5-2010

Use of Proteomics Tools to Investigate Protein Expression in Azospirillum brasilense

Gurusahai K. Khalsa-Moyers
University of Tennessee - Knoxville, gsahaik@comcast.net

Recommended Citation

Khalsa-Moyers, Gurusahai K., "Use of Proteomics Tools to Investigate Protein Expression in Azospirillum brasilense. " PhD diss., University of Tennessee, 2010.
https://trace.tennessee.edu/utk_graddiss/712

To the Graduate Council:

I am submitting herewith a dissertation written by Gurusahai K. Khalsa-Moyers entitled "Use of Proteomics Tools to Investigate Protein Expression in Azospirillum brasilense." 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.

Gladys Alexandre, Major Professor

We have read this dissertation and recommend its acceptance:

Cynthia B. Peterson, Dale A. Pelletier, Jennifer L. Morell-Falvey

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

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

To the Graduate Council:

I am submitting herewith a dissertation written by Gurusahai Kaur Khalsa-Moyers entitled "Use of Proteomics Tools to Investigate Protein Expression in *Azospirillum brasilense*". I have examined the final electronic copy of this dissertation for form and content and recommended that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Life Sciences.

Gladys B. Alexandre___
Major Professor

We have read this dissertation
and recommended its acceptance:

Dale A. Pelletier___________

Cynthia B. Peterson_______

Jennifer L. Morell-Falvey_______

Accepted for the Council:

____Carolyn Hodges___
Vice Provost and
Dean of Graduate School

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

Use of Proteomics Tools to Investigate Protein Expression in *Azospirillum brasilense*

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

Gurusahai Kaur Khalsa-Moyers

May, 2010

Dedication

To my daughter, Kelsey

For her never-ending encouragement and her happy smile.

Acknowledgements

I would like to first acknowledge my advisor, Gladys Alexandre, for her time and energy and for commenting on what seemed like endless versions of this work. I would also like to acknowledge the members of my committee for their service: Dr. Dale A. Pelletier, Dr. Cynthia Peterson, and Dr. Jennifer Morrell-Falvey.

I am grateful to a number of people for instruction and help with implementing all techniques used in this work. Dr Dale Pelletier and members of his lab provided laboratory facilities and instruction. Microbial growth assistance was provided by Tse-Yuan Lu for *R. palustris* cultures, and DNA manipulation assistance was provided by Linda Foote. Patricia Lankford helped tremendously with all protein preparation work ranging from western blot procedures to trypsin digestion and all stages of mass spectrometry. I am grateful for her ever-positive attitude and encouragement and immense amount of help. I would also like to acknowledge Amber Bible from Gladys Alexandre's lab for help with growth of all *Azospirillum* cultures.

I would like to thank the members of the Organic and Biological Mass Spectrometry group at Oak Ridge National Laboratory. I would like to specially thank Keiji Asano, who was always available for help with any and all mass spectrometer questions and issues. I am grateful to Greg Hurst for filling in when things were not going so well, and for his help with mass spectrometry experiments, data analysis and Microsoft Access. I would like to thank Miriam Land for help with BLAST searches.

I would like to acknowledge the support of the Genome science and technology program at the University of Tennessee, Knoxville. And finally, this research was sponsored by the Genomics:GTL program, Office of Biological and Environmental Research, U.S. Department of Energy, under contract No. DE-AC05-00OR22725 with Oak Ridge National Laboratory managed and operated by UT-Battelle, LLC.

Abstract

Mass spectrometry based proteomics has emerged as a powerful methodology for investigating protein expression. “Bottom up” techniques in which proteins are first digested, and resulting peptides separated via multi-dimensional chromatography then analyzed via mass spectrometry provide a wide depth of coverage of expressed proteomes. This technique has been successfully and extensively used to survey protein expression (expression proteomics) and also to investigate proteins and their associated interacting partners in order to ascertain function of unknown proteins (functional proteomics). *Azospirillum brasilense* is a free-living diazotrophic soil bacteria, with world-wide significance as a plant-growth promoting bacteria. Living within the rhizosphere of cereal grasses, its diverse metabolism is important for its survival in the competitive rhizospheric environment. The recently sequenced genome of strain Sp245 provided a basis for the proteome studies accomplished in this work. After initial mass spectrometer parameter optimization studies, the expressed proteomes of two strains of *Azospirillum brasilense*, Sp7 and Sp245, grown under both nitrogen fixing and optimal growth (non nitrogen fixing) conditions were analyzed using a bottom up proteomics methodology. Further proteome studies were conducted with *A. brasilense* strain Sp7 in order to ascertain the effect of one chemotaxis operon, termed Che1. In this study, proteomic surveys were conducted on two bacterial derivative strains, created earlier, which lacked either a forward signaling pathway or an adaptation pathway. The proteomic surveys conducted in this work provide a foundation for further biochemical investigations. In order to facilitate further investigation and a movement into functional proteomics, a set of destination vectors was created that contain a variety of tandem affinity tags. The addition of tandem affinity tags to a protein allow for generic purification schemes, and can facilitate future studies to investigate proteins of interest discovered in the first expression proteomic surveys of *A. brasilense*. Taken together, this dissertation provides a valuable data set for investigation into the physiology of *A. brasilense* and further provides biochemical tools for analysis of the functional protein interactions of *A. brasilense* cells.

Table of Contents

List of Tables

List of Figures

List of Attachments

Chapter 1. Introduction

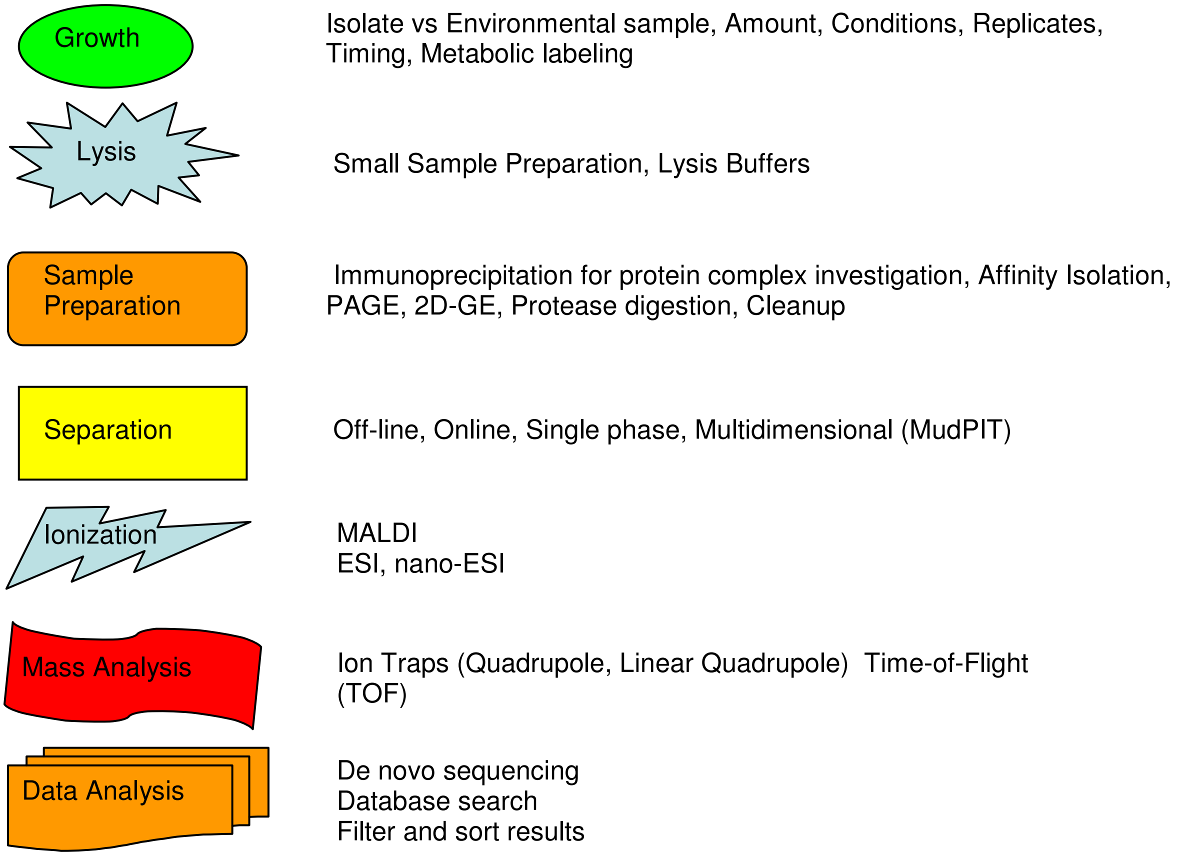
Proteomics tools

Biochemical techniques provide a wealth of information regarding the physiology of individual cells in terms of individual proteins and their structure, function and location within a cell, and participation within protein complexes and functional pathways. Combining this knowledge with systems biology information can serve to broaden the perspective of cell physiology under specific conditions. Systems biology approaches include acquisition of genome structure (genomics), which gives a picture of what is possible for a cell, as well as elucidation of which genes are transcribed in a given situation (transcriptomics). Further, systems biology approaches include a number of techniques that investigate the average population of proteins present within a cell as well as their modifications and possible function. The “expression proteome” of a cell consists of the entire set of proteins expressed in cell, while the “functional proteome” examines the protein-protein interactions on a genome-wide scale in an attempt to dissect functional pathways [1]. From these definitions, it follows that proteomics is defined as the study and identification of the full complement of proteins expressed in a cell under a given set of conditions. Due to its sensitivity and its ability to identify a large percentage of the proteins present in complex mixtures, mass spectrometry has emerged as a very powerful experimental tool for investigation of proteomes.

Shotgun proteomics is a term derived from DNA shotgun sequencing [2] and represents a methodology in which proteins in a mixture are digested with proteases to yield a very complex peptide mixture. Peptides within this mixture are then separated and analyzed using mass spectrometry, allowing corresponding proteins from which they were derived to be identified following analysis [3-5]. In tandem mass spectrometry experiments, parent peptide ions are first isolated and then fragmented through collision with an inert gas, giving rise to raw fragmentation spectra containing unique information about the parent peptide sequence [3]. Using database searching algorithms, individual peptide sequences can be derived from comparison of the fragmentation pattern of parent peptide ions with theoretical spectra derived from an available database [3, 6, 7]. When sequence information is not available, other algorithms deduce peptide sequences or parts of peptide sequences from the spectral information and then compare these “tags” to protein sequences derived from larger databases composed of sequence information from multiple closely related organisms to make protein identifications [8]. The primary goal of this dissertation was to employ a mass spectrometry-based proteomics work flow using shotgun sequencing methods to explore the expression proteome of *Azospirillum brasilense* cells.

Proteomics work flows

For most biologists, mass spectrometry represents a “black box” from which large amounts of data emerge. Within this black box are contained a number of different steps which make up an entire proteomics work flow as represented in Figure 1.1. At each step of the proteomic work flow, decisions need to be made and optimizations done in order to ensure the most useful and instructive data output. First and foremost, if investigating proteomes from bacterial isolates or eukaryotic cell lines, decisions about growth need to be made. First, the size of the culture needed for total proteome investigation or for



Growth

Isolate vs Environmental sample, Amount, Conditions, Replicates, Timing, Metabolic labeling

Lysis

Small Sample Preparation, Lysis Buffers

Sample Preparation

Immunoprecipitation for protein complex investigation, Affinity Isolation, PAGE, 2D-GE, Protease digestion, Cleanup

Separation

Off-line, Online, Single phase, Multidimensional (MudPIT)

Ionization

MALDI
ESI, nano-ESI

Mass Analysis

Ion Traps (Quadrupole, Linear Quadrupole) Time-of-Flight (TOF)

Data Analysis

De novo sequencing
Database search
Filter and sort results

Figure 1. 1 Steps in proteomics work flows
A number of individual steps are involved in a proteomics work flow. Decisions and optimizations at each step of the work flow are made based on the questions being asked. Amount of material needed for individual experiments dictates the culture size and conditions. Mass spectrometry platforms chosen include such decisions as separation techniques (online or off-line), ionization techniques (typically MALDI or ESI), and type of mass spectrometer used for analysis. Once data is obtained, choices must then be made about search algorithms used and types of statistical analysis applied. From this analysis, filtering levels must be chosen that are applicable to each specific data set to ensure significance of results.

isolation of protein complexes needs to be determined. For example, nitrogen fixing cultures had to be grown in a relatively small volume in order to minimize exposure to oxygen, which is detrimental to nitrogenase function. Culture volumes were therefore limited to 250 ml media in a 1liter flask. *Azospirillum brasilense* nitrogen fixing cells will grow only to an optical density of 0.1-0.2 under these conditions, yielding a wet cell pellet weight of ~0.05g. Resolubilization of this pellet in 250 µl buffer yields ~1 mg/ml protein concentration, which after clean-up and exchange into mass spectrometry compatible buffers will yield enough protein lysate at a concentration of ~ 1mg/ml for only 2-3 technical replicates. As a result, 2 or more biological replicates must be combined in order to have enough protein for multiple technical replicate mass spectrometry runs. More typically, larger cultures of bacterial isolates ranging from 500 ml to 1.5 L are grown for expression proteomes. A 500 ml *A. brasilense* non nitrogen fixing culture grown to mid-log phase (OD_{600} ~ 0.6) yielded a wet cell pellet weight of ~0.3 g. After resolubilization in 1.5 mL buffer, a protein concentration of ~1mg/ml was obtained and enough lysate was obtained for 6-7 technical replicates.

Additional decisions about growth include the culture media to be used for growth, the possible treatments or growth conditions to be applied to each culture if doing proteome comparisons, the number of replicates needed to establish statistical validity, and the amount of time required for growth, especially if doing a time course study. Optimization of each of those parameters is unique to the organism under study and the research question being addressed which will ultimately dictate the set of conditions and the type of experiments being performed. An additional choice at this initial step in the process can involve addition of labeling agents such as “heavy” nitrogen, ^{15}N [9], or “heavy” labeled amino acids [10], to the culture medium in order to shift the mass of proteins within a given sample. This labeled sample can then be mixed with a control sample during sample preparation steps, and the resulting mix used for direct comparisons of relative amounts of individual proteins per growth state or for ascertaining non-specific interactions between proteins [11, 12].

The choice of growth conditions combined with the desired downstream analysis dictates the lysis method used to retrieve the proteins from the cell for further analysis. For example, small sample sizes benefit from single tube lysis procedures such as those developed by Thompson et al [13]. Buffer compositions for lysis buffers need to be chosen carefully to facilitate downstream processing, as illustrated by the need for specialized or optimized lysis buffers to accommodate downstream processing in immunoprecipitation experiments or affinity isolations of single proteins. For instance, detergents found within some chemical lysis buffers can interfere with the ability of tagged proteins to bind to affinity resins (Qiagen, Valencia CA). Protein concentrations of samples need to be optimized through choice of lysis methodology and subsequent downstream concentration and clean-up steps, since both steps can lead to sample loss. Although mass spectrometry can detect proteins present at low femtomole levels within a sample [14], reproducibility of spectra acquired and ability to reliably detect and identify all peptides and thus proteins present suffers when the concentrations of proteins are low [15].

Further sample preparation methods may involve a number of steps. Complex protein mixtures separated in chromatographic steps in which peptides are eluted directly into the mass spectrometer (online chromatography) by methods such as Multidimensional Protein Identification Technology (MudPIT) [16-20] used in Chapters 2-4 of this work, may require only digestion and sample clean-up as further sample preparation steps. Since online separation methods minimize sample loss and can also minimize contamination from outside sources [21], this is an excellent methodology to use for small sample sizes. Liquid chromatographic separations coupled directly to the mass spectrometer can include both single dimensional separation, usually reverse phase, and multiple dimension separation [22]. For simple samples derived from a single purified protein or protein complex, a single dimension of separation may yield enough depth of coverage of the available proteins. However, if a more complex sample is being investigated, such as the subset of membrane proteins [22-24] or the total soluble proteome [19, 25-33], then one or more separation steps are demanded as part of the sample preparation in order to reduce sample complexity and improve the depth of proteome coverage. Reproducibility of chromatographic separation is of great interest, with a number of studies investigating the degree of reproducibility of both chromatographic retention times and subsequent mass spectra derived at a specific retention time in replicate runs [34, 35]. This may be especially significant when comparing two different growth states because retention times of the same peptide identified in both mass spectrometric runs can be used as validation of data results [35].

Alternatively, when a sample has a great deal of complexity, off-line separations in which chromatography is performed and fractions collected for later mass spectrometry analysis can also be performed. A single dimension separation of proteins via SDS-PAGE gel electrophoresis or multi-dimensional separation via 2-dimensional gel electrophoresis (2D-GE) are common steps in sample preparation [36]. This type of separation allows the proteins present to be directly visualized which may also provide a visual quantification of change in protein expression levels. At the same time, gel separation methods are limited by the protein concentrations present within the sample because of limits of detection of protein stains used in gels. Inherent physical properties of proteins may also prove to be a limiting factor in this separation methodology since gel electrophoresis tends to be biased for proteins within a specific pI range or molecular weight range [36]. Once visualized, the protein bands can be excised and in-gel digests performed before extracting peptides for subsequent mass spectrometry analysis. Due to the hydrophobic nature of membrane proteins and difficulties with solubility in liquid separations, separation by gel electrophoresis before analysis is the preferred methodology for initial separation of membrane proteins [23]. However, from personal experience, recovery of digested peptides from gels is poor. As a result, optimization of growth and lysis procedures, as well as subsequent preparation steps for maximal protein expression, is important for sufficient data output [37]. Off-line protein separation can also be accomplished through High Performance Liquid Chromatography (HPLC) separations with subsequent examination of the proteins present in individual fractions by mass spectrometry. This is a commonly used methodology for examination of intact proteins [38], or investigation of modifications such as phosphorylation [39] because it can allow enrichment of the desired protein or of specific protein modifications in a mixture, thus optimizing its concentration and subsequent mass spectrometry.

Once the samples are prepared, the “black box” of mass spectrometry, including ionization techniques, mass analysis and mass detection, can be applied. A number of excellent reviews [4, 5, 40-43] cover each of these elements of mass spectrometry, so they will be only briefly mentioned here. Investigation of large biomolecules was facilitated by the development of "soft ionization techniques" of Matrix Assisted Laser Desorption Ionization (MALDI) [44] and electrospray ionization (ESI) [45], which allowed for transfer of biomolecules into the gas phase without fragmentation. These two ionization techniques are fundamentally different. When using a MALDI source, proteins or peptides are first mixed with a matrix material, spotted onto a plate, and dried. The plate is placed onto the MALDI source in a vacuum and a laser pulse is used to ablate both matrix and analyte molecules from this solid surface to form ions, which then enter the mass analyzer. Electrospray ionization is accomplished through application of voltage to a needle at atmospheric pressure through which preformed ions in a liquid phase are transformed into the gas phase and subsequently sprayed into the mass analyzer.

The most commonly used mass analyzers in proteomics experiments are Time-of-Flight (TOF) and ion trap mass analyzers [5]. MALDI ionization sources are most often coupled with TOF analyzers, which measure the mass-to-charge ratio (m/z) of the ions present by detecting the time it takes an ion to traverse the length of a flight tube under a vacuum [5]. A high voltage is applied to the MALDI source and ions enter the flight tube and travel down the length of the tube to the detector, with larger molecules traveling at slower rates than smaller ones. Mass is determined by the length of time it takes the ion to traverse the flight tube. ESI, and its lower flow-rate sister, nano-ESI, are most often coupled with ion trapping mass analyzers, either linear ion trapping quadrupoles or three-dimensional trapping quadrupoles [5]. Quadrupole instruments trap ions in either 2 or 3 dimensions through the application of radio frequency (rf) and DC voltages. Detection of ions stored within the trap is accomplished by ramping voltages such that ions of decreasing mass-to-charge ratio become unstable and exit the trap where they are then detected as they hit an electron multiplier tube. See March [46, 47] and Schwartz [48] for thorough reviews of ion trap operation and theory. Although there are a number of other mass analyzers, including hybrid configurations of the above-mentioned mass analyzers, they will not be mentioned here.

Tandem mass spectrometry (MS/MS) is a powerful technique for investigating protein sequence. Ion trapping instruments lend themselves well to tandem mass spectrometry experiments because of ease in trapping and isolating ions of interest although a number of other mass analyzers can also perform MS/MS. During tandem mass spectrometry experiments in an ion trap, the parent ion of interest is isolated within the trap and then fragmented, often through a process known as collision-induced dissociation (CID) [49]. In a CID experiment (Figure 1.2), the isolated parent ions are subjected to collision with a neutral inert gas. Energy from repeated collisions is transferred to the ion causing it to vibrate. When enough energy is accumulated, the peptide backbone bonds break, primarily at amide bonds resulting in b and y type ions (Figure 1.2). The resulting fragment ions are then measured and compiled to give a series of peaks representing the measured mass-to-charge (m/z) ratio of the product ions [49]. The distance between the individual peaks in a spectrum represents the mass of an amino acid, thus giving rise to sequence information, as represented in Figure 1.2.

