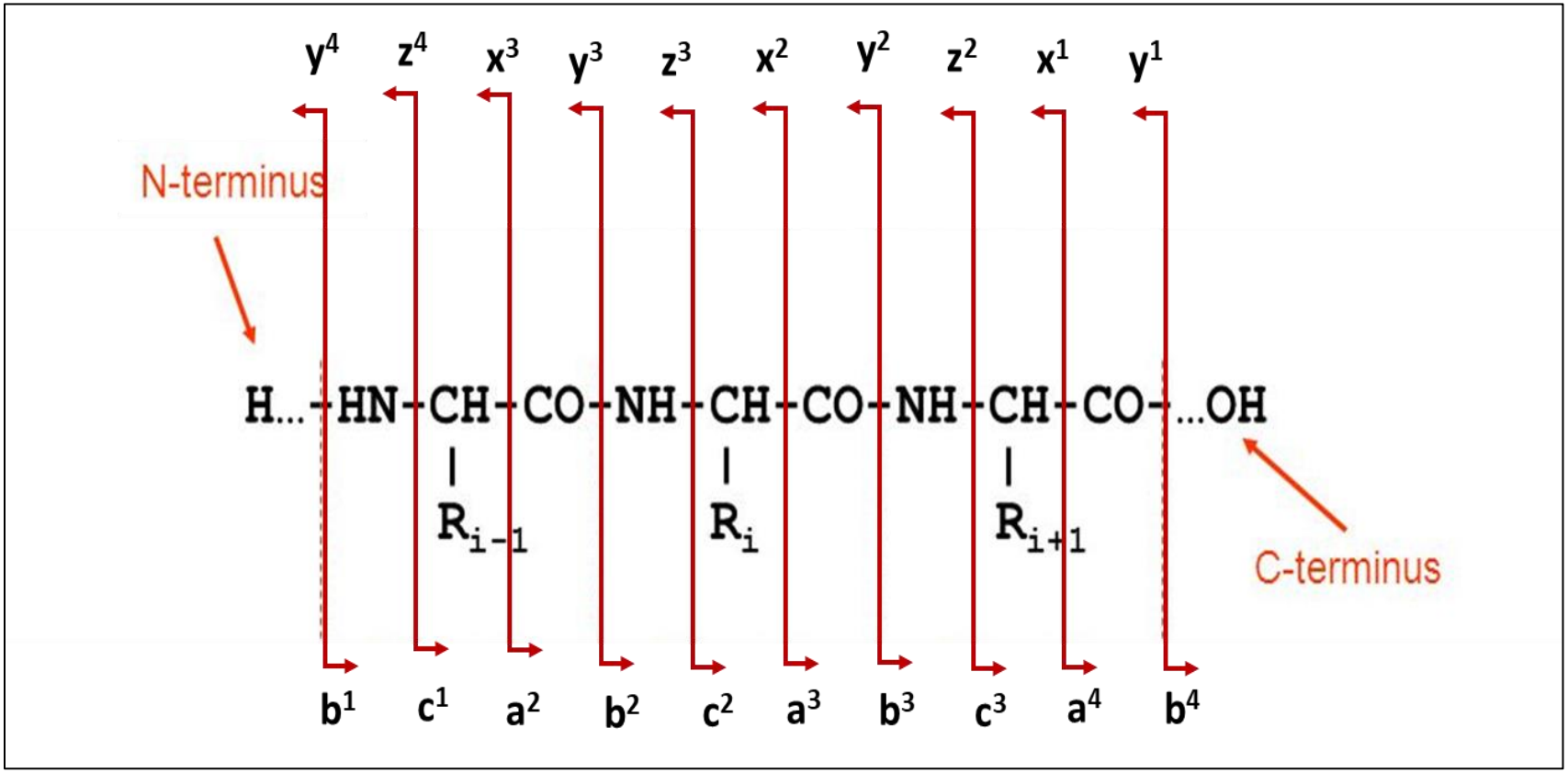


Figure 2.8: Nomenclature for naming the ions formed from peptide backbone cleavages

expression or activity level present in the genome. Since all the protein based interpretations are based on this proteomic database, it Is imperative that genome of the species under study must be completely sequenced. In the absence of sequence information, other protein validation strategies are employed which will be discussed in broader detail in chapter 4.

The first step in database searching involves the *in-silico* digestion of a protein fasta file using the same endoproteinases that were used during sample preparation (like trypsin). In the next step, theoretical spectra are generated by the computational algorithm with *b* and *y* type ions. Then, a pattern matching strategy is employed where experimental and theoretical spectra are matched with one another. The matched spectra are assigned a cross-correlation score called XCorr. In calculating XCorr, the top candidate peptide sequences are subjected to a "Barcode" analysis, where *in-silico b* and *y* ion series are predicted and assigned the highest intensities. This creates an *in-silico* barcode. At the same time, the actual measured spectrum is preprocessed where the top peaks are selected and binned in 10 m/z windows. In the next step the *in -silico* and the preprocessed spectrum slide across each other. This procedure is repeated for all the peptide candidates and the top match is calculated amongst these and the top XCorr is reported in the final output. Each XCorr is assigned a value called deltCN that denotes the distinct match with the next best match. Finally, if both XCorr and deltCN pass a user defined threshold, the peptide spectrum match (PSM) is considered as a valid identification. Each PSM matches to a single peptide which is then computationally assembled to a protein candidate. There are several open source search programs that perform peptide sequencing to identify confident PSM's, including SEQUEST [52], X! Tandem [84], MS-GF+ [85] , MASCOT [80] , MyriMatch [86], Comet [87] , Andromeda [88], OMSSA [82] etc.

2.6.3 Controlling False Discovery Rate:

False discovery rate (FDR) is the rate at which false positive identifications are made. FDR is used to ascertain the authenticity of MS measurements using the database search strategy. The FDR is calculated using the following equation.

$$\frac{2 * \text{Decoy ID's}}{\text{Total ID's}} * 100$$

Decoy ID's correspond to the protein sequences that do not belong to the sequenced organism under study. These decoy proteins can be either the reverse protein sequences or random sequences from an unrelated organism. This method of FDR calculation proposed by Elias and Gygi [89] assumes that, for a given number of decoy hits that pass a pre-defined threshold, there are equal number of false hits in the target sequences. Since false hits are added to target hits in the denominator, the number of false hits are doubled and hence a factor of 2 is added to the numerator containing decoy hits. The use of reverse or a randomized protein database that is concatenated to the actual protein database increases the total computational time but it helps in removing false positives and provides a FDR percentage (for example 1% FDR means 99% confidence that proteins were correctly identified) which is valuable in ascertaining the quality of MS data as well as the protein database [90].

With the ever-increasing size of the database used in metaproteomic investigations, using the standard filtering thresholds may result in substantial loss of PSM's especially those

corresponding to low abundant peptides. In such cases, manual validation of PSM's are carried out which will be discussed in chapter 6.

2.6.4 Assembling Peptides to Proteins:

Once the PSM's are identified, the final computational task is to assemble the candidate peptides to get a list of protein identifications. DTASelect [91], IDPicker [92], PeptideShaker [38] are commonly used protein assembly algorithms used for this task. Here, within a user defined threshold of FDR and a minimal requirement of either one or two distinct peptides (a distinct peptide is basically a distinct sequence of amino acid) for a protein call, the algorithm assembles peptides back to protein and reports them.

2.7 Quantitative Evaluation of MudPIT Data:

Semi quantitative proteomics, also called label-free proteomics, refers to an approach where the intensities of the entire repertoire of proteins (as peptides) present in the sample is detected by LC-MS/MS. The very nature of this approach leads to variation in identification of peptides [93]. This variation is influenced by many factors such as run to run variation, fragmentation and detection differences. Hence, the data is normalized to estimate the relative abundances of proteins identified in MudPIT analysis. There are several approaches for data normalization [94-96], and one such approach uses the normalized spectral abundance factor (NSAF), discussed below, to gauge relative protein abundances present in the sample [94] .

Computational analysis of data from MudPIT runs yields several variables that can be used for calculating protein abundances. These include the total number of proteins detected (N) and

the total number of fragmentation spectra matched to a specific protein which is also referred to as spectral count (*SpC*). The spectral abundance factor (SAF) for a particular protein K is calculated as $SAF_K = (SpC/L)_K$ where L is the length of the protein. SAF allows ranking of proteins in each run by its abundance. Now, a larger protein might give rise more tryptic peptides. Hence, to compare relative abundances of a protein in different mass spectrometry runs, the normalized spectral abundance factor is calculated.

$$\mathbf{NSAF_K} = \frac{(SpC/L)_K}{\Sigma^{n}_{i=1} (SpC/L)_i}$$

The denominator in this equation (which is the summation of spectral counts of all the proteins divided by their lengths) reconciles run-to-run variation in total spectral counts. NSAF is a common method for normalizing protein abundances under different growth conditions. All the MudPIT data described in this dissertation were normalized using NSAF. The resulting NSAF values are then balanced by multiplying them with a common factor and converted to normalized spectral counts (*nSpC*) and used for reporting purposes.

Chapter 3 - Optimization of Salt Pulse Step Elution Conditions for Improved Depth and Enhanced Coverage of Unique Peptides in Multi-dimensional LC-MS/MS Proteome Measurements

Text and figures were taken from: **Ramsunder Iyer**, Richard J. Giannone, Rose S. Kantor; Jill F. Banfield; Robert L. Hettich. Optimization of Salt Pulse Step Elution Conditions for Improved Depth and Enhanced Coverage of Unique Peptides in Multi-dimensional LC-MS/MS Proteome Measurements. *Proteomics, Manuscript in Preparation*

Ramsunder Iyer's contributions to this work included: Experimental design, performed all the proteomics sample preparation and mass spectrometry runs, data analysis, wrote, edited and revised the manuscript.

3.1 Separation by Two-Dimensional Chromatography to Enhance Unique Peptide Measurements:

A deep and comprehensive analysis of a complex proteome sample requires a high-performance mass spectrometric approach coupled with an efficient and robust separation protocol for resolving the peptide mixture prior to detection. The main goal of this study was to analyze complex peptide separation profiles on samples of varying complexities to enhance unique peptide identifications, and then optimize the approach for an online multi-dimensional chromatographic strategy. The method focuses on a modified MudPIT (multidimensional protein identification technology) salt pulse scheme, which exploits multiple shallow salt fractions of ammonium acetate concentrations in the first stages of chromatographic separations. In order to test the range and robustness of the modified scheme, we tested its performance characteristics with a range of complex environmental samples ranging from an isolate of *Caenorhabditis elegans*, a model mixture of six environmental microbes *(Saccharomyces cerevisiae, Escherichia coli, Clostridium thermocellum, Ignicoccus hospitalis,*

Nanoarchaeum equitans and Streptomyces eurocidicus and referred to as *6iso*), and a complex environmental microbial community sample capable of degrading thiocyanate. In general, we obtained a dramatic improvement in proteome coverage with this scheme (28% and 41% increase in peptide counts for *6iso* and *C-elegans* respectively) achieved by a more uniform peptide elution along the balanced salt pulses of the MudPIT experiment. This approach was designed to reduce incidences of high density peptide co-elution that overwhelm the analytical capacity of the mass spectrometer, thus impacting the overall depth of measurement. The enhanced depth in proteomic measurements obtained by this method demonstrates the potential of the shallower salt pulse scheme for multidimensional analysis of environmental samples of varying complexities.

3.2 Current Status and Limitations in Peptide Chromatography:

Multidimensional liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) is a common analytical platform for the identification and quantification of peptides and proteins in complex samples. MudPIT is a widely used workflow in which proteolytic peptides are separated online with a biphasic column containing Strong Cation Exchange (SCX) and Reverse Phase (RP), and then measured by High Performance tandem mass spectrometry (MS/MS) [50]. There are several other methods that can be used to separate complex peptide samples for LC-MS/MS analysis such as hydrophilic interaction liquid chromatography (HILIC), perfluorinated reversed phase chromatography, RP-RP with high pH and low pH elution, and mixed bead ion exchange chromatography etc. [97-100]. Multidimensional separations can be carried out either in online or offline modes. Offline separations offer flexibility in higher

resolution separations and enable facile re-analysis of samples, but require more sample handling and potential losses. Online approaches afford ease of automation and reduction in contamination and sample loss, but often require compromised chromatographic operating conditions [101]. Recently, a comprehensive analysis of the yeast proteome was demonstrated in just over an hour using a hybrid mass spectrometer (Orbitrap fusion) with a high MS/MS acquisition speed [102]. However, analysis of complex proteomes in such a short duration may be subject to random sampling events which may result in some irreproducibility in the peptide measurements, especially with previous generation mass spectrometers with sub-optimal duty-cycles. Also, the spectra collected in this shortened run enhance qualitative identifications but may confound quantitative evaluations for samples with increased complexity. While there continues to be a growing development in high scanning speed MS instrumentation, the coupling of robust multi-dimensional chromatographic separation schemes that can complement these types of mass spectrometers is still an area of active research.

To this end, MudPIT has been extensively employed because of its relative ease in sample handling and its ability to identify thousands of proteins in metaproteomic samples of relatively high complexities [103, 104]. Although there are a variety of multidimensional LC approaches, 2D-LC employing SCX and RP continues to be a widely-used method for peptide fractionation. This is due to the ability to exploit two distinct peptide chemical properties (charge and hydrophobicity), thus offering reasonable orthogonality. The approach consists of applying a series of discrete salt steps (referred to herein as salt pulses) in which the concentration of ammonium acetate (a volatile salt) is increased over subsequent steps. The salt step elution causes a fraction of peptides to be displaced according to their charge and trapped on the RP

column. The peptides are then eluted from the RP column using an organic phase gradient, usually acetonitrile. The use of volatile salts in a discontinuous gradient causes less overlap between adjacent fractions, thus reducing the carry over [105, 106]. However, a major bottleneck in SCX-RP MudPIT strategy is its relatively low throughput; many of the experimental protocols provide one measurement per day in order to achieve reasonable depth. A separation strategy that more selectively optimizes salt steps in a MudPIT experiment would not only help improve overall measurement depth, but could also be useful for enhancing the throughput of these measurements.

The utility of MudPIT for discovery proteomics was first demonstrated in the proteome of *Saccharomyces cerevisiae [50, 51, 107]*. Since then, the method has been extensively used for samples differing in range and complexity. The typical protocol generally starts with 25 mM of ammonium acetate and increases by 25-50 mM increments, topping out at 500 mM in the last pulse [50, 60-62, 108]. Each salt pulse translocates a sub-population of peptides to be further separated by reversed phase (RP) HPLC and then detected by tandem mass spectrometry. This analysis typically consists of eleven to twelve strong cation exchange steps followed by a ~ 2-hour reverse phase gradient, thus constituting an approximate runtime of 22-24 hours. Modifications to this elution profile strategy have also been proposed where the initial concentration of 25 mM for the volatile salt has been further scaled down [100, 109, 110]. These variations in salt concentration and reverse phase gradient time durations were chosen based on sample complexity and a desired level of resolution, but have not been explored in systematic detail. More recently, Yates *et al.* reported a modified micro MudPIT scheme for SCX-RP consisting of 39 strong cation exchange steps. In this study, they divided the initial

concentrations of ammonium acetate into smaller incremental windows, starting at 10 mM and ramping up at the later stage to a final concentration of 500 mM [111]. The protein and peptide counts of this new shallower scheme were equivalent to the standard 24 hr MudPIT employing wider and equally spaced windows ranging from 50 mM to 500 mM. Using this method, they were able to reduce the instrument operation time by 40 % by having a shorter 18.5 min reversed phase gradients.

These findings prompted us to focus on a systematic examination of the critical aspects of the peptide separation metrics for these enhanced 2D-LC approaches. We chose to focus on a few important questions that shed light on the utility of shallower scheme for improving the depth of proteomic analysis. (1) Is the shallower scheme able to resolve more unique peptides (peptides having distinct sequences of amino acids) per salt pulse in the chromatographic space, thus contributing to their improved detection by the mass spectrometer? (2) What is the chemical nature of the peptides eluting in the earlier fractions in the shallower schemes, and do they provide biased separations or simply increase the unbiased separation resolution? (3) Can the shallower scheme be optimized to enhance total peptide identifications and thus improve the overall depth of quantitative proteomics for complex environmental samples?

Initial testing on a *Caenorhabditis elegans* sample in our laboratory revealed that smaller incremental windows in ammonium acetate concentration led to a remarkable level of distinctness in peptide elution profiles in the first 2-3 salt pulses. Based on this, we proposed that shallowing the first series of salt pulses would have the greatest impact on unique peptide separations and thus subsequent MS detection. This approach is able to identify more peptides by evenly spacing them across each salt pulse resulting in their better detection by the mass

spectrometer without overwhelming its analytical capacity. This should result in better proteome coverage and thus a more robust quantitative proteomic analysis, especially relative to the generally accepted MudPIT salt pulse scheme used today by numerous laboratories. To this end, we have designed a salt step elution profile strategy to achieve maximum unique peptide elution per salt pulse. We optimized and evaluated the performance of this approach on three different samples of varying complexities using two different MS platforms.

3.3 Materials and Methods:

3.3.1 Protein Extraction and Enzymatic Digestion:

Three different sample types were used for this study. The first was an isolate of *C-elegans*; the second was a mixture of six environmental microbes herein referred to as *6iso* that included *Saccharomyces cerevisiae, Escherichia coli, Clostridium thermocellum, Ignicoccus hospitalis, Nanoarchaeum equitans* and *Streptomyces eurocidicus*. The third sample used for the study was a complex microbial community capable of degrading thiocyanate. **The results pertaining to the thiocyanate degrading microbial community sample will be discussed in chapter 4.** All samples were prepared separately as follows: 1 mg of microbial sample was excised and thawed for cell lysis and protein extraction. These samples were first boiled for 5 min in 1 mL of lysis buffer containing 100 mM Tris-HCl, pH 8.0, 4% w/v SDS (sodium dodecyl sulfate), and 10 mM dithiothreitol (DTT). The suspension was vortexed and sonicated with a Branson ultrasonic cell disruptor (20% amplitude for 2 min, 10 s pulse with 10 s rest). The resulting crude protein extract was precleared via centrifugation at 21000 g and quantified by the BCA assay (Pierce Biotechnology, Waltham, MA). An aliquot consisting of ~1 mg of protein was subjected to TCA

precipitation and subsequent digestion with trypsin using the method described previously [103]. The resulting peptides were quantified by the BCA assay and stored at −80 °C until use.

3.3.2 Nano 2D LC-MS/MS Measurement:

Proteolytic peptide samples were analyzed via an online nano 2D LC–MS/MS system interfaced with either LTQ-Velos Pro or hybrid LTQ-Orbitrap-Elite MS (ThermoFisher Scientific). A 25 µg aliquot of peptides was loaded onto a biphasic silica back-column which was first packed with ∼4.5 cm strong cation exchange (SCX; Lune by Phenomenex) followed by ~4.5 cm reverse phase (C18; Kinetex by Phenomenex). Back-columns were washed offline after sample loading with solvent A (95% HPLC grade water, 5% acetonitrile, 0.1% formic acid (FA) for 20 min, followed by a 25-min gradient of solvent B (70% acetonitrile, 30% HPLC grade water, 0.1% FA). Each back-column was coupled in-line with an in-house pulled, reverse-phase (∼12 cm) packed nanospray emitter and analyzed by eleven or twelve-step MudPIT (multidimensional protein identification technology), as described previously [112]. The salt pulse schemes with their corresponding reverse phase gradient profiles are described in the results and discussions section. Technical quadruplicates for *6iso*, technical duplicates for *C-elegans* and technical triplicates were performed for the community sample. Both the instruments (LTQ Velos Pro and LTQ-Orbitrap-Elite) were operated in a data-dependent mode. For LTQ-Velos Pro, all data-dependent MS/MS was performed in LTQ (top twenty), 1 microscan for both full and MS/MS scans; normalized collision energy 35% and dynamic exclusion time of 15 seconds. For LTQ-Orbitrap-Elite, MS1 was performed in Orbitrap and data dependent MS/MS was performed in LTQ (top twenty), 1

microscan for both full and MS/MS scans; normalized collision energy 35% and dynamic exclusion time of 30 seconds.

3.3.3 Data Analysis:

MS/MS spectra were analyzed using the following software protocol. A decoy database of reversed protein sequences and common contaminants was appended to the two target databases under consideration The fragmentation spectra were searched with the Myrimatch v2.1 algorithm [86] against the appropriate databases. For LTQ Velos Pro, the following configuration parameters were used: fully tryptic peptides with any number of miscleavages, an average precursor mass tolerance of 1.5 *m*/*z*, a fragment mass tolerance of 0.5 *m*/*z*, a static cysteine modification (+57.0214 Da), an N-terminal dynamic carbamylation modification (+43.0058 Da), and a dynamic methionine oxidation modification (+15.9949 Da). For Orbitrap Elite, all search configurations remained same except for the mono precursor mass tolerance, which was set to 10 ppm. Peptide identifications were filtered with IDPicker v3.1 [92] to achieve peptide-level FDR of < 1% (maximum *Q* value < 2%). At the protein level, a minimum of two distinct peptides plus one additional peptide were required per protein call and a minimum two spectra per protein. Protein abundances for a subset of proteins were estimated using normalization of spectral counts as described previously [94, 113]. Briefly, to account for the fact that larger proteins tend to contribute more peptide/spectra, spectral counts were divided by protein length to obtain a spectral abundance factor (SAF). These SAF values were then normalized against the sum of all SAF values in the run, allowing the comparison of protein

levels across individual replicates. These values were then balanced and converted to normalized spectral counts (nSpC).

3.3.4 *De novo* Sequencing:

De novo sequencing was performed using the software DeNovoGUI that provides a graphical user interface and parallelization of the PepNovo+ algorithm [114, 115]. We used the same mass tolerances and PTM parameters as done for the database searching. The default CID fragmentation and tryptic cleavage was specified. The resulting *de novo* peptide suggestions were filtered by a PepNovo+ score threshold above 100 for high-quality identifications as described previously [42]. *De novo* was performed on the thiocyanate degrading microbial community sample and will be discussed in chapter 4.

3.4 Results and Discussion:

3.4.1 The Problem of Front Loading:

The first objective of our study was to improve the depth of proteomic measurements within a 22-hour time frame. For this, we evaluated two different salt pulse schemes: one was a standard salt pulse scheme that was routinely used in our laboratory that consisted of equally spaced incremental windows in ammonium acetate starting at a concentration of 25 mM. We refer to this scheme as the “conventional 22 h” scheme. Peptide elution using this separation profile demonstrated a peculiar problem of front loading. **Figure 3.1A** reveals the total unique peptides of *6iso* from each salt pulse for the conventional 22 h scheme. We defined unique

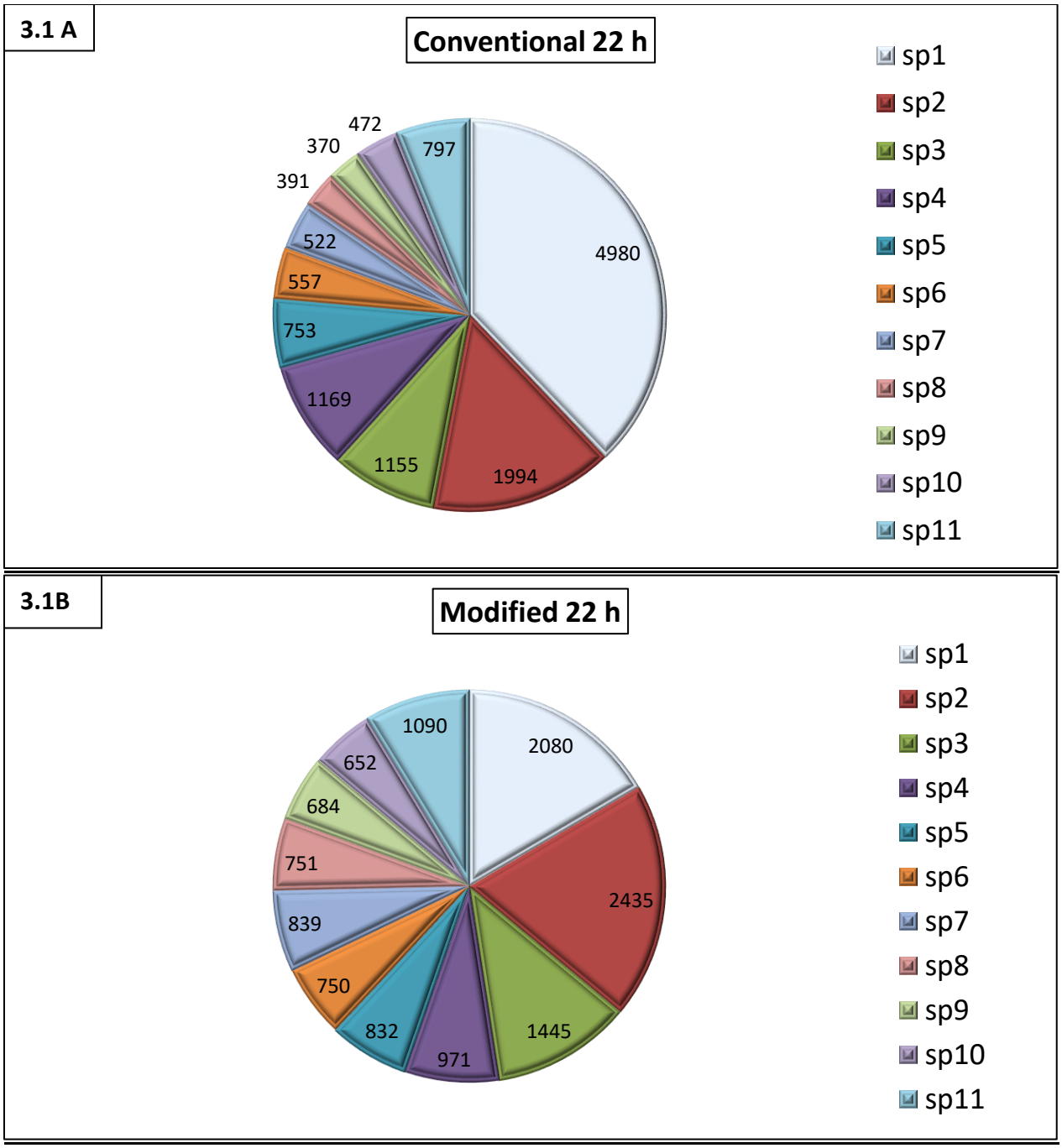


Figure 3.1: Distribution of unique peptides across each salt pulse (sp) for the Conventional 22 h (3.1A) and Modified 22 h (3.1B) schemes for 6iso sample.

Peptides are highly front loaded in the conventional scheme as indicated by the first salt pulse while they are more uniformly distributed in the modified scheme.

peptides as those peptides which differed from each other in amino acid sequence and were detected in only one salt pulse. The peptides observed in the conventional separation scheme indicate a clear egregious front loading of unique peptides. This is especially evident in the first salt pulse which had an overwhelmingly large unique peptide elution profile relative to the other later pulses. This observation prompted us to split the initial salt pulse of 25 mM into smaller incremental windows. After some testing and optimization, we defined a modified scheme (referred to as modified 22 h) that combined the shallow initial pulses starting at a lower concentration of 10 mM with larger steps later in the scheme. The salt steps of the modified scheme were designed to have smaller incremental windows in the initial pulses to allow for more efficient chromatographic separation of the eluting unique peptides. Both the conventional and modified schemes employ a total of eleven salt steps, each followed by a RP gradient with a total acquisition time of 2 h per salt step or 22 h per sample. The ammonium acetate concentrations and RP gradient times of 22 h tested schemes are summarized in the **Table 3.1** and the gradient profile for these 22 h salt pulse schemes are summarized in the **Table 3.2.**

3.4.2 Proteome Metrics of Conventional 22 h vs. Modified 22 h Schemes:

The optimized chromatographic separation scheme, which employs shallower salt pulses at the onset of the analysis was designed to enhance the distribution of peptides across the RP gradient. This should enable increased proteome coverage of complex environmental samples within the same time frame of 22 hours that constitutes a typical MudPIT setup. The *6iso*

Table 3.1: The gradient elution profiles for tested MudPIT schemes

MudPIT schemes	Solvent A	Solvent C or Solvent D	Post salt pulse wash with Solvent A	0-50 % Solvent B	**Total time per pulse**
22 hr scheme (Conventional and Modified)	5 mins	5 mins	5 mins	105 mins	**120 mins**
13 hr scheme (Modified)	3 mins	4 mins	1 min	52 mins	**60 mins**

Solvent A: 95% HPLC grade water, 5% acetonitrile, 0.1% formic acid.

Solvent B:70% acetonitrile, 30% HPLC grade water, 0.1% formic acid.

Solvent C: 50 mM Ammonium Acetate in Solvent A

Solvent D: 500 mM Ammonium Acetate in Solvent A

sample was used as initial benchmark to test the performance of the modified scheme. The total number of peptides, proteins and spectral counts identified in this sample, as measured by the LTQ Velos Pro, are presented in **Table 3.3**. We observed a dramatic improvement in peptide counts (28%) when we switched from conventional 22 h to modified 22 h scheme. This was mainly because the modified 22 h scheme better distributed the complexity of unique peptides across multiple fractions, as evident in **Figure 3.1B**. This is in stark contrast to the conventional 22 h scheme, where the first salt pulse had an overwhelmingly large peptide elution profile (**Figure 3.1A**). In addition to using the *6iso* sample, we initially tested the performance matrix of conventional 22 h and modified 22 h schemes on an isolate sample of *Caenorhabditis elegans.* These samples were run in technical duplicates and we found a striking improvement

Table 3.2: Ammonium acetate concentrations and time durations for the conventional 22 h, modified 22 h and modified 13 h schemes.

Conventional 22 h		Modified 22 h		Modified 13 h	
NH_4Ac concentration (in mM)	**Time duration of each RP gradient (in hours)**	**NH_4Ac concentration (in mM)**	**Time duration of each RP gradient (in hours)**	**NH_4Ac concentration (mM)**	**Time duration for each salt step (in mins)**
25	2	10	2	10	60
37.5	2	15	2	15	60
50	2	20	2	20	60
62.5	2	25	2	25	60
75	2	30	2	30	60
87.5	2	40	2	40	60
100	2	50	2	50	60
125	2	75	2	75	60
175	2	100	2	100	60
250	2	300	2	150	60
500	2	500	2	250	60
				500	120
Total Runtime	22 hrs	**Total Runtime**	22 hrs	**Total Run time**	13 hrs

Table 3.3: Overview of proteomic results from samples measured by the conventional 22 h and modified 22 h schemes for 6iso sample

	Conventional 22 h				Modified 22 h			
Run	Run 1	Run 2	Run 3	Run 4	Run 1	Run 2	Run 3	Run 4
Spectral count	131912	119729	144153	150728	164774	172895	176144	150951
Peptide counts	19409	19016	22217	20841	27324	26695	24998	25488
Protein counts	3697	3972	4149	4007	4699	4583	4417	4350

Table 3.4: Overview of proteomic results from samples measured by the conventional 22 h and modified 22 h schemes for C-elegans sample

	Conventional 22 h		Modified 22 h	
Run	Run 1	Run 2	Run 1	Run 2
Spectral counts	119680	116851	141285	149259
Peptide counts	16546	16834	25493	21431
Protein counts	5943	6019	7198	6904

(41 %) in peptide counts. The total number of peptides, proteins and spectral counts identified in this sample, as measured by the LTQ Velos Pro, are presented in **Table 3.4.**

We next compared the persistence of protein and peptide identification across the technical replicates **(Figure 3.2)**. For *6iso*, 2874 proteins of the total 5014 (57%) were reproducibly identified across all four technical replicates of the conventional 22 h scheme. The reproducibility improved in the modified 22 h scheme, where 3416 proteins out of the total 5557 proteins (62%) were repeated across all four technical replicates **(3.2A).** The overlap at the peptide-level was expectedly less with 38% and 44% across conventional and modified schemes respectively for *6iso* **(3.2B).** Thus, the inventory of reproducible peptide and protein

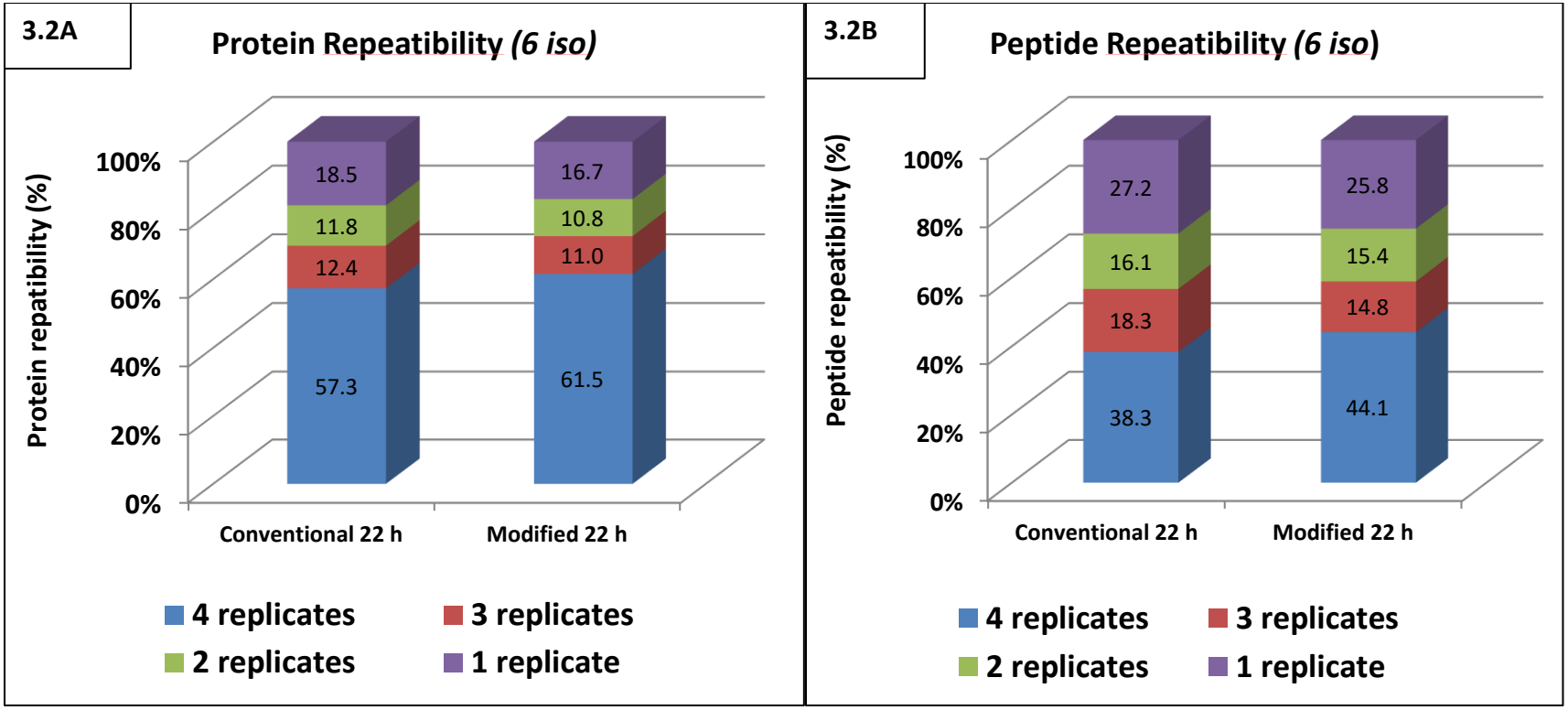


Figure 3.2: Percentage of total proteins (3.2A) and peptides (3.2B) repeating across technical replicates for 6iso sample.

