8-2014

Development of an experimental and computational platform for enhanced characterization of modified peptides and proteins in environmental proteomics

Ritin Sharma
University of Tennessee - Knoxville, rsharma3@utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Ritin Sharma entitled "Development of an experimental and computational platform for enhanced characterization of modified peptides and proteins in environmental proteomics." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Life Sciences.

Robert L. Hettich, Major Professor

We have read this dissertation and recommend its acceptance:

Kurt Lamour, Alison Buchan, Cynthia Peterson, Loren Hauser

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

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

Development of an experimental and computational platform for enhanced characterization of modified peptides and proteins in environmental proteomics

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

Ritin Sharma
August 2014

DEDICATION

This dissertation is dedicated to the memory of
Divesh Thimmaiya who was a great friend, a brilliant researcher and a constant motivator who left a big impression on me in the short period of 4 years I had known him.
(04.19.1982 - 10.05.2008)

ACKNOWLEDGEMENTS

First of all, I would like to thank my adviser Dr. Robert L. Hettich for being an excellent mentor who guided me through the ups and downs of graduate school and provided challenging scientific problems to address in a world class research environment. He has always led by example and has never asked any student to work more hours than he himself has put every week. I cannot remember a single meeting with him in last five years or so, when he has not started with a positive statement, thereby never letting me lose morale when the going was tough. I cherish the relationship I have built with him and I look up to his guidance in every sphere of life.

I would like to thank my esteemed committee members Dr. Cynthia Peterson, Dr. Loren Hauser, Dr. Kurt Lamour and Dr. Alison Buchan for their time and valuable guidance during the course of my PhD. They helped me think in a broad perspective, which is important as I progress in my career and have to face moments where I am put against people from different scientific backgrounds.

I would like to thank Dr. Karuna Chourey for teaching experimental approaches for biological mass spectrometry and her constant support during the course of my graduate studies. I thank Dr. Brian Dill for his help in understanding mass spec instrumentation, teaching the GELFrEE approach and for his motivation throughout my PhD. I thank Dr. Paul Abraham and Dr. Rachel Adams for their help in PTM project via sequence tagging methods. I thank Dr. Rich Giannone for his guidance in developing experimental methods for soil proteomics. I am thankful to Dr. Greg Hurst for the numerous helpful discussions with respect to fundamental mass spectrometry and his guidance in troubleshooting mass spec instrumentation. I thank Dr. Nathan VerBerkmoes for valuable discussions with respect to soil proteomics. I thank Keiji Asano for making our life easier in the mass spec lab and for being always available to resolve any issues in the operation of lab. I thank Becky Maggard for her help in administrative work at ORNL and with printing of posters. I thank all the organic and biological mass spectrometry group members for their support during my time at ORNL especially Zhou Li and Xiaoxin Liu with whom I shared my office space.

I would also like to acknowledge and thank the Genome Science and Technology program, without which I would not have been here. I thank Dr. Albrecht Von Arnim for his support and

guidance during the course of program. I thank the GST staff, Kay Gardner, Terrie Yeatts and Roger Gray for their assistance throughout my time in the graduate school.

I thank all my friends especially Sumit Goswami, Sangeetha Rajagopalan, Migun Shakya, Pintu Masalkar, Ansul Lokdarshi, Sukanya Iyer, Tripti Bhaskar and Swapna Purandare who made living in Knoxville fun and were a big support during the prelim exam.

A special thanks to my parents who worked very hard to give me the best education. During a time and environment, when everyone else around them constantly nagged them that it was a waste of money to spend most of their income on our education, they did not back down. It is their hard work and support which has allowed me to pursue my graduate studies.

Lastly, I want to thank my beloved wife Ritika Sehgal, whose love and care has made it possible for me to complete this dissertation. Her relentless support and constant motivation always egged me on and helped me realize my goals.

ABSTRACT

Over the last decade, mass spectrometry based proteomics has been established as the front-runner in systems-level protein expression studies. However, with the field progressing into research of more and more complex samples, novel challenges have been raised with respect to efficient protein extraction and computational matching. In this dissertation, various aspects in the proteomics workflow, including experimental and computational approaches, have been developed, optimized and systematically evaluated. In this work, some of the critical factors with respect to proteomics sample preparation, like available biomass, detergent removal methods, and intact protein fractionation to achieve deeper proteome measurements were evaluated. The presented work will help the broader scientific community to carefully design proteomics experiments especially in biomass limited samples.

A second major area of focus in this dissertation is comprehensive characterization of post-translational modifications (PTMs) in different biological systems. PTMs are critical for functioning of both the prokaryotic and eukaryotic species and this dissertation will highlight some of the experimental strategies to explore the diversity of PTMs in microbial isolates via application of alternate protease and multiple fragmentation schemes. The PTM discovery approach will be further extended into a complex eukaryotic model trees species *Populus trichocarpa* using recently developed sequence tagging methods to carryout broad scale PTM search and a complete blind PTM search.

Although the work presented in this dissertation mainly revolves around prokaryotic and eukaryotic species involved in environmental proteomics, the general considerations outlined in this work can be extended to every proteomics pipeline. Thus this dissertation will benefit the scientific community in carefully designing experiments before embarking on any research project involving mass spectrometry.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1- Introduction to MS-based proteomics and current challenges in metaproteomics studies

---

1.1 Historical perspective of mass spectrometry in biological sciences

The field of mass spectrometry (MS) was born when Sir J. J. Thomson constructed the first mass spectrometer in 1912 and used it for separation of neon isotopes. For the next couple of decades, the field of mass spectrometry made great strides in isotope measurements through the pioneering works of Dr. Francis Aston and Dr. Arthur Dempster [1-3]. The onset of World War II in 1939 revealed the real potential of MS in exact mass calculation and isotope separation. However the field remained focused on solving problems in physics and petroleum industry.

With major breakthroughs appearing in the field of biology, spear-headed by the elucidation of double helical structure of DNA by Watson and Crick, scientists had a more detailed picture of major biomolecules present in biological systems [4]. This development prompted some researchers to venture into the relatively unknown area of biological MS to measure these biomolecules. In the year 1959, MS was applied for peptide and oligonucleotide sequencing and later in 1962 for understanding nucleotide structures [5, 6].

However, the lack of appropriate ionization methods limited the field to small peptide sequencing, and it was not until the development of fast-atom bombardment method in 1981 that peptides and small proteins were being measured by MS without the requirement of digestion [7]. The field gained significant momentum towards measurement of intact proteins with the introduction of electrospray ionization by Dr. John Fenn in 1988 [8-11]. As the ESI mode could

easily be coupled to liquid chromatography, it allowed for simultaneous measurement of protein mixtures. During the similar time frame, Hillenkamp in Germany and Tanaka in Japan developed Matrix-assisted Laser Desorption/Ionization (MALDI) to be used with Time-of-Flight analysis of biomolecules [12-15]. The pioneering work of Fenn and Tanaka was recognized by the Royal Swedish Academy of Sciences and they were awarded Noble Prize in Chemistry for the development of methods to progress the field of biological MS in the year 2002.

1.2 Introduction of mass spectrometry for systems level characterization of proteins

The early 2000's ushered in a paradigm shift in the way scientific research, in particular molecular biology, could be carried out. This was fueled by the publication of draft human genome, which induced widespread interest in characterizing biological activity from a holistic view, rather than looking at the individual molecular players [16, 17].

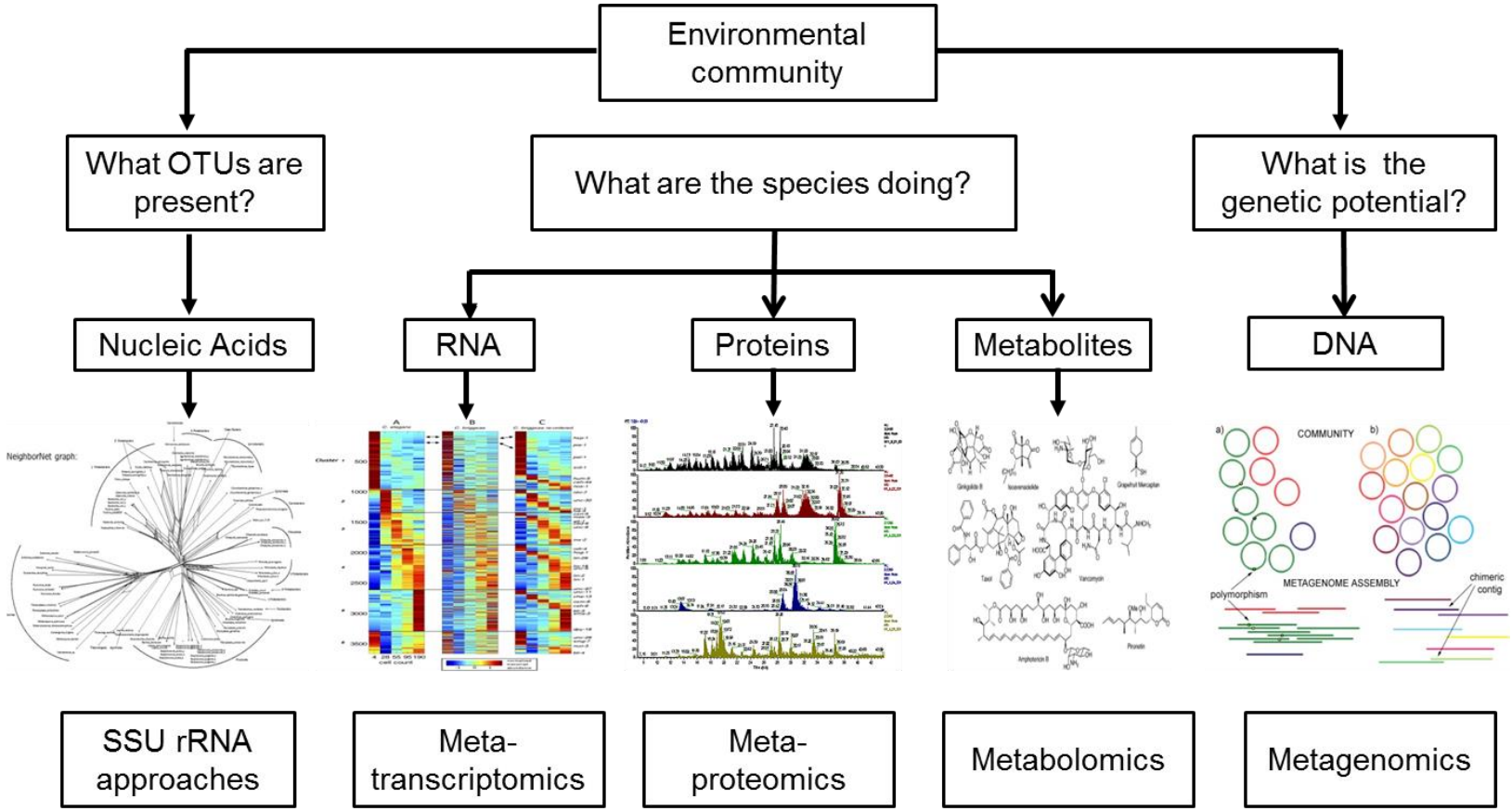
The publication of human genome lead to the development of numerous "OMICS" approaches to do "Systems Biology". As the name suggests, systems biology is directed at addressing biological questions at a holistic level rather than the traditional reductionist approach [18]. The word "system" in systems biology can have multiple connotations, as it could refer to a single tissue or an organ type, a single individual or a homogenous population of one species, or an entire community made up of different species.

The "OMICS" approaches in systems biology relate to three major levels of information processing in biology. A "Genomics" approach aims to identify species composition by looking at Operational Taxonomic Units (OTUs) via small subunit ribosomal RNA (SSU rRNA) methods as well as the total genetic potential in a given species defined by DNA sequencing

techniques [19]. While “genomics” provide a *static picture* of what all can be expressed by a species, all the other OMICS approaches look into the functional signature of biological system. As per the central dogma of molecular biology, information encoded in genes gets transcribed into mRNAs, which are then translated into proteins that carry out inherited functions. The proteins then interact with metabolites to carryout cellular function, as depicted in the overall scheme of **Figure 1.1**.

Thus, all the other OMICS approaches apart from genomics provide a dynamic picture of a cell by measuring both qualitative and quantitative information for mRNAs (transcriptomics) as well as the proteins (proteomics) or metabolites (metabolomics) levels. [20, 21] While it is important to know the mRNA levels to understand information translated from DNA to proteins, it is apparent that mRNAs are not the actual biomolecules that are involved in structural maintenance and housekeeping of cellular environment. Except for a small subset of specialized RNAs, they are the intermediates, which code for proteins that carryout cellular function alongside different metabolites. Therefore proteomics, which targets measurement of the entire suite of expressed proteins in a biological sample, provides a comprehensive, spatial and functional viewpoint of the cellular activities at systems level.

Before MS-based proteomics came to forte; the most common approach for proteome measurements was two Dimensional –Polyacrylamide Gel Electrophoresis or 2D-PAGE [22-24]. The 2D-Differential Gel Electrophoresis was a further advancement over the 2D-PAGE and employed fluorescent dyes for visualization and quantification. The basic principle of both these methods was to separate complex protein mixtures in two dimensions, first by



Adapted from Figure 3 in *Gut*. 2008 **57**(11):1605-15

Figure 1.1 OMICS approaches for systems biology research at the community level

the net charge and second by their molecular weight. Next by superposing two 2D-Gels from the samples across different conditions, spots were visually inspected for differential expression. The fluorescent dyes used in 2D-DIGE experiments would provide quantitative value to this differential expression. Even though 2D-DIGE was superior to any of the previous methods, it suffered from problems of reproducibility, amount of time taken for visual inspection and scope of only a two-way comparison [23].

Earlier mass spectrometric studies focused on measuring the molecular masses of isolated intact proteins or artificial peptides. By combining molecular weight based protein separation and liquid chromatography coupled to MS, the field evolved to measure a subset of proteome by intact protein measurements and associated tandem MS, which forms the approach termed as the "Top-Down" proteomics [25, 26]. The top-down approach not only measures intact protein mass but can also provide information on various isoforms as well as post-translational modifications on a protein **(Figure 1.2)**. However, the top-down proteomics approach is limited to measuring simple protein mixtures, due to the inefficient separation of complex samples by liquid chromatography and the inability of mass spectrometers to accurately measure high molecular weight proteins.

Converse to the top-down method of protein identification, a second approach aptly called the "bottom-up" proteomics measures proteolytic peptides from a cell and computationally maps these peptides to proteins to produce identifications [27, 28]. "Bottom-up" proteomics provides some distinct advantages over "Top-Down" proteomics including high throughput, ability to detect all types of proteins irrespective of their molecular weight since it is a peptide centric approach with high-resolution separations and relatively easy to interpret mass spectra **(Figure 1.2)**.

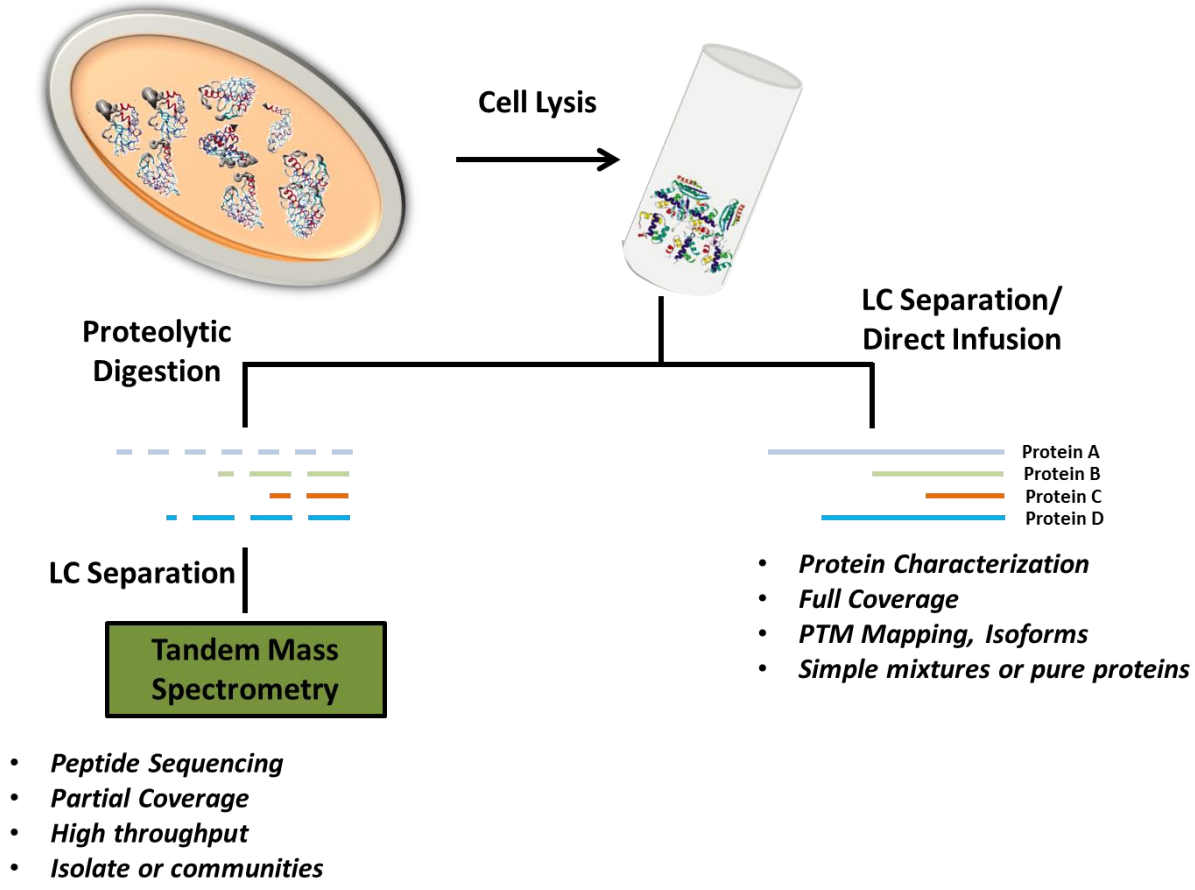


Figure 1.2 "Top-Down" and "Bottom-Up" approach for proteomics

Within the bottom-up proteomics, there are two distinct modes of protein identification, namely peptide mass fingerprinting and shotgun proteomics [29, 30]. While the peptide mass fingerprinting is more commonly used with TOF mass analyzer (often avoiding the need/use of tandem MS measurements), the shotgun proteomics approach has mainly been employed with ion-trap mass analyzers. A detailed explanation of different mass analyzers will be presented in Chapter 2 of this dissertation. The shotgun proteomics approach, named after a similar technique used for DNA sequencing, works by digesting peptides and separating them via liquid chromatography methods. The liquid chromatography separation can either be coupled to a mass spectrometer in what is known as online – LC-MS/MS or be separate from the mass spectrometer, termed as the offline LC-MS/MS. The Multidimensional Protein Identification Technology (MudPIT) scheme, which is explained in detail in Chapter 2, is a robust online 2D-LC-MS/MS approach for shotgun proteomics and is widely used for both discovery proteomics and hypothesis driven proteomics experiments [30].

1.3 Mass spectrometry of complex systems – the realm of community proteomics

With the rise of cheap and high throughput DNA sequencing methods, it is now relatively easy to determine the genetic potential of not just one species, but an entire ensemble of microbial players in the environment especially those systems which have low level of species complexity. This whole community sequencing approach, or *metagenomics* as it is commonly referred to, has propelled the field of mass spectrometry into surveying the functional proteome at a higher level of resolution, giving birth to *metaproteomics* [31].

The field of metaproteomics has progressed steadily to provide in-depth information on community architecture, functional activity, symbiotic and parasitic behavior of community members as well as their role in manipulating the surrounding environment [32-35].

One of the first studies in metaproteomics investigated an uncultured acid-mine drainage (AMD) system to decipher microbial species abundance with respect to functional activity, and identified key proteins involved in biofilm formation in this harsh environment [36]. While the initial work used a label-free approach to determine protein abundance, a more recent study by this same team used stable isotope probing (SIP) for quantitative determination of protein flux changes in the AMD system [37].

Some of the other systems that have been investigated by metaproteomics include nutrient-rich and nutrient-limited ocean ecosystems, microbial bioremediation activity in different contaminated sites, and understanding the host-microbe interactions, for instance in termite hindgut and crop rhizospheres [38-42].

One particular area of research that has benefitted significantly by metaproteomics is the human microbiome. It is known that human body is a host to far more microbial cells compared to the human cells, and this suggests that a human body expresses an order of magnitude more microbial genes than human. In light of these observations, there has been a tremendous interest in understanding why a human body provides safe haven to such a large number of microbial species, and does an imbalance in this harmonious relationship leads to certain type of human diseases? [43-46]

Metaproteomics studies revealed a detailed map of species diversity in human infant GI tract as well as core proteome in an adult human intestine [47]. Metaproteomics studies have provided an

insight into the role of human and microbial proteins in keeping this delicate symbiotic relationship in balance [48].

The metaproteomics studies have been a boon to the researchers who work on uncultivable microbial members (i.e. those that cannot be artificially grown in lab). Though the number of factors that can be varied in environmental systems is very limited compared to well-defined lab grown cultures, the analysis of unaltered native environmental samples may explain the real physiological role of individual members in the aggregate.

1.4 Challenges in community proteomics

Since the metaproteomics measurements heavily rely on metagenomes for proteome mapping, it is very important that the sequencing experiments produce enough high confidence reads that can be assembled into well-defined genomes. This is one of the most challenging steps in metagenomics and can lead to the creation of huge protein databases with millions of sequences which may or may not accurately reflect microbial diversity. As evident, deeper metagenomics measurements leads to a larger predicted protein database, which impacts the computational searches in multiple ways. Firstly, a large metagenome increases the search space which, in turn, increases the computational time to accomplish a proteomics search. Secondly, since most of the proteomics searches use a reverse database strategy to calculate a False Discovery Rate (FDR), which is a quantifiable metric to ascertain the accuracy of peptide calls, a larger reverse database means higher probability of matching a false hit. A more detailed explanation of the impact of metagenome quality on proteomics measurements will be given in Chapter 4.

A third challenge with metagenomics is the high occurrence of closely related sequences across genomes. Since metaproteomics provides more information via bottom-up approaches, which are

peptide-centric, it is predicated on the task of assigning sequenced peptides back on to the proteins. Therefore, the additive effect of large size and high redundancy in metagenome mandates that the metaproteomics measurements are carried out on instruments that afford high scanning speed, high resolution and high mass accuracy. Fortunately, the instrumentation challenge for metaproteomics measurement has been solved to some extent, thanks to commercially available, next generation instruments, such as the ThermoScientific LTQ-Orbitrap-Elite and LTQ-Orbitrap-Fusion.

The challenge, however, remains in more extensive and cleaner extraction of proteome, which has better exclusion of the complex matrix from which the community systems originate, be it soil, groundwater or gut microbiome **(Figure 1.3)**. Therefore, there has been a great need for an optimized and efficient sample preparation work flow for metaproteomics.

1.5 Proteomics to understand post-translation modifications

Post-translational modifications can be defined as alterations in protein structure by either addition of chemical moieties or deletion of amino-acids, which occur either at the time of translation or after translation. Most of the PTMs are reversible in nature, while some are permanent. For example, addition of a chemical group like a phosphate or acetate is a reversible protein modification, while post-processing of the polypeptide via N-terminal cleavage is not [49, 50].

Whether temporary or permanent, all PTMs on a protein serve unique and distinct functions. A protein can have multiple level of the same PTM on an amino-acid, for example, mono-phosphorylation, di- phosphorylation etc., or multiple types of modifications on an amino-acid, like methylation and acetylation. It is the combination of stoichiometry and diversity of PTMs

(a) Biofilm from an acid-mine drainage system. Green color due to high metal content

(b) Groundwater sample from Rifle aquifer, Colorado

(c) *Olavius algarvensis* symbiotic system

(d) Soil systems

Figure 1.3 A snapshot of different sources for community metaproteomics. *Each of these biological systems has some common interfering components as well post some unique challenges to extract cleaner proteomes.*

(Image source: https://maple.lsd.ornl.gov/mspipeline/)

that determines the final outcome of a protein activity. Since any type of addition or deletion of chemical moiety from a protein results in a mass difference, mass spectrometry is the best tool to investigate PTMs. While the extent of PTMs in biological systems is very wide, major strides have been made in measurement of phosphorylated and acetylated proteomes in multiple organisms [51-56]. This was aided by development of robust sample preparation methods which used selective enrichment of acetylated and phosphorylated peptides from whole cell lysates [57, 58]. Computational search algorithms were altered not just to search for amino acid-based peptide fragment masses, but also to include mass shifts caused by specific PTMs for peptide-spectrum matching. However, this type of enrichment based approach to detect PTM is currently limited to a very few modifications, typically including acetylation, phosphorylation, and methylation.

1.6 Scope of the dissertation

This dissertation will focus on the two critical components which determine the depth of mass spectrometry measurements for protein and PTM identification: (a) Efficient sample preparation method for deeper proteome measurement and (b) optimized informatics pipeline for comprehensive PTM identification. In addition to these two major aspects, this dissertation will also describe some of the biological insights obtained from our optimized experimental approaches.

Chapter 2 of this dissertation provides in-depth explanation of biological mass spectrometry workflow. Herein, fundamental understanding of different ionization modes, choice of mass spectrometry platform, peptide sequencing, and spectral matching will be described.

Chapter 3 of this dissertation will describe the development of optimized experimental strategy for improved sample preparation, taking into consideration the starting biomass of biological sample. It will also introduce the concept of solution-based intact protein fractionation and its integration in routine MudPIT analysis for enhanced proteome coverage. The findings from the work described in this chapter provide foundation for experimental workflow used in remaining chapters.

Chapter 4 of this dissertation will evaluate sample preparation methods with respect to microbial soil and groundwater communities. It will highlight some of the key parameters to examine when the total protein identification are low and below expectation, and will describe strategies to compare the quality of raw MS data with respect to the quality of metagenomes.

Chapter 5 of this dissertation will describe some of the challenges associated with PTM identification. It will give an overview on different fragmentation modes to assist in PTM identification as well look into different search pipelines and their integration for comprehensive PTM identification in microbial isolates.

Chapter 6 of this dissertation will focus on PTM identification in *P. trichocarpa* genome. In this chapter, implications of using a sequence tagging approach vs. a normal database searching mode will be discussed. Results from using a sequence tagging approach in true blind PTM mode and a broad PTM search mode will be provided for three different organ types in *P. trichocarpa*.

Chapter 7 of this dissertation will extend the work of chapter 6 to provide a biological insight into a natural microbial community that was grown in lab. Using techniques developed in Chapter 3, this work will describe proteomics sample preparation of biomass limited samples and

will provide first proteomics validation of alternate coding by a STOP codon by a member of this microbial community.

Chapter 8 summarizes the major accomplishments of this dissertation research and provides a framework to carryout future proteomics experiment for protein or PTM identification in diverse sample types. It concludes by highlighting some of the key areas that need further attention for comprehensive protein identification.

Chapter 2 - Experimental and computational approaches for MS-based proteomics

2.1 General overview of proteomics experiment workflow

In this chapter, the overall approach and major considerations for examining biological samples for MS based shotgun or bottom-up proteomics will be described.

One of the bottom-up proteomics strategies that is widely acceptable and provides high quality proteome coverage is called the Multi-dimensional Protein Identification Technology (MudPIT) [30]. The MudPIT approach was originally developed by Dr. Michael Washburn and Dr. John Yates at The Scripps Research Institute. In consideration of the work scope for this dissertation, this approach was chosen and optimized to match the research needs of characterizing normal and modified peptides/proteins that provide metabolic information for a variety of microbiology studies.

An efficient sample preparation method for discovery proteomics should be able to quickly disrupt microbial biomass without any bias towards a specific biological system and/or cell architecture, should lead to complete denaturation of proteins without degradation, and employ reagents that are mass spec friendly. However, in reality it is hard to employ a protocol that fits all the before mentioned metrics, and therefore researchers have to circumnavigate these challenges to get best results. The described workflow below is a generic protocol which was modified, based on research requirements and those specific elements are discussed as appropriate in other chapters.

A typical MudPIT experiment begins by lysing biological material (microbial culture, cell pellets, plant tissue etc.) to rupture the cellular membranes and release the proteins into solution phase. The most commonly used modes of lysis include chemical disruption, including usage of harsh detergents and chaotropes or mechanical means like sonication and bead-beating, or combination of both chemical and mechanical treatments [59]. Once the proteins are in solution, they can be chemically denatured and digested into peptides using a protease.

Once the protein sample is digested to the peptide level, it is ready to be loaded onto a chromatographic column. But before doing so, the peptide solution is cleaned-up (usually termed "desalting" for removal of residual salts, detergents, impurities that can negatively impact MS measurements). Desalting can be done either offline (i.e. prior to the loading of peptides on a chromatographic column) or online (i.e. after the peptides are loaded on the chromatography column). The offline desalting method usually employs a Seppak cartridge in a process called Solid-Phase Extraction (SPE). The peptide solution is loaded onto the Seppak cartridges that are packed with C18 resin. The cartridge is first washed with 100% H_20, 0.1% formic acid solution to remove any salts and then peptides are eluted from the cartridge using 100% acetonitrile (AcN), 0.1% formic acid solution. As the last step prior to loading the peptides on a chromatographic column, a solvent exchange step is carried out which puts back peptides into acidified aqueous media from the initial organic media. This ensures that the peptides are in a protonated state, which is required for MS measurements in positive mode.

The peptides are loaded on to a Strong-Cation Exchange (SCX) column which is connected to a Reverse-Phase (RP) column. The biphasic column is interfaced to a mass spectrometer which is operated in a data-dependent mode. Peptides are eluted from SCX back column by step-wise increasing concentration of ammonium acetate and then separated by gradient elution using

organic phase from the Reverse-phase column. The RAW MS and MS/MS data thus obtained from mass spectrometer is searched against predicted proteome by a MS data search program resulting in a list of identified peptides which were mapped on to their respective proteins. This complete pipeline starting from biological material to protein inference constitutes the MudPIT approach of MS-based proteomics **(Figure 2.1)**.

An alternative to Seppak clean-up is on-column cleaning using biphasic back column i.e. a RP-SCX back column. Since there is evidence of peptide loss with the usage of SPE, the on-column approach was adopted for the studies in this dissertation. A more detailed explanation of this scheme will be given in appropriate chapters.

2.1.1 Sample Consideration

Proteomics sample preparation protocols can accommodate a wide variety of sample types. In a simplistic scenario, a clean frozen cell pellet is preferred. There are various ways of quantifying workable amounts for proteomics, including wet cell pellet weight, total number of cells, or protein concentration. In general, knowing the protein concentration is the best route to decide on downstream sample preparation steps. Ideally, a 1 mg – 2 mg total protein amount will be suited for most of the proteomics sample preparation protocols, but one can go as low as 10 μg of total protein amount to detect a partial proteome. When it comes to environmental samples, the starting protein amount is empirically decided, since most of the protein estimation methods fail due to interferences. For soil samples, a typical starting amount varies from 5-100 gm of soil. Similarly, for groundwater samples, it can vary from 5 ml to 20 ml of starting volume for proteomics sample preparation. While working with environmental samples, it is best to carry out a test run and then decide if the chosen amount provides good quality MS/MS data or not.

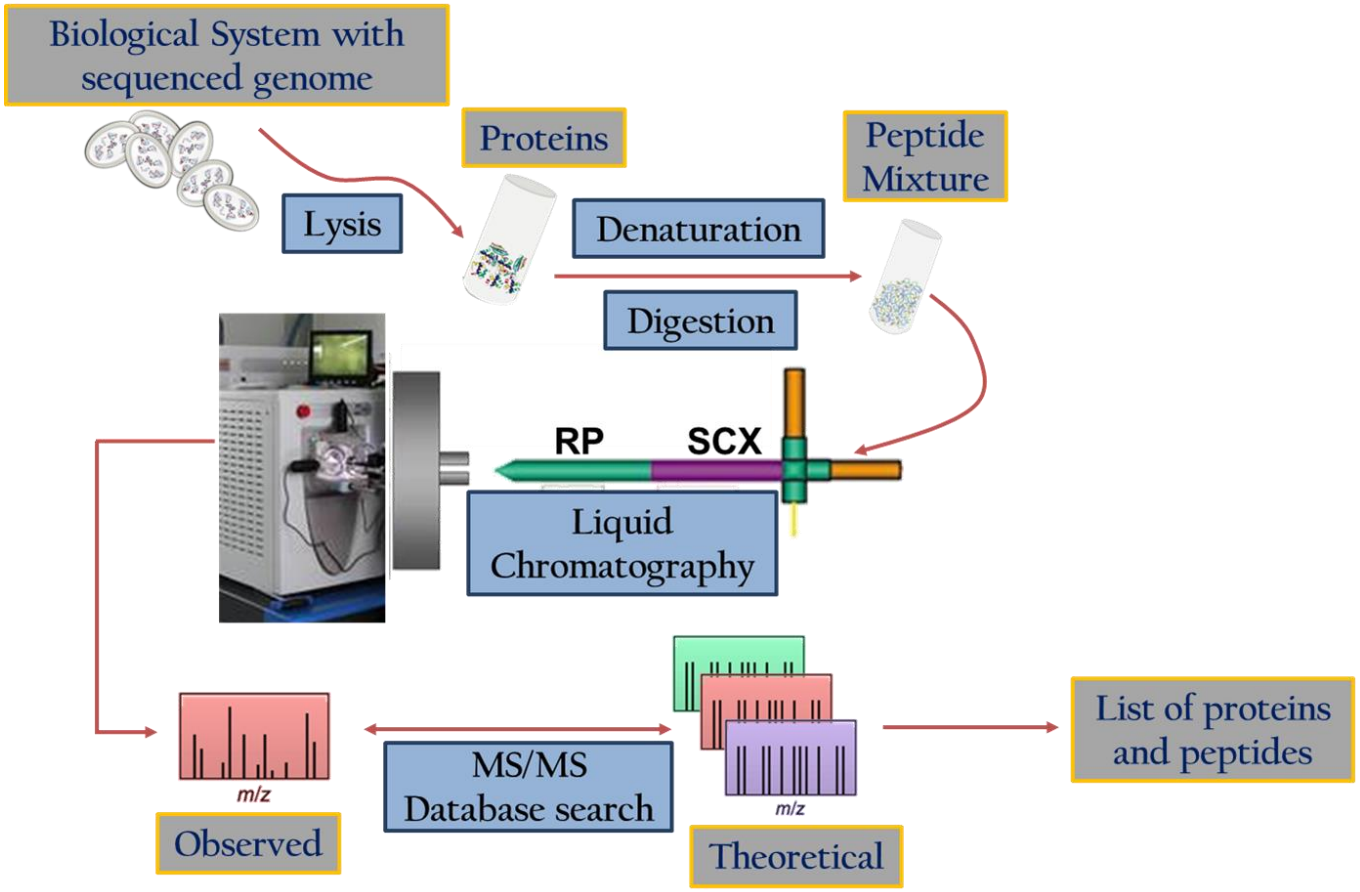


Figure 2.1 Schematic diagram of the MudPIT workflow

Many times, it can be an iterative procedure, where different sample prep methods are tried using different amount of starting material. Samples should be frozen soon after extraction to ensure minimum proteome change.

2.1.2 Protein Estimation

The starting protein amount and the peptide amount to load for each mass spec run was determined by Bicinchoninic Acid Assay (BCA) [60]. The assay employs reduction of Cu^{2+} to Cu^{1+} by proteins in an alkaline solution, followed by detection of cuprous cation using bicinchoninic acid. First, a standard curve was obtained by measuring optical density of bovine serum albumin prepared in serial dilution and then the optical density of sample in question was measured. Using standard curve as the reference, the protein concentration of biological sample was determined.

Protein estimation experiments in this dissertation were performed by Pierce BCA Assay Kit (Thermo Scientific). The 2 mg/ml ampoule of BSA (provided in the kit) was serially diluted to give a concentration range of 31.25 µg to 2 mg/ml by addition of HPLC H_20 or any compatible buffer. Next, two to three dilutions of biological sample were made. A BCA working reagent was prepared by mixing 50 parts of Reagent Bottle A with 1 part of Reagent Bottle B both provided in the kit. 1 ml of the working reagent was added to each of the BSA standard and the sample. Next the samples were incubated for 30 minutes in a water bath at 37 °C. Following the incubation, samples were cooled to room temperature and their absorbance was measured using a spectrophotometer (Biomate 3, Thermo Scientific, Waltham MA) at a wavelength of 562 nm. While the BCA assay was designed for protein quantitation, we also employed it for peptide quantification, since the color formation in the assay is not only dependent on protein structure

but is also affected by peptide bonds, presence of cysteine, cystine, tryptophan and tyrosine residues which are present in peptides as well. Although the accuracy is less defined than that for proteins, quantification of peptides provided a means to systematically control the amount of sample loaded onto the column in an effort to standardize the measurements.

2.1.3 Cell Lysis and protein denaturation

Most of the microbial and environmental samples investigated in this dissertation were solubilized with SDS, which is a denaturing anionic surfactant. It works by disrupting hydrophobic interactions in cell membrane, thereby compromising cellular integrity and releasing most of the proteins into solution [61]. Biological samples were solubilized by addition of SDS solution (SDS dissolved in 100 mM Tris-HCl buffer pH 8.0) keeping the final concentration of SDS less than 4% and 10 mM DTT.

The samples were either boiled in SDS solution for 10-15 minutes at 95 °C (environmental samples) or treated at 60 °C (microbial samples). The samples were cooled to room temperature and then 100% TCA solution was added to achieve a final concentration of 20-25% TCA. The tubes were than kept overnight at low temperature (on an ice-bath in a fridge or at -80 °C), and the TCA precipitated sample was thawed on ice the next day. The tubes were spun at 21,000 g for 10-15 minutes. The supernatant was discarded and the pellet washed twice using chilled acetone by centrifugation at 21,000 g. The resulting protein pellet was then treated with 8M Urea, which further denatures proteins. The denatured proteins were reduced by addition of 10 mM DTT and alkylated by addition of iodoacetamide. The addition of DTT helped in breaking of disulfide bonds and IAA blocked the released free thiol groups so that they cannot reform disulfide bonds. An alternate approach for denaturing proteins was to add 6M guanidine in 50

mM Tris-HCl buffer to the TCA precipitated protein pellet and heat the tube for 30 minutes at 60 °C.

2.1.4 Protein Digestion

The next phase in the MudPIT approach is to optimally generate peptides that will be sequenced by mass spectrometry. Protein digestion can be achieved both enzymatically and chemically. The most commonly used chemical for protein digestion is Cyanogen Bromide, which specifically cleaves after methionine residues and is best suited for digestion of membrane proteins. Some other chemical treatments include formic acid, hydrochloric acid, acetic acid and hydroxylamine [62].

In spite of the availability of many low cost chemicals, enzymatic digestion is the most prevalent method in the field of bottom-up proteomics. Amongst numerous proteases available for protein digestion, trypsin is the most widely used endoprotease. Trypsin acts as a catalyst in the hydrolysis of peptide bonds and specifically cleaves on the C-terminal side of lysine and arginine residues unless these two amino-acids are followed by a proline [63]. Some of the favorable aspects of using trypsin for proteolytic digestion include low cost and optimal length of tryptic peptides. On an average, tryptic peptides are between 10 and 20 amino acids depending on the frequency of lysine and arginine residues in a protein. This molecular weight range is optimal for measurements with the majority of mass analyzers which typically scan in the range of 400-2,000 m/z.