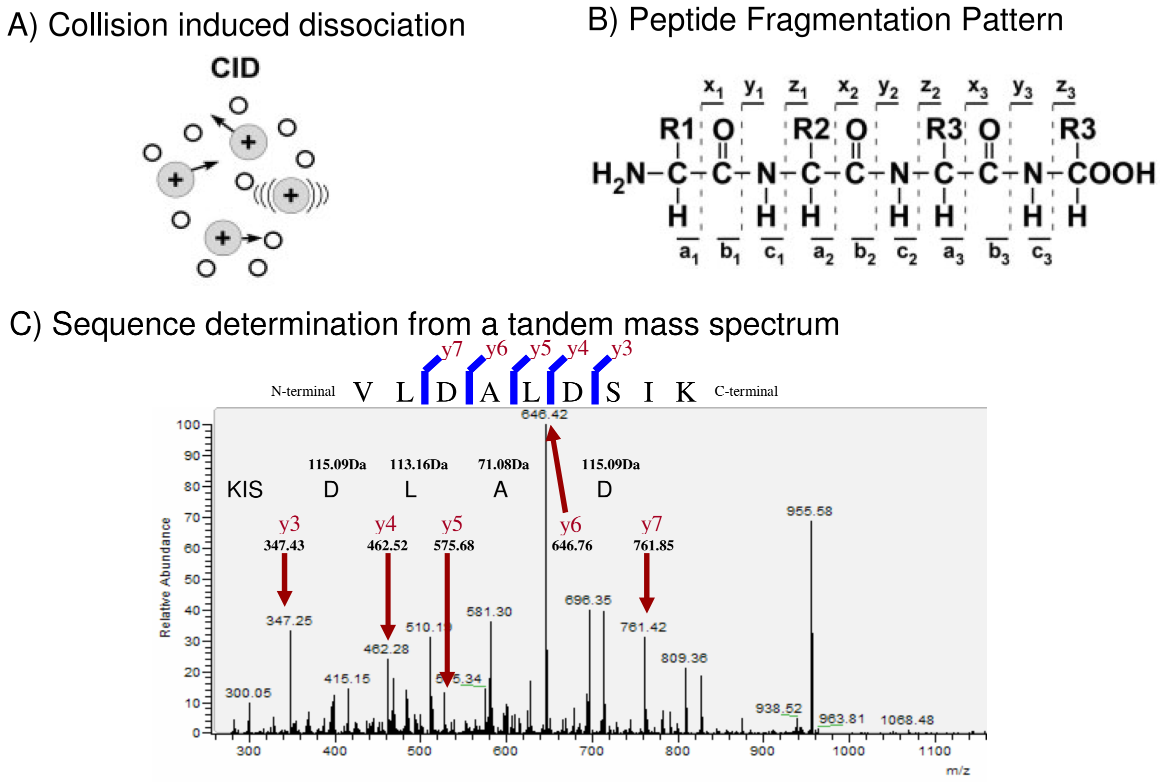


Figure 1. 2 Creation of fragment ions in tandem mass spectrometry A) CID energy transfer process, B) Fragmentation patterns of peptides C) MS/MS spectra of peptide VLDALDSIK from carbonic anhydrase.

A) Analyte molecules, represented by large filled circles, collide with an inert gas, represented by small open circles, causing a buildup of energy in the molecule that causes rupture of the weakest bonds of the analyte molecule. B) Fragmentation patterns of peptide molecules are predicted by common breakage points. In CID experiments, the peptide ion breaks most commonly at the peptide bond, creating “y” and “b” type ions, represented above by the dotted lines above which are y1, y2, or y3 and below which are b3, b2, or b1. The direction of the slash above the dotted line indicates which ion fragment retains the charge, thus making it “visible” to the mass spectrometer. Thus y-type ions retain charge on the C-terminal end and b-type ions on the N-terminal end. C) Sequence determination is accomplished through examination of the distance by which individual peaks are separated from one another. The above spectrum shows“y-type” ions created through a CID process. The first visible peak in the spectrum at 347.25 m/z is the y3 ion made up of amino acids lysine, isoleucine and serine. The next peak used to identify the peptide sequence is at 462.28 m/z (labeled above as y4). The fragment ion contributing this peak is made of amino acids K, I, S, and aspartic acid, D. The mass of aspartic acid is 115.09 Da (shown above the D), and thus the y3 and y4 peaks are separated by 115 Da due to the addition of the D amino acid in the fragment ion sequence. The y5 ion at 575.34 m/z is composed of KISD and leucine (L). The molecular weight of leucine is 113.16 Da and thus the peaks of the y4 and y5 fragment ions are separated by 113 Da. Fragment ion peaks for each additional peak labeled in the spectrum above are separated by the individual weights of the additional amino acids. When search algorithms are assigning peaks, the experimental spectra such as that shown above are compared to theoretical spectra containing peptide fragment ions derived from sequences within the database. The distance of the peaks from one another allows the search algorithm to assign sequence information to the experimental spectra complete with a score that represents how well the experimental spectrum matches the theoretically derived spectrum.

Once raw MS data has been acquired, interpretation of individual spectra can be done manually or by automated processing strategies. A number of software programs have been developed to extract information from raw mass spectra, thus facilitating fairly rapid analysis of MS and MS/MS results. For those organisms which have no genome sequence available, *de novo* programs such as GutenTag [8], DirecTag [50] or Lutefisk [51] derive sequence information directly from the raw spectral data, with the distance between peaks in MS/MS spectra representing the mass of individual amino acids in a sequence. Sequence tags consisting of 3 or more amino acids identified by distances between abundant peaks in the experimental spectrum can then be compared to a large database, such as the NCBI non-redundant database [52], to make possible protein predictions. When genomic sequence information for an organism is available, protein databases can be developed based on open reading frames which predict proteins from genomic sequence. Database search algorithms such as SEQUEST [3] or MASCOT [53] can then be used to compare raw spectra to theoretical spectra derived from *in silico* digest of the protein sequences within the database, with scores given for each available match. Programs such as DTASelect [54] then filter and sort the data output from the search algorithm, putting final protein lists together in an easily readable format.

Regardless of the type of experiment done, the quality of raw data collected will certainly affect the ability of the search algorithms to make true positive identifications. When surveying an entire proteome of a cell, a vast amount of data is acquired and the possibility for false positive identifications is high. Studies have been done both to address the issues of optimization of data quality being input to search algorithms [15, 34, 55-61] and of optimization of filter levels to ensure quality data output after searches are performed [62].

In a tandem mass spectrometry experiment, a number of parameters can be manipulated by the experimenter to ensure an optimal data output. Since ion trap instruments were used in the remainder of this work, we will only consider parameters manipulated in these instruments. In an ion trap, these parameters include amount of time required to fill the ion trap [57, 60], scan rates [34], collision energy required for fragmentation of the parent ion [57], parent ion isolation m/z window, and number of averaged scans (microscans) contributing to the final scan [58-60]. The ion trap fill time, scan rates, collision energy, and number of averaged scans have been investigated, with optimized ranges of operation being determined. The experiments investigating the optimum number of averaged scans [58-60] have been controversial, and the width of the isolation m/z window chosen for isolation of a parent ion has not been investigated at all.

In Chapter 2 of this thesis, two tandem mass spectrometry parameters, number of microscans and width of the isolation window selected when isolating the parent ion, will be discussed in more detail. The isolation m/z window width and the number of averaged scans (microscans) were optimized for experiments done on both a linear ion trapping quadrupole instrument (LTQ) and a 3-dimensional trapping quadrupole instrument (LCQ). We began our optimization with a mixture of known proteins in which we could get definitive identifications, and then tested our parameters using a total proteome preparation of the soil bacterium, *Rhodopsuedomonas palustris*. Results from this work were then applied to subsequent investigation of the soluble proteome of two different strains of *Azospirillum brasilense*, described more thoroughly below.

Mass spectrometry based proteomics in bacterial systems

The last few years have seen an explosion of mass spectrometry-based proteomics experiments. A simple search in PubMed using the words "microbial proteomics" returns 418 articles, with over 80 of those articles being published in the last year. The goal of many earlier proteomics experiments in bacteria and archaea was to simply ascertain the protein complement of growing cells, often with the added goal of contributing to the genome annotation of newly sequenced strains [25, 26, 63, 64]. More recent studies have not only characterized the complement of proteins expressed within growing cells, but also have included comparisons of the cells grown under different conditions, such as comparing the proteome expression profiles of *R. palustris* cells grown under different metabolic states [30] or the comparison of the proteomes of *Nostoc punctiforme* cells grown under nitrogen and non nitrogen fixing conditions [29]. Other studies seek to determine proteins involved in pathways contributing to changes in bacterial function by detecting changes in expression levels of proteins when microbes are exposed to different growth environments, such as identifying a putative set of proteins involved in nodulation [26] or pathogenicity [65] or metabolism of alcohol [66].

There are a number of ways to perform comparative analysis of proteins within samples, either using labeling techniques or using label-free techniques. Labeling of samples for subsequent mass spectrometric analysis allows for direct comparison of individual protein abundance between two different growth states by comparison of the mass shift induced in labeled peptides with their unlabeled counterpart. Labeling can be accomplished at different points in the sample preparation pathway. For instance, mass shifts can be introduced into proteins by introducing a heavy isotope during growth such that all proteins synthesized during growth will incorporate the heavy isotope [9, 10]. Alternatively, proteins can be labeled after harvest by addition of labeling agents such as heavy isotope ^{18}O to samples. In this instance, proteins are proteolytically digested, dried and resolubilized in "heavy" water containing a heavy isotope of oxygen ^{18}O. Two molecules of heavy oxygen are incorporated into the carboxy groups at the ends of the peptide, thus shifting the peptide mass by 4 Da [67]. Also late in the preparation pipeline, at the final step in preparation, peptides can be labeled with the isobaric reagent, iTRAQ®, which does not cause a mass shift, but instead releases a signature reporter ion on fragmentation, allowing for quantification of abundance in the labeled sample [68]. Recently, this methodology was used to quantify changes in 721 proteins in *Nostoc punctiforme*, both filament and heterocyst, upon shifting to nitrogen fixation [29]. The changes in individual protein abundances were detected via presence of reporter ion in the MS/MS scans. The proteins showing changes in abundance were mapped back to metabolic pathways, and allowed confirmation of existing pathway data as well as discovery of novel participants in metabolic pathways affected by a shift to nitrogen fixation. Although this study is an excellent example of the use of iTRAQ technique, our lab has found the reagents expensive and difficult to use, with the further complication of inefficient labeling of proteins.

An interesting example of a labeling technique for bacterial cultures termed I-DIRT [9] includes growth of a test culture in media containing heavy isotope components of ^{15}N nitrogen alongside a control culture that does not contain the heavy isotope. The two cultures are harvested and mixed together in equal weights, and the mixed cultures are then analyzed using mass spectrometry. Use of heavy nitrogen allows all proteins within the test sample to be labeled in such a way as to cause a shift in mass in the test culture. When mixed with a non-labeled culture and analyzed via mass spectrometry, twin peaks will be observed for peptides present in both samples, one from the non-labeled control, and a second mass-shifted peak from the labeled test sample. Calculation of the peak area for each of these spectra gives the measure of relative abundance of the peptide within the heavy labeled test sample in direct comparison to relative abundance within the non-labeled control. Recently this labeling technique was applied to a global analysis of *E. coli* and *R. palustris* cultures with protein detection done using MudPIT [11]. In addition to providing information about individual proteins and their interacting partners after immunoprecipitation experiments, quantitation of all proteins detected was done to assess the global effect of expressing tagged proteins off a plasmid. Direct comparison of individual peptide areas can be done to provide an accurate comparison of relative abundance in different cultures.

However, in spite of the success of labeling techniques in quantification of relative abundances of proteins between two samples, one labeled and one unlabeled, the labeling reagents are expensive, and data analysis of mixed samples is often time consuming and quite complicated [69]. Label-free techniques have the advantage of two cultures being grown in media with similar components, such that any possible interference from heavy isotopes during growth is avoided. In addition, label-free approaches can offer the advantage of relatively uncomplicated data analysis. In shotgun proteomics experiments analyzed by using MudPIT [20], spectral count of peptides has been shown to correlate well with relative abundance of a protein present within an individual sample [70]. Spectral abundance factors (SAF) can be calculated by taking into account how often a given peptide is sampled by the mass spectrometer (spectral count or SpC) divided by the total length of the protein. Dividing by the length of the protein accounts for the probability of higher spectral counts due to a higher number of peptides present with increasing length of the protein [71]. Since different replicates will have different numbers of identified peptides and proteins, normalization of SAF (NSAF) for each individual run allows for comparison of relative protein abundance across samples [72]. In Chapter 3 of this work we have used label-free shotgun proteomics to investigate the baseline proteomes of *A. brasilense,* a newly sequenced organism. Further, NSAF values have been calculated and used to investigate the expression levels of proteins detected when isolates were grown under both nitrogen fixing conditions and non-nitrogen fixing conditions (see below). While chapter 3 investigates the effect of differing environmental conditions, chapter 4 of this work utilizes the same proteomic survey method to examine the effect of mutations in one specific chemotaxis-like pathway (see below) on the overall proteome expression in *A. brasilense* Sp7 cells.

Functional proteomics

By nature, expression proteomics is a discovery-based technique, simply elucidating the proteins that are present at a given time under specific growth conditions. This survey can in turn produce insights into physiological function and can lead to identification of target proteins that need further functional exploration. One methodology for exploration of protein function includes examination of the interacting partners of a particular protein. Proteins often function within a cell as part of a "molecular machine" with associated proteins all serving to contribute to the overall outcome of the "machine". One such example includes DNA-directed RNA polymerase, in which 5 subunits associate and function together (along with a number of associated proteins) to transcribe DNA into messenger RNA. Thus the function of individual proteins can often be elucidated by determining the molecular machinery (protein complex) with which it participates [1]. Classic gold-standard biochemical techniques for determining protein interacting partners on a large scale include yeast 2 hybrid techniques and phage display techniques [1]. These, however, are limited to direct pairwise interactions and investigations are often dictated by what is currently known. Mass spectrometry-based techniques, on the other hand, are not limited to detection of only pairwise interactions, but instead can detect all proteins present within a complex [16]. Recently developed techniques include strategies for tagging proteins with an affinity epitope such that proteins and their associated interacting partners can be isolated on a genome-wide scale, thus facilitating development of large protein interaction networks for model organisms such as *E. coli* [73-75] or *S. cerevisiae* [16, 76-80], as well as in soil bacterium *R. palustris* [81]. A strategy in which a protein of interest in fused to an affinity purification tag provides an efficient means with which to isolate a protein and associated interacting partners [82]. Since single affinity tags can lead to isolation of a large number of non-specifically interacting proteins, tandem affinity tagging systems have been developed in which two affinity tags are fused to either the C-terminal or N-terminal end of a protein, thus allowing for purification of proteins to near homogeneity

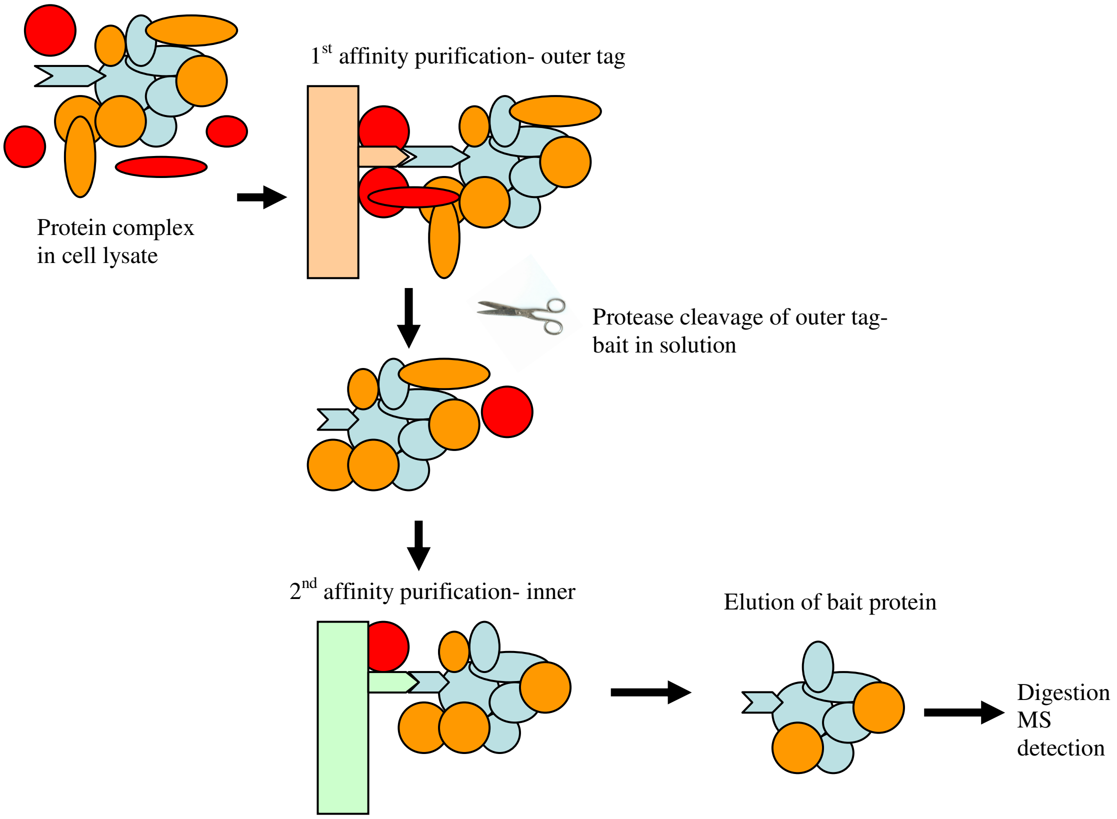

Protein complex in cell lysate
1st affinity purification- outer tag
Protease cleavage of outer tag-bait in solution
2nd affinity purification- inner
Elution of bait protein
Digestion
MS
detection

Figure 1. 3 Tandem affinity purification steps
Proteins tagged with a tandem affinity tag can be purified in a two step-purification process. Cell lysate containing a tagged bait protein is affinity purified using first via the outer tag. In this purification step, the bait protein is tightly bound to the resin, and a number of non-specific interacting proteins are present in the mixture bound to both the bait protein and to the affinity resin. The outer tag is then cleaved off, leaving the bait protein and its associated interacting partners free in solution. The second affinity purification is done using the remaining inner affinity tag. After incubation, washes help to remove a majority of non-specifically bound proteins. Gentle elution and subsequent proteolytic digestion then prepare the proteins and interacting partners for detection via mass spectrometry. All steps are performed using physiological buffers.

and decreasing the numbers of non-specifically interacting proteins [75, 76, 80, 83, 84]. Inclusion of a protease cleavage site between the tags allows for purification strategies done under physiological conditions, as represented in Figure 1.3. When fused to a tandem affinity tag containing a protease cleavage site between the two motifs, the protein of interest ("bait" protein) is first purified by immunoprecipitation using the outermost high affinity tag. For this initial purification, physiological buffers are used, thus preserving non-covalent interactions between protein interacting partners. The innermost tag is then cleaved off through incubation with an appropriate protease, again in physiological buffers, leaving the tagged protein and its associated interacting partners free in solution. A second immunoaffinity step is then performed, concentrating the protein and its associated interacting partners and further washing non-specifically interacting proteins away. The innermost affinity tag is chosen such that gentle elution conditions may be employed from this affinity purification step. The use of physiological buffers in all steps of the purification strategy, combined with the gentle elution conditions, serve to facilitate preservation of non-covalent interactions between the protein of interest and its interacting partners, thus preserving protein complexes that may be lost when harsher purification and elution conditions are employed [1]. Disadvantages to affinity tags include possible interference of the tag with protein folding or structure, or interference of the tag with interaction sites [85]. Further, since proteins contain a wide variety of sequences and structures, expression levels can differ with different tag moieties or a protein can be insoluble with one tag but not another [85]. In order to increase functional proteome coverage, it is therefore desirable to have a set of tags with which to tag proteins. With that in mind, in Chapter 5 of this work, we present development of a set of destination vectors based on pBBR-Dest42 created earlier [81]. The advantages of this vector set include the broad host range expression afforded by the pBBR replicon of the parent vector and ease of cloning through the inclusion of a Gateway cloning cassette in the parent vector. The new vector set includes a choice of C-terminal tags based on tandem affinity tags that have been previously proven to function efficiently to isolate protein complexes in bacterial and eukaryotic systems. RNA polymerase subunits were then expressed as a fusion with each individual tag and used as bait proteins in immunoprecipitation experiments with generic protocols. Prey proteins that co-immunoprecipitated with the bait were then detected via mass spectrometry as a proof of concept for the vector set. Results are presented in Chapter 5.