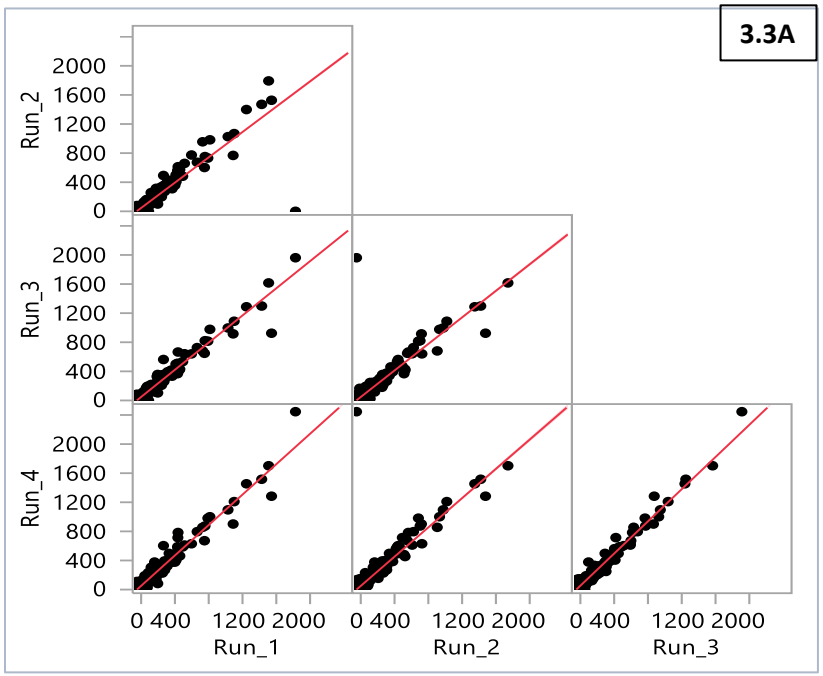
The inventory of repeatable peptide and proteins across technical replicates remained consistent as we moved from one separation protocol to another.

identifications across technical replicates remained consistent from one separation protocol to another. Since data-dependent acquisition introduces a level of stochasticity in discovery proteomics, it is imperative that new separation strategies should be as robust as possible against random sampling events. Since this percentage remained relatively constant, if not somewhat improved, we concluded that the modified scheme did not contribute any further to random sampling events.

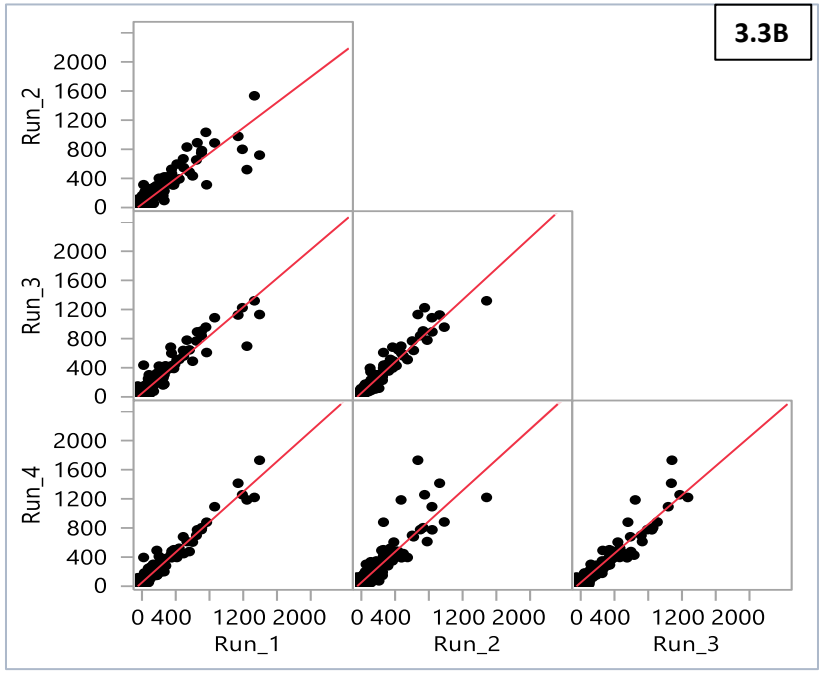
We further sought to assess the overall reproducibility of the conventional 22 h and modified 22 h schemes by looking at semi quantitative values of the normalized spectral counts (nSpC). We found a high degree of correlation within the technical quadruplicates of the *6iso,* as depicted by the scatter plot matrix and Pearson's correlation analysis of nSpC for both the schemes (**Figure 3.3)**. Thus, the new proteins were consistently detected as we switched from the conventional to the modified scheme.

To further illustrate the improvement in dynamic range and better sampling of low abundant proteins for *6iso* sample, we performed a statistical comparison of spectral counts of high and low abundant proteins for both the schemes. Here, we wanted to check if the newly designed scheme was able to improve the detection of low abundant proteins. As expected, the modified 22 h scheme facilitated increased sampling **(Figure 3.4)**, Specifically, we found 304 additional proteins in the modified scheme that had a spectral count of twenty or lower.

In general, highly abundant, proteotypic peptides, i.e. those peptides, easily and most persistently identified by LC MS/MS-based proteomic experiments, tend to mask the less proteotypic ones [116]. However, the more efficient separation strategy employed in the



	Run 1	Run 2	Run 3	Run 4
Run 1	1			
Run 2	0.9221	1		
Run 3	0.9808	0.9042	1	
Run 4	0.9810	0.8909	0.9880	1



	Run 1	Run 2	Run 3	Run 4
Run 1	1			
Run 2	0.9339	1		
Run 3	0.9635	0.9752	1	
Run 4	0.9822	0.9304	0.9679	1

Figure 3.3: Scatter plot matrix and Pearson's correlation values of normalized spectral counts (nSpc) for conventional 22 h (3.3A) and modified 22 h (3.3B) schemes for the 6iso sample.

A high degree of correlation within the technical quadruplicates was found indicating the high reproducibility of the newly designed scheme.

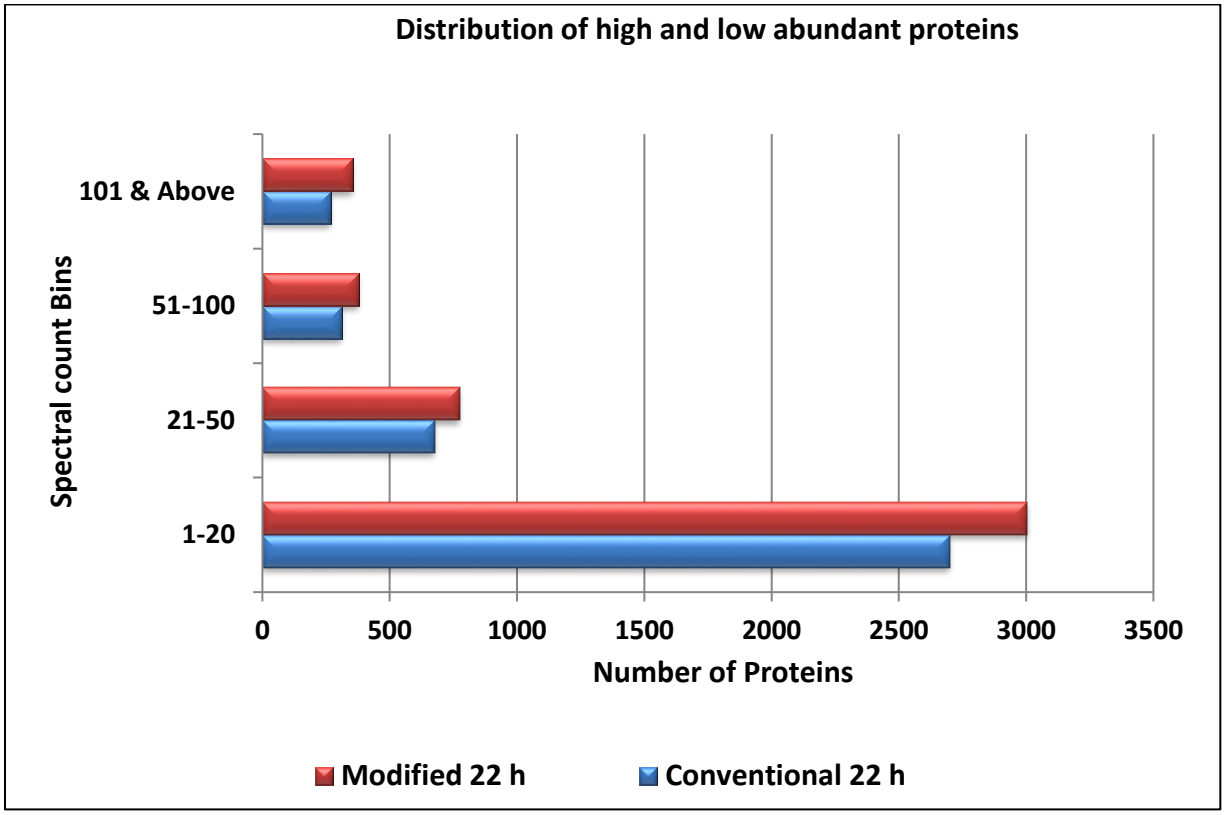


Figure 3.4: Proteome metric improvement in the modified 22 h scheme.

Figure depicts the binning of proteins based on their spectral counts. Proteins having spectral counts in the range 1-20 were considered to be low abundant. The modified 22 h scheme facilitated increased sampling of 304 low abundant proteins.

modified scheme improves the visibility of low abundance proteotypic peptides, improving their sampling rate/reproducibility and thus providing additional information about proteins identified in the experiment.

3.4.3 Gravy Index of Peptides:

We compared the distributions of identified peptides from the conventional 22 h and modified 22 h platforms in different hydrophobic and hydrophilic ranges. This was mainly done to assess the physiochemical properties of eluting peptides between the two tested schemes. A gravy score, which is measure of peptide/protein hydrophobicity, can be used to determine if the peptide is polar or non-polar. A negative gravy score indicates that the peptide is hydrophilic and positive score corresponds to hydrophobic ones [117, 118].

We hypothesized that the new scheme would result in the elution of more hydrophilic peptides in the initial pulses when the concentration of ammonium acetate in lower (0-50 mM). This in turn would enhance the elution of hydrophobic peptides at the latter pulses when salt pulse concentrations are higher (51-200 mM and 201-500 mM). We pooled all the non-redundant peptides of both the schemes and subjected them to analysis. The figure denotes the distribution of hydrophilic **(3.5A & 3.5C)** and hydrophobic peptides **(3.5B & 3.5D)** under different concentrations of ammonium acetate for *C.elegans* and *6iso* samples.

After the application of a salt pulse, the gradient starts with an aqueous phase and gradually shifts towards the organic phase. Peptides that are highly hydrophilic should be displaced initially in the gradient flow when the aqueous phase content is high; otherwise their detection would be averted by the hydrophobic peptides that are eluted during the organic phase. We

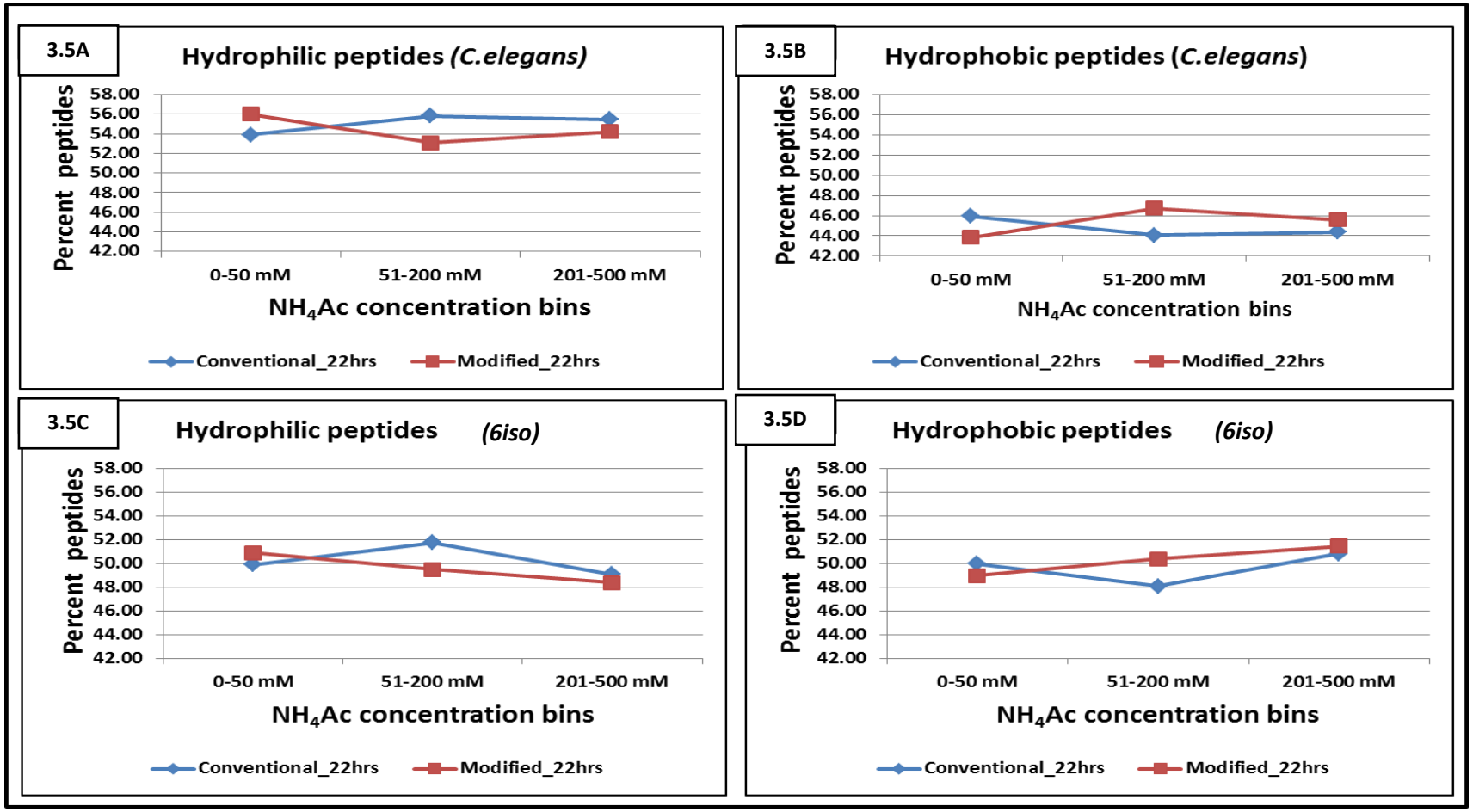


Figure 3.5: Hydrophilic (3.5A & 3.5C) and hydrophobic (3.5B & 3.5D) peptide elution profiles for conventional 22 h and modified 22 h schemes for C.elegans samples and 6iso samples.

Peptides were grouped under three different bins of ammonium acetate concentrations. The modified scheme resulted in the elution of more hydrophilic peptides in the initial pulses when the concentration of ammonium acetate in lower (0-50 mM).

reasoned that the shallower scheme would displace an optimum concentration of peptides across the reverse phase in each salt pulse, thus causing the hydrophilic peptides to be sampled more effectively. On the other hand, having a higher salt pulse (as in the conventional scheme) would displace a large sub-population of peptides, and many of the hydrophilic peptides may not get access to the short aqueous phase and hence would escape detection. The conventional 22 h scheme has only three salt pulses under 50 mM. The modified 22 h scheme on the other is broken down into seven salt pulses under 50 mM. As indicated in the figures below **(3.5A & 3.5C)**, there is a slight enhancement in the sampling of hydrophilic peptides in the modified 22 h scheme.

3.4.4 Percentage and Numerical Gains in Total Peptides between the Conventional 22 h and Modified 22 h Schemes:

The 2D analysis of tryptic digests of *C-elegans* resulted in the cumulative identification of 47,943 redundant peptide sequences across the eleven pulses of the conventional 22 h scheme. The modified 22 h scheme revealed 62,656 peptides. Similarly, for *6iso* it was 70,735 and 103,179 for the conventional 22 h and modified 22 h respectively -an increase of roughly 24% and 46% for *C-elegans* and *6iso* respectively. These drastic shifts in peptide counts prompted us to analyze the depth in peptide detections across individual fractions. **Figure 3.6** depicts the percentage and numerical gains in peptide counts per salt pulse as we move from the conventional 22 h to the modified 22 h scheme for *6iso* **(3.6A)** and *C-elegans* **(3.6B)** samples. The shift of the bar towards the negative axis indicates that the conventional scheme gave more peptide identifications for that particular salt pulse while a shift towards the positive axis indicates increased identification rates in the modified scheme. As shown, the modified scheme

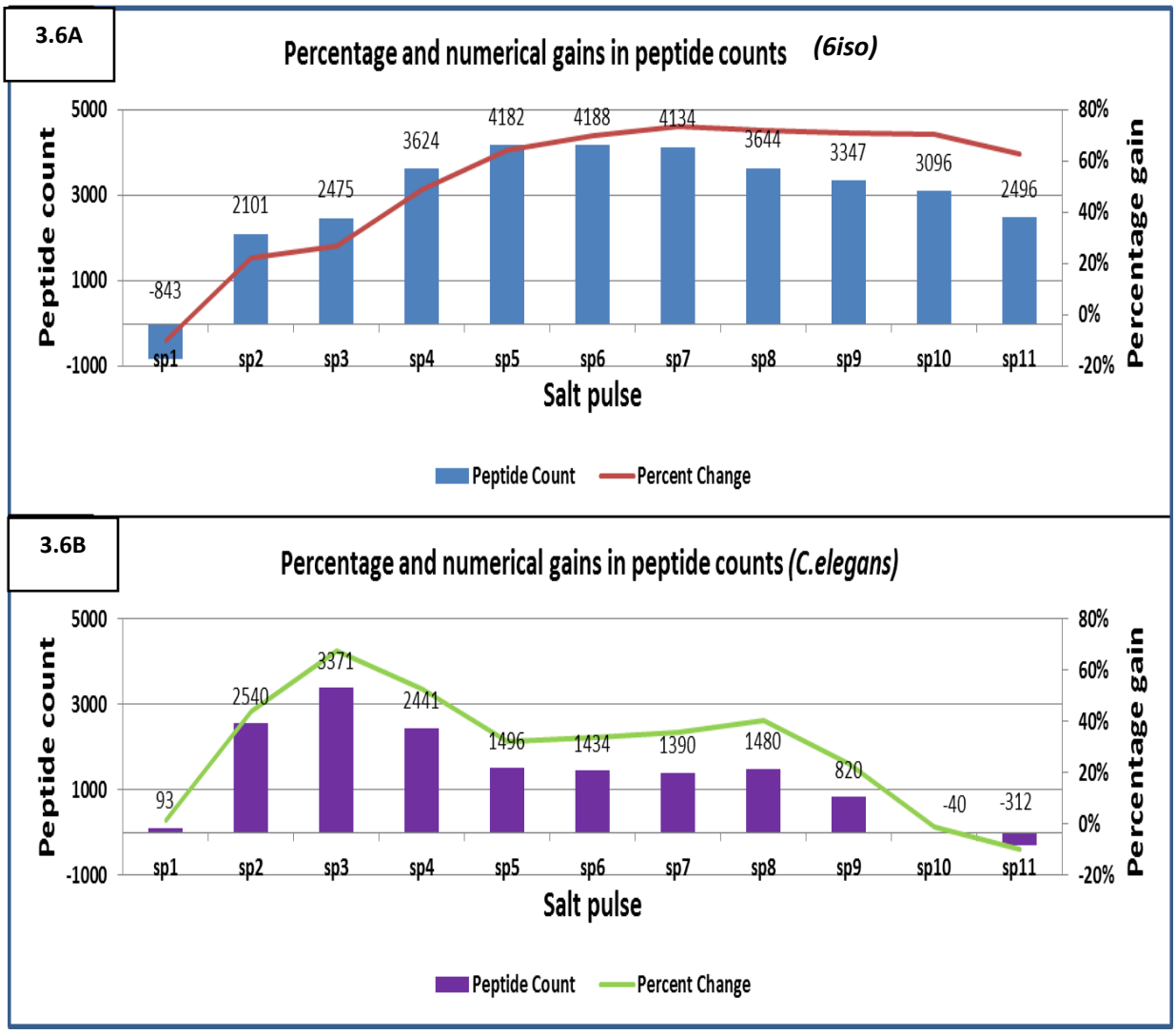


Figure 3.6: Numerical gains in peptide counts between the conventional 22 h and modified 22 h schemes for 6iso (3.6A) and C.elegans (3.6B) samples.

. The negative bar indicates that conventional scheme gave more peptides for that particular salt pulse. As indicated, the modified scheme outperformed the conventional scheme in most of the salt steps

outperformed the conventional scheme in all the salt pulses except the first one. This was expected because the conventional scheme was highly front-loaded in the first salt pulse and was broken into smaller incremental windows in the modified scheme, resulting in more uniform peptide elution profiles.

In a 2D LC-MS/MS analysis employing SCX and RP, the displacement of peptides from the charged SCX resin to the hydrophobic reverse phase resin is carried out by a volatile salt. The conventional scheme, which employs a higher concentration of ammonium acetate upfront, causes the displacement of a larger sub-population of peptides to the reverse phase resin for gradient separation and concurrent MS/MS. This increased complexity impacts overall peptide identification, as the mass spectrometer generally has a fixed duty cycle that can limit its capability to detect all peptides eluting at any given time. Thus, a larger subpopulation will likely lower the probability that all will be sampled. The modified scheme, however, causes a more uniform displacement of peptides in each salt pulse thus increasing the probability of complete sampling and leading to the gains observed here.

3.4.5 Design of Shorter MudPIT Schemes for Improved Sample Throughput:

The second objective of our study was to evaluate the impact of shallower scheme for improving the throughput of MudPIT measurements. The enhanced chromatographic separation, by virtue of the shallower salt pulse scheme, should result in a more even distribution of peptides across the reverse phase, thus permitting the use of a steeper RP gradient to perhaps achieve the same level of depth as a 22 h MudPIT. This in turn should reduce the overall instrument operation time, allowing for increased throughput. To test this,

we designed a twelve-step modified salt-pulse scheme with each pulse followed by a RP gradient of 75 min, 60 min or 45 min thus employing a total runtime of 15 h, 13 h and 9.5 h respectively. These shortened time schemes were chosen based on the desire to achieve higher throughput in MudPIT measurements without impacting the overall proteome metrics.

6iso sample was used as a benchmark to evaluate these results on the LTQ Velos Pro mass spectrometer. The ammonium acetate concentrations and RP gradient times of 13 h tested schemes are summarized in the **Table 3.1** and the gradient profile for these 13 h salt pulse schemes are summarized in the **Table 3.2**. The protein and peptide counts of the technical quadruplicates of *6iso* for the conventional 22 h and modified 13 h schemes are illustrated in **figure 3.7**. Amongst the shorter MudPIT schemes, the modified 13 h scheme gave near identical peptide counts to the conventional 22 h scheme and worse than the 22 h modified scheme. We also studied the distribution of protein groups within the technical replicates for conventional 22 h and modified 13 h methods **(Figure 3.8)**. Here, proteins were binned based on the number of distinct peptides by which they were identified. We found a high degree of overlap in protein group distribution within the conventional 22 h and modified 13 h schemes for proteins identified by at least 5 distinct peptides – likely the more abundant proteins in the sample. Together, the four conventional 22 h MudPIT runs identified a total of 30428 distinct peptides and the four modified 13 h MudPIT runs identified 28052 distinct peptides. The extent of enhancement in duty cycle achieved in the modified 13 h scheme is similar to that achieved by Yates *et. al.* in their modified MudPIT strategy [111]. Thus, reducing the time by 9 hours using the shallower scheme does not have much of a negative impact on sampling of proteomic measurements, as the counts remained more or less the same.

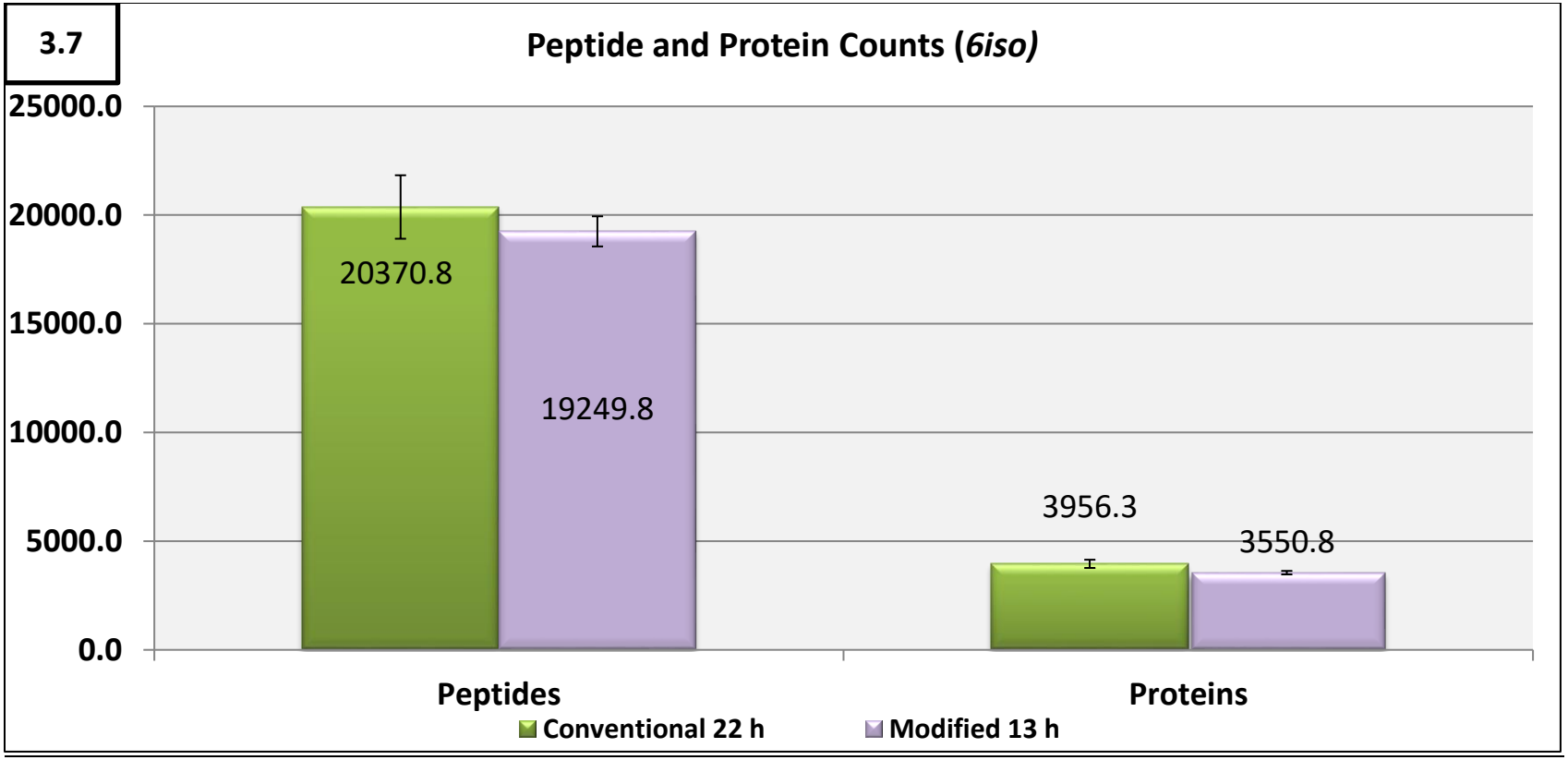


Figure 3.7: Peptide and protein counts of 6iso tryptic peptides shown on a linear scale for conventional 22 h and modified 13 h schemes.

Error bars indicate standard deviation (s.d); n=4 technical replicates.

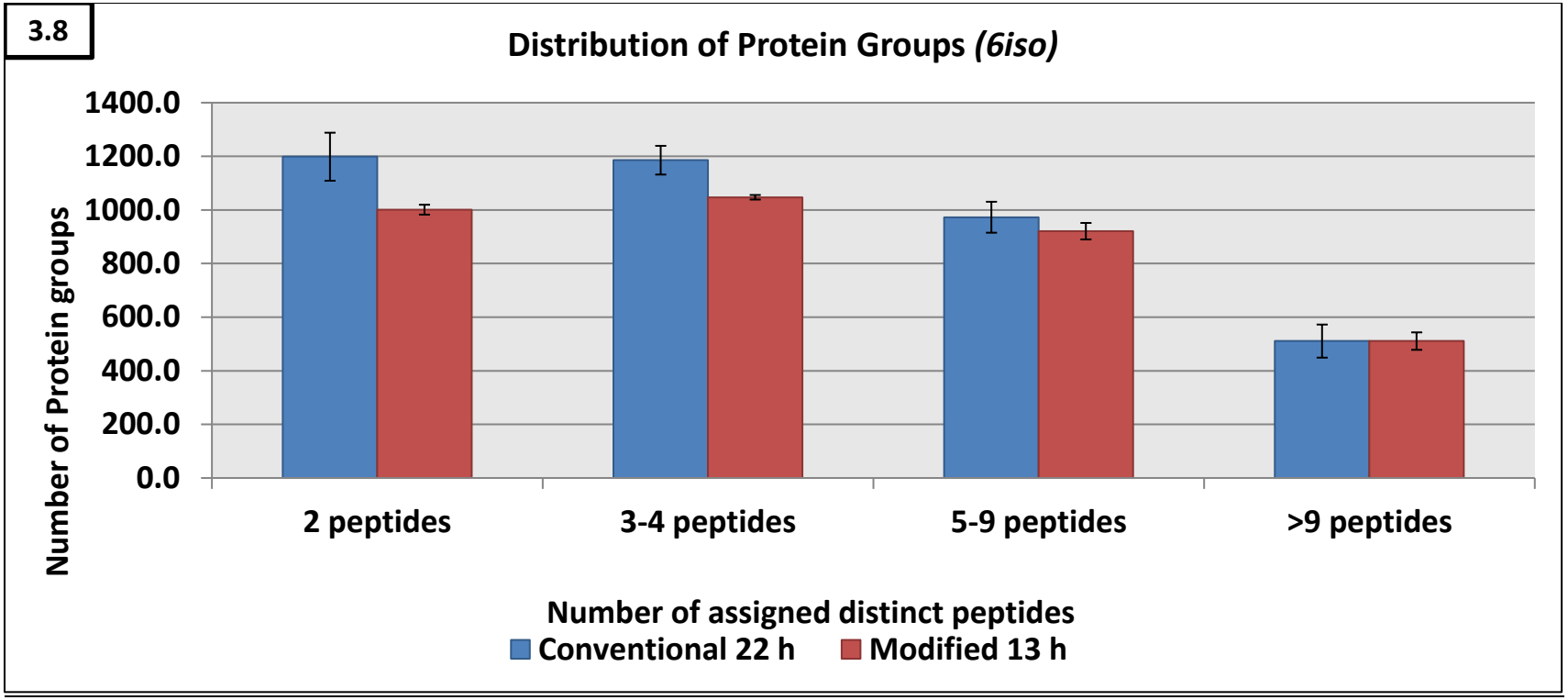


Figure 3.8: Distribution of Protein groups for 6iso sample

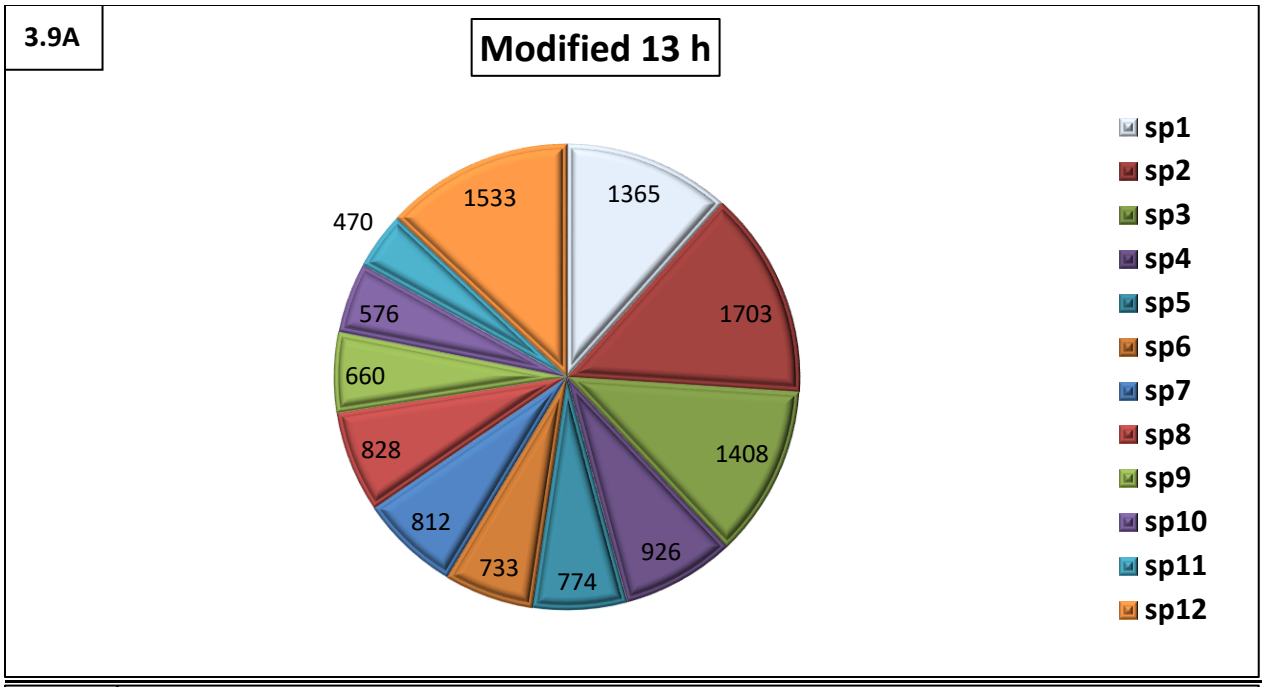
Distribution of protein group identifications during successive replicate analyses of tryptic peptides from 6iso sample, as classified by the number of distinct peptides that characterized each protein. Listed are proteins that were identified by 2 distinct peptides, those identified by 3 or 4 peptides, by 5 through 9 peptides and by 9 or more peptides. Error bars indicate standard deviation (s.d); n=4 technical replicates. A high degree of overlap in protein group distribution within the conventional 22 h and modified 13 h schemes for proteins identified by at least 5 distinct peptides – likely the more abundant proteins in the sample

3.4.6 Distribution of Unique Peptides Across Modified 22 h and Modified 13 h Schemes:

To evaluate the elution profile of peptides across individual fractions for the shorter 13 h schemes, we concatenated all the unique peptides of *6iso* from each salt pulse and assessed their distribution across the salt pulses **(Figure 3.9)**. The modified 13 h **(3.9A)** schemes showed a more even distribution of peptides across all the fractions. This observation was similar to the modified 22 h scheme (**figures 3.1B and 3.9B)** and is in stark contrast to the conventional 22 h scheme where the first salt pulse had an overwhelmingly large peptide elution profile (**figure 3.1A**). For example, it takes only 5 pulses to achieve 75% peptide identification in the conventional scheme whereas, it takes 8 to 9 pulses to achieve 75% in the modified 22 h and modified 13 h respectively. Since the mass spectrometer has a somewhat fixed duty cycle, increased uniformity in the chromatographic separation obtained in the modified schemes allowed us in achieving deeper proteome coverage.