In this study, trypsin was used in a 1:20 enzyme to substrate ratio with respect to protein concentration. Denatured samples were subjected to trypsin digestion initially for 4 hours, and then followed with another round of trypsin digestion overnight. For the samples prepared via

guanidine denaturation, the sample was first diluted six fold with Tris buffer, as trypsin is rendered inactive in high concentration of guanidine. Depending on the denaturing agent, trypsin digestion was either carried out at 37 °C (for guanidine-HCl) or 25 °C (for Urea).

2.1.5 Sample Clean-up and column preparation:

For samples prepared by guanidine, the removal of residual salts and detergents was carried out via Reverse-Phase solid phase extraction method. For the desalting procedure, Sep-Pak Plus and Sep-Pak Light C18 cartridges (Waters Corporation) were used. First the cartridges were equilibrated by passing 100% HPLC water, 0.1% formic acid with a syringe. This step helped in removal of any contaminants present on the reverse-phase resin. Next, the sample was passed through the cartridge at the rate of 1-2 drops per second using a syringe. To ensure maximum binding, the flow-through was injected again using a syringe. Finally the peptides were eluted into three, 2 ml Eppendorf tubes using an organic solvent (100% AcN, 0.1% formic acid). The peptides were then solvent exchanged back into aqueous phase using a speedvac which is required for protonation of peptides.

The samples prepared by urea denaturation were transferred to a 10 kDa molecular weight cut-off filter (MWCO). An acid-salt solution (4M NaCl, 2% formic acid) was added to the sample to give a final concentration of 200 mM NaCl, 0.1% formic acid. This step was required as the salt helps in efficient removal of peptides from MWCO filter and reducing surface adsorption of peptide while the acid helps in protonation of peptides. The filter was spun at 4,500 g for 10 min, which resulted in elution of tryptic peptides while all the higher molecular weight cellular debris was retained on the filter. Further desalting of the sample was done after loading it on to the chromatographic column.

Depending on the desalting procedure, peptides can be either loaded on a uni-phasic SCX only column (guanidine method) or a bi-phasic SCX-RP column (Urea method). The back column assembly on which peptides are loaded comprises of fused silica (150 μm inner diameter - i.d. 150, 360 μm outer diameter - o.d., Polymicro Technologies, Phoenix, AZ), two ferrules, one union and one in-line filter (Upchurch Scientific, Oak Harbor, WA). **(Figure 2.2)**

For each mass spec run, ~25 μg of peptides were loaded on the column. The back-column was first packed with a 3-4 cm of SCX resin (Luna 5 μm particle size, 100 Å pore size, Phenomenex, Torrance, CA) using a pressure cell. The pressure cell allows controlled loading of low volume solutions and LC resins from micro-centrifuge tubes into a fused silica column by application of high pressure in the range of 400-800 psi. The column was washed with Solvent A (95% H_20, 5% AcN, 0.1% formic acid) for few minutes if SCX was the only phase used in the back column. For the bi-phasic columns, the column was washed with methanol for few minutes, and then 3 cm of Reverse Phase resin (Aqua 5 μm particle size, 125 Å pore size, Phenomenex, Torrance, CA) was packed behind the SCX resin. Once the column was packed with SCX and RP resins, the column was washed with Solvent A (95% HPLC H_20, 5% AcN, 0.1% formic acid) and at last ~25-50 μg of peptides were loaded with the aid of a pressure cell. Once the peptides were loaded, the samples were washed for 45 minutes by first equilibration (15 minutes) using Solvent A and then five ramps of 100% Solvent A to 100% Solvent B (95% AcN, 5% H_20, 0.1% formic acid) at a flow rate of 200 μl/min. This step was crucial to remove trace SDS, urea and other interfering substances from the column.

The back-column with peptides loaded on a SCX or SCX-RP column was connected to a front-column, which was either an in-house pulled nanospray emitter (100 μm i.d. 150, 360 μm o.d., Polymicro Technologies, Phoenix, AZ) or a PicoFrit column (100 μm i.d., New Objective,

Waltham, MA). The front-column was packed with 12-15 cm of C18 resin (Aqua 5 µm particle size, 125 Å pore size, Phenomenex, Torrance, CA) using a pressure cell.

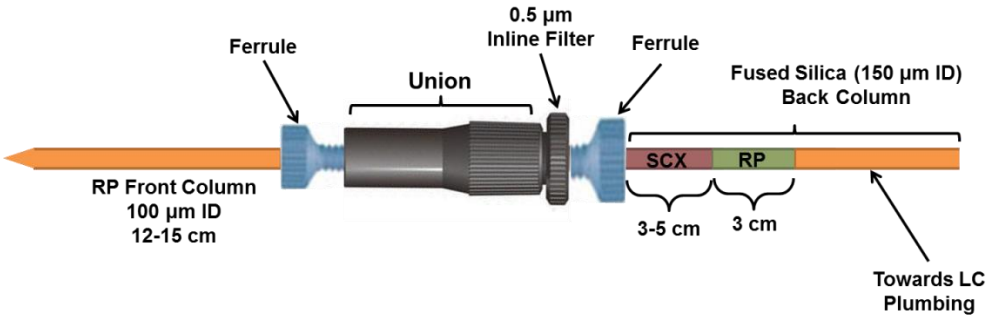


Figure 2.2 Schematic diagram of back-column assembly

2.2 Liquid Chromatography

All the MudPIT experiments described in this dissertation used high performance liquid chromatography (HPLC) with a U3000 quaternary HPLC pump (Dionex, San Francisco, CA). The HPLC pump and subsequent spectra acquisition both were under control of the Xcalibur software (Thermo Scientific). **Figure 2.3** provides a schematic of column assembly used in all the online 2D-LC-MS/MS experiments. HPLC solvent A and solvent B flowed from the pump into a 100 µm id fused silica, where a desired positive voltage (3-4.5 kV) was applied at the first Microtee junction with a gold electrode to help in generation of micrometer sized droplets from the electrospray tip.

The first Microtee junction was connected to a second Microtee junction via a 10 cm long fused silica column (100 µm id). The second tee junction split the HPLC flow into two streams, one

going into the waste via a smaller 50 µm i.d. fused silica creating back pressure and the second going into the back-column/front-column assembly which was interfaced to the mass spectrometer mounted on a nanospray source (Proxeon, Denmark).

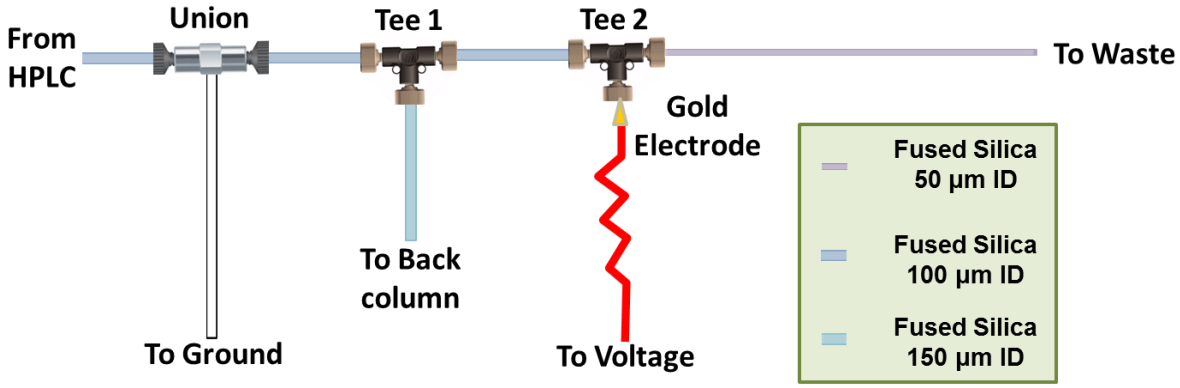


Figure 2.3 Schematic diagram of MudPIT plumbing used for online 2D-LC-MS/MS

Under the control of Xcalibur, a typical MudPIT experiment comprised of 11 salt pulses of increasing amounts of 500 mM Ammonium acetate (Solvent D). In most of the experiments discussed in this dissertation, either salt pulses with 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 60% and 100% Solvent D or 5%, 7%, 10%, 12%, 15%, 17%, 20%, 25%, 35%, 60% and 100% Solvent D were used for SCX separation. While the first scheme of salt pulses was mainly used with samples cleaned via Seppak, the latter was deemed fit for on-column cleaned samples. From previous experience in our lab, it was determined that short increments in step gradients at low percentage of ammonium acetate provides better resolution in chromatography

for peptides loaded on bi-phasic back columns. Each salt pulse except for the last one was followed by a 120 minute reverse-phase gradient from 100% solvent A to 50% solvent B. The last salt pulse used a 150 minute time period for reverse phase gradient to go from 100% A to 100% B.

The online scheme works by first shifting peptides released by a step-gradient of ammonium acetate from SCX resin in back-column to the long RP resin in front-column. These peptides are further separated along the increasing organic phase in the reverse-phase gradient. Thus at any one time only a subset of peptides are being measured by the mass spectrometer.

2.3 Ionization modes

The two main modes of ionization for transferring charged analyte from liquid or solid phase into a gas phase ion are Matrix-assisted Laser Desorption/Ionization (MALDI) and Electrospray Ionization (ESI) [9, 12]. MALDI employs the energy of laser beam to dislodge analyte on a solid surface and create gas phase ions, while ESI uses high voltage to desolvate analyte-containing liquid droplets to gaseous ions. While both the methods are widely used, each has its own advantages and disadvantages. The MALDI approach mainly generates ions that carry +1 charge, while the ESI generates multiply charged molecules. The MALDI scheme is more easily coupled with Time-of-Flight (TOF) instruments, while the ESI is more amenable to ion-trap instruments. Due to the instrument limitation associated with each ionization methods, MALDI is suited for intact protein measurements, as it has very high mass range. On the other hand, ion-traps equipped with on-line HPLC and ESI are more suited for the bottom-up proteomics research outlined in this dissertation.

2.3.1 Principle of Electrospray Ionization:

ESI is a soft ionization technique, in which the pre-charged analyte solution is passed through a very thin fused silica column held at high electric potential with respect to the entrance lens of the MS. The analyte solutions have an ionization agent which provides the charge to the analyte, for e.g. Na^+ or K^+ or in our case protons originating from the addition of formic acid. The presence of excess positive ions in solution leads to discharge of droplets that carry excess positive charge. So, as the positively charged analyte passes the emitter tip, it is aerosolized into a "Taylor Cone" [64] **(Figure 2.4).** During transmission along the Taylor Cone, the ions

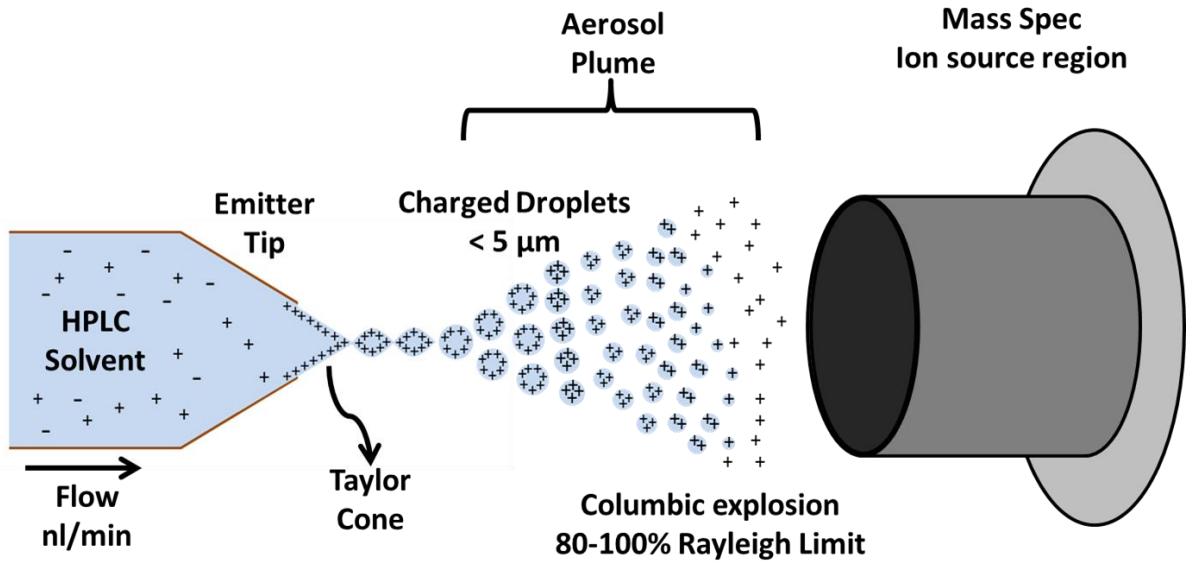


Figure 2.4 Generation of gas-phase ions in electrospray ionization

experience opposing columbic repulsion and surface tension forces within the liquid droplet. This fine stream of charged droplets is constantly losing solvent in flight due to evaporation and therefore shrinking in volume. Just before reaching the Rayleigh limit, where charge balances surface tension, these droplets further breakdown, giving rise to secondary droplets and the

process continues [65]. When these charged droplets reach 5-10 nm or less, they give rise to gas phase ions via either the "Charge Residue Model (CRM)" or the "Ion Evaporation Model (IEM)" or both [66].

According to the CRM model originally proposed by Dole, the sequence of evaporation and columbic explosion ultimately leads to an extremely small charged droplet containing only one analyte [67]. When the last droplet in this sequence loses its solvent, the residual charge is retained by the analyte in gas phase. As per the IEM model proposed by Iribarne and Thomson, the repetitive columbic fissions and concomitant solvent loss, ultimately leads to a very small droplet which has electric field strong enough to emit gas phase ions from its surface [68].

All the proteomics experiment carried out in this dissertation were performed by nanospray ionization which shares the same principle as electrospray ionization, except for the flow rate is reduced to the nL/min range from the μL/min range employed in ESI. This slower flow rate leads to reduction in droplet size and enhanced desolvation, which increases sensitivity for analyte measurement.

2.4 Mass Spectrometry instrumentation

2.4.1 Mass Analyzer:

The ESI mode of ionization perfectly aligns with wide variety of mass spectrometers that have ion-traps or quadrupoles as one of their mass analyzer. This combination includes LCQ, LTQ, Orbitrap and Velos series of instruments from Thermo Scientific. The linear trapping quadrupole or LTQ is one of the basic 2D linear ion trap mass spectrometers. The linear ion trap component of LTQ includes four precision-machined and aligned hyperbolic rods, making two pairs with one pair having slits to eject ions **(Figure 2.5)**. LTQ employs variable dc/rf potential to focus

ions and is therefore dynamic in nature. The quadrupole ion trap can hold or eject ions with desired m/z values by manipulating applied radio frequency (RF) and direct current (DC) potentials on the four rods. When the AC voltages applied to the rods become equal to the resonance frequency of the ion, which is dependent on its mass, the ion gains kinetic energy and thus can be ejected from the trap towards the ion detection system. The

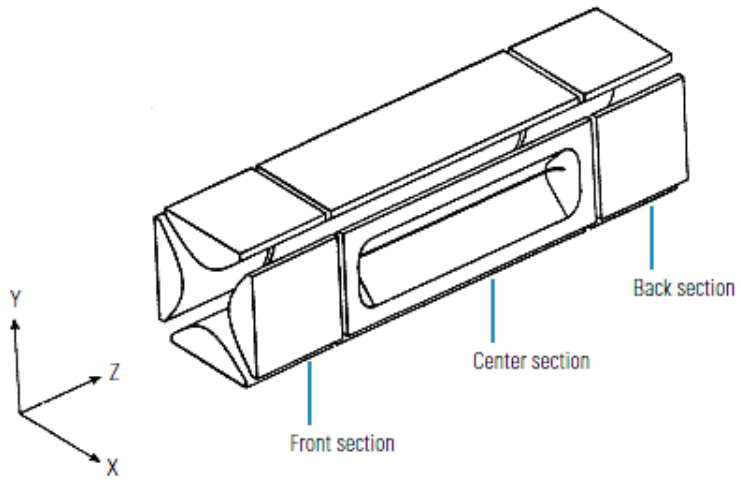


Image Source: Thermo Scientific LTQ Series Hardware manual

(http://www.thermoscientific.com/en/product/ltq-xl-linear-ion-trap-mass-spectrometer.html)

Figure 2.5 LTQ-XL Linear Trapping Quadrupole rod assembly

stability and trajectory of ions inside an ion-trap is described by Mathieu's equation, which is a second order differential equations given below

$$\frac{d^2u}{d\varepsilon^2} + (a_u - 2q_u cos2\varepsilon)u = 0 \qquad \text{Equation 2.1}$$

In equation 2.1 u represents x or y direction.

Solving this equation, gives two dimensionless parameters

$$a = \frac{4QU}{mr_0^2 2\omega^2}$$

where U is the DC voltage

and

$$q = \frac{2QV}{mr_0^2 \omega^2}$$

where V is the RF voltage. [69]

Both the Mathieu parameters also depend on mass to charge (*Q/m*) of the ion in motion, distance r_0 between the rods and the oscillation frequency ω.

Plotting a vs. q gives a stability diagram, which is a graphical representation of all the solutions to Mathieu's equation, and reveals that ions can occur in only two conditions inside a trap/quadrupole (a) periodic but unstable and (b) periodic and stable. As can be seen from **Figure 2.6**, ions have a very narrow zone in which they can remain stable inside the trap. Any offset to ac or dc voltage beyond the stability zone can either cause ions to hit the trap walls and be annihilated or lead to ejection from the slits in the traps.

LTQ instruments have two ion detection devices, each comprising of a conversion dynode which is concave metal surface and an electron multiplier that are placed orthogonally to the trap. The ions are ejected radially and hit the conversion dynode, emitting secondary particles. The concave surface of conversion dynode focuses these secondary particles and sends them to electron multipliers at high speed. When these high energy secondary particles strike the inside

of electron multipliers, they eject electrons. The ejected electrons further strike the multiplier surface, ejecting more electrons. The cascading effect results in a measurable current at anode which is directly proportional to the number of striking secondary particles at cathode. The data system records this current.

One of the newer instruments in the proteomics research is the LTQ Velos which is dual cell linear ion trapping instrument **(Figure 2.7)** [70]. The LTQ Velos features a novel ion-

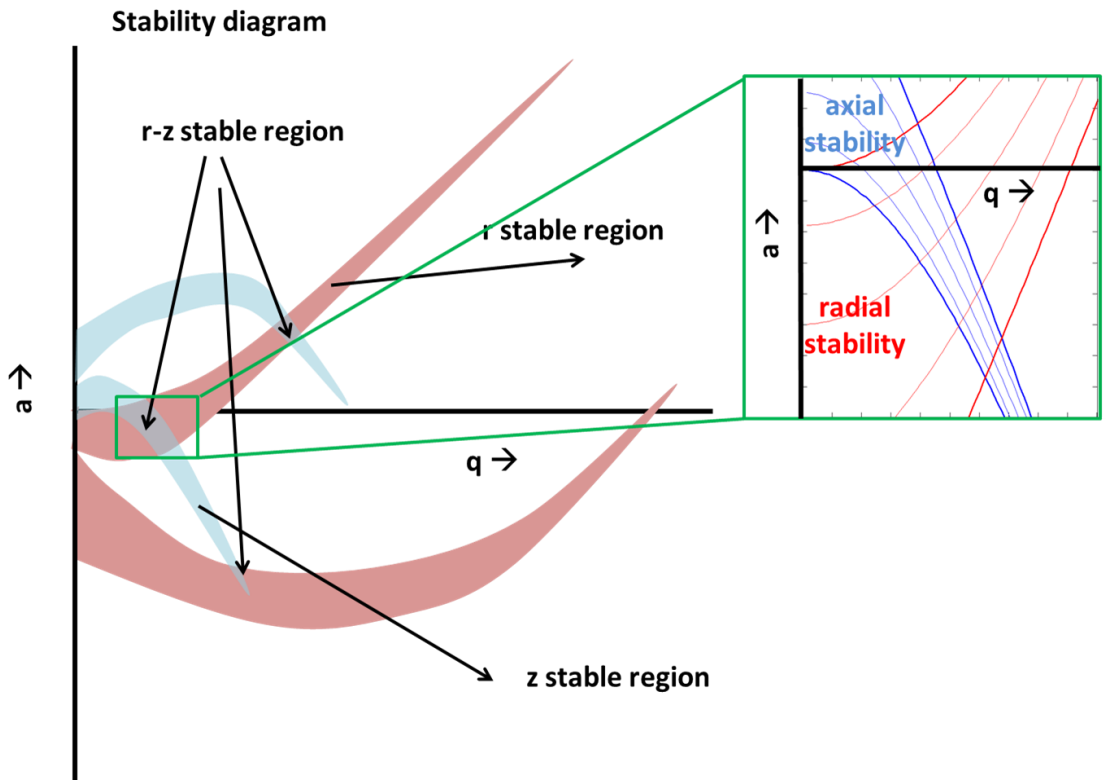


Figure 2.6 Stability diagram describing ion motion in an ion-trap

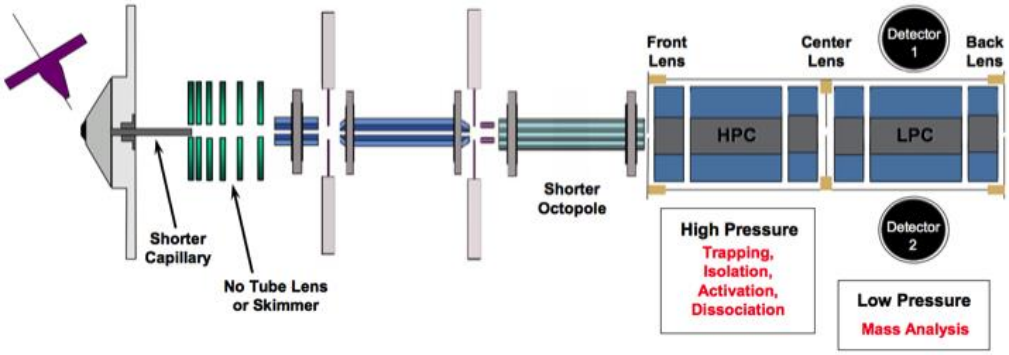
transmission pipeline compared to LTQ series instrument, thereby enhancing transfer efficiency. The dual traps, one operated at high pressure to improve ion trapping efficiency by 90% while the second trap coupled to ion detection system is operated at low pressure to enhance the measurement resolution.

The Orbitrap Mass Analyzer is a type of an ion trap in which ions are held in an electrostatic field and unlike conventional ion traps, there is no use of RF voltages for ion storage. [71] The Orbitrap mass analyzer shares the principle of Fourier Transform Ion Cyclotron Resonance (FTICR) mass analyzer, without using massive architecture to hold the superconducting magnet, which is the essence of FTICR. At the core of an Orbitrap is a system of an axially symmetrical mass analyzer, which consists of spindle shaped inner electrode enclosed by a barrel shaped outer electrodes. Ions travel from the linear ion trap into a curved gas filled ion trap (C-trap) **(Figure 2.7)**. The presence of nitrogen bath gas considerably slows down the ions and the ions are diverted orthogonally into the space between the outer and inner electrodes via a fine slot. Once inside the analyzer, ions oscillate around the central electrode due to the presence of applied electrostatic field.

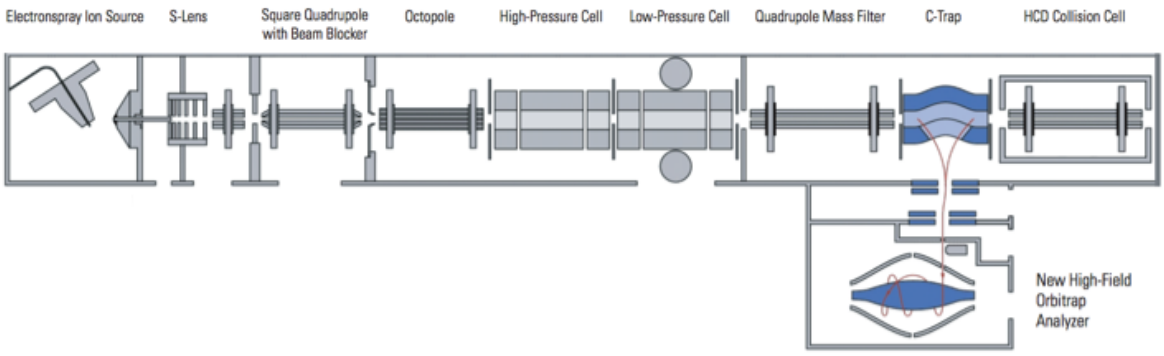
The motion of the ions around the central electrode is a simple harmonic motion that is explained by the following equation:

$$\omega = \sqrt{\frac{kq}{m}}$$

where ω is the frequency of axial oscillation, k is the potential between the electrodes (held constant), q is the charge and m is the mass of ion. Therefore, the frequency of ion oscillation is only dependent on the mass to charge ratio of the ion, and the coherent motion of ions carrying same mass to charge induces electric current that can be measured. This type of current measurement, called image current detection technique, was successfully used in FTICR mass spectrometry before the advent of Orbitrap analyzer. The image current is amplified by a differential amplifier and its output is converted from analog to digital. The digitized output is



(a) LTQ-Velos



(b) LTQ–Orbitrap Elite

Figure 2.7 Block diagram of LTQ-Velos and LTQ-Orbitrap Elite mass spectrometer.

(Image Source: Thermo Scientific LTQ-Series and Orbitrap Series Hardware manual, http://www.thermoscientific.com/en/products/liquid-chromatography-mass-spectrometry-lc-ms.html)

Fast-Fourier transformed, which converts recorded time-domain signal to a mass to charge spectrum.

The three instruments described above are among the primary work-horses in the field of mass spectrometry based proteomics. As the type of samples analyzed has become increasingly complex, demand for instruments that can measure ions with high sensitivity, high mass accuracy and high speed has also increased.

A more clear distinction between these analyzers can be made by looking at some of the performance metrics like mass accuracy, mass resolution, dynamic range and scanning speed. These metrics vary with each MS instrument and are crucial to know before embarking on a research project.

(a) Mass Accuracy:

This determines the ability of a mass spectrometer to measure as close as possible to the correct theoretical mass of analyte. It is measured in Da or parts per million (ppm).

$$Mass\ Accuracy = \frac{Observed\ m/z - Theoretical\ m/z}{Theoretical\ m/z}$$

(b) Mass Resolution/Resolving Power:

This is the ability of mass analyzer to differentiate between two neighboring peaks which have slightly different m/z values. It is a dimensionless quantity and typically is obtained by measuring peak width at 50% or 10% intensity. Mass resolution depends on the m/z of ion species in question and can be different for different analytes.

(c) Dynamic Range:

This is the ability of mass spectrometer to simultaneously measure the most abundant and the least abundant components in a sample, and is given as a ratio between the two. A high dynamic range means that the analyzer can distinguish between a wide range of ion abundance.

(d) Mass Range:

This is the difference between the largest and smallest m/z values over which the mass analyzer can accurately measure mass. While the MALDI instruments have very high mass range, ion trap instruments have limited mass range since there is only a narrow window of m/z values where applied RF frequency can contain the ions.

(e) Scanning speed:

This is the time taken by the mass analyzer to measure m/z values over a mass range.

(f) Sensitivity:

This is defined as the slope of a plot of analyte concentration vs. instrument response, which is generally listed as raw or relative intensity. A simpler assessment of instrument sensitivity is signal to noise ratio

Table 2.1 provides these performance metrics for the MS instruments employed for carrying out research in this dissertation.

Table 2.1 Performance metrics of MS instruments used in this dissertation

Parameter	LTQ-XL	LTQ-Orbitrap XL	LTQ-Orbitrap Elite
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
Mass Resolution	0.05 FWHM 1000-2000	7,500 - > 100,000 at m/z 400	15,000 - > 240,000 at m/z 400
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
Dynamic Range		>4,000 within a single scan guaranteeing specified mass accuracy	>5,000 within a single scan guaranteeing specified mass accuracy
MS/MS Sensitivity	25:1 Signal-to-Noise Ratio	100:1 Signal-to-Noise Ratio	100:1 Signal-to-Noise Ratio

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

FWHM = Full Width at Half Maxima

2.5 Data acquisition in mass spectrometry

Since a mass spectrometer measures charged species, irrespective of their origin, that have molecular masses within the mass range of analyzer, it is very important to operate the instrument in a mode that provides meaningful data rather than a random sampling of analyte that is given to the instrument. Data acquisition on a mass spectrometer can be performed in two ways (a) Data-dependent and (b) Data-independent mode. The two modes differ on the manner in which the mass spec is operated to carry out tandem scans after a full scan has been performed. In any given full or parent scan (MS1), there are a large number of peaks that can be subjected to dissociation for further sequencing. The data-independent mode works by allowing the instrument to carry out tandem scans (MS2) on each and every peak that is present in given m/z window. On the other hand, the data-dependent scan mode allows only a designated number of top N peaks in the MS1 spectra to be considered for MS2 scans. Each of the two methods has its own advantages and disadvantages. While the data-independent mode provides for complete sequencing of every available m/z value in parent scan, it has greatly reduced throughput. On the other hand, the data-dependent acquisition has fast duty cycle, as only top N peaks are sampled per MS1 where N is an integer between 5-20. By applying dynamic exclusion setting of 30 sec – 60 sec, peaks that have already been sampled for MS2 fragmentation are not considered again till the duration of filter. In this way, mass spec is able to sample more peaks on chromatographic time scale, but even then there is a loss of information for some of the less abundant peaks. On Thermo Scientific instruments, these two modes are employed via Xcalibur software package. In this dissertation, all the mass spectrometry runs were performed by data-dependent acquisition, using an isolation width of 0.5 m/z and dynamic exclusion of 0.02 Da for Orbitrap measurement and 1.5 Da for ion-trap measurement.

2.6 Tandem Mass Spectrometry and peptide sequencing

As mentioned above, the data-dependent mode allows for tandem mass scanning on only the most abundant peaks from the parent scan. While the parent scan measures the mass to charge ratio of the intact peptide, the tandem scan gives information about the composition of the peptide. This is achieved by fragmenting the intact peptide and then measuring the m/z values of all the fragment ions. By reconstituting the fragment ions, we can deduce the original sequence of the peptide ion. The most common approach of fragmenting peptides is by colliding fast moving ions with a neutral target gas such as nitrogen or helium in a process termed collision-induced dissociation (CAD) or collision-activated dissociation (CAD). CAD is a low energy dissociation method (1 – 100 eV) which involves transfer of some of the translational energy of fast moving ion into internal energy following an inelastic collision with the target gas, thereby leading to its dissociation[72]. Since CAD is a low energy approach, a single collision may not generate enough internal energy to cause dissociation and therefore multiple collisions with the target gas are required to achieve fragmentation. Protonated peptides follow charge-directed or charge-remote fragmentation pathways, as explained by the mobile proton model [73]**.** As per this model, peptides exists in a heterogeneous population with protons residing on multiple locations, preferable sites being the amino-terminal and side chains of positively charged amino acids like arginine, lysine, and histidine. When the proton is sequestered from these favorable sites to less favorable sites like on the peptide backbone, it can lead to weakening of the amide bond and subsequent fragmentation. In charge-directed fragmentation, the proton is transferred to amide carbonyl oxygen on the peptide backbone. This makes the electropositive carbon of the protonated carbonyl susceptible to nucleophilic attack by nearby carbonyl, leading to dissociation at the site of proton. This type of fragmentation gives different degrees of backbone

cleavage, with majority of fragment ions being *'b'* and *'y'* type **(Figure 2.8)**. On the other hand, in charge-remote fragmentation, as the name suggests, the cleavage occurs at a site farther from the proton. This type of fragmentation requires higher energy, with peptide charge being less than or equal to number of arginine residues in the sequence [73].

Since most of the proteomics experiments employ trypsin digestion, it results in generation of large number of doubly charged peptide species, which carry mobile protons and therefore are

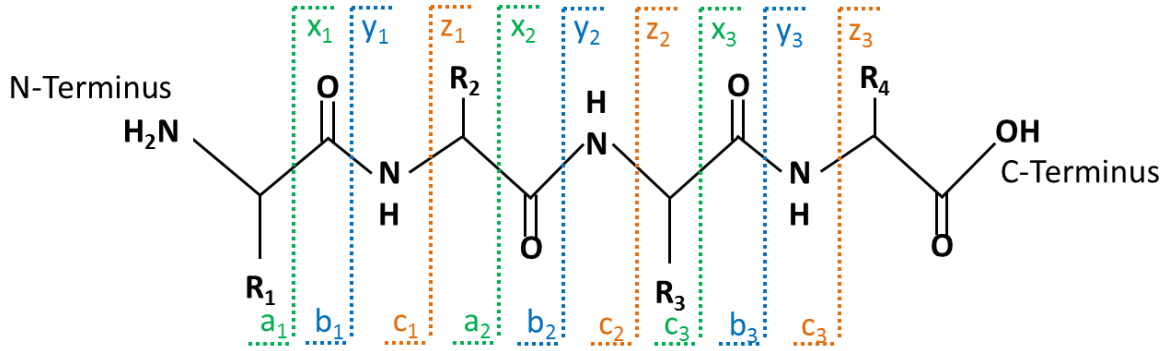


Figure 2.8 Type of fragment ions produced via peptide backbone cleavage

more amenable to charge-directed fragmentation pathway giving rise to *'b'* and *'y'* ions. The peptide cleavage can occur on multiple locations on the peptide backbone and depending on which part of the peptide retains the charge, can be classified as a, b, c type of ions in which charge is retained on the N-terminus, or they can be named x, y, z type of ions if the charge is retained on the C-terminus **(Figure 2.8)**.

A second form of collision type fragmentation method is High energy Collision Dissociation (HCD), which is beam-type CAD dissociation available in Orbitrap instruments. In HCD, the fragmentation takes place in a dedicated collision cell at the back end of the mass spectrometer instead of the ion-trap [74]. Following dissociation, ions are transported back into the Orbitrap

mass analyzer for measurement. Compared to CAD mode, HCD offers high resolution and high mass accuracy peptide sequencing, since the measurements are carried out in the Orbitrap but it takes a hit in the duty cycle due to slower speed. HCD mode employs relatively higher energy for fragmentation than CAD and therefore provides greater sequence coverage. One other mode of peptide fragmentation is Electron Transfer Dissociation (ETD) which will be discussed in Chapter 5.

2.7 Database searching of MS/MS data

The last step in a proteomics experiment is to sequence and identify peptides that are measured by the mass spectrometer. To do so, tandem spectra acquired by mass spec are computationally matched against the predicted proteome of the species under investigation. The predicted proteome is a fasta formatted database of amino acid sequences of total genes irrespective of their expression or activity level present in the genome. Therefore, for any proteomics experiment to provide deep coverage, the genome of species under study must be completely sequenced. The first step in database searching is to do *in-silico* digestion of proteins using trypsin. Next theoretical spectra are generated for each *in-silico* generated tryptic peptide with the *'b'* and *'y'* fragments having 100% intensity value on y axis and fixed m/z value on x-axis. Once this is done, then it is a simple case of pattern matching where experimental spectra are matched against theoretical spectra. Each matched spectra is given a cross-correlation score called XCorr, and is assigned a value called deltCN that reflects how distinct it is to the next best match. If both the XCorr and deltCN values pass a user defined threshold, the peptide-spectrum match (PSM) is retained; in cases where these criteria are not satisfied, the PSM is rejected. Each PSM matches to a single peptide and a computational program assembles these peptides into proteins. The most commonly used search programs that carryout peptide sequencing to find

confident PSMs includes SEQUEST, Mascot, and MyriMatch [75-77]. To ascertain the accuracy of mass spec measurements and database search methods, a reverse database or shuffled database is generally concatenated to the forward database. Although using both forward and reverse databases in search increases total computational time it helps in removal of false positives and provides a false discovery rate for the proteomics measurement, which is valuable in ascertaining quality of both MS data and the predicted proteome [78].

The FDR can be calculated using the following equation,

$$\frac{2 * False\ Positives}{True\ Positive + False\ Positives} * 100$$

in which false positives can refer to matched reverse proteins, peptides or spectra and true positives means identified forward proteins, peptides or spectra. A factor of 2 is added to reverse hit as they are false hits and therefore any match to them should penalize FDR.

Once the peptides are sequenced, the next computational task is to get a list of protein identifications. DTASelect and IDPicker are two commonly used programs that carry out this task [79, 80]. Within a user specified threshold for FDR and minimal requirement for protein call like a 1 peptide hit which requires matching of 1 unique peptide within a protein will suffice for positive identification or a 2 peptide hit which require matching of 2 unique peptide from a protein for positive identification, these program assemble peptides back at protein level and report them.

Chapter 3 - Coupling a Detergent Lysis/Cleanup Methodology with Intact Protein Fractionation for Enhanced Proteome Characterization

---

Text and figures were taken from: **Sharma R**, Dill BD, Chourey K, Shah M, VerBerkmoes NC and Hettich RL. Coupling a detergent lysis/cleanup methodology with intact protein fractionation for enhanced proteome characterization. *Journal of Proteome Research.* 2012**(11)** 6008-6018

Ritin Sharma's contributions included: Experimental design, performed all the proteomics sample preparation and mass spectrometry runs, data analysis, wrote, edited and revised the manuscript.

3.1 Application of detergents in proteomics sample preparation

Current research in environmental microbiology is heavily focused on a more comprehensive understanding of what microbial species are present in specific ecosystems, how they co-exist and cooperate/compete for resources, the range of their genetic potential, and how they thrive in their natural habitats [33]. This is driven by two primary biological factors; 1) microbes do not exist in isolation and, thus, a community-level understanding gives a more representative picture of their metabolic activities as compared to an isolated culture in a lab; [31] and 2) many of the microbes found in community samples are not amenable to culture under lab conditions [81].

To this end, Liquid Chromatography-Mass Spectrometry (LC-MS)-based proteomics has become a powerful method for characterization of global protein changes, either at the single organism or at the consortium/community level (referred to as metaproteomics) [31]. This approach provides

an unprecedented level of biological detail that has not been attainable with any other approach. For example, early studies have provided in-depth knowledge of protein abundance for numerous model organisms, such as *E. coli, Saccharomyces cerevisiae,* or *Drosophila melanogaster* [30, 82, 83]. Recent work in this field has shifted towards analyzing more complex microbial communities, such as those found in acid-mine drainage systems, oceans, and even human gut microbiota [36, 84, 85]. By taking a global approach to investigate community metabolic function, metaproteomics can reveal gene product information for the active community members, and can highlight pathways that are either activated or deactivated by environmental factors. More broadly, proteome research, whether focused on microbes, eukaryotic tissues, or cell lines, is heavily impacted by numerous experimental factors (protein biomass, preparation method, instrumentation, and proteome bioinformatics techniques) that need to be carefully evaluated for optimum proteome coverage and depth. Key challenges for metaproteomics include intractable matrices, low microbial biomass, interference from compounds within the environmental sample (such as humic acids in soil), strain variation, and inter-species and intra-species similarities. To overcome these challenges, efficient ways of extracting clean proteomes from such samples is critical and therefore, traditional sample preparation methods for microbial isolates have to be either modified or completely replaced by more advanced methods for microbial communities.

Detergents and surfactants have been used extensively in the field of protein biochemistry [86]. Over the last few years, detergents and surfactants have been heavily employed in proteomic sample preparations due to the high efficiency with which they disrupt membranes and solubilize proteins. One of the most commonly used detergents is sodium dodecyl sulphate (SDS), which is very efficient in disrupting membranes and solubilizing a wide range of protein types, as

evidenced by applications in polyacrylamide gel electrophoresis for GeLC-MS experiments [87]. Other detergents, such as sodium-3-[(2-methyl-2-undecyl-1,3-dioxolan-4-yl)-methoxyl]-1-propanesulfonate (RapiGest) and 3-[3-(1,1-bisalkyloxyethyl)pyridin-1-yl]propane-1-sulfonate (PPS), have also been evaluated for proteomics sample preparation [88]. While the use of SDS effectively facilitates cellular lysis, it poses a major impediment for LC-MS/MS experiments. In particular, the presence of SDS negatively affects the efficiency of trypsin digestion and hinders the resolving power of reverse phase liquid chromatography [89, 90]. The introduction of SDS into the mass spectrometer during electrospray can lead to a variety of problems, such as ion suppression or accumulation inside the ion source [91].