Azospirillum brasilense

Rhizospheric habitats are defined by the soil regions surrounding plants in which are found abundant root secretions in a concentration gradient that varies inversely with increasing distance from the plant [86]. This habitat offers a great variety of nutrients that serve as a source of both carbon and energy for bacterial cells that live within the rhizosphere. Plant root exudates include organic acids, amino acids, sugars, vitamins, and enzymes [86]. The specific composition of the nutrients found within a given soil environment is dependent upon the plant growing therein. Further, this composition changes as a function of distance from the root [86]. Motile bacteria such as *Azospirillum* species with ability to utilize a number of different carbon and energy sources possess a competitive advantage in these environments. Interaction with the plant root hair is dependent upon the ability of the bacteria to show biased movement, or chemotaxis, to the root exudates [87]. Additionally, the bacterial surface composition facilitates interaction with the root hairs and further mediates formation of bacterial aggregates which can facilitate nitrogen fixation capabilities [88]. The ability to fix nitrogen and to show chemotaxis to optimal environments are important adaptation strategies to the challenge of living in the rhizosphere where competition for limited nutrients is fierce.

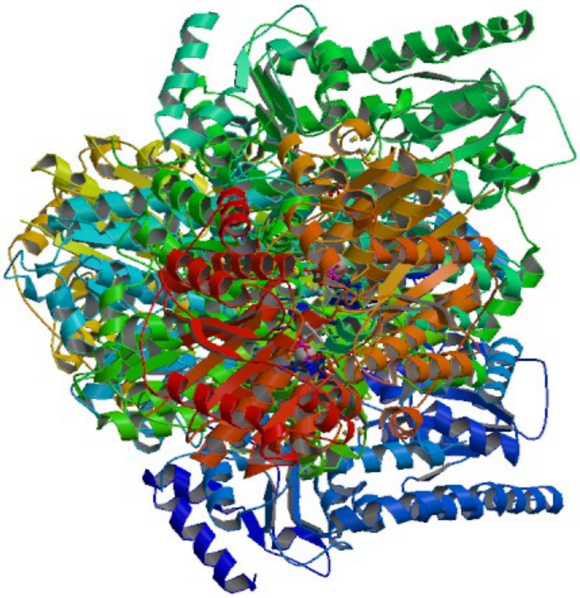
Azospirillum brasilense are free-living soil bacteria that are found within the rhizosphere of grasses and cereals in tropical and sub-tropical climates. They are a member of the alpha subclass of proteobacteria, and are capable of promoting plant growth upon inoculation of the roots of cereal grasses and grains [89]. *Azospirillum* colonizes the root hair zone where easily metabolizable nutrients such as organic acids and low molecular weight sugars are abundant [86]. *Azospirillum* species can use a number of carbon and nitrogen sources. They are motile, and show taxis to root exudates, but with primary taxis response to optimal low oxygen levels [90], allowing them to adapt well and outcompete other bacterial species found within the rhizosphere. Further, although they do not survive well in nutrient poor environments, they are able to store carbon for later use in the form of poly-3-hydroxybutyrate (PHB) granules and also to form cyst-like structures that facilitate survival under less than optimal conditions [86]. Although they do fix nitrogen, they do not release large amounts of ammonia into the soil, so their plant growth promotion is thought to be primarily due to the secretion of plant-growth promoting hormones such as indole-3-acetic acid (IAA) that increase volume and number of root hairs, thus leading to better uptake of nutrients and creating a stronger, healthier, bigger plant [91]. Because of their physiological adaptability and positive effects on growth of cereal grains, *Azospirillum* species are of considerable interest as inocula for cereal crops world-wide [91, 92].

In chapter 3 of this work, the total proteomes of two different strains of *A. brasilense*, Sp245 and Sp7, grown under both nitrogen fixing and non nitrogen fixing conditions are investigated. *A. brasilense* strain Sp245 was originally isolated from wheat roots [89] and has been shown to colonize both the outer surface as well as the interior of wheat root hairs [93]. *A. brasilense* strain Sp7 was originally isolated from the rhizosphere of *Digitaria decumbens* (Brazil) [94], and colonizes only the outer surface of root hairs [93]. Since each strain occupies a different ecological niche, it is likely that each will respond differently to environmental challenges. Exploring the complement of expressed proteins within each strain under both nitrogen fixing and non nitrogen fixing conditions will contribute to understanding the differences in metabolic profiles and physiology of these strains.

Nitrogen fixation

Although air contains almost 80% nitrogen which diffuses into the soil, it is not in a form that is readily available for plant uptake. Therefore, nitrogen-fixing soil bacteria, either free-living or endophytic, are essential for converting nitrogen gas to a more readily available form. This is accomplished through the activity of the prokaryotic-specific nitrogenase enzyme complex encoded by the *nif* operon, which encodes dinitrogenase reductase (also known as Fe protein) and dinitrogenase (also known as FeMoCo or FeMo protein) [91]. Dinitrogenase, a heterotetramer of two subunits containing a molybdenum-iron (FeMo) cofactor (Figure 1.4), is responsible for reduction

A) PDB 1M34



B. PDB 1M1Y

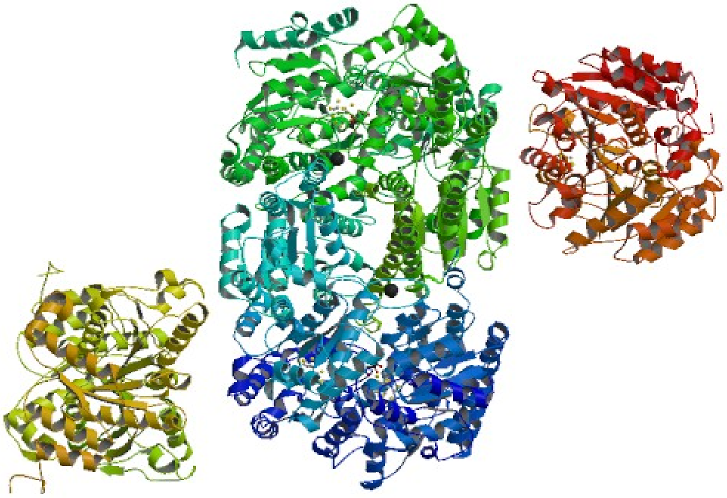


Figure 1. 4 Ribbon structure of nitrogenase enzyme complex from *Azotobacter*
A) Protein data bank (www.rscb.org) accession number 1M34 Nitrogenase complex from Azotobacter Vinelandii stabilized by ADP-tetrafluoroaluminate B) Protein data bank accession number 1M1Y Chemical crosslink of nitrogenase MoFe protein and Fe protein

of nitrogen to ammonia, which is then assimilated into the amino acids and proteins within the bacterial cell through the action of glutamine synthetase. Dinitrogenase reductase, a homodimer containing an iron (Fe) cofactor (Figure 1.4), serves to donate electrons from ferredoxin to dinitrogenase [91]. This energetically expensive process is tightly regulated on several levels from translation of nitrogenase structural components to post-translational modification of the nitrogenase enzyme complex [95, 96]. The nitrogen regulation system (Ntr operon) regulates nitrogenase expression based on general nitrogen metabolism within the cell. There are four proteins within the Ntr system system: dual function uridylyltransferase/uridyl-removing enzyme GlnD, P_{II} protein GlnB, and the two component signaling system composed of histidine kinase, NtrB, and its cognate response regulator, NtrC [97]. GlnD functions as an intracellular sensor of nitrogen status through response to glutamine levels in the cell. Based on these levels, GlnD can function either to add an uridyl group to GlnB or to remove it. GlnB also serves as an intracellular nitrogen sensor for the cell based on both direct sensing of alpha-ketoglutarate and on indirect response to glutamine levels through its modification state [98, 99]. In low nitrogen conditions, GlnD transfers an uridyl group to GlnB causing it to dissociate from histidine kinase, NtrB. This dissociation in turn activates NtrB, allowing transfer of a phosphate group to NtrC. NtrC-P can then bind to upstream activation sequences in concert with sigma-54 RNA polymerase (also called RpoN or NtrA) to stimulate transcription of all genes containing sigma-54 promoter sites (Figure 1.5). Therefore, the proteomes of *Azospirillum* species grown under nitrogen fixing

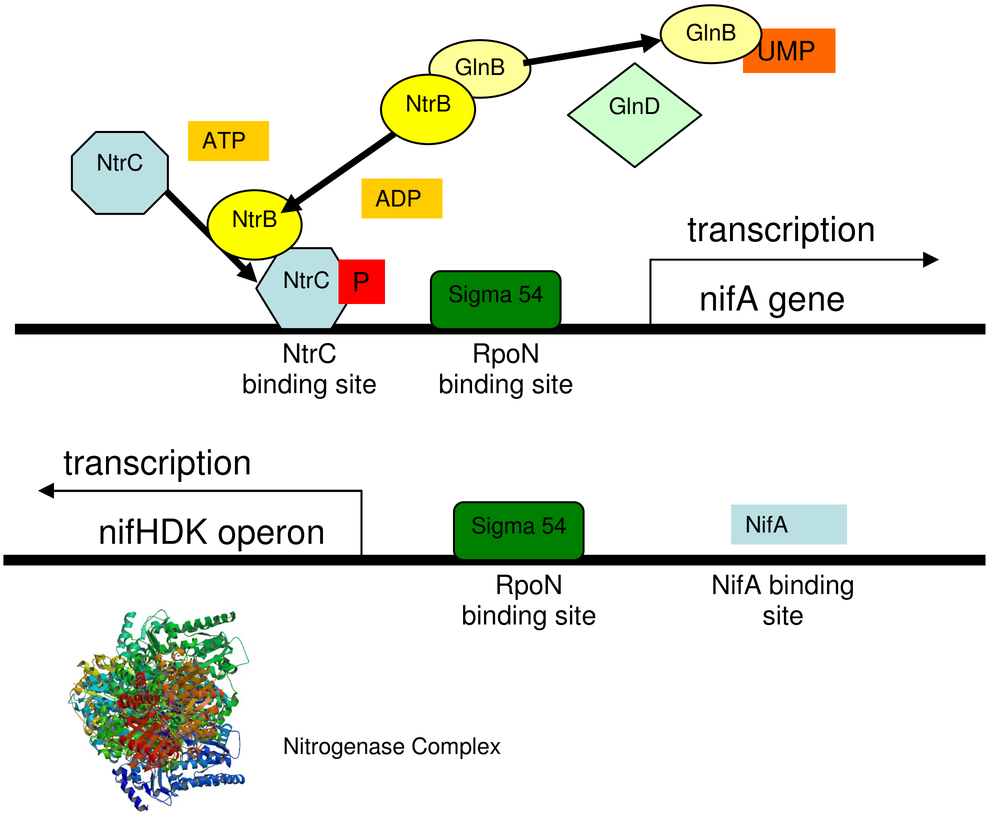


Figure 1. 5 Levels of regulation in expression of nitrogenase structural proteins and related synthesis proteins

Due to the energetically expensive process of nitrogen fixation, several layers of regulation are inherent in expression of the structural components of the nitrogenase enzyme complex as well as in expression of related proteins. GlnB and NtrB exist in a complex in the cell. When GlnD senses low nitrogen content, it serves to uridylylate GlnB, causing a dissociation of the GlnB-NtrB complex. NtrB is then free to bind with transcription factor NtrC-P to effect transcription of transcription factor nifA. NifA and sigma 54 binding to consensus sites upstream of the nifHDK operon then facilitates transcription of the structural components of the nitrogenase complex as well as genes related to nitrogenase synthesis.

conditions exhibit a different profile than those grown under non nitrogen fixing conditions, where sigma-54 RNA polymerase is not active.

Once transcribed, the activity of the nitrogenase complex is further regulated by post translational modifications that are also tied not only to nitrogen sensing systems within the cell, but also to oxygen levels as oxygen is a potent inhibitor of nitrogenase activity [98]. When combined nitrogen is not readily available, *A. brasilense* fixes atmospheric nitrogen under microaerophilic conditions, preferring an oxygen content of about 0.5% [100]. Both high levels of fixed nitrogen and higher oxygen content results in ADP-ribosylation of the Fe protein of nitrogenase, dinitrogenase reductase, effectively turning the enzyme off [91, 101]. A return to optimal oxygen content or unavailability ofa sufficient concentration of fixed nitrogen results in removal of ribosylation, restoring nitrogenase function and ultimately resulting in nitrogen fixation [98].

Much of nitrogenase experimental research has been performed in alpha-proteobacterium *Azotobacter vinelandii* due to the ease with which nitrogenase can be purified from this species [97]. *Azotobacter* controls transcription of the nitrogenase protein through associations between NifA and NifL proteins [96, 97]. NifA is a sigma-54 dependent transcriptional activator required for transcription of nitrogenase structural genes. It has an N-terminal GAF domain, a central sigma-54-dependent activator domain and a C-terminal DNA binding domain [97]. A variable region between the central domain and the C-terminal DNA binding domain can either contain a CXXXXC motif, making the NifA protein oxygen sensitive as it does in *Azospirillum* species, or can lack the CXXXXC motif, making it oxygen-resistant as it is in *Azotobacter* species [97]. In those species in which NifA is oxygen resistant, NifL is a partner protein to NifA, expressed in a 1:1 stoichiometry [97]. NifL has an N-terminal PAS domain which functions as a redox sensor, and a Histidine kinase-like C-terminal domain that binds ATP and ADP, and serves to inhibit NifA activity in response to ADP and fixed nitrogen levels [97]. In both *Azotobacter* and *Azospirillum*, NifA (and thus NifL in *Azotobacter*) is constitutively expressed, but *Azospirillum* species do not have a *nifL* gene within their genome. Expression levels of NifA in Azospirillum are moderated by oxygen and ammonia content [91]. Activity levels of NifA are controlled through interaction with GlnB (not NifL as in *Azotobacter* species), perhaps as an indicator of nitrogen status of the cell [91]. Due to these differences, *A. brasilense* has served as a model for nitrogen fixation in *Azospirillum* species for a number of years. In Chapter 3 of this work, we explore the expression of nitrogenase and associated proteins for both Sp7 and Sp245 strains as well as examine the overall changes in the expression proteome during nitrogen fixation in *Azospirillum brasilense* cells.

Earlier experiments suggested the presence of an alternative nitrogenase containing an iron-only cofactor in *A. brasilense* strain Cd (a close relative of Sp7) [102], but additional genetic evidence has not been forthcoming. Interestingly, the genome sequence of strain Sp245 has indicated the presence of an alternative nitrogenase which is proposed to use vanadium as a cofactor instead of molybdenum. A number of bacterial species have been shown to have alternative nitrogenase systems in which the nitrogenase uses vanadium as a cofactor for dinitrogenase (V-Fe protein), or alternatively uses only iron [103-105]. However, most alternative nitrogenases are hierarchically expressed, and so require total chelation of metals from the growth media, or alternatively, deletion of the genes for nitrogenase structural components containing molybdate co-factors, in order to see expression of the alternative molybdenum-free nitrogenase. With this in mind, cultures of *A. brasilense* Sp245 were grown with and without added molybdate, with vanadium added to the culture in the absence of added molybdate. Due to the difficulties inherent with total chelation of metal from all media components and glassware, we elected to simply add vanadium to our cultures in order to test for expression of this alternative nitrogenase within the proteome studies. Data from both the nitrogen fixing cultures grown with vanadium and with molybdate are included in Chapter 3 of this work, and comparisons are made between protein expression levels in both vanadium and molybdate nitrogen fixing cultures with those in control cultures grown under non nitrogen fixing conditions.

Chemotaxis

Motile bacteria have a complex and highly sensitive sensory system that allows them to sense and respond to minute changes in their environment [106]. Biased swimming in response to chemoeffector gradients in motile bacteria is known as chemotaxis, where increases in chemoattractants result in biased swimming up the concentration gradient while increased concentrations of chemorepellants has the opposite effect [106]. In a homogeneous environment, the swimming behavior of bacteria is erratic, with regular tumbling resulting in random changes of direction [106].

The best known and most well characterized signaling system in bacteria is the *E. coli* chemotaxis system [107], a large signaling complex localized to the poles of the *E. coli* cell. Signal transduction in this system is mediated by a two-component system consisting of histidine kinase, CheA, and cognate response regulator, CheY, whose sequences are conserved throughout all bacterial species [106]. Sensory input to the signaling system is accomplished through binding of chemoeffectors to membrane spanning proteins known as methyl-accepting chemotaxis proteins, or MCPs. MCPs contain variable periplasmic sensing domains, with conserved cytoplasmic signaling domains [106, 108]. Clusters of MCPs at one or both poles of an *E. coli* cell bind chemoeffectors and then undergo a conformational change, working cooperatively to amplify and transmit environmental signals [108-110]. This conformational change in MCP proteins is transmitted via docking protein CheW to histidine kinase CheA, which then undergoes autophosphorylation [109, 110] (Figure 1.6). The phosphate group is subsequently transferred to response regulator CheY, which then diffuses through the cytoplasm and binds to flagellar motor switch protein, FliM, to change the direction of flagellar rotation [109]. In *E. coli*, counterclockwise rotation of the flagellar bundles allows formation of a single bundle, and results in smooth swimming, while clockwise rotation causes the bundles to dissociate resulting in tumbling behavior and thus a random change in direction [106].

Environments are not usually static in nature, so sensing and signaling systems must be reset in order to respond to new and ever-changing situations. Two modifications provide a re-set of the *E. coli* chemotaxis system [108, 109]. Phosphatase CheZ dephosphorylates CheY, thus resetting the forward signaling pathway and allowing CheY to diffuse back to the large receptor complex at the opposite pole [106]. S-Adenosylmethionine (SAM)-dependent methyl-transferase CheR and methyl-esterase CheB work together to modulate the level of methylation of glutamate residues within

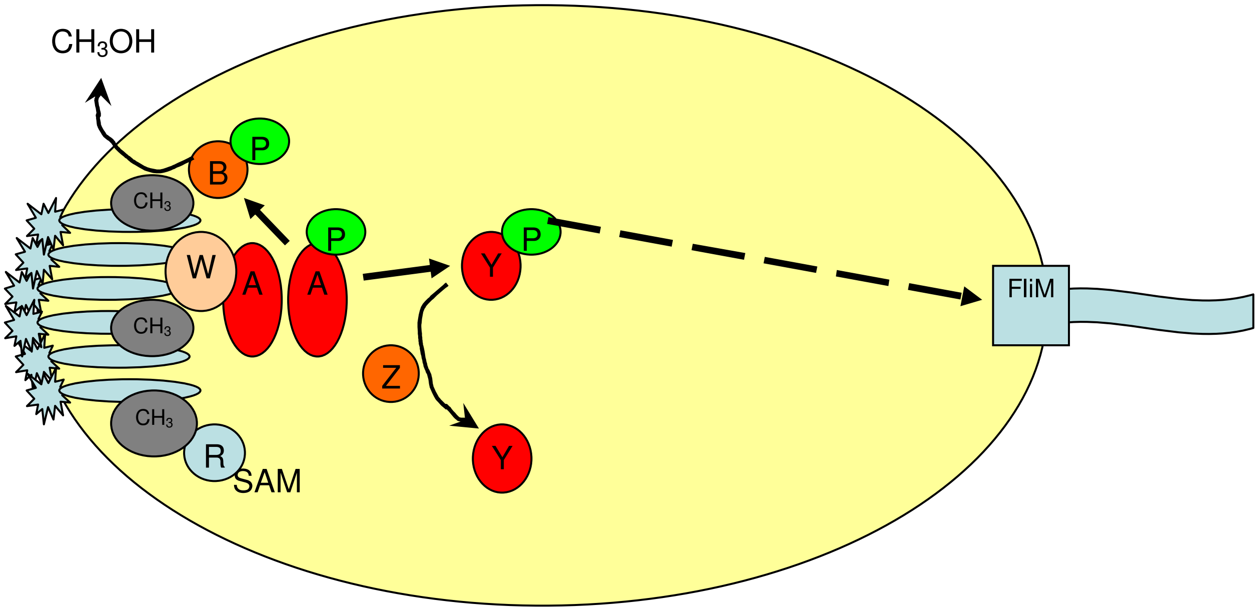


Figure 1. 6 Prototypical chemotaxis pathway in *E. coli*
Membrane spanning methyl-accepting chemotaxis proteins (MCPs) cluster at the pole of the *E coli* cell, represented above by the light blue stars with tails at the pole of the cell. Binding of effectors to these MCPs causes signal transmission to CheA histidine kinase, represented above by red ovals labeled with "A", through adaptor protein CheW, represented above by tan circle labeled with "W". CheA autophosphorylates on histidine residues and then transfers this phosphate (green circle labeled "P") to an aspartate residue on cognate response regulator CheY (red oval labeled "Y"), which then binds to molecular motor protein FliM to effect a change in direction of the flagella. Adaptation to signal is accomplished through the action of phosphate CheZ (orange circle labeled "Z") which de-phosphorylates CheY and causes it to diffuse back to the complex at the opposite pole from the flagella, and by constitutively active methyltransferase CheR (light blue circle labeled "R"), and methylesterase CheB (orange circle labeled "B"). CheR uses a methyl group from donor S-adenosylmethionine (SAM) to methylate conserved glutamate residues in the cytoplasmic tails of the MCPs. CheB is activated to remove methyl groups from the MCP (represented here by the methanol) by phosphorylation by CheA. Demethylation of receptors can be determined through measuring methanol release from the cells. Levels of methylation of the receptors determine signaling response.

conserved motifs in cytoplasmic tails of MCP sensory receptor proteins [106, 109] (Figure 1.6). CheR is constitutively active, methylating glutamate residues and removing negative charge to form methyl esters [106, 109]. Methylated receptors are more likely to facilitate autophosphorylation of CheA upon binding of a chemoeffector, and thereby activate the forward signaling pathway that results in a change in swimming direction [106]. CheB functions as a methylesterase, removing methyl groups from glutamate residues and thus restoring negative charge in the cytoplasmic tails of the MCP receptor [109]. It has also been shown to remove amine groups from glutamine to create glutamate, providing additional methylation sites for CheR [106]. CheB is phosphorylated, and thus activated, by phosphotransfer from CheA. Upon CheB phosphorylation, methylation levels of receptor cytoplasmic tails decreases, making the signaling system less likely to react to changes in chemoeffector concentration, with the end result being a memory function of sorts, allowing the cell to swim smoothly for a longer period, giving time for the system to return to pre-stimulus state [109].