3.5 Conclusions:

The study presented here describes an optimized approach for multidimensional chromatographic separations in complex proteome measurements, specifically for SCX-RP separation of peptides. The high level of distinctness obtained in the shallower scheme demonstrates a more uniform elution of different peptides. We found that the modified 22 h scheme provides significant increases in measurement depth by virtue of improved spacing of unique peptide sub-populations per salt pulse. This translates to deeper total proteome coverage. Given that dynamic range is a critical issue in proteomics, the shallower scheme can



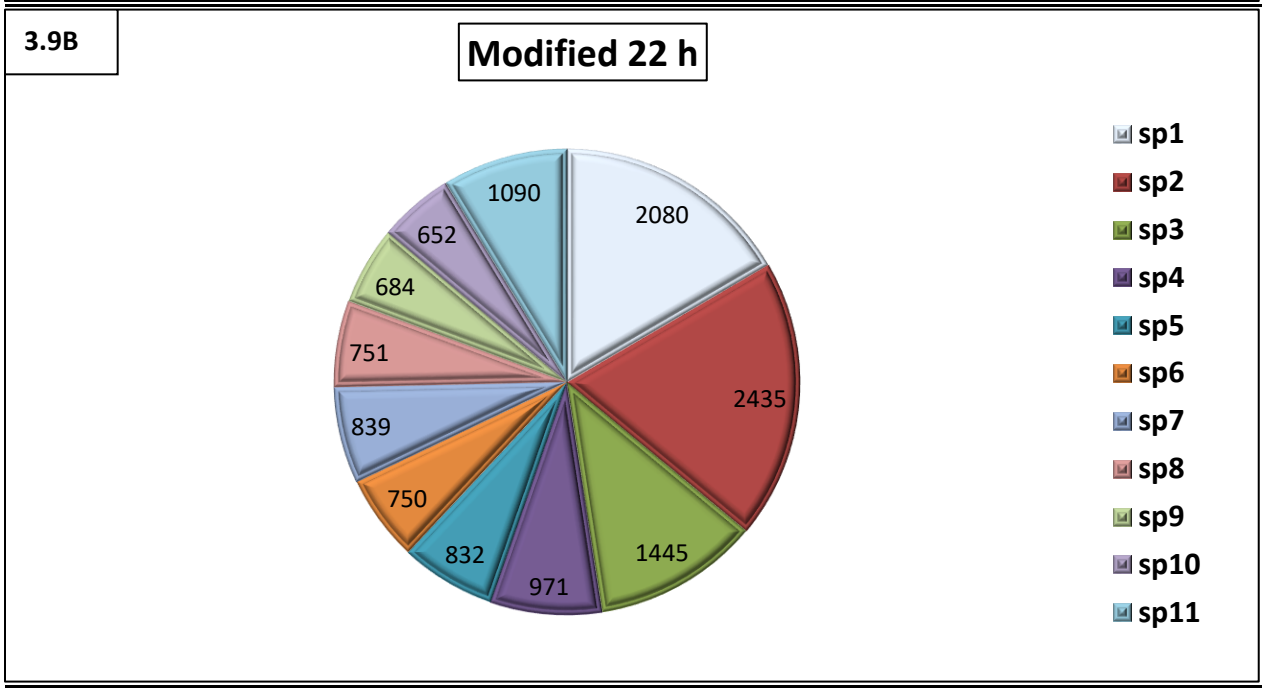


Figure 3.9: Distribution of unique peptides across each salt pulse (sp) for modified 13 h (3.9A) and modified 22 h (3.9B) schemes.

Both the schemes showed even distribution of peptides across all fractions as indicated by uniform slices.

better sample low abundant peptides as there is reduced peptide density especially during the initial pulses. The modified 22 h scheme provides superior proteome coverage without compromising the total measurement time of 22 h that constitutes a typical MudPIT setup. In fact, the modified 13 h approach is comparable with the much longer conventional approach, thus yielding a significant increase in measurement throughput. Though the proposed 2D separation method described here was demonstrated using an online mode, it can easily be extended to offline fractionation techniques.

The utility of modified scheme was further demonstrated on a complex community sample, the results of which are discussed in the next chapter

Chapter 4 - Evaluating the Impact of Multiple Search Engines and *De Novo* Sequencing Algorithms for Obtaining Deeper Proteome Coverage in Complex Environmental Samples

4.1 *De Novo* Sequencing as an Alternative Tool for Peptide Validation-Case Study Using a Thiocyanate Degrading Microbial Community Sample:

In order to further demonstrate the utility of modified 22 h scheme, we chose a "real world" thiocyanate (SCN^-) degrading microbial community sample obtained from a bioreactor[119, 120]. This sample was sent to us as a part of the collaborative initiative with Dr. Jillian F. Banfield's group at UC, Berkeley and Dr. Susan T.L. Harrison at the University of Cape Town, South Africa. The sample was enriched in several genera that have been previously detected in or isolated from SCN^- and CN^- degrading microbial communities. These included *Thiobacillus, Mesorhizobium, Sphingomonas, Sphingopyxis, Comamonas, Rhodanobacter,* and *Microbacterium* [121-124]. The thiocyanate sample was analyzed in technical triplicate for both Conventional 22 h and modified 22 h schemes on a high-resolution mass spectrometer (LTQ-Orbitrap Elite) to assess whether enhanced ion mass accuracy/resolution coupled with our newly designed chromatographic scheme impacts proteomic identification in a complex sample. Though data-dependent acquisition settings for this instrument were comparable to those used with the LTQ Velos Pro analysis of the *6iso* and *C elegans* samples (described in the materials and methods section of chapter 3), the Orbitrap employed here allowed us to exclude unassigned and singly-charged analytes from targeted fragmentation as they are less informative with regard to database searching [125]. The database used in this study was constructed from metagenomic assemblies described in Kantor *et al* [120].

The results of peptides, proteins and spectral counts of all the technical replicates are summarized in **table 4.1.**

Table 4.1: Overview of proteomic results from samples measured by the conventional 22 h and modified 22 h schemes for the Thiocyanate community sample

	Conventional 22 h			Modified 22 h		
Run	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3
Spectral counts	70108	70322	69537	68775	66413	65961
Peptide counts	20609	22359	22251	20137	14208	21121
Protein counts	6788	7295	7005	6346	5313	6667

Surprisingly, the conventional 22 h scheme gave comparable peptide counts to the modified 22 h scheme. But a closer analysis of the peptides eluting in the individual salt pulses revealed that the modified 22 h scheme outperformed the conventional 22 h scheme from pulse 4 to pulse 11 **(Figure 4.1).** Also, inspection of the total spectral counts of both the schemes revealed that they were similar (Mean=70,001, SD=386 for the conventional 22 h and Mean=67050, SD=1511 for the modified 22 h). This prompted us to investigate why the modified 22 h scheme did not exhibit the expected increased performance. To this end, we sought to examine whether the lack of depth in the modified scheme was due to peptide identification rather than *peptide detection*, as multispecies samples have exceptional complexity and heterogeneity, which creates a bottleneck for peptide detection using database searching approaches [39, 126, 127].

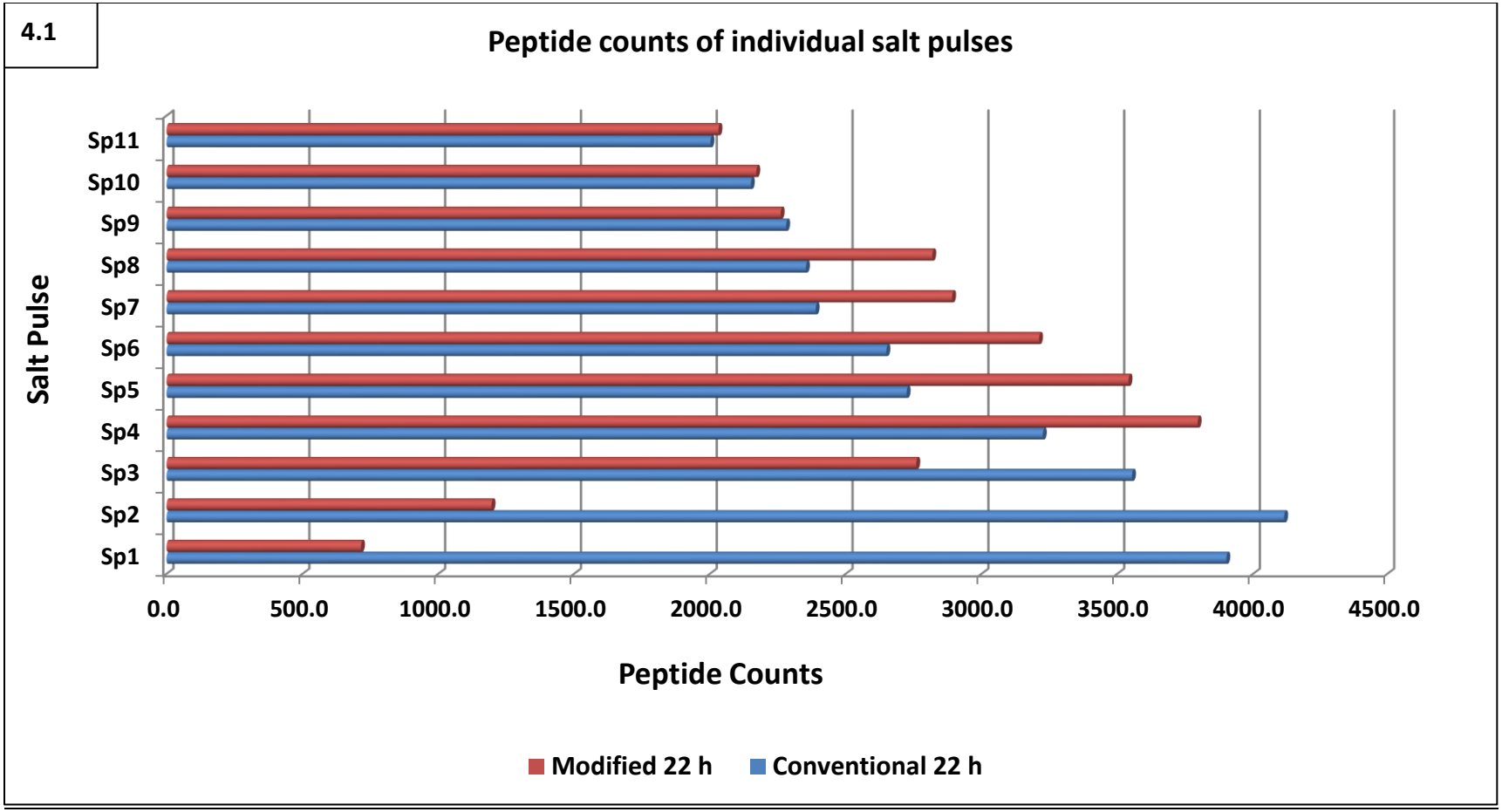


Figure 4.1: Distribution of concatenated peptides across individual salt pulses for technical triplicates of the thiocyanate community sample.

The modified 22 h scheme outperformed the conventional 22 h scheme from pulse 4 to pulse 11.

For this, we used *De novo* sequencing, which is an alternative tool used to infer peptide sequences directly from the MS/MS spectra and does not require a protein sequence database. The utility of *de novo* sequencing in metaproteomic investigations has been demonstrated previously [42, 128]. Thus, we performed *de novo* sequencing on the thiocyanate metaproteome using DeNovoGui [114] which has a graphic user interface and enables the user to run parallelized versions of PepNovo+ sequencing algorithm[115]. We retrieved *de novo* peptide suggestions for the technical triplicates of conventional 22 h and modified 22 h schemes **(Figure 4.2)** and found that the total number of collected spectra (i.e. those having PepNovo+ scores ranging from zero to maximum) was similar for both the schemes (blue bars). However, high confident spectra (PepNovo+ score above threshold 100) were observably greater (~20,000) in the modified scheme (red bars), even though the same to slightly less spectra were collected on average. These data imply that a larger proportion of high quality peptide spectra were measured in the modified scheme compared to the conventional, 38% vs. 31% respectively, an increase that would no doubt add to the current peptide and protein count of the modified scheme irrespective of database quality. Although, the current metagenomic approaches provide a glimpse into community dynamics of high abundant organisms, they provide limited information for community members present at low frequencies [129]. Hence, the high quality unassigned spectra obtained in the thiocyanate community sample, combined with the information from **Figure 3.4** for the *6iso* sample (chapter 3) where the modified scheme has superior measurement depth for lower abundance peptides, suggests that the expected gains in measurement performance here might be masked by the incomplete metagenome information of lower abundant species in the thiocyanate

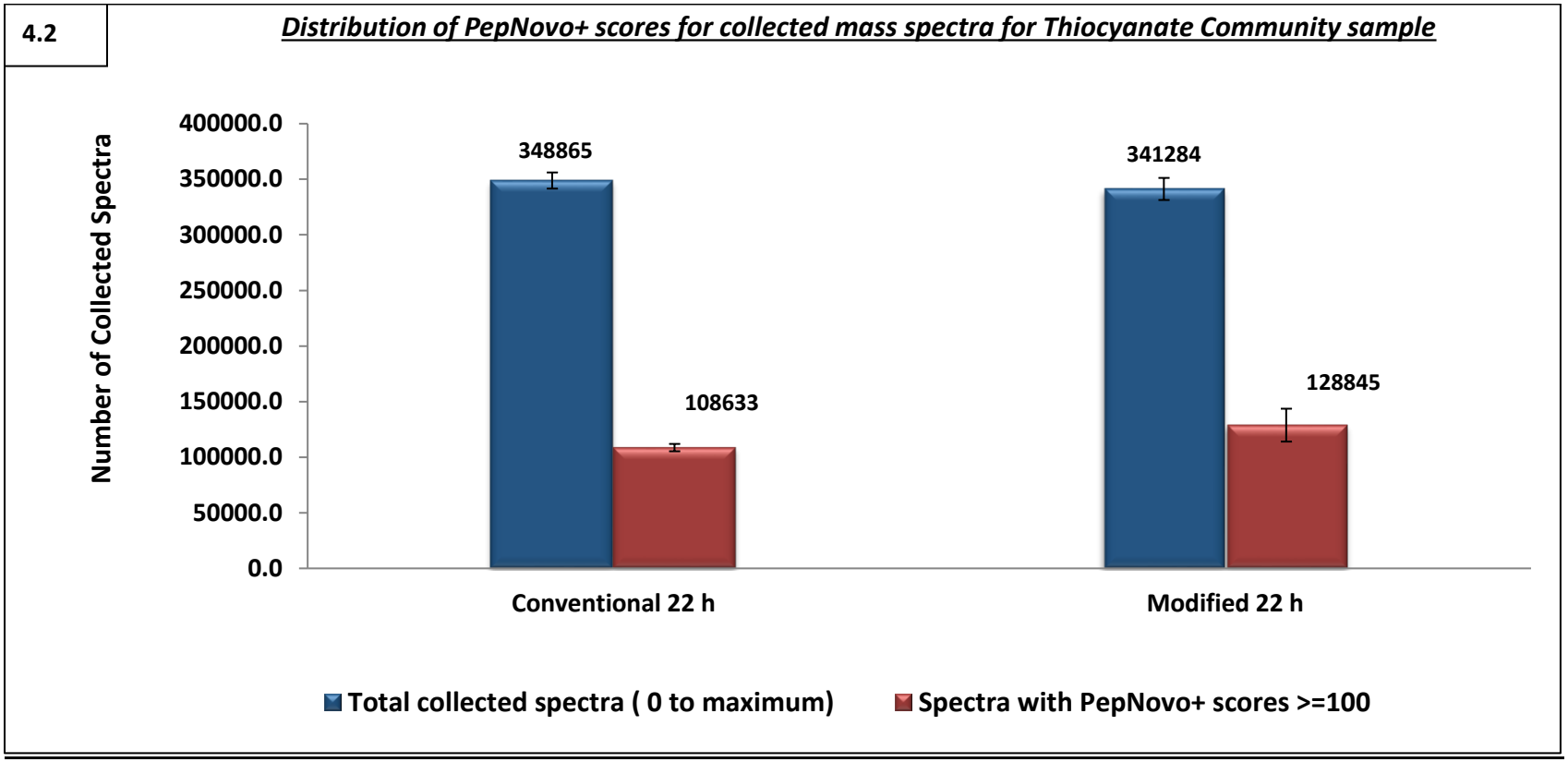


Figure 4.2: Distribution of PepNovo+ scores for collected mass spectra for the thiocyanate community sample

Error bars indicate standard error (SE); n=3 technical replicates. The total number of collected spectra (0 to maximum) was similar for both the schemes (blue). However, high confident spectra (scores greater than or equal to 100) were observably greater in the modified scheme (red) indicating that a large portion of high quality spectra in this newly designed scheme are not matching to the corresponding database.

metaproteome.

In a complex community sample, success in achieving accurate protein identification and enhanced proteome coverage is heavily reliant upon the protein database that is constructed from the metagenomic data. In comparison to the single cell type/microbial isolate databases which are well-curated, a large portion of high quality spectra in a metaproteomic study remain unassigned due to poor matches with the proteomic database. This is mainly due to high complexity and heterogeneity of the sample containing thousands of proteins belonging to hundreds of different microbial species having varied abundances and often sharing a high level of homology. Also, the restricted availability of (predicted) peptide sequences matching the experimental spectra makes peptide to spectrum matching an arduous task [126] . *De novo* sequencing offers an extended approach to common database searching strategy and can be used as an alternative validation tool for peptide identification. By using this tool, we were able to demonstrate the overall enhanced measurement depth of our new scheme (~20,000 more HQ peptide spectra) for complex community samples, which would otherwise have been missed if only database search results were taken into consideration. This also reiterates the use of *de novo* sequencing as a powerful tool for improving the depth and reliability of metaproteomic identifications.

The observations with respect to *thiocyanate* metaproteome lead to two important questions that formed the premise for the subsequent sections discussed in this chapter. 1) How to map *de novo* peptides back to metaproteome database? because peptide sequences though important give little relevant information unless one finds the corresponding protein sequences. 2) Instead of using one search engine, can a battery of multiple iterative search

engines be used for improving the reliability and depth of metaproteomic data? This approach has been shown to be suitable when working with well-established systems like human metaproteomes, but has not been verified with complex environmental samples where sample availability can be sometimes limited and metagenomic assembly is usually poor.

4.2 Computational Bottlenecks Associated with Database Search Strategy for Complex Metaproteomes:

Protein sequence database searching is the current gold standard in the proteomics community for accurate peptide and protein identifications. This is mainly because the search software requires limited computational space to interrogate the experimental spectra from a database having a pre-defined protein list derived from the metagenomic assembly of the sample under consideration. However, database search algorithms may contribute to low identification rates (low sensitivity) [130, 131] and elevated false discovery rates (low accuracy) [132]. Thus, improving the performance of database search algorithms is still an area of active research.

The database search approach essentially presents two competing challenges for protein identification; accuracy and sensitivity. Accuracy is measured by false discovery rate, which is defined as the percentage of false identification in a pool of total identifications above a particular score threshold (FDR is discussed in detail in chapter 2). Accuracy or the confidence in protein identification is accomplished by increasing the score threshold. Thus, in order to account for both sensitivity and accuracy, new scoring functions have been developed that separates the true and false identifications more efficiently [133, 134].

Another bottleneck associated with the database search strategy is the search speed. To maintain an acceptable search speed, filtration methods are usually adopted that quickly shortlist a subset of proteins or peptide candidates based on cross-correlation scores which are further evaluated using more advanced scoring functions. However, this stringent filtration method often excludes real peptides. In shotgun proteomic studies, these real peptides often belong to those proteins that are present in the sample at a very low level and only represented in the MS/MS data by one of two spectra [135]. Alternatively, there are novel peptides present in the experimental spectra and their characterization is averted simply because the proteins belonging to these peptides were derived from low abundant species in the community and were missed during genome assembly.

4.3 Combining Results from Multiple Search Engines:

There are several open source algorithms that can be used to match the experimental spectra to the corresponding database to identify confident peptide-spectrum matches. These include SEQUEST [52], X! Tandem [84], MS-GF+ [85] , MASCOT [80] , MyriMatch [86], Comet [87] , Andromeda [88], OMSSA [82] etc. However, computational analysis of MS/MS spectra represents a significant challenge in that only a fraction (typically <30%) of all acquired MS/MS spectra can be successfully interpreted as peptides using database search strategy. This can be partially explained by the differences in the scoring schemes implemented in the search tools that rank candidate database peptides with the given experimental spectrum. A standard scoring scheme for annotating the fragmentation spectra has not yet been designed and each database-searching algorithm implements similar, but slightly different strategies to identify peptides. For example, SEQUEST calculates a correlation score between a normalized MS/MS

spectrum and a unit-intensity fragmentation model. Mascot and OMSSA calculate the likelihood of matching peaks using Poisson, and hypergeometric distributions. X! Tandem calculates a score like SEQUEST, but considers the distribution of those scores to calculate an expect value (*E* value) like BLAST. Newer search engines like MyriMatch utilize intensity based information encoded in the spectrum and rank peaks into intensity bins to infer peptides.

These differences in database search engines for scoring peptides has led to integration of multiple search engines for improved assessment of complex datasets, thus yielding a similar yet somewhat distinct set of peptide at the same FDR criterion [136-138]. The utility of multiple search engines has been successfully demonstrated for well-established systems like human metaproteomes [139], but has not yet been evaluated with complex multi-species environmental samples where biomass availability can sometimes be limited and metagenomic assembly can be poor.

4.4 *De Novo* Sequencing- Advantages and Limitations:

The ability to infer peptide sequences directly from the MS/MS spectra confers an indisputable advantage to the process of *de novo* sequencing. This is essentially useful for metaproteomic specimens where public repositories often fail to cover the entire proteome. In such cases where metagenome sequences may be difficult to obtain, *de novo* sequencing allows for the identification of peptides even if the peptide sequence information is incomplete or partially not available [140]. Although, *de novo* sequencing is a computationally intensive process, the availability of modern computer architecture has rendered it as a powerful and cost-effective method for peptide validation. The complementation of *de novo* sequencing with conventional

database searching methodologies can support the validation of borderline peptide identifications. In some cases, *de novo* sequencing can also prove to be a better alternative to database searching algorithm (for example, homologous proteins from different taxonomic domains). Here, database search programs are more prone towards certain protein identifications by assigning them greater probabilities. *De novo* sequencing on the other hand can counter such biases as it relies on spectrum information alone [42].

However, *de novo* detection is frequently hampered by common *de novo* sequencing errors such as inversions and amino acid substitutions. Thus, proper evaluation of *de novo* inferred peptides for further downstream analysis is important to avoid false positives [127]. Also, the essential step of peptide mapping to protein sequences is not performed by *de novo* sequencing software.

The utility of *de novo* sequencing for peptide identifications has been demonstrated previously for metaproteomic specimens [42]. However, extension of this methodology to infer taxonomic and functional information is still an un-explored domain. This step will be crucial in deepening the extent of meta-information that can be obtained from these complex samples.

4.5 PepExplorer- A Pattern Recognition Tool for Mapping *de novo* Sequencing Results:

PepExplorer is a post-processing *de novo* sequencing software that allows protein inference via statistical mapping of *de novo* sequencing results to a corresponding protein sequence database [141]. PepExplorer uses extensive pattern recognition to gather a list of homologous proteins derived from *de novo* sequencing data to allow for biological interpretation of data. In

order to achieve this, it utilizes a radical basis functional neural network that takes into account the *de novo* sequencing scores and peptide lengths to select probable protein candidates that belong to the target decoy database. PepExplorer is designed to handle the outputs from several commercially available and open source *de novo* tools like PepNovo, pNovo+, PEAKS etc. Similarly, the software accepts a series of database formats for input analysis.

The focus of this study was to develop an integrated metaproteomic workflow consisting of multiple iterative search engines coupled with de novo sequencing algorithms for a comprehensive and in-depth characterization of complex environmental samples.

4.6 Materials and Methods:

4.6.1 Sample Types:

Three different sample types were used for this study. The first was an isolate of *Desulfovibrio desulfuricans* (ND132); the second was a mixture of six environmental microbes herein referred to as 6iso that included *Saccharomyces cerevisiae, Escherichia coli, Clostridium thermocellum, Ignicoccus hospitalis, Nanoarchaeum equitans* and *Streptomyces eurocidicus*. The third sample used for the study was a biofilm sample inhabiting a site of extreme acid mine drainage (AMD) collected from Richmond Mine, Iron Mountain, CA.

4.6.2 Protein Extraction and Enzymatic Digestion:

ND132 and *6iso* samples were prepared separately as follows: 1 mg of microbial sample was excised and thawed for cell lysis and protein extraction. These samples were first boiled for 5 min in 1 mL of lysis buffer containing 100 mM Tris-HCl, pH 8.0, 4% w/v SDS (sodium dodecyl

sulfate), and 10 mM dithiothreitol (DTT). The suspension was vortexed and sonicated with a Branson ultrasonic cell disruptor (20% amplitude for 2 min, 10 s pulse with 10 s rest). The resulting crude protein extract was precleared via centrifugation at 21000 g and quantified by the BCA assay (Pierce Biotechnology, Waltham, MA). An aliquot consisting of ∼1 mg of protein was subjected to TCA precipitation and subsequent digestion with trypsin using the method described previously [103]. The resulting peptides were quantified by the BCA assay and stored at −80 °C until use.

For the AMD sample, whole-cell protein fraction was extracted using methods similar to earlier studies [20, 142]. 1 g of biofilm was gently mixed with 5 ml of H_2SO_4 at pH 1.1, the suspended biofilm was then centrifuged at 12000 *g* for 20 min at 4°C, the supernatant was then discarded. A washing step was incorporated to avoid interference of near-molar concentrations of metals with protein extraction. The pellet was re-suspended in 6 ml of 20 mM Tris-SO_4, pH 8.0 and the cells were lysed by sonication on ice (5 × 1 min using a microprobe at high power with 1 min breaks to avoid excessive heating). 5 ml of 0.4 M Na_2CO_3 was then added at pH 11 and the whole-cellular protein fraction was separated from non-lysed cells and extracellular polymers by centrifugation (6000 *g*, 20 min, 4°C). Finally, proteins were precipitated by the addition of 1:10 volume trichloroacetic acid and incubated at 4°C for at least 3 h. The mixture was centrifuged for 10 min at 20 000 *g* at 4°C and after washing the pellet with methanol at 4°C, it was centrifuged at 20 000 *g* at 4°C again, after which the pellet was air-dried and frozen at −80°C. The protein extract was subsequently denatured and reduced with 6 M Guanidine/10 mM DTT, and digested into peptides using sequencing grade trypsin (Promega, Madison, WI).

4.6.3 Nano 2D LC-MS/MS Measurement:

Peptides were de-salted, concentrated and analyzed either via 22 h nano-2D-LC MudPIT (for ND132 and *6iso* samples) or 24 h nano-2D-LC MudPIT (for AMD sample) consisting of strong cation exchange and reversed phase. Raw spectra were collected on either an LTQ-Orbitrap ELITE mass spectrometer (for ND132 and *6iso* samples) or an LTQ-Orbitrap XL mass spectrometer (for AMD sample), [Thermo Fisher Scientific, San Jose, CA] with technical duplicates for each sample. The LTQ-Orbitrap-Elite was operated in a data-dependent mode. MS1 was performed in Orbitrap and data dependent MS/MS was performed in LTQ (top twenty), 1 microscan for both full and MS/MS scans; normalized collision energy 35% and dynamic exclusion time of 30 seconds. The LTQ-Orbitrap-XL was run as follows: full scans of the peptide mixture entering the mass spectrometer from the LC column were taken in the Orbitrap at 30 K resolution followed by five data-dependent MS/MS spectra acquired in the LTQ; two microscans were averaged for both full and MS/MS scans; normalized collision energy 35% and dynamic exclusion was set at one.

4.6.4 Data Analysis:

Raw files were converted to MGF peak lists. The resulting peak list files were identified using four different database search algorithms; OMSSA version 2.1.9 [82], X!Tandem version X! Tandem Vengeance (2015.12.15.2) [143], MS-GF+ version Beta (v10282) [144] and Comet version 2016.01 rev. 2 [87]. The search was conducted using SearchGUI version 3.2.7 [145].

Protein identification was conducted against a concatenated target/decoy version of the three samples under consideration (ND132, *6iso* and AMD having 3452, 16,321 and 191778 target

sequences respectively). The decoy sequences were created by reversing the target sequences in SearchGUI. The identification settings were as follows: Trypsin, Specific, with a maximum of 2 missed cleavages 10.0 ppm as MS1 and 0.5 Da as MS2 tolerances. Peptides and proteins were inferred from the spectrum identification results using PeptideShaker version 1.16.4 [38]. Peptide Spectrum Matches (PSMs), peptides and proteins were validated at a 1.0% False Discovery Rate (FDR) estimated using the decoy hit distribution.

4.6.5 *De Novo* Sequencing:

Non-validated spectra i.e. those spectra that did not find a valid hit during database searching above were again exported as MGF peak lists and were subjected to *de novo* sequencing. For *de novo* sequencing, the software DeNovoGUI (version 1.2.0) was used [114] that provides a graphical user interface and parallelization of the PepNovo+ algorithm [115]. The fragmentation method used was CID_IT_TRYP in PepNovo+ that accounts for CID fragmentation and tryptic cleavage. The PepNovo+ algorithm was used in multithreaded mode via a locally available computing cluster. Finally, the *de novo* peptide suggestions were filtered by a PepNovo+ score threshold above 100 for high-quality identifications for further downstream analysis.

4.6.6 *De novo* Results Parsing via PepExplorer:

The High scoring *de novo* peptides (having PepNovo+ score threshold ≥ 100) together with the target decoy database of the three samples were filtered through PepExplorer which is integrated into the *PatternLab for Proteomics (v4.0.0.68)* environment [146]. The identity cutoff of 0.95 and minimum peptide length of 8 amino acids was used to select confident *de novo* peptides matching to the target-decoy database. This peptide length was similar to the

sequences of database derived peptides. The neural network of PepExplorer was run in a simplified mode by copying and pasting the high scoring *de novo* peptide suggestions in the provided text box.

4.7 The Computational Pipeline:

Figure 4.3 depicts the schematic illustration of the designed computational pipeline used for analyzing the metaproteomic data. Briefly, the raw spectra obtained from MudPIT measurements were converted to MGF peak lists which were then subsequently analyzed against either one (MS-GF+) or four (OMSSA, X!Tandem, MS-GF+ and Comet) different database search algorithms. The numerical gains in peptide and protein counts were then compared to quantify the gains observed when using multiple search engines (Steps 1 through 3). Subsequently, the spectra of non-validated PSM's i.e. those that did not find a valid peptide hit during database searching, were subjected to *de novo* sequencing (steps 4 and 5). The high scoring *de novo* predicted peptides were then mapped to the target decoy database via PepExplorer to identify novel protein candidates (steps 6 through 8).

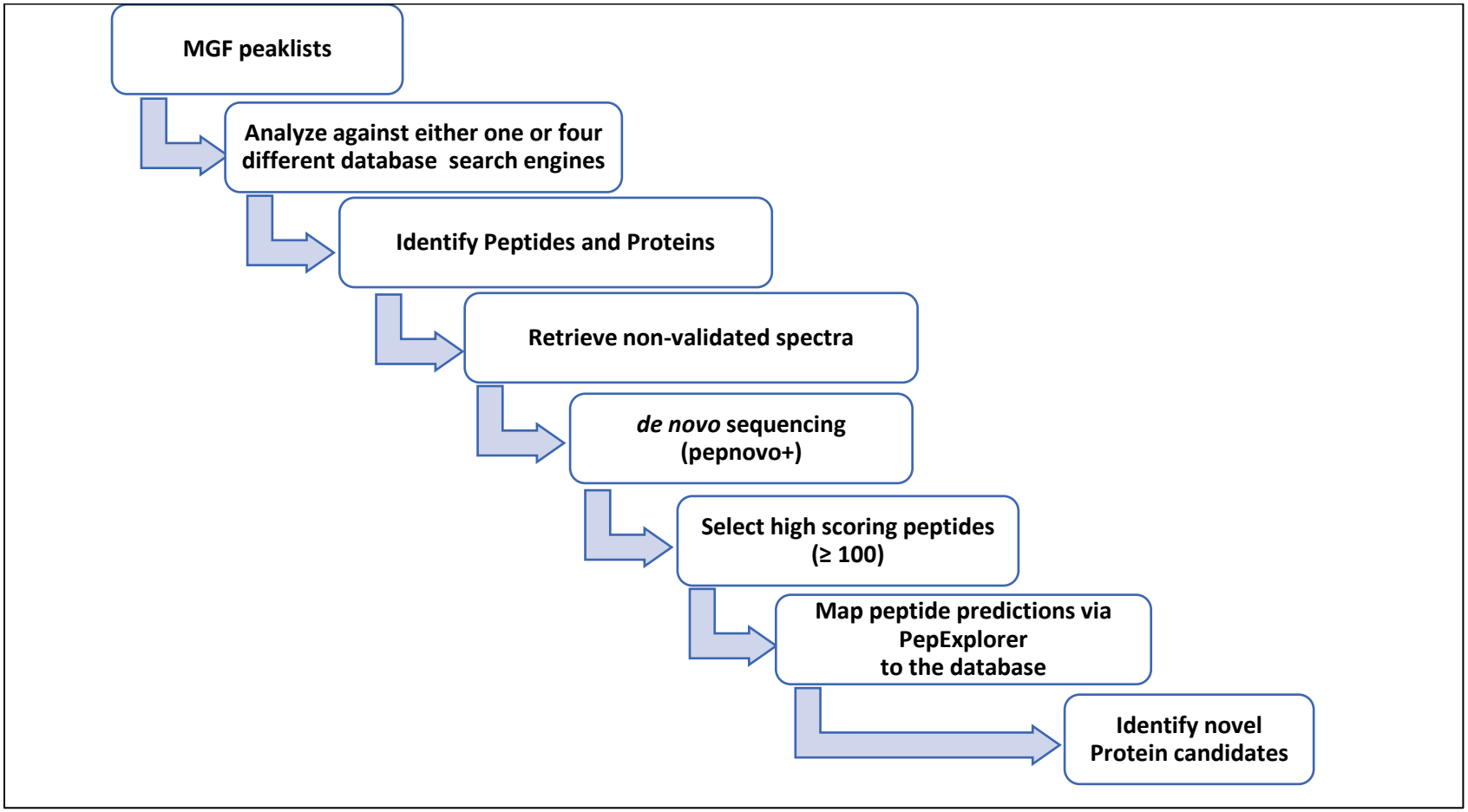


Figure 4.3: Schematic illustration of the designed computational pipeline used for analyzing the metaproteomic data.