To increase the compatibility of SDS with LC-MS/MS based experiments, numerous methods have been proposed to remove SDS from a sample prior to LC-MS analysis. These methods include alternate chromatography techniques, such as hydrophobic interaction chromatography, size exclusion chromatography or ion exchange chromatography, or organic solvent precipitations like ethyl acetate extraction, acetone precipitation, Trichloro acetic acid (TCA) precipitation, or Chloroform/Methanol/Water (CMW) precipitation. Recent methodologies to remove SDS include the Filter-aided Sample Preparation (FASP) [92] approach, a spin-column-based detergent removal (DRS) [93], and strong-cation exchange (SCX) liquid chromatography [94].

A related and also critically important aspect to consider during sample preparation using detergent based lysis/solubilization is the starting amount of protein in the sample, which often dictates the choice of sample preparation method that can be used. To our knowledge, there is very limited information about the efficacy of SDS lysis/solubilization and cleanup at varying protein amounts. The FASP method has been evaluated and compared to SDS-PAGE and tri-

fluoroethanol methods at low (150 ng) and high (50 µg) protein amounts [95]. In an effort to extend this work, we sought to compare four commonly used SDS removal methods with respect to starting protein amount. The four SDS removal methods evaluated here include TCA precipitation followed by cold acetone wash, CMW precipitation, DRS, and FASP. Each of these methods has specific advantages for proteomic sample preparation [96, 97]. The TCA precipitation protocol is inexpensive, straightforward, and can be used for diverse types of samples. In fact, the applicability of TCA for protein extraction has been demonstrated in complex biological samples such as soil [98]. The major disadvantage of this approach is that a reasonable amount of protein starting material is needed to obtain a visible and stable pellet during the wash steps. The CMW precipitation protocol has been used extensively for lipid extraction [99] and has been employed for protein extraction [100, 101]. However, this approach requires somewhat more complicated sample handling (precise pipetting/extraction) and involves use of a hazardous chemical (chloroform) that is difficult to use and dispose of in many standard laboratories. The DRS method employs a proprietary resin for one-step detergent removal in a spin column format [93]. A recent study by Bereman *et al.* evaluated a modified DRS protocol versus FASP, and, under the conditions tested, found the former to give higher protein identification compared to FASP [102]. Note that this study used a low concentration of SDS (0.1%), and a single protein concentration was tested. Additionally, detergent removal was carried out differently for the two methods: at the protein level for FASP and at the peptide level for DRS. The FASP method is a recent protocol which uses molecular weight cut-off (MWCO) filters to capture proteins and remove incompatible MS reagents, such as SDS and urea. The filter acts as a mini reactor for efficient trypsin digestion and enables simple peptide elution. The method has been successfully used for mammalian cell cultures, yeast, and paraffin embedded

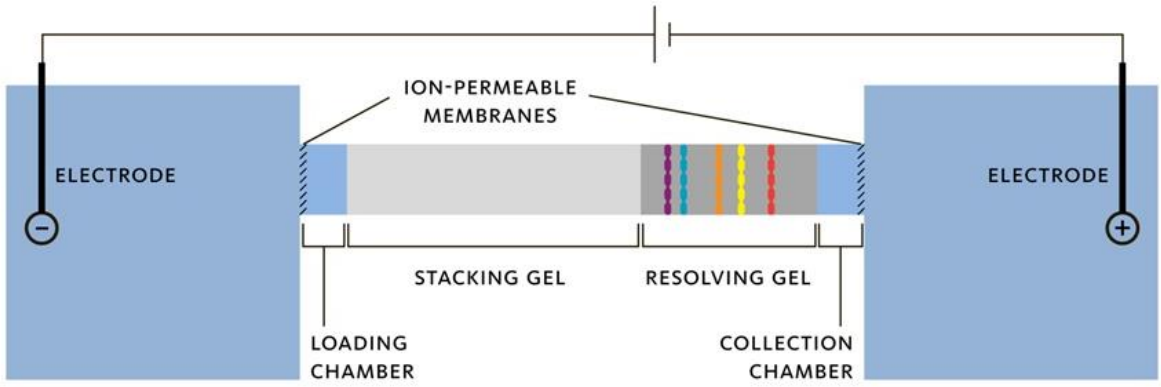
tissue samples with remarkable results, but little information is available for its performance with microbial systems [92, 103, 104]. Therefore, we evaluated the performance of these four SDS clean-up methods on *E. coli,* specifically varying the starting amount of protein from 10 µg to 1 mg.

3.2 Introduction to intact protein fractionation – The GELFrEE approach

The increased lysis/solubilization efficiency of the SDS-based lysis approach generates a complex sample rich with a variety of proteins, and thus it is natural to consider coupling this approach with an additional separation methodology that can enhance proteomic measurements. For example, reduction of complexity at the protein level can be accomplished by chromatographic fractionation prior to proteolysis, generating fractions which can then be incorporated into the typical two dimensions of an on-line chromatographic separation at the peptide level. The fractionation process can enhance the measurement of low abundance proteins from the mixture by providing fractions with decreased complexity, thereby giving these proteins more opportunity to be identified in the LC-MS experiment. Protein fractionation typically targets the physicochemical properties of molecular weight or isoelectric point. The most commonly used molecular weight-based separation of proteins is by SDS-PAGE and proteins can be further separated by pI in 2D-PAGE. Proteins are thus separated into discrete bands/spots, which are then cut out and digested, a process which is commonly referred to as GeLC-MS [105]. However, GeLC-MS has some known complications, including incomplete recovery of peptides from gel slices, added potential for keratin contamination, and added sample handling steps of in-gel digestion techniques. A recent development for extending proteome coverage is molecular weight-based fractionation using Gel-eluted Liquid Fraction Entrapment Electrophoresis or GELFrEE technology (Protein Discovery), which yields liquid fractions

collected at pre-defined times **(Figure 3.1)** [106]. This technology has shown great promise in high throughput top-down proteomics [107].

To further develop advanced separation methods for enhanced proteome measurement depth in microbial samples, we systematically evaluated the efficacy of complementary molecular weight based in-solution fractionation of proteome samples using the GELFrEE device. We analyzed three different sample types with increasing complexity (*E. coli*, a five microbial isolate mixture (5MM) and a natural microbial community from an environmental ground water sample) via LC-MS/MS based proteomics measurements of whole cell lysates, in conjunction with their intact protein fractionated counterparts. (**Figure 3.2**)



(Image Source: www.expedeon.com/Portals/0/product%20manuals/G8100_Manual_v1-4.pdf)

Figure 3.1 A schematic diagram of GELFrEE fractionation system

The sample is loaded into the loading chamber and the voltages are applied which results in migration of proteins based on their molecular weight. The low molecular weight proteins elute first into the collection chamber followed by proteins with high molecular weight. The instrument is paused at pre-defined specific time intervals to collect liquid fractions from the collection chamber.

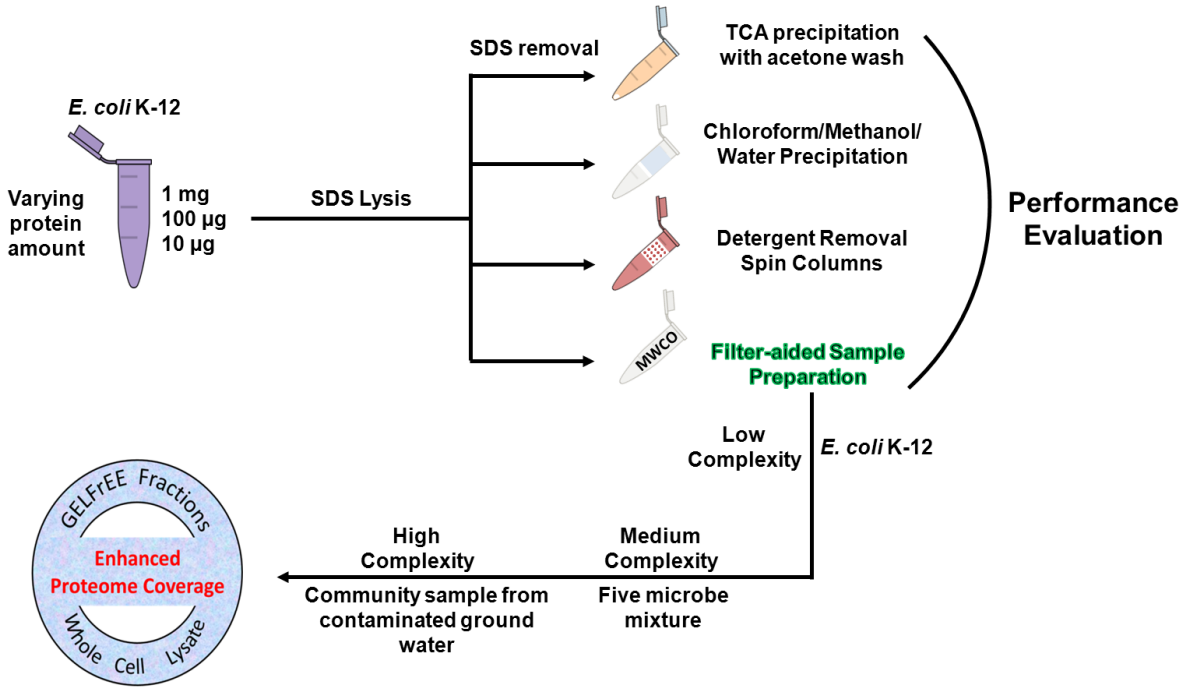


Figure 3.2 A schematic overview of the experimental design used in this study.

E. coli K-12 with three different protein concentrations were lysed by SDS and the SDS was removed with four different detergent removal methods. Next, the advantage of detergent based lysis and removal was utilized by combining a MW based GELFrEE fractionation and unfractionated analysis scheme on three increasingly complex biological samples.

3.3 Materials and Methods

Microbial samples

The bacterial strains *Shewanella oneidensis* MR-1, *Shewanella putrefaciens* CN-32, and *Pseudomonas putida* F1 were cultivated aerobically under constant agitation (250 rpm) at 30°C in Luria-Bertani (LB) medium (pH 7.2). The bacterial strain of *E. coli* K-12 was cultivated aerobically under constant agitation (250 rpm) at 37°C in LB medium (pH 7.2). When each

culture reached an O.D. of ~0.8, it was centrifuged (1000 rpm for 10 min), and the resulting pellet washed with PBS. The pellet was flash frozen in liquid nitrogen with long term storage at – 80°C. A cell pellet of *Geobacter uraniireducens* was a generous gift from Dr. Derek Lovely, University of Massachusetts, Amherst, MA. *Rhodopseudomonas palustris* CGA010 cell pellet was graciously donated by Dr. Dale Pelletier, ORNL. The ground water microbial community sample was obtained from the DOE Integrated Field Research Challenge site at Rifle Colorado, provided by Dr. Kenneth Williams (Lawrence Berkeley National Lab).

Cell lysis

Protein Biomass Study: The protein concentration for cell cultures and the ground water sample were determined via an RC DC protein assay (Bio-Rad). A small aliquot from each stock culture was lysed by SDS-containing GELFrEE buffer with 50 mM DTT, and the RC DC assay was performed on the subsequent lysate. From this, a total protein amount in each stock culture was determined, and aliquots of *E. coli* K-12 corresponding to 1 mg, 100 µg or 10 µg total protein amount were removed. These aliquots were lysed using GELFrEE Sample Buffer (10% SDS starting concentration, 2% effective SDS concentration in solution) and 50 mM DTT by heating at 60 °C for 10 minutes. A similar procedure was adopted for the 5MM sample and Rifle ground water sample used in the study.

Fractionation and whole cell lysate complementary study: 500 µg of *E. coli* K-12 was lysed using SDS containing GELFrEE sample buffer (final concentration 2% SDS) and 50 mM DTT by heating at 60 °C for 10 minutes. For the 5MM sample, five microbes i.e. *P. putida* F1 (50%, 500 µg), *R. palustris* CGA010 (25%, 250 µg), *S. oneidensis* MR1 (20%, 200 µg), *S. putrefaciens* CN-32 (4%, 40 µg) and *G. uraniireducens* (1%, 10 µg) were combined (total protein amount of

1 mg) after cellular lysis via sonication. Following combination, cells were solubilized in GELFrEE sample buffer and 50 mM DTT and heated for 10 minutes at 60 °C (total volume 300 µL, of which 200 µL was taken up for LC-MS/MS experiment). The ground water sample corresponding to 500 µg total protein was lysed using GELFrEE sample buffer and 10 mM DTT and heated at 60 °C for 10-15 minutes.

GELFrEE fractionation

The 500 µg samples of *E. coli* K-12 and ground water were loaded onto a single channel of a GELFrEE 8100 mid-mass cartridge after cellular lysis. The lysate from 1 mg of the 5MM sample was divided equally and loaded into two channels of GELFrEE 8100 mid-mass cartridge. Following this, 12 fractions (~ 150 µL each) were collected at specified time intervals (57.5, 59.5, 61.5, 64.5, 66.5, 68.5, 71.5, 76.5, 83.5, 93.5, 108.5, 128.5 min). Fractions from the two channels of 1 mg of 5MM sample were pooled (total 300 µL), and 200 µL was used for further downstream sample preparation, while the remaining volume was kept for analysis via SDS-PAGE. For the fractions from the *E. coli* K-12 and ground water, all the fraction volume (150 µL) was used for downstream sample preparation.

SDS clean-up procedures

TCA precipitation: TCA (100%) was added to 150 µl of SDS lysed *E. coli* K-12 cells to a final concentration of 25% (w/v) TCA. The TCA/protein solution was kept overnight on ice at 4 °C, then the precipitate was centrifuged at 20800 *g* for 10 min. The supernatant was discarded and the pellet was washed twice with chilled acetone by centrifugation at 20800 *g* and finally air-dried. The air-dried and acetone washed pellet was re-suspended in 6M guanidine.

Chloroform/Methanol/Water precipitation: The method was an adaptation from the protocol by Wessel and Flugge [100]. To the 1 mg, 100 µg and 10 µg *E. coli* K-12 lysate (150 µl volume), 600 µL of methanol, followed by 150 µL chloroform and then 450 µL water was added with brief vortexing after each solvent addition. The sample was centrifuged for 15 min at 20800 *g* and the top layer was removed. A further 600 µL aliquot of methanol was added without vortexing, and spun at 20800 g for 15 min. Following this step, the supernatant was carefully removed without disturbing the pellet. The pellet was air-dried and re-suspended in 6M guanidine.

Detergent removal spin column: The 150 µl of *E. coli* K-12 lysate was loaded on to a 0.5 mL column size Pierce detergent removal spin column and used per manufacturer's guidelines. Briefly, SDS lysed samples were loaded on the spin column and incubated for a couple of minutes at room temperature. Finally, the SDS-free sample was eluted by spinning the column for 2 min at 1,500 *g*.

Filtration/FASPKits: To test the filter assisted sample preparation method we digested *E. coli* K-12 lysates via a commercial FASP^{TM} Protein Digestion kit (Expedeon Inc., San Diego, CA), as well as a method similar to the original paper describing use of spin filters [95]. *E. coli* K-12 lysate (150 µl) was loaded on to a 10kDa MWCO filter (VWR centrifugal filters with Polyether sulfone membrane), and 200 µL of 8M urea was added on top of the filter, which was then vortexed and centrifuged at 20,800 *g*. The urea wash step was repeated twice. Following urea spins, 200 µL of 100mM Tris; 10 mM $CaCl_2$ buffer (pH 8) was added. The filter was again vortexed and centrifuged at 20,800 *g*. This step was repeated twice. The sample was then subjected to trypsin digestion. The commercial FASPKits employed a 30kDA filter and use a 50 mM ammonium bicarbonate buffer in place of the Tris-$CaCl_2$ buffer used for VWR filters. The

filters were operated as per the protocol for handling GELFrEE fractions described in the instruction sheet.

Protein Digestion: Except for samples cleaned up using the Filtration/FASPKits method, protein lysates from other cleanup methods (i.e. TCA, CMW and DRS method) were subjected to chemical denaturation by 6M guanidine and reduction by10mM DTT. Following this, the samples were diluted 6 fold and trypsin (Promega, Madison, WI) was added (1:20 Trypsin: Sample ratio). Digestion was carried out for 4 hours at 37°C, followed by addition of another aliquot of trypsin for overnight digestion at 37°C. Samples were desalted using C-18 Solid Phase extraction (Sep-Pak Lite, Waters). *E. coli* K-12 samples cleaned up via the filtration method (protein biomass variation study) were digested on the MWCO filter using two aliquots of trypsin (first for 4 hours and then for overnight) at room temperature. The peptides were eluted by 0.5 M NaCl and desalted using C-18 Solid Phase extraction (Sep-Pak Lite, Waters). The FASPKits samples (fractionation and whole cell lysate complementary study) were digested on the MWCO filter using two aliquots of trypsin, described as above for filtration prep at room temperature. The resulting peptides were collected in a fresh tube by eluting them with 0.5 M NaCl. The FASPKit samples were desalted on the column by loading the sample on Reverse Phase (RP)-Strong Cation Exchange (SCX) back column and washing from high water to high organic three times over 25 minutes.

LC-MS/MS

Digested samples were pressure cell-loaded onto a 3 cm SCX only fused silica back column (150-µm i.d.) back column. For the FASPKits prepped samples, 3 cm of C18 reverse phase resin (Aqua, 300 Å pore size, Phenomenex) was added up-stream from the SCX resin. The back

column was connected to a 15-cm-long 100-μm-i.d. C18 RP resin PicoFrit column (New Objective) and placed in-line with a U3000 quaternary HPLC (Dionex, San Francisco, CA). The SCX LC separation was performed with either eleven salt pulses (peptides from whole cell lysates, protein biomass studies with 100 μg and 1 mg E. coli), six salt pulses (peptides from 10 μg E. coli samples) or three salt pulses (peptides from GELFrEE fractions) containing increasing concentration of 500mM ammonium acetate. Each salt pulse was followed by a 2 hr. reverse phase gradient from 100% Solvent A (95% H20, 5% AcN and 0.1% formic acid) to 60% Solvent B (30% H_20, 70% AcN and 0.1% formic acid). The LC eluent was directly nanosprayed into either a Thermo Scientific LTQ Orbitrap-XL mass spectrometer (protein biomass variation study) or a LTQ-XL mass spectrometer (GELFrEE fractions and unfractionated lysate study) with an ionization voltage of 4-4.2 kV. During LC, the mass spectrometer was operated in data-dependent mode and under the control of the Xcalibur software (Thermo Scientific). The data-dependent acquisition used the following parameters: collision-induced dissociation of 5 parent ions were performed following every full scan; 2 micro-scans were averaged for every full MS and MS/MS spectrum; a 3 m/z isolation width; 35% collision energy was used for fragmentation; and a dynamic exclusion repeat of 1 with duration of 60 seconds.

Peptide and protein identifications

The raw MS/MS data from *E. coli* K-12 samples was searched against an *E. coli* K-12 predicted database having 4170 proteins (downloaded January 20, 2011 from Department of Energy-Joint Genome Institute (DOE-JGI) server) including common contaminants (trypsin, keratin etc.). The raw MS/MS data from the 5MM runs was searched against a concatenated database (23098 proteins) of the predicted proteomes of *P. putida* F1, *S. putrefaciens* CN-32, *S. oneidensis* MR-1, *R. palustris* CGA009, *G. uraniireducens* and common contaminants (downloaded January 8,

2008 from DOE-JGI server). The ground water samples were searched against a microbial database (totaling 30758 proteins plus common contaminants) assembled from eight *Geobacter* isolates (downloaded January 1, 2008 from DOE-JGI server) [40], the dominant microbial species in these samples [40]. All searches were done by SEQUEST using search parameters described previously [108]. The output DTA files were then filtered and sorted using DTASelect algorithm using the following parameters: fully tryptic peptides only with ΔCN of at least 0.08 and cross-correlation scores (Xcorr) of at least 1.8 (+1), 2.5 (+2), and 3.5 (+3). FDR were calculated using a concatenated forward-reverse database on selected runs.

Bioinformatics

The whole proteomes of all the microbes (i.e., of *P. putida* F1, *S. putrefaciens* CN-32, *S. oneidensis* MR-1, *R. palustris* CGA009, *G. uraniireducens* and *E. coli* K-12) were analyzed by PSORTb v3.0.2 [109]. Pathway analysis was done by Pathway Tools v15.0 [110]. For the analysis, we imported Ecocyc [111] and other databases (PGDBs) corresponding to microbes involved in our study into Pathway tools. Following the import, Gene IDs corresponding to protein identifications from the MS/MS search were mapped on the pre-built pathways. For the GRAVY analysis, protein sequences corresponding to protein identifications from MS/MS data were imported into the PROMPT tool (Protein Mapping and Comparison Tool) [112]. All the Venn diagrams were made using Venny (http://bioinfogp.cnb.csic.es/tools/venny/index.html) or Venn Diagram Plotter (PNNL; http://omics.pnl.gov).

3.4 Effect of protein amount on efficacy of different detergent clean-up methods

In order to evaluate the performance of the detergent removal methods, SDS-lysed *E. coli* K-12 at different protein amounts were processed with four commonly used approaches: TCA

precipitation, CMW method, DRS method and FASP **(Figure 3.2)**. After determining the protein concentration of the starting *E. coli* K-12 culture, aliquots of 1 mg, 100 µg and 10 µg total protein were lysed for proteomics experiments.

Based on the results, the FASP method performs slightly better at the high protein (1 mg) amount and strikingly superior relative to the other three methods at the low protein conditions (10 µg) The four-way Venn diagram in **Figure 3.3** shows unique and common proteins identified after combining results from the two technical replicates of *E. coli* K-12 lysates with varying protein amount and employing different detergent clean-up methods. At the highest protein amount tested, 1 mg of total protein, ~51% of the identifications are common to all the methods; this number decreases to ~45% with 100 µg total protein and then takes a sharp dip to just 10% with 10 µg total protein. For the 10 µg *E. coli* K-12 lysate, the number of proteins uniquely identified from the DRS and the CMW methods are insignificant compared to the unique proteins identified by the FASP method. Surprisingly, the TCA precipitation protocol failed to identify a single unique protein at the 10 µg level.

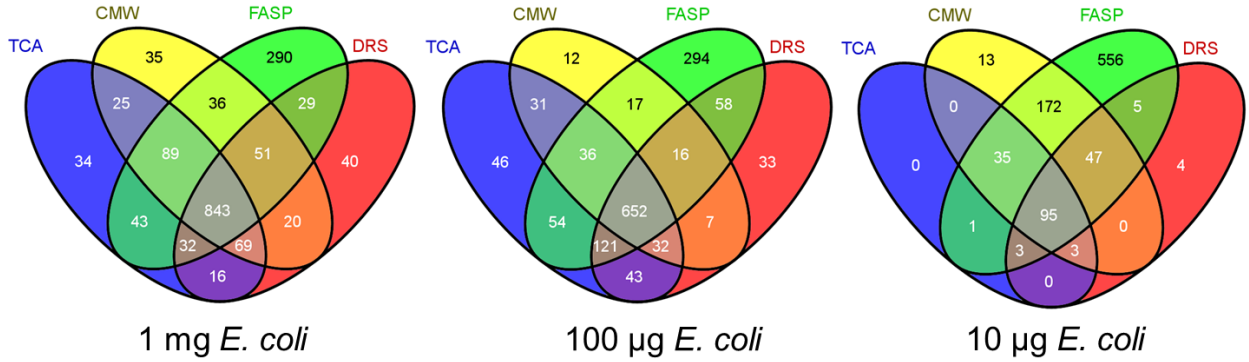
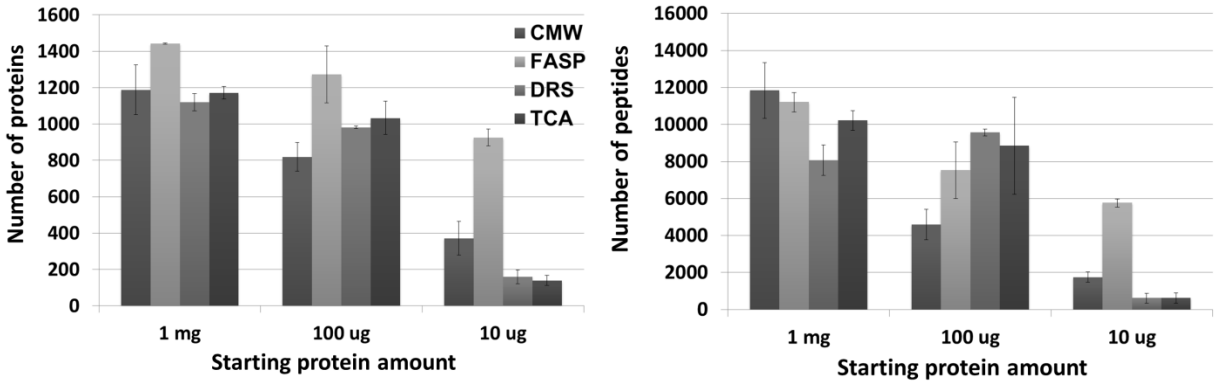
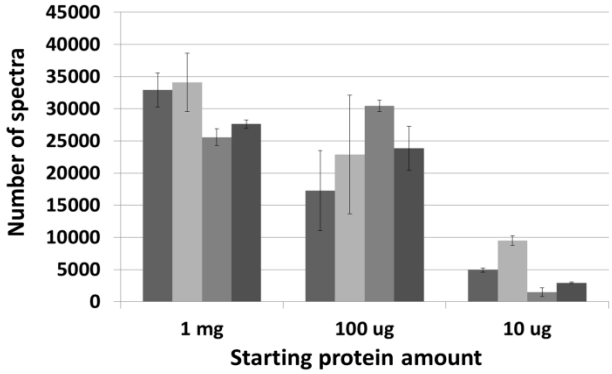


Figure 3.3 Unique and common proteins identified in *E. coli* K-12 lysate after SDS lysis / removal using four different methods and three different protein amounts

The number of peptides and spectra identified by each SDS clean-up method showed a similar trend to the protein identifications, except for a few variations. As shown in the **Figure 3.4a** and **3.4b**, in the case of the DRS method, as expected, the average number of proteins is lower for a 100 µg sample than a 1 mg sample, but surprisingly the average number of peptides follows the reverse trend. The 100 µg sample of the *E. coli* K-12 prepped by the DRS method yields more peptides compared to a 1 mg protein sample. The number of proteins, peptides and spectra identified by the FASP method shows an almost linear trend for differing protein amounts, but a very sharp decrease is observed for the other three SDS clean-up methods when the starting protein amount is reduced from 100 µg to 10 µg. Also, it is notable that the FASP method does not always yield an improvement in peptide and spectral identification rates for each starting protein amount, even though more proteins are identified at each level. From the 1 mg sample, the highest number of peptides was identified by the CMW method, followed by the FASP, TCA and DRS methods. For the 100 µg sample, both the DRS method and the CMW method resulted in more identified peptides then the FASP method, with the least identified by the TCA method. However, for the samples with 10 µg starting protein, FASP was clearly superior in the number of peptides identified. The number of spectra identified by each method also reflects the performance of each method in recovering sample. The FASP method identifies more spectra than the remaining three methods for the 1 mg and the 10 µg protein samples but is in between DRS and CMW method for the 100 µg protein sample (**Figure 3.4c**). Even though the FASP method yields slightly higher protein identifications at the 1 mg protein level, this is above the threshold limit recommended for the FASP kits (400 µg). The high protein amount can lead to clogging of the filter, which can be more problematic for proteomes extracted from biofilms and



(a) (b)



(c)

Figure 3.4 Average numbers of (a) proteins (b) peptides and (c) spectra identified by LC-MS/MS in E. coli K-12 lysate with three different protein amounts (1 mg, 100 µg and 10 µg) and prepared by four different SDS clean-up methods.

For calculating the average protein count, peptide count and spectral count, data is taken from the summary table of individual DTASelect.txt outputs, including those assigned to contaminants. Error bars denote 1 standard deviation.

other microbial communities with thick matrices. Even for microbial isolates, we performed a 10 min spin at 20800 g after the SDS lysis to pellet the cellular debris. In discovery driven proteomics experiments, a sample preparation method should be efficient in extracting proteins from all the cellular compartments in an unbiased manner. Extracting membrane proteins is more difficult than the cytosolic proteins, so we investigated whether the four SDS clean-up methods are preferentially distinct in recovering membrane proteins. In order to evaluate differences in the cellular localization profiles of samples resulting from biased protein recovery of each method, the PsortB bacterial localization prediction program was used to predict localization of the identified proteins. **Figure 3.5a** shows the prediction of all the proteins identified from the 1 mg *E. coli* K-12 sample, while **Figure 3.5b** represents the proteins identified uniquely by each SDS clean-up method. The FASP method identified a higher number of proteins for each bacterial location, but it is the number of unique protein identifications which clearly demarcates the efficacy of the FASP method over the other three methods. FASP identified 60 proteins which are predicted to localize to the membrane, compared to 4, 2, and 5 proteins identified by DRS, CMW, and TCA methods, respectively. The number of cytoplasmic proteins identified by FASP is also significantly higher.

To further investigate soluble versus hydrophobic protein identifications, the GRAVY scores for all the identified proteins were calculated. A positive GRAVY score indicates that a protein is hydrophobic in nature, while a negative GRAVY score indicates a protein with a hydrophilic nature.

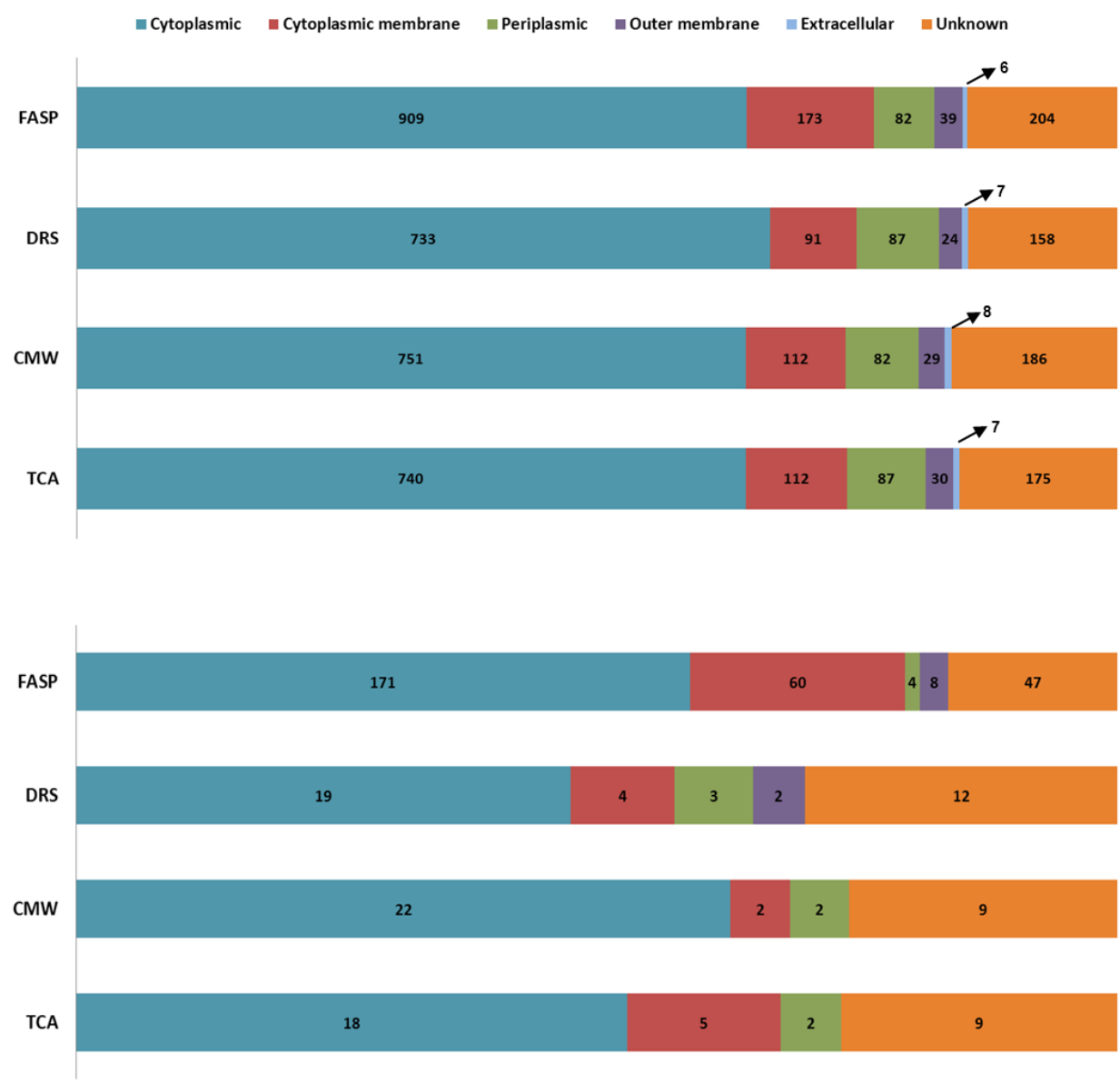
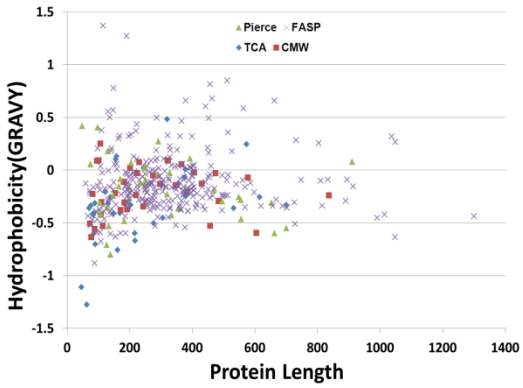


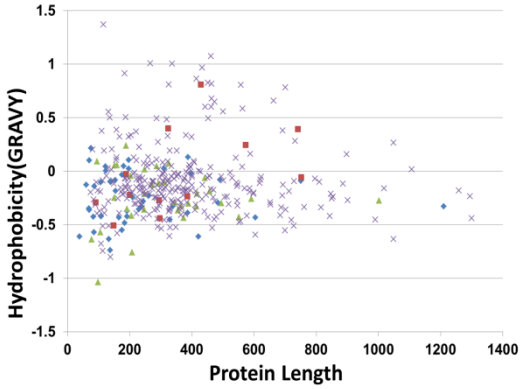
Figure 3.5 Predicted localization of proteins identified by LC-MS/MS after SDS clean-up method in a 1 mg *E. coli* K-12 sample (a) total number of proteins identified by each method (b) proteins uniquely identified by each method.

Figure 3.6 a, b and c shows GRAVY scores of all the unique proteins identified by each SDS clean-up method in 1 mg, 100 µg, and 10 µg *E. coli* K-12 samples. The unique proteins identified by FASP trend more towards the positive side of the GRAVY scale compared to other three methods, suggesting that FASP performs better in retaining hydrophobic proteins compared to other three protocols.

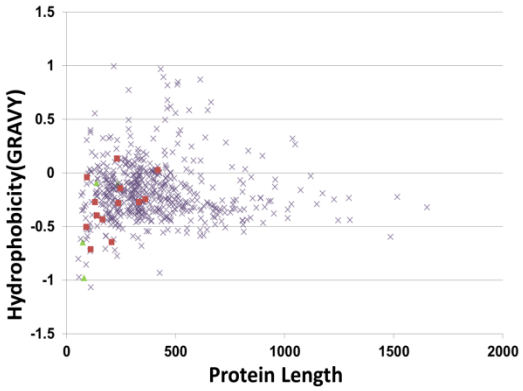
The results suggest that the FASP method is superior to the other three methods at the protein concentrations tested. At a 1 mg protein concentration, all the methods are at a baseline level of ~1200-1300 protein identifications, with FASP faring slightly better. However, the clearest advantage of using FASP is under sample limitation conditions. FASP outperforms the other three SDS clean-up methods at a 10 µg protein concentration. While the other methods identify ~120 proteins, FASP performs 7 times better by identifying ~900 proteins. We think the considerable gain in performance using FASP over the precipitation methods is that FASP does not involve a precipitation step as in TCA and CMW protocols, which can result in loss of an invisible pellet when working with minimal material. Furthermore, the DRS method employs a relatively large resin volume, providing for surface area-associated protein loss. Another advantage of FASP is that all sample preparation steps take place in a small reaction container, which greatly aids in efficient trypsin digestion and minimal protein contact with other surfaces. However, if one wants to evaluate samples with very high protein biomass or samples that are associated with a complex matrix, FASP is unlikely to perform as well. For example, preparation of soil samples using FASP have not been successful, because even after SDS solubilization and centrifugation to remove insoluble material, the supernatant is rich with low density particulate matter. These particulates can easily clog the FASP filter, making it impossible to extract the



(a)



(b)



(c)

Figure 3.6 Hydrophobicity and protein length distribution of uniquely identified proteins by each SDS clean-up method in *E. coli* K-12 samples (a) 1 mg, (b) 100 µg and (c) 10 µg.

proteome. In such cases, TCA precipitation is the preferred method of SDS removal, as it has already been shown to work with soil samples.

3.5 Coupling in-solution intact protein fractionation with a 2D LC-MS/MS experiment

After evaluating different SDS clean-up methods, we sought to examine how to use additional protein fractionation to deepen the proteome coverage. To do so, we employed intact protein separation based on protein molecular weight using GELFrEE technology. The GELFrEE approach shares the principles of SDS-PAGE, fractionating proteins based on their molecular weight, but instead of producing separate bands as in SDS-PAGE, GELFrEE yields proteins in liquid fractions. The separation of proteins into distinct liquid fractions instead of gel bands reduces sample loss and improves protein recovery. Lower molecular weight proteins elute first in the GELFrEE cartridge and subsequent fractions have proteins in order of increasing molecular weight. The performance of the GELFrEE approach was tested on three different biological samples with increasing complexity, i.e. an *E. coli* lysate, 5MM sample, and a ground water microbial community sample (**Figure 3.2**). After solubilization in the SDS sample buffer, the samples were divided into two equal aliquots, one for unfractionated 2D-LC-MS/MS and the other for GELFrEE fractionation prior to 2D-LC-MS/MS. While fractionation of the proteome reduces sample complexity, there is a concomitant reduction of total protein amount available for downstream sample processing. For example, a 500 µg protein biomass sample, when fractionated into twelve fractions, will on average yield only ~ 40 µg proteins per fraction, assuming no protein loss. Therefore, SDS removal from such fractions needs to be performed using a method which can give best results with a low amount of protein. As shown earlier, FASP stands out as the method of choice for the 10 µg – 100 µg range of protein amounts, and, thus, was utilized for preparation of all the GELFrEE fractions. The unfractionated sample was

run in three replicates on an LTQ mass spectrometer, using a 24 hr. 2D-LC-MS/MS method, while the twelve GELFrEE fractions were run individually using a shorter 6 hr. 2D-LC-MS/MS method, thereby keeping the total instrument time almost same for each scheme. For both schemes, we used similar protein amounts per sample type: the *E.coli* K-12 sample and ground water sample utilized 500 µg total protein for each scheme, while 1 mg of 5MM sample was utilized for each of the fractionation and whole cell lysate approach. As seen in **Figure 3.7**, there is significant overlap in identifications between the fractionated and whole lysate schemes, however, the total identifications are boosted by the contributions of identifications unique to each method, in particular for the more complex community samples.