Many motile bacteria show a wide variety of chemotaxis components, and can encode a number of chemotaxis operons within their genome. Multiple chemotaxis operons are present in almost half of sequenced motile bacteria that have a CheA homolog [111]. Although these operons are orthologous to chemotaxis operons, not all of them are essential for chemotactic behavior, nor are they all expressed at the same time. For instance, *Rhodopseudomonas palustris* encodes 3 complete chemotaxis operons [112], but chemotactic behavior has not been studied in this species. *R. sphaeroides* has 3 partial chemotaxis operons encoding 4 CheA proteins, 6 CheY, 2 CheB, and 9 MCPs. Two of the three operons are essential for chemotactic behavior, with one being more highly expressed in aerobic conditions, and the other in anaerobic conditions [106]. *Pseudomonas aeruginosa*, on the other hand, has 5 chemotaxis operons, with two operons being essential for chemotaxis, while a third operon acts to optimize chemotactic response [113]. A fourth chemotactic-like operon in *P. aeruginosa* plays a role in cell aggregation and colony morphology through mediation of cyclic di-GMP levels [114]. Hybrid molecules, such as CheV or CheC proteins of *B. subtilis* containing both response regulator and docking functions [115], or phosphatase activity, respectively, are common in more complex chemotaxis pathways [116]. Receptor sensing molecules can be either transmembrane or wholly cytoplasmic [111]. Methylation patterns may be variable on cytoplasmic tails of MCPs depending upon the environmental stimulus being sensed, with output response based on the particular methylation pattern [106]. Some CheY-like response regulators control gene expression of genes involved in developmental pathways and have no influence on microbial motion at all [116]. Taken together, the variety of components and functions of chemotaxis signaling pathways can enhance the signaling and response capability of the bacteria with multiple pathways [111].

The genome sequence of *A. brasilense* Sp245 revealed 4 chemotaxis-like operons and 45 MCPs, suggesting the capability of sensing multiple chemoeffectors, and subsequently responding to signals in multiple ways. Characterization of one of these four chemotactic pathways (herein called Che1) indicated that it has a supportive role in chemotaxis, while also contributing to regulation of other cellular responses such as cell length at division and clumping behavior [117, 118]. Mutants were constructed in which either the forward signal transduction pathway (CheY) or the adaptation pathway (CheB and CheR) was removed from the genome. Those mutants lacking a CheY, and thus unable to send a signal in the forward signaling direction, exhibited a phenotype of being considerably shorter than wild type cells and formed clumps under high aeration conditions. In contrast, those mutants lacking both a CheB methylesterase and a CheR methyltransferase were longer, and were impaired in clumping behavior [117]. In order to more fully examine the effects of this chemotaxis pathway on cell function, whole cell proteomics of these mutants were analyzed and results compared to those of wild-type cells. The overall goal of these comparisons, presented in Chapter 4, was elucidation of possible pathways affected by the Che1 operon.

The overall goal of this dissertation is the use of proteomic technology to investigate the response of *Azospirillum* species to both changes in growth conditions, and to manipulation of a single signal transduction pathway. The first section of this dissertation involves optimization of mass spectrometer parameters in order to maximize protein identifications while minimizing the duty cycle of the mass spectrometer. In the second section of this dissertation, the optimized mass spectrometry parameters are used in the investigation of the total proteome expression of two different strains of newly sequenced bacteria, *Azospirillum brasilense*, grown under both nitrogen fixing and non-nitrogen fixing conditions. Further investigations include examination of the effect of mutations introduced in components of a chemotaxis-like operon on the proteome expression of *A. brasilense* species Sp7. Mass spectrometry based proteome analyses comparing bacterial isogenic mutant strains lacking components involved in signal transduction and adaptation pathways were compared, and results are presented in Chapter 4. And finally, a set of vectors with broad host range expression capability that offers the advantage of ligation-free cloning and choice of C-terminal tandem affinity tags was developed, and is presented in Chapter 5. This set of vectors can be used to characterize the function of proteins found to be of interest in nitrogen fixation or to be a consequence of alteration in the Che1 chemotactic pathway.

Chapter 2. Optimization of Tandem Mass Spectrometry Methods

Introduction

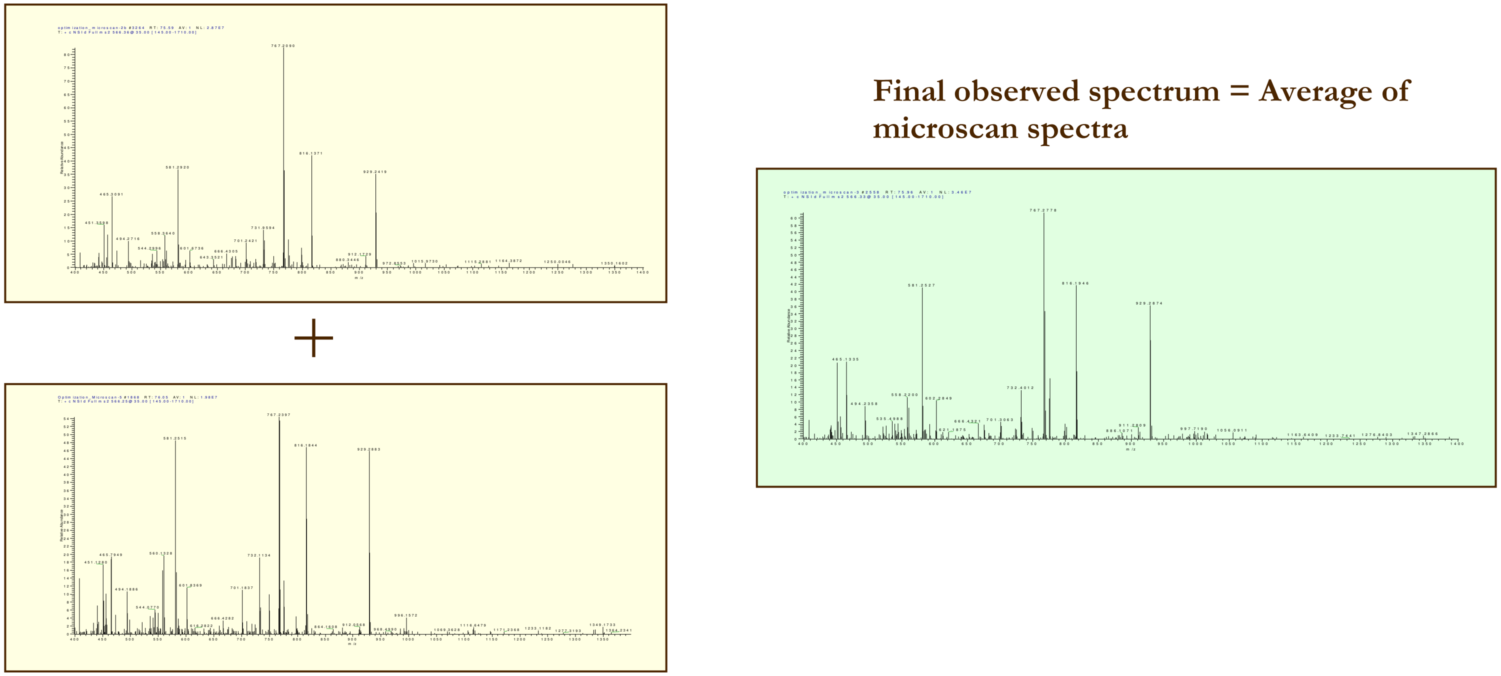
As discussed in Chapter 1, shotgun proteomics has become a very powerful and commonly used methodology for investigating proteome structure, whether a survey of all proteins present within a cell lysate, an investigation of proteins and their interacting partners, or quantitative or semi-quantitative comparisons of concentrations of proteins present under different conditions [4, 40, 119, 120]. Use of a "bottom up" tandem mass spectrometry (MS/MS) approach in which proteins are first digested into peptides and then peptides are detected within a mass spectrometer allows for elucidation of the sequences of peptides, and subsequent identification of proteins present in the entire protein complement of a cell [72]. In tandem mass spectrometry (MS/MS) experiments performed in an ion trapping mass spectrometer, the mass spectrometer isolates a selected ion through application of appropriate voltages, with a specific mass-to-charge (m/z) window being selected for isolation. The isolated parent then undergoes fragmentation through collision with a target gas to create a set of fragment ions that are scanned out of the trap to yield an MS/MS spectrum, as shown in figure 1.2. The spacing of peaks based on the difference in mass of individual side chains of amino acids within the fragment ion spectrum provides information on the amino acid sequence of the parent ion. Single or even multidimensional chromatography in line with the mass spectrometer provides separation of extremely complex mixtures of peptides such that subsets of peptides are eluted and injected directly into the mass analyzer. This on-line separation allows for detection of less abundant peptides in a complex mixture, resulting in acquisition of a greater number of peptides per individual protein in the mixture, and thereby increasing the depth of analysis. Database searching algorithms such as SEQUEST [3] or Mascot [53] can then match this experimental spectrum to theoretical spectra derived from the database sequences, providing a score for the top matches. Other programs such as DTASelect [54] filter and sort this output data to yield a final output of peptide identifications which in turn contribute to individual protein identifications.

Identification of the correct peptide sequence from this raw spectrum depends upon the ability of a database search algorithm such as SEQUEST [3] to compare a sub-optimal experimental spectrum to a theoretical computer-generated (and thus "optimal") spectrum and find the very best match available. This raises the question of what factors contribute to making an "optimal" experimental spectrum. What spectral qualities will contribute to the maximum number of true positive identifications while at the same time ensuring the minimal number of false identifications? Certainly, the score thresholds specified in filtering programs like DTASelect [54] for accepting a positive identification can be changed such that the number of false positives is below an acceptable threshold [62]. However, raising score thresholds comes at the cost of rejecting a small proportion of peptides that are true positives along with those that are falsely deemed to be positive identifications. Additionally, different scoring algorithms, such as MASPIC [121] or other search algorithms such as Mascot [53], DBDigger [7], MyriMatch [6], or InsPect [122] can be used and the results compared to improve the confidence level of positive identifications. Interestingly, earlier studies have indicated that Mascot and SEQUEST perform comparably when used on spectral data acquired on a linear ion trapping quadrupole instrument [119].

The above mentioned options are certainly viable, but the parent ion and fragmentation spectra collected by each mass spectrometer provide the raw data input to these search algorithms. If an instrument is not behaving optimally, manipulation of the search algorithm can only yield minimal gains due to a dearth of good quality spectra. Parameters by which the spectral quality may be evaluated include the ion intensity of the parent ion and the intensity and distribution of the fragment ions resulting from collision-induced dissociation of this parent ion. A number of experimental parameters can contribute to the final quality of the spectra obtained. Sample preparation methods resulting in low concentrations of proteins can result in less than optimal spectra due to the low intensity or absence of fragment ion peaks and can interfere with relative quantitation of sample amounts between two different growth states [15]. Separation methods can either reduce or increase the concentration of peptides available to the mass analyzer, but can result in the presence of co-eluting peptides that may be sampled by the mass analyzer in a given sampling cycle [123]. Instrument parameters in an ion trap mass spectrometer such as the ion injection time, collision energy, scan rate, number of averaged scans (microscans) that make up the raw spectrum, and width of the mass-to-charge (m/z) isolation window can all affect the quality of the spectra obtained. Longer ion injection times can result in more ions entering the trap, which in turn can result in greater ion intensities for the parent ion, perhaps leading to better fragmentation spectra [57, 58]. However, a longer ion injection time can also result in too many ions in the trap causing a decrease in trapping efficiency due to space charge effects [124]. Low scan rates result in greater ion intensities for high m/z ions and low intensities for low m/z ions, while the opposite is true for high scan rates [57]. Collision energy has a variable effect on MS/MS spectra that is dependent on the settings of other parameters [57].

Data acquisition in a tandem mass spectrometry experiment in an ion trap instrument begins with filling the trap with ions being sprayed from the ion source at that point in the chromatographic run. The instrument software ramps voltages so that ions of decreasing m/z values become unstable and are scanned out of the trap to a detector. It then catalogs those ions that are present in the full MS scan. When an instrument is operated in data-dependent mode, a parent ion is chosen for isolation based on user-selected criteria addressing the issues of whether the ion has been scanned recently within a specified amount of time (termed dynamic exclusion), and how much of a window on either side of the peak of the ion is to be isolated. Each parent ion is made up of a series of peaks resulting from isotopologues including one or more “heavy” isotopes of each element. For instance a population of ions will have some percentage that contains one ^{13}C atom with all the rest of the carbon atoms being ^{12}C, thus this small population of ions will be mass shifted by 1 Da. Likewise, some percentage will contain two ^{13}C atoms, shifting the mass by 2 Da, and so on. Larger molecules thus have more possible isotopologues, resulting in a wide distribution of individual isotopologue peaks with an approximate Gaussian distribution, where the isotopologue with the most abundant mass is at the centroid of the peak. The goal of isolating a parent ion for fragmentation includes isolating the entire isotopic packet of that parent ion, accomplished through setting an isolation m/z “window” centered around the central mass peak. Data is typically collected using preset isolation window widths of 2 or 3 m/z units (1 or 1.5 m/z respectively on each side of the peak), which can be set within the software controlling the mass spectrometer. A wider isolation width of 5 m/z units has the potential to yield more total ions, more intensity and anecdotally better MS/MS spectra (Verberkmoes, personal communication). However, co-isolation of other peptides may be a potential negative consequence of the larger mass-to-charge isolation window (McDonald, personal communication). In our lab an increase in identifications has been noted with a larger mass-to-charge isolation window of 5 m/z. Here we test isolation widths of 2, 3, 4 or 5 m/z to ascertain the effect of varying this parameter on the number of peptide and protein identifications.

Once the parent ion is isolated in the trap, fragmentation via collision induced dissociation (CID) yields an experimental fragmentation spectrum that can be matched to predicted spectra derived from a database. A number of studies have been done to investigate the reproducibility of spectra between replicate runs [59, 125]. Mass spectra collected from the same peptide in different replicate runs may have widely variant intensities for each peak collected, and may vary widely in the presence of low abundance fragment spectral peaks [59]. The differences in spectral quality due to variable ion intensities can be compensated for by collecting a number of spectra of the same ion (microscan), and averaging all of these spectra together to produce a final single raw output spectrum. When acquiring data, the mass spectrometer takes one or more individual “microscans” of repeated fills of the trap. The spectra obtained from each microscan are then averaged to create the final displayed MS scan as illustrated in Figure 2.1. More microscans result in better representation of the ions present in the trap, which should ultimately result in a better probability of obtaining higher quality spectra, with



Final observed spectrum = Average of microscan spectra

Figure 2. 1 Representation of spectral averaging (number of microscans) to obtain a final observed spectrum
The ion trap is filled and a parent ion is isolated and subjected to collision induced dissociation to create a fragment ion spectrum which is recorded by instrument software, but not displayed. This process is repeated a number of user-specified times and the peak intensities and positions of the individual spectra are averaged to create a final fragment ion spectrum that is displayed and recorded. Increasing the number of microscans can result in a cleaner spectrum with greater signal to noise ratio and a greater number of fragment ions, but the trade-off is the amount of time required to do multiple scans.

better ion intensities and a greater number of fragment ion peaks. Averaging a number of microscans normalizes the intensities of each ion present in the trap at a given time for both parent ion (MS) scans and for resulting fragment ion (MS/MS) scans [60], making the experimental spectrum more closely resemble the theoretical spectrum used by database search algorithms. A more populated fragmentation spectrum resulting from an increased number of microscans can potentially lead to a more confident identification as the experimental spectrum approaches a more optimal appearance.

However, earlier studies using different numbers of microscans have led to conflicting results. When separating peptides from a single protein bovine serum albumin (BSA) digest and using an ion trapping quadrupole mass spectrometer (LCQ, Thermo-Finnigan) for mass analysis, Wenner et al [60] determined that a window of optimal microscan numbers existed, with either 3, 5, 7 or 9 microscans resulting in a higher number of peptide identifications than those obtained using either higher or lower numbers of microscans. Experiments done by Venable et al [59] also indicated that increasing the number of microscans when acquiring tandem mass spectrometry data from protein mixtures on an LCQ leads to less total identifications of peptides, but yields a better separation of true positive identifications from false positive identifications. However, subsequent studies by Moberg and colleagues indicate that increasing the number of microscans has no significant effect on the identifications made when collecting tandem mass spectrometry data on a linear ion trapping quadrupole (LTQ) instrument [57].

Instrument software allows a choice of the number of microscans performed to make up the final spectrum. Previous data from the group indicates that performing five microscans for each tandem mass spectrum on the LCQ results in better data quality than performing three microscans or one microscan (data not shown). Data quality output from increasing microscan numbers in the LTQ have not been investigated in our group. As stated earlier, an increased number of microscans should yield better ion statistics and thus should improve both sensitivity and specificity of data-dependent LC/MS-MS analysis of peptides. This increased sensitivity and specificity should in turn decrease the number of false negatives and false positives returned by our data analysis pipeline. However, performing an increased number of microscans requires an increased amount of time, and thus results in collection of fewer total spectra with the attendant possibility of missing some peptides. Based on previous data from the group, we compare data quality from two, three or five tandem mass spectrometry microscans on the LCQ ion trapping quadrupole instrument. The LTQ linear two dimensional quadrupole ion trap mass spectrometer has larger ion storage capacity and higher trapping efficiency, trapping 20-40 times as many ions as does the LCQ [48]. A higher number of ions in the trap should provide better ion statistics, and thus better data quality [48]. Here, for the LTQ, we compare the data quality obtained from one microscan versus two or three microscans.

The overall goal of experiments in this chapter is to optimize the mass spectrometry operation parameters of isolation width m/z window and number of microscans employed to obtain an MS/MS spectrum in order to maximize the number of true positive identifications. A systematic examination of results obtained when tandem mass spectrometer parameters are varied will lead to a more optimized operation of the mass spectrometer, which is essential for maximizing the number of identifications made in any given experiment. Subsequent use of these parameters in proteomics experiments performed in chapters 3 and 4 of this work will allow for more accurate determination of the protein components present in a given proteomic experiment.

Materials and Methods

Chemicals

All chemicals were obtained from Sigma Chemical Company (St. Louis, Mo) unless otherwise stated. Acetonitrile, and HPLC-grade water were purchased from Burdick and Jackson (Muskegon, MI), 98% formic acid (FA) from EM Science (an affiliate of MERCK KgaA, Darmstadt, Germany), and sequencing-grade trypsin from Promega (Madison, WI).

Preparation of extended protein standard mix (EPSM)

The extended protein standard mixture was prepared as described earlier [126]. Briefly, twenty proteins listed in Table 2.1 were mixed in equimolar amounts and suspended in 6 M guanidine hydrochloride with 10 mM DTT. This stock solution was frozen at -80°C for later digestion with trypsin as related below. Once digested, the EPSM digest was concentrated by vacuum centrifugation to a final concentration of 1 mg/ml, aliquoted into individual tubes, and stored at -80°C for later analysis.

Table 2. 1 Proteins included in the Extended Protein Standard Mix (EPSM)

Protein	**Organism**
Carbonic Anhydrase II	*Bos taurus*
Conalbumin (ovotransferrin)	*Gallus gallus*
Concanavalin A	*Canavalia ensiformis*
Cytochrome c	*Bos taurus*
Deoxyribonuclease I	*Bos taurus*
Lysozyme c	*Gallus gallus*
Beta-lactoglobulin B	*Bos taurus*
Ribonuclease A	*Bos taurus*
Thyroglobulin	*Bos taurus*
Serum albumin	*Homo sapiens*
Serum albumin	*Bos taurus*
Alcohol dehydrogenase E	*Equus caballus liver*
Alcohol dehydrogenase I	*Saccharomyces cerevisiae*
Alpha-amylase	*Bacillus subtilis*
Beta-amylase	*Ipomoea batatas*
Apomyoglobin	*Equus caballus*
Hemoglobin A	*Equus caballus*
Hemoglobin B	*Equus caballus*
(Apo)-transferrin	*Bos taurus*

Preparation of *R. palustris* standard proteome

Cell Growth

R. palustris was grown photoheterotrophically as described earlier [30] under anaerobic conditions in 1.5 L of defined mineral medium with 12.5 mM sodium phosphate, 12.5 mM potassium phosphate,10 mM succinate, 0.1% ammonium sulfate and 2 mg/L p-aminobenzoic acid. Nitrogen gas replaced the air in the head space, and cultures were grown at 30°C with a light source to mid-log phase (OD_{660nm} ~ 5.0) while mixing the culture with a stir bar. Cells were harvested by centrifugation at 5000 xg in an S-17 rotor in a Sorvall centrifuge for 20 minutes at 4°C, washed twice with cold 50 mM Tris buffer, and pellets were frozen at -80°C for later lysis.