4.8 Results and Discussion:

4.8.1 Comparison of Proteomic Results of Three Different Samples Analyzed Against Database Search Algorithms:

Figure 4.4 illustrates the comparison among peptides **(4.4A**) and proteins **(4.4B)** identifications obtained by searching the acquired MS spectra against either four different search engines (OMSSA, X!Tandem, MS-GF+ and Comet and named as "Combined") or one search engine (MSGF+) using a <1% FDR threshold. Combining the search results from four different search engines lead to 13%, 18% and 36% increase in peptides and 6%, 12% and 14% increase in protein counts for ND132, *6iso* and AMD samples respectively.

Further assessment of the spectral ID's indicated that 34%, 11% and 32% of high quality spectra for ND132, 6iso and AMD samples were converted to peptide identification via database searching. This implied that a large portion of high-confident spectra were not matching the database. This is a serious issue in community proteomics, arising mainly due to the restricted availability of predicted peptide sequences matching the experimental spectra. Hence, the spectra that did not find a valid hit during database searching were subjected to *de novo* sequencing.

4.8.2 Taxonomic Attribution of Peptides via Unipept:

One of the recent advancements in metaproteomics research is the biodiversity analysis of complex metaproteome samples using the Unipept web application [147]. Unipept has the lowest common ancestor (LCA) algorithm [148] that allows the taxonomic identification of

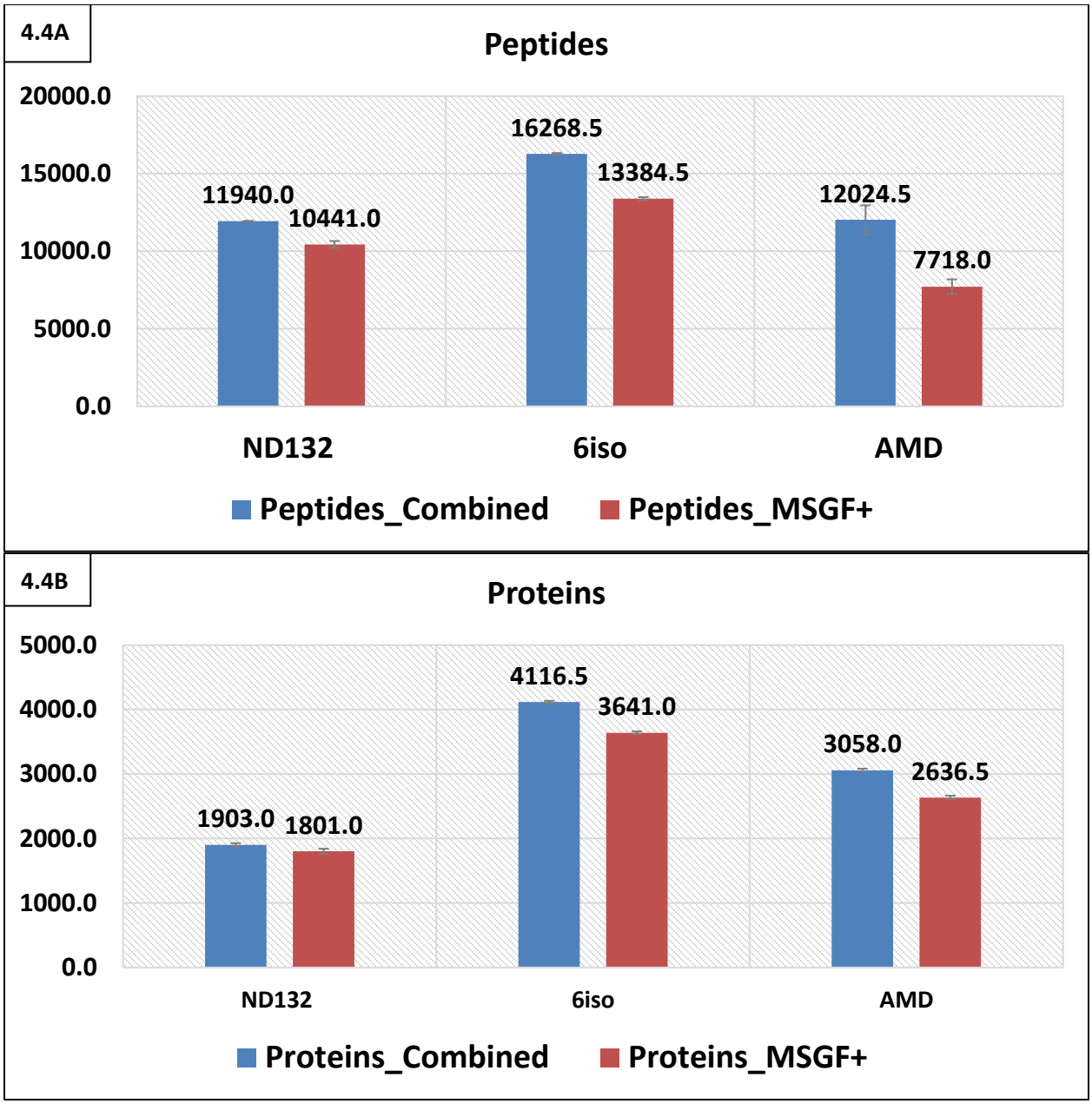


Figure 4.4: Peptide (4.4A) and protein (4.4B) counts respectively of the three samples searched against either four search algorithms ("Combined") or one search algorithm (MSGF+).

Error bars indicate standard deviation (s.d); n=2 technical replicates.

peptides detected in MS measurements. LCA is based on the principle that a particular peptide or a protein sequence is assigned to a species only if does not match with any other species contained in the Uniprot sequence database. On the other hand, if a peptide or a protein sequence is shared amongst different species that belong to the same genus, then the sequence is unambiguously assigned only at the genus level. Thus, unique peptides derived from shotgun MS measurements can be grouped in different taxonomic levels based on the LCA algorithm. In general, widely conserved sequences are assigned to higher order taxa (such as class or phylum).

Unipept supports the biodiversity analysis of metaproteomic samples using the tryptic peptides generated during shotgun MS/MS experiments. The utility of Unipept for biodiversity analysis of metaproteome specimens has been demonstrated previously [46, 149]. In the current study, instead of tryptic peptides detected using database search algorithm, we used the *de novo* predicted peptides as inputs for Unipept and evaluated their taxonomic distribution. Since, all the three samples used in the study were digested using trypsin, the *de novo* predicted peptides were also tryptic in nature and served as ideal candidates for Unipept analysis. We only used high quality *de novo* predicted peptides having a score ≥ 100 and amino acid length ≥ 8 for downstream analysis. The resulting peptide sequences were imported on Unipept (https://unipept.ugent.be/datasets) to deduce taxonomic information about the *de novo* sequencing peptides. Peptides were subjected to 'Multi peptide' analysis by setting the following parameters; 'equate I and L residues' and 'filter duplicate peptides'.

Figure 4.5 illustrates the taxonomic attribution of 6sio **(4.5A)** and AMD **(4.5B)** metaproteomic data obtained via Unipept. 1,114 out of 16,177 and 10,625 out of 108,115 peptides obtained via

de novo sequencing were taxonomically grouped by Unipept. The overwhelmingly large number of *de novo* predicted peptides obtained for 6iso sample compared to the AMD sample was mainly due to the different instrument platforms used for acquiring MS/MS spectra (Orbitrap-ELITE for 6sio having faster scanning speed compared to Orbitrap-XL for AMD sample).

In the next step, we evaluated the common taxonomic annotations between combined database search (OMSSA, X!Tandem, MS-GF+ and Comet) and *de novo* sequencing i.e. the total number of phylum, class, family, genus and species common between combined database search and *de novo* sequencing **(Figure 4.6)**. For example, there were thirty-three different phyla detected by combined database search for the AMD sample, out of which twenty-one were also identified by *de novo* sequencing (highlighted by red box). We found that *de novo* sequencing data from 6iso sample is able to account for almost all the phyla, class, family and genus detected by database searching. For the AMD sample, the resolution obtained is less and is partially confined to phyla and class and diminishes further as we move to family, genus and species.

However, it must be noted that peptide attributions to different taxonomic bins via Unipept could lead to misassignments. This is evident from the *6iso* data where species level resolution identified twenty-six different species (instead of six) out of which eleven were also identified by *de novo* sequencing. Assignment of peptides to more than six species could be due to strain-level variation occurring in microbial species. It can also be an experimental artifact (peptides belonging to Keratin which is a common contaminant protein in mass spectrometry experiments). Since the species composition of *6iso* sample was already known, this allowed the quantification of incorrect assignments present in this sample. However, in case of complex

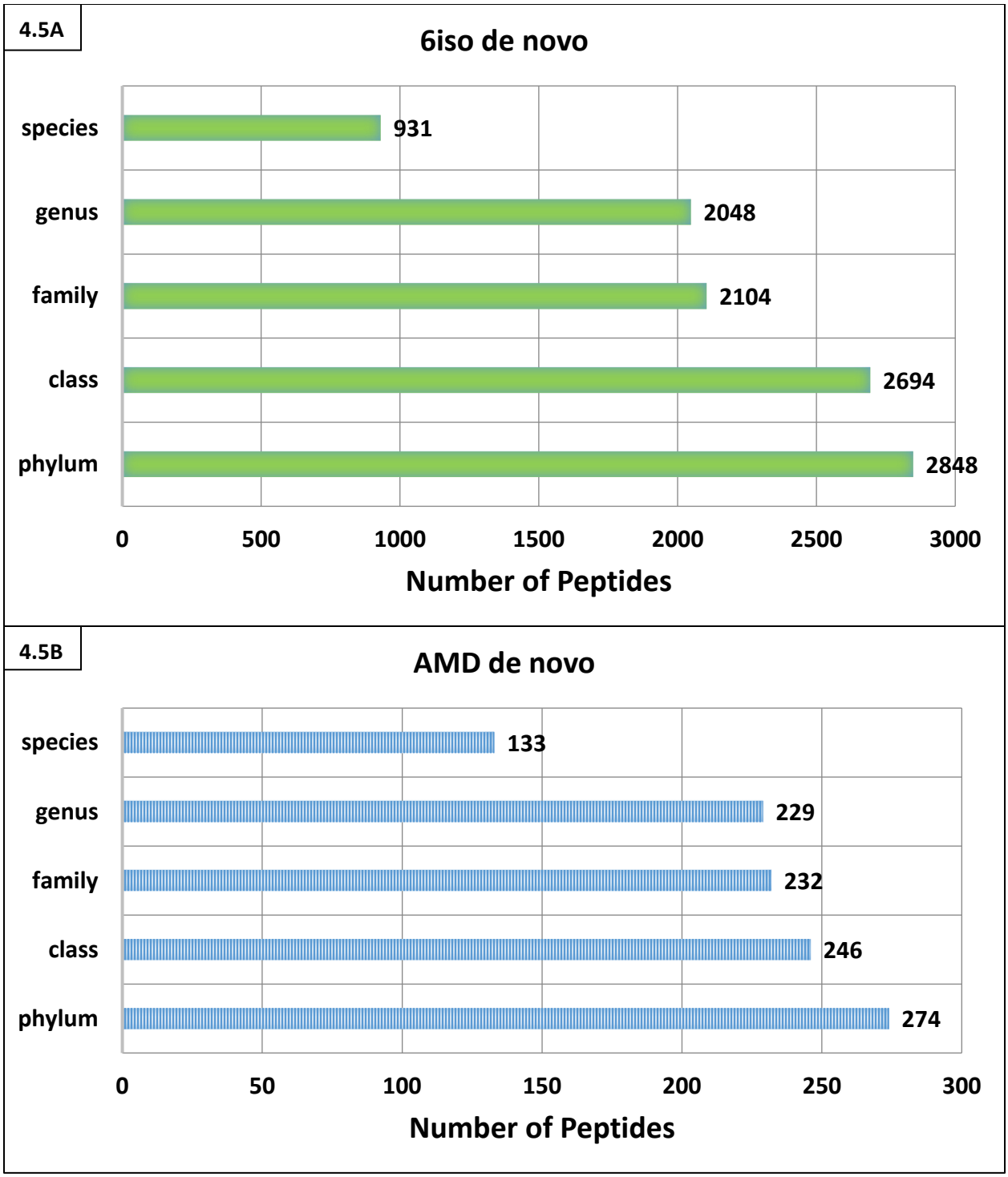


Figure 4.5: Taxonomic attribution of 6iso (4.5A) and AMD (4.5B) metaproteomic data obtained via de novo sequencing and filtered via Unipept.

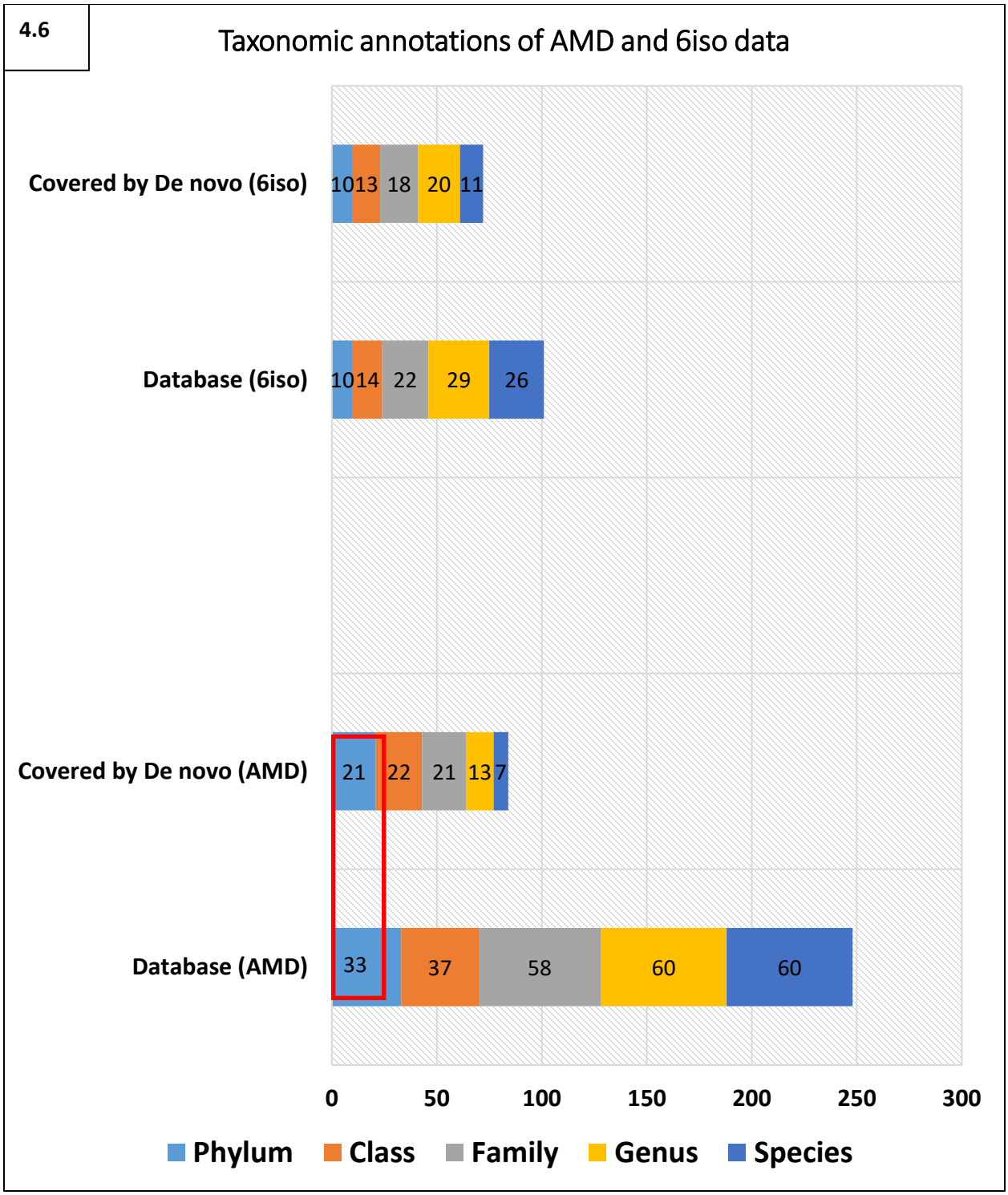


Figure 4.6: Taxonomic annotations common between combined database search algorithms and de novo sequencing for 6iso and AMD metaproteomic data.

heterogeneous sample like AMD, caution must be exercised before extending the conclusions obtained via Unipept and further validation thresholds must be applied to control the resolution of taxa identification before attributing the peptides to different taxonomic bins. This also reiterates the utility of *de novo* sequencing as a complimentary but not a 'stand-alone' approach for inferring the taxonomic diversity of metaproteomic data.

4.8.3 Protein Gains after PepExplorer Data Integration for AMD Sample:

In the next step, the high scoring *de novo* peptides (having PepNovo+ score threshold ≥ 100) were filtered through PepExplorer together with the target decoy database. **Figure 4.7** depicts the sequential increase in protein counts for the AMD sample. It should be noted that peptides with an identity cutoff of 0.95 and minimum length of eight amino acids that matched to a given protein sequence in the target decoy database were considered to be true hits. Combining the proteins from 'Combined database search' and PepExplorer mapping of *de novo* predicted peptides to the database lead to an increase of 1873 proteins for the two technical replicates of the AMD sample.

PepExplorer also generates a dynamic interactive report that allowed us to extract detailed information about the protein of interest. **Figure 4.8** illustrates the graphical user interface of the results browser that is composed of two panels. The upper panel displays information pertaining to the protein such as the number of peptide alignments, spectral counts i.e. the number of *de novo* spectra that matched to the protein, the unique spectra associated with the protein and the coverage chart that displays the regions of the proteins which were mapped by the *de novo* predicted peptides. When a protein is selected, detailed information is displayed in

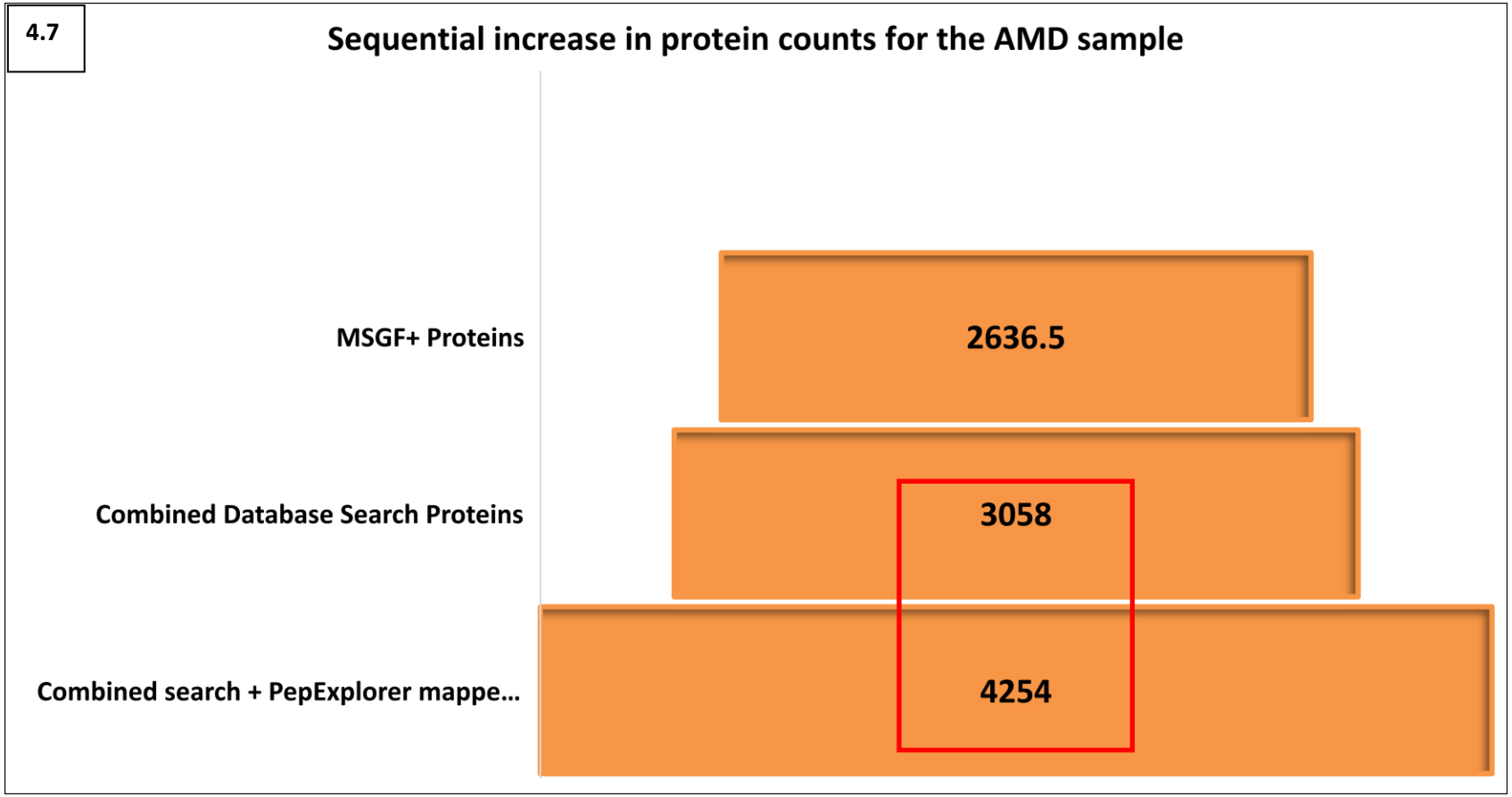


Figure 4.7: Average number of proteins detected in the AMD sample using a single database search engine (MSGF+), combined database search engines and summation of proteins obtained using the combined database search engines and Pepexplorer mapping of de novo predicted peptides to the database.

Combining the proteins from 'Combined database search' and PepExplorer mapping of de novo predicted peptides lead to an average increase ~1200 (1873 in total) proteins for the two technical replicates of the AMD sample (highlighted by red box).

4.8

	ID	Alignments	SpecCounts	Unique	CoverageChart
0	2007_11_V4_scaffold_192_3	14	14	9	
1	2007_11_V4_scaffold_2_66	10	10	0	
2	937.3.1095_trim_clean_scaffold_3021_3	10	10	0	
3	P7_C10m_GS2_1106_0711_scaffold_2483_1	10	10	0	
4	P48_C10m_GS1.5_0807_0711_scaffold_75_19	10	10	0	
5	2007_11_V4_scaffold_56_36	7	7	0	
6	CNXI.2065.1.1739_trim_clean_scaffold_4299_2	7	7	0	
7	951.3.1101_trim_clean_scaffold_3869_3	7	7	0	
8	2007_11_V4_scaffold_2473_1	6	6	0	
9	P7_C10m_GS2_1106_0711_scaffold_5871_1	6	6	0	
10	2008_10_V3_scaffold_216_41	6	6	0	
11	2007_11_V4_scaffold_115_17	6	6	0	
12	2007_11_V4_scaffold_115_15	6	6	0	
13	P48_C10m_GS1.5_0807_0711_scaffold_126_39	6	6	0	
14	P48_C10m_GS1.5_0807_0711_scaffold_2654_1	6	6	0	
15	CNXI.2065.1.1739_trim_clean_scaffold_3935_1	6	6	0	

Upper Panel

ScanNo	Z	DeNovoSeq	DBSeq
3378	1	SAGSNSMELKR	SAGSNSMELKR
2872	1	NAASVASLMLT	NAASVASLMLT
4343	1	VTLGPK	VTLGPK
3807	1	LKLSDLGR	LKLSDLGR
3502	1	NAGLEGSVVVQK	NAGLEGSVVVQK
715	1	TALADAVK	TALADAVK
2474	1	TLVVNKLR	TLVVNKLR
5581	1	AGDGTTTATVLAHA	AGDGTTTATVLAHA
4059	1	LKLENLK	LKLENLK
5553	1	TLKVEGDQK	TLKVEGDQK
4404	1	GLKLENLK	GLKLENLK
6230	1	AGLEGSVVVQK	AGLEGSVVVQK
3849	1	ALEEPLR	ALEEPLR
9352	1	ASVASLMLT	ASVASLMLT

Lower Panel

Figure 4.8: Graphical user interface of the PepExplorer results browser

the lower panel of all alignments that mapped to it, such as scan number, spectral count (Z), sequence provided by *de novo* sequencing tool and the corresponding sequence found in the database. Since we used a stringent identity cutoff (0.95), the *de novo* peptide sequence and database peptide sequence was in complete alignment with one another and only such proteins were considered in the candidate list for downstream analysis.

Double clicking on the row of interest in the upper panel generates a graphical coverage report that shows the extension of area covered by predicted peptides **(Figure 4.9)**.

4.8.4 Functional Annotation of PepExplorer Inferred Protein List for the AMD Sample:

Finally, we focused on the functional annotation of *de novo* sequencing data. For this we used GhostKOALA [150] which is KEGG's (Kyoto Encyclopedia of Genes and Genomes) internal annotation tool for KEGG Orthology (KO) assignment. GhostKOALA utilizes the GHOSTX program [151] for searching the query sequences against the non-redundant protein database. The GHOSTX program searches the given query sequence against the non-redundant database and the resulting multiple hits (Paralogs) are combined and only the top scoring hits are displayed. The K number assignment in GhostKOALA is based on the sum of GHOSTX normalized scores. The GHOSTKOALA server can be accessed freely via the Kegg.jp website (http://www.kegg.jp/ghostkoala/). The query list of protein sequences can be submitted as FASTA files using the weblink. GhostKOALA is an email based web server meaning that the job request for K number assignment can be confirmed by clicking on the link in the automatically sent mail. Thus, the 1873 PepExplorer filtered proteins for the AMD sample were submitted to

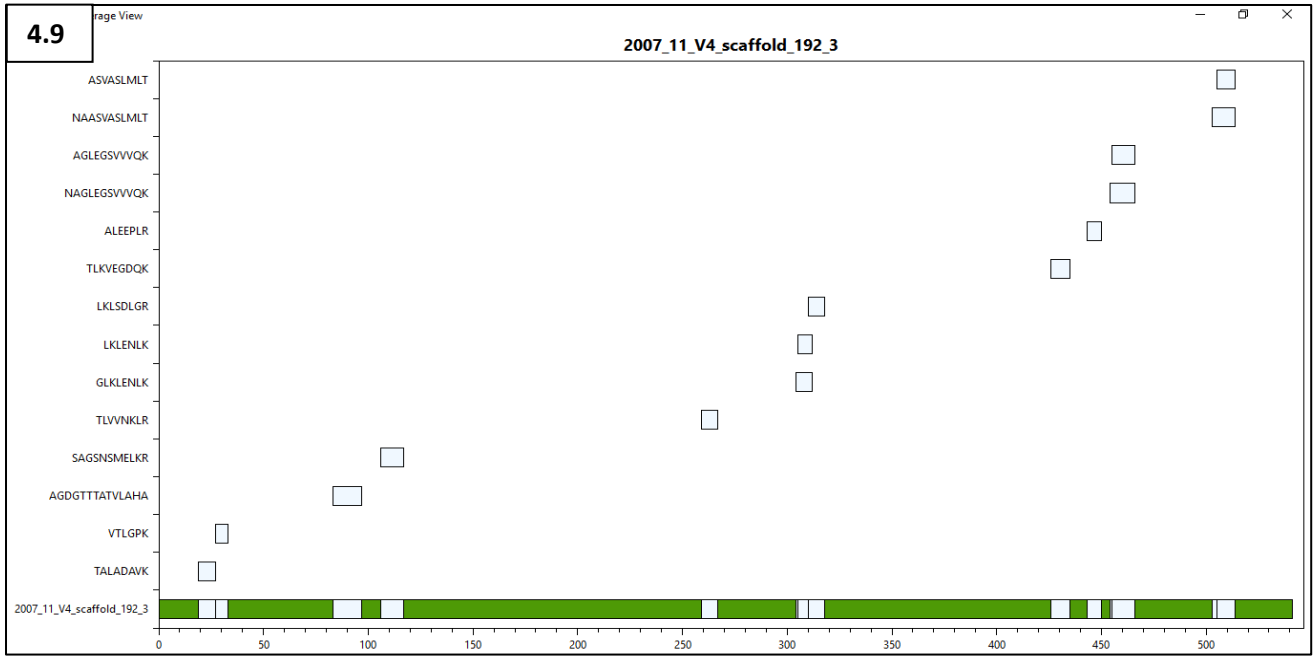


Figure 4.9: The graphical report of the protein sequence coverage showing the extension of area covered by the predicted peptides

GhostKOALA. The K number assignments of each protein were then grouped into different molecular functions. 1222 out of 1873 (65%) proteins had K number assignments. **Figure 4.10** depicts the functional annotation of AMD metaproteomic data obtained using GhostKOALA. Proteins involved in genetic information and processing were the most predominant members in the dataset. The GhostKOALA results page also gives information on the taxonomic composition of the query data. Here, each query protein is assigned a taxonomic category according to the best hit gene in the cd-hit cluster supplemented version of the non-redundant pangenome dataset. Thus, the taxonomy data of the 1873 proteins were downloaded and analyzed to look for the most abundant phyla **(figure 4.11)**. Of the five most abundant phyla, *Nitrospirae* was found to be the most predominant. Subsequently, the Nitrospirae phyla was further filtered down to look for the most abundant genus. Of these, Leptospirillum was found to be the predominant genus.

Leptospirillum are a group of obligate aerobic bacteria that play a crucial role in industrial bioleaching and biooxidation of toxic metals. The predominance of this bacterial genera in the AMD sample is consistent with previous findings [152]. They are also important contributors to the acid mine drainage process especially in toxic metal accumulation sites (like the Iron Mountain in northern California). Organisms of *Leptospirillum* inhibit the deep underground mines enveloped in thick biofilms which are pink in color and float on the surface of water flowing in the mine [153, 154].

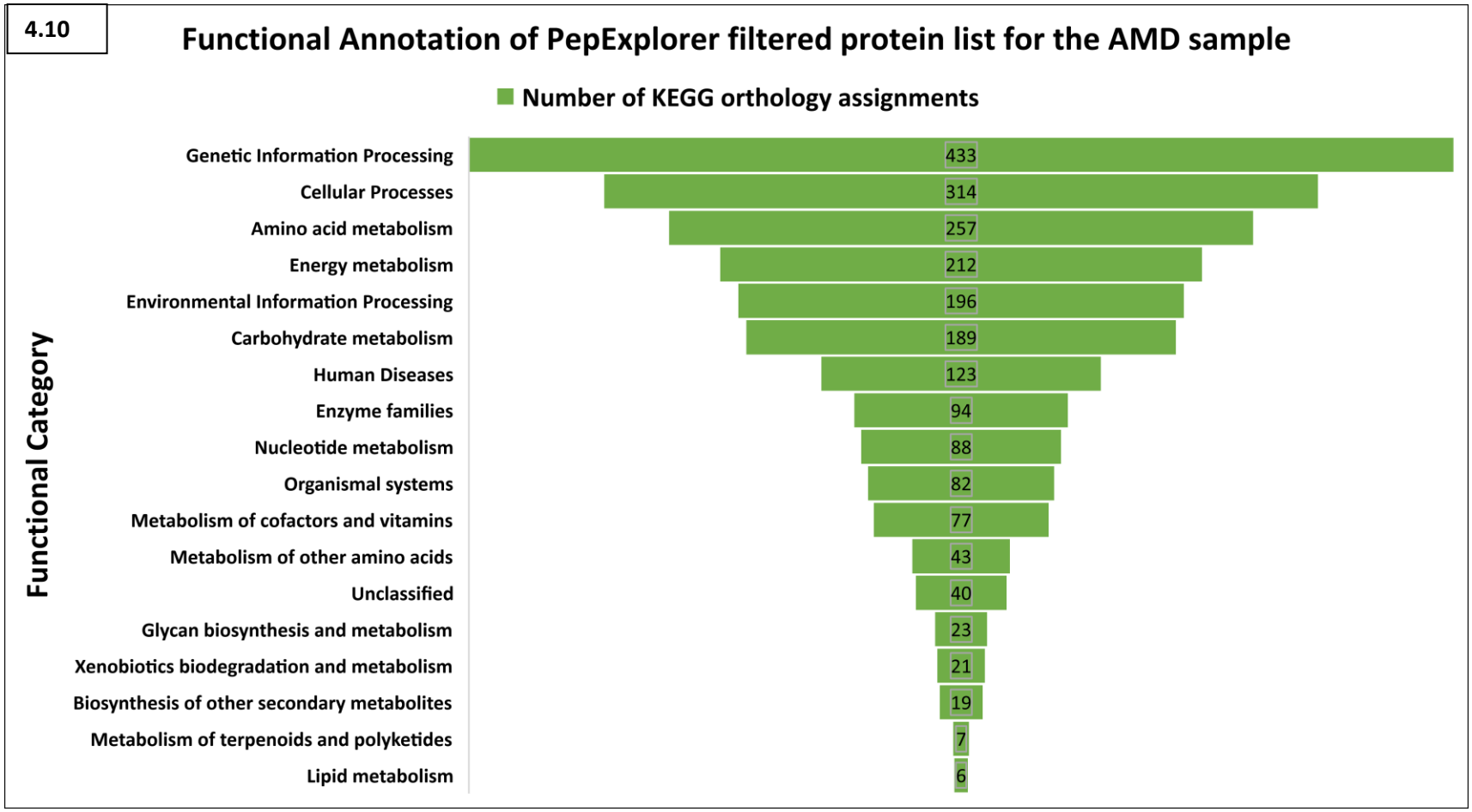


Figure 4.10: Functional annotation of AMD metaproteomic data obtained using GhostKOALA.