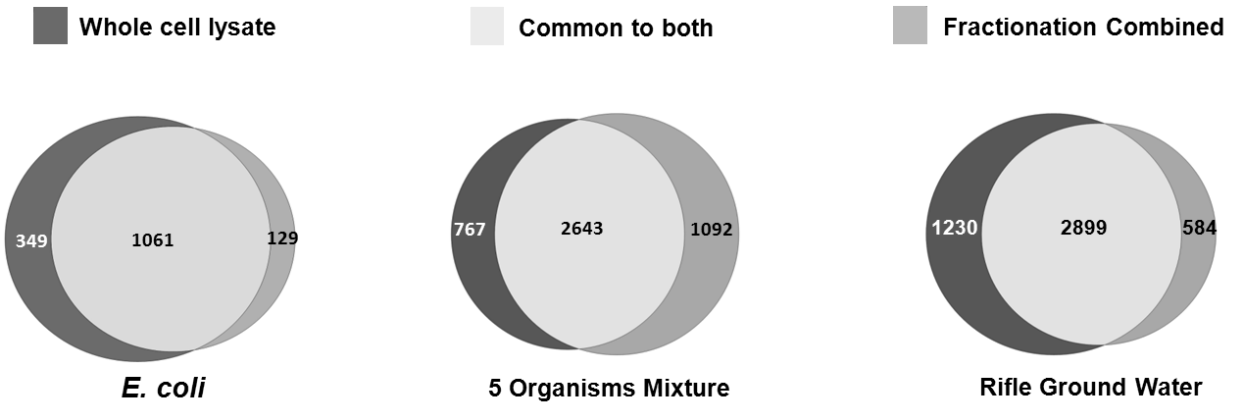


Figure 3.7 Unique and common proteins identified by fractionation and whole cell lysate proteomic methods in three biological samples of increasing complexity

The GELFrEE approach uniquely identifies 129, 1092, and 584 proteins in the whole lysate *E.coli* K-12, 5MM sample, and ground water samples, respectively. Similarly, the unfractionated approach identifies 349 unique proteins in the *E. coli* sample, 767 unique proteins in 5MM sample and 1230 unique proteins in ground water sample. Although repeated technical replicates

can increase total identifications, the whole cell lysate scheme missed many proteins identified by the fractionation approach, even after three technical replicates.

The GELFrEE fractionation scheme yielded 12 fractions collected at different time points, with earlier fractions having low molecular weight proteins, while the later fractions are enriched for high molecular weight proteins. As can be seen in **Figure 3.8**, the Coomassie blue stained 1D gel of GELFrEE fractions from 5MM sample shows that the fractions are well resolved and fairly distinct. The overlap between fractions is a function of sample loading, as shown by Tran & Doucette [113]. In our study, we used either one channel (*E. coli* and ground water sample) or 2 channels (5MM sample) of 8% GELFrEE mid mass cartridge, which fractionates proteins ranging from 3.5 kDa to 150 kDa with optimized resolution at 35-150 kDa.

Table 3.1 shows the number of total proteins as well as the unique proteins identified by each GELFrEE fraction for all the sample types. From this table, it is clear that each GELFrEE fraction contributes to the total protein identification and by manipulating fraction collection time one can get optimum proteome coverage for a given sample. As reflected in the SDS-PAGE gel (**Figure 3.8**), the GELFrEE fractions are enriched in a specific molecular weight zone, and we confirm the same via mass spectrometry. We also observed that the GELFrEE fractionation scheme is unbiased towards the pI of the proteins present in the proteome (**Figure 3.9**). On-line LC-MS/MS is a robust and efficient approach for bottom-up proteomics. However, due to the high complexity of samples, simplification of protein mixtures is a necessary step prior to protein digestion. Our study shows that by analyzing a whole cell lysate in conjunction with fractionation in a tandem approach, we can obtain deeper proteomic coverage of biological samples.

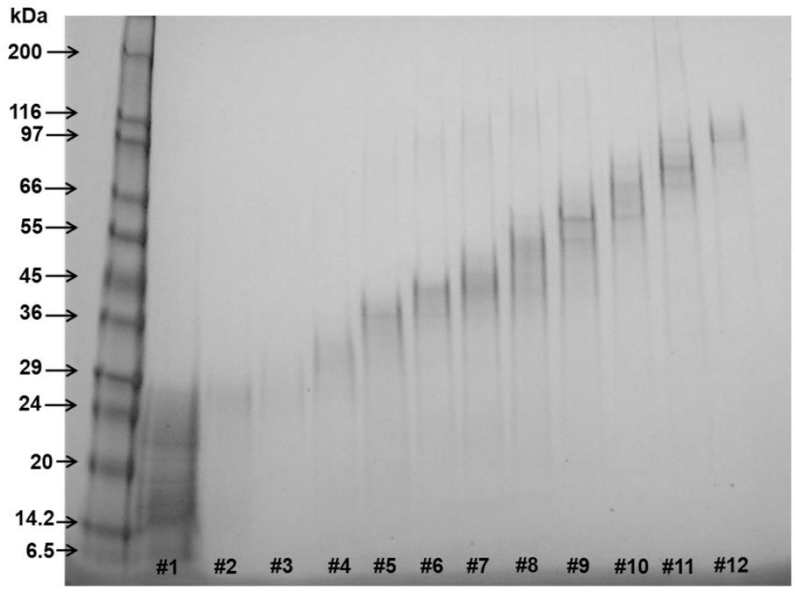


Figure 3.8 Fractionation of 5MM sample using the 8% GELFrEE cartridge.

A 500 µg aliquot of 5MM sample was fractionated into 12 fractions. The figure shows an SDS-PAGE gel image of GELFrEE fraction numbers 1 to 12 ran on a 4-20% gel

Table 3.1 Total proteins and unique proteins identified by each GELFrEE fraction for the three sample types in our study

Fraction No.	*E. coli* K-12		5MM sample		Rifle ground water	
	Total number of proteins identified in each fraction	Number of proteins unique to fraction	Total number of proteins identified in each fraction	Number of proteins unique to fraction	Total number of proteins identified in each fraction	Number of proteins unique to fraction
1	447	119	339	39	40	-
2	538	76	488	31	271	21
3	402	26	785	138	422	43
4	475	32	684	74	449	42
5	484	40	1039	286	869	185
6	280	16	1146	264	896	155
7	288	41	1103	155	1335	387
8	154	15	1429	146	946	113
9	256	26	1070	67	1217	177
10	257	17	997	12	1076	181
11	119	2	781	33	561	25
12	156	13	638	193	597	135

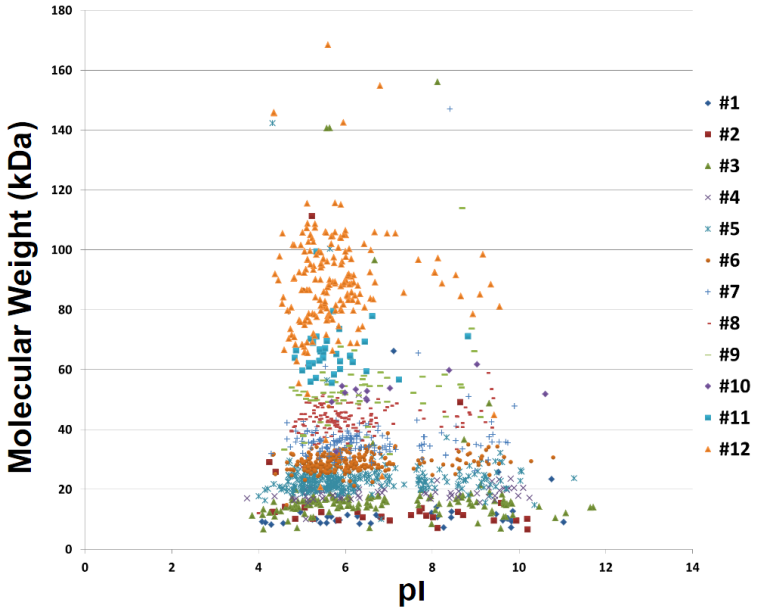
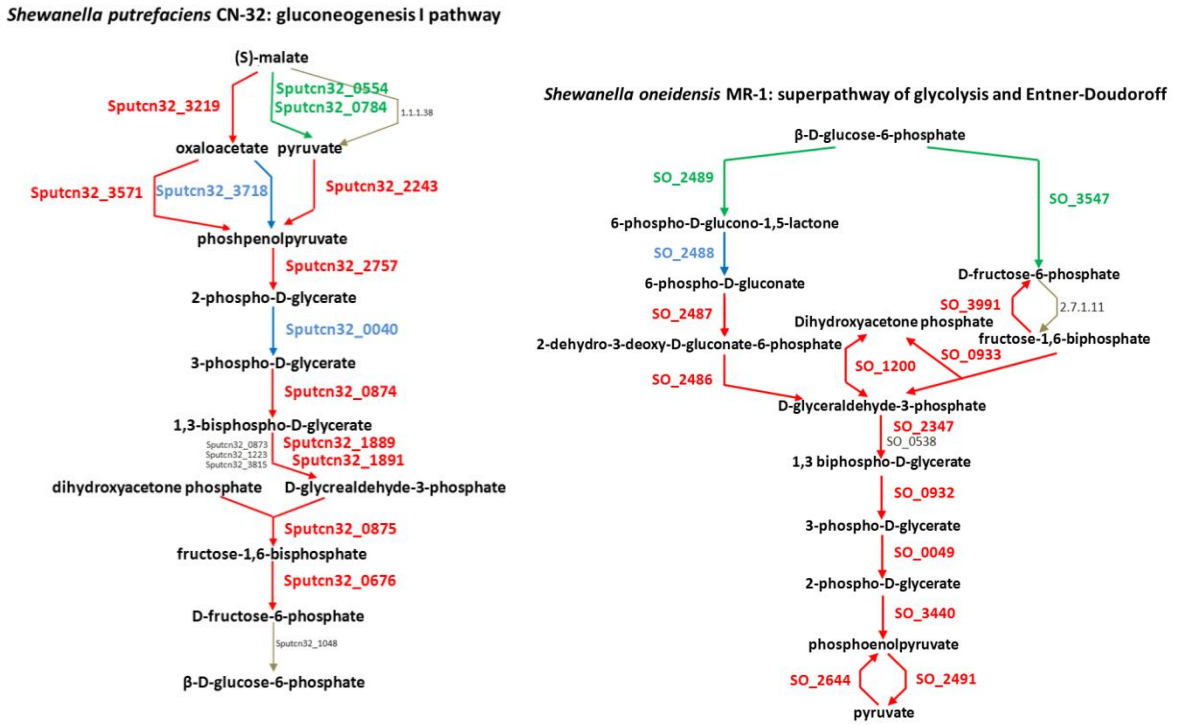


Figure 3.9 Distribution of unique proteins identified in MS run of each GELFrEE fraction from a mixture of 5 microbial isolates

In **Figure 3.10 a and b**, we show two representative metabolic pathways from *S. putrefaciens CN-32* and *S. oneidensis MR-1*, microbes which were part of the 5MM sample. The near completeness of enzyme identifications in these pathways is possible because of the complementary nature of the fractionation and whole cell lysate approach. The complementarity of the tandem fractionation and whole cell lysate approach provides greater biological information. By focusing on the gluconeogenesis pathway (Figure 8a) of *S. putrefaciens* CN-32, the two-prong scheme reveals that malate can either be transformed into oxaloacetate (Gene Sputcn32_3219) or into pyruvate (Gene Sputcn32_0554, Sputcn32_0784). If we had used only the fractionation scheme, we would have obtained only partial information. Similarly, proteomics reveals expression of two genes Sputcn32_3571 and Sputcn32_3718 that mediates the conversion of oxaloacetate into phosphoenolpyruvate.



(a) **(b)**

Figure 3.10 Pathway mapping of identified proteins in (a) *S. oneidensis* MR-1 and (b) *S. putrefaciens* CN-32.

Both the GELFrEE fractionation and whole cell lysate approaches complement each other by providing more biological information as reflected in selected pathways. Red represents proteins identified by both the approaches; blue represents proteins identified only by the fractionation method while the green represents proteins identified only by the whole cell lysate approach. Gray denotes proteins that were not identified by any method.

Thus, if one gene is knocked out, the pathway is still functional, as the second gene can take over the role of the first gene. This information would not have been evident if we had not used fractionation scheme in our proteomics experiment.

Likewise, complementary information is evident for the glycolysis pathway (Figure 8b) of *S. oneidensis* MR1. Here the whole cell lysate scheme identifies gene SO_2489 which convert β-d-glucose-phosphate to 6-phospho-D-glucose-1,6 lactone and gene SO_3547 that converts β-d-glucose-phosphate to D-fructose-6-phosphate. The fractionation scheme provides the connection for conversion of 6-phospho-D-glucose-1, 6 lactone to 6-phospho-D-gluconate by SO_2488. Therefore, it is evident that both possible routes are functionally active in the glycolysis pathway of *S. oneidensis* MR1. The increase in pathway information can enable more informed decisions for other genetic or molecular biology studies.

3.6 Conclusions

Systematic investigations suggest that the choice of detergent removal method heavily relies on the amount of protein in a sample. At high protein amounts, the performance of different SDS removal methods is fairly comparable. However, when the sample is limited with low protein amounts, the experimental approach needs to be carefully planned by taking into consideration all the pros and cons associated with different SDS removal methods. For example, we demonstrated that for an *E. coli* K-12 sample with a low protein amount, FASP significantly outperforms the other three SDS removal methods. However, the FASP method should be further optimized and evaluated with microbial samples from more complex matrices. We also investigated benefits of coupling whole lysate 2D-LC-MS/MS with intact protein fractionation followed by short 2D-LC-MS/MS on collected fractions. Our study demonstrates a definite merit

in protein-level fractionation prior to digestion and downstream peptide separations. This method not only has the potential to increase the total number counts of identified proteins, but also would favorably impact a more targeted approach, when one is only interested in proteins belonging to specific molecular weight range of the proteome.

Chapter 4 - Improving protein extraction for optimal coverage of complex environmental samples

---

4.1 Environmental proteomics – Determining the role of native microbial communities in heterogeneous and complex background

As the name suggests, "environmental proteomics" refers to the identification and characterization of expressed proteins from biomass extracted from its native condition [114, 115]. Therefore, in such studies, the origin of the sample is not a lab grown culture, but rather a guided or a random sampling of the environmental material, which captures a proportion of the total microbial biodiversity in a complex background matrix.

An increasing interest in environmental proteomics is largely due to three main drivers; Firstly, measuring protein expression in environmental samples provides a more accurate representation of the functional activities played by the participating microbial species; secondly, most of the microbial species are not amenable to lab culture and are therefore poorly characterized; and thirdly, *in-situ* proteomics provides information on cooperation and competition among different species within a microbial community with respect to environmental changes and nutrient flux [116-118].

A brief introduction to community proteomics was provided in Chapter 1 and in this chapter we will focus on methods development for enhanced proteomics coverage of a microbial community present in high organic material-containing prairie soils from Konza, KA.

4.2 Characterizing carbon cycling by microbial consortia in response to rainfall variation in native prairie soils

Rapid industrial growth in the last two centuries and our dependence on fossil fuels has significantly perturbed some of the basic universal climate parameters like global temperature, total carbon content, and the percentage of greenhouse gases in the atmosphere; most of which had remain fairly constant for hundreds and thousands of years [119, 120]. This rapid change in climate has brought about unexpected droughts in one part of the globe simultaneous with enormous floods in the other parts of the world. Climate change is further expected to alter precipitation levels that can have disturbing effects on grassland ecosystems which harbors a vast amount of global carbon [121-123].

In this regard, the Department of Energy initiated a major collaborative study to investigate carbon cycling by microbial species present in the native soil systems. In this particular project, the overall goal was to identify the key factors influencing carbon cycle in grasslands when there is a change in natural precipitation regime.

Prairies are natural grassland ecosystems of North America which occupy ~25% of the area in US, and store large amounts of carbon (~33%), and therefore, are an excellent system to conduct this study. The Konza prairie site is a native tall grass prairie preserve managed jointly by the Nature Conservancy and the University of Kansas, and allows researchers to conduct specific altered precipitation experiments. Further, the site has documented evidence of watershed management since 1977 [124, 125]. In this chapter, we will provide some key findings from proteomics measurement of soil system from this site, mainly focusing on the quality of metagenomes and the experimental challenges.

4.3 Experimental procedures to extract proteins from soil samples

The major challenge in extracting microbial proteomes from environmental samples is the complex and interfering background matrix, which also contributes to the difficulty in ascertaining the total protein content of the sample. Information quantifying the available biomass is critical to determine the choice of sample preparation method, as discussed in Chapter 3. However, the presence of unknown interfering substances as well as humic acids, metal ions and chelating agents in the environmental samples precludes accurate determination of protein concentration by any known method [126]. Due to these matrix effects, soils are among the most challenging system to process for proteomics. This difficulty is further complicated by the tremendous heterogeneity in different soil types. Gentle cell lysis methods are unable to break open microbial cells which are unevenly distributed in soils and vigorous and prolong cell lysis methods have a danger of losing proteins, since proteins are known to stick to soil particles [127, 128].

4.3.1 Konza Prairie soil samples and protein extraction

For this study, we received 24 samples from the Konza prairie reserve that were acquired at different time-intervals following different sets of rainfall treatments. **Table 4.1** provides details of the 24 samples and the various treatments. The collaborative project took advantage of the Rainfall Manipulation Plot (RaMP) infrastructure at the Konza Prairie Long-Term Ecological Research site in the Flint Hills region in NE Kansas. The RaMPs provide a long-term replicated (n=6 per treatment) field experiment that aims to mimic predicted shifts in precipitation intervals, while keeping the total precipitation volumes unchanged. In the course of our current program,

Table 4.1 Summary of samples analyzed by proteomics for Konza prairie sediments

Sample ID	Plot	Treatment	RainTime
3	Ambient-4	Ambient	Pre-June
8	Delay-10	Delay	Pre-June
10	Delay-12	Delay	Pre-June
12	Ambient-15	Ambient	Pre-June
15	Ambient-4	Ambient	Pulse-June
20	Delay-10	Delay	Pulse-June
22	Delay-12	Delay	Pulse-June
24	Ambient-15	Ambient	Pulse-June
27	Ambient-4	Ambient	Post-June
32	Delay-10	Delay	Post-June
34	Delay-12	Delay	Post-June
36	Ambient-15	Ambient	Post-June
39	Ambient-4	Ambient	Pre-Sept
44	Delay-10	Delay	Pre-Sept
46	Delay-12	Delay	Pre-Sept
48	Ambient-15	Ambient	Pre-Sept
51	Ambient-4	Ambient	Pulse-Sept
56	Delay-10	Delay	Pulse-Sept
58	Delay-12	Delay	Pulse-Sept
60	Ambient-15	Ambient	Pulse-Sept
63	Ambient-4	Ambient	Post-Sept
68	Delay-10	Delay	Post-Sept
70	Delay-12	Delay	Post-Sept
72	Ambient-15	Ambient	Post-Sept

Seasons – June and September

Treatments – Ambient and Delay rainfall treatments

Three time points relative to rainfall – Pre, Pulse and Post

our collaborators sampled soils before (pre), during (pulse), and after (post) rainfall events from experimental units that represent ambient rainfall (Ambient) as well as experimental units that experienced a 50% increase in the dry intervals between precipitation events and fewer but larger rainfall resulting in same precipitation volume simulating "droughty" conditions (Delay). Proteomics preparation for the 24 soil samples was done in two phases, with the first phase performed at the Lawrence Berkeley National Laboratory (LBNL) and the second phase performed at Oak Ridge National Laboratory (ORNL).

Prior to starting with 24 actual samples, we sought to use best route for proteomics measurements. Therefore, a test soil sample which was from the same site as the actual samples was prepped with three different protocols. The first experimental protocol used SDS lysis, followed by guanidine denaturation and clean-up via solid phase extraction as described in Chapter 2. The second protocol employed differential centrifugation, which aims to separate microbial cells with soil particles by progressive centrifugation speeds [129]. These samples were followed with urea denaturation, trypsin digestion and on-column desalting as described in chapter 2. The last method employed SDS lysis, followed by urea denaturation, trypsin digestion and on-column desalting.

The main difference between the three methods was the amount of starting material. While the SDS-Urea and SDS-guanidine method were only able to use 30 grams of soil sample, the differential centrifugation method was able to utilize 100 grams of material. The results from this initial study suggested that the differential centrifugation approach yields the highest protein coverage followed by the SDS-Urea approach and then by the SDS-guanidine method. However, in spite of the better performance of differential centrifugation approach, it posed a major challenge for peptide loading. This method resulted with peptides in a very high volume (~ 2 ml)

solution, which slowed loading onto the chromatography column and also increased probability of back-column clogging.

Therefore, it was decided to use best features of the top two methods. The first part of sample preparation was performed by our collaborators wherein differential centrifugation displaced microbial cells from soil particles and cell lysis was performed by boiling the sample in SDS. Proteins were precipitation from the lysate by the addition of TCA and the resulting pellet was washed with acetone. Frozen protein pellets were shipped to ORNL and the pellet was reconstituted in 8M urea. Remaining steps of sample preparation including digestion and sample clean-up were performed as described earlier, ensuring that the final volume for peptide loading was less than a ml.

By using the two-step strategy, full 100 grams of soil sample was utilized. As already mentioned, since it is not feasible to accurately determine the protein concentration in environmental samples, the reading from colorimetric assay were not taken into consideration for proteomics sample preparation. Most often, BCA assay on soil samples results in immediate coloration, suggesting possible interference by unknown small molecules or chelating agents rendering the observed protein concentration meaningless.

The samples were prepped as detailed in Chapter 2 from the acetone step onwards, except that an additional sonication step was added to ensure complete denaturation of samples and separation of proteins from the fine soil particles.

The sonication step was introduced in the sample preparation workflow right after suspending protein pellets in 8M urea. Samples in 8M urea were sonicated using a pulse of 5 sec ON| 10 sec OFF| 20% amplitude cycle for 2 minutes. The remaining steps were identical to the generalized

SDS-TCA Urea prep protocol described in Chapter 2. Protein digestion was carried out using trypsin, and samples were loaded onto a 2D back-column (3 cm SCX, 3 cm RP).

For the determination of loading volume for each MS run, BCA assay was performed at the peptide level. But for some of the samples, the BCA color change was almost instantaneous, suggesting a possible interference. Therefore, in such cases the resulting peptide solution was equally divided into three aliquots and one aliquot was utilized for MS measurements. The desalting was performed on the column using 45 minutes column-wash, as described in Chapter 2.

4.3.2 Mass Spectrometry measurements and database searching

All MS measurements were performed on a ThermoScientific LTQ-Orbitrap Elite using the top 20 data-dependent method. For each MS1 in Orbitrap at 15,000 resolution, top twenty peaks were selected for fragmentation via CAD in the ion-trap. Each Konza soil sample was measured once, since the samples included biological replicates.

Representative MS runs were searched first using multiple metagenomes to determine the suitability of the metagenome for all the samples and provide some performance metrics for MS runs. Three different databases were made available for proteomics searches, with two being metagenomes (F12B and F14TB) and the third one being a compendium of 160 microbial isolates relevant to soil systems.

The F14TB was a metagenome constructed from total DNA extracted from the native soil following incubation with bromodeoxyuridine (BrdU), while the F12B was constructed from the DNA extracted after a precipitation step to keep only the DNA that had BrdU incorporated.

4.4 Impact of metagenome size on the depth of proteome identification coverage

While a casual observation with respect to any MS run would suggest that the percentage of identified proteins should increase with an increase in the size of the protein search database. In reality, what matters is how accurately the predicted proteome represents the native microbial community. For the Konza prairie sediments, the size of metagenomes used for the MS searches ranged from 700,000 protein sequences on the low end to 7.8 million protein sequences on the high end. After the visual inspection of total ion chromatograms (TIC) and base-peak chromatograms of all the 24 soil sediments, 16 samples were found to have sufficient high quality spectra that can generate decent proteome coverage. While visual inspection of the raw MS datasets (TICs) can sometimes be misleading, computational programs provide more quantitative metrics which can differentiate between a good proteomics measurement versus a poor proteomics measurement.

Since the source of metagenomics for Konza soil sediments was not an exact match with the source for metaproteomics, we decided to use three different metagenomes. Due to the uncertainty in the accuracy of given metagenomes to correctly match with metaproteomics data, we decided to start with only 3 samples at the beginning. The three samples which made the test search were picked after visual inspection of their base peak chromatograms.

As shown in **Table 4.2**, three different metagenomes were used to explore their feasibility for database searching of all 24 samples. The samples considered for test search include Sample 10 (sample from Pre-June with delayed rainfall treatment), Sample 15 (sample from Pre-June season with ambient rainfall treatment) and Sample 27 (sample from Post-June with ambient rainfall treatment). A snapshot of base peak chromatograms of selected salt pulse from Sample

10 and Sample 27 is shown in **Figure 4.1** for reference. In **Figure 4.1**, the base peak chromatograms show a reasonable density of high abundance peaks, suggesting that the mass spectrometer acquired a high number of tandem scans which should match peptide sequences.

The results tabulated in **Table 4.2**, show poor performance of the three MS runs in terms of total proteins and peptides identified against all the three metagenomes. Among the three samples, Sample 27 gave the best result of 51 proteins with default criteria of two unique peptides per matched protein against the F12B database.

The low number of protein identifications for all the three samples that showed decent quality TICs in visual inspection prompted further investigation. The basic question was whether the proteomics sample preparation failed to extract proteins from these soil systems, or whether the size/accuracy of the metagenome was problematic for matching raw MS/MS spectra to peptides?

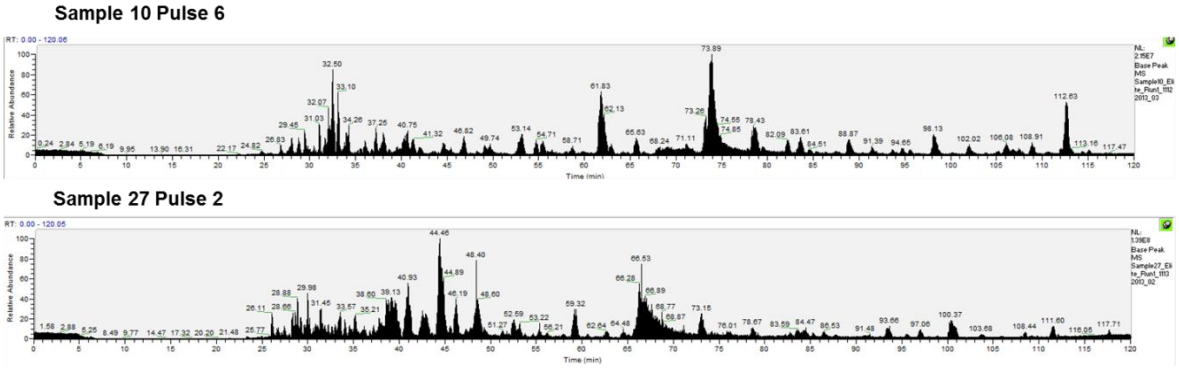
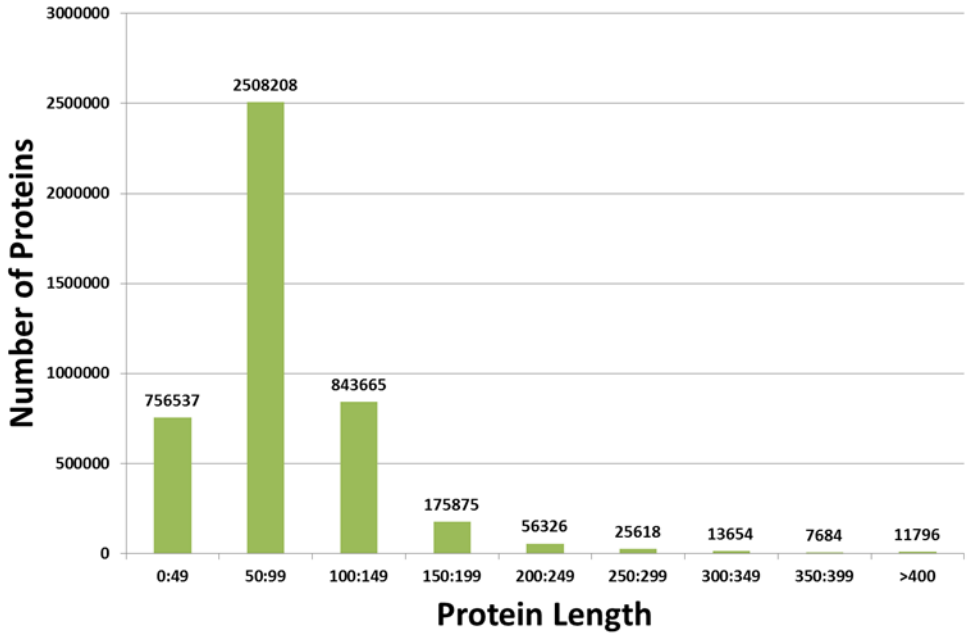
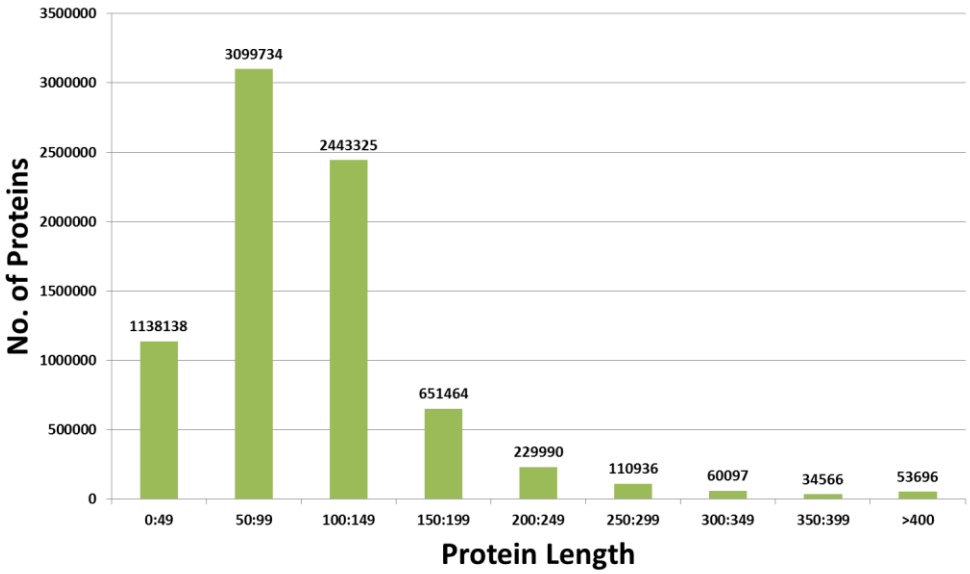


Figure 4.1 Representative salt pulses from the two samples showing the quality of base peak chromatograms.

In order to resolve these questions, sequence level analysis of the two metagenomes was carried out. As shown in **Figure 4.2**, the metagenomes provided for proteomics search were highly fragmented. Almost 75% of the protein sequences in both F14TB and F12B database were less



(a)



(b)

Figure 4.2 Distribution of number of proteins with respect to protein length for (a) F12B database and (b) F14TB database

Table 4.2 Proteomics results from selected runs of Konza soil sediments

Sample ID	Metagenome Identifier	Database Size	Proteins	Protein Groups	Peptides	Protein FDR	Filtering Criteria	Q Value
Sample 10	F14TB	7,821,946	14	13	27	0%	2 unique/1 non-unique peptide	2%
Sample 10	F14TB	7,821,946	145	63	77	0%	1 unique/1 non-unique peptide	2%
Sample 10	F14TB	7,821,946	795	132	175	0%	1 unique/1 non-unique peptide	5%
Sample 10	F14TB > = 150 aa	1,140,749	17	15	33	0%	2 unique/1 non-unique peptide	2%
Sample 10	F14TB > = 150 aa	1,140,749	443	52	70	0.90%	1 unique/1 non-unique peptide	2%
Sample 10	F14TB > = 150 aa	1,140,749	496	101	144	0.81%	1 unique/1 non-unique peptide	5%
Sample 10	F14TB > = 300 aa	148,358	50	32	72	0%	2 unique/1 non-unique peptide	2%
Sample 10	F14TB > = 300 aa	148,358	83	39	77	2.41%	1 unique/1 non-unique peptide	2%
Sample 10	F14TB > = 300 aa	148,358	103	57	106	1.94%	1 unique/1 non-unique peptide	5%
Sample 15	F14TB > = 150 aa	1,140,749	19	18	63	0%	2 unique/1 non-unique peptide	2%
Sample 15	F14TB > = 150 aa	1,140,749	77	69	114	0%	1 unique/1 non-unique peptide	2%
Sample 15	F14TB > = 150 aa	1,140,749	120	105	179	1.67%	1 unique/1 non-unique peptide	5%
Sample 27	F12B	4,339,163	51	45	129	0.00%	2 unique/1 non-unique peptide	2%
Sample 27	F12B	4,339,163	347	259	343	0.58%	1 unique/1 non-unique peptide	2%
Sample 27	F12B	4,339,163	498	398	540	3.21%	1 unique/1 non-unique peptide	5%
Sample 27	F12B >= 100 aa	1,134,618	41	41	124	0%	2 unique/1 non-unique peptide	2%
Sample 27	F12B >= 100 aa	1,134,618	170	139	222	1.18%	1 unique/1 non-unique peptide	2%
Sample 27	F12B >= 100 aa	1,134,618	269	228	370	4.46%	1 unique/1 non-unique peptide	5%
Sample 27	160 Isolates	744,221	20	15	56	0%	2 unique/1 non-unique peptide	2%
Sample 27	160 Isolate genomes	744,221	167	49	90	0%	1 unique/1 non-unique peptide	2%
Sample 27	160 Isolate genomes	744,221	272	96	176	1.47%	1 unique/1 non-unique peptide	5%

Default filtering criteria is shaded green

less than a 100 amino acids.

So, the database size was systematically reduced, first by eliminating all the sequences that were less than 100 amino acids, and then extending the truncation to remove sequences which were less than 300 amino acids. But the search results suggested that shortening the database did not aid in rescuing peptide-spectrum matches. As shown in **Table 4.2**, on the one hand, there was a marginal increase in protein identifications for sample 27 with the reduction in F14TB database, while on the other hand there was a decrease in protein identification for Sample 27 with the reduction in F12B database. Therefore, reduction in database size was ineffective in increasing the proteome coverage. We further relaxed the search parameters to allow one-peptide hits, but that also did not significantly raise the total protein identification. The best result obtained from the test search was ~800 proteins from Sample 10 against Full F14B database of 7.8 million sequences, using 1 peptide hit at a relaxed Q value of 5%. Therefore, it was decided that these metagenomes are not sufficiently curated for searching the remaining samples, and that the available data should be mined further to examine spectral quality and scores of detected peptide-spectrum matches.

4.5 Evaluating the quality of MS/MS data using ScanRanker

While the results from the test searches gave poor peptide identification metrics, it was essential to search for the root cause of this failure. We decided to evaluate the spectral quality of the MS measurements by using ScanRanker, which assigns a numeric score to every collected MS/MS spectra based on certain features present in the tandem scan [130].

The ScanRanker program developed by Tabb group at Vanderbilt University enables quantitative evaluation of tandem scans via a sequence tagging approach. A sequence tag is a short 3-5 amino

acid long peptide sequence that can be directly inferred from a tandem scan with or without a need of protein database. By applying sequence tagging schema, the program works *de-novo* and is not dependent on protein database. The program thus helps in the identification of high quality spectra that are not picked up by regular database searching algorithm. For each tandem scan, ScanRanker evaluates several parameters for the generation of a sequence tag including, spectral intensity, relatedness between two sequence tags from the same scan, the probability of inferred sequence tag to occur by chance, and the distribution of fragment m/z values forming the tag. After evaluating individual variables associated with the generation of sequence tags, a composite score is assigned for each tandem spectra. The final score of each tandem spectra is further normalized to account for the variation in scores across the entire salt pulse and the complete MudPIT run. The final score for a tandem scan vary from a low negative value (which defines a low quality spectrum) to a low positive value. In our experience, we have found that any score above zero is a high quality tandem scan that should provide for a peptide sequence match.

A representative example of ScanRanker analysis for one of the Konza soil sample is shown in **Figure 4.3**. The x-axis represents the normalized ScanRanker score, while the y-axis is the total number of MS/MS scans acquired during the entire MudPIT run. As evident, there are a large number of high quality spectra in this plot (blue bars with score above 0), out of which only a handful of tandem scans (red bars) are identified by the database search algorithm. For example, the total number of tandem scans that have a ScanRanker score of 2 is 3821, but only 71 of these 3821 spectra are mapped to a peptide sequence, which is a mere ~2% of the total scans. This highlights the fact that the proteomics measurements were of better quality than the depth of identifications provided by the search results.

We further evaluated the results of MyriMatch database search algorithm without parsing it through IDPicker filtering [77, 80]. The MyriMatch algorithm generates a list of peptide-spectrum matches by regular database searching and assigns cross-correlation scores to each PSM. The Idpicker program then parses these PSMs and maps them to proteins after applying user-defined filtering criteria like the one-peptide or two-peptide rule, FDR threshold, and the extent of protein grouping, etc.

We applied zero filtering at the Idpicker stage and allowed all the tandem scans that were picked by MyriMatch to be considered for further analysis. This provided an opportunity to evaluate the discrimination between the forward and reverse hit just by knowing the XCorr and the confidence in matched PSM via Q-value.

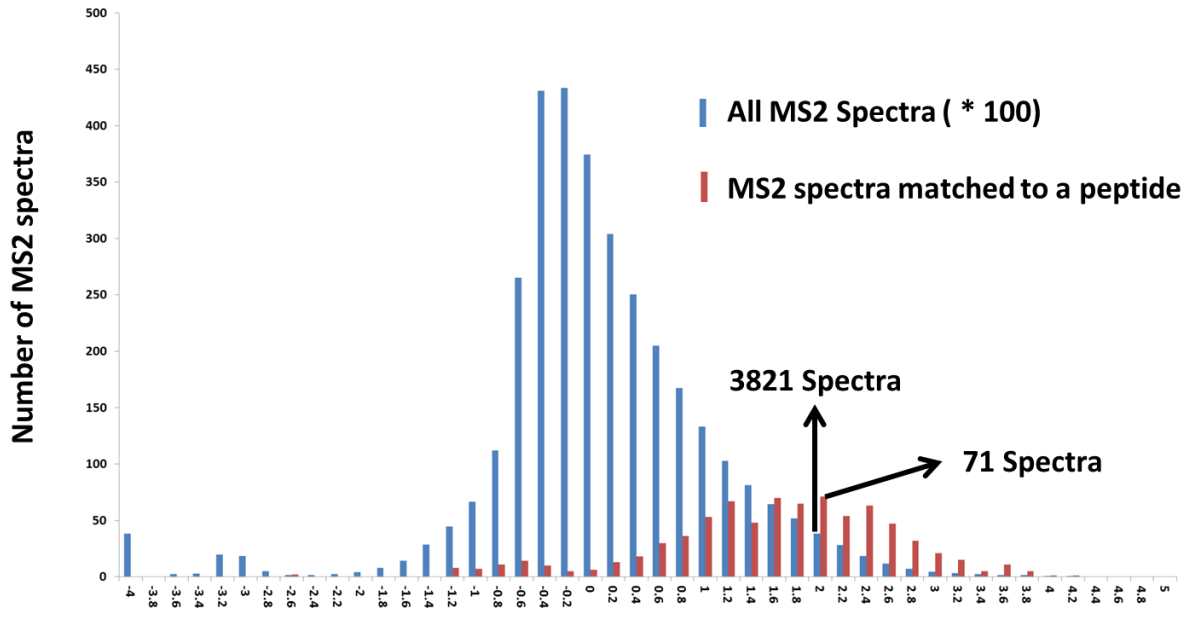


Figure 4.3 Plot of ScanRanker score with respect to the total number of MS/MS scans acquired from Sample 10. *The red bars represent the total number of experimentally identified scans (1 unique peptide, 5% Q-value) for Sample 10 by MyriMatch database searching against F14TB database (~7.8 million sequences)*

Figure 4.4 highlights our observation for three distinct systems going from the simplest on the right hand side to the most complicated system on the left hand side. As shown in the top rightmost panel of **Figure 4.4**, database search of *C. elegans* reveals that proteomics can clearly distinguish between the forward and the reverse database of this genome. The high XCorr values predominantly come from the forward hits. In comparison, the Rifle groundwater (top middle) shows slightly lower discrimination in the forward and reverse database matching. But still, at a database size of 2.2 million sequences, there are a considerable number of sequences that reflect microbial population in this sample and those are picked up by database searching algorithm. These results indicate that this is a fairly high quality metagenome assembly, which was known from other independent sources. In contrast to the other two systems, the metagenome from Konza prairie system is very poorly assembled/curated. This is evidenced by the proteomics not being able to distinguish between the forward and the reverse hits.