Cell lysis

Cell pellets were thawed and resuspended in ice-cold buffer (50 mM Tris, 10mM $CaCl_2$, pH 7.4). After resuspension, cells were lysed by sonication (Misonix Sonicator 4000, Newtown, CT, 4 x 30 sec, 40% power, with 30 second cooling periods) and cell debris cleared by centrifugation at 18,000 xg for 10 minutes. Total protein concentration was then determined through bicinchoninic acid (BCA) assay (Pierce Protein Research, a division of Thermo Scientific) following manufacturer's protocols. Supernatant was then aliquoted and frozen at -80°C.

Trypsin Digestion

Trypsin digestion was done following protocols described earlier for *S. oneidensis* with minor modifications [126]. Briefly, one proteome aliquot (approximately 6 mg of protein) was denatured by addition of 6M Guanidine to the lysate with subsequent incubation at 60°C for 1 hour. Guanidine was then diluted 6-fold through addition of 5 volumes of 50 mM Tris/10 mM $CaCl_2$ pH 7.4, and 20 µg sequencing-grade trypsin was then added for every mg of protein. Digestion was done for 18 hours at 37°C with rotation. More sequencing-grade trypsin was added (20 µg/mg protein) and digestion continued at 37°C with rotation for an additional 5 hours. A final reduction step was done through addition of 20 mM DTT followed by incubation for 1 hour at 60°C. Samples were then desalted using a Sep-Pak plus C-18 solid-phase extraction column (Waters, Milford, MA), and extracted with acetonitrile/0.1% FA. Samples were concentrated through centrifugal evaporation (Savant Instruments, Holbrook, NY), and solvent-exchanged to water with 0.1% FA. Samples were then filtered, aliquoted and frozen at -80°C until MS analysis.

LC/LC-ES-MS/MS Analysis:

EPSM Chromatography

Six technical replicates of EPSM were analyzed. Stock solution of the digested EPSM at a concentration of 1 mg/ml was diluted 1:10 in Switchos Buffer A (100% HPLC-grade water, 0.1 % formic acid (FA)) for a final concentration of 0.1 mg/ml. Tips were pulled using a P-2000 laser puller (Sutter Instruments, Novato, CA) from 100µ ID fused silica to a 5µ tip opening, and packed via pressure cell (New Objective, Woburn, MA) with 15cm Aqua C-18 5µm reverse phase (RP) packing material (Phenomenex, Torrance, CA) to form the resolving column. Fifty microliters of the 0.1 mg/ml solution was injected via FAMOS autosampler onto a reverse-phase trapping column. Peptides were then eluted from the trapping column to the resolving column, which was placed directly in-line with either an LCQ-DecaXP quadrupole ion trap mass spectrometer (Thermo- Fisher Scientific, Waltham, MA) or an LTQ linear ion trap mass spectrometer (Thermo-Fisher Scientific, Waltham, MA). Peptides were eluted from the resolving column by gradient elution from 100% Buffer A (100% HPLC-grade water, 0.1 % FA) to 100% Buffer B (70% acetonitrile, 0.1% FA) over a time period of 120 minutes. Experiments were performed in random order, with blank chromatography runs between each experimental condition in order to minimize carry-over between runs.

Proteome Chromatography

Proteome samples were analyzed in triplicate via 5-step multidimensional protein identification technique (MudPIT) [17, 19] performed with an Ultimate HPLC (LC Packings, a division of Dionex, San Fransisco, CA) coupled to either an LCQ-Deca quadrupole ion trap mass spectrometer (Thermo- Fisher Scientific, Waltham, MA) or an LTQ linear ion trap mass spectrometer (Thermo-Fisher Scientific, Waltham, MA). A triphasic split column was used for separation of peptides. The front resolving column was constructed as described for the EPSM above and was placed in-line with the mass spectrometer. Back columns consisted of 150µ ID fused silica, and were loaded first with 3 cm of Luna 5µm strong cation exchange (SCX) resin (Phenomenex, Torrance, CA) followed by 3cm of Aqua 5 µm C-18 reverse phase resin (Phenomenex). Proteome aliquots, each 10µL of approximately 1 mg/ml concentration, were loaded directly onto the back column via pressure cell. The back column was then coupled to the front resolving column.

Five individual chromatography steps were set up using Xcaliber software. Initial desalting and elution of peptides to the strong cation exchange resin was done with a 120 minute gradient elution of 100% buffer A (95% water, 5% acetonitrile, 0.1% FA) to 100% buffer B (70% acetonitrile/ 0.1% FA). Three subsequent MudPIT [17, 20] chromatography steps included 2 minute salt pulses of 10, 20, and 40% buffer C (500 mM ammonium acetate, 0.1% FA) to elute subsets of peptides from SCX resin to the resolving column. These subsets of eluted peptides were then resolved by 120 minute gradient elution from 100% buffer A to 100% Buffer B. The final MudPIT chromatography step consisted of a 20 minute salt pulse of 100% buffer C followed by a 120 minute gradient from 100% buffer A to 100% buffer B. Columns were washed with 10 minutes of 100% buffer B followed by 50% buffer B, then re-equilibrated to 100% buffer A before each run.

Mass spectrometry

Data acquisition was under the control of Xcaliber software (Thermo Fisher Scientific), with data being collected in data-dependent mode. Dynamic exclusion was enabled, with repeat count of 1, repeat duration of 0.5 minutes and exclusion duration of 1.0 minute. Mass spectrometry parameters that were held constant in all experiments are listed in Table 2.2. For the LCQ, a full MS scan of ions with mass-to-charge ratios of 400 to 2000 m/z was performed, followed by 3 data-dependent tandem mass spectrometry (MS/MS) scans, while one full MS scan followed by 6 data-dependent MS/MS scans were performed for the LTQ. Tandem mass spectrometry (MS/MS) parameters were varied according to the 6 experimental conditions listed in Table 2.3.

Table 2. 2 Mass spectrometry operation parameters held constant

	LCQ parameters	LTQ parameters
Normalized Collision energy	35.0%	35.0%
No. scan events	4 – 1 full, 3 data dependent	7 – 1 full, 6 data dependent
Activation Q	0.250	0.250
Activation time	30.00	30.00
No. parental microscans	2	1
Dynamic Exclusion	Enabled	Enabled
Repeat count	1	2
Repeat duration	0.50 min	0.30 min
Exclusion list size	50	100
Exclusion duration	60 sec	180 sec

Table 2. 3 Variable tandem mass spectrometer experimental parameters

LCQ variable parameters			LTQ variable parameters		
Experiment	Isolation width	No. of Microscans	Experiment	Isolation width	No. of Microscans
1	2 m/z	3	1	2 m/z	2
2*	3 m/z	3	2*	3 m/z	2
3	4 m/z	3	3	4 m/z	2
4	5 m/z	3	4	5 m/z	2
5	3 m/z	2	5	3 m/z	1
6	3 m/z	5	6	3 m/z	3

Based on results from experiments done on extended protein mixture, only experimental conditions 2, 4, and 5 from Table 2.3 were performed on proteome mixtures. MS/MS parameters were varied randomly for each run in order to ensure statistical validity to data collected and to minimize effects of instrument variability.

Data Analysis

Data from 6 independent EPSM replicates and 3 proteome replicates for each experimental condition was pooled and searches were performed on both pooled data and individual independent experimental runs using both SEQUEST [3] and DBDigger [7]. EPSM data was searched against two different databases, one containing the protein sequences for the 20 proteins included in the extended protein standard mix (EPSM) plus common contaminants with the *R. palustris* proteome database as a distractor. The second database contained reversed sequences of the EPSM plus common contaminants appended to the forward sequences. Proteome data was searched against a concatenated forward and reverse *R. palustris* database including 36 contaminants (http://compbio.ornl.gov/rpal_proteome/databases). Data was then filtered and sorted using DTASelect [54] with the following parameters: Xcorr thresholds of 1.8 for ions with a +1 charge state, 2.5 for +2 ions, 3.5 for +3 ions, deltaCN of 0.08, requiring 2 peptides for positive identification of a protein. Using results from the reverse database searches, threshold levels were then determined which would yield a 95% confidence level for each experimental condition, and data was re-filtered.

Statistics were computed on both the percentage of identifications (number of true identifications/ total spectra collected) and the number of true identifications using individual data files. Means for each individual experimental category (6 replicates in each category) for both percentage of total peptide identifications and number of true identifications were calculated, and compared using one-way ANOVA tests. Pairwise comparisons of mean numbers of true identifications for the EPSM data were performed with Tukey's F tests to determine which means were significantly different from the others. All statistical analysis was performed using SAS data analysis software, with p-values less than 0.05 being considered significant.

ROC (Receiver Operating Characteristics) curves were generated from the EPSM data using R statistical package. Data for each set of experiments was pooled and files of true positive identifications and true negative identifications were created from the DTASelect.txt output files for each experiment using in-house software. Receiver Operating Characteristc (ROC) curves were created by plotting the number of true negative identifications versus the number of true positive identifications for peptides identified. Individual plots were done for those peptides identified as having charge states of plus one, plus two and plus three. Plots were not normalized, but were instead used for reporting the maximum number of known true identifications for each data set. Isolation window widths of two, three, four, and five m/z were compared in each ROC curve. Microscan ROC curves compared one, two, and three microscans for the LTQ data, and two, three and five microscans for the LCQ.

Results

Initial qualitative analysis, done by comparing the highest scoring spectra among the isolation width experiments and among the microscan experiments indicated no discernible visual difference in spectral quality among MS/MS spectra acquired using different isolation width windows or those acquired with different microscan numbers for either mass spectrometer tested. Representative spectra from the peptide sequence DLILQGDATTGTDGNLELTR derived from protein conconavalinA collected using different isolation widths and different numbers of MS/MS microscans are shown in Figures 2.2 and 2.3, respectively. Examination of these spectra shows similar intensity levels for most abundant peaks, similar signal-to-noise ratios and a similar population of fragment ion peaks, indicating that changing isolation width windows or number of microscans does not have a significant effect on the appearance of the individual spectrum.

Effect of varying data acquisition isolation width

LCQ quadrupole 3D ion trapping mass spectrometer results

Receiver Operating Characteristic (ROC) curves, discussed more thoroughly below, were created by establishing a table containing the number true positive identifications and the number of false positive identifications for a given threshold level based on those proteins known to be in the EPSM mixture. Curves were created by plotting the number of false positive identifications on the x-axis versus the number of true identifications on the y-axis. Each point on the curve corresponds to a plot of these numbers at a given threshold level. Examination of ROC curves for data collected on the LCQ mass spectrometer suggests an isolation window width of 2 m/z units is best for those peptides having a +2 or +3 charge state, while for peptides having a +1 charge state

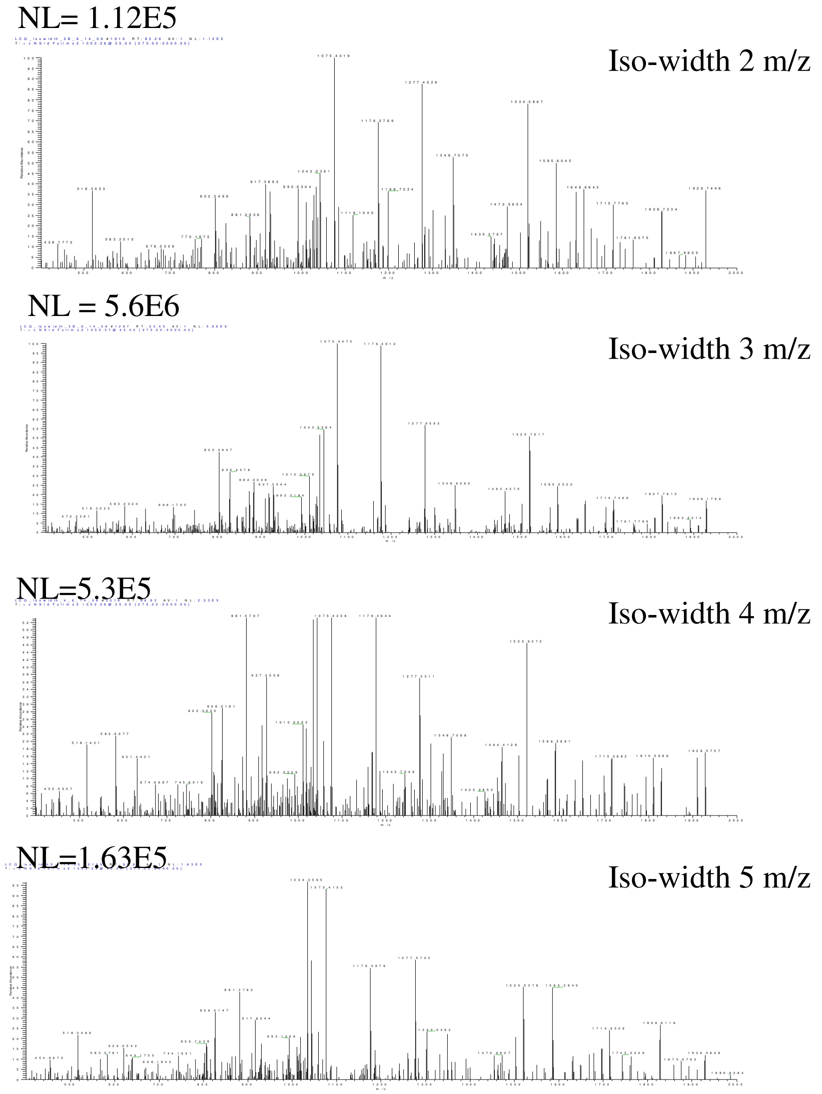


Figure 2. 2 Representative fragmentation spectra from peptide sequence DLILQGDATTGTNLELTR derived from protein conconavalinA collected using different isolation widths
NL = normalization level, Iso-width = isolation width. Examination of the fragmentation spectra collected at each of the above isolation widths on the LCQ shows a well-populated spectrum with similar intensity levels and signal-to-noise ratios for each of the experimental spectrum.

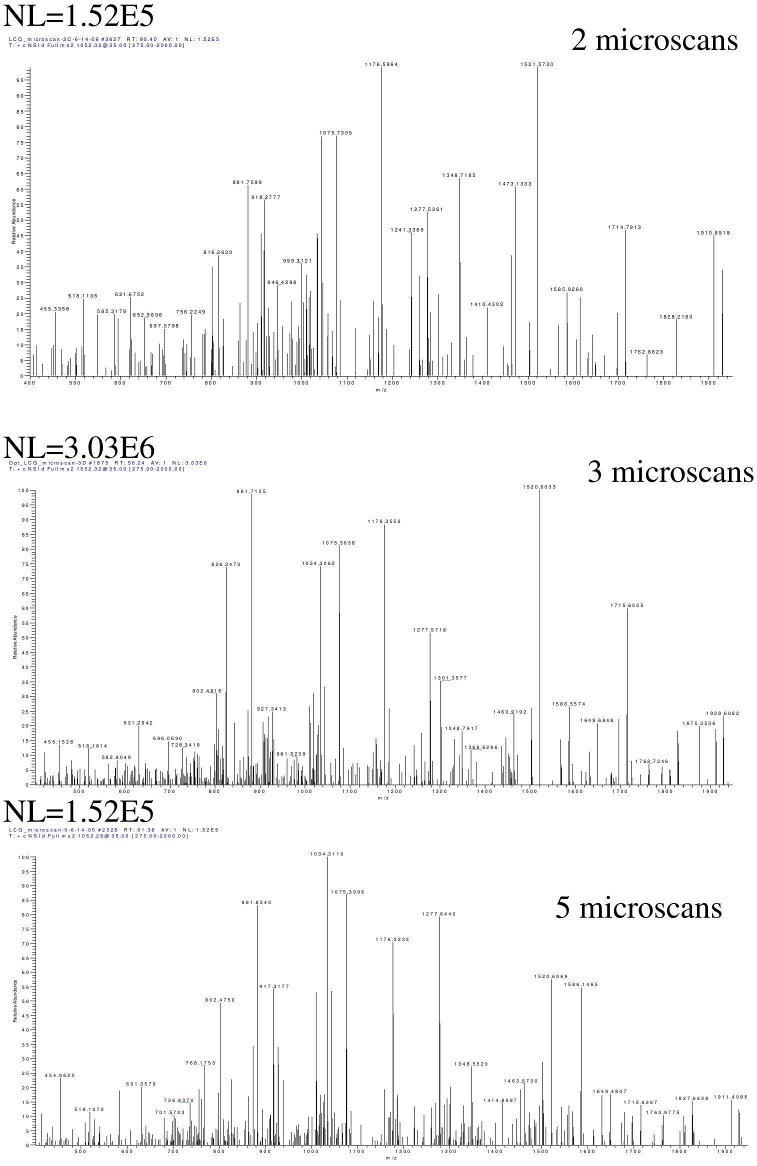
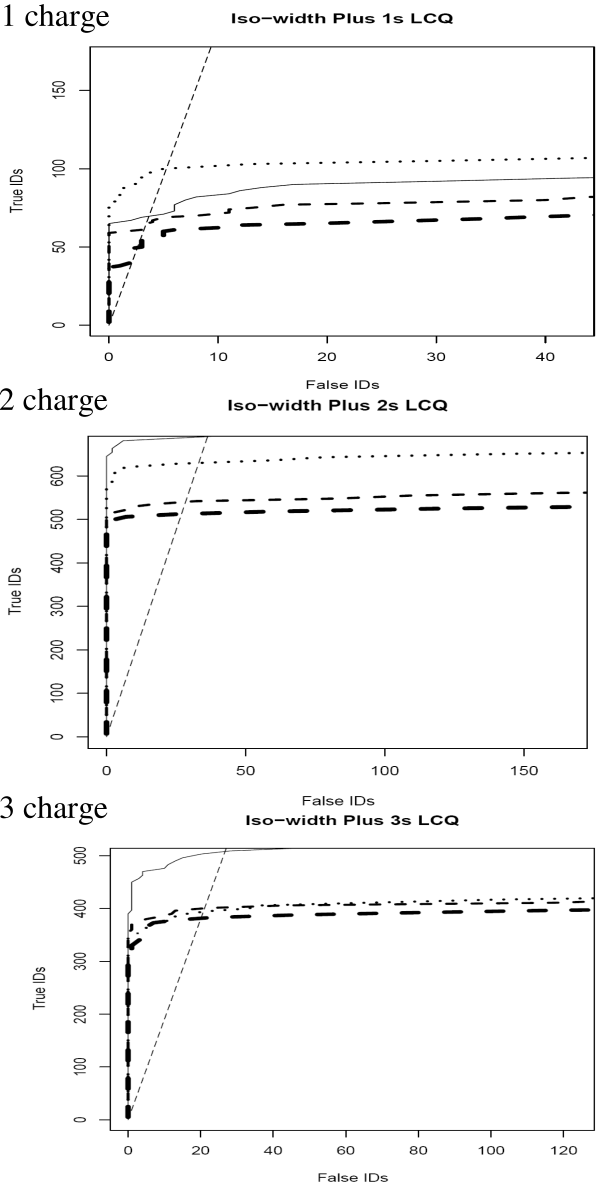


Figure 2. 3 Representative fragmentation spectra from peptide sequence derived from protein concanavalin A collected using different microscans

NL = normalization level. Each individual spectrum shown above is well-populated with fragment ions, and each have similar intensity levels as noted by height of the peaks with similar normalization levels. Additionally signal strength of fragmentation peaks is well above noise levels.


A) Peptides with +1 charge

B) Peptides with +2 charge

C) Peptides with +3 charge

Figure 2. 4 ROC curves of LCQ isolation width EPSM data
Data from 6 replicates of each EPSM isolation width window experiment was pooled and the number of false identifications (x-axis) versus the number of true identifications (y-axis) was plotted for peptides detected in each charge state. A) Data from peptides of +1 charge state, B) Data from peptides of +2 charge state, and C) Data from peptides of +3 charge state. The solid line represents an isolation window of 2 m/z units, dotted line represents 3 m/z units, dashed line is 4 m/z units, and heavy dashed line is 5 m/z units. The angled line represents a 95% confidence threshold for each plot. The point at which the straight dashed line crosses the curve represents threshold levels at which identifications are made with 95% confidence. From the above plots, it appears that an isolation window width of 2 m/z is best for + 2 and +3 peptides, while an isolation width of 3 m/z units is best for +1 peptides. Isolation widths of 4 or 5 m/z units do not appear to be significantly different for any charge state.

an isolation window of 3 m/z units gives a greater number of true peptide identifications for the same number of false peptide identifications (Figure 2.4). An isolation width window of 5 m/z units appears to give the lowest number of true positive identifications for all charge states. Taken together the ROC curves suggest that an isolation window width of 2 or 3 m/z units is best for data collection on the LCQ instrument, while an isolation width of 5 m/z units can possibly be detrimental to results obtained on the LCQ.