(description in text)

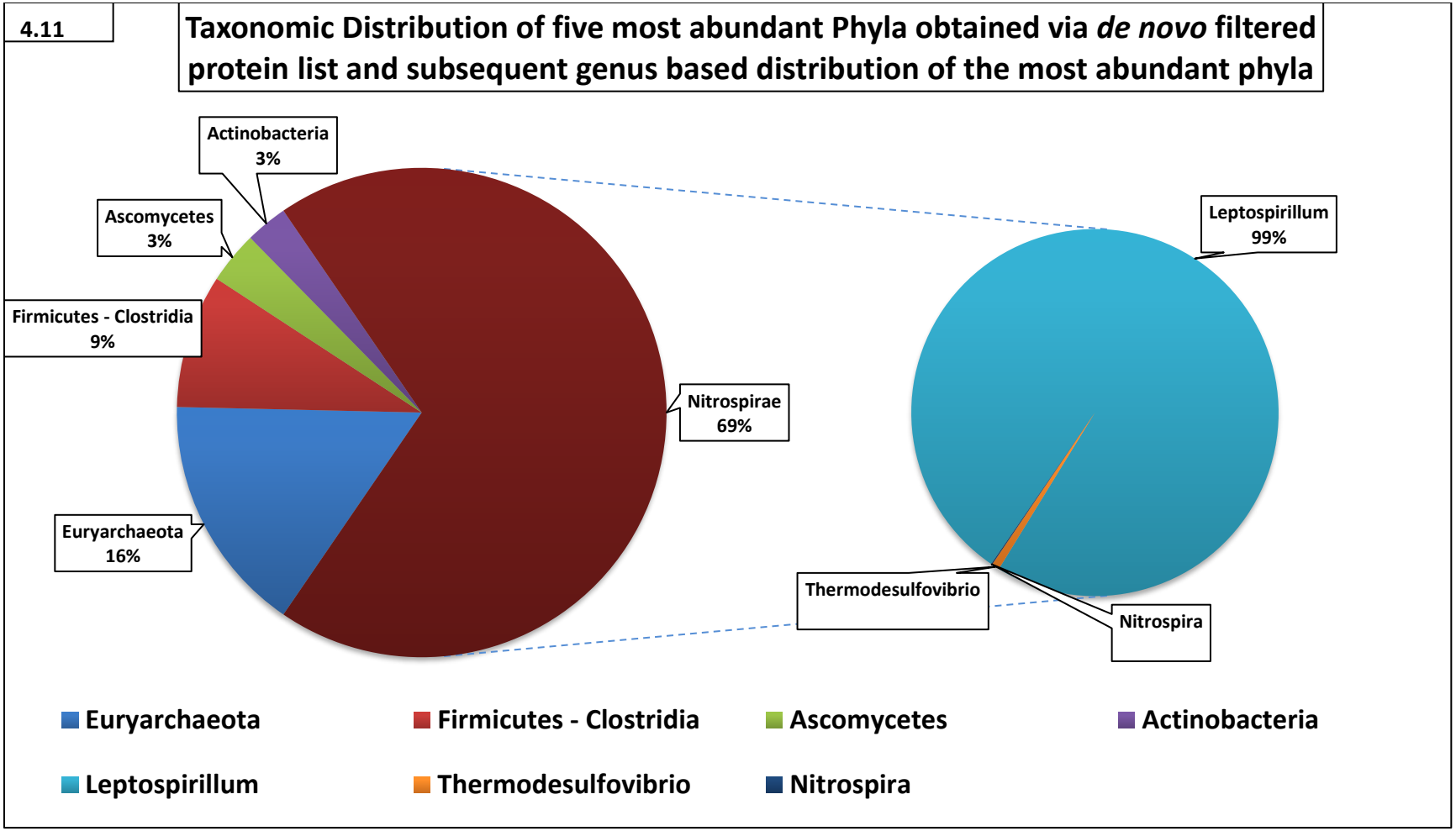


Figure 4.11: Taxonomic distribution of five most abundant phyla and subsequent genus based distribution of the most abundant phyla for the PepExplorer filtered protein list.

4.9 Conclusions:

The study presented here describes a two-pronged computational strategy for improving the reliability and depth of metaproteomic identifications. The results presented with respect to the thiocyanate community sample provided the initial impetus, as it highlighted the restrictions associated with database search results for metagenomes. A formidable effort is already underway to develop bioinformatic strategies that can tackle the complicated landscape in metaproteomic data analysis for multispecies samples (such as microbial communities) [44, 126, 140, 155, 156]. The combined database search strategy comprising of multiple iterative search engines increased the PSM's, peptides and proteins when analyzing environmental samples of varying complexities. This allowed us to extract as much information as possible using the well-established database search strategy before embarking on other complimentary methods like *de novo* sequencing. Using our approach, we have shown that a significant number of *de novo* sequencing peptides could be matched back to the original protein sequence database especially for multispecies samples like AMD. Earlier studies on metaproteomic data analysis were solely focused on presenting the number of PSM's and peptides and did not infer taxonomic or functional information [42]. In the current study, we have demonstrated the utility of *de novo* sequencing for improving the depth of meta information that can help in providing additional details while understanding complex ecosystems. We have shown a strategy for using a stringent filtering criteria for selecting *de novo* predicted peptides and mapping these peptides back to the database to identify novel protein candidates thus increasing the depth of biological information extracted from these complex datasets. Software enabling LCA analysis of metaproteome data (Unipept) can also be

used for *de novo* predicted peptides and can provide reliable results at the species level. But proper filters with specified thresholds (like high quality *de novo* peptides with significantly longer amino acid sequences) must be set to reduce false positives.

In conclusion, the proposed computational strategy will be useful for all researchers involved in complex microbiome studies and provides a template for improving the reliability and analysis depth of metaproteomic investigations.

Chapter 5 – Applications of Metaproteomics for Investigation of Microbial Dynamics in Ground Water Specimens

---

Part of the text and figure 5.8 was taken from: Rose S. Kantor, Robert J. Huddy, **Ramsunder Iyer**, Brian C. Thomas, Christopher T. Brown, Karthik Anantharaman, Susannah Tringe, Robert L. Hettich, Susan T. L. Harrison, and Jillian F. Banfield. Genome-Resolved Meta-Omics Ties Microbial Dynamics to Process Performance in Biotechnology for Thiocyanate Degradation. *Environmental Science & Technology* **2017** *51* (5), 2944-2953 DOI: 10.1021/acs.est.6b04477

Ramsunder Iyer's contributions to this work included: Sample preparation and experimental design for mass spectrometry measurements, data acquisition and analysis for the proteomics section.

5.1 Introduction:

The coupling of high performance multi-dimensional liquid chromatography and tandem mass spectrometry for deep proteome characterization of microbial proteins from complex environmental samples (metaproteomics) has set the platform for a new era in scientific discovery for complex microbial communities. Coupled with that, the other 'metaomics' methods like metagenomics and metatranscriptomics are increasingly being used to understand crucial processes and potential weak points in biotechnology. One such process is the nitrogen and phosphorus removal or bulking in wastewater treatment [157-159]. The genome-resolved metaomics approach has played an important role in industrial waste water treatment [160-163]. Using high-throughput sequencing of DNA and RNA followed by spectral interrogation of these communities has provided a holistic approach to identify key species involved in the wastewater treatment. Further enhancement in the activities and abundance of

these organisms under varying conditions could provide crucial insights for designing and operations of these systems.

One of the widespread contaminants found in gold mining effluents is Thiocyanate (SCN^-). It is found in especially high concentrations (up to 4000 mg/L). Its bioremediation is crucial as it can affect human health and aquatic organisms. [164-166]. Notably, SCN^- is inhibitory toward iron- and sulfur-oxidizing microorganisms used in bio-oxidation processes at some gold mines (such as BIOX) and therefore must be removed before wastewater can be recycled within a mining site or discharged into the environment. Chemolithoautotrophic bacteria use SCN^- as their source of energy and thus can help in its biodegradation [167-171]. The initial degradation products of this process are ammonium, carbon dioxide and reduced sulfur compounds. An industrial-scale process known as Activated Sludge Tailings Effluent Remediation (ASTER, Outotec, South Africa) successfully treats SCN^- containing wastewater at several gold mines.

A long running SCN^- fed bioreactor (known as SCN^- stock reactor) at the University of Cape-Town, South Africa has been inoculated with the sludge obtained from the ASTER process. The reactor consists of diverse microbial communities that have previously been studied and consist of several abundant *Thiobacillus species* that are involved in SCN^- degradation due to the presence of an operon in the genomes of these autotrophic bacteria. Results also suggested the presence of other heterotrophs and the potential to remove nitrogen by *Thiobacillus species* [172]. However, questions regarding the stability of the community under different SCN^- loadings, expression of observed metabolic potential and the significance of inter-organism interactions for nitrogen removal still remain unanswered.

In the current study, time-series genome resolved metagenomics followed by metaproteomics of samples from the final time point was used to track changes in the microbial community. For this, the newly inoculated SCN^- bioreactor was operated under increasing loadings. In order to enrich for organisms that can use and remove the nitrogen produced by SCN^- degradation, a control reactor with the same inoculum was fed with ammonium sulfate ($NH_4(SO_4)_{1/2}$) and molasses to mimic the breakdown products of SCN^-. The study was used to describe the microbial community structure, protein expression, and replication rates in both reactors during the experiment. The analysis interconnects the community membership and changes in reactor functions. The metaproteomics study highlights the key organisms and metabolic pathways active under high SCN^- loadings.

5.2 Materials and Methods:

5.2.1 Sample Types:

Two continuous stirred tank reactors were inoculated with homogenized biofilm and planktonic samples from the long-running SCN^- stock reactor at the University of Cape Town [172]. One was fed KSCN (Experimental reactor) and the other was fed ammonium sulfate[$NH_4(SO_4)_{1/2}$] (Control reactor) at equivalent nitrogen loadings to mimic thiocyanate degradation end-products. Feed concentrations were increased over 218 days, and feeds also included molasses (150 mg/L) and KH_2PO_4 (0.28 mM) as supplemental nutrients. The reactors reached a final loading rate of 1.43 mmol/h SCN^- or $NH_4(SO_4)_{1/2}$ at 12 hours HRT. Concurrent with the metagenomic sampling, three samples for metaproteomics were taken. These included a biofilm sample from the experimental bioreactor and a biofilm and planktonic sample from the

control bioreactor. Biofilm samples were scraped off the reactor wall (the interiors of both reactors accumulated biofilm) while planktonic samples were taken by filtering liquid from inside the reactors. **Figure 5.1** depicts the schematics of samples chosen for metaproteomic characterization in the study.

5.2.2 Protein Extraction and Proteomic Analysis:

Proteins were extracted as described previously [173] and ~1 mg of protein was subjected to trichloroacetic acid precipitation and subsequent digestion with trypsin. Proteolytic peptides were analyzed via an online nano 2D LC–MS/MS system interfaced with hybrid LTQ-Orbitrap-Velos MS (ThermoFisher Scientific). The samples were run in technical triplicates. A 25-µg aliquot of peptides was loaded onto a biphasic column consisting of reverse phase followed by strong cation exchange and analyzed by eleven step MudPIT (Multidimensional protein identification technology) as described previously [112] . The instruments were operated in a data-dependent mode. MS1 was performed in Orbitrap and data dependent MS/MS was performed in LTQ (top twenty), 1 microscan for both full and MS/MS scans; normalized collision energy 35% and dynamic exclusion time of 30 seconds. MS2 mass spectra were analyzed using the following software protocol: Thermo RAW files were converted to mzML peaklists by ProteoWizard msConvert [174] and database searches used Myrimatch [86], with the de-replicated set of genomes as the database. Configuration parameters were as follows: fully tryptic peptides with any number of missed cleavages, an average precursor mass tolerance of 1.5 m/z, a mono precursor mass tolerance of 10 ppm, a fragment mass tolerance of 0.5 m/z, a static cysteine modification (+57.0214 Da), an N-terminal dynamic carbamylation modification

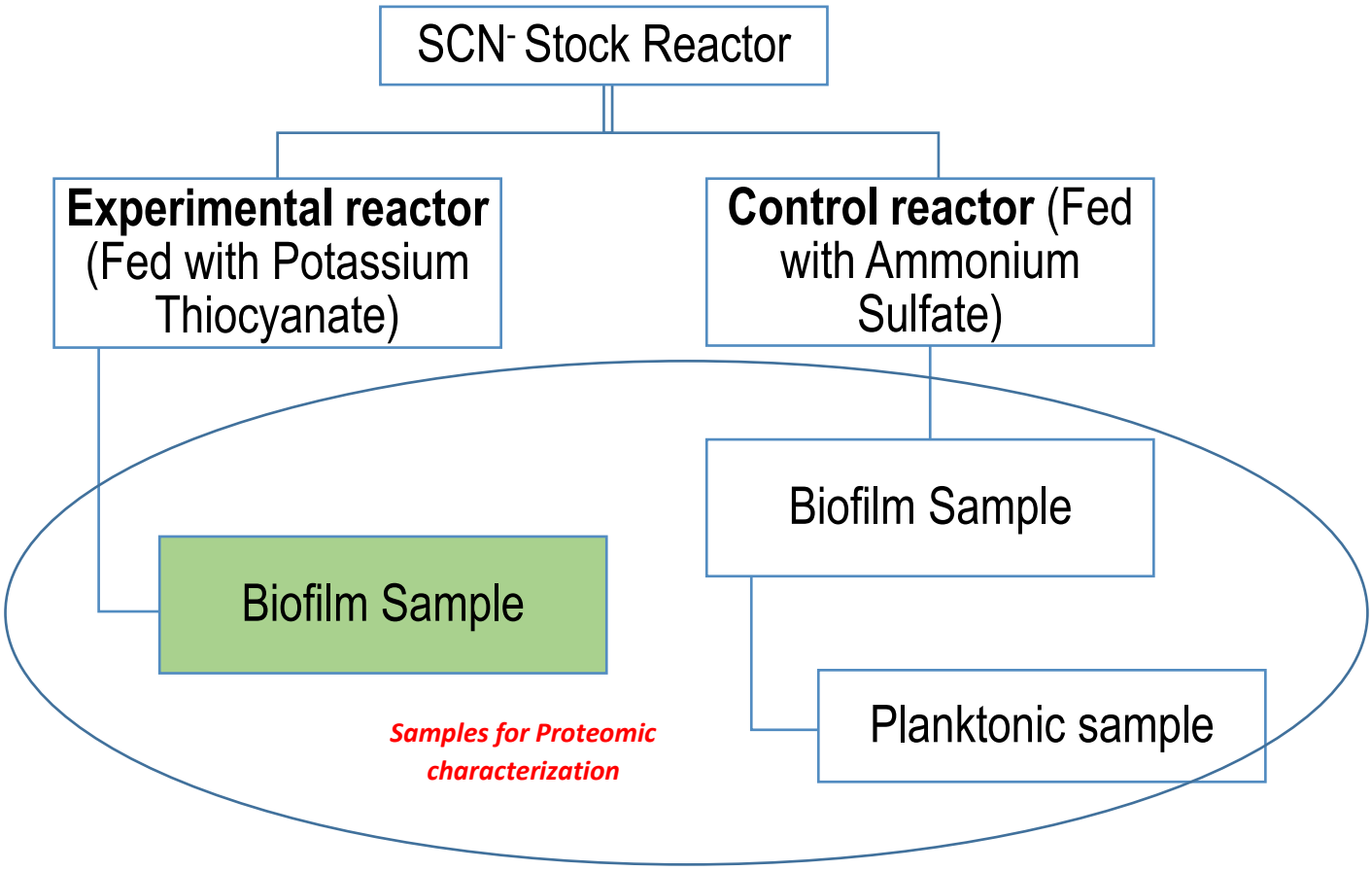


Figure 5.1: Schematics of the sample types chosen for the thiocyanate study (description in text)

(+43.0058 Da), and a dynamic oxidation modification (+15.9949 Da). Peptide identifications were filtered with IDPicker v3.1 [92] to <1% peptide FDR (at the peptide level: maximum Q value <2%, minimum one spectra per peptide, and minimum one spectra per match; at the protein level: minimum two distinct peptides, minimum one additional peptide, and minimum two spectra per protein). ScanRanker [175] was used to assess the spectral quality.

5.2.3 Metaproteome Bioinformatics:

Two different versions of the metaproteome database were provided by collaborators for proteomic searching **(figure 5.2)**. The first database was constructed by assembling the metagenome from a single sample from the experimental biofilm; the assembly accounted for 87.5% of all reads in that sample (allowing 3 mismatches per 250 bp read).

The second database was a compilation of 15 metagenomes from two samples (planktonic inoculum and SCN^- biofim), dereplicated at the level of individual genomes. Each of the 15 samples was assembled independently, the data was binned into genomes, and then the genomes from different samples were compared and dereplicated at 98% nucleotide identity). The database accounted for 90.4% of all reads in the sample (allowing 3 mismatches per 250 bp read).

5.3 Results and Discussion:

5.3.1 Impact of Database Versions on Protein/Peptide Identifications:

In order to study the impact of database versions on the depth of proteomic analysis, we used the collected mass spectra of the experimental biofilm sample (sample highlighted in green in

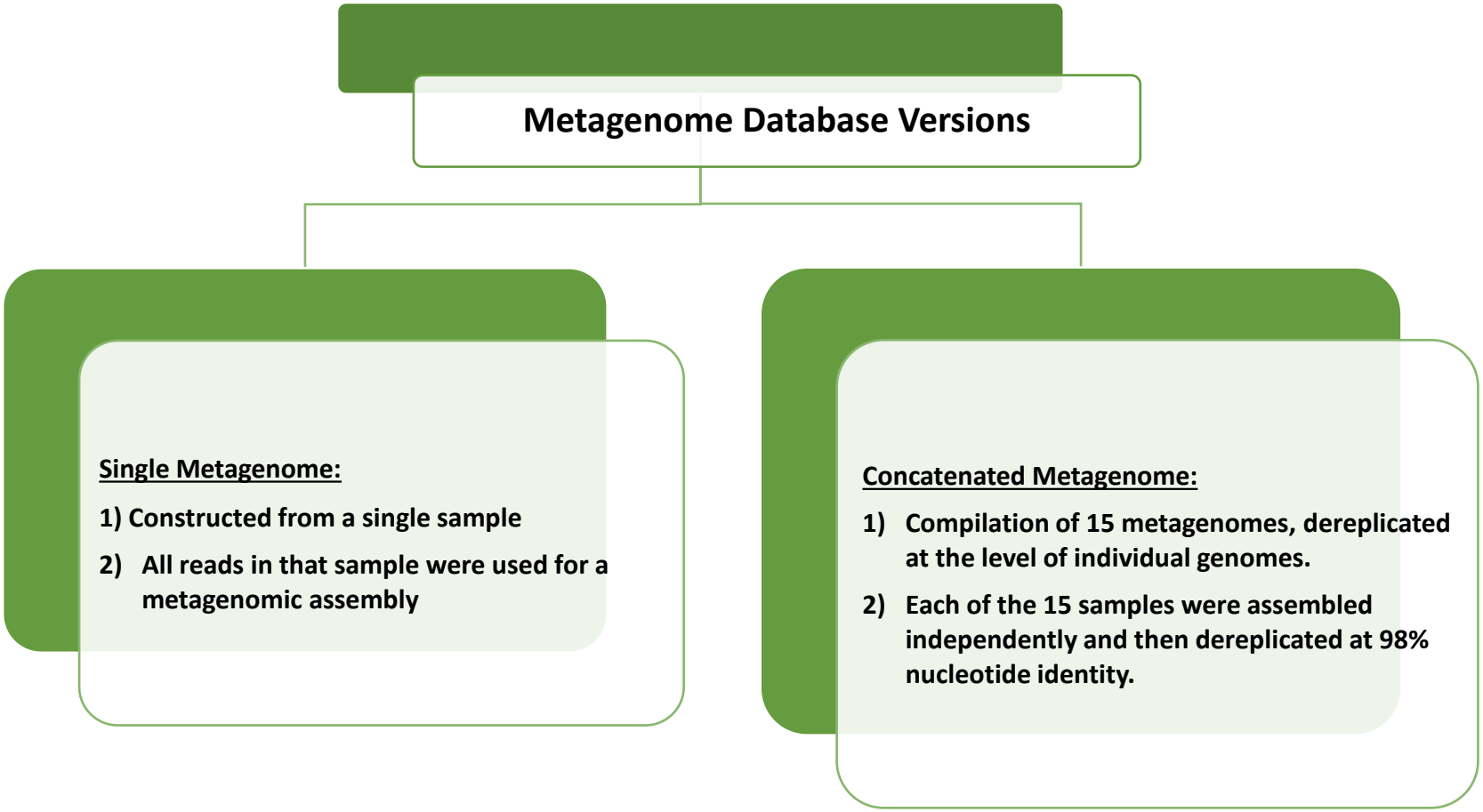


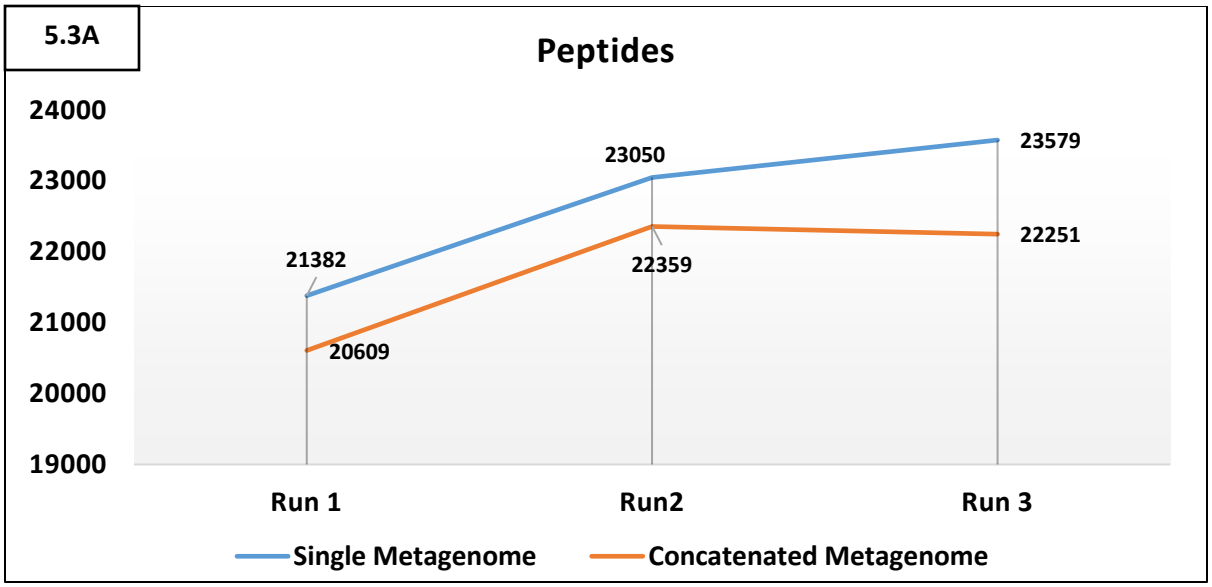
Figure 5.2: Schematic illustration of the database versions used in the thiocyanate study.

Figure 5.1) and searched it against the single and concatenated metagenomic assemblies. **Figures 5.3A and 5.3B** depict the peptide and protein counts respectively of each replicate when the data was searched with single and concatenated metagenomes. Searching the data with single metagenome assembly yielded an average total of 22670 peptides (SD = 1146.7) and 6619 proteins (SD = 241.1) and 73088 spectra (SD = 469.9) for the three technical replicates. Concatenated metagenomic assembly on the other hand identified an average total of 21739 peptides (SD = 980.2), 7029 proteins (SD = 254.4) and 70001 spectra (SD = 385.8) for the three technical replicates.

The concatenated metagenomic assembly identified more proteins in all the technical replicates **(figure 5.3B)**. However, peptide counts were observably less for the concatenated assembly **(figure 5.3A)** than the single metagenome. This discrepancy can be explained by the process of protein assembly used in IDPicker [92].

IDPicker has an easily comprehensible approach for assembling peptides to proteins in shotgun proteomic experiments. A 'distinct peptide' in IDPicker is defined as a distinct sequence of amino acids. In particular, post-translational modifications and charged states are disregarded. However, different isobaric sequences do count as distinct. For example, peptides ISBARIK and ISBARLK will be counted as distinct from one another. The peptide counts shown in **figure 5.3A** are all distinct peptides.

The process of protein assembly is explained by the bipartite graph shown below **(figure 5.4)**. The boxes in red indicate protein accessions and those in blue indicate peptide sequences, while those in green indicate clusters. Now, proteins are identified as belonging to the same



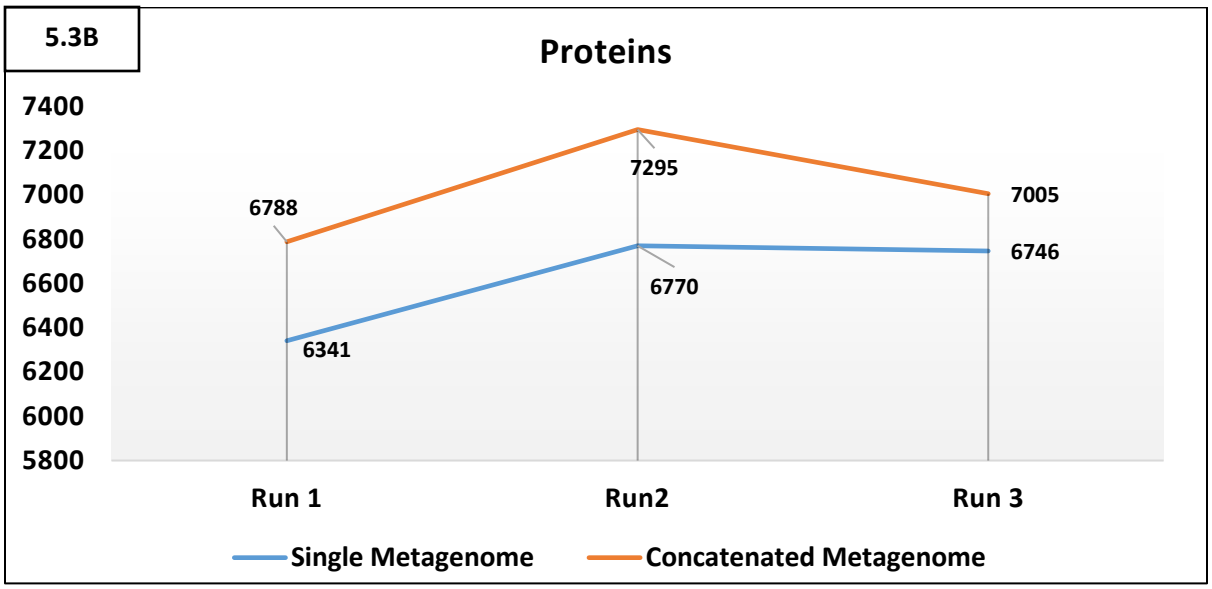


Figure 5.3: Peptide (5.3A) and protein (5.3B) counts respectively of each replicate when the thiocyanate data was searched with single and concatenated metagenomes.

Searching the data with single metagenome assembly yielded an average total of 22670 peptides (SD = 1146.7) and 6619 proteins (SD = 241.1) and 73088 spectra (SD = 469.9) for the three technical replicates. *Concatenated metagenomic assembly on the other hand identified an average total of 21739 peptides (SD = 980.2), 7029 proteins (SD = 254.4) and 70001 spectra (SD = 385.8) for the three technical replicates.*

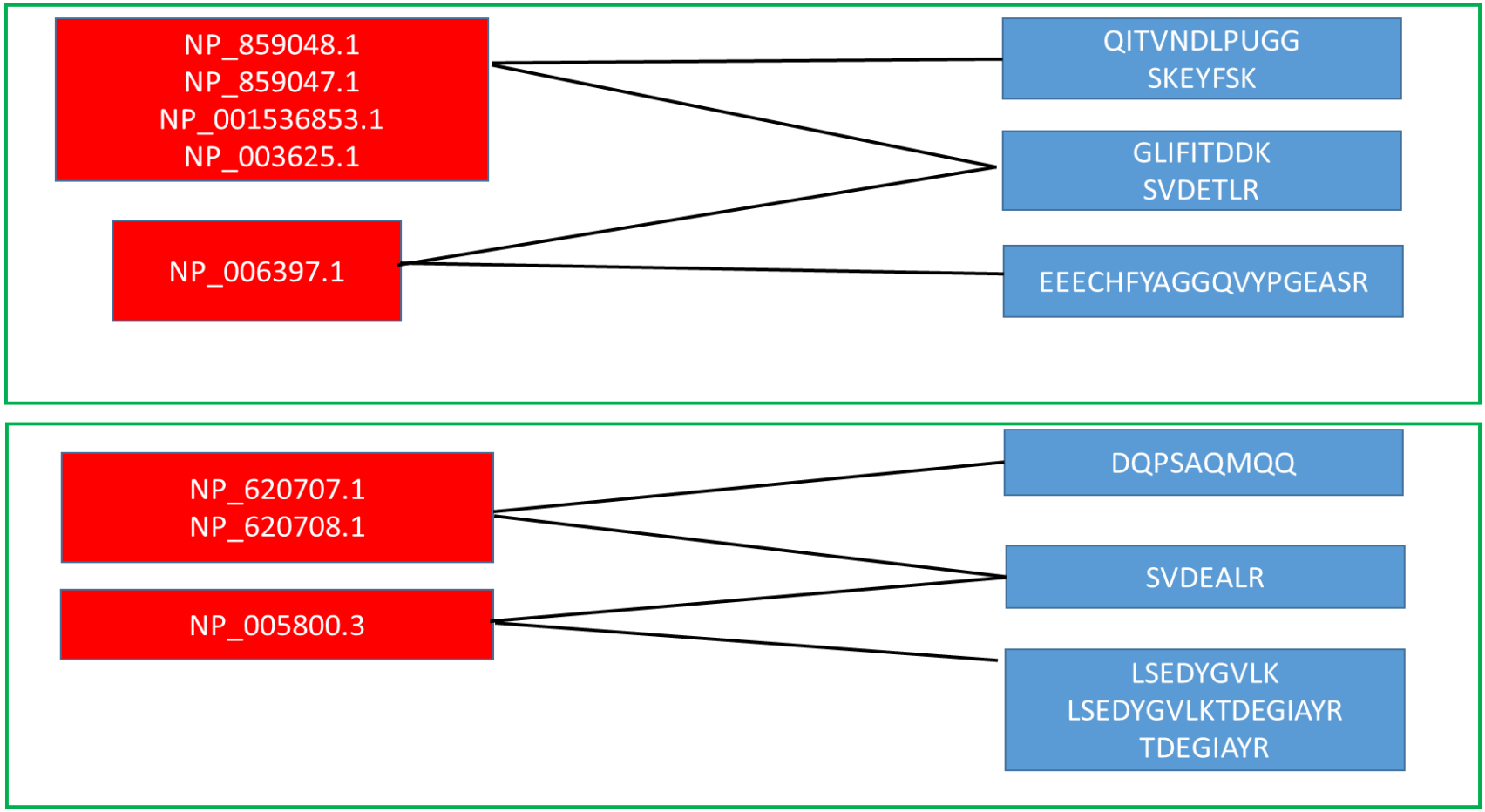


Figure 5.4: Bipartite graph explaining the process of protein assembly in IDPicker.

group when they are evidenced by exactly the same sequence of peptides. In the above case, the first green box has two protein groups that are identified by three different subsets of peptide groups. However, the second protein group has a peptide that is unique to it (peptide having sequence EEECHFYAGGQVYPGEASR). Clusters are a way to organize peptide and protein groups into associated subsets. Now since there is no connection between first two protein groups and last three peptide groups, two subsets are created.

In the current study, while assembling peptides to proteins in IDPicker, we used a criterion of two distinct peptides for a protein to be called as a true hit **(section 5.2.2)**. In case of single metagenomic assembly, there were several proteins that were found to have more than two distinct peptides (three or more) and thus, in spite of the fact that there were more distinct total peptides, they matched to fewer total proteins. On the contrary, while assembling peptides to proteins using the concatenated metagenomic assembly, there were several proteins that had just two distinct peptides matching to them. Hence less number of total distinct peptides were matching to more number of total proteins.

5.3.2 Peptide Redundancy Between the Single and Concatenated Metagenomes:

Next, we calculated the total number of peptide sequences shared between single and concatenated metagenomes. The 2D LC MS/MS analysis of tryptic digests of the experimental biofilm sample resulted in cumulative identification of 33,515 distinct, non-redundant tryptic peptide sequences using the single metagenome assembly. The concatenated metagenomic assembly on the other hand identified 32,311 distinct, non-redundant peptides. The Venn diagram in **figure 5.5** depicts the peptide redundancy i.e. the total number of shared peptides

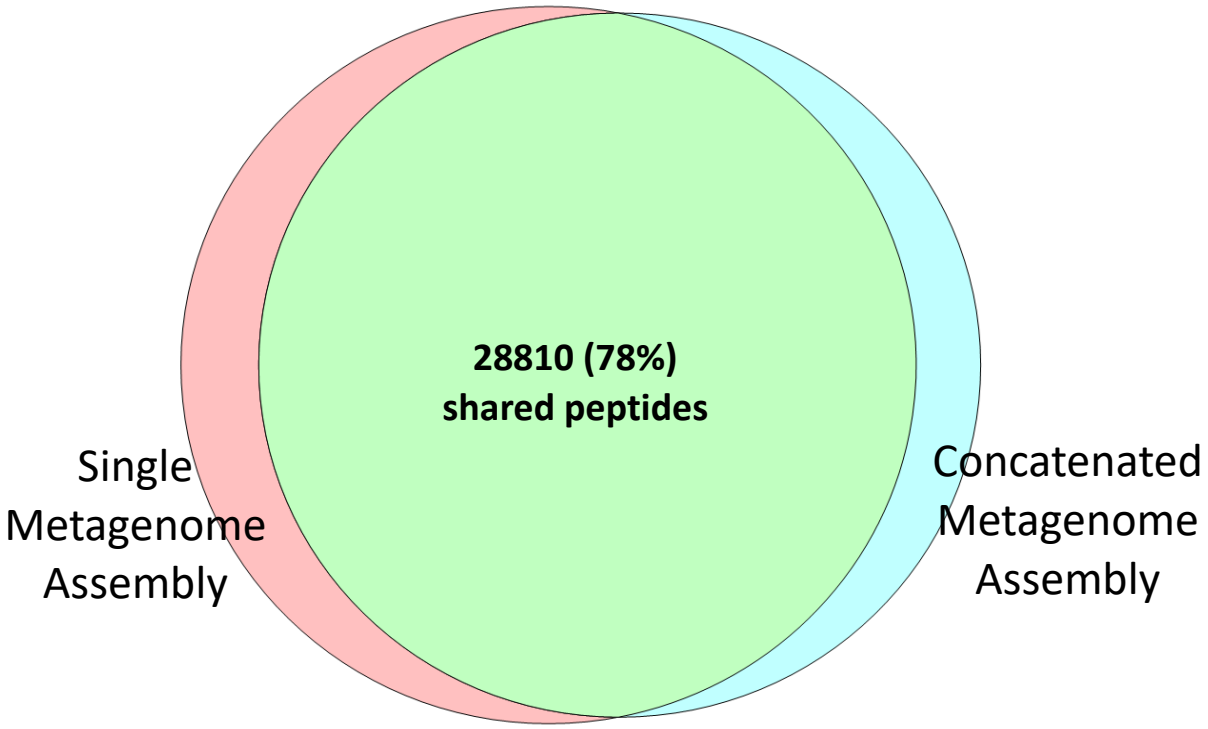


Figure 5.5: Venn diagram depicting the peptide redundancy i.e. the total number of shared peptides between the single and concatenated metagenomic databases.

between the single and concatenated metagenomic databases. A total of 28,810 peptides (78%) were shared between the two databases with 4705 (13%) and 3501 (9%) unique to single and concatenated metagenomes respectively.

5.3.3 Spectral Quality Assessment of Single and Concatenated Metagenomic Assemblies:

The success in achieving accurate protein identifications and a deeper proteome coverage in a complex metaproteomic specimens is heavily reliant on the quality of a predicted protein sequence database that is constructed from metagenomic data. As compared to single cell type/microbial isolate which are well assembled and curated, a larger portion of high-quality spectra in a metaproteomic study remain unassigned due to the incompleteness of the proteomic database. To quantify this, we used a spectral quality assessment tool, ScanRanker [175] which assigns scores for all of the collected spectra which in turn can be used to evaluate the quality of the database. Using ScanRanker scores, a distribution of total collected spectra including unassigned and assigned microbial spectra was plotted for all the three samples under consideration (control-planktonic, experimental-biofilm and control-biofilm). **Figures 5.6A, 5.6B and 5.6C** represent the distributions of ScanRanker scores for collected mass spectra of control-planktonic, experimental-biofilm and control-biofilm respectively searched with single metagenomic assembly. **Figures 5.7A, 5.7Band 5.7C** denote the distributions of ScanRanker scores for collected mass spectra of control-planktonic, experimental-biofilm and control-biofilm respectively searched with concatenated metagenomic assembly.