The bottom-panel in **Figure 4.4** provides further information on the importance of accurate metagenomes for environmental samples. It illustrates Q-values, a score is assigned by MyriMatch for each PSM that varies from 0 to 1, and wherein 0 refers to a high confidence accurate match and 1 refers to a completely random match. By default, Idpicker imports the MyriMatch data passing 0.25 Q-value. Note that the total number of MS/MS scans in Konza Sample 10 below this threshold is almost zero. The simple *C. elegans* system yields majority of the tandem scans mapping to the forward database with a Q-value less than 0.25. In comparison, the groundwater sample has an overall reduction in total MS/MS scans, but assigns the bulk of them to the forward database at low Q-value. However, in Konza soil sediments the majority of tandem scans have a Q-value above 0.75, indicating that there is no distinction among the forward and reverse hits.

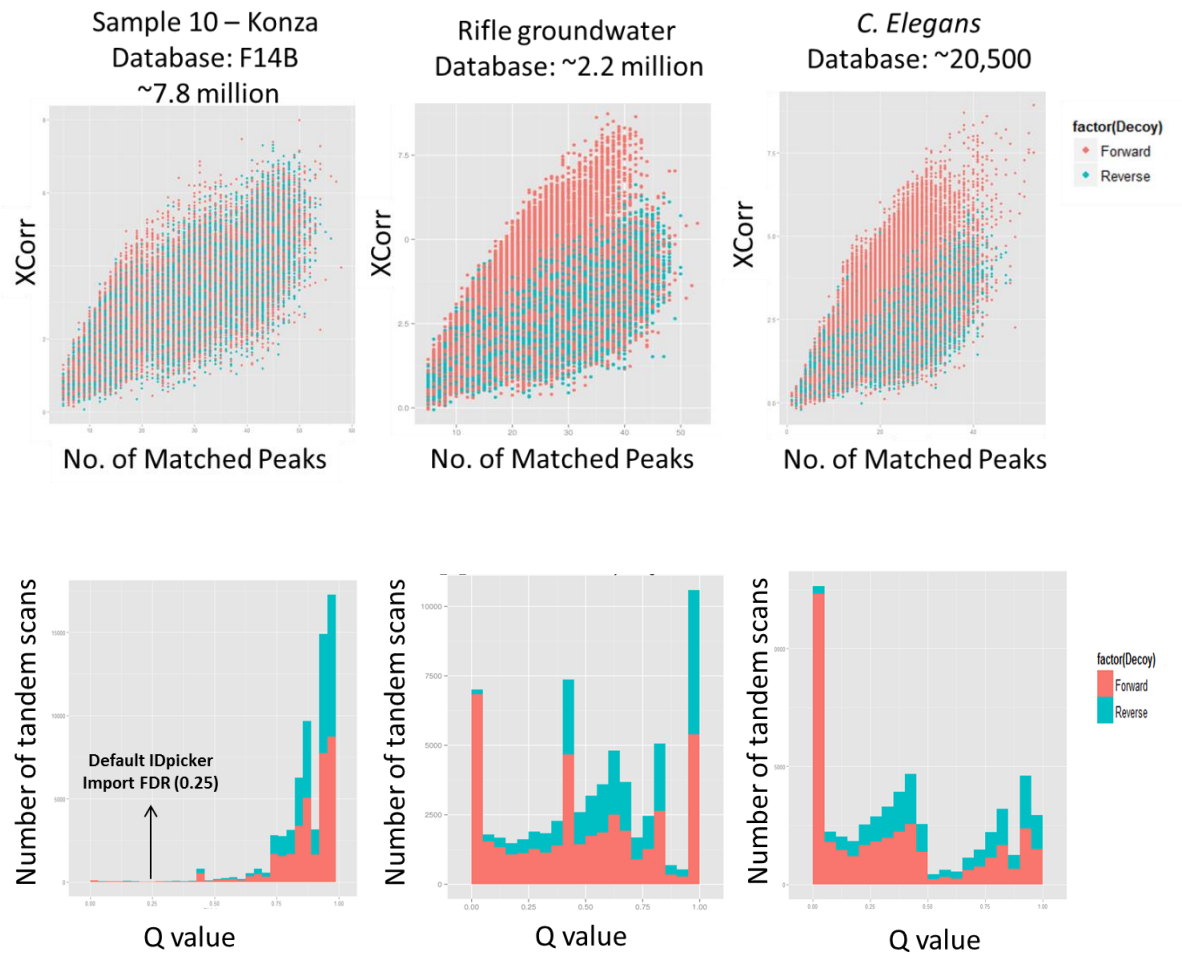


Figure 4.4 Quality check of tandem scans with respect to XCorr and Qvalue.

Top panel shows plot of XCorr vs. number of matched peaks in a tandem scan. In general, the value of XCorr is directly proportional to the number of matched peaks in a tandem scans apart from several other factors. The bottom panel represents the number of tandem scans on y-axis with respect to their Q-value. A lower Q-value means higher confidence in the PSM.

.

The in-depth analysis of proteomics dataset from the selected samples provided valuable feedback to our collaborators involved in metagenomics and an updated metagenome based on expression data is under progress for more accurate and refined proteome matches.

4.6 Conclusions

Environmental proteomics pose numerous challenges that are not encountered in simple biological systems. While we resolved some of the experimental challenges by adapting our current proteomics methods, in systems with so large microbial heterogeneity and complex sample media, it is not likely that one approach will work in all types of conditions. The *de-novo* assessment of tandem mass spectrometry data is a definite boon for the field of environmental proteomics, especially in the absence of high quality metagenomic data, and permits evaluation of the raw MS/MS datasets, thereby pointing to the likely problem-area of inaccurate or poorly assembled metagenomes.

Chapter 5 - Challenges in identification of modified peptides: Using alternate proteases and fragmentation methods for comprehensive identification of post-translational modifications in bacterial isolates

5.1 Mass spectrometry in identification of modifications and substitutions on peptide sequences

The complexity of the proteome of any organism goes beyond the interplay of twenty amino-acids that form the primary sequence of hundreds and thousands of proteins that are present. To date, more than 140 natural non-proteinogenic amino acids and their derivatives have been characterized in different organisms, along with twenty basic amino-acids coded by the genetic machinery [131]. It is therefore apparent that the majority of proteins that are translated in an organism can be modified either at the time of translation or after it. The huge diversity of PTMs, of which we only know a handful, decorates the tertiary and quaternary structure of proteins, altering their behaviors by rendering them active or inactive, changing their turnover time, enabling/disabling protein-protein or protein-ligand interaction, and aiding in translocation [132-134]. Therefore, the analysis of PTMs is very crucial to decipher the complete picture of protein machinery in action and making meaningful inferences.

The extent of PTMs is thought to increase with the increase in complexity of the organism as one goes from lower eukaryotes to higher eukaryotes, such as plants or animals. Though the presence of PTM challenges the central dogma of molecular biology, which states that one gene codes for one protein which carries out a single function, it helps in understanding how most of the species

are able to carry out an order of magnitude more functions than the total number of genes encoded in their genome.

The most common PTMs that have been widely studied include phosphorylation, acetylation, methylation, and glycosylation [132]. Since any change to protein sequence, whether it is the addition or a subtraction of a chemical moiety, causes a mass shift, mass spectrometry has been a method of choice to investigate modified proteins and peptides. The field of PTM identification has moved from PTM identification on single proteins in early 1990's to simple mixtures in the early 2000s to complex mixtures and proteome level studies in recent years [135, 136]. The focus of most of the proteome level PTM discovery research has been eukaryotic systems, including the major model systems of human cell lines, mouse, yeast, the fruit fly, and *Arabidopsis thaliana* [51, 53, 137, 138]. However, majority of studies in these systems have focused on one type of modification, most commonly phosphorylation followed by acetylation. Using enrichment strategy, these studies targeted phosphopeptides or acetylated peptides, and then employed mass spectrometry to identify and localize the site of modification [139].

Only a limited number of studies have used prokaryotic systems for phosphoproteomics discovery, and to the surprise of many researchers, low level of PTM regulation has been identified in microbial species [140-142]. **Table 5.1** highlights some of the recent work done in PTM discovery of phosphoproteins in unicellular prokaryotic systems. The presence of phosphorylated proteins in prokaryotes suggests that microbial species utilize PTM as a method of regulation and this hint at the existence of other major modifications, like acetylation and methylation, which have not been studied in prokaryotes at system level.

Table 5.1 Global and phosphorylation site-resolved studies on bacterial phosphoproteins based on gel-free methods

Organism	Strain	Genome size	No. of P-proteins	No. of P-events	P-Ser (%)	P-Thr (%)	P-Tyr (%)	Publication
B. subtilis	168	4245	78	103	69	20.5	10	Macek et al., 2007 76
E. coli	K12-MG1655	4289	79	105	68	23.5	8.5	Macek et al., 2008 77
L. lactis	IL1403	2266	63	73	46.5	50.5	3	Soufi et al., 2008 78
H. salinarum	R1	2886	69	81	86	12	1	Aivaliotis et al., 2009 79
K. pneumoniae	NTUH-K2044	5814	81	117	n.d.	n.d.	n.d.	Lin et al., 2009 82
P. aeruginosa	PAO1	5565	23	55	52.7	32.7	14.5	Ravichandran et al., 2009 83
P. putida	PNL-MK25	5532	40	53	52.8	39.6	7.5	Ravichandran et al., 2009 83
S. coelicolor	A3(2)	7897	40	46	34	52	14	Parker et al., 2010 80
M. tuberculosis	H37Rv	3918	301	516	60	40	n.d.	Prisic et al., 2010 84
S. pneumoniae	D39	2246	84	163	47	44	9	Sun et al., 2010 81

Macek B, Mijakovic I *Proteomics*. 2011;**11**(15):3002-11.

Therefore, in this work we coupled CAD and ETD fragmentation based 2D-LC-MS/MS experiments on two model bacterial species, namely *Pseudomonas putida* F1 and *Shewanella putrefaciens* CN-32 digested with alternate proteases, to characterize the range of modifications. The modifications targeted here include phosphorylation of serine, threonine and tyrosine, mono-methylation, di-methylation and tri-methylation of lysine and mono-methylation and di-methylation of arginine and acetylation of lysine.

5.2 Materials and Methods

5.2.1 Sample Lysis

Microbial cultures corresponding to 500 µg total protein were used for sample preparation. First, sample was lysed by addition of SDS buffer and boiling for 10 minutes. The cellular debris was removed by brief centrifugation at 20800 g for 5 minutes. SDS from the supernatant (150 µl) was removed by Filter-aided Sample Preparation (FASP) method using FASPkits (Protein Discovery). Briefly, supernatant was loaded on to a FASPkit centrifugal molecular weight cut-off filters. 8M Urea was added on top of the filter and the tube spun for 20 minutes. The process was repeated thrice and the flow through was discarded each time. Following urea washes, sample was treated with iodoacetamide (IAA) for 15 min in dark to block cysteines. After IAA treatment, the filter was equilibrated with ammonium bicarbonate buffer.

5.2.2 Protein Digestion

Since our experimental design did not include any enrichment-based strategy, therefore to achieve maximum sequence coverage to increase the probability of finding a modified peptide, we used multiple proteases for protein digestion. Specifically we employed LysC and GluC, in addition to the commonly used protease trypsin.

5.2.3 Understanding the proteolytic activity of LysC and GluC endoprotease

LysC is serine protease produced by multiple bacterial species but *Lysobacter enzymogenes* is the most commonly used host for the commercial production. *Pseudomonas aeruginosa* is another bacterial species that is used for commercial production of LysC. The enzyme is produced as a 48 kDa inactive proenzyme which matures to a 33 kDa active enzyme. The enzyme shows exceptional cleavage specificity of only cleaving on C-terminal side of a lysine residue. Apart from being highly specific, the enzyme is more robust to chemical denaturation and retains activity even in solutions containing 8M urea which gives it another advantage over trypsin for protein digestion [143].

Staphylococcus aureus Protease V8 which is commonly referred as GluC is a secreted serine endoprotease which shows activity over wide range of pH from 3.5 to 9.5 but shows maximal activity at pH 4.0 and pH 7.8. The enzyme specificity is dependent on the buffer composition in which proteins are present. While, GluC preferentially cleaves after carboxylic acid group of glutamate in wide variety of buffers including Tris-HCl, ammonium bicarbonate and phosphate buffers, however, in presence of phosphate buffers it also cleaves after aspartic acid residue at pH 7.8. GluC is a highly stable enzyme and can tolerate up to 6M Urea, 5.5 M guanidine-HCl and 0.5% SDS [144].

Both LysC and GluC yield longer proteolytic peptide sequences, since they have higher specificity of peptide cleavage compared to trypsin. The advantage of having longer sequences that can carry higher charge state will become clear in the next section where alternate fragmentation modes in mass spectrometry are explained.

5.2.4 Liquid chromatography and mass spectrometry

Digested samples were pressure-cell loaded onto a 150-µm i.d. back column packed with 3 cm of C18 reverse phase column followed by 3.5-5 cm of strong cation exchange column (SCX Luna, 5 µm particle size, 100 Å pore size, Phenomenex). The back column was connected to a 15-cm-long 100-µm-i.d. C18 RP PicoFrit column (New Objective) and placed in-line with a U3000 quaternary HPLC (Dionex, San Francisco, CA). The SCX LC separation was performed with eleven salt pulses containing increasing concentrations of ammonium acetate. Each salt pulse was followed by a 2 hr. reverse phase gradient from 100% Solvent A (95% H_20, 5% AcN and 0.1% formic acid) to 60% Solvent B (30% H_20, 70% AcN and 0.1% formic acid). The LC eluent was directly nanosprayed into a Thermo Scientific LTQ Orbitrap-XL mass spectrometer with ETD. During LC, the mass spectrometer was operated in data-dependent mode and under the control of the Xcalibur software (Thermo Scientific). For each full scan, top 5 peaks were selected for MS/MS fragmentation via CAD (35% collision energy) or ETD mode (100 ms activation time).

5.2.5 Electron Transfer dissociation (ETD) as a fragmentation mode in mass spectrometry

As the name suggests, the ETD scheme employs a transfer of an electron from a radical anion to the positively charged peptide, which leads to a localized point charge fragmentation of the N-Cα peptide bond, forming *c*- and *z*- ion pairs. The ETD mode of fragmentation was developed by Syka *et. al.* in 2004 based on the Electron Capture Dissociation (ECD) approach described by Zubarev *et. al.* in 1988, which at that time was applicable only to Fourier-Transform Ion Cyclotron Resonance mass spectrometers (FTICR) [145, 146].

For ETD, the generation of electrons takes place in an external source, typically using polycyclic aromatic hydrocarbon molecules like fluoranthene. ETD introduces low energy electron transfer from the anion to a multiply charged positive peptide species, thereby converting it into a radical cation, which is highly unstable and immediately dissociates via multiple pathways, mainly leading to the formation of an even electron *c*- type ions and an odd electron *z*- type ions.

One of the most important characteristics of ETD fragmentation is that it is a non-Ergodic mode of fragmentation, which preserves labile PTMs and is therefore highly suited for PTM discovery. While ETD fragmentation is not dependent on amino-acid composition of a peptide (except for proline) and therefore yields full sequence coverage, its efficiency does depend on the precursor charge of the peptide. In general, ETD fragmentation gives enhanced sequence coverage for peptides with a charge state of +3 and higher, as compared to CAD, which works best with doubly charged ions [147]. Since tryptic digests typically consist of +2 charge state peptides, using alternate proteases enabled generation of longer peptides, which can carry higher charge states, thereby facilitating ETD fragmentation.

5.2.6 Computational methods for PTM analysis

The data from ETD/CAD runs on two microbial isolates *P. putida* F1 and *S. putrefaciens* CN-32 digested with two alternated proteases (GluC and LysC) was searched using multiple informatics programs. The RAW files were searched for potential modifications using Myrimatch and by COMPASS (GUI version of OMSSA or Open Mass Spectrometry Search Algorithm) [148]. The data from the two search programs, MyriMatch and COMPASS were then filtered using IDPicker and Protein Herder respectively [77, 80].

The PTMs included in search were phosphorylation of Ser, Thr and Tyr, mono-methylation, di-methylation, tri-methylation of Lys, mono-methylation, di-methylation of Arg and acetylation of Lys.

5.2.7 Basic principles of OMSSA and MyriMatch search algorithms

Open Mass Spectrometry Search Algorithm (OMSSA) is a fast database search program developed at NIH, which employs a probability based method to calculate a score between the observed fragment ions and *in-silico* derived peptide spectrum map. Using tandem mass spectrometry data as input, OMSSA removes background noise, extracts m/z values and compares these values against theoretical m/z values within a user defined tolerance window. The hits are then scored by statistical and probabilistic methods [149].

MyriMatch is another database search algorithm developed by the group of Dr. David Tabb at Vanderbilt University which not only considers the number of matched fragment ions to the theoretical spectra like most of the available database search programs, but also uses the intensity of the matched ions to give a final score to a peptide-spectrum match. By taking in both the total number of matched fragment ions and their intensities, MyriMatch gives a better discrimination between a random match and a true match [77].

5.3 A pilot study using trypsin digestion to evaluate ETD/CAD fragmentation

The three isolates (*E. coli*, *P. putida* F1 and *S. putrefaciens* CN-32) digested with trypsin for a short duration (1 hour at 37 °C) gave reasonable results in terms of protein identification with CAD fragmentation, but performed drastically poorer for ETD fragmentation. The number of PTMs identified also reflected the same.

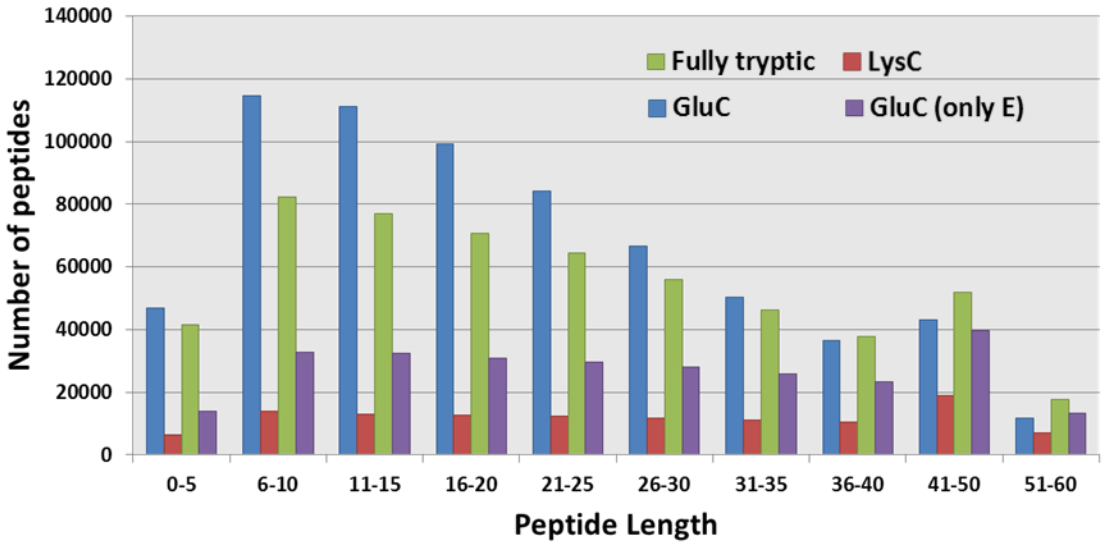
The results are summarized in **Table 5.2**. As seen in the table, a CAD measurement of *S. putrefaciens* CN-32 identified 843 proteins while the ETD run only identified126 proteins. Similarly, the total number of modified spectra, i.e. the number of spectra that can possibly validate presence of PTMs in a sample, was 1,110 from CAD run, as compared to only 82 for ETD of *S. putrefaciens* CN-32.

The reason for the low number of identified proteins with ETD scheme was further examined. Since ETD works best at higher charge states, which are directly proportional to the peptide length, we performed an *in-silico* digestion of our microbial species with trypsin and the other alternate proteases.

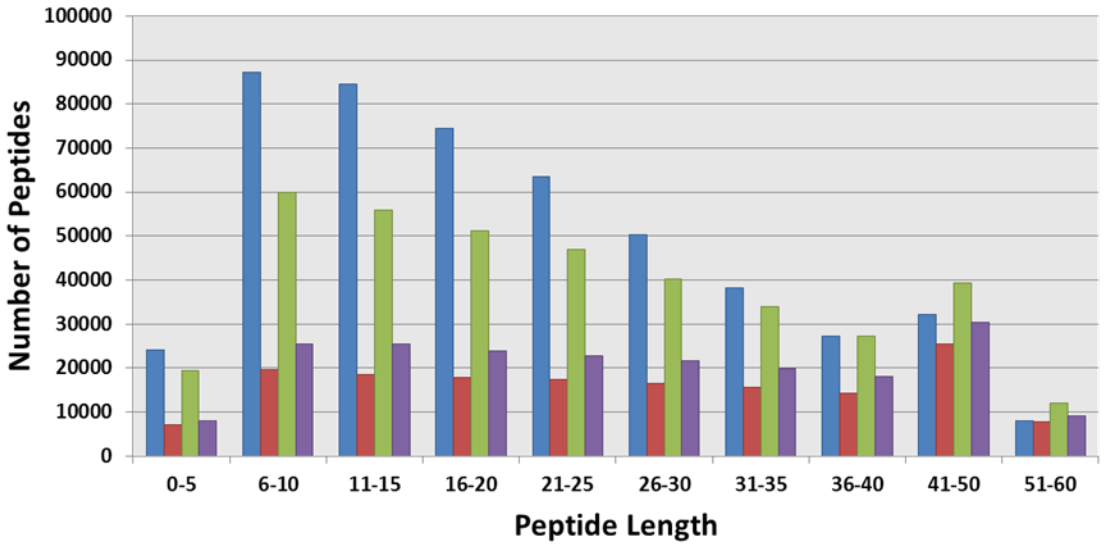
As shown in **Figure 5.1 a and b**, majority of tryptic peptides fall in the range of 6-20 amino acids, while the number of LysC and GluC (only E) digested peptides are equally distributed across the length starting from 6 amino acids to 40 amino acids. Therefore, it is logical to assume that with LysC or GluC digested proteins, there is equal likelihood of longer peptides to be sampled for fragmentation, as compared to trypsin. As expected, tryptic peptides shorter than 20 amino acids are dominant, therefore, there is a very low probability of longer peptides to undergo fragmentation. In general, longer peptide sequences have higher propensity to accommodate additional charge. As discussed before, ETD is more suited for $> +3$ charge state species compared to CAD mode. Therefore, charge state of MS identified peptides from trypsin digests was also examined **(Figure 5.2a and b).** As is evident in **Figure 5.2**, using trypsin not only resulted in smaller peptides theoretically, but experimentally derived data suggest high abundance of +2 charge state compared to any other charge state. This further explained the poor performance of ETD measurements with trypsin as the protease.

Table 5.2: The total number of proteins and the total number of modified spectra identified for both the fragmentation schemes for each organism using trypsin digestion.

Organism	Protein IDs CAD	Total Modified Spectra	Protein IDs ETD	Total Modified Spectra
S. putrefaciens CN-32	843	1110	126	82
P. putida F1	1027	2808	168	101
E. coli K-12	694	1697	404	311

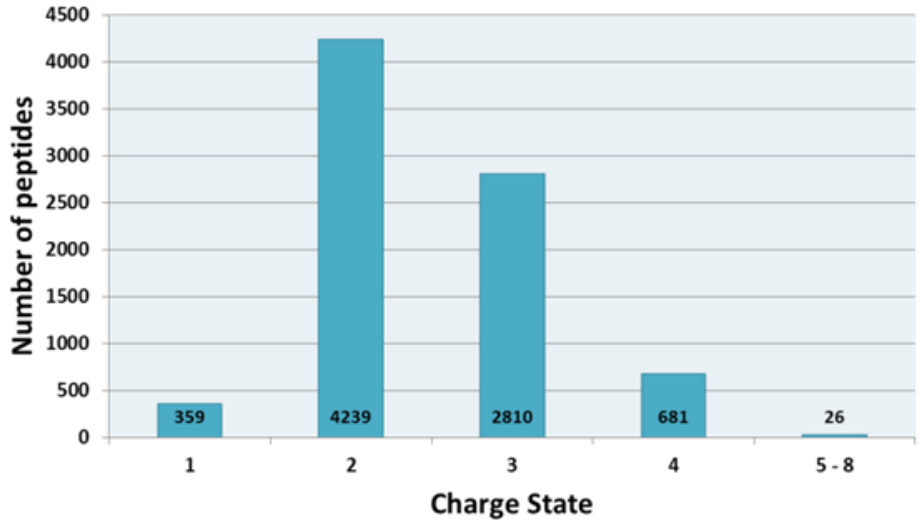


(a)

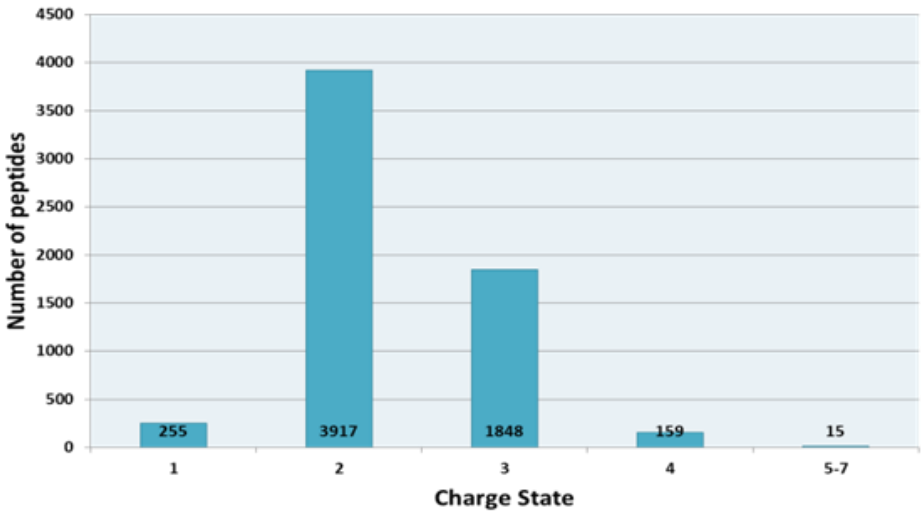


(b)

Figure 5.1 *In-silico* digestion of microbial isolates using various proteases (a) *P. putida* F1 and (b) *S. putrefaciens* CN-32



(a) *P. putida* F1



(b) *S. putrefaciens* CN-32

Figure 5.2 Charge state distribution of tryptic peptides identified by CAD fragmentation in (a) *P. putida* F1 and (b) *S. putrefaciens* CN-32

5.4 Using alternate proteases to boost PTM identifications

Results from the MS runs and the *in silico* digestion revealed trypsin as the least effective protease for PTM discovery, especially by ETD fragmentation scheme.

Therefore, further work was carried out using LysC and GluC as the alternative proteases to trypsin. Results from the MS run of samples digested with alternate proteases, showed MS runs employing CAD fragmentation provide more protein identifications compared to MS runs with ETD fragmentation **(Table 5.3)**. Though, a high degree of overlap was observed in terms of proteins commonly identified by both CAD and ETD, the CAD mode significantly superseded identification of unique proteins compared to that by ETD. This can be attributed to a couple of factors. Firstly, the informatics pipeline in dealing with CAD spectra is very well established compared to ETD data. Secondly, the total number of spectra collected from a single salt pulse in CAD mode was on average 4,000 more than that of a salt pulse in ETD mode. So adding up this number over a 24 hour run, we collected approximately 40,000 more spectra from CAD run compared to ETD run. The lower number of collected spectra in ETD mode is directly proportional to the duty cycle of the instrument, which in turn is directly related to activation time we used for ETD fragmentation. On the informatics front, several informatics pipelines were considered to mine for broad-scale PTM search. Every informatics pipeline has two components where the first part is peptide sequencing and the second part is peptide mapping. The protein identification process comes after the database searching algorithms have finished generating a list of peptide-spectrum matches (PSM). The protein identification/peptide mapping programs just assemble identified PSMs and map them onto proteins using user-defined thresholds. The process of peptide sequencing is the most time consuming and is also dependent on the type of MS/MS fragmentation employed.

Table 5.3 Summary of protein identification in the two microbial isolates using alternate proteases and different search pipelines

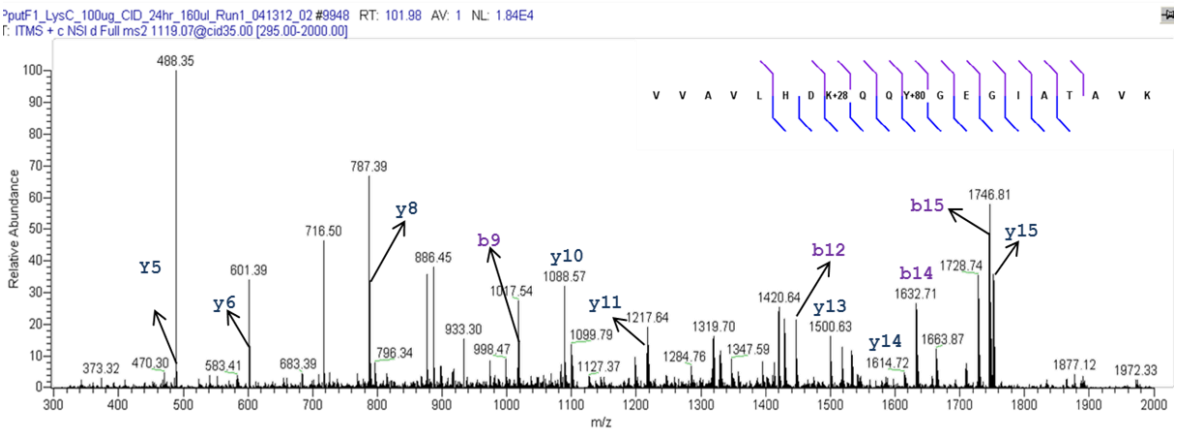
Organism	Protease	Fragmentation	Search Engine	Peptide Assembler	Protein Groups	Total Proteins	FDR
P. putida F1	LysC	CAD	OMSSA	Protein Herder	1075	1078	2%
P. putida F1	LysC	CAD	OMSSA	IDPicker	937	940	1.69%
P. putida F1	LysC	CAD	MyriMatch	IDPicker	904	907	1.53%
P. putida F1	LysC	ETD	OMSSA	Protein Herder	618	621	2%
P. putida F1	LysC	ETD	OMSSA	IDPicker	546	549	1.82%
P. putida F1	LysC	ETD	MyriMatch	IDPicker	610	612	1.63%
S. putrefaciens CN32	LysC	CAD	OMSSA	Protein Herder	1101	1105	2%
S. putrefaciens CN32	LysC	CAD	OMSSA	IDPicker	1154	1158	1.71%
S. putrefaciens CN32	LysC	CAD	MyriMatch	IDPicker	1148	1153	1.89%
S. putrefaciens CN32	LysC	ETD	OMSSA	Protein Herder	624	627	2%
S. putrefaciens CN32	LysC	ETD	OMSSA	IDPicker	613	616	1.92%
S. putrefaciens CN32	LysC	ETD	MyriMatch	IDPicker	675	689	2.03%
P. putida F1	GluC	CAD	OMSSA	Protein Herder	900	901	2%
P. putida F1	GluC	CAD	OMSSA	IDPicker	935	937	1.69%
P. putida F1	GluC	CAD	MyriMatch	IDPicker	988	989	2.02%
P. putida F1	GluC	ETD	OMSSA	Protein Herder	498	500	2%
P. putida F1	GluC	ETD	OMSSA	IDPicker	504	505	1.98
P. putida F1	GluC	ETD	MyriMatch	IDPicker	570	577	1.73%
P. putida F1	Trypsin	CAD	MyriMatch	IDPicker	1025	1027	1.92%
P. putida F1	Trypsin	CAD	OMSSA	IDPicker	1035	1037	1.91%
P. putida F1	Trypsin	CAD	OMSSA	Protein Herder	1003	1003	2.00%
P. putida F1	Trypsin	ETD	MyriMatch	IDPicker	167	168	1.18%
P. putida F1	Trypsin	ETD	OMSSA	IDPicker	168	170	1.17%
P. putida F1	Trypsin	ETD	OMSSA	Protein Herder	111	112	2.00%
S. putrefaciens CN32	Trypsin	CAD	MyriMatch	IDPicker	837	843	1.64%
S. putrefaciens CN32	Trypsin	CAD	OMSSA	IDPicker	811	816	1.70%
S. putrefaciens CN32	Trypsin	CAD	OMSSA	Protein Herder	813	813	2.00%
S. putrefaciens CN32	Trypsin	ETD	MyriMatch	IDPicker	125	126	1.56%
S. putrefaciens CN32	Trypsin	ETD	OMSSA	IDPicker	165	167	1.18%
S. putrefaciens CN32	Trypsin	ETD	OMSSA	Protein Herder	80	80	2.00%

In general, computational time is increased exponentially with the increase in number of variable modifications, and the number of missed cleavages.

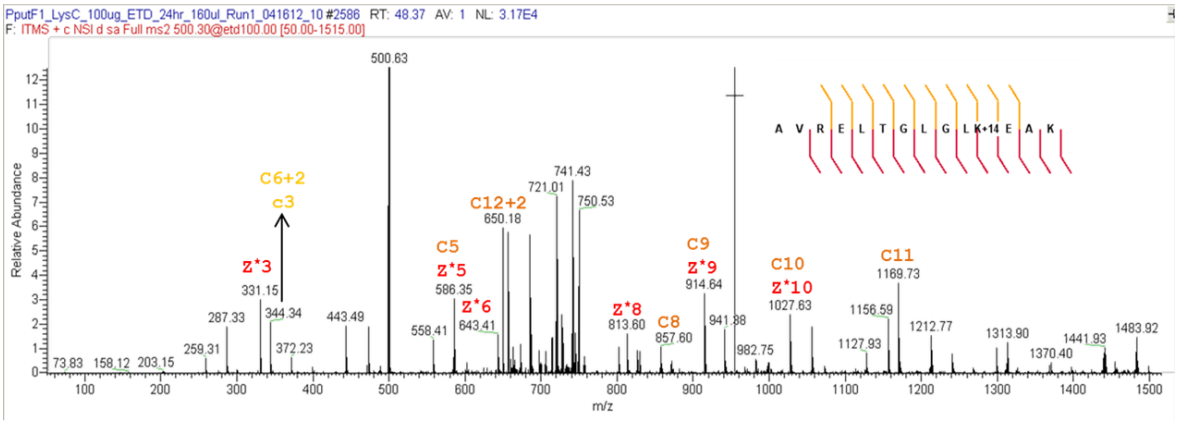
Also to note is the quality of spectra generated from an ETD fragmentation compared to a CAD fragmentation. ETD spectra are most often dominated by one charge reduced precursor ion peak followed by several small peaks at 5-10% level of the dominant peak.

As can be seen from the representative CAD spectra **(Figure 5.3 a)** of phosphorylated and dimethylated peptide, there is not only high complimentary between b and y ion pairs, but most of the assigned peaks are above 20% relative intensity. In contrast, the ETD spectra **(Figure 5.3 b)** of a monomethylated peptide from an ETD fragmentation of precursor ion at 500.3 m/z shows very low intensity level for most of the peaks. Just like the b and y ion complementarity in CAD scan, the ETD scan shows a very high complementarity in c and z ion pairs. However, unlike the peaks in a CAD scan, in an ETD scan almost all of the assigned ion pairs have their relative intensity below 12% of the maximum peak, which is the charge reduced precursor ion (cropped in this figure). The reduced spectral quality may have contributed to decreased identification from ETD runs. This observation also highlights the importance of using a high mass accuracy and high resolution instrument when using ETD as the fragmentation method.

Next the complementarity between peptides identified by ETD and CAD was evaluated. Since the two modes work differently for peptide fragmentation, it was expected that the two modes should identify a distinct set of modifications and **Figure 5.4** illustrates this observation in form of a Venn diagram. As can be seen, irrespective of the search pipeline, there is a high percentage of unique peptides identified by the two MS/MS mode in both *P. putida* F1 and *S. putrefaciens* CN-32.