However, statistical analysis of average peptide identifications from 6 individual replicates of isolation width m/z window data filtered using threshold values that returned a 95% confidence level did not support the suggested optimal isolation window widths derived from the ROC curves. One-way ANOVA analysis returned no significant difference between the average total number of identifications obtained (p = 0.9916). Percentages of peptides identified were computed by dividing the number of identifications made by the total number of spectra collected, giving a measure of the efficiency of the search algorithm in assigning sequence identification to the raw spectra. The average percentage of peptides identified per total spectra collected also showed no significant difference (p = 0.9054).

Since the EPSM samples are not entirely reflective of real-world total proteome samples, the possibility remains that significant differences can be discovered in identification of spectra collected from total proteome samples. For this reason, we elected to test isolation window widths of 2, 3, and 5 m/z using a total proteome sample. An isolation window width of 5 m/z yields a greater number of peptide and protein identifications (Table 2.4), but not significantly more than an isolation window width of

Table 2. 4 numbers of protein and peptide identifications from total proteome spectra collected using differing isolation widths

LCQ

	Protein Identifications		Peptide identifications	
Isolation window width	Average	Standard Deviation	Average	Standard Deviation
2 m/z	336	47	1613	326
3 m/z	401	33	2089	186
5 m/z	502	39	2927	281

LTQ

	Protein Identifications		Peptide identifications	
Isolation window width	Average	Standard Deviation	Average	Standard Deviation
2 m/z	1084	14	8020	352
3 m/z	1239	85	9684	939
5 m/z	1176	36	8446	756

3 m/z. In comparison, an isolation window width of 2 m/z yields a significantly lower number of peptide and protein identifications than windows of either 3 m/z or 5 m/z (p = 0.0433 for proteins, p = 0.0213 for peptides). Taken together, the above data suggest that an isolation window width of 3 m/z or 5 m/z will yield a higher number of identifications than an isolation window width of 2 m/z for an LCQ mass spectrometer.

LTQ linear ion trapping quadrupole mass spectrometer

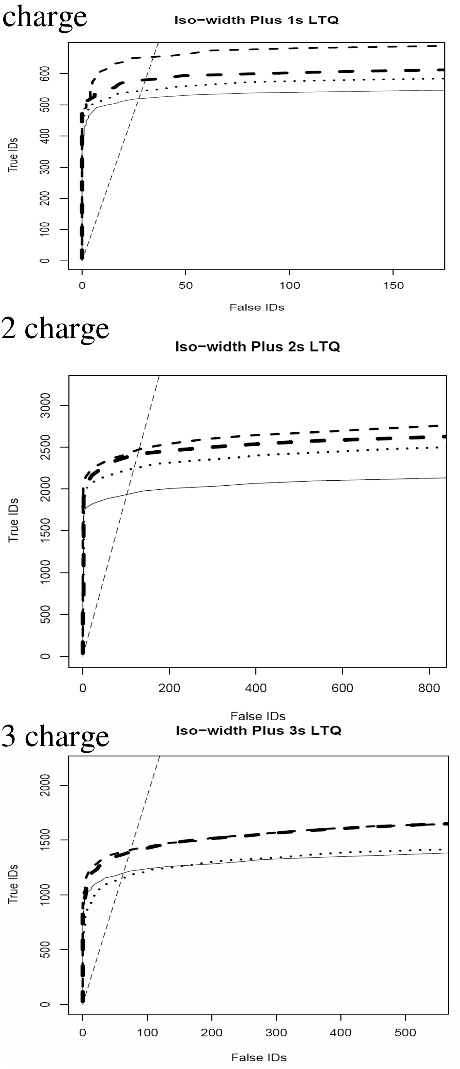
In contrast to the LCQ data results, examination of the ROC curves for the pooled EPSM LTQ isolation width data (Figure 2.5) suggests that an isolation window width of 2 m/z yields a lower number of positive peptide identifications for peptides of all charge states. However, isolation window widths of 3, 4 or 5 m/z units appear to be equivalent in the number of positive peptide identifications. As visualized in the ROC curves, no significant differences emerged in mean percentage identifications (p = 0.6432)

or the mean number of true identifications (p=0.2436) in the EPSM isolation width studies for the LTQ. When further studies were done on proteome samples in the LTQ, no significant differences emerge for any isolation width tested (Table 2.4) for either protein identifications (p = 0.1103) or peptide identifications (p = 0.1608).

Effect of varying tandem mass spectrometry (MS/MS) microscans

LCQ quadrupole 3D ion trapping mass spectrometer results

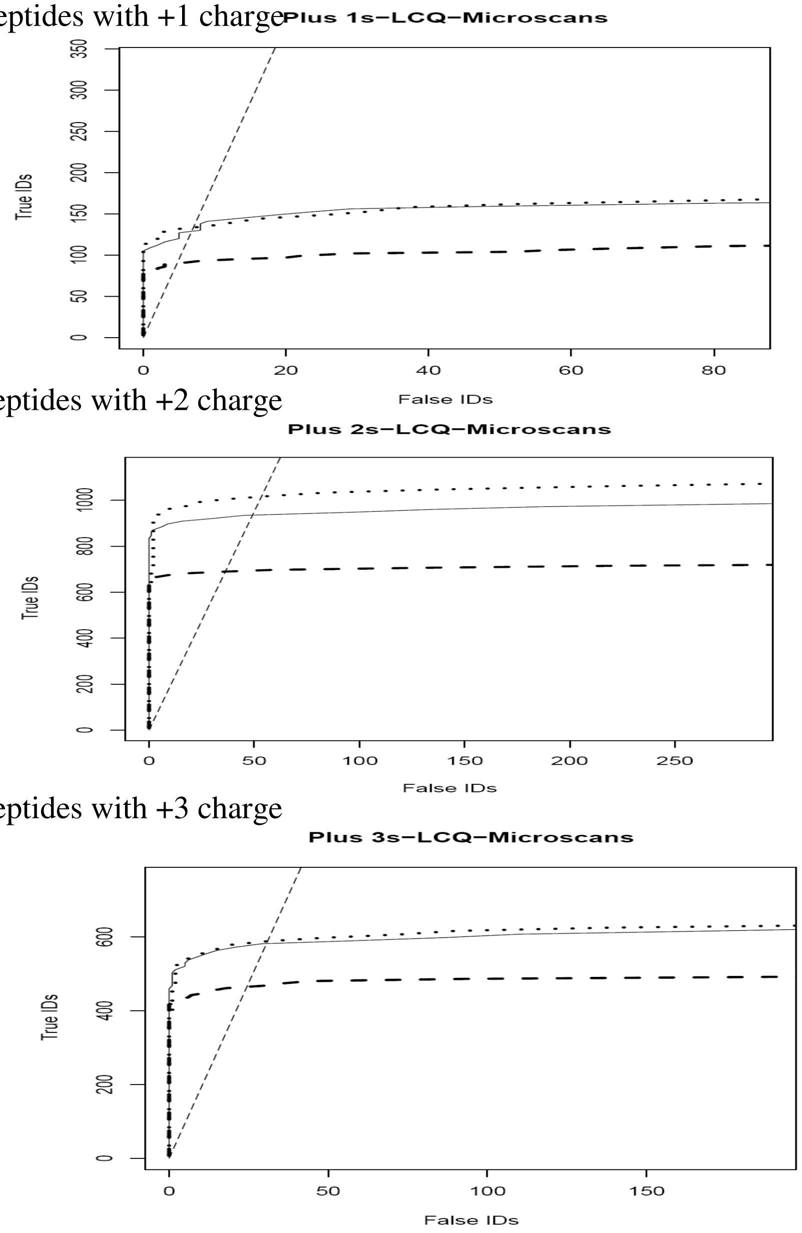
For the LCQ, ROC curves indicate that the use of two or three microscans does not seem to yield a significantly different number of true positive peptide identifications (Figure 2.6). However, a higher number of microscans appears to negatively impact the


A) Peptides with +1 charge

B) Peptides with +2 charge

C) Peptides with +3 charge

Figure 2. 5 ROC plots of LTQ isolation width EPSM data
Data from 6 replicates of each LTQ EPSM isolation width window experiment was pooled and the number of false identifications (x-axis) versus the number of true identifications (y-axis) was plotted for peptides detected in each charge state. A) Data from peptides of +1 charge state, B) Data from peptides of +2 charge state, and C) Data from peptides of +3 charge state. The solid line represents an isolation window of 2 m/z units, dotted line represents 3 m/z units, dashed line is 4 m/z units, and heavy dashed line is 5 m/z units. The angled line represents a 95% confidence threshold for each plot. The point at which the straight dashed line crosses the curve represents threshold levels at which identifications are made with 95% confidence. From the above plots, it appears that an isolation window width of 2 m/z negatively impacts the total number of identifications that can be made, regardless of charge state, while an isolation width of 3, 4, or 5 m/z units do not appear to be significantly different for any charge state.



A) Peptides with +1 charge

B) Peptides with +2 charge

C) Peptides with +3 charge

Figure 2. 6 ROC curves of pooled LCQ EPSM microscan data
Data from 6 replicates of each EPSM microscan experiment was pooled and the number of false identifications (x-axis) versus the number of true identifications (y-axis) was plotted for peptides detected in each charge state. A) Data from peptides of +1 charge state, B) Data from peptides of +2 charge state, and C) Data from peptides of +3 charge state. The solid line represents 2 MS/MS microscans, dotted line represents 3 MS/MS microscans and the dashed line represents 5 MS/MS microscans. The angled line represents a 95% confidence threshold for each plot. From the above plots, it appears that 5 MS/MS microscans has a negative effect on the number of identifications that can be made, while 2 or 3 microscans give the maximal number of identifications.

number of peptides identifications obtained from the protein standard mix. The statistical analysis of the microscan data, however, tells a different story. A significantly lower mean number of average percentage identifications was found for 2 microscans (p = 0.0078) while there was no significant difference between mean number of identifications of 3 and 5 microscans. The average number of true identifications also shows a significant difference between the mean numbers of identifications obtained using different microscans, with 5 microscans giving significantly lower number of IDs than either 2 or 3 microscans. Taken together this data suggests that 2 MS/MS microscans negatively impacts the ability of the search algorithm to identify peptides from raw spectra. On the other hand, 5 MS/MS microscans does not affect the ability of the search algorithm to identify peptides from raw spectra, but instead results in a lower number of identifications due to the increased amount of time required to collect the spectra. From this data, 3 MS/MS microscans appear to be the optimal number of microscans for the LCQ.

Further testing was done using 3 and 5 MS/MS microscans for the proteome data in order to determine if the increased amount of time to collect the scans would impact the total number of protein identifications made. Average peptide and protein counts are presented in Table 2.5, and box plots of LCQ proteome data in Figure 2.7. Although there is a slight decrease in the number of protein and peptide identifications with 5 MS/MS microscans versus 3 MS/MS microscans, this difference is not statistically significant for peptides (p = 0.5711) or for protein identifications (p = 0.2046).

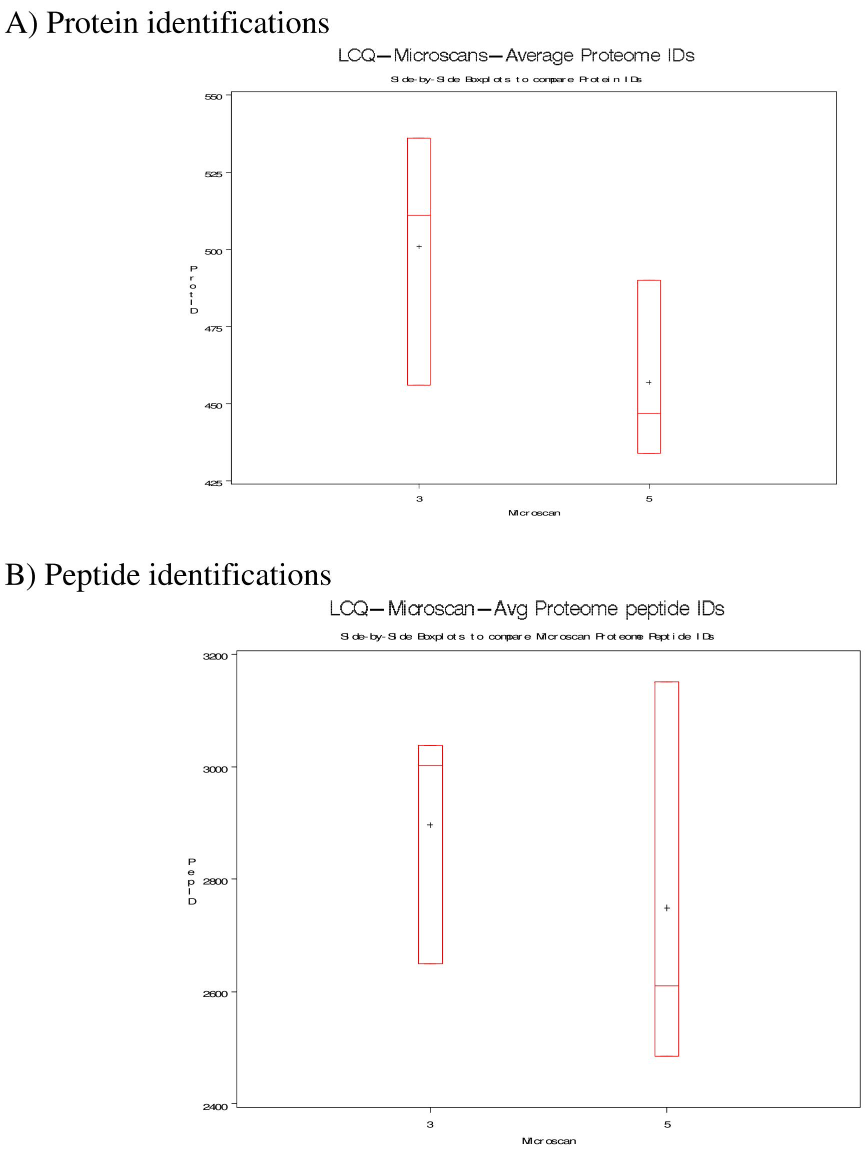
Table 2. 5 Average numbers of protein and peptide identifications from total proteome spectra collected using differing MS/MS microscan settings

LCQ

	Protein Identifications		Peptide identifications	
No. of MS/MS Microscans	Average	Standard Deviation	Average	Standard Deviation
3	404	33	2089	186
5	370	16	2042	263

LTQ

	Protein Identifications		Peptide identifications	
No. of MS/MS Microscans	Average	Standard Deviation	Average	Standard Deviation
1	1237	78	9638	774
2	1239	85	9684	939



A) Protein identifications

B) Peptide identifications

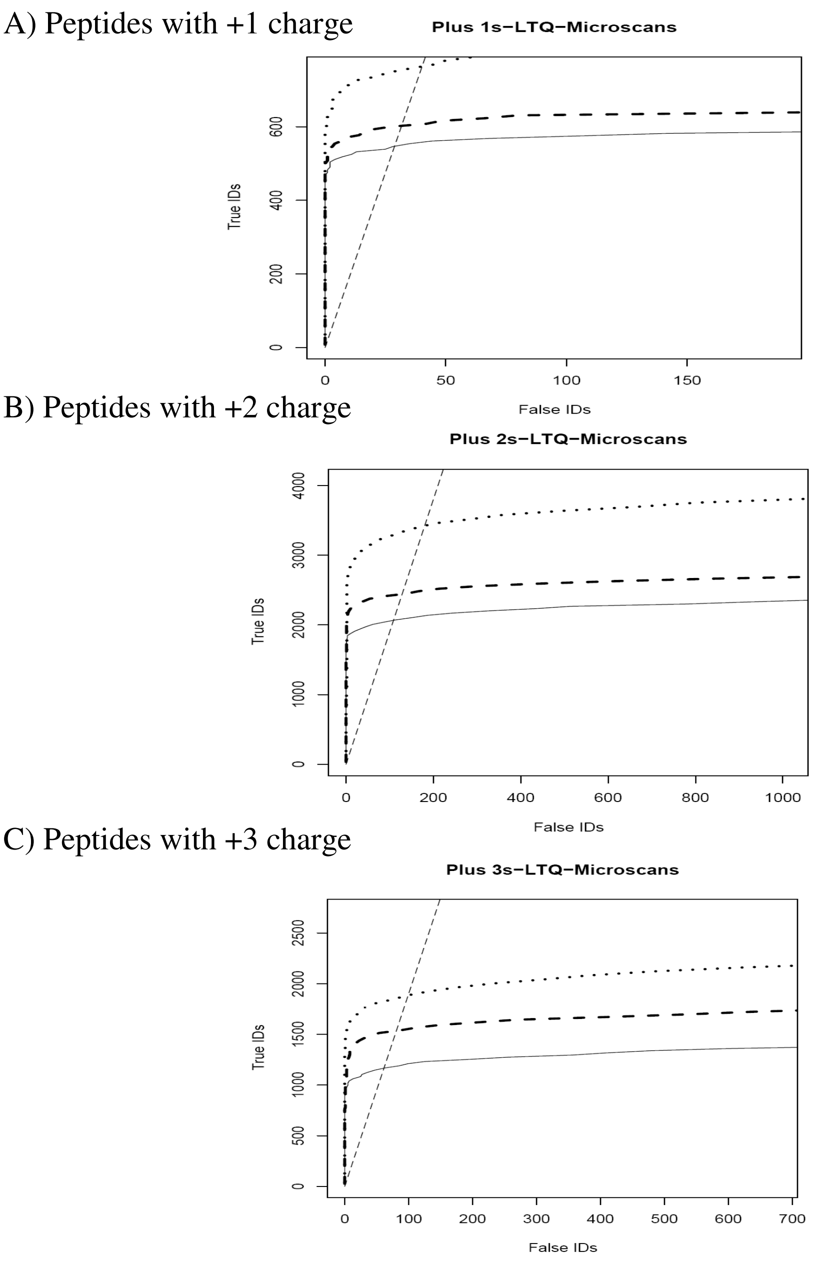
Figure 2. 7 Box plot representation of statistical results from pooled LCQ proteome microscan data Data from 6 individual replicates was analyzed via one-way ANOVA and means and standard deviation determined. The x-axis represents the number of MS/MS microscans (either 3 or 5) tested on LCQ proteome samples. The box encloses two standard deviations within the data range, while the star in the center represents the mean, and the line represents the median. A) Mean numbers and standard deviation of total numbers of protein identifications. B) Mean numbers and standard deviation of total numbers of peptide identifications. Statistical analysis showed that there is no significant difference between the groups in the numbers of protein (p = 0.2046) or peptide (p = 0.5711) identifications obtained when proteome samples are tested.

LTQ linear ion trapping quadrupole mass spectrometer

The LTQ microscan data showed similar results to those for the LCQ. Examination of ROC curves derived from EPSM samples run on the LTQ (Figure 2.8) shows that 2 microscans gives a noticeably higher number of true positive peptide identifications for peptides in all charge states, although the difference is less pronounced for peptides with a charge state of +3. Again, the number of microscans showed significant differences in average percentage identifications ($p < 0.0001$). But here, all three average percentage IDS for each microscan are significantly different from one another. There are also significant differences in the numbers of peptide identifications. The average number of identifications from 2 microscans overlaps with both 1 and 3 microscans. There is no statistically significant difference between the number of identifications found in data collected using 2 microscans and that found using 3 microscans. Nor is there a significant difference between the numbers of identifications found using 2 microscans versus 1 microscan. However, 3 microscans give significantly lower numbers of identifications than 1 microscan ($p = 0.0024$).

Due to the lower number of peptide identifications obtained with 3 MS/MS microscans, this experimental condition was not included in proteome studies.