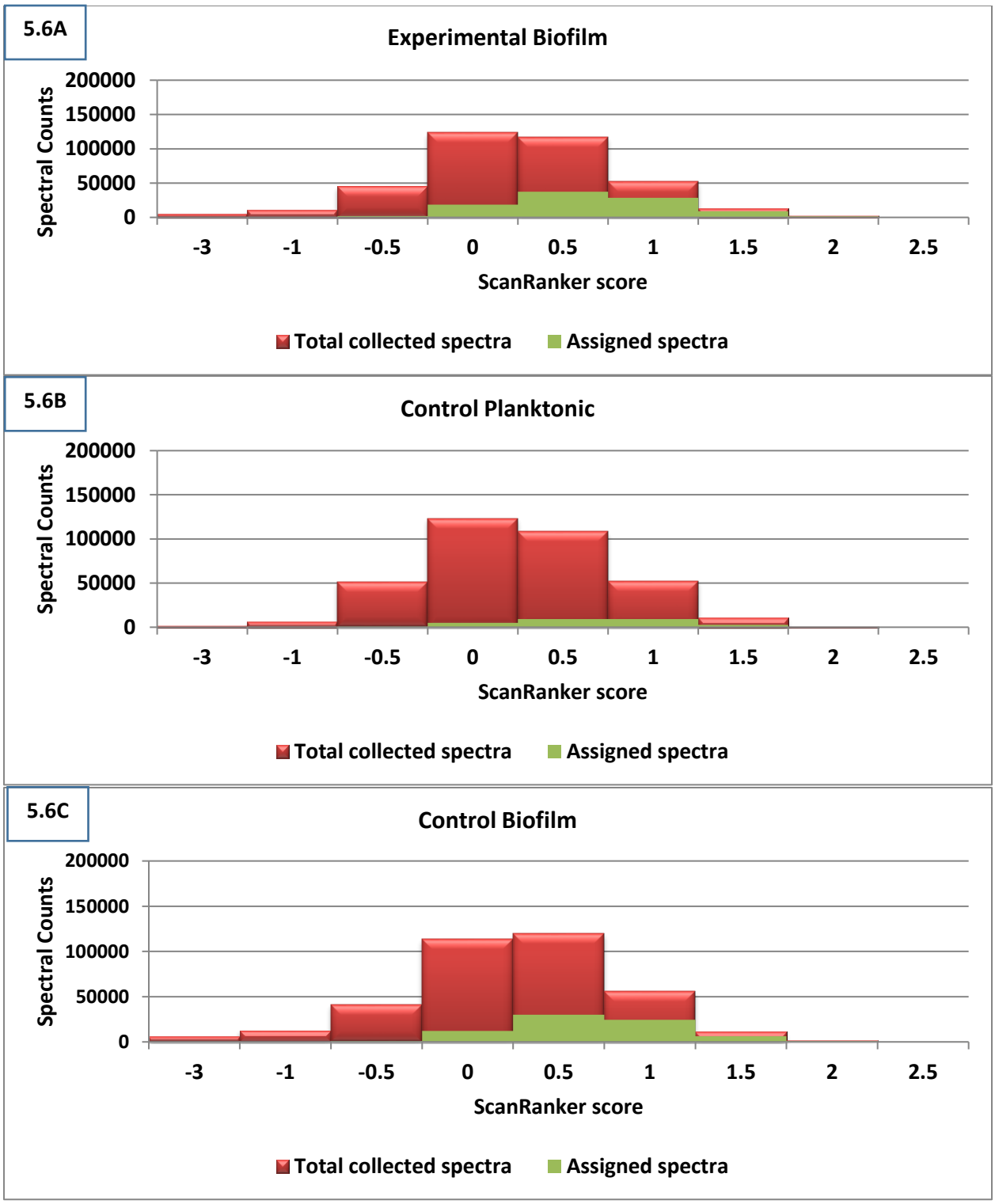


Figure 5.6: Stacked histograms representing the distributions of ScanRanker scores for collected mass spectra for control planktonic (5.6A), experimental biofilm (5.6B) and control biofilm (5.6C) respectively searched with single metagenomic assembly.

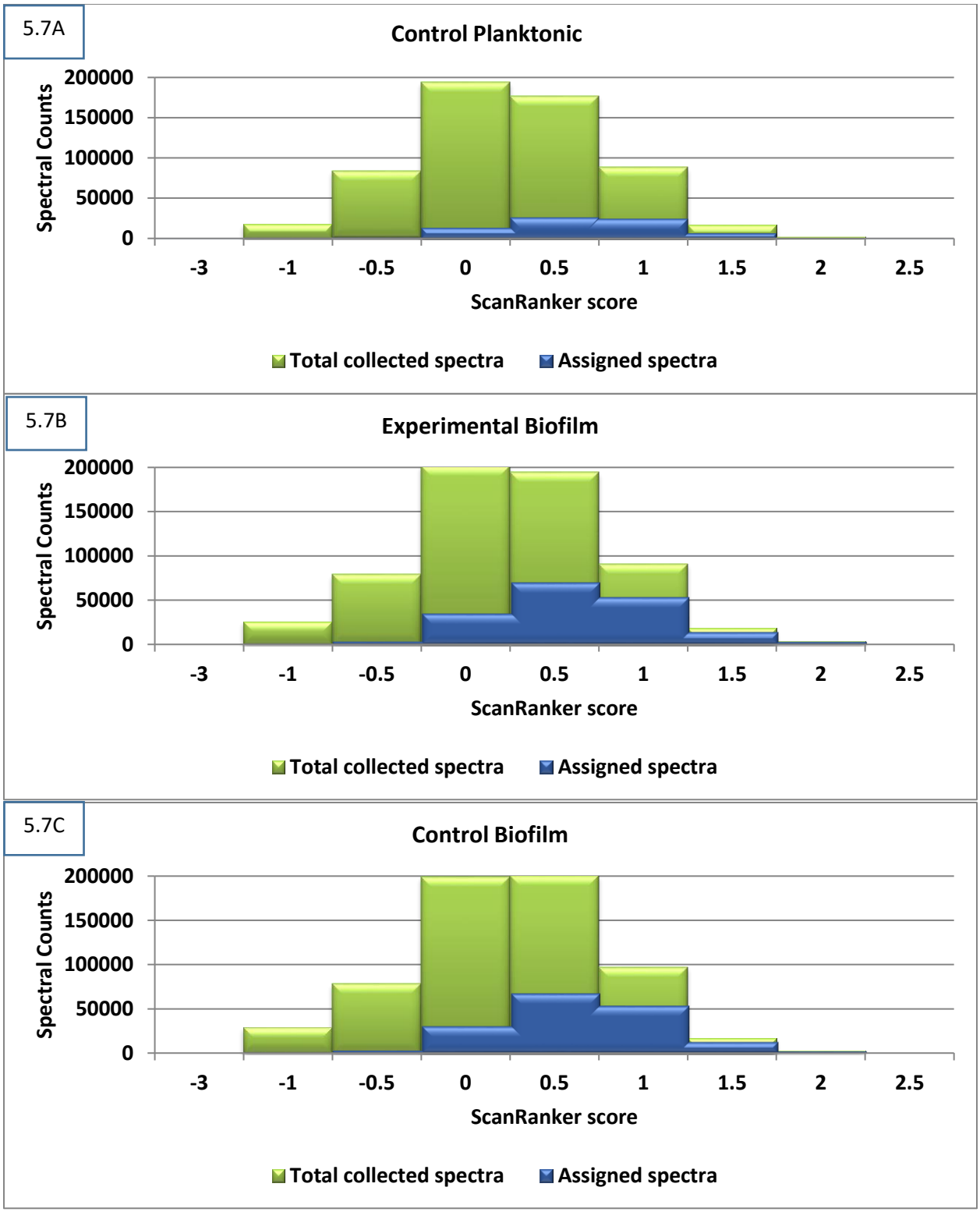


Figure 5.7: Stacked histograms representing the distributions of ScanRanker scores for collected mass spectra for control planktonic (5.7A), experimental biofilm (5.7B) and control biofilm (5.7C) respectively searched with concatenated metagenomic assembly.

In general, peptide identifications start occurring at a ScanRanker score of -0.5 and above. Spectra having scores below that range can be accounted for by experimental noise, implying that lower-quality spectra reside at the lower end of the distribution. Although somewhat variable for microbial isolates, we typically note that ~60% of collected mass spectra can be assigned to peptides for an organism with a completely sequenced genome (without accounting from PTMs, sequence variants, and other unknown contaminants). However, due to the increased complexity of these samples, as well as the fact that the metagenomic databases used here are incomplete, approximately 8%, 27% and 21 % of total collected spectra were assigned for control planktonic, experimental biofilm and control biofilm respectively searched with single metagenomic assembly. There was a substantial improvement in case of concatenated metagenomic assembly where 15%, 34% and 32% of the total collected spectra were assigned to control planktonic, experimental biofilm and control biofilm respectively (Note that the percentage values described here are the ones that had a ScanRanker score of -0.5 and above. Scores below that threshold were accounted for experimental noise and were discarded). The level of identification obtained using concatenated metagenome is comparable to previous studies using the same type of analysis on infant gut metaproteomes paired to metagenomic datasets [103]. Thus, meticulous construction of databases as in case of concatenated metagenomes that combines several samples can play an important role in boosting the coverage and depth of metaproteome data.

5.3.4 Expression of Genes Involved in SCN^- Degradation Confirmed by Proteomics:

The concatenated metagenomic assembly which was dereplicated at the level of individual genomes yielded a non-redundant bacterial set of 144 draft quality genomes. Besides this, eukaryotic, mitochondrial, chloroplast, phage and plasmid genomes were also recovered. This dereplicated set of metagenomic data was subsequently used for proteomic searches.

Figure 5.8 depicts the expression of genes involved in SCN^- degradation. The arrows denote the average of unique spectral counts across technical replicates ≥ 2 for a minimum of one subunit of enzyme involved in SCN^- degradation. There are two known types of SCN^- hydrolases and four genomes, *Thiobacillus_1, Thiobacillus_3, Thiobacillus_4, and Afipia_1*, contain one of these two known hydrolases [167, 176]. Proteomics data were able to support the activity of these organisms in SCN^-degradation, sulfur oxidation and carbon fixation. A recently described SCN^- operon from *Thiobacillus* spp.[172] contains a set of proteins whose corresponding peptide sequences were confirmed by shotgun proteomics. Specifically, proteomics analysis was able to reveal the expression of genes in this operon in at least one of the three *Thiobacillus* genomes. The SQR-like protein which is one of the members of this operon had the highest count of unique spectral hits for all three *Thiobacillus* genomes. The CbiM-like protein hypothesized to be involved in cobalt metabolism was the only protein from the operon that was not detected by proteomics. Consistent with previous studies the *Thiobacillus* genomes encode the potential for autotrophic growth on sulfur compounds [177]. These sulfur oxidation genes for several of the *Thiobacilli* was confirmed by proteomics.

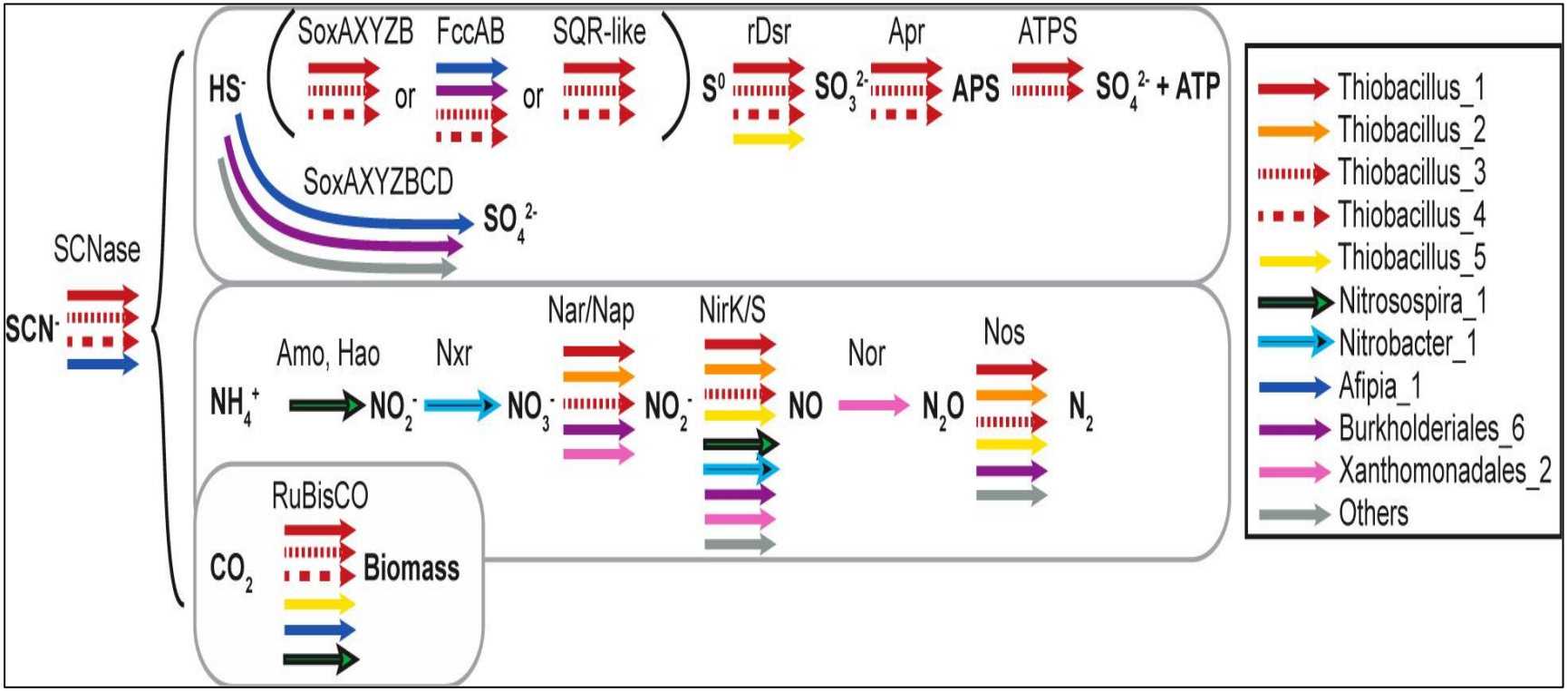


Figure 5.8: Metaproteomics in SCN– reactor showing expression of genes involved in SCN– degradation and by product breakdown.

Each arrow indicates that the average of unique spectral counts across two technical replicates was ≥2 for at least one subunit or component of the enzyme complex involved in SCN⁻ degradation.

Published in: Rose S. Kantor; Robert J. Huddy; Ramsunder Iyer; Brian C. Thomas; Christopher T. Brown; Karthik Anantharaman; Susannah Tringe; Robert L. Hettich; Susan T. L. Harrison; Jillian F. Banfield; Environ. Sci. Technol. ***2017,*** *51, 2944-2953. DOI: 10.1021/acs.est.6b04477.* *Copyright © 2017 American Chemical Society*

Another important feature observed in community dynamics was the increase in relative abundance of *Burkholderiales*_6. Although, *Burkholderiales*_6 genome lacks any genes for SCN^- degradation, but previous studies have isolated and characterized a strain of *Burkholderia phytofirmans* capable of thiocyanate degradation with acetate as a carbon source [178]. The genome of *Burkholderiales_6* contains genes encoding the sox pathway, and the corresponding proteins were detected by proteomics. Overall, the results suggest a transition from autotrophic to mixotrophic/heterotrophic thiocyanate degradation at high thiocyanate loadings and after long periods of reactor operation.

Since SCN^- degradation releases nitrogen in the form of ammonium, possible mechanisms for nitrogen cycling and removal to N_2 was investigated. Two predicted nitrite oxidizers, *Nitrobacter*_1 and *Nitrobacter*_2, were present at low abundances in the $NH_4(SO_4)_{1/2}$, in the SCN^- reactor their abundance was so low that their genomes did not assemble. However, proteins for nitrite oxidation corresponding to one of these genomes were detected in samples from both reactors.

For the *Thiobacilli*, several denitrification-related complexes were detected with proteomics. The *Burkholderiales*_6 organism also likely contributed to denitrification. All denitrification genes except *norB* were identified in proteomics from the SCN^- reactor biofilm. The limited detection of NorB was presumed to be an extraction bias due to numerous transmembrane domains in these proteins [179, 180].

5.4 Conclusions:

The current study shows that meticulous construction of concatenated metagenomes by combining several samples plays a crucial role in boosting the metaproteome coverage. This can provide critical biological insights into the expression patterns of microbial functionality in the community. Proteomics data reveal the presence of several *Thiobacillus* spp., as well as *Burkholderiales*_6 that couple sulfur oxidation to denitrification. Their abundances, and proteomic evidence suggest the role of these organisms to be the key players in the denitrification process.

This work also highlights the applicability of bioinformatics tools to gain a mechanistic understanding of contaminant degradation by a microbial community, to assess community stability, and ultimately, to inform engineering design choices. This study represents a step toward the use of metaproteomics for waste water treatment. The level of resolution achieved using metagenomics combined with metaproteomics enabled not only phylogenetic classifications and diversity of community members, but also confirmed the expression of key proteins involved in the degradation process. The data set and analysis provide valuable information that can be used to generate primers or probes for on-site measurements.

Chapter 6 - Applications of Metaproteomics to Investigate the Expression of Multi-heme Cytochromes in Methane Seep Sediments

---

Part of the text and figures 6.1, 6.2 and 6.3 was taken from: Connor T. Skennerton, Karuna Chourey, **Ramsunder Iyer**, Robert L Hettich, Gene W. Tyson, Victoria J Orphan. Methane-fuelled syntrophy through extracellular electron transfer: uncovering the genomic traits conserved within diverse bacterial partners of ANME archaea. *MBio, 2017. 8(4).*

Ramsunder Iyer's contributions to this work included: Data analysis and generating figures for the proteomics section.

6.1 Introduction:

One of the dominant mechanisms for controlling methane flux in ocean sediments is mediated by syntrophic association between methanotrophic archaea (ANME) and deltaproteobacterial sulfate-reducing bacteria (SRB) [181, 182]. This process is called reverse methanogenesis which is also referred to as anaerobic oxidation of methane, and is a major source of methane removal in ocean sediments. Initial ecological studies described the association between ANME archaea and SRB [181, 182]. Since then, significant efforts have been placed into understanding this syntrophic partnership using various molecular techniques. Of these, metaproteomic characterization of natural samples and reactor systems has played a crucial role in revealing the mechanism of association between SRB and ANME archaea.

Recent molecular evidence supports the hypothesis which is based on extracellular electron transfer (EET) mediated by multi-heme cytochromes that pass the electrons produced during

methane oxidation by the ANME archaea directly to their bacterial partners. [183, 184]. The mechanism of EET has been studied previously in metal reducing organisms like *Shewanella* and *Geobacter* [185, 186]. Subsequently, comparative genomics on ANME-2 revealed the presence of very large multi-heme cytochromes [183, 187] that are similar in size to outer membrane cytochromes of *Shewanella* and *Geobacter.* The majority of research to date has mainly focused on the metabolism of ANME lineages; however, little is known on the diversity and metabolic potential of the associated SRB partners in methane seeps. The syntrophic partners of ANME archaea come from a number of environmental clades of *Deltaproteobacteria,* of which SEEP-SRB1 is most common bacteria. Homologs of SEEP-SRB1 cytochromes are present only in some other cultured *Deltaproteobacteria,* predominantly in the *Desulfuromonadales* that are known metal reducers, including genome bins from the sediment samples. Additionally, *Ca. Desulfofervidus auxilii [188]*, several related *Thermodesulfobacterales*, *Anaeromyxobacter*, and *Desulfurivibrio alkaliphilus* AHT2 [189] contain homologs of these cytochromes. This operon consists of a core set of four genes: A six-bladed beta propeller fold protein, the two cytochromes (16 heme and 26 heme), and a peptidyl-prolyl cis-trans isomerase protein **(figure 6.1)**. Surrounding this core of the operon, there are other proteins that vary in composition and number but often consists of smaller cytochromes, proteins containing beta propeller fold proteins and other membrane proteins of unknown function.

In the current study, metaproteomics was used to assess the expression of this operon *in situ* at three different sites along the west coast of North-America that contain methane seeps. These included Santa Monica Mounds, Eel river basin and the hydrate ridge. This study highlights the applicability of metaproteomics to reveal the expression of proteins involved in extracellular

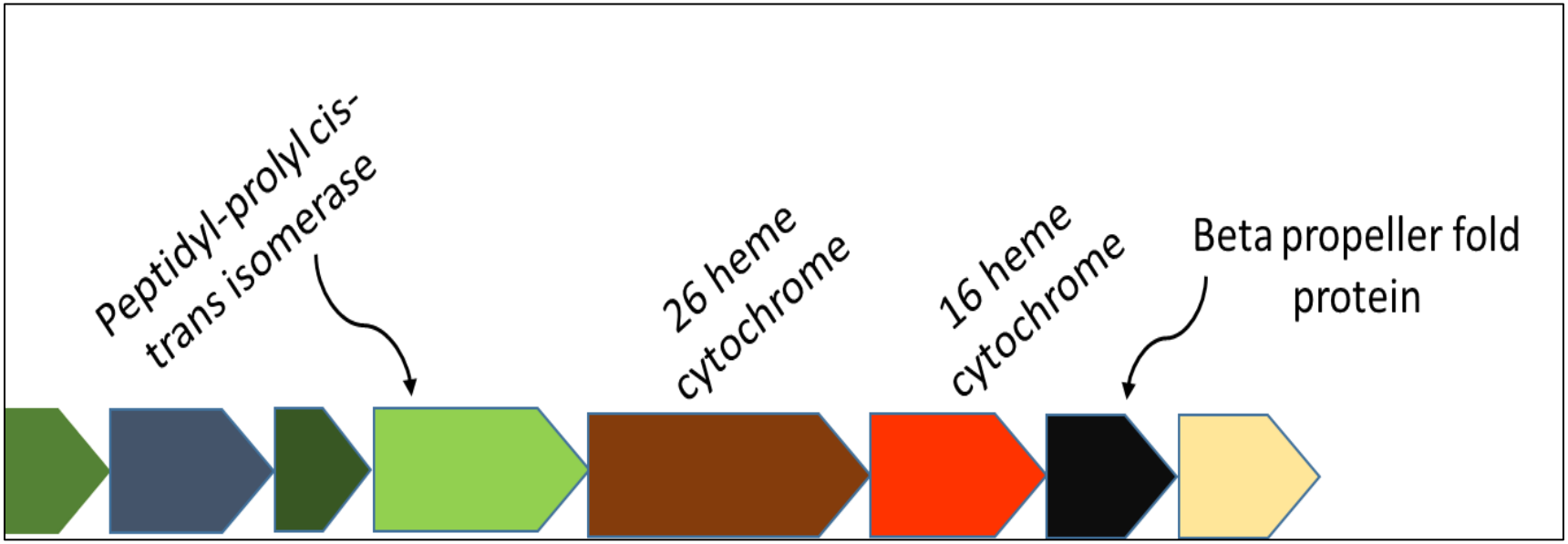


Figure 6.1: Representative operon structure from organisms containing large multiheme cytochromes found in SEEP-SRB1.

The operon consists of a core set of four genes: A six-bladed beta propeller fold protein, the two cytochromes (16 heme and 26 heme), and a peptidyl-prolyl cis-trans isomerase protein. Surrounding this core of the operon, there are other proteins that vary in composition and number but often consists of smaller cytochromes, proteins containing beta propeller fold proteins and other membrane proteins of unknown function.

electron transfer from ANME archaea to their SRB partners thus mediating the process of reverse methanogenesis.

6.2 Materials and Methods:

6.2.1 Sample Collection:

Five sediment samples were collected for metaproteomics from three locations **(figure 6.2)**. A 20-cm push core (PC48) was collected on 26 July 2005 from the Eel River Basin on dive T-863 of R/V Western Flyer using ROV Tiburon (coordinates: 40º 48.6631 N, 124º 36.7437 W; 520 m water depth) and divided into two sections of 10 cm (0 - 10cm and 10 - 20 cm). A second 20 cm push core was collected on 15th February 2005 on dive T-796 of R/V Western Flyer using ROV Tiburon from a mound a few hundred meters north-west of the venting mound in Santa Monica basin (coordinates: 33º 47.9748 N, 118º 38.796 W; 826 m water depth). This push core was divided into 4 cm segments; the 0-4 cm horizon and 8-12 cm horizon were used for proteomics. The final sample was the 3730-sediment collected from Hydrate Ridge that was also used for metagenomic sequencing.

6.2.2 Cellular Lysis, Protein Extraction and Sample Preparation:

Partially thawed seep sediments (5 gm) were suspended in 10 ml of detergent based lysis buffer, and subjected to heat-assisted cellular lysis as described previously [173]. The suspension was cooled down on bench top and centrifuged in fresh tubes for 5 min at 8000 g to settle the sediment. Resulting clear supernatant was aliquoted into fresh tubes and amended with chilled 100% trichloroacetic acid (TCA) to a final concentration of 25% (vol/vol) and kept at

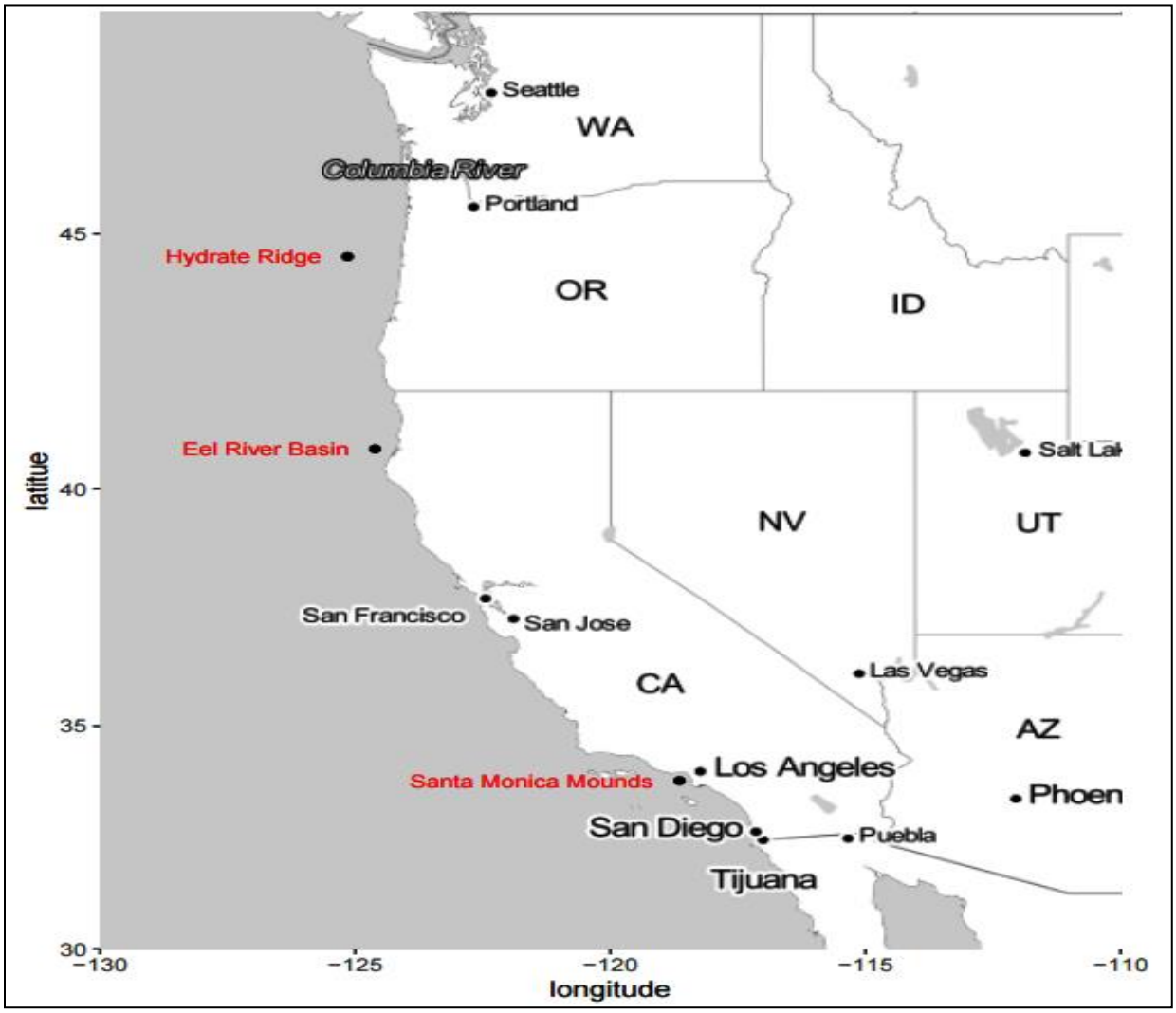


Figure 6.2: Metaproteomic Sample collection sites.

Eel River Basin (coordinates: 40º 48.6631 N, 124º 36.7437 W; 520 m water depth), Santa Monica basin (coordinates: 33º 47.9748 N, 118º 38.796 W; 826 m water depth) and Hydrate Ridge.

-20°C overnight. The residual sediment was discarded. Following overnight TCA precipitation, the supernatant was centrifuged at 21,000 x g for 20 min to obtain a protein pellet. The pellet was retained and washed thrice with chilled acetone [190] (air dried and solubilized in 6 M guanidine buffer (6 M guanidine; 10 mM dithiothreitol [DTT] in Tris-CaCl2 buffer (50mM Tris; 10mM CaCl2, pH 7.8) and incubated at 60°C for three hours with intermittent vortexing. An aliquot of 25 µl was utilized for protein estimation, carried out using the RC/DC protein estimation kit (Bio-Rad Laboratories, Hercules, CA, USA) as per the manufacturer's instructions. Remaining protein sample was diluted six-fold using Tris-CaCl2 buffer and trypsin was added (40 ug/1-3 mg total protein) based on protein estimation results. Proteins were digested overnight at 37 ^{0}C with gentle mixing and resulting peptides were reduced by addition of DTT (10 mM) and desalted using seppak column and solvent exchanged [191]. Peptides were stored at 80°C until MS analysis. All chemicals used in sample preparation and mass spectrometry analysis were obtained from Sigma Chemical Co. (St. Louis, MO), unless mentioned otherwise. Sequencing-grade trypsin was acquired from Promega (Madison, WI). High performance liquid chromatography- (HPLC-) grade water and other solvents were obtained from Burdick & Jackson (Muskegon, MI), 99% formic acid was purchased from EM Science (Darmstadt, Germany).

6.2.3 Nano 2D LC-MS/MS Measurement:

Peptide mix (100 ug peptide) was pressure loaded onto a biphasic resin packed column [SCX (Luna, Phenomenex, Torrance, CA) and C18 (Aqua, Phenomenex, Torrance, CA)] as described earlier [191, 192]. The sample column was connected to the C18 packed nanospray tip (New

Objective, Woburn, MA) mounted on Proxeon (Odense, Denmark) nanospray source as described earlier [193]. Peptides were chromatographically sorted using the Ultimate 3000 HPLC system (Dionex, USA) over a course of 24h. The HPLC system was connected to the LTQ Velos mass spectrometer (Thermo Fisher Scientific,Germany), which was employed for peptide fragmentation and measurements via the Multi-Dimensional Protein Identification Technology (MuDPIT) approach as described earlier [191-193]. The peptide fragmentation and measurements was carried out in data dependent mode, using Thermo Xcalibur software V2.1.0. Each full scan (1 microscan) was followed by collision-activated dissociation (CID) based fragmentation using 35% collision energy of 10 most abundant parent ions (2 microscans) with a mass exclusion width of 0.2 m/z and dynamic exclusion duration of 60 s.

6.2.4 Bioinformatic Data Analysis:

For protein identifications, the raw spectra were searched against three databases of varying sizes via Myrimatch v2.1 [86] using parameters described previously [103]. The first was composed of 16 genomes identified in this study belonging to the *Desulfobacterales* and *Desulfuromonadales*; the second contained all predicted c-type cytochromes from these genomes containing four or more heme binding motifs (CxxCH amino acid sequences) and the open reading frames of *Desulfuromonadales* sp. C00003107; finally, the third database contained the core four proteins from the cytochrome operon from all of the SEEP-SRB1 and *Desulfuromonadales* genomes recovered in this study. Static cysteine and dynamic oxidation modifications were not included in the search parameters. Identification of at least two peptides per protein (one unique and one non-unique) sequence was set as a prerequisite for

protein identification. Common contaminant peptide sequences from trypsin and keratin were concatenated to the database along with reverse database sequences. The reverse database sequences were used as decoy sequences to calculate false discovery rate (FDR) which was maintained at < 1% for peptide to spectrum identification. For downstream data analysis, spectral counts of identified peptides was normalized as described before [94] to obtain the normalized spectral abundance factor (NSAF), also referred to as normalized spectral counts (nSpc). Average of nSpc from duplicate runs was used to obtain relative abundance values of expressed proteins across different samples. Normalization of spectra helps to account for differences in protein length, variations in MS analysis of samples thereby providing information on relative abundance of protein in a given samples and across samples in a given study.

6.3 Results and Discussion:

The expression of the SEEP-SRB operon was assessed using semi-quantitative metaproteomics *in situ* at methane seeps at the Santa Monica Mounds, Eel River Basin and Hydrate Ridge along the west coast of North America **(Figure 6.2)**. While absolute quantification of expressed proteins was beyond the scope of the current study, a semi quantitative mass spectrometry approach was taken to assess the relative abundance of expressed proteins and has been used in a number of previous protein expression analyses of complex samples [37, 194, 195]. The raw mass spectra were matched against a large concatenated predicted proteomic database containing sixteen genomes that belonged to *Desulfobacteraceae* SEEP-SRB1, *Desulfobulbaceae* SEEP-SRB4 and *Desulfuromondales* genomes recovered from the seep sediments. Although the

number of total identified proteins were modest (367 proteins across all three sites), among these were informative proteins that belonged to Wood Ljungdahl pathway (carbon monoxide dehydrogenase, formylmethanofuran-tetrahydromethanopterin N-formyltransferase), nitrogen fixation and sulfate reduction pathways (Sat, AprAB, DsrAB and DsrC). Expression of proteins belonging to Wood-Ljungdahl pathway (reductive acetyl-CoA pathway) was in line with SEEP-SRB1 genomic studies. This pathway is involved in carbon fixation and in agreement with previous observations from lipid biomarkers and ^{13}C-bicarbonate labeling studies [196]. Similarly, metaproteomic evidence for the expression of nitrogen fixation pathways is supported by genomic data, where previous ecophysiological studies of AOM (Anaerobic oxidation of Methane) consortia in methane seep sediments have demonstrated differences in nitrogen utilization among different ANME partners, including direct or indirect involvement in nitrogen fixation [197, 198] and nitrate utilization [199] suggesting that N_2 can be used as a biosynthetic nitrogen source. Finally, The SEEP-SRB1 genomes contain the genes required for sulfate reduction pathway, including sulfate adenylyl transferase (Sat), APS reductase (AprAB), dissimilatory sulfite reductase (DsrAB) and the sulfur carrier protein, DsrC (31) the expression of which was also confirmed by metaproteomics.