(a)



(b)

Figure 5.3 Representative spectra of a modified peptide measured by (a) CAD and (b) ETD

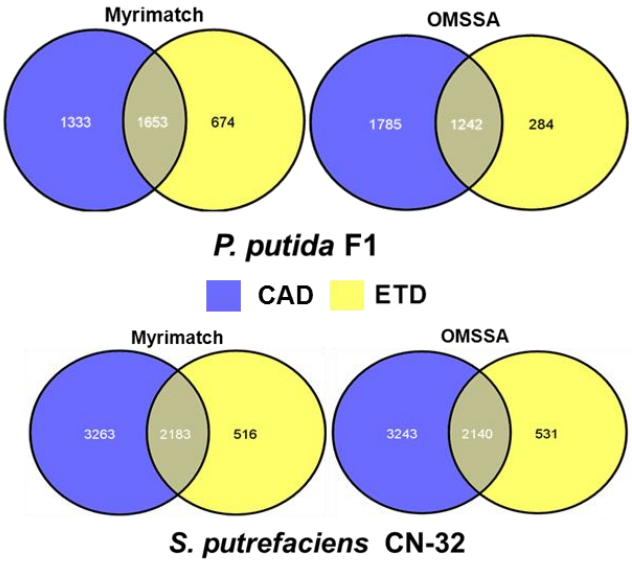


Figure 5.4: Unique and common peptides (total i.e. modified and unmodified) found by ETD/CAD fragmentation

Also to be noted is that there are a large number of peptides that are commonly identified by both modes, but overall CAD provides significantly more unique peptide identifications compared to ETD, and which in turn result in increased protein identification for CAD over ETD. Another interesting observation is that while for *S. putrefaciens* CN-32 the number of peptides is fairly similar between the two search algorithms, there is a slight variation in the number of identified peptides for *P. putida* F1 runs with the use of different search pipelines. Due to the complimentary nature of both the fragmentation modes, there was an increase in the overall sequence coverage of each organism. For example, we observed 17% and 9% increase in amino acid coverage of *P. putida* F1 and *S. putrefaciens* CN-32 samples digested by LysC respectively. The total proteins identified for each species represented all the major COG categories and the usage of different search algorithm did not significantly alter protein identifications. More than

90% of protein identifications for the same fragmentation scheme were common to the different informatics pipelines. While the total protein identifications were not significantly impacted by change in informatics pipeline, a significant variation was observed in the number of modified peptides for each informatics pipeline **(Figure 5.5)**. Between the two microbial species, the percentage of overlapping modified peptides was higher for *S. putrefaciens* CN-32 compared to *P. putida* F1. An analysis of modified peptides revealed that methylation is the most dominant PTM in microbial species **(Table 5.4).** As far as the search engine is concerned, since both

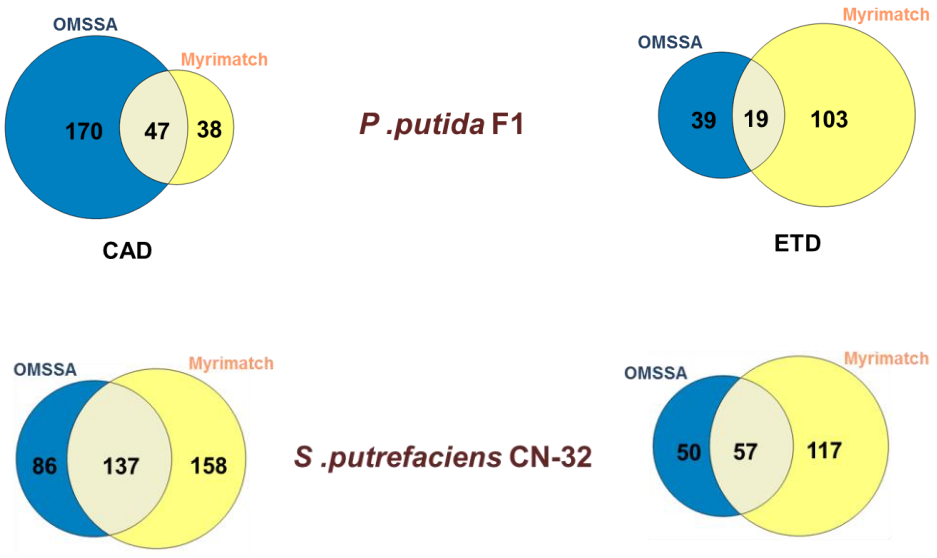


Figure 5.5 Overlap between modified peptides with respect to peptide fragmentation and search algorithm using LysC digestion.

Table 5.4 Distribution of PTM containing peptides in the two microbial species by ETD and CAD fragmentation

Organism	Protease	MS/MS	Search Engine	Assembler	Phospho	S	T	Y	Acetyl	K	Trimethyl	K	Dimethyl	K	R	Monomethyl	K	R
P. putida F1	*LysC*	CAD	MyriMatch	IDPicker	**9**	5	3	2	**7**	8	**11**	13	**12**	9	4	**14**	12	4
P. putida F1	*LysC*	ETD	MyriMatch	IDPicker	**15**	6	5	4	**19**	23	**16**	20	**28**	19	14	**31**	21	14
S. putrefaciens CN-32	*LysC*	CAD	MyriMatch	IDPicker	**44**	19	28	8	**31**	38	**26**	31	**52**	41	21	**52**	42	17
S. putrefaciens CN-32	*LysC*	ETD	MyriMatch	IDPicker	**35**	10	17	12	**33**	39	**17**	21	**31**	18	16	**32**	21	14
P. putida F1	*LysC*	CAD	OMSSA	Protein Herder	**40**	20	22	11	**8**	8	**16**	17	**35**	19	20	**42**	28	16
P. putida F1	*LysC*	ETD	OMSSA	Protein Herder	**9**	4	5	3	**1**	1	**3**	4	**10**	7	3	**17**	15	4
S. putrefaciens CN-32	*LysC*	CAD	OMSSA	Protein Herder	**36**	17	13	7	**8**	9	**21**	21	**14**	9	9	**21**	19	6
S. putrefaciens CN-32	*LysC*	ETD	OMSSA	Protein Herder	**16**	5	8	3	**1**	1	**6**	6	**7**	6	2	**13**	9	7
P. putida F1	*LysC*	CAD	OMSSA	IDPicker	**28**	15	15	8	**23**	28	**34**	34	**26**	21	12	**34**	19	20
P. putida F1	*LysC*	ETD	OMSSA	IDPicker	**7**	3	5	2	**5**	5	**15**	15	**14**	10	10	**22**	14	13
S. putrefaciens CN-32	*LysC*	CAD	OMSSA	IDPicker	**86**	41	42	26	**31**	33	**55**	62	**85**	68	28	**90**	73	32
S. putrefaciens CN-32	*LysC*	ETD	OMSSA	IDPicker	**23**	12	10	7	**19**	19	**27**	29	**41**	32	18	**37**	34	10
P. putida F1	*GluC*	CAD	MyriMatch	IDPicker	**34**	18	11	6	**34**	34	**22**	23	**63**	40	36	**63**	41	38
P. putida F1	*GluC*	ETD	MyriMatch	IDPicker	**20**	7	6	7	**13**	14	**13**	14	**23**	16	12	**43**	25	21

OMSSA and MyriMatch identify considerable number of unique PTMs, they are complimentary in nature and identification by both search engines gives more confidence that these modifications are real and not picked up by chance. Some of the modified peptides identified by LysC digestion via CAD and ETD are highlighted in **Table 5.5.** These peptides were identified with high spectral counts and their annotation suggests that they are among the most dominant proteins in these species. This shows that non-enrichment based PTM investigation in microbial species will probably capture only the most abundant proteins with PTM.

5.5 Conclusions

Our MS measurements using different proteases and fragmentation methods on experimental side coupled with multiple search programs on computational side showed that even the simplest form of life like unicellular microbial species exhibit a wider range of PTM diversity than previously thought. The ETD/CAD approach provides higher number of peptide identification and ~50% of identified peptides by CAD and almost 70% of identified peptides by ETD are common between the two fragmentation methods.

However, when it comes to PTM discovery, both the fragmentation schemes are highly complementary, which works to our advantage since it is highly challenging to identify PTMs in microbial systems. Therefore combining the two fragmentation methods lead to overall increase in identification of the PTM bearing peptides. The only downside of broad-scale PTM search in microbial species is that, even with multiple fragmentation and multiple proteases the measurements are limited compared to enrichment based methods. Another aspect to be kept in mind with ETD identified PTMs is that they need manual validation in comparison to CAD identified PTMs, which can be validated just by computational scoring. After evaluating two

Table 5.5 Representative modified peptides identified by ETD/CAD and different search pipelines using LysC digestion.

Phosphopeptides					
Organism	**Gene ID**	**Sequence**	**Search**	**MS/MS**	**Function**
P. putida F1	Pput_F1:640585633	VVAVLHDKQQ**Y**[80]GEGIATAVK (+2)	Myrimatch	CAD	Extracellular ligand-binding receptor
P. putida F1	Pput_F1:640586410	MIKKCLFPAAGYG**T**[80]RFLPATKAMPK (+3)	Myrimatch	CAD	UTP-glucose-1-phosphate uridylyltransferase
P. putida F1	Pput_F1:640588753	D**S**[80]NGQPAAISGAA**Y**[80]RSSSSK (+3)	OMSSA	CAD	Carbohydrate-selective porin OprB
P. putida F1	Pput_F1:640584899	QPAVIAEIKKASP**S**[80]K (+1)	Myrimatch	CAD	Indole-3-glycerol-phosphate synthase
S. putrefaciens CN-32	Sput_CN32:640497543	DVLPMVDGEIASGLRGGAELSARQQEMLAL**S**[80]DTLVAELK (+3)	Myrimatch	CAD	(Acyl-carrier-protein) phosphodiesterase
S. putrefaciens CN-32	Sput_CN32:640498186	SAKKRALQ**S**[80]EK (+2)	Myrimatch	ETD	ribosomal protein S20
S. putrefaciens CN-32	Sput_CN32:640499538	LLKALGANLVL**T**[80]EGAK (+3)	Myrimatch	ETD	cysteine synthase A
Methylated Peptides					
Organism	**Gene ID**	**Sequence**	**Search**	**MS/MS**	**Function**
P. putida F1	Pput_F1:640584932	AVRELTGLGL**K**[14]EAK (+2)	Myrimatch	CAD	ribosomal protein L7/L12
P. putida F1	Pput_F1:640584932	AVRELTGLGL**K**[14]EAK (+3)	Myrimatch	CAD	ribosomal protein L7/L12
P. putida F1	Pput_F1:640584932	AVRELTGLGL**K**[14]EAK (+2)	OMSSA	CAD	ribosomal protein L7/L12
P. putida F1	Pput_F1:640584932	AVRELTGLGL**K**[14]EAK (+3)	OMSSA	CAD	ribosomal protein L7/L12
P. putida F1	Pput_F1:640584932	AVRELTGLGL**K**[14]EAK (+3)	OMSSA	ETD	ribosomal protein L7/L12
P. putida F1	Pput_F1:640584932	AVRELTGLGL**K**[14]EAK (+3)	Myrimatch	ETD	ribosomal protein L7/L12
S. putrefaciens CN-32	Sput_CN32:640500489	VSSKLGEGIDTI**K**[14]GLLKDK (+2)	OMSSA	CAD	ribosome small subunit-dependent GTPase A
S. putrefaciens CN-32	Sput_CN32:640498489	GKALDEDL**R**[14]**R**[28]RQGEEQNK (+3)	Myrimatch	ETD	outer membrane chaperone Skp (OmpH)

distinct microbial systems with ETD, the gain with ETD approach is not as significant as we expected. However, this could also be due to the wrong choice of biological systems for ETD, since unlike enrichment-based PTM discovery studies in human cell lines that routinely find found several thousand modified peptides, PTM discovery in prokaryotes even with enrichment based methods have at best only found few hundred modified proteins [150]. Therefore, ETD being a slower process may have missed sampling low abundance modified peptides. To validate this hypothesis, a mixture of known peptides with PTMs should be subjected to ETD analysis at different concentrations. This can provide us some metrics with respect to limit of detection of ETD for PTMs.

Chapter 6 - Global survey of post-translational modification in a complex eukaryotic model plant system: *Populus trichocarpa*

6.1 Introduction to *Populus trichocarpa*: Systems level analysis of a complex eukaryote

Populus trichocarpa (black cottonwood) is a woody perennial plant and is the first tree species whose genome was completely sequenced in 2006 [151, 152]. *P. trichocarpa* represents a model system for biofuel research, owing to its smaller genome size, rapid juvenile growth, and ease of clonal propagation [153]. To modify/control traits that help in biofuel production, like drought tolerance, lignocellulosic content, resistance to pest, etc., one needs a clear understanding of the cellular processes underlying these phenotypes. Like many eukaryotic systems, plants are also programmed to use post-translational modifications (PTM) to alter these processes [154]. However, the complexity of their genome and myriad of PTMs available at a plant's disposal makes it highly challenging to characterize these modifications at global level. Poplar proteomics is further complicated by the fact that it has undergone whole genome duplications and therefore, has large number of paralogs within its genome [155]. It is reported that poplar genome exhibits single nucleotide polymorphism (SNP) every 200 base pairs [80]. This makes it highly challenging to ascertain diversity and extent of modifications and mutations in *P. trichocarpa*.

To address the complexity associated with identifying PTMs in eukaryotic systems, MS-based proteomics provides a unique capability unparalleled by any other analytical approach of surveying PTMs at global level [156]. Over the years, PTM discovery for the most common modifications, such as phosphorylation and acetylation, has become relatively straightforward

using enrichment based methods [157, 158]. When it comes to plants species, significant progress has been made in identifying phosphorylation sites in *Arabidopsis thaliana,* and databases like PhosphAT are a valuable starting point for probing the phosphorylation status of this plant [159]. But not much is known for most of the other plant species that are important commercially, like rice and wheat, or those that serve as a model for bioenergy research like poplar and switch grass. Even for less complex species, there remain challenges when searching for multiple-modifications or conducting a complete "blindPTM" search where all possible mass shifts on a protein are taken into consideration. One of the major challenge in carryout such complex searches is the lack of computational methods to perform multi-modification searches within a reasonable time.

6.2 Sequence tagging for PTM detection in discovery proteomics

A recent development in comprehensive PTM identification has been the onset of sequence tagging approaches that improves the speed of multiPTM or blindPTM searches by an order of magnitude compared to traditional database searching methods. These sequence tagging approaches fall in category of *de-novo* sequencing programs that infer peptide sequence directly from MS/MS data without requiring a protein database [160, 161].

Sequence tagging methods first try to define a minimal peptide sequence or a tag that can be confidently identified from the MS2 spectrum without any database searching methods. Next, the *de-novo* generated tag is further developed by addition of amino acids on the flanking ends of the tag. Further, any modifications are also considered if they lead to a peak matching in the MS2 spectra. This approach saves computational time by omitting matching of experimental spectra with theoretical spectra obtained by placing mass shifts on each candidate amino-acid.

In this work, we use the sequence tagging approach built into the DirecTag-TagRecon platform to search for pre-defined mass shifts on proteins, termed as PreferredPTM, in conjunction with a blindPTM search on the three organ types namely root, stem and leaf from *P. trichocarpa* [124, 125]. Proteomics measurements for this large scale PTM identification study were carried out on state of the art LTQ-Orbitrap Elite mass spectrometer using high mass accuracy on both the parent and the fragment ions.

6.3 Material and Methods

6.3.1 Sample preparation

Root, leaf and stem tissues from *P. trichocarpa* were ground under liquid nitrogen and then subjected to boiling at 95 °C for 10-15 minutes in 4% SDS solution made in 100 mM Tris buffer (pH 8.0) and 10 mM DTT. Based on protein estimation results using the BCA colorimetric assay, *P. trichocarpa* biomass corresponding to 3 mg total protein after SDS lysis was subjected to overnight TCA precipitation. Next day, the samples were spun at 21,000 g for 20 minutes. The supernatant was discarded and the pellet was washed twice with chilled acetone by centrifuging at 21,000 g for 10 minutes each time. The supernatant was carefully discarded taking care of the protein pellet. Residual acetone was air dried by leaving the tubes open for few minutes. Proteins were denatured and reduced using 250 µL of 8M Urea and 5mM DTT. To ensure complete denaturation, Poplar samples in 8M urea were briefly sonicated (20% amplitude, 5s ON, 10s OFF). Disulfide bond formation was blocked by addition of 20 mM iodoacetamide and incubation for 15 minutes in dark. An initial trypsin digestion was carried out for 4 hours at room temperature using 1 vial of 20 µg trypsin (1:75 [w:w], Promega Inc,) in 250 µL of Tris

buffer (100 mM Tris, 10 mM $CaCl_2$, pH 8.0) added to denatured samples. A second trypsin digestion was carried out overnight at room temperature using 1 vial of trypsin in 500 µL of Tris buffer (100 mM Tris, 10 mM $CaCl_2$, pH 8.0). 50 µL of an acid-salt solution (4M NaCl, 2% formic acid in HPLC H_20) was added to the digested samples. The addition of salt helped in reducing peptide adsorption to the MWCO filter while the acid helped in protonation of peptides. The tube was briefly vortexed and the contents of the tube transferred to a 10 kDa MWCO filter. The peptides were eluted from the filter by spinning at 4500 g for 30 minutes. BCA assay on peptides was performed to determine loading amount for each sample and the aliquots were stored at -80 °C until further use.

6.3.2 LC-MS/MS

Approximately 25-50 µg of peptides were pressure-loaded onto an integrated, self-packed 3 cm Reverse Phase (RP) resin (Aqua, 300 Å pore size, Phenomenex, Torrance, CA, USA) and 3 cm Strong Cation Exchange (SCX) resin in a 150 µm inner diameter fused silica back column. The peptides were desalted on the column by washing from solvent A (95 % HPLC H_20, 5 % AcN, 0.1 % formic acid) to solvent B (30 % HPLC H_20, 70 % AcN, 0.1 % formic acid) 3 times over a period of 25 min. The desalted back column was connected to a 15 cm-long 100 µm i.d. C18 RP resin PicoFrit column (New Objective, Woburn, MA, USA) and placed in line with a U3000 quaternary HPLC (Dionex, San Francisco, CA, USA). The SCX-RP LC separation was carried out using eleven salt pulses with increasing concentrations of a 500 mM ammonium acetate solution. Each of the first ten salt pulses was followed with 120 minute RP gradient from 100 % solvent A to 50 % solvent B, while the last salt pulse used 150 minute RP gradient from 100 % solvent A to 100 % solvent B. The LC eluent from the front column was directly nanosprayed into an LTQ-Orbitrap Elite mass spectrometer (Thermo Scientific). The mass spectrometer was

operated in a data-dependent mode under the control of Xcalibur software (Thermo Scientific). The following parameters were used for the data-dependent acquisition: High energy collision dissociation was carried out for top 10 parent ions in the Orbitrap (FWHM resolution of 15000) following a full scan in the Orbitrap at 30 000 resolution, a 3 m/z isolation width, 30 % collision energy; and a dynamic exclusion repeat count of 1 with a duration of 30 s.

6.3.3 Peptide Identification

Raw MS/MS data was searched against a predicted proteome of *P. trichocarpa* (v3 2012, downloaded from http://www.phytozome.net/cgi-bin/gbrowse/poplar, containing primary and alternate spliced gene models, mitochondria and chloroplast proteins. Overall the database size was 73,013 sequences including common contaminants like trypsin and keratin. Regular database searching was conducted using MyriMatch v 2.1 against a forward-reverse concatenated database and false discovery rates were estimated at the peptide level. MyriMatch configuration file included: fully tryptic peptides, +57 Da static modification for carbamidomethylation on Cysteine, +43 Da dynamic modification on N-terminus for carbamylation, +16 Da dynamic modification for Methionine oxidation.

Peptide sequence tagging was performed by DirectTag algorithm and the tags were developed to infer PTMs via TagRecon algorithm. [124, 125] The DirectTag algorithm generates 50 sequence tags (3 amino-acid long) along with their score per MS/MS scan, and stores them in tab-delimited text file. TagRecon then matches each sequence tag to a peptide sequences in the subset protein database (i.e. proteins identified by MyriMatch searching) and develops the tag by addition of amino acids to its flanking region both in the MS/MS spectra and the subset protein database.

In this work, TagRecon was operated in two modes: (a) PreferredPTM mode in which following mass shifts were provided in the configuration file to search: mono-methylation (14.0156 Da) of lysine and arginine, di-methylation (28.0312 Da) of lysine and arginine, tri-methylation (42.0468 Da) of lysine, acetylation (42.0105 Da) of lysine, phosphorylation (79.9663 Da) of serine, threonine and tyrosine, methionine oxidation (15.9949 Da) and N-terminal carbamylation (43.00582 Da). A static modification for C-carbamidomethylation (57.0214 Da) was also included in the config file. (b) In the second mode, TagRecon was set to search for all possible modifications as blindPTMs. The config file only included a static modification of C+57 Da, M +16Da and N-terminal carbamylation. All the other mass shifts were reported as BlindPTMs and were analyzed later to determine their identity. A peptide was set to have a maximum of 2 dynamic modifications.

6.3.4 Protein and PTM Inference

IDPicker 3.0 was used to assemble proteins from MyriMatch and DirectTag-TagRecon searching at 2% peptide level FDR. To further eliminate false positive PTM containing peptides, a set of attestation rules as described in Abraham *et al.* were applied to PTM containing MS/MS spectra. [77, 80, 150]

Rule 1: Spectra mapping to contaminant proteins were removed.

Rule 2: Spectra containing only C +57Da, M +16Da or N-terminal +43Da were eliminated.

Rule 3: Spectra showing K or R modification on C-terminal tryptic cut-site were eliminated.

Rule 4: The mass accuracy on parent ion giving rise to PTM containing spectra should be within $\pm$ 10 ppm.

Rule 5: If the same MS/MS spectrum was reported by both MyriMatch and TagRecon, then the XCorr of modification containing spectrum should be improved by 10% or more as compared to that in the regular database searching method.

Rule 6: A distinct modified peptide should have at least 3 spectral counts across two technical replicates.

Rule 7: The modified peptide in blindPTM search mode cannot be a mutation or be explained as a sample handling artifact.

6.4 Proteome measurement in the three organ types from *P. trichocarpa*

Black cottonwood, being a dioceous plant, exhibits a very high level of proteome complexity, as well as high genome redundancy. Transcriptomics and proteomics have shown that different organ types in Poplar have high number of unique proteins expressed alongside a set of proteins that present in all organ types.[162, 163] This heterogeneity in protein expression further expands the level of post-translational modifications both qualitatively and quantitatively by *P. trichocarpa.* Since our work relies on unenriched samples, we interrogated *P. trichocarpa* proteome for qualitative distribution of PTMs in leaf, stem and root. PTM searches involving unenriched samples rely on very comprehensive proteome coverage. We therefore, used LTQ Orbitrap Elite instrument, which has the capability of acquiring spectra both at high speed and at a high resolution. [164] We also employed HCD fragmentation to generate MS/MS spectra with complete sequence information to aid in PTM identification.[74]

Using high resolution and high mass accuracy on both the parent and the top 10 HCD generated fragment ions, we collected approximately 2 million spectra from leaf, stem and root samples of *P. trichocarpa* ran in technical duplicates via a 22.5 hour 2D LC-MS/MS method. This extensive

sampling gives an opportunity to mine for PTMs without enriching and thereby not limiting the scope of PTM diversity. Due to a very high sequence redundancy in *P. trichocarpa* genome, all the predicted proteins were grouped into clusters at 90% sequence identity cut-off via UClust.[165] Next all the proteins identified by mass spec were mapped to seed proteins of the clustered *P. trichocarpa* genome to provide accurate representation on the depth of the proteomics measurements. As showed in **Table 6.1**, our deep proteomics measurements resulted in identification of ~4000-5800 protein clusters in Poplar genome within each organ type. Proteomics measurements of stem tissue gave the highest number of protein identifications (~5800 protein clusters), followed by leaf and root. The replicate measurements were highly reproducible, as evident from the scatter plot of adjusted NSAF values of proteins that were common in both the replicates **(Figure 6.1)** [166]. The R^2 value which relates to the tightness of the two measurements was more than 0.9 in all the three organ types, suggesting very low variation in protein abundance for commonly identified proteins. However, a slightly higher variation in total protein count was observed for root samples. A closer inspection of mass spec data reveals that the measurements covered almost all the major pathways present in *P. trichocarpa* genome. Out of the total of ~73,800 proteins present in the search database, 22,820 genes have been mapped in PopCyc which represent proteins that are involved in specific functional pathways [167]. The unmapped proteins are either involved in non-enzymatic processes or not sufficiently characterized. Our measurements on three different organ types of Poplar identified ~8900 genes and half of these were mapped by PopCyc in major functional pathways.

As can be seen from **Figure 6.2**, identified proteins map to almost all the important functional pathways, except for the Brassinosteroid signaling pathway in *P. trichocarpa*. By color coding

Table 6.1 Summary of MyriMatch protein identification from the three different organ types of *P. trichocarpa*

Organ	Replicate	Clusters	Protein Groups	Proteins	Peptides	Total Unique Clusters 90% Seq. Identity
Leaf	**Replicate 1**	3422	4921	10308	24,922	4,424
Leaf	**Replicate 2**	3072	4390	9277	21,973	3,990
Stem	**Replicate 1**	4201	6392	12925	33,823	5,758
Stem	**Replicate 2**	4270	6501	12992	35,828	5,837
Root	**Replicate 1**	3653	5560	11294	30,848	4,997
Root	**Replicate 2**	3039	4560	9483	24,229	4,109

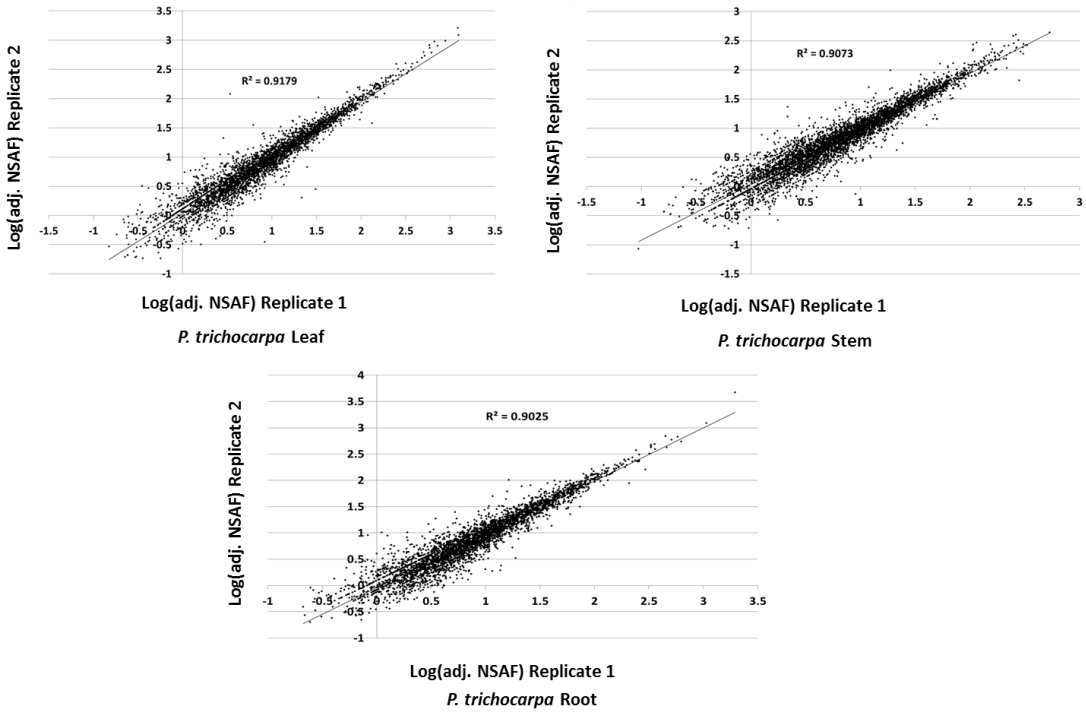


Figure 6.1 Reproducibility between replicate measurements.

Plot of NSAF values from proteins that were common between the two replicates.

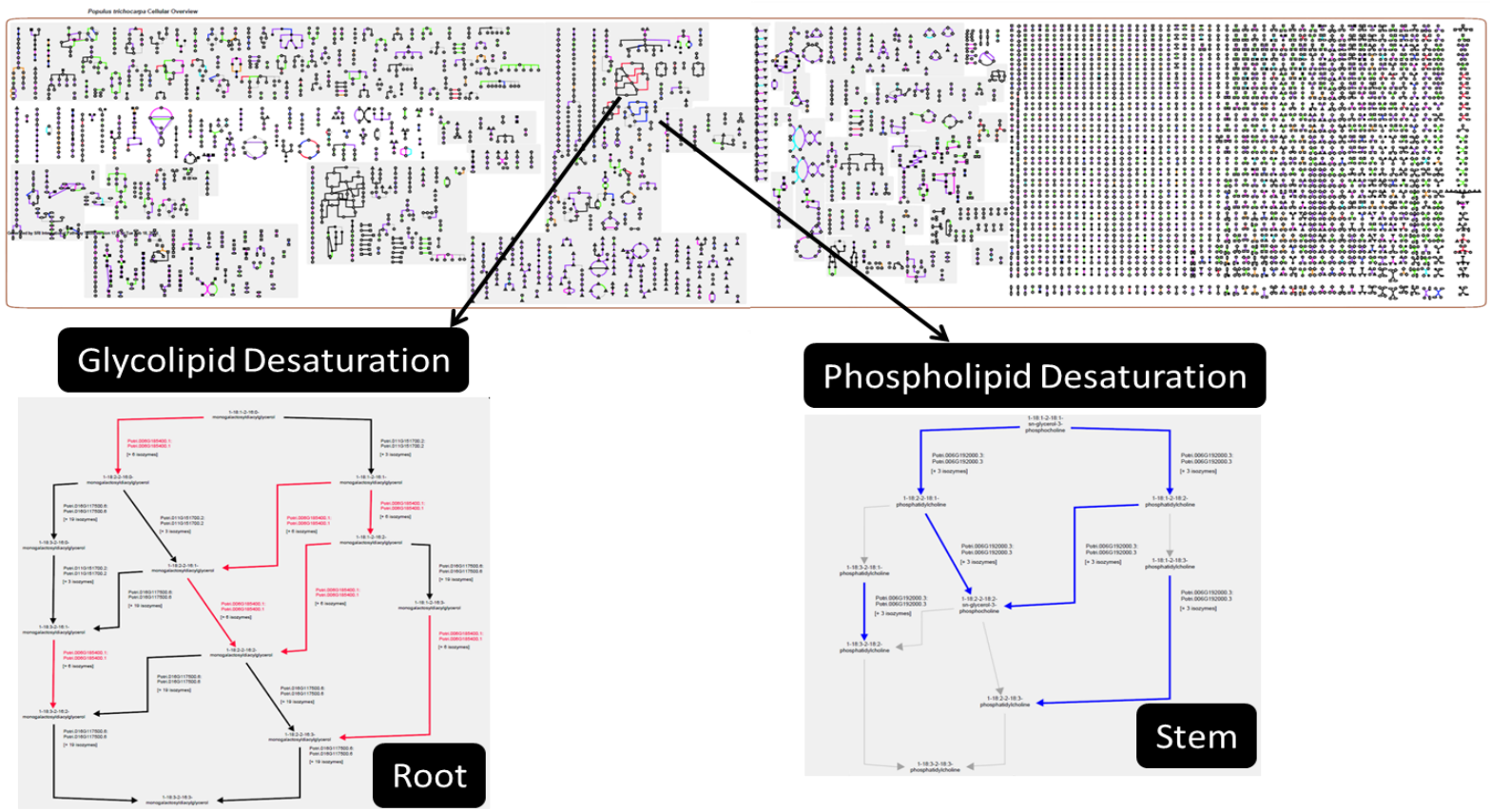


Figure 6.2 Overlay of MS-identified proteins (3471 proteins out of a total of 8262) from the three organ types (Leaf, Stem and Root) on pathway map of *P. trichocarpa*.

Bottom panel: Representative pathways that were uniquely present only in root and the stem tissue. (Colored arrows: MS identified proteins; Black arrows: Not identified by MS)

proteins that are unique and common across different organ types, one can highlight pathways that are confined to a specific organ type. For instance, proteins corresponding to the glycolipid desaturation pathway were exclusively found in the root sample while proteins present in the phospholipid desaturation pathway were only identified in the stem sample.

6.5 Targeted PTM identification in *P. trichocarpa* leaf, stem and root

We investigated methylation, acetylation and phosphorylation signatures in *P. trichocarpa*. A high-high strategy which means that both the precursor and tandem scan were performed with high mass accuracy and high resolution, allowed detecting high confidence PTM identification without using any enrichment strategy. Our results revealed methylation to be the most prominent modification out of the three types of modifications examined. Within methylation, mono-methylation of lysine residue was the most dominant modification. Ser/Thr phosphorylation was the most common phosphorylation event, while our measurement identified very few phospho-tyrosine peptides. Since a great amount of work has been done in the area of phosphoproteomics, including that in *Arabidopsis thaliana*, we chose phosphorylation as a key PTM for validation of our results. This was further bolstered by the presence of several bioinformatics tools designed for phosphopeptide prediction [168].

Out of a total of 97 phosphoproteins identified in the three organ types of *P. trichocarpa*, 15 (15.4%) of them had their nearest homolog in Arabidopsis with experimentally confirmed phosphopeptides in PhosphAt database, while almost 92% of *P. trichocarpa* homologs in Arabidopsis had presence of high confidence predicted phosphopeptides. 36 of the 97 proteins were represented in PopCyc, suggesting they are enzymes (mainly kinase, transferase, ATPase) that play a role in modulating reaction pathways. The remaining 61 phosphoproteins included 5

ribosomal proteins and 2 proteins involved in Light-Harvesting Complex of photosystem II, both of which are known to be key components of phospho-transfer. The criteria of high mass accuracy on precursor ion ($\pm$ 10 ppm) and minimum of 3 spectral counts along with biological relevance of our phosphopeptides further gives strength to the notion that enrichment based strategy is not always required to detect post-translation modifications.

Our measurements across three organ types identified 138 phosphosites mapping to 124 phosphopeptides, which were further mapped to 97 proteins. Of the 138 phosphosites, 28 sites were unique to leaf, 54 sites were unique to stem, and 30 sites were unique to root samples. Only 9 sites were commonly identified in all the three organ types.

A limited overlap in phosphorylation site is not surprising, since each of the three different organs not only have distinct external morphology, but are also internally composed of different tissue types. Therefore, these organs are supposed to have distinct signaling events to carry-out their specific functions.

As shown in **Table 6.2**, the most abundant acetylated proteins belonged to histone family, and there was a considerable overlap in acetylated proteins from the three organ types. The identification of histone proteins as acetylation targets further bolster the efficacy of sequence tagging approach in PTM discovery since histone family proteins are well known in literature to contain acetylation.[169, 170]. Apart from the proteins in the histone family, our measurements identified several other key proteins to bear acetylated peptides **(Table 6.2).** In the list was ATP synthase, which is an important enzyme in all the cellular organisms for ATP production. In plants, one of the major tasks assigned to leaves is energy production via photosynthesis and

Table 6.2 Most abundant acetylated peptides in the three organ types from *P. trichocarpa*

Organ Type	Peptide Sequence	Position	SpC Leaf	SpC Stem	SpC Root	Seed Protein	Function of Arabidopsis Homolog
Stem	GAKGLLTSK	**3**	-	**7**	-	Potri.018G032000.1	histone H2A protein 9
Leaf/Root	GSKPAPFSDIGKR	**3**	**13**	-	**13**	Potri.006G169400.1	voltage dependent anion channel 4
Leaf/Root	KGSITSVQAIYVPADDLTDPAPATTFAHLD ATTVLSR	**1**	**33**	-	**13**	Potri.008G126600.1	ATP synthase alpha/beta family protein
Leaf/Stem	KPAAAEKAPAEK	**7**	**41**	**46**	-	Potri.004G091400.1	Histone superfamily protein
Stem	KPAEKKPAAAEK	**5**	-	**8**	-	Potri.004G091400.1	Histone superfamily protein
Root	KPIILMPRR	**1**	-	-	**15**	Potri.001G148800.1	pleckstrin homology (PH) domain-containing protein
Leaf/Stem/Root	KQLATKAAR	**6**	**24**	**9**	**12**	Potri.001G016700.1	Histone superfamily protein
Leaf/Stem	SKDVIEEGQTHTK	**2**	**15**	**11**	-	Potri.004G176300.1	plasma membrane intrinsic protein 3
Root	SLKGELETVIELK	**3**	-	-	**92**	Potri.008G131100.1	MLP-like protein 43

therefore, detection of acetylation in ATP synthase in leaf samples may be crucial for activity of this enzyme [171, 172].

One of the acetylated peptides, "SL**K**GELETVIELK" was identified uniquely with a very high abundance (92 Spectral Counts) in root samples mapped to poplar protein Potri.008G131100.1. A BLAST search of this protein against RefSeq database matched it to Major-Latex Protein (MPL) family. The MLP proteins were previously found to dominate in root tissue in cotton, and are known to be involved in plant defense, ligand binding and growth. In light of these functional roles, acetylation might be critical for this enzyme [173].

Compared to phosphorylation and acetylation, methylation was the most dominant PTM expressed in all the three organ types. Our PreferredPTM search strategy identified a total 231 distinct methylated peptides, of which the leaf sample had 135 methylated peptides, the stem sample had 106 methylated peptides, and the root sample had 78 methylated peptides. Out of a total 231 distinct methylated peptides, only 18 of the methylated peptides were common among the three organ types, 80 were uniquely identified in leaf, 40 were uniquely identified in stem and 41 were uniquely identified in root.