Proteome data, given in Table 2.5, shows very little difference between average numbers of proteins and peptides identified from data collected with 1 MS/MS microscan and that collected with 2 MS/MS microscans. For a more complex soluble proteome sample, no significant difference in mean number of protein ($p = 0.9061$) or peptide ($p=0.9858$) identifications are found between 1 and 2 microscans (Figure 2.9), suggesting that

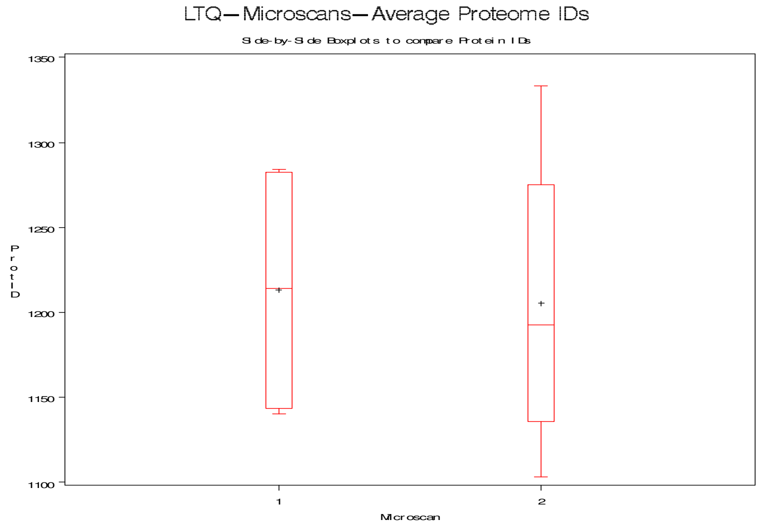


A) Peptides with +1 charge

B) Peptides with +2 charge

C) Peptides with +3 charge

Figure 2. 8 ROC curves representing LTQ pooled EPSM microscan data
A) Number of EPSM peptides identified of +1 charge state, B) Number of +2 charge state EPSM peptides identified, and C) Number of +3 charge state EPSM peptides identified. Solid line represents data collected using 1 MS/MS microscan, dotted line represents data collected with 2 MS/MS microscans, and dashed line with 3 MS/MS microscans. From the above plots, 2 microscans appears to be the clear winner for all charge states.



B) Peptide identifications

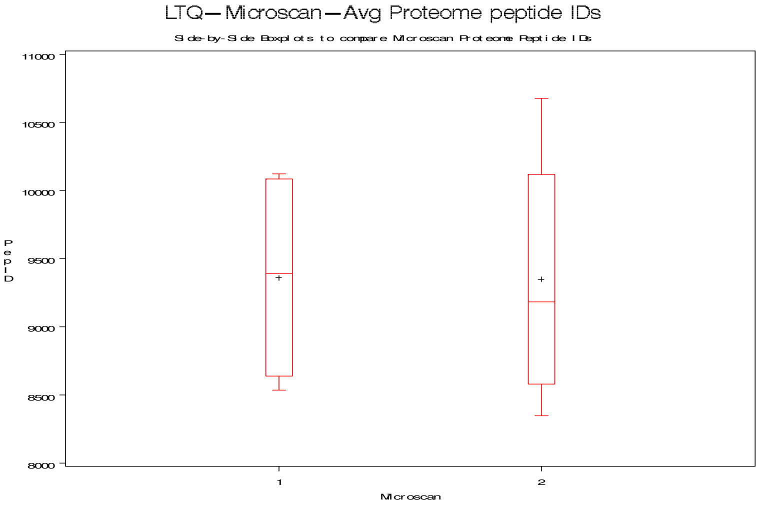


Figure 2. 9 Box plots representing statistical analysis of LTQ microscan proteome data
Data from 6 individual replicates was analyzed via one-way ANOVA and means and standard deviation determined. The x-axis represents the number of MS/MS microscans (1 or 2 microscans) tested on LTQ proteome samples. The box encloses two standard deviations of the data range, while the star in the center represents the mean, and the line represents the median. A) Mean numbers and standard deviation of total numbers of protein identifications. B) Mean numbers and standard deviation of total numbers of peptide identifications. As with the LCQ data, statistical analysis showed that there is no significant difference between the groups in the numbers of protein (p = 0.9061) or peptide (p = 0.9858) identifications obtained when proteome samples are tested.

SEQUEST is not significantly affected by improvements in spectral appearance derived from spectral averaging of LTQ scans.

Discussion

The advantage to using a small set of known proteins for initial analysis is that it provides a method for establishing true positive identifications from false positive identifications. Initial searches against a database containing both forward and reversed protein sequences from each protein included in the EPSM mix allowed for determination of specific data filter levels that would ensure a 95% false positive rate. It also allowed for determination of numbers of true and false identifications for each threshold level, used in creating the ROC curves discussed below. The disadvantage to using a small set of known proteins mixed at equimolar concentrations is that this mixture does not represent a real world sample. Although it can narrow the window of test conditions, an EPSM can not necessarily give an accurate indication of results in extremely complex samples where protein concentrations vary across a wide range. For this reason two levels of tests were done, with the second level including investigation of total protein identifications obtained from total soluble proteome preparation of *R. palustris*.

In order to determine appropriate test levels for proteome experiments, Receiver Operating Charateristic (ROC) curves were compiled from initial protein standard data. Our ROC curves were constructed by plotting the number of true positive peptide identifications versus the number of false positive identifications at a given threshold level. They were initially used to define the performance of radio receivers in terms of signal-to-noise ratios, and are now commonly used to ascertain the efficacy of medical tests [127]. They have also been used in comparisons of performance of search algorithms when searching spectra derived from known protein standard mixes [7]. The advantage of using a protein standard mix for creating ROC curves is the presence of a known set of proteins. Unlike medical tests where the detection of a true positive test result can never truly be "known," a true identification of a peptide from a defined set of proteins has a greater chance of being a true positive, while false identifications can easily be determined.

When constructing ROC curves from EPSM data, the data was pooled to provide a greater statistical representation of peptide identification numbers. On the ROC plot, the vertical axis is number of true positive identifications, while the horizontal axis is the number of false positive identifications represented by the number of peptide identifications made to proteins known not to be included in the EPSM mixture (Figure 2.10). The curves are constructed by determining the number of true positives and the number of false positives at a number of different threshold levels. Identification numbers are separated by charge states of the parent ion because tandem mass spectra fragment ions derived from parent ions of different charge states are identified with different levels of efficiency. Each point on each line of the ROC curves represents the number of true positives versus the number of false positives for a given threshold level. Ideally, the curve would be a straight ninety-degree angle, where 100% of the peptides present would be identified positively before any false positives are identified. Comparisons between curves are made by observing which curve is on top of another, with the curve on top indicating a higher number of overall peptide identifications [7].

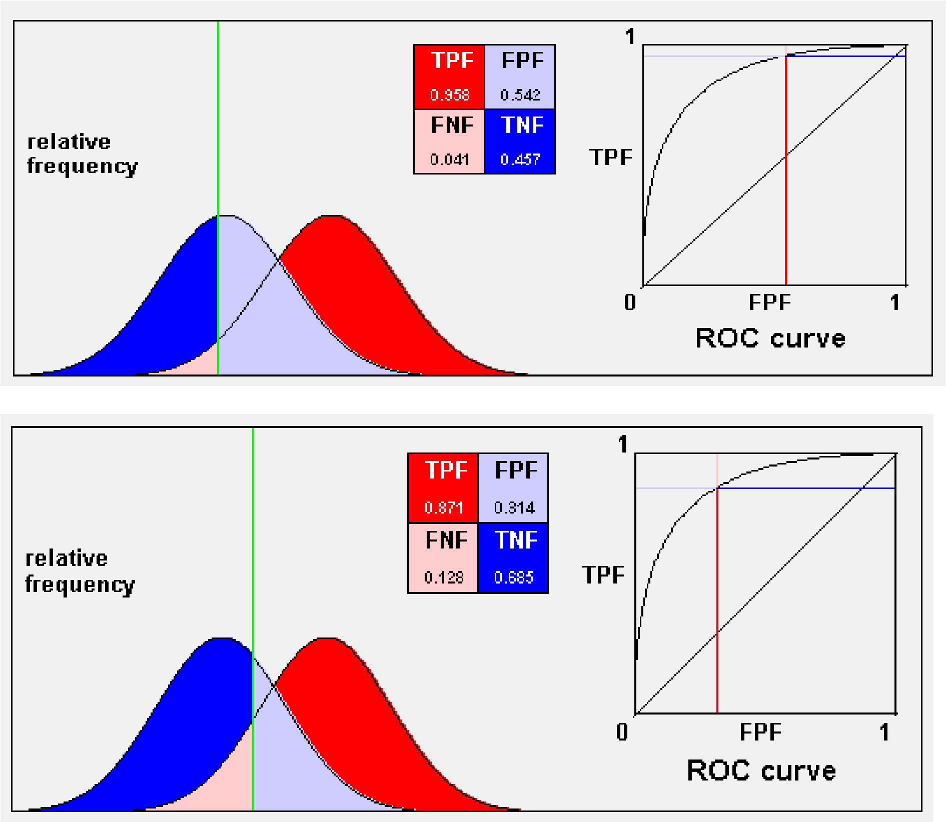


Figure 2. 10 Development of an ROC curve (Adapted from http://www.anaesthetist.com/mnm /stats/roc/)

True negative identifications (TNF, dark blue curve) and true positive identifications TPF, red curve) are represented by Gaussian distribution curves on the left side of the figure above. In the above illustration, the blue curve represents the total number of false negatives while the red curve represents the total numbers of true positive identifications in a given data set. Light blue represents those identifications which are incorrectly identified as true (FPF, false positive fraction). Pink represents those true identifications incorrectly classified as false (FNF, false negative fraction). Overlap of the two curves is indicative of the ability to discriminate between true and false identifications. The green line represents a given threshold. ROC curves on the left are derived from moving the threshold level (green line) and plotting a point representing an x-axis value of false positives and a y-axis value of true positives for that given threshold level.

For this study the curves were not normalized due to technical difficulties in changing the R script from which they were derived. A second level of statistical analysis was then done for each experimental condition at one specific threshold level of 95% false positive rate to determine if results at that threshold level were statistically significant. The isolation m/z width and the number of microscans tested using a proteome sample were based on a combination of results obtained from analysis of pooled EPSM data sets as represented by the ROC curves and from further statistical analysis of individual EPSM sample data sets.

Tandem mass spectrometry is performed through isolating and fragmenting a selected ion of a desired mass-to-charge ratio (m/z). Radio frequency (RF) voltage is applied to isolate the parent ion, but there are a number of factors to consider in that isolation. The isolation window width should be chosen such that the entire isotopic packet of the ion of interest is isolated while at the same time precluding isolation of a co-eluting ion of a similar m/z. If the isolation window width is too small, the entire isotopic packet of the ion of interest may not be collected, which may in turn significantly affect the ability to identify the isolated peptide. On the other hand, if the isolation window width is too large, the risk of selecting a co-eluting peptide of similar m/z increases. Co-eluting peptides could lead to the presence of additional fragment ions and/or an increase in noise levels, and could thus result in an inability to correctly identify the peptide sequence of the parent ion of interest. Choice of an isolation window of 2 m/z remains common because this is the default isolation window in the XCaliber software, while isolation window widths of 3 m/z are common in literature. In contrast, isolation widths of 5m/z have been used in the OBMS group for proteome analysis, with greater identification numbers reported for this larger m/z isolation window. We therefore chose a range of isolation window from 2 m/z units to 5 m/z units for our initial EPSM tests in order to determine the effect of varying the width of this window. Our results from the set of isolation window width tests indicated that although the simpler EPSM mixture appeared to have significant differences in ROC curves (figures 2.4 and 2.5 for the LCQ and LTQ, respectively), statistical analysis did not support that conclusion. More complex proteome mixtures, however, did show a significant decrease in the numbers of identifications obtained with an isolation window width of 2 m/z for the LCQ (Table 2.4) but no significant differences in identifications for any isolation width in the LTQ (Table 2.4). A narrow isolation window width may lead to a lesser ability to isolate parent ions in the LCQ, especially if other isolation parameters are not functioning optimally, due to the lower amount of ions in the trap than in the LTQ linear ion trapping quadrupole instrument.

When considering the effect of varying the number of averaged scans designated to give a final spectrum, both the quality of the acquired final spectrum and the amount of time required to acquire it must be taken into consideration. Mass spectra are obtained by a scan function consisting of an ion injection time, fixed time length prescan followed by an additional variable ion injection time, set by the ion density found in the prescan to result in a specified number of ions in the trap. The final observed spectrum is a result of averaging a number of microscans (Figure 2.1). Higher numbers of microscans result in a spectrum with more normalized peaks, greater signal-to-noise ratio and a greater number of fragment ion peaks, but also increase the amount of time needed to perform each scan. Although the appearance of the spectra may be cleaner with more microscans, the appearance may not make a difference when it comes to identifying the peptide which gave rise to it. In fact, when examining spectra from the identical peptide collected using different experimental conditions (Figure 2.3), no discernible difference in spectral quality is noted.

The additional time required for more microscans may instead hinder peptide identifications because less time is available to collect spectra during each point in the chromatographic run. For instance, if you have a scan function lasting for a time period of 125 msec, and you are performing 4 microscans per observed spectrum, you have a total spectral collection time of 500 msec per spectrum, thus limiting the number of spectra you can collect to two spectra per second. This longer scan time may in turn decrease the number of identifications that can be made. We found that increasing the time required for doing additional microscans did have a negative impact on the number of identifications acquired, with a far greater difference noted in data obtained on the LTQ mass spectrometer than the LCQ mass spectrometer.

But the question remains, what number of microscans can produce the happy medium of adequate spectra plus maximum identifications? Does an increased number of microscans increase the search algorithm's ability to distinguish true and false identifications? Earlier studies have yielded conflicting results when attempting to answer these questions on either a 3-dimensional (3D) trapping quadrupole or a linear trapping quadrupole instrument. When varying only the number of microscans while leaving all other mass spectrometry parameters constant, the number of microscans does seem to result in significant differences in the number of identifications [60] or in better scores for each identified peptide [59]. However, when varying a number of parameters simultaneously and considering interactions between these parameters in a multi-parametric statistical modeling approach [34, 57, 58], the number of microscans in a 3D trapping quadrupole instrument does not emerge as a significant factor in the number of identifications acquired [57, 58]. In our tests, the number of microscans was held constant for the parent ion, and settings were based on the standards used in the industry (typically 3 microscans for a full scan in LCQ, and 1 microscan for full scans in the LTQ). The number of microscans used to collect the tandem (MS/MS) mass spectra were varied according to experimental parameters shown in Table 2.3. Further, microscan numbers for the LCQ quadrupole ion trap were larger than those tested for the LTQ linear ion trap due to the larger number of ions that can be trapped within a linear trap. Performing a higher number of microscans while acquiring spectra in the LCQ ion trap can help to compensate for the lower ion population in the trap. The LTQ ion trap has a much higher ion population, resulting in better ion statistics, and thus a lower number of microscans should be sufficient for obtaining pretty spectra as well as maximum identifications [56]. Since only 1 microscan should be required, more spectra can be collected due to the decreased amount of time required to collect spectra. Just as was done for the isolation width data, initial tests for the microscan data were performed using a known set of proteins mixed in equimolar amounts, and results used in testing a more complex proteome mixture.

Analysis of microscan data indicated that significant differences emerged in data collected using different numbers of MS/MS microscans in the LCQ (Figure 2.7). Even though 5 microscans gave the lowest numbers of identifications (Table 2.5), statistical analysis did not show a significant difference in number or percentage of identifications for the LCQ microscan data. However, 2 microscans resulted in a significantly lower percentage of spectra identified for the total spectra collected, suggesting that a lower number of MS/MS microscans negatively impacts the ability of search algorithms to assign identifications to the total numbers of spectra being collected. In other words, even though there are more spectra being collected due to the shorter collection time, less of these spectra are being identified. A different picture emerged for the LTQ microscan data, however, with significant differences noted between all experimental conditions for the EPSM mixture (Figure 2.9). Performing only 1 MS/MS microscan when acquiring data on the LTQ did result in higher numbers of identifications (Table 2.5), most likely due to the lesser amount of time required to collect the spectra, but the difference between higher numbers obtained with 1 microscan versus the numbers obtained with 2 microscans proved to be statistically insignificant.

Conclusions

Isolation width window studies showed no significant differences between experimental conditions tested for either protein standard mix or proteome samples. A smaller m/z isolation window slightly increased the number of identifications in the LCQ, while a slight decrease was noted in data from the LTQ for smaller m/z isolation windows. Opening the width of the isolation window to 5 m/z units did not change the number of identifications obtained in this study for either LCQ or LTQ data. The decreased probability of identifications due to co-eluting peptides was not seen in this study. However, determination of interference in the MS/MS from fragmentation of both the parent ion of interest and a co-eluting peptide with a larger m/z window may only be noticed with manual processing of individual spectra. It is possible that automated processing with the available software simply can not distinguish the presence of a co-eluting peptide, perhaps making a correct identification regardless of the presence of additional fragment ions or an increase in signal-to-noise ratio. It is also possible that a false positive identification is made in this instance, or that no identification is made at all. The question of interference of co-eluting peptides with the ability of the search algorithm to identify them could possibly be answered by testing a less complex mixture of only one protein, or a mixture of two similar peptides. Since no significant difference was noted in identifications obtained in isolation width experiments, it is recommended that the industry standard of an isolation width of 3 m/z units be maintained.

Microscan studies also yielded no significant differences for proteomic studies. Less microscans on the LCQ are not beneficial for peptide or protein identification, perhaps due to the limited number of ions in the trap resulting in an MS/MS spectrum that contains more noise or does not have enough fragment ions present for identification. In contrast, performing more microscans in an LTQ MS/MS experiment does not significantly increase the number of peptide (and thus protein) identifications obtained. The greater amount of time required for additional scans results in less spectra collected. If the same percentage of spectra are being identified regardless of the number of microscans, then a lower number of spectra will result in less identifications just due to simple percentage identified. Like other studies before this one, no significant differences in number of identifications obtained were noted in the lower number of microscans tested for the proteome samples. Thus, the number of MS/MS microscans used for the LCQ should be kept at 3, while 1 MS/MS microscan is sufficient for maximal identifications in the LTQ.

Chapter 3. Proteomic Investigations of *A. brasilense* strains Sp245 and Sp7 under nitrogen fixing and non-nitrogen fixing conditions

Introduction

Azospirillum brasilense, introduced in Chapter 1, are free-living soil bacteria with plant growth promoting capacities, and as such are important organisms in the study of plant-microbe interactions. The physiological versatility of *A. brasilense* gives them a competitive advantage in utilization of the limited nutrients found within the rhizosphere. Because of the positive effect on the growth of grasses when soil is inoculated with *Azospirillum* species, genetic studies have been carried out on *Azospirillum* species for a number of years. The genome structure has been found to be highly mobile, containing a number of phage sequences integrated into the genome of all *Azospirillum* species [128]. All species carry a number of plasmids [129, 130]. Earlier studies revealed the presence of 5 megareplicons, 2 circular and 3 assumed to be linear, within all *A. brasilense* strains studied [130], but the number of plasmids are variable between strains and also between different cultures of the same strain. A plasmid of ~142-kb with a mass of ~85-MDa from one culture of *A. brasilense* Sp245 has recently been characterized, but is not present in the recently sequenced Sp245 strain [131], thus demonstrating the plasmid variability among different strains or even different cultures of the same strains.

One variant of *A. brasilense* strain Sp245 has been sequenced by the Joint Genome Institute (JGI) and the genome sequence can be found at http://genome.ornl.gov/microbial/abra/19sep08/. This draft genome sequence is 7.5 Mb in 70 contigs of 20 reads or greater, contains 68.4 % GC content, and has 7043 candidate protein-encoding gene models available. From this draft sequence, a protein sequence database consisting of 6927 potential protein coding sequence translations has been derived.

A. brasilense strain Sp7, in contrast, contains a smaller genome of ~6.3Mb, with variability in the plasmid sizes seen between it and the Sp245 genome structure [130]. Additionally, Sp7 has one essential plasmid of 151.3 kb with a mass of 90 kDa, named pRHICO or p90, which contains genes necessary for exopolysaccharide (cell wall) biosynthesis, as well as for synthesis and export of polysaccharides. Other genes involved in beta-lactamase expression for resistance to ampicillin are also present on the pRHICO plasmid [132, 133]. A number of genes on this plasmid have been shown to be present within megareplicons of the Sp245 strain. Strain Sp7 is closely related to strain Sp245, but its genome sequence is not yet available. Searching the NCBI nucleotide database (http://www.ncbi.nlm.nih.gov) using the term "*Azospirillum brasilense* Sp7" returns 67 results of individual genes sequenced from the *A. brasilense* Sp7 strain.

A BLASTx [134] search comparing the translated coding sequences to those in the Sp245 genome revealed all but 1 Sp7 gene sequence to be present in the Sp245 genome with the remaining genes having 90-100% similarity to Sp245 sequences at a protein level. Therefore, in this work we have used the Sp245 database in the analysis of data derived from the Sp7 species. A small percentage of expressed proteins may go undetected in the Sp7 strain data by using the Sp245 genome for searches, but the large degree of similarity in 95% of the translated Sp7 coding sequences to those in the Sp245 genome suggests that those proteins detected are likely to be accurate. Using the same protein databases for both species also facilitates direct comparisons of protein expression by each strain.

Protein products from a variety of genes are involved in nitrogen fixation. These genes encode not only the structural components of the nitrogenase itself (discussed in Chapter 1), but also those genes involved in synthesis of both the nitrogenase and the iron-molybdate cofactor, in electron transport to the nitrogenase, and in regulation of nitrogen fixation [96]. The *nif*HDK operon is about 5500bp and encodes the basic structural components for the molybdenum nitrogenase. Additional related *nif* genes can be found both upstream and downstream of this cluster [135]. Altogether, these genes are found clustered in a 30 kb DNA region in the genome [98]. The *fix* genes proposed to be involved in electron transfer to the nitrogenase components [136-138] are found in close proximity to nitrogenase structural component genes within the genome. Since *Azospirillum brasilense* has served as a model organism for nitrogen fixation among soil bacteria of genus *Azospirillum*, nitrogen fixation genes have been well characterized [91].