However, the identification of multiheme cytochrome proteins, which are present at low-moderate levels in these systems, was challenging. For complex systems, metaproteomics easily identifies the most abundant proteins, but has limited dynamic range and thus the lower abundance proteins often do not have adequate fragment ions or signal intensity to pass the standard threshold filters [200, 201]. In particular, the use of a large database to search for low abundant proteins adds to the difficulty in identifying less abundant proteins, since overall

peptide identification metrics (spectral matching, scoring criteria, and false discovery rates) are driven by the higher abundance peptides/proteins [202]. To better enable the search criteria for scouting lower abundance proteins, the database complexity was reduced by generating a smaller database of 2246 proteins derived from the sixteen genomes of SEEP-SRB1 and SEEP-SRB4 clades. This database comprised of proteins having four or more heme motifs and the full genome of *Desulfuromonadales* bacterium C00003107 (consisting of ~2000 proteins). Using this approach, we could detect several multi-heme cytochromes, all having less than 10 heme motifs **(Table 6.1).** These cytochromes might be members of a second operon, also widely distributed in SEEP-SRB1 that contain two cytochromes with 11 or 12 heme binding motifs and are related to the OmcX gene in *Geobacter* species [203].

For a more refined search, we further reduced the database complexity by assembling another database of only multiheme cytochromes, derived from the sixteen genomes of SEEP-SRB1 and SEEP-SRB4 clades. Protein expression was detected in all three sites from members of SEEP-SRB1a, SEEP-SRB1c and two members of the *Desulfomonadales* **(Table 6.2)**. Not all genes from each operon were detected at all sites, with Santa Monica and Eel River Basin having more matches than Hydrate Ridge. To more definitively support these limited database searches, peptide and protein identification was confirmed by manual validation of acquired peptide mass spectra from representatives of all the four members that constitute this operon **(Figure 6.3 A-F)**. These results show that the cytochromes are expressed *in situ* in syntrophic partnership with ANME archaea but at levels that are lower than enzymes from the sulfate reduction pathway. The reasons for this are unclear, it may be that extraction and detection were more difficult, cells may not require that many copies of the protein, or they may be

Table 6.1: Multiheme cytochromes containing less than 10 heme motifs searched with a database of 2246 proteins derived from the sixteen genomes of SEEP-SRB1 and SEEP-SRB4 clades.

This database comprised of proteins having four or more heme motifs and the full genome of Desulfuromonadales bacterium C00003107 consisting of ~2000 proteins.

Accession	Description	Organism	Number of CxxCH Motifs	Normalized Spectral Counts (nSpC)
OEU45386.1	hypothetical protein BBJ60_11710	*Desulfobacterales bacterium* S7086C20	9	110.2
OEU48827.1	cytochrome C	*Desulfobulbaceae bacterium* S5133MH15	8	140.7
OEU55250.1	cytochrome C	*Desulfobulbaceae bacterium* S3730MH12	8	140.7
OEU78971.1	hypothetical protein BA873_03805	*Desulfobulbaceae bacterium* C00003063	6	319.7
OEU81657.1	hypothetical protein BA865_04060	*Desulfobacterales bacterium* S5133MH4	5	296.7
OEU78922.1	hypothetical protein BA873_11840	*Desulfobulbaceae bacterium* C00003063	5	198.5
OEU56220.1	hypothetical protein BA862_03850	Desulfobulbaceae bacterium S3730MH12	5	198.5

Table 6.2: Multi-heme cytochromes detected in each replicate run across all three sites (Eel River Basin, Santa Monica Basin and Hydrate Ridge).

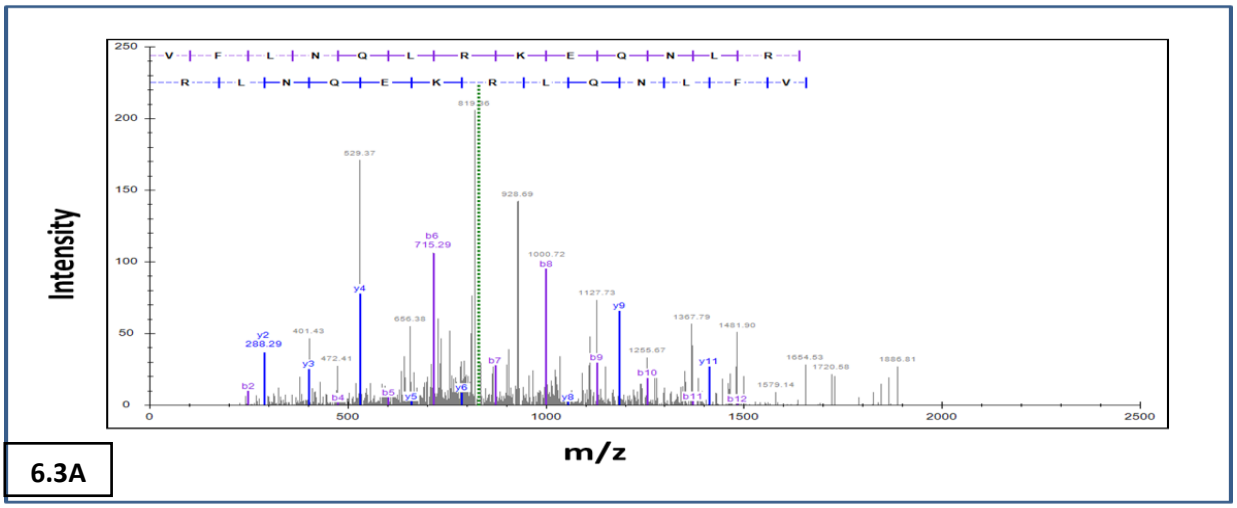
The raw mass spectra were searched with a refined database of only multiheme cytochromes derived from the sixteen genomes of SEEP-SRB1 and SEEP-SRB4 clades. Numbers in centi-meters (cm) represent the depth at which the samples were collected and used for proteomic characterization.

Sample Name	Site	Number of Multiheme cytochrome proteins
Run_1_0-6 cm	Hydrate Ridge	10
Run_2_0-6 cm	Hydrate Ridge	11
Run1_0-4 cm	Santa Monica	20
Run2_0-4 cm	Santa Monica	10
Run3_0-4 cm	Santa Monica	11
Run1_8-12 cm	Santa Monica	12
Run2_8-12 cm	Santa Monica	9
Run1_0-10 cm	Eel River	11
Run2_0-10 cm	Eel River	12
Run3_0-10 cm	Eel River	12
Run1_10-20 cm	Eel River	7
Run2_10-20 cm	Eel River	13

Figure 6.3 (A-F): Manual validation of acquired mass spectra

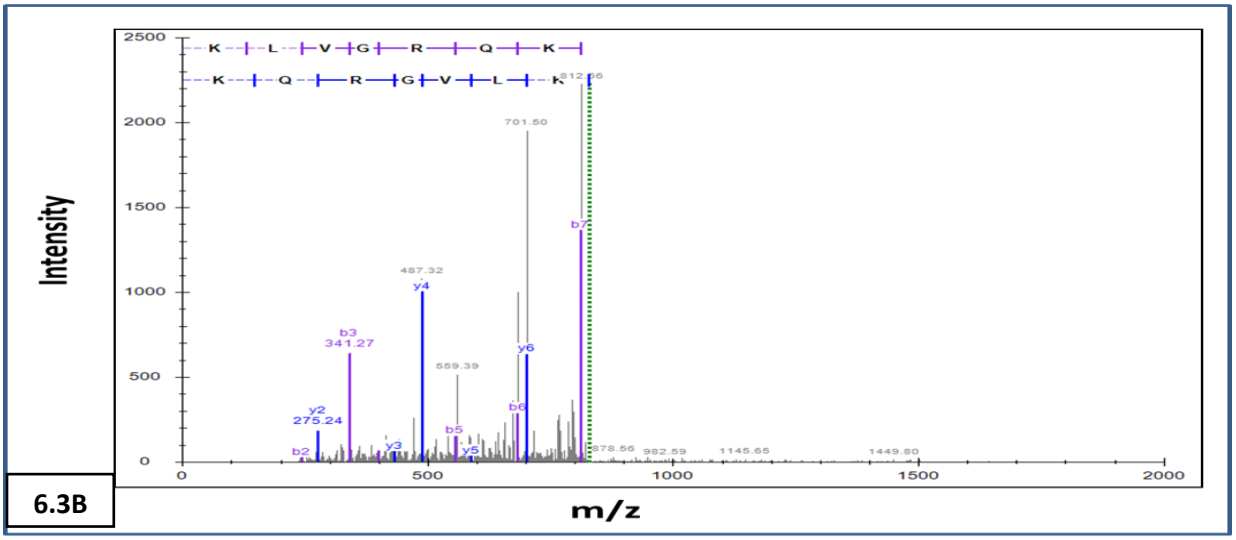
Figures A through F depict spectra of atleast one peptide of all the proteins belonging to the SEEP-SRB operon involved in extracellular electron transfer.

Peptidylprolyl isomerase



OEU55094.1 hypothetical protein BA871_14950 [Desulfuromonadales bacterium C00003096]
Spectrum from Hydrate Ridge 0-6 cm sediment sample

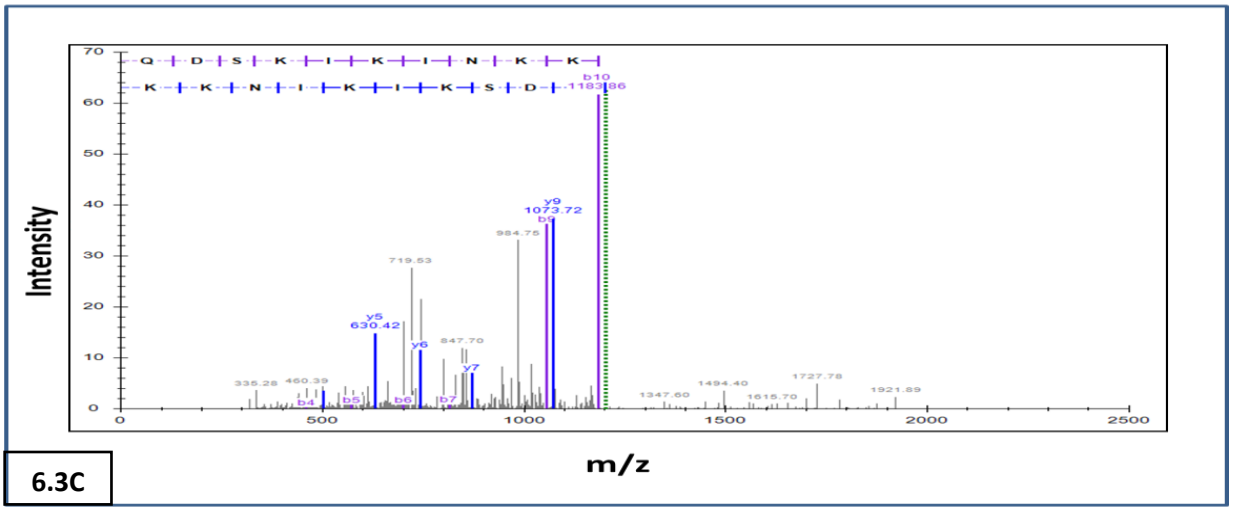
Peptidylprolyl isomerase



OEU44960.1 hypothetical protein BBJ60_04640 [Desulfobacterales bacterium S7086C20]
Spectrum from Santa Monica 0-4 cm sediment sample

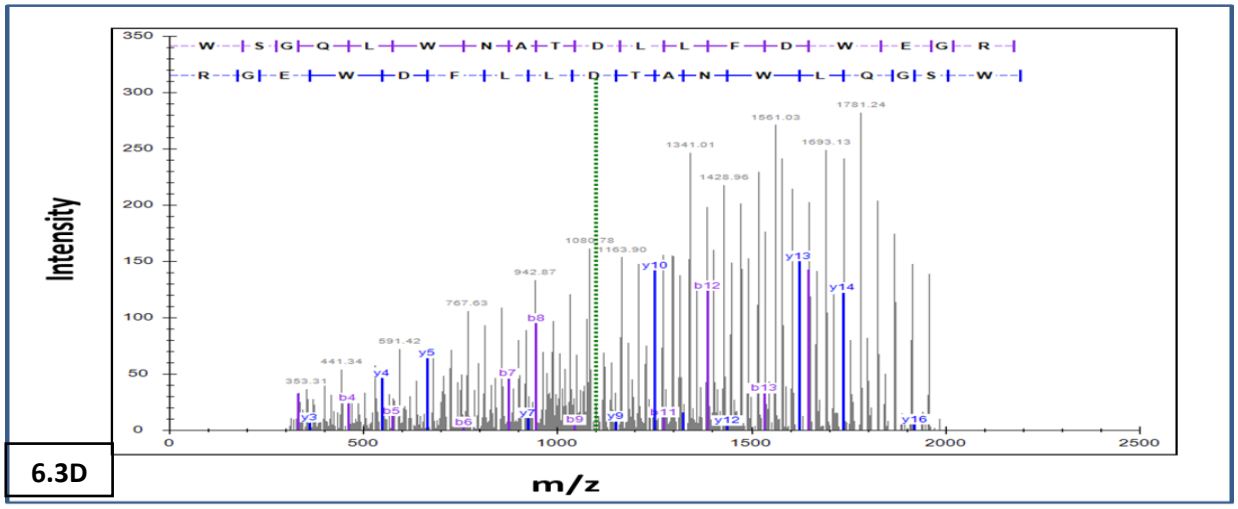
Figure 6.3 continued

Foldase protein



OEU80493.1 hypothetical protein BA872_08985 [Desulfobacterales bacterium C00003060]
Spectrum from Hydrate Ridge 0-6 cm sediment sample

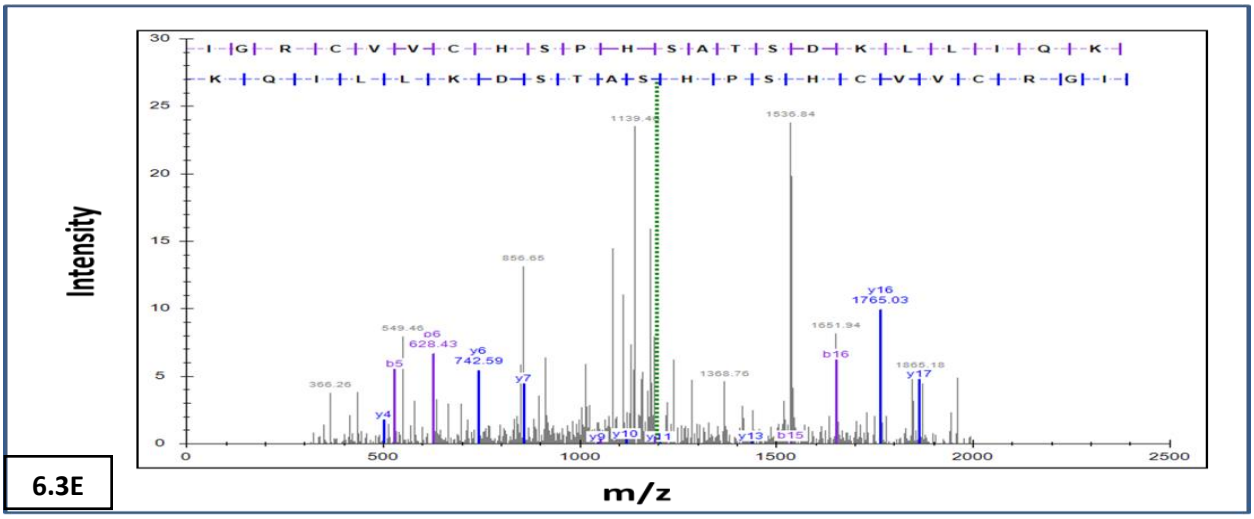
NHL repeat protein



OEU69591.1 hypothetical protein BA864_05830 [Desulfuromonadales bacterium C00003093
Spectrum from Santa Monica 0-4 cm sediment sample]

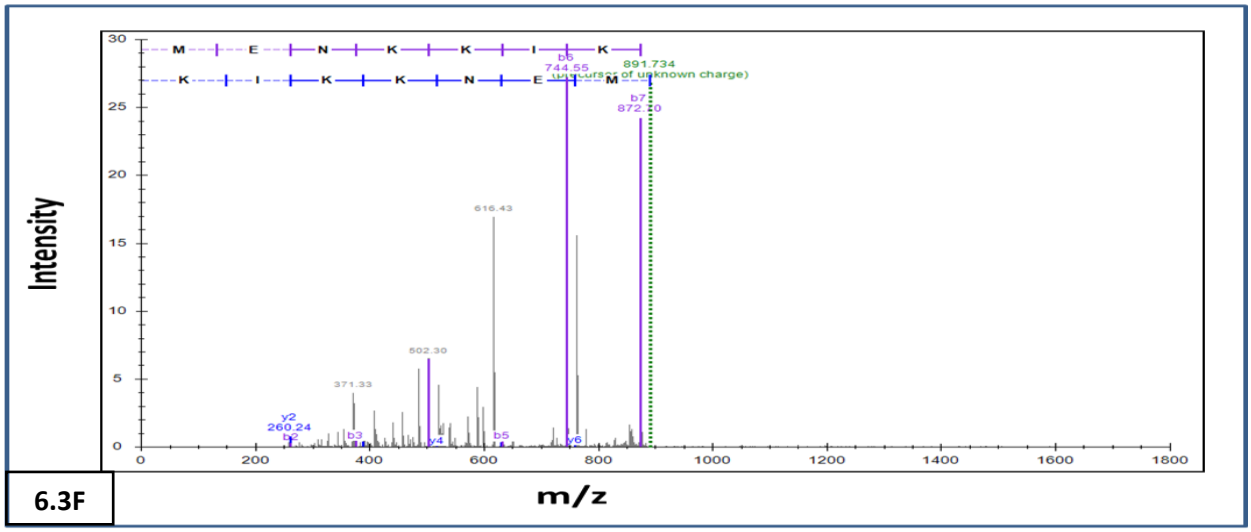
Figure 6.3 continued

26 heme cytochrome



OEU67620.1 hypothetical protein BBJ57_09125 [Desulfobacterales bacterium PC51MH44]
Spectrum from Hydrate Ridge 0-6 cm sediment sample

16 heme cytochrome



OEU53798.1 hypothetical protein BA868_10085 [Desulfobacterales bacterium C00003106]
Spectrum from Santa Monica 0-4 cm sediment sample

Figure 6.3 continued

involved in a separate physiology not related to the ANME syntrophy. However, their presence in SEEP-SRB1 and phylogenetically unrelated Ca. Desulfofervidus auxilii warrants further analysis as a possible component of the ANME-SRB syntrophy.

6.4 Conclusions:

Some archaea known as anaerobic methanotrophs possess the ability to convert methane to carbon dioxide when they are growing in partnership with the sulfate reducing bacteria. This is achieved through the process of extracellular electron transfer (EET) via multi-heme cytochromes to pass electrons generated during methane oxidation by the archaea to their bacterial counterparts. In the proposed study, metaproteomics was used to gauge the expression of the operon involved in EET across three different sites along the pacific coast.

Besides this, the study also highlights the bottleneck associated with protein identifications when searching the data with very large databases and how to circumvent that issue. While assembling peptides to proteins, each peptide is ranked depending on its score and only the top ranked peptides and its proteins are filtered and eventually displayed by the protein assembly software. In such cases, lower abundance proteins often do not pass the standard threshold filters as their identification is masked by the high abundant peptides/proteins present in the sample. The three-tiered database approach used in the current study where database complexity was sequentially reduced enabled us to identify several proteins involved in ANME-SRB syntrophy that might otherwise have been missed if only large database were taken into consideration. Finally, manual inspection of the peptide mass spectra helped us to further validate the peptides that belong to SEEP-SRB operon.

Chapter 7 - Conclusions, Current Trends and Future Perspectives

7.1 Conclusions from this Dissertation Work:

Proteomic technology has made remarkable strides in the past decade, and high throughput mass spectrometry has emerged as the most preferred method for detailed characterization of the protein components in the biological systems. These studies have proved the versatility of mass spectrometry based proteomics as an emerging and powerful method for solving complex biological questions. The dominance of mass spectrometry can be attributed to several factors. The instrument's incomparable ability to acquire quantitative as well as qualitative information about complex samples and the quest for continuous improvement in sensitivity, throughput and coverage has made mass spectrometry based proteomics a routine method for addressing a wide range of biological problems.

The broad goal of this dissertation was to provide an enhanced experimental and computational tool kit that can benefit researchers carrying out discovery proteomics on complex environmental samples that can also be translationally extended to other sample types. We have attempted to design an integrated pipeline that incorporates both experimental and bioinformatic components into one framework to address a broad range of questions in the field of high-throughput proteomics. We have also thought extensively about the computational bottlenecks associated with large scale discovery proteomic studies, which include peptide redundancy, concatenated metagenomic builds, *de novo* sequencing etc. and provided a mechanistic framework for addressing these challenges. Finally, we have presented

two 'Real World' scenarios where knowledge gained from experimental and computational strategies have proved fruitful in addressing specific biological problems.

In chapter 3 of the dissertation, we outlined some key aspects of peptide separation via two-dimensional chromatography and how it can be optimized to obtain deeper non-redundant peptide identifications that translates to deeper proteome coverage. This work highlights the significance of shallowing the initial windows of ammonium acetate concentration that resulted in improved spacing of peptides across the reverse phase resin which in turn improved the detection capability of mass spectrometer having a fixed scanning speed. We further demonstrated the utility of our approach by using two different samples of varying complexities and thus demonstrated the success of our separation strategy. In this chapter, we also evaluated the efficacy of this separation strategy for improving the throughput of complex proteome measurements. Our results showed that reducing the time duration for data acquisition using the newly designed chromatographic separation method does not alter the overall proteomic depth. This study will be handy to researchers who want to have a quick glance of the protein compliments present in their sample before embarking on a much-detailed experimental campaign.

In chapter 4, we focused on real world environmental samples and demonstrated the outcome of revised experimental procedure described in chapter 3. Here, we introduced the applicability of alternative computational platform (*de novo* sequencing) for interrogating high-throughput mass spec data and how it can be useful in addressing the complex landscape of proteomic characterization of environmental samples. Subsequently, we described an integrated computational platform consisting of multiple iterative search engines and *de novo* sequencing

algorithms and proved its worthiness on samples of varying complexities. In the process, we also reported a few other computational tools (Unipept, PepExplorer etc.) that can prove handy in providing reliable results during metaproteomic investigations. Lastly, we highlighted the significance of using stringent filters to reduce false positive identifications during data analysis.

In chapter 5, we highlighted the applicability of proteomic tools to gain a mechanistic understanding of contaminant degradation by the microbial community in groundwater samples. Our results also show a definite merit in using concatenated metagenomes in boosting the coverage and reliability of metaproteomic results. Besides this, we demonstrated the functionality of protein assembly software and explained how it impacts total peptide and protein identifications in shotgun proteomic experiments. Further, a definite advantage of using sequence tagging approach (ScanRanker) was examined and described to gauge the extent of raw spectra that was covered by the single and concatenated metagenomic builds.

In chapter 6, we transitioned to another complex environmental ecosystem for identification of multi-heme cytochromes and demonstrated the applicability of metaproteomics to determine the expression of an operon involved in extracellular electron transfer ultimately contributing to the process of reverse methanogenesis in oceanic sediments. We also highlighted another bottleneck associated with protein identification when searching the mass spectra with very large databases. To circumvent this issue, we described a three-tiered database search workflow followed by manual validation of select few mass spectra as an alternative approach for scouting very low abundant peptides. This method though cumbersome, provides a

mechanistic framework when looking for specific proteins of interest in complex biological specimens.

In summary, the work described in this dissertation is a comprehensive study on the experimental and computational aspects of proteomic workflows and will serve as a valuable resource to all those MS researchers who are inclined towards understanding the nuts and bolts of discovery proteomic assays.

7.2 Current Trends in Mass Spectrometry Based Proteomics:

With its humble origins in basic research, the field of mass spectrometry has rapidly expanded and is the forefront in several domains that includes biology, chemistry, physics, clinical medicine and even space exploration. This trend is expected to continue in the future with the availability of high throughput state of the art mass spectrometers that can cater the needs of both researchers and commercial outfits where portable mass spectrometers are increasingly deployed which are easy to operate and require very little technical expertise.

In the realm of biology, MS is being increasingly used in clinical laboratory settings and has wide range of applications from toxicological studies to personalized medicine. There is a growing belief that high throughput mass spectrometric immunoassays could one day replace ELISA (Enzyme linked immunosorbent assay). The ability of ELISA to detect specific antibodies with high degree of sensitivity and precision still makes it a standard choice for routine biochemical assays, but there are cases where clinically important proteins remain undetectable via ELISA. These include sequence variants and post-translational modifications [204]. Also, ELISA procedures often require extensive sample preparation to deplete the high abundant

protein/peptides. The ability of MS to detect protein isoforms even in complex biological matrices could prove worthy in such cases. Another distinct advantage of MS over ELISA is its ability to detect several analytes simultaneously with a high degree of sensitivity. The surge of MS based methods in clinical laboratories is expected to continue for bacterial identifications, imaging tissue sections (MALDI), diagnostic testing (for a panel of analytes that can include metabolites as well as peptide/proteins) and functional assays. Improvements in MS workflows are coupled with corresponding advancements in software architecture and information management. The development of laboratory information management systems (LIMS)[205] that features enhanced data tracking and exchange interfaces has made inter laboratory communications much more accessible. This has brought MS based detection into a larger and more diverse landscape where the transition from analyte quantification in biological samples to disease diagnosis and subsequent treatment has become much more streamlined.

In terms of separation and detection of complex samples, two-dimensional chromatography coupled with tandem mass spectrometry has made tremendous progress since its inception in 2001. The overall acquisition time has also gone down with the introduction of next generation mass spectrometers like Q-Exactive and Orbitrap fusion. This is evident from a recent publication where the entire yeast proteome was measured in an hour having the same proteome coverage as the 24 hr MudPIT procedure [102]. However, analysis of complex proteomes in such a short duration may be subject to random sampling events which may result in some irreproducibility in the peptide measurements. Also, the spectra collected in this shortened run enhance qualitative identifications but may confound quantitative evaluations for samples with increased complexity. Additionally, the newer instrument platforms are more

expensive than previous generation mass spectrometers. In labs where there is scarcity of funds, replacing old instruments with newer state of the art mass spectrometers may not be always feasible. Hence there continues to be growing trend in coupling a better separation strategy to MS for obtaining deeper proteome coverage especially for complex samples. These include hydrophilic interaction liquid chromatography (HILIC), perfluorinated reversed phase chromatography, RP-RP with high pH and low pH elution, and mixed bead ion exchange chromatography etc. [97-100] to name a few.

Another growing trend in MS based proteomics is the concept of data sharing. This is becoming a common scientific practice which is in line with other 'omic' fields like transcriptomics and genomics. This crucial switch has been mainly triggered by open access journals that require data to be made publicly available so that researchers across the globe can get access to it. Additionally, the advent of user friendly resources and tools to support simpler data dissemination so that experimental researchers can easily upload their data to central repositories. Several noteworthy repositories have been developed which include MassIVE (http://massive.ucsd.edu/), jPOST (http://jpost.org/), the Human Proteome Map (http://www.humanproteomemap.org/) , ProteomicsDB (https://www.proteomicsdb.org/) and Chorus (https://chorusproject.org/). Also, the ProteomeXchange (PX) Consortium [206] supports many mass spectrometry based data analysis tools that are readily available for download. This consortium undergoes periodic upgrades and new bioinformatic tools are routinely made available. With the development of centralized repositories for data storage the growing field of MS is now filled with opportunities for those wanting to use this knowledge for making novel discoveries from pre-existing data. One of the main caveats in discovery

proteomic studies is with regards to data analysis and extracting relevant biological information from these complex measurements. Often, a researcher is looking for the expression of a subset of proteins in the dataset. In such cases, a wealth of the remaining data stays unanalyzed. By sharing the data in repositories, others can analyze them and thus further optimize their scientific value.

The growing trends in mass spectrometry based proteomics described above that ranges from experimental improvements to bioinformatic tools, places MS in the forefront of biological research. However, it also poses an additional responsibility on the researches working in this domain. Since the debacle of SELDI mass spectrometry in biomarker discovery of ovarian cancer [207], new techniques need to be scrutinized thoroughly before being brought forth for commercial usage. This puts further onus on MS researchers like us to work more diligently at the fundamental level so that new progress whether experimental or computational, is backed by sufficient replicable data so that such disasters don't happen in the future.

7.3 Future Outlook, Perspectives and Demands:

Given the enormous complexity of the metaproteome and the wide dynamic range of protein compliments in microbial communities, metaproteomic characterization faces a multitude of challenges at different stages of the analytical workflow. These include enhanced cell lysis and protein extraction protocols that are applicable to most complex sample matrices like soils, mass spectrometers having high mass accuracies and faster scanning speeds, high-throughput computing clusters capable of detecting a wide variety of post-translational modifications in

these complex specimens and the development of dedicated software that can integrate the metaproteomics data with other 'omic' studies for better visualization.

Currently, a lot of laboratories use databases downloaded from public repositories (like NCBI and Uniprot) for data analysis. We expect this trend to change in the future with the availability of cheaper sequencing technology. The construction of high quality metagenomic and metatranscriptomic databases from the same samples (example [208-210]) will allow the construction of customized and sample specific metaproteomic databases which are expected to markedly improve the reliability and scope of protein identifications. Further improvements in data validation would involve FDR estimations. Since metaproteomic investigations involve many proteins having hypothetical and unannotated sequences, the FDR is not always correct. In such cases semi-supervised machine learning algorithms can be leveraged [211]. Metaproteomics using database searching is currently restricted to identification of 5-30% of spectra. A recent study estimated that about 30% of spectra belong to solvent and background components, [212] which leaves an additional 40% spectra that remain unidentified. Spectral libraries that can store and cluster spectra can be used as an alternative strategy to handle these spectra. *De novo* sequencing which has been extensively discussed in this dissertation could also be put into use to rescue myriads of unidentified spectra. Further refinement in *de novo* sequencing can be achieved by using alternative proteases such as Lys-C or Arg-C which gives rise to longer peptides and thus more robust identifications. As already highlighted before, we expect a surge in holistic data integration consisting of metagenomics, metaproteomics and metatranscriptomic studies. Given the high-dimensionality of these

datasets, improved data integration and analysis tools will be designed that can give a better understanding of system functionality and perturbations.

Most of the challenges outlined here and corresponding solutions proposed are already under implementation, hence we expect the landscape of metaproteomics to flourish and become a major focus of systems biology in the coming decade.

In conclusion, the field of metaproteomics has opened up new avenues for detailed characterization of microbes that exist in natural ecosystems. However, the field is still at a nascent stage and is restricted to few laboratories having access to high throughput instrumentation and computing clusters. We expect the field to gain further traction with reduced instrumentation costs similar to DNA sequencing technology and its active promotion by the research community as a superb and complimentary technique to other meta-omic methods especially in context of integrated omic analysis of microbial consortia. For this, more active collaborations need to occur between industries and academic institutions coupled with active workshops that can attract graduate student with interests in this exciting field of biology.

References

1. De Laeter, J.R., *Applications of inorganic mass spectrometry*. Vol. 3. 2001: John Wiley & Sons.

2. Chace, D.H., T.A. Kalas, and E.W. Naylor, *The application of tandem mass spectrometry to neonatal screening for inherited disorders of intermediary metabolism.* Annu Rev Genomics Hum Genet, 2002. **3**: p. 17-45.

3. Gygi, S.P., et al., *Proteome analysis of low-abundance proteins using multidimensional chromatography and isotope-coded affinity tags.* J Proteome Res, 2002. **1**(1): p. 47-54.

4. Patterson, K.Y. and C. Veillon, *Stable isotopes of minerals as metabolic tracers in human nutrition research.* Exp Biol Med (Maywood), 2001. **226**(4): p. 271-82.

5. Mano, N. and J. Goto, *Biomedical and biological mass spectrometry.* Anal Sci, 2003. **19**(1): p. 3-14.

6. Wilmes, P. and P.L. Bond, *Metaproteomics: studying functional gene expression in microbial ecosystems.* Trends Microbiol, 2006. **14**(2): p. 92-7.

7. Fenn, J.B., *Electrospray ionization mass spectrometry: How it all began.* J Biomol Tech, 2002. **13**(3): p. 101-18.

8. Fenn, J.B., et al., *Electrospray ionization for mass spectrometry of large biomolecules.* Science, 1989. **246**(4926): p. 64-71.

9. Guo, B., *Mass Spectrometry in DNA Analysis.* Analytical Chemistry, 1999. **71**(12): p. 333-337.

10. Silva, A.M.N., et al., *Post-translational Modifications and Mass Spectrometry Detection.* Free Radical Biology and Medicine, 2013. **65**: p. 925-941.

11. Pitt, J.J., *Principles and applications of liquid chromatography-mass spectrometry in clinical biochemistry.* Clin Biochem Rev, 2009. **30**(1): p. 19-34.

12. Chakravarti, B., B. Mallik, and D.N. Chakravarti, *Proteomics and systems biology: application in drug discovery and development.* Systems Biology in Drug Discovery and Development: Methods and Protocols, 2010: p. 3-28.

13. de Godoy, L.M., et al., *Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast.* Nature, 2008. **455**(7217): p. 1251-4.

14. Kelleher, N.L., et al., *Top Down versus Bottom Up Protein Characterization by Tandem High-Resolution Mass Spectrometry.* Journal of the American Chemical Society, 1999. **121**(4): p. 806-812.

15. Moore, J.B. and M.E. Weeks, *Proteomics and systems biology: current and future applications in the nutritional sciences.* Adv Nutr, 2011. **2**(4): p. 355-64.

16. Cho, C.R., et al., *The application of systems biology to drug discovery.* Curr Opin Chem Biol, 2006. **10**(4): p. 294-302.

17. Villoslada, P., L. Steinman, and S.E. Baranzini, *Systems biology and its application to the understanding of neurological diseases.* Ann Neurol, 2009. **65**(2): p. 124-39.

18. Baker, B.J. and J.F. Banfield, *Microbial communities in acid mine drainage.* FEMS Microbiol Ecol, 2003. **44**(2): p. 139-52.

19. Singer, S.W., et al., *Characterization of cytochrome 579, an unusual cytochrome isolated from an iron-oxidizing microbial community.* Appl Environ Microbiol, 2008. **74**(14): p. 4454-62.

20. Ram, R.J., et al., *Community proteomics of a natural microbial biofilm.* Science, 2005. **308**(5730): p. 1915-20.

21. Wilmes, P., et al., *Community proteogenomics highlights microbial strain-variant protein expression within activated sludge performing enhanced biological phosphorus removal.* ISME J, 2008. **2**(8): p. 853-64.