Pathway analysis revealed several pathways in *P. trichocarpa* with all the proteins bearing one or more modifications. **Figure 6.3** highlights modified proteins from the three organ types i.e. Leaf, Stem and Root on pathway map and the bottom panel shows some of the representative pathways in which majority of proteins bear PTMs. An in-depth pathway analysis reveals the merit of using unbiased broad PTM search strategy. As shown in **Figure 6.4**, gluconeogenesis I pathway, which is involved in energy metabolism using non-sugar sources, has enzymes that are

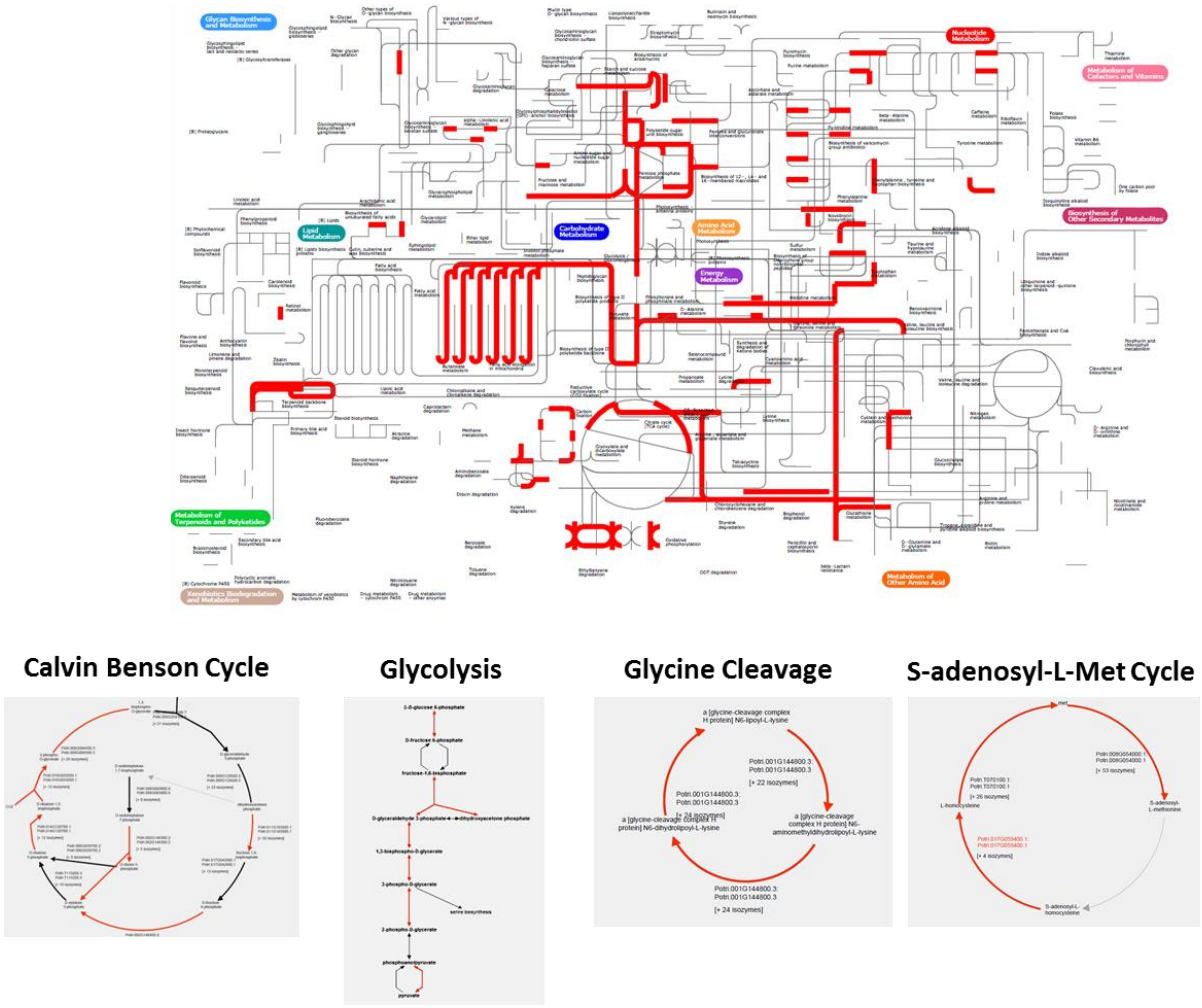


Figure 6.3 Overlay of PTM bearing *P. trichocarpa* proteins on cellular network. *Bottom panel shows some of the representative pathways in which most of the steps are mediated by enzymes bearing PTMs.*

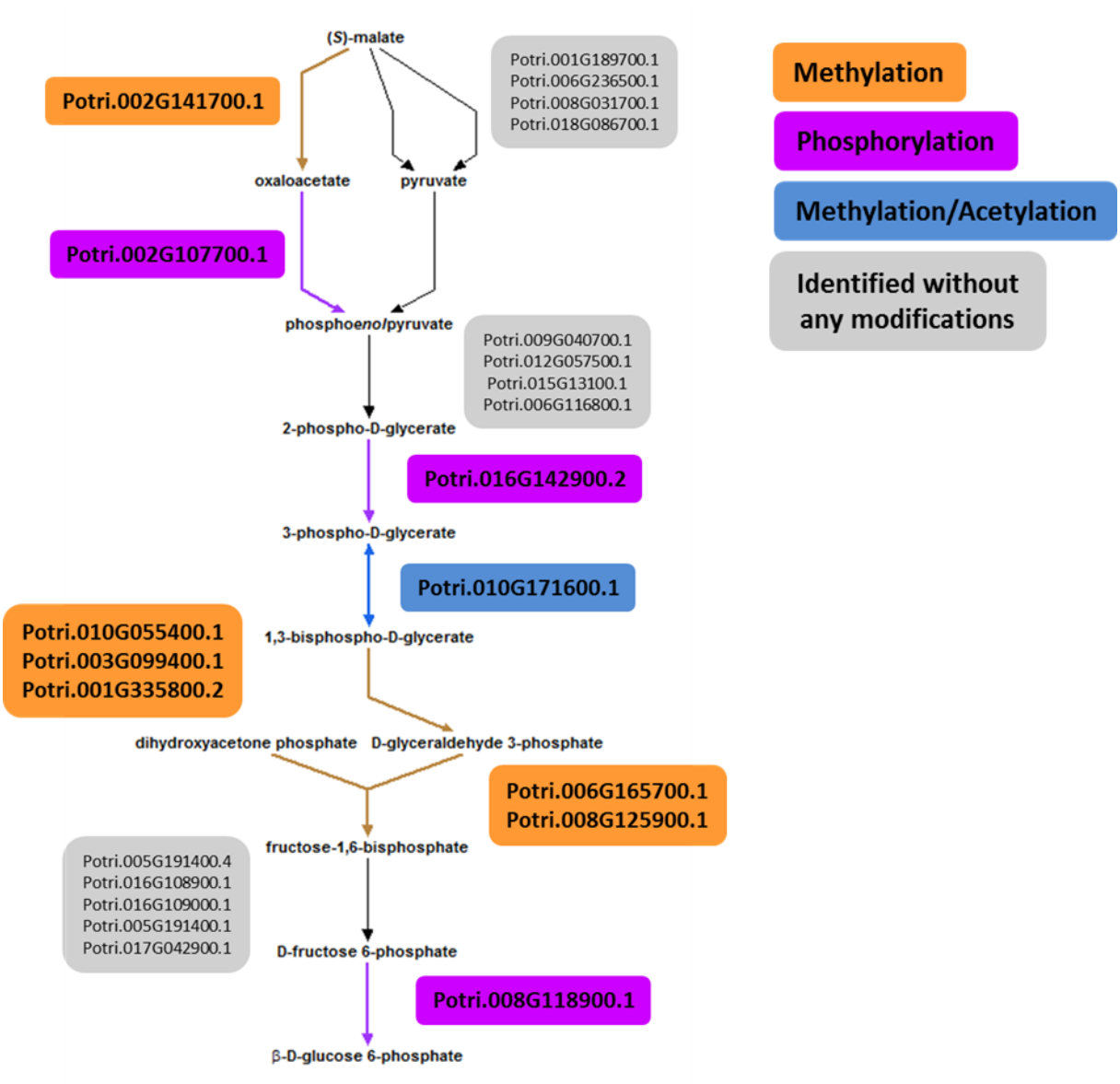


Figure 6.4 Gluconeogenesis I pathway in *P. trichocarpa* showing enzymes bearing multiple type of modifications.

modified by one or multiple modification. An enrichment based strategy will miss out on information with respect to combinatorial action of PTMs in plant physiology.

6.6 Blind PTM search in *P. trichocarpa* genome

Blind PTM search mode is an exhaustive search where full range of mass shifts is explored to match as many ions as possible from tandem mass spectra. Since there is no imposition of residue type and mass range, conventional database search methods can employ true blindPTM search mode. However, the computational pipeline used in our work overcomes this challenge by building upon sequence tags that are generated independent of search database. In our study, almost 50% of the collected spectra contained some type of modification (static or dynamic) in blindPTM mode, as compared to 25-30% of modified spectra in PreferredPTM mode. This highlights the fact in a given mass spec run, there are far more peptide matching spectra than that are assigned, and one of the major cause of unassigned spectra is not taking into account the variety of PTMs that decorate peptides. The Poplar genome is rich with single amino-acid polymorphisms, and studies at genetic and proteomic levels have shown this behavior. Our blindPTM search strategy further confirmed this finding that a majority of mass shifts were in fact amino-acid mutations, while the remaining ones posed significant challenge for characterization.

Table 6.3 illustrates some of the most abundant blindPTM mass shifts that were identified form our proteomics measurement. As reported in this table, large number of mass shifts corresponds to amino-acid mutations, while several other mass shifts are not annotated. One of the most abundant mass shift in blindPTM search mode was that of +40Da mass shift on glycine, which

Table 6.3 Representative most abundant blindPTM identified in the three organ types of *P. trichocarpa*

Organ	Peptide	Position	Residue	Mass Shift	SpC1	SpC2	Total SpC	Comment
Root	FEKEAAEMNKR	3	K	14	4	54	58	Methylation
Leaf	VGGTNHSHATQDLYDSIAAGTYPEWK	2	G	-3	57		57	
Leaf	VIERFEKEAAEMNKR	7	K	14	20	36	56	Methylation
Stem	KPAAAEKAPAEK	7	K	42	31	17	48	Acetylation, Trimethylation
Root	FIKDYAHVADAIEPVK	7	H	-24	29	18	47	His --> Xle; His --> Asn
Root	GHYTEGAEMIDSVLDVVR	9	M	-18		46	46	Met --> Xle
Stem	FIKDYAHVADAIEPVK	7	H	-24	23	22	45	His --> Xle; His --> Asn
Stem	IGLAGLAVMGQNLALNIAEK	2	G	40		43	43	Gly --> Pro; Sample Handling
Leaf	MAELCGFDTLDSLIDATVPK	5	C	112	26	15	41	N-methylmaleimide, DMPO
Root	IFSNPAIAAEEPWYGIEQEYTLLQK	14	Y	-16	18	23	41	Tyr --> Phe
Stem	FEYVDNVQPAEMISGGPQVISHVSPPK	21	S	27	15	21	36	Ser --> Asn, Ethylamino, Formyl, Ser --> Asp
Stem	QLVQEALNTHQFSTAPK	3	V	-17	21	14	35	
Leaf	QFNGLIDVYRK	2	F	-17	15	19	34	Phe --> Met
Leaf	FIKDYAHVADAIEPVK	7	H	-24	14	18	32	His --> Xle; His --> Asn
Root	GPELLTMWFGESEANVR	9	F	16	21	11	32	Oxidation
Root	QLVQEALNTHQFSTAPK	3	V	-17	18	14	32	
Leaf	AGEKLGLDVKTISVPNPR	10	K	60	16	15	31	Hydroxytrimethyl
Leaf	VAILGAAGGIGQPLAMLMK	2	A	40	20	11	31	
Root	SGFEGPWTANPLIFDNSYFK	2	G	40	14	17	31	Gly --> Pro; Sample Handling
Stem	LPSPTFNIAQLQQSFSQR	5	T	-13	15	16	31	
Stem	QFNGLIDVYKK	2	F	-17		31	31	Phe --> Met
Leaf	ISLNEQLLNHVTTLSR	8	L	-42	16	14	30	Xle --> Ala
Root	IVGCIPQILNPNPDAMSK	4	C	14	16	14	30	Methylation
Leaf	EQIFEMPTGGAAIMR	14	M	-48	18	11	29	Met --> Hsl; Dethiomethyl
Root	LLVPLVSAFRYEGEEVNTILAK	19	I	34	8	20	28	Xle --> Phe
Stem	IVGCIPQILNPNPDAMSK	4	C	14	14	14	28	Methylation
Root	MLQSYLGAKNFQR	10	N	-26	14	13	27	
Stem	NGPSMMPGGSFEAFK	12	E	-14	16	11	27	Glu --> Asn; Glu --> Asp
Stem	TWGGRPENVQDAQETLLIR	11	D	-44	7	20	27	Asp --> Ala
Root	LSGDAASVDIATPQNLQRLK	17	Q	-85	19	7	26	
Root	QFNGLIDVYRK	2	F	-17	15	10	25	Phe --> Met
Root	LSAPNFYYDLEDILAGEVAK	2	S	27	12	12	24	Ser --> Asn, Ethylamino, Formyl, Ser --> Asp
Stem	MLQSYLGAKNFQR	10	N	-26	10	14	24	
Stem	FEKEAAEMNK	3	K	14	10	13	23	Methylation
Stem	LKEVEAVCNPIITAVYQR	8	C	112		22	22	N-methylmaleimide, DMPO

Also corresponds to Gly→ Pro mutation. Since such a mutation will be very drastic for tertiary structure of protein, it was not clear why this mutation/mass shift was prominent. Furthermore, this mass shift was present in all the three organs, and was almost always present on glycine when it was the 2nd residue in the peptide sequence. A thorough literature survey revealed that the usage of acetone in sample preparation can also lead to formation of +40 Da adduct on glycine, which provides a more reasonable explanation then a Gly → Pro mutation [174]. This observation highlights the danger of over-interpretation of Blind PTM data and need for thorough investigation in understanding meaning of all the mass shifts.

6.7 Conclusions

Our proteomics pipeline detected more than 10,000 proteins from the poplar genome. To accurately represent MS measurements and biological significance, these 10,000 proteins were reduced to ~4000-6000 protein clusters based on 90% sequence identity. Further, these detected proteins were mined for PTMs via sequence tagging approach, which proved viable, if used with careful manual validation. The PreferredPTM mode can rapidly and accurately decipher the extent of major PTMs, like phosphorylation, methylation and acetylation, without the need of time-consuming and expensive enrichment based methods. The blindPTM search provides an elaborate list of modified peptides that require in-depth assessment to distinguish them from sample preparation induced artifacts and true biologically relevant modifications.

Chapter 7 - Strategies to delineate alternate coding sequences in a microbial community

---

Portions of text are adapted from: Hanke A, Hamann E, **Sharma R**, Geelhoed JS, Hargesheimer T, Kraft B, Meyer V, Lenk S, Osmers H, Wu R, Makinwa R, Hettich RL, Banfield JF, Tegetmeyer HE and Strous M. Recoding of the stop codon UGA to glycine by a BD1-5/SN-2 bacterium and niche partitioning between Alpha- and Gammaproteobacteria in a tidal sediment microbial community naturally selected in a laboratory chemostat. *Frontiers in Microbiology* **5** (231): 2014

Ritin Sharma's contributions included: The sample extraction, bioconductor experiments and metagenomics studies were carried out by research group led by Marc Strous. The microbial samples were provided to ORNL and the experimental approach for proteomics was designed by Ritin Sharma. Ritin Sharma carried out all MS measurements and performed computational searches against predicted protein database provided by the collaborators. Ritin Sharma compiled the table of proteins and peptides which were mined for biological significance by the Strous group. To decipher alternate codon usage by microbial members in this system, the *in-silico* translation of proteins to different version of UGA was done by the Strous group and the predicted database was searched with MS data by Ritin Sharma and the final list of unique peptides with alternate codon usage was prepared. Ritin Sharma also wrote the proteomics method section of the paper.

7.1 Introduction to the microbial diversity in Wadden Sea tidal flat

The Wadden Sea along the northern European coast is the largest tidal system worldwide and has been a UNESCO world heritage area since 2009. It receives nutrients, mainly in the form of nitrate, phosphate, and silicate from a large catchment area in northern Europe, stimulating growth of algae and other microorganisms in the surface water. Tidal pumping of this water through the permeable sediments of tidal flats leads to the continuous removal of nutrients by

highly active indigenous benthic microbes. For example, the measured *in situ* denitrification rates are very high, up to 60 µmol m^{-2} h^{-1} [175].

Over the past decades, the biogeochemistry and the microbial diversity of the Wadden Sea tidal flat have been studied intensively. The microbial community in the upper oxic tidal flat sediments was shown to be dominated by populations of Gammaproteobacteria [176] and Flavobacteria [177, 178]. The Gammaproteobacteria belonging to the 'insertiae sedis' clade make up for an important part (ca. 39.6 %) of all Gammaproteobacteria found on Janssand. This clade also includes the ubiquitous marine SUP05 cluster and many bacterial symbionts of marine invertebrates. These bacteria are known to be facultative aerobic chemolithoautotrophs that can use inorganic electron donors such as sulfide and hydrogen to perform denitrification [179, 180]. Only a very few species from this clade have been isolated in pure culture. The Flavobacteria on the other hand are well known to be involved in the degradation of macromolecules such as polysaccharides [181].

7.2 Approaches to study microbial communities in native and controlled environment

To understand the overall function of individual bacterial species in complex natural microbial communities, microbiology has traditionally depended on the isolation of target microbes in pure culture. Because such isolation is often unsuccessful, metagenomic genome reconstruction [33, 182, 183] and single cell genomics[184, 185] have been used to unravel the metabolism of key community members without their isolation. Metagenomic genome reconstruction can yield near-complete [117, 182] and, in some cases, complete genomes [186, 187] without the amplification artifacts that plague single cell genomics studies. Extensive genome reconstruction for coexisting organisms is possible, even for quite complex consortia. However, sequencing

methods alone lack the ability to probe metabolic function and provide limited insight into microbial interactions.

Both with conventional pure cultures and single cell methods, the target microorganism is, literally, isolated from its natural context and it is often difficult to understand the ecological niche of the isolated microorganism. For this, it would be ideal to study the organism in the context of its natural habitat. Because of the dynamics and complexity of natural communities this is generally not straightforward.

Engineering of a natural ecosystem in the laboratory can enable experiments that include uncultivated members. This approach has been used to study biofilm communities [188]. A consortium-cultivation approach could also be used to select for simplified versions of complex natural microbial communities, simplifying recovery of genome sequence information for the abundant populations via metagenomics. Enrichment also has the advantage that it brings some populations to high enough abundance levels to enable study by proteomics. This allows for characterization of overall community metabolic function, interactions, and provides a route to unravel the contribution of each of the individual members. Once an engineered system is available, key environmental factors that define the ecological niche of selected populations can be identified by manipulating the applied conditions.

To achieve a significant substrate turnover at low, near *in situ* substrate concentrations, habitat engineering depends on continuous culture cultivation, e.g. a chemostat. The low substrate concentrations prevent the selection of R-strategists. Further, in a continuous culture the applied conditions are stable (or dynamic, if desired) and reproducible for an indefinite amount of time.

In the present study we simulated the environmental conditions of a tidal flat sediment in a simplified form. The most significance difference between the simulated and the natural environment was that in the natural habitat, microbes grow as thin biofilms on sand grains, whereas in the simulated habitat, cells grew in suspension. Nitrate and nitrite were supplied as the main electron acceptors and oxygen was supplied twice daily, mimicking tidal cycling. The carbon and energy source consisted of a mixture of glucose, amino acids and acetate, in a ratio similar to decaying biomass, the main carbon and energy source in situ. After 23 days, the resulting community was shown to be dominated by representatives of phylogenetic clades also detected in the tidal flat sediments, including a member of the enigmatic bacterial BD1-5/SN-2 clade which lacks cultivated representatives and was previously predicted to translate the stop codon UGA into glycine. Proteomic analysis of the simplified microbial community enabled the experimental validation of this prediction.

7.3 Materials and Methods

7.3.1 Sample collection and establishment of chemostat

Sediment samples were taken from an intertidal flat in the central German Wadden Sea known as "Janssand" located south of the Eastern Friesian Island Spiekeroog (N: 053° 44' 151" / E: 007° 41' 945"). For direct metagenomic sequencing one sample was collected on October 24 2009 (0-5 cm depth), and three samples on March 23 2010 (0-2cm depth). The turbid, sand free supernatant (the "cell extract") was filled in 800 ml portions into 1 l transfusion bottles (Ochs Glasgerätebau, Bovenden/Lenglern, Germany) and pH was set to 8.1-8.4 with 1 M NaOH solution. The cell extract was made anoxic by alternately applying vacuum to 0.3 bar and argon to 1.2 bar, 3 times each. Each bottle was supplemented with $NaNO_3$ stock solution to reach a

final concentration of 0.1 mM serving as electron acceptor. 50 mg/l cycloheximide (AppliChem, Darmstadt, Germany) was added and the cell extract was incubated at 4 °C for 24 h to kill predatory eukaryotic organisms. After that, the cell extract was used for inoculation of the continuous culture.

7.3.2 Proteomics sample preparation

Based on the protein estimation results from cultures, an aliquot corresponding to 300 µg total protein was used for proteomics sample preparation via the Filter-aided Sample Prep method (FASP) [92], as described in Chapter 2. Protein digestion was carried out first for 4 h at 37 °C using trypsin (Promega) in 1:20 protease to protein ratio. A second aliquot of trypsin was added following first 4 hours and sample was incubated at 37 °C for an overnight digestion. Peptides were collected in a fresh tube after washing the filter with two washes of 50 mM ammonium bicarbonate and final addition of 0.5 M NaCl and spinning at 14 000 g. The pH of resulting peptides solution was adjusted to < 3 by addition of formic acid.

7.3.3 2D - LC-MS/MS

Approximately 25 µg of peptides were pressure-loaded onto an integrated, self-packed 3 cm Reverse Phase (RP) resin (Aqua, 300 Å pore size, Phenomenex, Torrance, CA, USA) and 3 cm Strong Cation Exchange (SCX) resin in a 150 µm inner diameter fused silica back column. The peptides were desalted on the column by washing from solvent A (95 % HPLC H20, 5 % AcN, 0.1 % formic acid) to solvent B (30 % HPLC H20, 70 % AcN, 0.1 % formic acid) 3 times over a period of 25 min. The desalted back column was connected to a 15 cm-long 100 µm-I.D. C 18 RP resin PicoFrit column (New Objective, Wobum, MA, USA) and placed in line with a U3000 quaternary HPLC (Dionex, San Francisco, CA, USA). The SCX-RP LC separation was carried

out by eleven salt pulses with increasing concentrations of 500 mM ammonium acetate solution. Each of the first ten salt pulses was followed with 120 minute RP gradient from 100 % solvent A to 50 % solvent B, while the last salt pulse used 150 minute RP gradient from 100 % solvent A to 100 % solvent B. The LC eluent from the front column was directly nanosprayed into an LTQ-Orbitrap Elite mass spectrometer (Thermo Scientific). The mass spectrometer was operated in a data-dependent mode under the control of Xcalibur software (Thermo Scientific). The following parameters were used for the data-dependent acquisition: collision induced dissociation was carried out for top 20 parent ions in the ion trap following a full scan in the Orbitrap at 30 000 resolution, a 0.5 m/z isolation width, 35 % collision energy was used for fragmentation; and a dynamic exclusion repeat count of 1 with duration of 30 s.

7.3.4 Database searching

The raw MS/MS data was searched using MyriMatch v2.1[77] against a predicted protein database (28 627 sequences with bin E, Figure 7.1 translated thrice: UGA as STOP codon, UGA coding for glycine, UGA coding for tryptophan) constructed from metagenome assembly, along with common contaminants (44 sequences) and reverse sequences. A second search was performed using MyriMatch v2.1 against a predicted database same as before, with the only difference being that the binE was translated 20 times with UGA coding for all twenty amino acids as well as UGA acting as a STOP codon. A fixed modification of +57.0214 Da for carbamidomethylation of cysteine and a +16 Da modification for oxidation of methionine and a +43 Da modification for N-terminal carbamylation were included as dynamic modifications in the search parameters. Identified peptides were then filtered at <1 % peptide level FDR and assembled into proteins (minimum of two peptides per protein) by IDPicker 3.[80]

7.4 Distribution of microbial population in an artificially regulated community

After 14 days of stable oxic/anoxic cycling (23 days after inoculation) the enriched microbial community was characterized by metagenomic sequencing and proteomic analysis. The N50 contig length of the assembly was 4.1 kb at 5.8x sequencing coverage and in total 13.2 Mb of unique sequence data was assembled. Three full length 16S rRNA genes were detected in the *de novo* assembly, as well as several fragments that were assigned to Roseobacter populations; these were combined into 2 additional full length genes by alignment to reference sequences of closely related bacteria. In parallel, iterative read mapping yielded two full length 16S rRNA genes that were >97 % identical to two of the assembled ones. By binning of contigs based on tetranucleotide frequencies five bins were generated, each with a distinct phylogenetic profile that aligned well with the recovered 16S rRNA gene sequences. **Figure 7.1** visualizes the position of each bin on the GC versus coverage plot of the assembly as well as the per-bin phylogenetic profile calculated. Based on read mapping, the binned populations were estimated to make up approximately 93 % of the overall community.

The most abundant population in the culture (corresponding to bin A) belonged to the genus Maritimibacter (Alphaproteobacteria). Bins B and C corresponded to two different Roseobacter populations, bin D to a Gammaproteobacterium insertiae sedis and bin E to a representative of the enigmatic bacterial BD1-5/SN-2 clade which lacks cultivated representatives.

In accordance with abundance based on metagenomic reads, proteomics measurement also showed a similar trend. As shown in **Figure 7.2**, highest numbers of proteins were identified from bin A (1435) followed by bin B (1230) and bin D (1183). bin C (783) and bin E (747) had the lowest number of identified proteins reflecting their lowest abundance in the community.

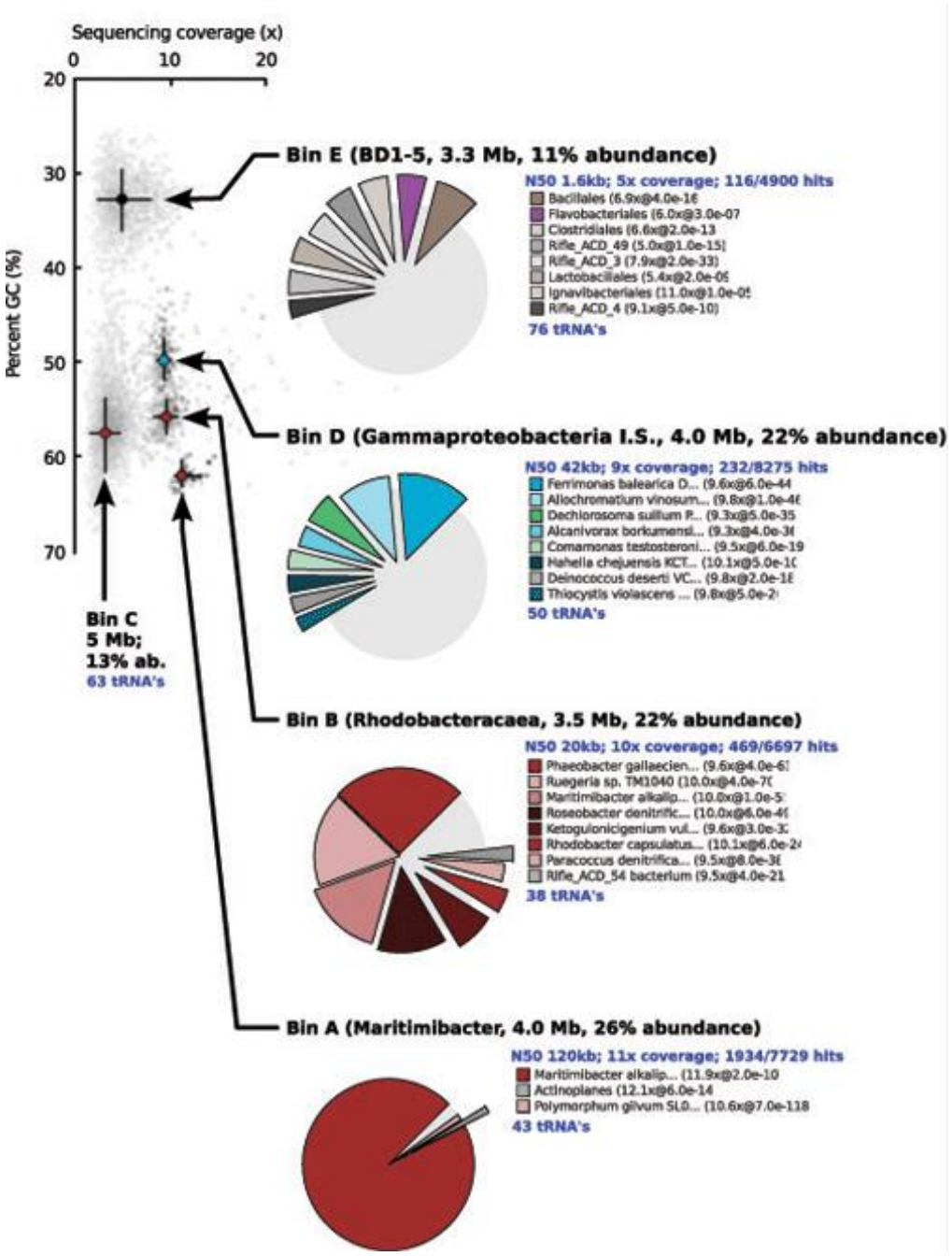


Figure 7.1 GC versus coverage plot showing scattering of the contigs into distinct "clouds", each associated with a different bin.

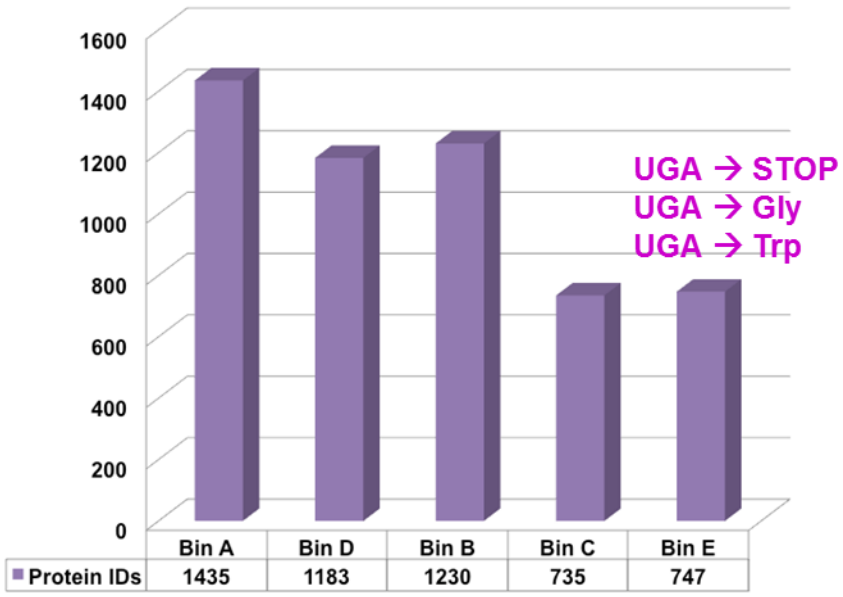


Figure 7.2 Total number of protein identified for each bin in two technical replicate measurements.

The number of proteins in bin E is inflated, as it represents three versions of same protein as described in later part of this chapter. The partitioning of metabolism over the different populations was inferred from the proteome. **Table 7.1** shows the average peptide coverage and detected proteins for distinct subsystems for each of the bins. Proteomic analysis suggested that all populations except BD1 5/SN 2 were competing for oxygen, whereas the denitrification pathway may have been characterized by a combination of competition and cross-feeding as can be seen from different steps of denitrification apparently performed by different organisms.

The expression of transporters, the enzymes involved in glycolysis, the citric acid cycle and respiratory complex I suggested that these four populations were also competing with each other for the supplied carbon substrates. The three Rhodobacterales presumably used the substrates as energy and carbon source. In addition, the Maritimibacter population (bin A) may also have used sulfide as additional energy source, as shown by the expression of a sulfide dehydrogenase. The

Table 7.1 Proteome coverage of binnable populations at t=23 days (.n.d. = not detected)

Bin		**A**	**B**	**C**	**D**	**E**
Taxonomic clade		Maritimi-bacter	Roseo-bacter-Clade	Roseo-bacter-Clade	Gamma-proteo-bacteria I.S.	BD1-5/ SN-2
Expressed subsystems				**# Expressed proteins detected**		
Ribosomal proteins (64)	(#)	10	12	4	48	27
Cell division & growth (30)	(#)	23	23	14	25	12
tRNA metabolism (34)	(#)	27	26	12	28	6
F0F1 ATP synthase (7)	(#)	7	7	2	6	2
Complex I (11)	(#)	8	5	8	6	n.d.
Complex IV (4)	(#)	3	3	3	2	n.d.
Oxygen stress (11)	(#)	6	5	4	10	6
Nitrate reductase (3)	(#)	n.d.	1	1	3	n.d.
Nitrite reductase (1)	(#)	1	1	n.d.	1	n.d.
Nitric oxide reductase (1)	(#)	1	1	n.d.	1	n.d.
Nitrous oxide reductase (1)	(#)	1	1	1	1	n.d.
Sulfide oxidation (13)	(#)	1	n.d.	n.d.	13	n.d.
Hydrogen oxidation (4)	(#)	n.d.	n.d.	n.d.	3	n.d.
Formate oxidation (5)	(#)	n.d.	n.d.	n.d.	5	n.d.
CO dehydrogenase (3)	(#)	3	n.d.	3	n.d.	n.d.
Calvin Cycle (3)	(#)	n.d.	n.d.	n.d.	3	n.d.
Citric acid cycle (23)	(#)	16	15	13	17	n.d.
Sugar metabolism (24)	(#)	18	14	9	11	n.d.
Amino acid metabolism (44)	(#)	30	32	26	30	4
Transporters				**# Expressed proteins detected**		
Sugars (9)	(#)	6	3	6	2	n.d.
Aminoacids (7)	(#)	5	4	3	3	n.d.
Di/tricarboxylates (4)	(#)	4	4	3	2	n.d.
Glycine-betaine (4)	(#)	2	1	2	1	n.d.
Oligopeptides (3)	(#)	2	2	2	1	n.d.
Purines (1)	(#)	1	1	1	n.d.	n.d.
Acetate (1)	(#)	n.d.	n.d.	n.d.	1	n.d.
Urea (1)	(#)	n.d.	1	1	0	n.d.
Transporters total (29)	**(#)**	**25**	**23**	**24**	**13**	**0**

activity of some genes involved in the reversed citric acid cycle in the Gammaproteobacterium suggested that this population may have mainly used the supplied organic molecules as a carbon source and not as an energy source; a partially reversed citric acid cycle is an indication for the inter-conversion of amino acids.

It may have obtained energy for growth by oxidizing sulfide to sulfate, formate to carbon dioxide, and hydrogen to water. The expression of the Calvin cycle enzymes was also significant, indicating that it may have grown partially autotrophically. Based on the genomic and proteomic evidence, sulfide, hydrogen and formate did not appear to be produced by any of the major populations presented in **Table 7.1**. Their consumption hinted at the presence of other, sulfate reducing and fermentative populations that remained below the detection limit of our metagenome.

7.5 Alternate coding by candidate division BD1-5/SN2

Candidate division BD1-5/SN-2 and the closely related candidate phylum SR1 are very common components of microbial communities in very different habitats such as the human gut, the oral cavity [189], subsurface aquifers [183] and marine sediments. At our sampling site almost 2 % of all 16S rRNA genes identified in 4 different metagenomes belonged to this division. Members of BD1-5/SN-2 have not been cultivated so far but a fermentative, possibly sulfur reducing lifestyle has been inferred from genomes reconstructed from metagenomic data and single cell genomes.

Both *in silico* analysis and proteomics showed very strong evidence for the use of an alternate genetic code by this bacterial division. As was previously described for the related division SR1, the opal stop codon, UGA, was found to be translated as glycine.[189] Alternate genetic codes are relatively rare among bacteria, but translation of UGA as tryptophan is known in multiple

organisms.[190, 191] The translation of UGA as glycine has now been reported in two Candidate Phyla, BD1 5/SN 2 and SR1.[189] The representation of non-standard coding may increase as more genomes are recovered via metagenomics and single cell methods. For in depth investigation of the possibility of alternate coding by STOP codon UGA, raw MS data was searched against a protein database made by *in-silico* translation of UGA in bin E proteins not only for glycine (stopG) and tryptophan (stopW), but to all twenty amino acids along with proteins from other bins. The search results showed that the UGA coding for Glycine was the most dominant alternate codon usage, while we also detected low levels of "mistranslation" for other amino acids. A single peptide was detected for tryptophan, alanine, aspartate and lysine using a $\pm$ 5 ppm strict cut-off on precursor ions. The alternate coding by BD1-5 is further illustrated in **Figure 7.3** which shows multiple sequence alignment of a representative protein

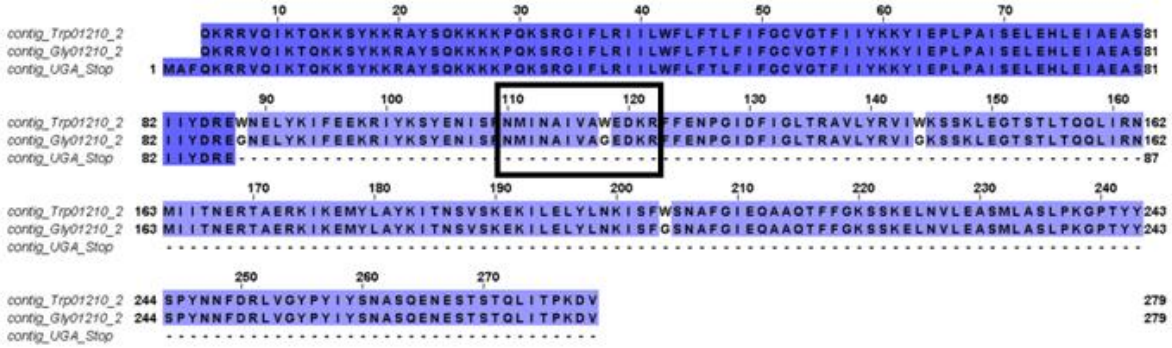


Figure 7.3 Multiple sequence alignment of a MS-identified representative protein in its three versions. *As evident from the figure any peptide that was identified before residue position 87 will not distinguish between the normal STOP codon, stopG and stopW database. But the highlighted peptide which was picked up by mass spectrometer was sufficient to inform whether UGA was terminating sequence, or coding for alternate amino-acids.*

from this clade. The sequence alignment shows three sequences of same bin E protein, one using UGA to code for glycine, second using UGA to code for tryptophan and third using UGA to terminate the protein sequence. The multiple sequence alignment shows that the majority of tryptic peptides are indistinguishable among the three versions of BD1-5/SR-1 phyla, however there are few regions in the protein where there are single amino-acid variations. Therefore, if mass spec can identify those peptides, we should be able to determine the codon usage in this protein. Our work identified 96 peptides that were uniquely matched to BD1-5 stopG proteins and 3 peptides mapped to BD1-5 stopW proteins within $\pm$10 ppm when using only three versions of BD1-5 or bin E proteome (normal STOP codon, stopG and stopW). Even with the larger database that had UGA in bin E proteins coding for all twenty amino acids, we identified 40 peptides from stopG database with precursor mass within $\pm$5 ppm and all of them had at least 2 spectral counts.The reduction in number of peptides with a larger database could be attributed to a more stringent filtering criteria applied by informatics pipeline to keep the FDR in check with the increase in the size of metagenome.

7.6 Conclusions

The deep proteome coverage of a bioconductor grown microbial community from Wadden Sea provided functional roles of individual microbial members in sustaining the community. While the proteomics measurements were in agreement with metagenomic abundance of microbial species, the high mass accuracy and high speed of MS platform provided the first experimental evidence of alternate coding by BD1-5/SR-1 phyla via proteomics. Our measurements confirm that the stop codon UGA in BD1-5 codes for a glycine residue and UGA also has a tendency to code for tryptophan. While the alternate genetic codes are very rare in prokaryotic species, the evidence of their occurrence is growing as more and more species are being sequenced. The low

level of mistranslations reported in this study may be an indicator that these are prerequisite for evolution of alternate genetic codes.

Chapter 8 - Conclusions and future outlook

8.1 Scientific impact of this dissertation work

The over-arching goal of this dissertation was to provide research community with better experimental and computational tool-sets to carry out discovery proteomics. The tremendous progress that mass spectrometry has made in last decade has only opened doors to newer and more complex challenges to be solved. The days of working on single protein through-out one's academic career are gone, and more and more people are embracing the holistic approach of systems biology.

Therefore, the aim of this dissertation work was to develop an integrated pipeline that encompasses both the experimental and computational components into one framework to address different questions in the field of mass spectrometry based proteomics. This work will help the proteomics community in choosing appropriate experimental methods for their needs based on key figure of merits that have been defined and explored in this research. Our work will provide impetus to those looking at broad-scale PTM investigation in biological systems and will caution them against over interpretation of mass spec data. This work will also present a case for using high mass accuracy and high resolution for data acquisition and its advantage in addressing specific biological questions.

In chapter 3 of this dissertation, we provided some key metrics that needs to be considered before starting on any proteomics campaign. Our work showed that starting biomass is an important consideration for selection of the most appropriate sample preparation method for

successful proteomics measurements. We further discussed pros and cons of different approaches for sample preparation involving SDS, and suggested guidelines of their usage under appropriate conditions. We also highlighted cases when the best described sample preparation method may not work and how to proceed forward.

In this chapter, we also evaluated the efficacy of GELFrEE technology in aiding proteomics. Our results showed that this solution based intact protein fractionation approach has definite merits, but it might cater more to the "top-down" and "targeted-proteomics" community rather than general global bottom-up approach. Although supplementing a whole cell lysate proteomics measurement with proteomics measurement of multiple fractions resulted in increased depth of proteome coverage, it came at a cost of instrument time. Therefore, the GELFrEE approach may not be promising for the labs that are limited by availability of instrument time.

In chapter 4 of this dissertation, we focused on environmental samples (mainly soils) and described challenges associated with their proteome extraction. This chapter highlighted some key steps that are required for efficient for protein extraction from soils, and stressed on the importance of accurate metagenome construction. In order to demonstrate it to the broader scientific community on how to gauge the success of their MS measurements, we evaluated and tested the usefulness of *de-novo* evaluation of spectral quality.

In chapter 5, we introduced ETD fragmentation in proteomics and the application of alternate proteases to help in identification of PTMs. We systematically evaluated different proteases starting with trypsin (which is the least expensive and most widely used) to more specific LysC and GluC digestion on two simple prokaryotic systems. Our results show definite merit in using alternate protease digestion along with regular database searching algorithms to find PTMs.

Applying dual fragmentation modes for the same systems helped in enhancing confident protein identifications and supplementing modified peptides. However, it appeared that using unenriched microbial systems with ETD was probably not an optimal choice, as we had limited success in identifying PTMs with ETD.