22. Mahowald, M.A., et al., *Characterizing a model human gut microbiota composed of members of its two dominant bacterial phyla.* Proc Natl Acad Sci U S A, 2009. **106**(14): p. 5859-64.

23. Becher, D., et al., *Metaproteomics to unravel major microbial players in leaf litter and soil environments: challenges and perspectives.* Proteomics, 2013. **13**(18-19): p. 2895-909.

24. Skennerton, C.T., et al., *Methane-Fueled Syntrophy through Extracellular Electron Transfer: Uncovering the Genomic Traits Conserved within Diverse Bacterial Partners of Anaerobic Methanotrophic Archaea.* MBio, 2017. **8**(4).

25. Lu, F., et al., *Metaproteomics of cellulose methanisation under thermophilic conditions reveals a surprisingly high proteolytic activity.* ISME J, 2014. **8**(1): p. 88-102.

26. Nuriel-Ohayon, M., H. Neuman, and O. Koren, *Microbial Changes during Pregnancy, Birth, and Infancy.* Front Microbiol, 2016. **7**: p. 1031.

27. Demirev, P. and T.R. Sandrin, *Applications of Mass Spectrometry in Microbiology*. 2016: Springer.

28. Warscheid, B., et al., *MALDI analysis of Bacilli in spore mixtures by applying a quadrupole ion trap time-of-flight tandem mass spectrometer.* Anal Chem, 2003. **75**(20): p. 5608-17.

29. Demirev, P.A., et al., *Top-down proteomics for rapid identification of intact microorganisms.* Anal Chem, 2005. **77**(22): p. 7455-61.

30. Jones, J.J., et al., *Strategies and data analysis techniques for lipid and phospholipid chemistry elucidation by intact cell MALDI-FTMS.* J Am Soc Mass Spectrom, 2004. **15**(11): p. 1665-74.

31. Russell, S.C., *Microorganism characterization by single particle mass spectrometry.* Mass Spectrom Rev, 2009. **28**(2): p. 376-87.

32. Wynne, C., et al., *Top-down identification of protein biomarkers in bacteria with unsequenced genomes.* Anal Chem, 2009. **81**(23): p. 9633-42.

33. Maron, P.A., et al., *Metaproteomics: a new approach for studying functional microbial ecology.* Microb Ecol, 2007. **53**(3): p. 486-93.

34. Gupta, N., et al., *Comparative proteogenomics: combining mass spectrometry and comparative genomics to analyze multiple genomes.* Genome Res, 2008. **18**(7): p. 1133-42.

35. Armengaud, J., E.M. Hartmann, and C. Bland, *Proteogenomics for environmental microbiology.* Proteomics, 2013. **13**(18-19): p. 2731-42.

36. Venter, J.C., et al., *Environmental genome shotgun sequencing of the Sargasso Sea.* Science, 2004. **304**(5667): p. 66-74.

37. Hettich, R.L., et al., *Metaproteomics: harnessing the power of high performance mass spectrometry to identify the suite of proteins that control metabolic activities in microbial communities.* Anal Chem, 2013. **85**(9): p. 4203-14.

38. Vaudel, M., et al., *PeptideShaker enables reanalysis of MS-derived proteomics data sets.* Nat Biotechnol, 2015. **33**(1): p. 22-4.

39. Muth, T., et al., *The MetaProteomeAnalyzer: a powerful open-source software suite for metaproteomics data analysis and interpretation.* J Proteome Res, 2015. **14**(3): p. 1557-65.

40. Michalski, A., et al., *A Systematic Investigation into the Nature of Tryptic HCD Spectra.* Journal of Proteome Research, 2012. **11**(11): p. 5479-5491.

41. Vaudel, M., et al., *Peptide identification quality control.* Proteomics, 2011. **11**(10): p. 2105-14.

42. Muth, T., et al., *Navigating through metaproteomics data: a logbook of database searching.* Proteomics, 2015. **15**(20): p. 3439-53.

43. Wright, J.C., R.J. Beynon, and S.J. Hubbard, *Cross species proteomics.* Methods Mol Biol, 2010. **604**: p. 123-35.

44. Jagtap, P., et al., *A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies.* Proteomics, 2013. **13**(8): p. 1352-7.

45. Kuhring, M. and B.Y. Renard, *Estimating the computational limits of detection of microbial non-model organisms.* Proteomics, 2015. **15**(20): p. 3580-4.

46. Tanca, A., et al., *Evaluating the impact of different sequence databases on metaproteome analysis: insights from a lab-assembled microbial mixture.* PLoS One, 2013. **8**(12): p. e82981.

47. Nesvizhskii, A.I. and R. Aebersold, *Interpretation of shotgun proteomic data: the protein inference problem.* Mol Cell Proteomics, 2005. **4**(10): p. 1419-40.

48. Kolmeder, C.A. and W.M. de Vos, *Metaproteomics of our microbiome - developing insight in function and activity in man and model systems.* J Proteomics, 2014. **97**: p. 3-16.

49. Allmer, J., *Algorithms for the de novo sequencing of peptides from tandem mass spectra.* Expert Rev Proteomics, 2011. **8**(5): p. 645-57.

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

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

52. Eng, J.K., A.L. McCormack, and J.R. Yates, *An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database.* J Am Soc Mass Spectrom, 1994. **5**(11): p. 976-89.

53. Tabb, D.L., J.K. Eng, and J.R. Yates, *Protein Identification by SEQUEST*, in *Proteome Research: Mass Spectrometry*. 2001, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 125-142.

54. Andersen, K.K., et al., *The role of decorated SDS micelles in sub-CMC protein denaturation and association.* J Mol Biol, 2009. **391**(1): p. 207-26.

55. Link, A.J. and J. LaBaer, *Trichloroacetic acid (TCA) precipitation of proteins.* Cold Spring Harb Protoc, 2011. **2011**(8): p. 993-4.

56. Bennion, B.J. and V. Daggett, *The molecular basis for the chemical denaturation of proteins by urea.* Proc Natl Acad Sci U S A, 2003. **100**(9): p. 5142-7.

57. Switzar, L., M. Giera, and W.M. Niessen, *Protein digestion: an overview of the available techniques and recent developments.* J Proteome Res, 2013. **12**(3): p. 1067-77.

58. Huynh, M.L., P. Russell, and B. Walsh, *Tryptic digestion of in-gel proteins for mass spectrometry analysis.* Methods Mol Biol, 2009. **519**: p. 507-13.

59. Smith, P.K., et al., *Measurement of protein using bicinchoninic acid.* Anal Biochem, 1985. **150**(1): p. 76-85.

60. Link, A.J. and M.P. Washburn, *Analysis of protein composition using multidimensional chromatography and mass spectrometry.* Curr Protoc Protein Sci, 2014. **78**: p. 23 1 1-25.

61. Fournier, M.L., et al., *Multidimensional separations-based shotgun proteomics.* Chem Rev, 2007. **107**(8): p. 3654-86.

62. Washburn, M.P., et al., *Analysis of quantitative proteomic data generated via multidimensional protein identification technology.* Anal Chem, 2002. **74**(7): p. 1650-7.

63. Zhang, X., et al., *Multi-dimensional liquid chromatography in proteomics--a review.* Anal Chim Acta, 2010. **664**(2): p. 101-13.

64. Tanaka, K., et al., *Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry.* Rapid Communications in Mass Spectrometry, 1988. **2**(8): p. 151-153.

65. Taylor, G., *Disintegration of Water Drops in an Electric Field.* Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences, 1964. **280**(1382): p. 383-397.

66. Smith, J.N., R.C. Flagan, and J.L. Beauchamp, *Droplet Evaporation and Discharge Dynamics in Electrospray Ionization.* The Journal of Physical Chemistry A, 2002. **106**(42): p. 9957-9967.

67. Sattler, K., et al., *Evidence for Coulomb Explosion of Doubly Charged Microclusters.* Physical Review Letters, 1981. **47**(3): p. 160-163.

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

69. Second, T.P., et al., *Dual-pressure linear ion trap mass spectrometer improving the analysis of complex protein mixtures.* Anal Chem, 2009. **81**(18): p. 7757-65.

70. Hu, Q., et al., *The Orbitrap: a new mass spectrometer.* Journal of Mass Spectrometry, 2005. **40**(4): p. 430-443.

71. Senko, M.W., et al., *A high-performance modular data system for Fourier transform ion cyclotron resonance mass spectrometry.* Rapid Commun Mass Spectrom, 1996. **10**(14): p. 1839-44.

72. Nibbering, N.M.M., *Fourier transforms in NMR, optical and mass spectrometry. A user's handbook. Alan G. Marshall and Francis R. Verdun Elsevier, Amsterdam (1990). xvi + 450*

pp, hardback $107.25; paperback $46.25. Rapid Communications in Mass Spectrometry, 1990. **4**(10): p. 462-462.

73. Michalski, A., et al., *Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer.* Mol Cell Proteomics, 2011. **10**(9): p. M111 011015.

74. Liu, A., et al., *Identification of two novel brominated contaminants in water samples by ultra-high performance liquid chromatography-Orbitrap Fusion Tribrid mass spectrometer.* J Chromatogr A, 2015. **1377**: p. 92-9.

75. Cooks, R.G., *Special feature: Historical. Collision-induced dissociation: Readings and commentary.* Journal of Mass Spectrometry, 1995. **30**(9): p. 1215-1221.

76. Crowe, M.C. and J.S. Brodbelt, *Infrared multiphoton dissociation (IRMPD) and collisionally activated dissociation of peptides in a quadrupole ion trap with selective IRMPD of phosphopeptides.* J Am Soc Mass Spectrom, 2004. **15**(11): p. 1581-92.

77. Kaczorowska, M.A. and H.J. Cooper, *Electron induced dissociation (EID) tandem mass spectrometry of octaethylporphyrin and its iron(III) complex.* Chem Commun (Camb), 2011. **47**(1): p. 418-20.

78. Zhang, Y., et al., *Effect of dynamic exclusion duration on spectral count based quantitative proteomics.* Anal Chem, 2009. **81**(15): p. 6317-26.

79. Sadygov, R.G., D. Cociorva, and J.R. Yates, 3rd, *Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book.* Nat Methods, 2004. **1**(3): p. 195-202.

80. Perkins, D.N., et al., *Probability-based protein identification by searching sequence databases using mass spectrometry data.* Electrophoresis, 1999. **20**(18): p. 3551-67.

81. Craig, R. and R.C. Beavis, *TANDEM: matching proteins with tandem mass spectra.* Bioinformatics, 2004. **20**(9): p. 1466-1467.

82. Geer, L.Y., et al., *Open mass spectrometry search algorithm.* J Proteome Res, 2004. **3**(5): p. 958-64.

83. Paizs, B. and S. Suhai, *Fragmentation pathways of protonated peptides.* Mass Spectrometry Reviews, 2005. **24**(4): p. 508-548.

84. Fenyo, D. and R.C. Beavis, *A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes.* Anal Chem, 2003. **75**(4): p. 768-74.

85. Kim, S. and P. Pevzner. *MS-GF+: universal database search tool for mass spectrometry*. in *8th Annual US HUPO Conference (USHUPO), San Francisco, CA*. 2012.

86. Tabb, D.L., C.G. Fernando, and M.C. Chambers, *MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis.* J Proteome Res, 2007. **6**(2): p. 654-61.

87. Eng, J.K., T.A. Jahan, and M.R. Hoopmann, *Comet: an open-source MS/MS sequence database search tool.* Proteomics, 2013. **13**(1): p. 22-4.

88. Cox, J., et al., *Andromeda: a peptide search engine integrated into the MaxQuant environment.* J Proteome Res, 2011. **10**(4): p. 1794-805.

89. Elias, J.E. and S.P. Gygi, *Target-decoy search strategy for mass spectrometry-based proteomics.* Methods Mol Biol, 2010. **604**: p. 55-71.

90. Aggarwal, S. and A.K. Yadav, *False Discovery Rate Estimation in Proteomics.* Methods Mol Biol, 2016. **1362**: p. 119-28.

91. Tabb, D.L., W.H. McDonald, and J.R. Yates, 3rd, *DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics.* J Proteome Res, 2002. **1**(1): p. 21-6.

92. Ma, Z.Q., et al., *IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering.* J Proteome Res, 2009. **8**(8): p. 3872-81.

93. Haqqani, A.S., J.F. Kelly, and D.B. Stanimirovic, *Quantitative protein profiling by mass spectrometry using label-free proteomics.* Methods Mol Biol, 2008. **439**: p. 241-56.

94. Paoletti, A.C., et al., *Quantitative proteomic analysis of distinct mammalian Mediator complexes using normalized spectral abundance factors.* Proc Natl Acad Sci U S A, 2006. **103**(50): p. 18928-33.

95. Ishihama, Y., et al., *Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein.* Mol Cell Proteomics, 2005. **4**(9): p. 1265-72.

96. Griffin, N.M., et al., *Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis.* Nat Biotechnol, 2010. **28**(1): p. 83-9.

97. Buszewski, B. and S. Noga, *Hydrophilic interaction liquid chromatography (HILIC)--a powerful separation technique.* Anal Bioanal Chem, 2012. **402**(1): p. 231-47.

98. Flieger, J., *Application of perfluorinated acids as ion-pairing reagents for reversed-phase chromatography and retention-hydrophobicity relationships studies of selected beta-blockers.* J Chromatogr A, 2010. **1217**(4): p. 540-9.

99. Yang, F., et al., *High-pH reversed-phase chromatography with fraction concatenation for 2D proteomic analysis.* Expert Rev Proteomics, 2012. **9**(2): p. 129-34.

100. Mommen, G.P., et al., *Mixed-bed ion exchange chromatography employing a salt-free pH gradient for improved sensitivity and compatibility in MudPIT.* Anal Chem, 2013. **85**(14): p. 6608-16.

101. Motoyama, A. and J.R. Yates, 3rd, *Multidimensional LC separations in shotgun proteomics.* Anal Chem, 2008. **80**(19): p. 7187-93.

102. Hebert, A.S., et al., *The one hour yeast proteome.* Mol Cell Proteomics, 2014. **13**(1): p. 339-47.

103. Xiong, W., et al., *Development of an enhanced metaproteomic approach for deepening the microbiome characterization of the human infant gut.* J Proteome Res, 2015. **14**(1): p. 133-41.

104. Verberkmoes, N.C., et al., *Shotgun metaproteomics of the human distal gut microbiota.* ISME J, 2009. **3**(2): p. 179-89.

105. Zhou, H., et al., *A fully automated 2-D LC-MS method utilizing online continuous pH and RP gradients for global proteome analysis.* Electrophoresis, 2007. **28**(23): p. 4311-9.

106. Winnik, W.M., *Continuous pH/salt gradient and peptide score for strong cation exchange chromatography in 2D-nano-LC/MS/MS peptide identification for proteomics.* Anal Chem, 2005. **77**(15): p. 4991-8.

107. Link, A.J., et al., *Direct analysis of protein complexes using mass spectrometry.* Nat Biotechnol, 1999. **17**(7): p. 676-82.

108. Washburn, M.P., *Coupled Multidimensional Chromatography and Tandem Mass Spectrometry Systems for Complex Peptide Mixture Analysis.* 2008: p. 243-259.

109. Florens, L. and M.P. Washburn, *Proteomic analysis by multidimensional protein identification technology.* Methods Mol Biol, 2006. **328**: p. 159-75.

110. Zhou, F., et al., *Online nanoflow RP-RP-MS reveals dynamics of multicomponent Ku complex in response to DNA damage.* J Proteome Res, 2010. **9**(12): p. 6242-55.

111. Webb, K.J., et al., *Modified MuDPIT separation identified 4488 proteins in a system-wide analysis of quiescence in yeast.* J Proteome Res, 2013. **12**(5): p. 2177-84.

112. Lochner, A., et al., *Label-free quantitative proteomics for the extremely thermophilic bacterium Caldicellulosiruptor obsidiansis reveal distinct abundance patterns upon growth on cellobiose, crystalline cellulose, and switchgrass.* J Proteome Res, 2011. **10**(12): p. 5302-14.

113. Neilson, K.A., et al., *Label-free quantitative shotgun proteomics using normalized spectral abundance factors.* Methods Mol Biol, 2013. **1002**: p. 205-22.

114. Muth, T., et al., *DeNovoGUI: an open source graphical user interface for de novo sequencing of tandem mass spectra.* J Proteome Res, 2014. **13**(2): p. 1143-6.

115. Frank, A. and P. Pevzner, *PepNovo: de novo peptide sequencing via probabilistic network modeling.* Anal Chem, 2005. **77**(4): p. 964-73.

116. Fonslow, B.R., et al., *Digestion and depletion of abundant proteins improves proteomic coverage.* Nat Methods, 2013. **10**(1): p. 54-6.

117. Farias, S.E., et al., *Quantitative improvements in peptide recovery at elevated chromatographic temperatures from microcapillary liquid chromatography-mass*

spectrometry analyses of brain using selected reaction monitoring. Anal Chem, 2010. **82**(9): p. 3435-40.

118. Bagag, A., et al., *Characterization of hydrophobic peptides in the presence of detergent by photoionization mass spectrometry.* PLoS One, 2013. **8**(11): p. e79033.

119. Kantor, R.S., et al., *Bioreactor microbial ecosystems for thiocyanate and cyanide degradation unravelled with genome-resolved metagenomics.* Environ Microbiol, 2015.

120. Kantor, R.S., et al., *Genome-Resolved Meta-Omics Ties Microbial Dynamics to Process Performance in Biotechnology for Thiocyanate Degradation.* Environ Sci Technol, 2017.

121. du Plessis, C.A., et al., *Empirical model for the autotrophic biodegradation of thiocyanate in an activated sludge reactor.* Lett Appl Microbiol, 2001. **32**(2): p. 103-7.

122. Felfoldi, T., et al., *Polyphasic bacterial community analysis of an aerobic activated sludge removing phenols and thiocyanate from coke plant effluent.* Bioresour Technol, 2010. **101**(10): p. 3406-14.

123. Huddy, R.J., et al., *Characterisation of the complex microbial community associated with the ASTER (TM) thiocyanate biodegradation system.* Minerals Engineering, 2015. **76**: p. 65-71.

124. Quan, Z.X., et al., *Bacterial community structure in activated sludge reactors treating free or metal-complexed cyanides.* Journal of Microbiology and Biotechnology, 2006. **16**(2): p. 232-239.

125. Kalli, A., et al., *Evaluation and optimization of mass spectrometric settings during data-dependent acquisition mode: focus on LTQ-Orbitrap mass analyzers.* J Proteome Res, 2013. **12**(7): p. 3071-86.

126. Muth, T., et al., *Searching for a needle in a stack of needles: challenges in metaproteomics data analysis.* Mol Biosyst, 2013. **9**(4): p. 578-85.

127. Muth, T., B.Y. Renard, and L. Martens, *Metaproteomic data analysis at a glance: advances in computational microbial community proteomics.* Expert Rev Proteomics, 2016. **13**(8): p. 757-69.

128. Benndorf, D., et al., *Functional metaproteome analysis of protein extracts from contaminated soil and groundwater.* ISME J, 2007. **1**(3): p. 224-34.

129. Podar, M., et al., *Targeted access to the genomes of low-abundance organisms in complex microbial communities.* Appl Environ Microbiol, 2007. **73**(10): p. 3205-14.

130. Bell, A.W., et al., *A HUPO test sample study reveals common problems in mass spectrometry-based proteomics.* Nat Methods, 2009. **6**(6): p. 423-30.

131. Kapp, E.A., et al., *An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis.* Proteomics, 2005. **5**(13): p. 3475-90.

132. Keller, A., et al., *Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search.* Anal Chem, 2002. **74**(20): p. 5383-92.

133. Brosch, M., et al., *Accurate and sensitive peptide identification with Mascot Percolator.* J Proteome Res, 2009. **8**(6): p. 3176-81.

134. Zhang, J., et al., *PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification.* Mol Cell Proteomics, 2012. **11**(4): p. M111 010587.

135. Cottrell, J.S., *Protein identification using MS/MS data.* Journal of Proteomics, 2011. **74**(10): p. 1842-1851.

136. Searle, B.C., M. Turner, and A.I. Nesvizhskii, *Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies.* J Proteome Res, 2008. **7**(1): p. 245-53.

137. Alves, G., et al., *Enhancing peptide identification confidence by combining search methods.* J Proteome Res, 2008. **7**(8): p. 3102-13.

138. Shteynberg, D., et al., *Combining results of multiple search engines in proteomics.* Mol Cell Proteomics, 2013. **12**(9): p. 2383-93.

139. Park, G.W., et al., *Integrated Proteomic Pipeline Using Multiple Search Engines for a Proteogenomic Study with a Controlled Protein False Discovery Rate.* J Proteome Res, 2016. **15**(11): p. 4082-4090.

140. Cantarel, B.L., et al., *Strategies for metagenomic-guided whole-community proteomics of complex microbial environments.* PLoS One, 2011. **6**(11): p. e27173.

141. Leprevost, F.V., et al., *PepExplorer: a similarity-driven tool for analyzing de novo sequencing results.* Mol Cell Proteomics, 2014. **13**(9): p. 2480-9.

142. Lo, I., et al., *Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria.* Nature, 2007. **446**(7135): p. 537-41.

143. Craig, R. and R.C. Beavis, *TANDEM: matching proteins with tandem mass spectra.* Bioinformatics, 2004. **20**(9): p. 1466-7.

144. Kim, S. and P.A. Pevzner, *MS-GF+ makes progress towards a universal database search tool for proteomics.* Nat Commun, 2014. **5**: p. 5277.

145. Vaudel, M., et al., *SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches.* Proteomics, 2011. **11**(5): p. 996-9.

146. Carvalho, P.C., et al., *PatternLab for proteomics: a tool for differential shotgun proteomics.* BMC Bioinformatics, 2008. **9**: p. 316.

147. Mesuere, B., et al., *Unipept: tryptic peptide-based biodiversity analysis of metaproteome samples.* J Proteome Res, 2012. **11**(12): p. 5773-80.

148. Huson, D.H., et al., *MEGAN analysis of metagenomic data.* Genome Res, 2007. **17**(3): p. 377-86.

149. Tanca, A., et al., *The impact of sequence database choice on metaproteomic results in gut microbiota studies.* Microbiome, 2016. **4**(1): p. 51.

150. Kanehisa, M., Y. Sato, and K. Morishima, *BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences.* J Mol Biol, 2016. **428**(4): p. 726-31.

151. Suzuki, S., et al., *GHOSTX: an improved sequence homology search algorithm using a query suffix array and a database suffix array.* PLoS One, 2014. **9**(8): p. e103833.

152. Denef, V.J., et al., *Proteomics-inferred genome typing (PIGT) demonstrates inter-population recombination as a strategy for environmental adaptation.* Environ Microbiol, 2009. **11**(2): p. 313-25.

153. Coram, N.J. and D.E. Rawlings, *Molecular relationship between two groups of the genus Leptospirillum and the finding that Leptospirillum ferriphilum sp. nov. dominates South African commercial biooxidation tanks that operate at 40 degrees C.* Appl Environ Microbiol, 2002. **68**(2): p. 838-45.

154. Schrenk, M.O., et al., *Distribution of thiobacillus ferrooxidans and leptospirillum ferrooxidans: implications for generation of acid mine drainage.* Science, 1998. **279**(5356): p. 1519-22.

155. Seifert, J., et al., *Bioinformatic progress and applications in metaproteogenomics for bridging the gap between genomic sequences and metabolic functions in microbial communities.* Proteomics, 2013. **13**(18-19): p. 2786-804.

156. Rooijers, K., et al., *An iterative workflow for mining the human intestinal metaproteome.* BMC Genomics, 2011. **12**: p. 6.

157. Speth, D.R., et al., *Genome-based microbial ecology of anammox granules in a full-scale wastewater treatment system.* Nat Commun, 2016. **7**: p. 11172.

158. Sekiguchi, Y., et al., *First genomic insights into members of a candidate bacterial phylum responsible for wastewater bulking.* PeerJ, 2015. **3**: p. e740.

159. Albertsen, M., et al., *Metagenomes obtained by 'deep sequencing' - what do they tell about the enhanced biological phosphorus removal communities?* Water Sci Technol, 2013. **68**(9): p. 1959-68.

160. Hug, L.A., et al., *Comparative metagenomics of three Dehalococcoides-containing enrichment cultures: the role of the non-dechlorinating community.* BMC Genomics, 2012. **13**: p. 327.

161. Lykidis, A., et al., *Multiple syntrophic interactions in a terephthalate-degrading methanogenic consortium.* ISME J, 2011. **5**(1): p. 122-30.

162. Nobu, M.K., et al., *Microbial dark matter ecogenomics reveals complex synergistic networks in a methanogenic bioreactor.* ISME J, 2015. **9**(8): p. 1710-22.

163. Taubert, M., et al., *Protein-SIP enables time-resolved analysis of the carbon flux in a sulfate-reducing, benzene-degrading microbial consortium.* ISME J, 2012. **6**(12): p. 2291-301.

164. Erdogan, M.F., *Thiocyanate overload and thyroid disease.* Biofactors, 2003. **19**(3-4): p. 107-11.

165. Speyer, M.R. and P. Raymond, *The acute toxicity of thiocyanate and cyanate to rainbow trout as modified by water temperature and pH.* Environmental Toxicology and Chemistry, 1988. **7**(7): p. 565-571.

166. Watson, S.J. and E.J. Maly, *Thiocyanate toxicity to Daphnia magna: modified by pH and temperature.* Aquatic Toxicology, 1987. **10**(1): p. 1-8.

167. Hussain, A., et al., *Cloning and expression of a gene encoding a novel thermostable thiocyanate-degrading enzyme from a mesophilic alphaproteobacteria strain THI201.* Microbiology, 2013. **159**(Pt 11): p. 2294-302.

168. Katayama, Y. and H. Kuraishi, *Characteristics of Thiobacillus thioparus and its thiocyanate assimilation.* Can J Microbiol, 1978. **24**(7): p. 804-10.

169. Katayama, Y., et al., *A thiocyanate hydrolase of Thiobacillus thioparus. A novel enzyme catalyzing the formation of carbonyl sulfide from thiocyanate.* J Biol Chem, 1992. **267**(13): p. 9170-5.

170. Sorokin, D.Y., et al., *Microbial thiocyanate utilization under highly alkaline conditions.* Appl Environ Microbiol, 2001. **67**(2): p. 528-38.

171. Katayama, Y., et al., *Cloning of genes coding for the three subunits of thiocyanate hydrolase of Thiobacillus thioparus THI 115 and their evolutionary relationships to nitrile hydratase.* J Bacteriol, 1998. **180**(10): p. 2583-9.

172. Kantor, R.S., et al., *Bioreactor microbial ecosystems for thiocyanate and cyanide degradation unravelled with genome-resolved metagenomics.* Environ Microbiol, 2015. **17**(12): p. 4929-41.

173. Chourey, K., et al., *Direct Cellular Lysis/Protein Extraction Protocol for Soil Metaproteomics.* Journal of Proteome Research, 2010. **9**(12): p. 6615-6622.

174. Chambers, M.C., et al., *A cross-platform toolkit for mass spectrometry and proteomics.* Nat Biotechnol, 2012. **30**(10): p. 918-20.

175. Ma, Z.Q., et al., *ScanRanker: Quality assessment of tandem mass spectra via sequence tagging.* J Proteome Res, 2011. **10**(7): p. 2896-904.

176. Arakawa, T., et al., *Structure of thiocyanate hydrolase: a new nitrile hydratase family protein with a novel five-coordinate cobalt(III) center.* J Mol Biol, 2007. **366**(5): p. 1497-509.

177. Beller, H.R., et al., *The genome sequence of the obligately chemolithoautotrophic, facultatively anaerobic bacterium Thiobacillus denitrificans.* J Bacteriol, 2006. **188**(4): p. 1473-88.

178. Vu, H.P., A. Mu, and J.W. Moreau, *Biodegradation of thiocyanate by a novel strain of Burkholderia phytofirmans from soil contaminated by gold mine tailings.* Lett Appl Microbiol, 2013. **57**(4): p. 368-72.

179. Hino, T., et al., *Structural basis of biological N2O generation by bacterial nitric oxide reductase.* Science, 2010. **330**(6011): p. 1666-70.

180. Matsumoto, Y., et al., *Crystal structure of quinol-dependent nitric oxide reductase from Geobacillus stearothermophilus.* Nat Struct Mol Biol, 2012. **19**(2): p. 238-45.

181. Orphan, V.J., et al., *Methane-consuming archaea revealed by directly coupled isotopic and phylogenetic analysis.* Science, 2001. **293**(5529): p. 484-7.

182. Boetius, A., et al., *A marine microbial consortium apparently mediating anaerobic oxidation of methane.* Nature, 2000. **407**(6804): p. 623-6.

183. Nauhaus, K., et al., *Environmental regulation of the anaerobic oxidation of methane: a comparison of ANME-I and ANME-II communities.* Environ Microbiol, 2005. **7**(1): p. 98-106.

184. Wegener, G., et al., *Metabolic Capabilities of Microorganisms Involved in and Associated with the Anaerobic Oxidation of Methane.* Front Microbiol, 2016. **7**: p. 46.

185. Schreiber, L., et al., *Identification of the dominant sulfate-reducing bacterial partner of anaerobic methanotrophs of the ANME-2 clade.* Environ Microbiol, 2010. **12**(8): p. 2327-40.

186. Scheller, S., et al., *Artificial electron acceptors decouple archaeal methane oxidation from sulfate reduction.* Science, 2016. **351**(6274): p. 703-7.

187. Losekann, T., et al., *Diversity and abundance of aerobic and anaerobic methane oxidizers at the Haakon Mosby Mud Volcano, Barents Sea.* Appl Environ Microbiol, 2007. **73**(10): p. 3348-62.

188. Krukenberg, V., et al., *Candidatus Desulfofervidus auxilii, a hydrogenotrophic sulfate-reducing bacterium involved in the thermophilic anaerobic oxidation of methane.* Environ Microbiol, 2016. **18**(9): p. 3073-91.

189. Sorokin, D.Y., et al., *Dethiobacter alkaliphilus gen. nov. sp. nov., and Desulfurivibrio alkaliphilus gen. nov. sp. nov.: two novel representatives of reductive sulfur cycle from soda lakes.* Extremophiles, 2008. **12**(3): p. 431-9.

190. Chourey, K., et al., *Environmental proteomics reveals early microbial community responses to biostimulation at a uranium- and nitrate-contaminated site.* Proteomics, 2013. **13**(18-19): p. 2921-30.

191. Thompson, M.R., et al., *Dosage-dependent proteome response of Shewanella oneidensis MR-1 to acute chromate challenge.* J Proteome Res, 2007. **6**(5): p. 1745-57.

192. Brown, S.D., et al., *Molecular dynamics of the Shewanella oneidensis response to chromate stress.* Mol Cell Proteomics, 2006. **5**(6): p. 1054-71.

193. Sharma, R., et al., *Coupling a detergent lysis/cleanup methodology with intact protein fractionation for enhanced proteome characterization.* J Proteome Res, 2012. **11**(12): p. 6008-18.

194. Bagnoud, A., et al., *Reconstructing a hydrogen-driven microbial metabolic network in Opalinus Clay rock.* Nat Commun, 2016. **7**: p. 12770.

195. North, J.A., et al., *Metabolic Regulation as a Consequence of Anaerobic 5-Methylthioadenosine Recycling in Rhodospirillum rubrum.* MBio, 2016. **7**(4).

196. Wegener, G., et al., *Assimilation of methane and inorganic carbon by microbial communities mediating the anaerobic oxidation of methane.* Environ Microbiol, 2008. **10**(9): p. 2287-98.

197. Dekas, A.E., R.S. Poretsky, and V.J. Orphan, *Deep-sea archaea fix and share nitrogen in methane-consuming microbial consortia.* Science, 2009. **326**(5951): p. 422-6.

198. Dekas, A.E., et al., *Spatial distribution of nitrogen fixation in methane seep sediment and the role of the ANME archaea.* Environ Microbiol, 2014. **16**(10): p. 3012-29.

199. Green-Saxena, A., et al., *Nitrate-based niche differentiation by distinct sulfate-reducing bacteria involved in the anaerobic oxidation of methane.* ISME J, 2014. **8**(1): p. 150-63.

200. Grover, H. and V. Gopalakrishnan, *Efficient Processing of Models for Large-scale Shotgun Proteomics Data.* Int Conf Collab Comput, 2012. **2012**: p. 591-596.

201. Reiter, L., et al., *Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry.* Mol Cell Proteomics, 2009. **8**(11): p. 2405-17.

202. Elias, J.E. and S.P. Gygi, *Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry.* Nat Methods, 2007. **4**(3): p. 207-14.

203. Aklujkar, M., et al., *Proteins involved in electron transfer to Fe(III) and Mn(IV) oxides by Geobacter sulfurreducens and Geobacter uraniireducens.* Microbiology, 2013. **159**(Pt 3): p. 515-35.

204. Thomas, L., *Emerging applications in clinical mass spectrometry*. 2016.

205. *What is a LIMS?* 2016.

206. Vizcaino, J.A., et al., *ProteomeXchange provides globally coordinated proteomics data submission and dissemination.* Nat Biotechnol, 2014. **32**(3): p. 223-6.

207. Whelan, L.C., et al., *Applications of SELDI-MS technology in oncology.* J Cell Mol Med, 2008. **12**(5A): p. 1535-47.

208. Hultman, J., et al., *Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes.* Nature, 2015. **521**(7551): p. 208-12.

209. Muller, E.E., et al., *Community-integrated omics links dominance of a microbial generalist to fine-tuned resource usage.* Nat Commun, 2014. **5**: p. 5603.

210. Roume, H., et al., *Comparative integrated omics: identification of key functionalities in microbial community-wide metabolic networks.* 2015. **1**: p. 15007.

211. Gonnelli, G., et al., *A decoy-free approach to the identification of peptides.* J Proteome Res, 2015. **14**(4): p. 1792-8.

212. Griss, J., et al., *Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets.* Nat Methods, 2016. **13**(8): p. 651-656.

Vita

Ramsunder (Sarvesh) Iyer was born in Nagpur (Maharashtra), India. He completed his Bachelors in Biotechnology, Chemistry and Zoology in 2006 and Masters in Biotechnology in 2008 from Hislop College, RTM-Nagpur University. After completing his Master's degree, he worked as a Technical Sales Officer in Lilac Medicare Pvt. Ltd. in Mumbai, India for one year. He then joined Eugeniks, a Cytogenetic and DNA Laboratory in Nagpur and worked there as a research associate under the supervision of Dr. Vinay Tule where he was chiefly involved in prenatal diagnosis from amniotic fluid and chorionic villus biopsy samples for detailed investigations of genetic disorders like sickle cell disease, β- thalassemia and trisomy (Down's syndrome, Edward's Syndrome, Patau's Syndrome etc.). After working at Eugeniks for three years, he enrolled in the UTK-ORNL Graduate School of Genome Science and Technology in Fall 2012 and later joined the lab of Dr. Robert L. Hettich to pursue doctoral studies in the field of Mass Spectrometry based Proteomics. He expects to receive his Ph.D. in August of 2017.