In chapter 6, we transition to a more complex eukaryotic plant system, *P. trichocarpa*, for PTM identification. In this chapter, we introduced HCD fragmentation and its utility in achieving complete sequence information, which is critical for PTM identification. Further, a distinct advantage in using a sequence tagging approach was examined and highlighted. Since current mass spec based PTM methods usually are only limited to handful of modifications, this approach can cater to a wide spectrum of researchers who are working on rare modifications.

In chapter 7, we applied our information from chapter 3 regarding experimental procedures for biomass limited samples for successful application in a natural microbial community. The advanced sample preparation workflow was further supplemented with high mass accuracy data acquisition, which helped in delineating alternate coding of STOP codons in this community. The conclusions from this chapter not only act as a proof-of-principle for the importance of knowing total biomass beforehand for proteomics sample preparation, but also provide researchers with the demonstrated value of high accuracy –high resolution mass spectrometry in the field.

The work described in this dissertation touches the complete breath of the salient aspects of proteomics workflows, and should be a valuable resource for practioners in this field, whether they have a stronger experimental background or computational bent.

8.2 Status of the field and remaining challenges

The field of biological mass spectrometry is at a critical juncture, wherein a plethora of mass spec instruments are available to cater to different research needs. The availability of high end, state of the art mass spectrometers in numerous labs and core facility around the country has made proteomics measurements highly accessible, though it is still not affordable to many.

Highly complicated samples are now being analyzed more commonly than what might have even been dreamed five years ago. System level biological mass spectrometry is now being applied not just in model systems for basic research, but has become quite common in drug discovery and development, biomarker discovery, crop improvement, and bioremediation. Some of the recent developments include, "*chemical proteomics*" which is the application of chemical probes to pick potential drug targets, "*redox proteomics"* which is the application of mass spectrometry to determine covalent modifications linked to redox metabolism, "Integrated personal omics profile or iPOP" which is the application of different omics approaches including proteomics for personalized medicine [117, 122, 147, 192].

While, all the novel applications where mass spectrometry finds itself tightly embedded in, is a good sign for the development of the field, but it also poses several challenges. Since the debacle of SELDI mass spectrometry in biomarker discovery for ovarian cancer, any new application of mass spectrometry in real world comes with increased scrutiny. This acts as an impetus for those like us who work at the fundamental level to provide as much information as possible with sufficient clarity, so that any new developments, whether on the experimental side or the computational side, can be replicated elsewhere.

In this regard, we recognize several challenges in the field of mass spectrometry, which should be addressed. While the proteomics of mammalian systems has become routine in many labs around the world, relatively slower progress has been made in the field of environmental proteomics. As mentioned already in this dissertation, there is no efficient way of accurately determining protein content in environmental samples and therefore, there is a great demand for better non-colorimetric methods for determination of *in-situ* protein concentration.

The online 2D-LC-MS/MS approach has made great progress since its introduction in 2001, but the turnaround time to complete a single measurement for complex samples is only down to 24 hours from 64 hours. Any reduction in the data acquisition time results in concomitant decrease in protein identifications. With the introduction of next gen mass spectrometers like Q-Exactive and Orbitrap Fusion, there is a hope that total time for data acquisition will be reduced considerably. This was bolstered by recent measurement of yeast proteome in 1 hour by the Coon lab, showing same protein coverage as one would get with a 24hr MudPIT [193]. But caveat of solely depending on new instruments is that each of these high end mass spec instruments are more expensive than their previous generation. Therefore, unless one is working in a lab where there is no scarcity of funds, it is very difficult to rapidly replace instruments. The requirement is therefore, to develop methods that can be incorporated in existing mass spectrometry platforms which can provide similar protein coverage at much shorter time scales. To achieve this goal, we will have to better evaluate chromatographic separations and help in designing intelligent data acquisition methods which go beyond the traditional data-dependent and data-independent modes. The development of inSeq algorithm and the application of high pH – low pH RP separations are positive step in this direction but they need to be evaluated on diverse sample types [194, 195].

A major thrust in recent years has been enhancement of quantitative proteomics via labeling methods, with a focus on multiplexing. While there are several labeling strategies in the field, none of them are universally applicable and, especially for environmental samples, are quite limited. Therefore, it will be highly valuable to develop labeling methods for environmental systems that will be able to determine relative protein expression levels, rather than just a digital "YES" or "NO" answer.

8.3 Future outlook

The outlook for mass spectrometry based proteomics is very promising. As mentioned before, the field is diving into new areas of research and is gradually being accepted by non-biologists. Since the beginning of the field of MS over a century, this field has measured ionic species across a wide gamete of analyte types. However, it is limited to either measuring a positively charged species or a negatively charged species, but never both in the same scan. Some earlier work by Dr. Scott McLukey in 2002 showed that it is possible to have ions of both polarities in the ion-trap in a single experiment, but they measured either positively charged ions or negatively charged ions at any given time [160]. Later in 2009 Chen *et al.* showed first dual polarity ESI-MS system which could be used for analysis of real world samples [161]. By making some fundamental changes in the development of next-gen mass spectrometers especially in the ion-detection system, it should be possible to determine ions of both polarities. Just as a concept, if an additional set of detectors are included and the geometry of ion-traps along with applied RF potential is manipulated in such a way that one set of detectors measure positively charged species and the second set of detectors measure negatively charged species, it should be possible to determine ratio between positively and negatively charged species. The advantage of such a scheme is that, while some biological molecules like to be protonated, there

are many other which prefer to be in basic conditions (like nucleic acids). Hence, if the MS instrument can measure both the polarities at the same time, we can answer some of the fundamental questions in protein-DNA and protein-drug interactions.

While novel mass spectrometry instruments will be a boon for the field, we also need better informatics pipelines that can assist in mining increasingly complex RAW MS datasets. A simpler way of putting this in perspective is that if a mass spectrometer can churn data that provides similar coverage by taking $1/10^{th}$ of the original time, the generated data will be that much more complicated than before and therefore, computational programs need to keep up with the developments taking place on the instrument sides. Hence, we need computational programs that can leverage high speed, high resolution and high mass accuracy of current mass spectrometers to provide greater proteome coverage. A significant development of informatics pipeline is needed in the field of quantitative proteomics.

A second area of active research will be to develop low-cost, benchtop mass spectrometers that can be put in every school and a hospital lab. The current mass spectrometry platforms are such that they need special care and are found in labs which can provide solid infrastructure and support. With mass spectrometry making tremendous progress in the field of clinical proteomics, we need better educational and training programs, so that instead of clinics sending their samples to MS core labs they are able to process basic mass spectrometry measurements in-house, thereby speeding patient care.

High schools are the basal level where we can modulate young minds to take up science and develop them into next generation of researchers. However, the expense and infrastructure associated with mass spectrometry lab has so far kept this analytical approach inaccessible to

large percentage of science students. Since the basic aspect of mass spectrometry is measurement of ions, which is a concept taught to students in schools, mass spectrometry is very much an alien subject to them. Therefore, just like microscopy that has helped our young generation to appreciate biology, mass spectrometry needs an outreach program. In this regard, we need to develop workshops that provide a platform to high school students to have first-hand experience with MS instrumentation and data analysis. By reaching out to them, we can mold more students to take up analytical studies in their graduate school.

References

1. Aston, F.W., *The Story of Isotopes.* Science, 1933. **78**(2010): p. 5-6.
2. Dempster, A.J., *A new Method of Positive Ray Analysis.* Physical Review, 1918. **11**(4): p. 316-325.
3. Dempster, A.J., *Thirty years of mass spectroscopy.* Sci Mon, 1948. **67**(3): p. 145-53.
4. Di Domenico, F., et al., *Redox proteomics analysis of HNE-modified proteins in DS brain: clues for understanding development of Alzheimer disease.* Free Radic Biol Med, 2014.
5. Biemann, K., G. Gapp, and J. Seibl, *APPLICATION OF MASS SPECTROMETRY TO STRUCTURE PROBLEMS. I. AMINO ACID SEQUENCE IN PEPTIDES.* Journal of the American Chemical Society, 1959. **81**(9): p. 2274-2275.
6. Biemann, K. and J.A. McCloskey, *Application of Mass Spectrometry to Structure Problems.1 VI. Nucleosides2.* Journal of the American Chemical Society, 1962. **84**(10): p. 2005-2007.
7. Morris, H.R., et al., *Fast atom bombardment: a new mass spectrometric method for peptide sequence analysis.* Biochem Biophys Res Commun, 1981. **101**(2): p. 623-31.
8. Fenn, J.B., *Electrospray ionization mass spectrometry: How it all began.* J Biomol Tech, 2002. **13**(3): p. 101-18.
9. Fenn, J.B., et al., *Electrospray ionization for mass spectrometry of large biomolecules.* Science, 1989. **246**(4926): p. 64-71.
10. Yamashita, M. and J.B. Fenn, *Negative ion production with the electrospray ion source.* The Journal of Physical Chemistry, 1984. **88**(20): p. 4671-4675.
11. Yamashita, M. and J.B. Fenn, *Electrospray ion source. Another variation on the free-jet theme.* The Journal of Physical Chemistry, 1984. **88**(20): p. 4451-4459.

12. 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.

13. Karas, M., et al., *Matrix-assisted ultraviolet laser desorption of non-volatile compounds.* International Journal of Mass Spectrometry and Ion Processes, 1987. **78**(0): p. 53-68.

14. Karas, M., D. Bachmann, and F. Hillenkamp, *Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules.* Analytical Chemistry, 1985. **57**(14): p. 2935-2939.

15. Karas, M. and F. Hillenkamp, *Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons.* Analytical Chemistry, 1988. **60**(20): p. 2299-2301.

16. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.

17. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.

18. Kitano, H., *Systems biology: a brief overview.* Science, 2002. **295**(5560): p. 1662-4.

19. Riesenfeld, C.S., P.D. Schloss, and J. Handelsman, *Metagenomics: genomic analysis of microbial communities.* Annu Rev Genet, 2004. **38**: p. 525-52.

20. Schaechter, M. and J.L. Ingraham, *What limits genomics, proteomics, transcriptomics?* Int Microbiol, 2002. **5**(2): p. 51-2.

21. Raamsdonk, L.M., et al., *A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations.* Nat Biotechnol, 2001. **19**(1): p. 45-50.

22. Unlu, M., *Difference gel electrophoresis.* Biochem Soc Trans, 1999. **27**(4): p. 547-9.

23. Van den Bergh, G. and L. Arckens, *Fluorescent two-dimensional difference gel electrophoresis unveils the potential of gel-based proteomics.* Curr Opin Biotechnol, 2004. **15**(1): p. 38-43.

24. Wilkins, M.R., et al., *From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis.* Biotechnology (N Y), 1996. **14**(1): p. 61-5.

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

26. Loo, J.A., C.G. Edmonds, and R.D. Smith, *Primary sequence information from intact proteins by electrospray ionization tandem mass spectrometry.* Science, 1990. **248**(4952): p. 201-4.

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

28. 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.

29. Henzel, W.J., et al., *Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases.* Proc Natl Acad Sci U S A, 1993. **90**(11): p. 5011-5.

30. 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.

31. Wilmes, P. and P.L. Bond, *The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms.* Environ Microbiol, 2004. **6**(9): p. 911-20.

32. 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.

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

34. Belnap, C.P., et al., *Quantitative proteomic analyses of the response of acidophilic microbial communities to different pH conditions.* ISME J, 2011. **5**(7): p. 1152-61.

35. Kleiner, M., et al., *Metaproteomics of a gutless marine worm and its symbiotic microbial community reveal unusual pathways for carbon and energy use.* Proc Natl Acad Sci U S A, 2012. **109**(19): p. E1173-82.

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

37. Pan, C., et al., *Quantitative tracking of isotope flows in proteomes of microbial communities.* Mol Cell Proteomics, 2011. **10**(4): p. M110 006049.

38. Sowell, S.M., et al., *Transport functions dominate the SAR11 metaproteome at low-nutrient extremes in the Sargasso Sea.* ISME J, 2009. **3**(1): p. 93-105.

39. Morris, R.M., et al., *Comparative metaproteomics reveals ocean-scale shifts in microbial nutrient utilization and energy transduction.* ISME J, 2010. **4**(5): p. 673-85.

40. Wilkins, M.J., et al., *Proteogenomic monitoring of Geobacter physiology during stimulated uranium bioremediation.* Appl Environ Microbiol, 2009. **75**(20): p. 6591-9.

41. Hongoh, Y., *Toward the functional analysis of uncultivable, symbiotic microorganisms in the termite gut.* Cell Mol Life Sci, 2011. **68**(8): p. 1311-25.

42. Wang, H.B., et al., *Characterization of metaproteomics in crop rhizospheric soil.* J Proteome Res, 2011. **10**(3): p. 932-40.

43. Brown, J., et al., *Translating the human microbiome.* Nat Biotechnol, 2013. **31**(4): p. 304-8.

44. Cox, M.J., W.O. Cookson, and M.F. Moffatt, *Sequencing the human microbiome in health and disease.* Hum Mol Genet, 2013. **22**(R1): p. R88-94.

45. Eloe-Fadrosh, E.A. and D.A. Rasko, *The human microbiome: from symbiosis to pathogenesis.* Annu Rev Med, 2013. **64**: p. 145-63.

46. Levy, R. and E. Borenstein, *Metagenomic systems biology and metabolic modeling of the human microbiome: From species composition to community assembly rules.* Gut Microbes, 2014. **5**(2).

47. Klaassens, E.S., W.M. de Vos, and E.E. Vaughan, *Metaproteomics approach to study the functionality of the microbiota in the human infant gastrointestinal tract.* Appl Environ Microbiol, 2007. **73**(4): p. 1388-92.

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

49. Baumann, M. and S. Meri, *Techniques for studying protein heterogeneity and post-translational modifications.* Expert Rev Proteomics, 2004. **1**(2): p. 207-17.

50. Walsh, C.T., S. Garneau-Tsodikova, and G.J. Gatto, Jr., *Protein posttranslational modifications: the chemistry of proteome diversifications.* Angew Chem Int Ed Engl, 2005. **44**(45): p. 7342-72.

51. Pinkse, M.W., et al., *Highly robust, automated, and sensitive online TiO2-based phosphoproteomics applied to study endogenous phosphorylation in Drosophila melanogaster.* J Proteome Res, 2008. **7**(2): p. 687-97.

52. Villen, J., et al., *Large-scale phosphorylation analysis of mouse liver.* Proc Natl Acad Sci U S A, 2007. **104**(5): p. 1488-93.

53. Yang, F., et al., *Phosphoproteomics profiling of human skin fibroblast cells reveals pathways and proteins affected by low doses of ionizing radiation.* PLoS One, 2010. **5**(11): p. e14152.

54. Li, X., et al., *Large-scale phosphorylation analysis of alpha-factor-arrested Saccharomyces cerevisiae.* J Proteome Res, 2007. **6**(3): p. 1190-7.

55. Choudhary, C., et al., *Lysine acetylation targets protein complexes and co-regulates major cellular functions.* Science, 2009. **325**(5942): p. 834-40.

56. Weinert, B.T., et al., *Proteome-wide mapping of the Drosophila acetylome demonstrates a high degree of conservation of lysine acetylation.* Sci Signal, 2011. **4**(183): p. ra48.

57. Beltran, L. and P.R. Cutillas, *Advances in phosphopeptide enrichment techniques for phosphoproteomics.* Amino Acids, 2012. **43**(3): p. 1009-24.

58. Zhang, K., S. Tian, and E. Fan, *Protein lysine acetylation analysis: current MS-based proteomic technologies.* Analyst, 2013. **138**(6): p. 1628-36.

59. Bodzon-Kulakowska, A., et al., *Methods for samples preparation in proteomic research.* J Chromatogr B Analyt Technol Biomed Life Sci, 2007. **849**(1-2): p. 1-31.

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

61. 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.

62. 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.

63. Vandermarliere, E., M. Mueller, and L. Martens, *Getting intimate with trypsin, the leading protease in proteomics.* Mass Spectrom Rev, 2013. **32**(6): p. 453-65.

64. 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.

65. Rayleigh, L., *XX. On the equilibrium of liquid conducting masses charged with electricity.* Philosophical Magazine Series 5, 1882. **14**(87): p. 184-186.

66. Wilm, M., *Principles of Electrospray Ionization.* Molecular & Cellular Proteomics, 2011. **10**(7).

67. Dole, M., et al., *Molecular Beams of Macroions.* The Journal of Chemical Physics, 1968. **49**(5): p. 2240-2249.

68. Iribarne, J.V. and B.A. Thomson, *On the evaporation of small ions from charged droplets.* The Journal of Chemical Physics, 1976. **64**(6): p. 2287-2294.

69. March, R.E., *An Introduction to Quadrupole Ion Trap Mass Spectrometry.* Journal of Mass Spectrometry, 1997. **32**(4): p. 351-369.

70. 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.

71. Hu, Q., et al., *The Orbitrap: a new mass spectrometer.* J Mass Spectrom, 2005. **40**(4): p. 430-43.

72. Wells, J.M. and S.A. McLuckey, *Collision-induced dissociation (CID) of peptides and proteins.* Methods Enzymol, 2005. **402**: p. 148-85.

73. Paizs, B. and S. Suhai, *Fragmentation pathways of protonated peptides.* Mass Spectrom Rev, 2005. **24**(4): p. 508-48.

74. Olsen, J.V., et al., *Higher-energy C-trap dissociation for peptide modification analysis.* Nat Methods, 2007. **4**(9): p. 709-12.

75. 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.

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

77. 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.

78. Nesvizhskii, A.I., *A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics.* J Proteomics, 2010. **73**(11): p. 2092-123.

79. 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.

80. Gu, C., et al., *Chemical proteomics with sulfonyl fluoride probes reveals selective labeling of functional tyrosines in glutathione transferases.* Chem Biol, 2013. **20**(4): p. 541-8.

81. Tyson, G.W. and J.F. Banfield, *Cultivating the uncultivated: a community genomics perspective.* Trends Microbiol, 2005. **13**(9): p. 411-5.

82. Brunner, E., et al., *A high-quality catalog of the Drosophila melanogaster proteome.* Nat Biotechnol, 2007. **25**(5): p. 576-83.

83. Corbin, R.W., et al., *Toward a protein profile of Escherichia coli: comparison to its transcription profile.* Proc Natl Acad Sci U S A, 2003. **100**(16): p. 9232-7.

84. Sowell, S.M., et al., *Environmental proteomics of microbial plankton in a highly productive coastal upwelling system.* ISME J, 2011. **5**(5): p. 856-65.

85. Kolmeder, C.A., et al., *Comparative metaproteomics and diversity analysis of human intestinal microbiota testifies for its temporal stability and expression of core functions.* PLoS One, 2012. **7**(1): p. e29913.

86. Helenius, A. and K. Simons, *Solubilization of membranes by detergents.* Biochim Biophys Acta, 1975. **415**(1): p. 29-79.

87. Shevchenko, A., et al., *In-gel digestion for mass spectrometric characterization of proteins and proteomes.* Nat Protoc, 2006. **1**(6): p. 2856-60.

88. Wu, F., et al., *Comparison of surfactant-assisted shotgun methods using acid-labile surfactants and sodium dodecyl sulfate for membrane proteome analysis.* Anal Chim Acta, 2011. **698**(1-2): p. 36-43.

89. Yu, Y.Q., et al., *Enzyme-friendly, mass spectrometry-compatible surfactant for in-solution enzymatic digestion of proteins.* Anal Chem, 2003. **75**(21): p. 6023-8.

90. Arakawa, T., et al., *Induced resistance of trypsin to sodium dodecylsulfate upon complex formation with trypsin inhibitor.* Journal of Protein Chemistry, 1992. **11**(2): p. 171-176.

91. Botelho, D., et al., *Top-down and bottom-up proteomics of SDS-containing solutions following mass-based separation.* J Proteome Res, 2010. **9**(6): p. 2863-70.

92. Wisniewski, J.R., et al., *Universal sample preparation method for proteome analysis.* Nat Methods, 2009. **6**(5): p. 359-62.

93. Antharavally, B.S., et al., *Efficient removal of detergents from proteins and peptides in a spin column format.* Anal Biochem, 2011. **416**(1): p. 39-44.

94. Sun, D., N. Wang, and L. Li, *Integrated SDS removal and peptide separation by strong-cation exchange liquid chromatography for SDS-assisted shotgun proteome analysis.* J Proteome Res, 2012. **11**(2): p. 818-28.

95. Liebler, D.C. and A.J. Ham, *Spin filter-based sample preparation for shotgun proteomics.* Nat Methods, 2009. **6**(11): p. 785; author reply 785-6.

96. Fic, E., et al., *Comparison of protein precipitation methods for various rat brain structures prior to proteomic analysis.* Electrophoresis, 2010. **31**(21): p. 3573-9.

97. Jiang, L., L. He, and M. Fountoulakis, *Comparison of protein precipitation methods for sample preparation prior to proteomic analysis.* J Chromatogr A, 2004. **1023**(2): p. 317-20.

98. Chourey, K., et al., *Direct cellular lysis/protein extraction protocol for soil metaproteomics.* J Proteome Res, 2010. **9**(12): p. 6615-22.

99. Folch, J., et al., *Preparation of lipide extracts from brain tissue.* J Biol Chem, 1951. **191**(2): p. 833-41.

100. Wessel, D. and U.I. Flugge, *A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids.* Anal Biochem, 1984. **138**(1): p. 141-3.

101. Ferro, M., et al., *Organic solvent extraction as a versatile procedure to identify hydrophobic chloroplast membrane proteins.* Electrophoresis, 2000. **21**(16): p. 3517-26.

102. Bereman, M.S., J.D. Egertson, and M.J. MacCoss, *Comparison between procedures using SDS for shotgun proteomic analyses of complex samples.* Proteomics, 2011. **11**(14): p. 2931-5.

103. Nagaraj, N., et al., *System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap.* Mol Cell Proteomics, 2012. **11**(3): p. M111 013722.

104. Wisniewski, J.R., P. Ostasiewicz, and M. Mann, *High recovery FASP applied to the proteomic analysis of microdissected formalin fixed paraffin embedded cancer tissues retrieves known colon cancer markers.* J Proteome Res, 2011. **10**(7): p. 3040-9.

105. Pflieger, D., et al., *Systematic identification of mitochondrial proteins by LC-MS/MS.* Anal Chem, 2002. **74**(10): p. 2400-6.

106. Tran, J.C. and A.A. Doucette, *Gel-eluted liquid fraction entrapment electrophoresis: an electrophoretic method for broad molecular weight range proteome separation.* Anal Chem, 2008. **80**(5): p. 1568-73.

107. Tran, J.C., et al., *Mapping intact protein isoforms in discovery mode using top-down proteomics.* Nature, 2011. **480**(7376): p. 254-8.

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

109. Yu, N.Y., et al., *PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes.* Bioinformatics, 2010. **26**(13): p. 1608-15.

110. Karp, P.D., et al., *Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology.* Brief Bioinform, 2010. **11**(1): p. 40-79.

111. Keseler, I.M., et al., *EcoCyc: a comprehensive database of Escherichia coli biology.* Nucleic Acids Res, 2011. **39**(Database issue): p. D583-90.

112. Schmidt, T. and D. Frishman, *PROMPT: a protein mapping and comparison tool.* BMC Bioinformatics, 2006. **7**: p. 331.

113. Tran, J.C. and A.A. Doucette, *Multiplexed size separation of intact proteins in solution phase for mass spectrometry.* Anal Chem, 2009. **81**(15): p. 6201-9.

114. Fathi, A., et al., *Quantitative proteomics analysis highlights the role of redox hemostasis and energy metabolism in human embryonic stem cell differentiation to neural cells.* J Proteomics, 2014. **101**: p. 1-16.

115. Liu, P., et al., *Identification of redox-sensitive cysteines in the Arabidopsis proteome using OxiTRAQ, a quantitative redox proteomics method.* Proteomics, 2014. **14**(6): p. 750-62.

116. Perluigi, M., A.M. Swomley, and D.A. Butterfield, *Redox proteomics and the dynamic molecular landscape of the aging brain.* Ageing Res Rev, 2014. **13**: p. 75-89.

117. Butterfield, D.A. and I. Dalle-Donne, *Redox proteomics: from protein modifications to cellular dysfunction and disease.* Mass Spectrom Rev, 2014. **33**(1): p. 1-6.

118. Colombo, G., et al., *Pathophysiology of tobacco smoke exposure: Recent insights from comparative and redox proteomics.* Mass Spectrom Rev, 2014. **33**(3): p. 183-218.

119. Charles, R., T. Jayawardhana, and P. Eaton, *Gel-based methods in redox proteomics.* Biochim Biophys Acta, 2014. **1840**(2): p. 830-7.

120. Pan, K.T., et al., *Mass spectrometry-based quantitative proteomics for dissecting multiplexed redox cysteine modifications in nitric oxide-protected cardiomyocyte under hypoxia.* Antioxid Redox Signal, 2014. **20**(9): p. 1365-81.

121. Rahaman, M.M., et al., *S-guanylation proteomics for redox-based mitochondrial signaling.* Antioxid Redox Signal, 2014. **20**(2): p. 295-307.

122. Li-Pook-Than, J. and M. Snyder, *iPOP goes the world: integrated personalized Omics profiling and the road toward improved health care.* Chem Biol, 2013. **20**(5): p. 660-6.

123. Roukos, D.H., *Dynamics of genome 'iPOP': predicting disease or 'narciss-ome'?* Expert Rev Mol Diagn, 2012. **12**(6): p. 545-8.

124. Briggs, J.M. and A.K. Knapp, *Interannual Variability in Primary Production in Tallgrass Prairie: Climate, Soil Moisture, Topographic Position, and Fire as Determinants of Aboveground Biomass.* American Journal of Botany, 1995. **82**(8): p. 1024-1030.

125. Knapp, A.K. and T.R. Seastedt, *Introduction: Grasslands, Konza Prairie, and long-term ecological research.* Long-Term Ecological Research Network Series; Grassland dynamics: Long-term ecological research in tallgrass prairie, 1998: p. 3-15.

126. Werner, T.P., N. Amrhein, and F.M. Freimoser, *Novel method for the quantification of inorganic polyphosphate (iPoP) in Saccharomyces cerevisiae shows dependence of iPoP content on the growth phase.* Arch Microbiol, 2005. **184**(2): p. 129-36.

127. Colzani, M., et al., *Quantitative chemical proteomics identifies novel targets of the anti-cancer multi-kinase inhibitor E-3810.* Mol Cell Proteomics, 2014.

128. Ku, X., et al., *A new affinity probe targeting VEGF receptors for kinase inhibitor selectivity profiling by chemical proteomics.* J Proteome Res, 2014.

129. Trochine, A., et al., *Trypanosoma cruzi chemical proteomics using immobilized benznidazole.* Exp Parasitol, 2014. **140C**: p. 33-38.

130. Wang, J., et al., *A quantitative chemical proteomics approach to profile the specific cellular targets of andrographolide, a promising anticancer agent that suppresses tumor metastasis.* Mol Cell Proteomics, 2014. **13**(3): p. 876-86.

131. Ambrogelly, A., S. Palioura, and D. Soll, *Natural expansion of the genetic code.* Nat Chem Biol, 2007. **3**(1): p. 29-35.

132. Khoury, G.A., R.C. Baliban, and C.A. Floudas, *Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database.* Sci Rep, 2011. **1**.

133. Strahl, B.D. and C.D. Allis, *The language of covalent histone modifications.* Nature, 2000. **403**(6765): p. 41-5.

134. Prabakaran, S., et al., *Post-translational modification: nature's escape from genetic imprisonment and the basis for dynamic information encoding.* Wiley Interdiscip Rev Syst Biol Med, 2012. **4**(6): p. 565-83.

135. Verma, R., et al., *Phosphorylation of Sic1p by G1 Cdk required for its degradation and entry into S phase.* Science, 1997. **278**(5337): p. 455-60.

136. Jedrzejewski, P.T. and W.D. Lehmann, *Detection of modified peptides in enzymatic digests by capillary liquid chromatography/electrospray mass spectrometry and a programmable skimmer CID acquisition routine.* Anal Chem, 1997. **69**(3): p. 294-301.

137. Ficarro, S.B., et al., *Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae.* Nat Biotechnol, 2002. **20**(3): p. 301-5.

138. Nuhse, T.S., et al., *Large-scale analysis of in vivo phosphorylated membrane proteins by immobilized metal ion affinity chromatography and mass spectrometry.* Mol Cell Proteomics, 2003. **2**(11): p. 1234-43.

139. Zhao, Y. and O.N. Jensen, *Modification-specific proteomics: strategies for characterization of post-translational modifications using enrichment techniques.* Proteomics, 2009. **9**(20): p. 4632-41.

140. Aivaliotis, M., et al., *Ser/Thr/Tyr protein phosphorylation in the archaeon Halobacterium salinarum--a representative of the third domain of life.* PLoS One, 2009. **4**(3): p. e4777.

141. Macek, B., et al., *Phosphoproteome analysis of E. coli reveals evolutionary conservation of bacterial Ser/Thr/Tyr phosphorylation.* Mol Cell Proteomics, 2008. **7**(2): p. 299-307.

142. Macek, B., et al., *The serine/threonine/tyrosine phosphoproteome of the model bacterium Bacillus subtilis.* Mol Cell Proteomics, 2007. **6**(4): p. 697-707.

143. Jekel, P.A., W.J. Weijer, and J.J. Beintema, *Use of endoproteinase Lys-C from Lysobacter enzymogenes in protein sequence analysis.* Anal Biochem, 1983. **134**(2): p. 347-54.

144. Drapeau, G.R., Y. Boily, and J. Houmard, *Purification and properties of an extracellular protease of Staphylococcus aureus.* J Biol Chem, 1972. **247**(20): p. 6720-6.

145. Speers, A.E. and B.F. Cravatt, *A tandem orthogonal proteolysis strategy for high-content chemical proteomics.* J Am Chem Soc, 2005. **127**(28): p. 10018-9.

146. Speers, A.E. and B.F. Cravatt, *Chemical strategies for activity-based proteomics.* Chembiochem, 2004. **5**(1): p. 41-7.

147. Sun, B. and Q.Y. He, *Chemical proteomics to identify molecular targets of small compounds.* Curr Mol Med, 2013. **13**(7): p. 1175-91.

148. Margarucci, L., et al., *Chemical proteomics-driven discovery of oleocanthal as an Hsp90 inhibitor.* Chem Commun (Camb), 2013. **49**(52): p. 5844-6.

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

150. Macek, B. and I. Mijakovic, *Site-specific analysis of bacterial phosphoproteomes.* Proteomics, 2011. **11**(15): p. 3002-11.

151. Tuskan, G.A., et al., *The genome of black cottonwood, Populus trichocarpa (Torr. & Gray).* Science, 2006. **313**(5793): p. 1596-604.

152. Wullschleger, S.D., et al., *Revisiting the sequencing of the first tree genome: Populus trichocarpa.* Tree Physiol, 2013. **33**(4): p. 357-64.

153. Jansson, S. and C.J. Douglas, *Populus: a model system for plant biology.* Annu Rev Plant Biol, 2007. **58**: p. 435-58.

154. Battey, N.H., H.G. Dickinson, and A. Hetherington, *Post-translational modifications in plants*. Vol. 53. 1993: Cambridge University Press.

155. Rodgers-Melnick, E., et al., *Contrasting patterns of evolution following whole genome versus tandem duplication events in Populus.* Genome Res, 2012. **22**(1): p. 95-105.

156. Wolfe, L.M., et al., *A chemical proteomics approach to profiling the ATP-binding proteome of Mycobacterium tuberculosis.* Mol Cell Proteomics, 2013. **12**(6): p. 1644-60.

157. Dal Piaz, F., et al., *Chemical proteomics reveals HSP70 1A as a target for the anticancer diterpene oridonin in Jurkat cells.* J Proteomics, 2013. **82**: p. 14-26.

158. Gyenis, L., et al., *Chemical proteomics and functional proteomics strategies for protein kinase inhibitor validation and protein kinase substrate identification: applications to protein kinase CK2.* Biochim Biophys Acta, 2013. **1834**(7): p. 1352-8.

159. Yount, J.S., M.M. Zhang, and H.C. Hang, *Emerging roles for protein S-palmitoylation in immunity from chemical proteomics.* Curr Opin Chem Biol, 2013. **17**(1): p. 27-33.

160. McLuckey, S.A., et al., *Oligonucleotide mixture analysis via electrospray and ion/ion reactions in a quadrupole ion trap.* Anal Chem, 2002. **74**(5): p. 976-84.

161. Chen, H.K., et al., *Synchronized dual-polarity electrospray ionization mass spectrometry.* J Am Soc Mass Spectrom, 2009. **20**(12): p. 2254-7.

162. Abraham, P., et al., *Defining the boundaries and characterizing the landscape of functional genome expression in vascular tissues of Populus using shotgun proteomics.* J Proteome Res, 2012. **11**(1): p. 449-60.

163. Abraham, P., et al., *Putting the pieces together: high-performance LC-MS/MS provides network-, pathway-, and protein-level perspectives in Populus.* Mol Cell Proteomics, 2013. **12**(1): p. 106-19.

164. Michalski, A., et al., *Ultra high resolution linear ion trap Orbitrap mass spectrometer (Orbitrap Elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes.* Mol Cell Proteomics, 2012. **11**(3): p. O111 013698.

165. Edgar, R.C., *Search and clustering orders of magnitude faster than BLAST.* Bioinformatics, 2010. **26**(19): p. 2460-1.

166. Florens, L., et al., *Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors.* Methods, 2006. **40**(4): p. 303-11.

167. Nag, A., et al., *Enhancing a Pathway-Genome Database (PGDB) to capture subcellular localization of metabolites and enzymes: the nucleotide-sugar biosynthetic pathways of Populus trichocarpa.* Database (Oxford), 2012. **2012**: p. bas013.

168. Yao, Q., et al., *Predicting and analyzing protein phosphorylation sites in plants using musite.* Front Plant Sci, 2012. **3**: p. 186.

169. Graff, J. and L.H. Tsai, *Histone acetylation: molecular mnemonics on the chromatin.* Nat Rev Neurosci, 2013. **14**(2): p. 97-111.

170. Zee, B.M. and B.A. Garcia, *Validation of protein acetylation by mass spectrometry.* Methods Mol Biol, 2013. **981**: p. 1-11.

171. Kane, L.A. and J.E. Van Eyk, *Post-translational modifications of ATP synthase in the heart: biology and function.* J Bioenerg Biomembr, 2009. **41**(2): p. 145-50.

172. Walker, J.E., *The ATP synthase: the understood, the uncertain and the unknown.* Biochem Soc Trans, 2013. **41**(1): p. 1-16.

173. Chen, J.Y. and X.F. Dai, *Cloning and characterization of the Gossypium hirsutum major latex protein gene and functional analysis in Arabidopsis thaliana.* Planta, 2010. **231**(4): p. 861-73.

174. Simpson, D.M. and R.J. Beynon, *Acetone precipitation of proteins and the modification of peptides.* J Proteome Res, 2010. **9**(1): p. 444-50.

175. Willem Michel Kieskamp, L.L., Eric Epping, Willem Helder, *Seasonal variation in denitrification rates and nitrous oxide fluxes in intertidal sediments of the western Wadden Sea.* Mar. Ecol. Prog. Ser., 1991. **72**: p. 145-151.

176. Lenk, S., et al., *Novel groups of Gammaproteobacteria catalyse sulfur oxidation and carbon fixation in a coastal, intertidal sediment.* Environ Microbiol, 2011. **13**(3): p. 758-74.

177. Llobet-Brossa, E., R. Rossello-Mora, and R. Amann, *Microbial Community Composition of Wadden Sea Sediments as Revealed by Fluorescence In Situ Hybridization.* Appl Environ Microbiol, 1998. **64**(7): p. 2691-6.

178. Eilers, H., et al., *Isolation of novel pelagic bacteria from the German bight and their seasonal contributions to surface picoplankton.* Appl Environ Microbiol, 2001. **67**(11): p. 5134-42.

179. Walsh, D.A., et al., *Metagenome of a versatile chemolithoautotroph from expanding oceanic dead zones.* Science, 2009. **326**(5952): p. 578-82.

180. Stewart, F.J., *Dissimilatory sulfur cycling in oxygen minimum zones: an emerging metagenomics perspective.* Biochem Soc Trans, 2011. **39**(6): p. 1859-63.

181. Kirchman, D.L., *The ecology of Cytophaga-Flavobacteria in aquatic environments.* FEMS Microbiol Ecol, 2002. **39**(2): p. 91-100.

182. Tyson, G.W., et al., *Community structure and metabolism through reconstruction of microbial genomes from the environment.* Nature, 2004. **428**(6978): p. 37-43.

183. Wrighton, K.C., et al., *Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla.* Science, 2012. **337**(6102): p. 1661-5.

184. Blainey, P.C., *The future is now: single-cell genomics of bacteria and archaea.* FEMS Microbiol Rev, 2013. **37**(3): p. 407-27.

185. Kalisky, T. and S.R. Quake, *Single-cell genomics.* Nat Methods, 2011. **8**(4): p. 311-4.

186. Albertsen, M., et al., *Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes.* Nat Biotechnol, 2013. **31**(6): p. 533-8.

187. Castelle, C.J., et al., *Extraordinary phylogenetic diversity and metabolic versatility in aquifer sediment.* Nat Commun, 2013. **4**: p. 2120.

188. Belnap, C.P., et al., *Cultivation and quantitative proteomic analyses of acidophilic microbial communities.* ISME J, 2010. **4**(4): p. 520-30.

189. Campbell, J.H., et al., *UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota.* Proc Natl Acad Sci U S A, 2013. **110**(14): p. 5540-5.

190. Yamao, F., et al., *UGA is read as tryptophan in Mycoplasma capricolum.* Proc Natl Acad Sci U S A, 1985. **82**(8): p. 2306-9.

191. McCutcheon, J.P., B.R. McDonald, and N.A. Moran, *Origin of an alternative genetic code in the extremely small and GC-rich genome of a bacterial symbiont.* PLoS Genet, 2009. **5**(7): p. e1000565.

192. Yuet, K.P. and D.A. Tirrell, *Chemical tools for temporally and spatially resolved mass spectrometry-based proteomics.* Ann Biomed Eng, 2014. **42**(2): p. 299-311.

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

194. Bailey, D.J., et al., *Instant spectral assignment for advanced decision tree-driven mass spectrometry.* Proc Natl Acad Sci U S A, 2012. **109**(22): p. 8411-6.

195. Kong, R.P., et al., *Development of online high-/low-pH reversed-phase-reversed-phase two-dimensional liquid chromatography for shotgun proteomics: a reversed-phase-*

strong cation exchange-reversed-phase approach. J Chromatogr A, 2011. **1218**(23): p. 3681-8.

VITA

Ritin Sharma was born in Udaipur (Rajasthan) in India. He graduated from The Study School, Udaipur with Senior Secondary School Certificate from Central Board of Secondary Education in 1999. He then completed his B.Sc. in Environmental Sciences, Chemistry and Botany from Mohan Lal Sukhadia University, Udaipur in 2003. After completing his Bachelor degree, he joined the Institute of Bioinformatics and Applied Biotechnology, Bangalore and earned his Post Graduate Diploma in Bioinformatics in 2004. Following the completion of his PG Diploma, he worked as a Project Trainee (2005-2006) in the laboratory of Dr. N. Srinivasan at Molecular Biophysics Unit, Indian Institute of Science – Bangalore, India. From 2006-2008, he worked in the Chemistry division of AstraZeneca India Private Limited as a Research Associate. After working in industry for a couple of years, he enrolled in the UTK-ORNL Graduate School of Genome Science and Technology in Spring 2009 and later joined the lab of Dr. Robert L. Hettich to pursue doctoral studies in the field of biological mass spectrometry. He expects to receive his Ph.D. in August of 2014.