While Sp245 can enter the intercellular space of root cells and live within this somewhat protected environment, Sp7 does not possess this ability, but instead forms clumps on the outer surface of the root hairs [93]. Since these strains are suited to grow in different environments, their physiological responses to external stimuli are likely to be slightly different as well. For instance, in response to heavy metal stimuli there are noticeable differences in physiological effects between strains [92]. Sp7 shows accumulation of poly-hydroxybutyrate (PHB) compounds within the cell in response to metal stress, while Sp245 does not. On the other hand, both strains show decreased production of indole-3-acetic acid (IAA) in response to metal stress, but the decrease is far greater for Sp245 than Sp7 [92]. Thus, comparison of entire proteome expression for these two strains grown under two different growth conditions should reflect the differences in physiology required for life in different ecological niches.

This study presents comparative proteomic investigations of 2 different strains of *A. brasilense,* Sp245 and Sp7, grown in minimal media under both nitrogen limited (nitrogen fixing) and nitrogen replete (non nitrogen fixing) conditions, with vanadium trichloride added to the media of one Sp245 nitrogen fixing culture. The goal of this research was to characterize the overall physiology of both strains under these defining growth conditions and to reveal putative differences and similarities. Bottom-up mass spectrometry employing a MudPIT separation technique [17, 19, 20] coupled to an LTQ linear quadrupole mass analyzer was used to investigate protein expression under each growth condition. Comparison of global changes in protein expression between these two growth conditions were investigated in more detail and indicate that while most of the proteins detected under both conditions are those involved in protein translation, ribosomal syntheses and biogenesis, nitrogen fixing cells expressed more proteins involved in nutrient synthesis and manipulation. In addition to adding information to complete genome sequence annotation, analysis of changes in the physiology of *A. brasilense* cells in presence or absence of combined nitrogen by proteomics provides further insight into how this diazotroph adapts to nitrogen-fixing conditions.

Materials and Methods

Cell growth

Overnight starter cultures (5 mL) were inoculated from fresh plates. Starter cultures were grown overnight at 27°C in a shaking water bath in minimal media containing malate as carbon source and ammonium chloride as nitrogen source. Cells were pelleted from starter cultures and washed with appropriate growth media. Base media for all cultures was minimal media (MMAB) [139] with 20 mM malate as carbon source, and ammonium chloride as nitrogen source where appropriate. All Sp7 strain cultures and Sp245 strain control (non nitrogen fixing) cultures were supplemented with molybdate, while one Sp245 nitrogen fixing culture was supplemented with molybdate only and a second culture supplemented with vanadium trichloride only. Starter cultures were resuspended with appropriate media and used to inoculate 500 ml cultures for control (non-nitrogen fixing) growth, 250 mL cultures for nitrogen fixing growth. Nitrogen fixation requires a great deal of energy and continuous optimal oxygen concentrations, so growth of nitrogen fixing cells is slower than those growing in nitrogen sufficient conditions. Cells grown under nitrogen fixing conditions exhibit a doubling time of 170 minutes while control (non nitrogen fixing) cells have a doubling time of 120 minutes [140]. Further, OD of cells grown under nitrogen fixing cultures never reaches high levels, tending to level off at or below an OD_{600} of 0.2 – 0.3 [140]. Therefore, each growth condition was optimized as follows: nitrogen fixing cultures were grown at 25°C without shaking to early log phase (OD_{600} = 0.1 - 0.2) to minimize exposure to high levels of oxygen Azospirillum spp. are microaerophilic diazotrophs, and as such, control cultures (non nitrogen fixing) were grown under optimum growth conditions (shaking and in presence of ammonium) at 25°C on an orbital shaker to mid-log phase (OD_{600} = 0.5 – 0.6). Cells were harvested by centrifugation at 8000 rpm for 10 minutes, washed twice with 50 mM Tris (pH 7.9), then pelleted by centrifugation at 8000 rpm for 10 minutes. Two separate cultures were grown under each condition, here termed biological replicates. Cell pellets from these two biological replicates were pooled for subsequent proteome preparation in order to ensure sufficient protein concentration allowing technical replicates (two mass spectrometry runs of the same sample) to be performed. Wet cell pellet weight was recorded and cell pellets were frozen for later analysis.

Proteome preparation for LC/LC-MS/MS

Frozen cell pellets (0.1 g for each sample) were resuspended at a rate of 500 µl lysis buffer/ 0.1 g wet cell pellet weight in lysis buffer of 6 M guanidine hydrochloride, 10 mM DTT solubilized in 50 mM Tris-HCl, 10 mM $CaCl_2$. Resuspended cells were then further lysed by sonication. Lysate was centrifuged at 18,000 xg for 20 minutes to clear cellular debris. Supernatant was collected, with 1 ml of lysate used for immediate tryptic digestion and the rest frozen at -80°C for later digestion. Protein concentration was not quantified due to interference of guanidine salt with commonly used protein quantification assays. Trypsin digestion was performed according to manufacturer's recommendations. Briefly, 10 mM DTT was added and lysate was incubated at 60°C for 1 hour. Lysate was then diluted 6-fold with trypsin digestion buffer (50 mM Tris-HCl, 10 mM $CaCl_2$, 10 mM DTT, pH 7.9) and 20 µg sequencing-grade trypsin (Promega, Madison, WI) was added to each sample. Samples were incubated overnight at 37°C with gentle rotation. An additional 20 µg of trypsin was added the following morning and samples were subsequently incubated for an additional 5-6 hours at 37°C with gentle rotation. Digestion was halted by addition of 5 µl formic acid to the 5 ml lysate. Samples were then desalted using Sep-Pak Plus C-18 solid phase extraction (Waters, Milford, MA) following manufacturer's recommendations, and subsequently concentrated and solvent-exchanged into 100% HPLC-grade H_2O, 0.1% formic acid using vacuum centrifugation (Savant, Thermo Scientific). Samples were aliquoted into 40 µL volumes and stored at -80°C until analysis.

LC/LC-MS/MS analysis

Instrument performance was evaluated before beginning experimental mass spectrometry proteomic runs through first running an *R. palustris* soluble proteome standard. Proteome standards were run according to the protocol described in Chapter 2, with a 5-step MudPIT setup [17, 19]. Parameters evaluated from the standard included chromatographic separation capacity, normalization levels of the mass spectra, evaluation of signal-to-noise ratio and numbers of proteins resulting from data searches as described in Chapter2. An additional *R. palustris* proteome standard was run at the end of experimental runs and results compared in order to ensure that the LTQ mass spectrometer was functioning optimally throughout the experimental runs.

Proteome samples were analyzed in quadruplicates via Multi-dimensional Protein Identification Technology (MudPIT) [17, 19, 20] with triphasic columns. Columns were individually packed using a pressure cell (New Objective, Woburn, MA). Back columns were loaded first with 3 cm of Luna 5µm strong cation exchange (SCX) resin (Phenomenex, Torrance, CA) followed by 3 cm of Aqua 5 µm C-18 reverse phase resin

(Phenomenex). Proteome aliquots (50 µl) were loaded directly onto the back column via pressure cell and subsequently coupled to the front column. Final protein concentration loaded onto the back column was evaluated by the intensity level of the chromatogram, and amount of protein loaded was adjusted in subsequent runs to achieve approximately equal loading amounts. Front columns were pulled to a tip with an inside diameter of 5 µm using a P-2000 laser puller (Sutter Instruments, Novato, CA), and packed with 17 cm of Aqua 5 µm C-18 reverse phase resin. This column acts as the resolving column for peptides eluted from the back column. For analysis, the combined columns were placed directly in-line with the linear trapping quadrupole (LTQ) mass spectrometer (ThermoFinnigan, San Jose, CA) using a Proxeon source.

Chromatographic separation was accomplished with an Ultimate HPLC system (LC Packings, a division of Dionex, San Francisco, CA) providing a flow rate of 100 µl/ minute which was split prior to the resolving column such that the final flow rate at the tip was ~300 nl/minute. Twelve two-dimensional (2D) chromatographic steps were done. An initial 1 hour gradient from buffer A (95% water, 5% acetonitrile, 0.1% FA) to buffer B (70% acetonitrile, 0.1% FA) bumped the peptides from the initial reverse phase column onto the strong cation exchange column. Subsequent cycles included 2 minute salt pulses with varying percentages of 500 mM ammonium acetate (10, 15, 20, 25, 30 35, 40, 45, 50, 60%) to first elute subsets of peptides according to charge, followed by a 2 hour gradient from buffer A to buffer B, to further separate peptides by hydrophobicity. The final chromatographic step consisted of a 20 minute salt pulse of 100% 500 mM ammonium acetate, followed by a 2 hour A-to-B gradient.

Data collection was controlled by Xcaliber software (ThermoFinnigan). Data was collected in data-dependent mode using parameters established in Chapter 2 with one full scan followed by 6 dependent scans, each with 2 microscans. Dynamic exclusion was employed with a repeat count of 1, repeat duration of 60 and exclusion list size of 300 and duration of 180 msec. Isolation mass width was set at 3 m/z units.

Data Analysis

The two best technical replicates were used in all data analysis. An *Azospirillum* protein database based on the newly sequenced Sp245 genome was constructed from CDS text files based on translations of the coding sequences called in the draft genome sequence (http://genome.ornl.gov/microbial/abra/19sep08/). A list of common contaminants was appended to the gene call sequences, and all coding sequences, including contaminant sequences, were reversed and appended to the forward sequences in order to serve as distractors. The protein database can be downloaded at https://compbio.ornl.gov/mspipeline/azospirillum. Due to the high degree of similarity of translated coding sequences available for the Sp7 strain to the translated coding sequences of the Sp245 strain, other databases were not appended for the Sp7 database searches. Further, additional databases were not included in the search due to the large increase in time required for searches of larger databases. All MS/MS spectra for both Sp245 and Sp7 strains were searched against this database using SEQUEST [3] specifying tryptic digestion, peptide mass tolerance of 3 m/z and a fragment ion tolerance of 0.5 m/z. Additionally, search parameters included two dynamic modifications: 1. methylation represented by a mass shift of +14 m/z on aspartate residues, and 2. deamidation followed by methylation represented by a mass shift of +15 m/z on glutamine residues. Output data files were sorted and filtered with DTASelect [54] , specifying XCorr filter levels of 1.8 for peptides with a charge state of +1, 2.5 for those with charge state +2 and 3.5 for charge state +3, minimum delta CN of 0.08, semi-tryptic status and 2 peptides per protein identification. DTASelect files in both html and text formats are posted at https://compbio.ornl.gov/mspipeline/azospirillum/study1/ status.html.

Results were imported into Microsoft Access for data analysis. Results for each technical replicate run of an individual growth state were combined to give an average number for that growth state, and combined results were used in all analysis. From the number of identifications in the reverse direction, peptide false positive rates were determined using the formula %FP = 2[No. reverse ID/ (no. reverse ID + no. real ID)] [18]. In order to determine relative abundance of a given protein in a sample, normalized spectral abundance factors (NSAF) were calculated for each individual protein k using the formula $NSAF_k = (SpC/L)_k / \Sigma (SpC/L)_{n,}$ where SpC is the total spectral count for all peptides contributing to protein k, L is the length of protein k, and n is the total number of proteins detected in the sample [72]. For those proteins found in both replicate runs, an average of two NSAF values was calculated. Pairwise comparisons of the calculated NSAF for each protein in each growth state were used to determine up- and down-regulation. Experiments done by Zybailov et al [24] indicated that a 1.4-fold increase in NSAF was significant for their yeast membrane proteome data set with 9 replicate analyses. With only two replicates, statistical significance is difficult to determine, so in order to ensure significance of up-regulation in this data set, data was filtered such that only a 5-fold or greater difference in the calculated NSAF was reported.

For analysis of fold-change of proteins between growth states, proteins were required to be present in both growth conditions, so the NSAF values could be compared directly across growth conditions. Because NSAF values are reflective of the abundance of a protein within a sample (discussed in chapter 1 of this work), comparison of the NSAF level of a protein in nitrogen fixing conditions with that of the same protein in non nitrogen fixing conditions gives an estimate of the degree of up-regulation or down-regulation of that protein when conditions change. Dividing the NSAF value of each single protein found in nitrogen fixing conditions by the NSAF value of that same protein in non nitrogen fixing control conditions gives an approximate fold-change value of that protein. Proteins present in only one growth state were not considered in this analysis because of the possibility that these proteins are expressed only in one growth state and not the other, making it impossible to determine if levels of up-regulation or down-regulation are appropriate.

Results

Overall proteome evaluation

The newly sequenced *Azospirillum brasilense* Sp245 genome contains 6927 candidate gene coding sequences. As shown in Table 3.1, our proteomic study identified a total of 1289 proteins (19%) in 2 combined technical replicates of Sp245 cultures grown under optimal growth conditions (in presence of ammonium and with shaking), herein termed control conditions and 1372 (20%) and 1269 (18%) proteins in Sp245 cultures grown under nitrogen fixing conditions with molybdate and vanadium, respectively. Reproducibility between technical replicate runs was 67% for both Sp245 control cultures and nitrogen fixing with molybdate, while reproducibility between

Table 3. 1Total numbers of protein and peptide identifications in each sample

Sample	Number of Protein Identifications			Number of Peptide Identifications	
	Replicate 1	Replicate 2	Combined Dataset	Replicate 1	Replicate 2
Sp245 Control	976	1177	1289	7220	10259
Sp245 Nitrogen Fixing - Mo	1155	1159	1372	9133	8324
Sp245 Nitrogen Fixing - V	1067	1171	1269	9220	6167
Sp7 Control	812	851	1117	6169	7540
Sp7 Nitrogen fixing	985	1017	1181	8076	7839

nitrogen fixing vanadium culture technical replicates was slightly higher at 71%. False positive rates at a peptide level for Sp245 cultures ranged from 1.4% for non nitrogen fixing optimally-growing cultures to 2.8% for nitrogen fixing cultures with molybdate. Sp7 cultures yielded slightly lower totals of 1117 (16%) and 1181 (17%) protein identifications for data from combined technical replicates of control optimal growth (non-nitrogen fixing) cultures and nitrogen fixing cultures, respectively. Optimal growth controls and nitrogen fixing Sp7 cultures had 77% and 68% technical replicate reproducibility respectively, with higher false positive rates at a peptide level of 4.3% and 2.3% for non nitrogen fixing and nitrogen fixing cultures respectively.

Normalized spectral abundance factor (NSAF) values were calculated for each protein detected. Proteins were grouped into functional categories and the NSAF values totaled for each protein within a given category. Percentages of the proteome devoted to each functional category based on the totaled NSAF values for each protein within the category were calculated and results are given in Table 3.2 for Sp245 and Table 3.3 for Sp7 strain. The number of proteins contributing to each category, and the percentage of those expressed proteins based on total spectral abundance (ΣNSAF) of all proteins within a given category is tabulated with results as shown. Visual representation of the percentage of the expressed proteins devoted to each functional category based on the totaled NSAF values (ΣNSAF) for each protein within a functional category is shown in the pie charts of Figure 3.1 for Sp245 cultures and Figure 3.2 for Sp7 cultures.

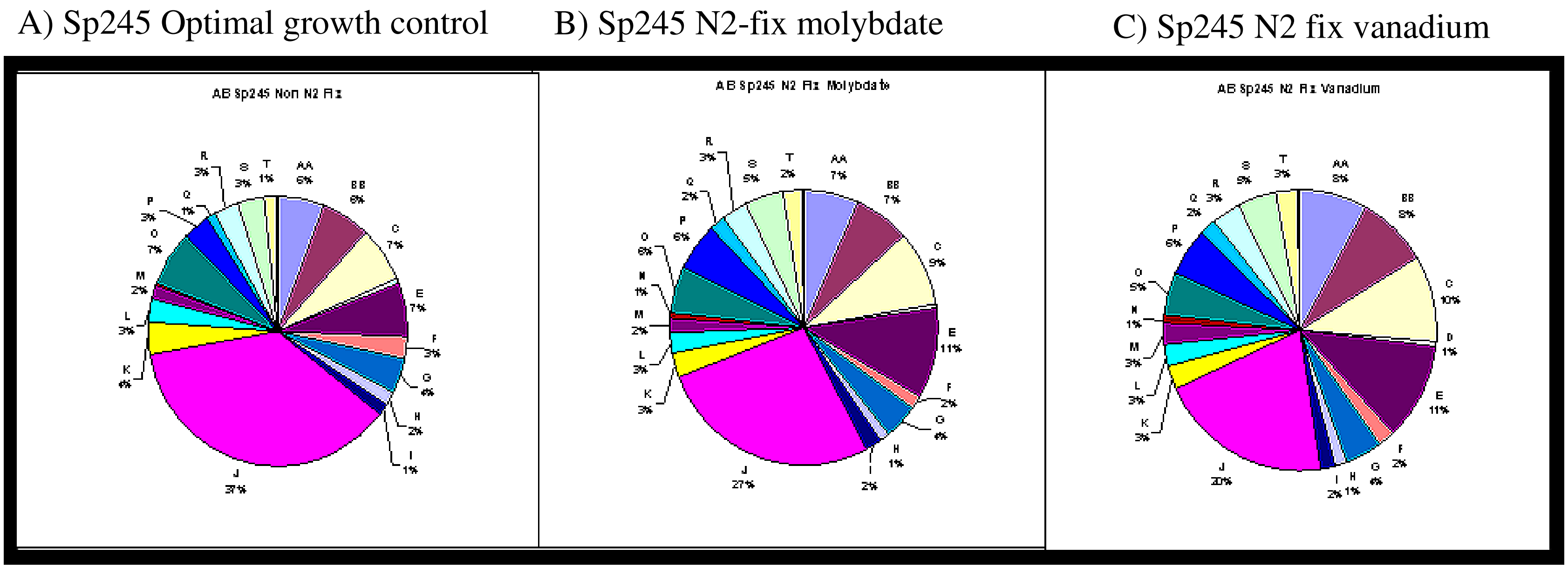
Table 3. 2 Numbers and relative abundance percentages (by NSAF) of proteins identified in each Sp245 sample by functional category

		Sp245	Sp245 control		Nitrogen Fix -Mo		Nitrogen Fix-V	
FC	Functional Category Description	No. Common Proteins^a	No. Proteins^b	ΣNSAF Percentage^c	No. Proteins^b	ΣNSAF Percentage^c	No. Proteins^b	ΣNSAF Percentage^c
B	Chromatin structure and dynamics	0	0	NA	1		1	
C	Energy production and Conversion	76	93	6.6%	115	8.8%	110	8.5%
D	Cell Division and Chromosome Partitioning	11	21	0.5%	19	0.4%	16	0.4%
E	Amino Acid Transport and Metabolism	99	133	6.5%	144	10.5%	137	11.0%
F	Nucleotide Transport and Metabolism	31	40	2.6%	42	1.9%	42	2.0%
G	Carbohydrate Transport and Metabolism	40	54	4.3%	56	4.0%	50	4.2%
H	Coenzyme Metabolism	29	60	1.5%	52	1.2%	44	1.1%
I	Lipid Metabolism	20	32	1.3%	33	1.8%	30	1.8%
J	Translation, Ribosomal Structure and Biogenesis	111	127	36.8%	122	26.9%	121	26.0%
K	Transcription	35	53	3.9%	49	2.9%	47	3.0%
L	DNA Replication, Recombination and Repair	19	41	2.7%	24	2.5%	28	3.0%
M	Cell Envelope Biogenesis, Outer Membrane	23	47	1.6%	41	1.5%	32	1.3%
N	Cell Motility and Secretion	15	26	0.3%	44	0.6%	28	0.4%
O	Posttranslational Modification, Protein turnover, Chaperones	49	58	6.7%	57	5.6%	60	5.2%
P	Inorganic Ion Transport and Metabolism	38	47	3.2%	57	5.7%	56	6.2%
Q	Secondary metabolites biosynthesis, transport and catabolism	20	28	1.2%	35	1.9%	28	1.9%
R	General Function Prediction Only	59	95	3.1%	94	3.0%	80	3.1%
S	Function Unknown	53	77	3.1%	94	4.8%	85	4.8%
T	Signal Transduction Mechanisms	21	32	1.4%	51	2.2%	43	2.3%
U	Intracellular trafficking and secretion	6	10	0.2%	8	0.2%	6	0.2%
V	Defense mechanisms	0	4	0.05%	1	0.01%	2	0.02%
AA	Hypotheticals with similarity or homology to known	45	79	5.8%	84	6.7%	81	6.3%
BB	Hypothetical	74	132	6.1%	149	6.7%	142	7.2%
	TOTAL	874	1289	99.5%	1372	99.8%	1268	99%

[a] Number of common proteins identified in all growth states

[b] Number of proteins identified belonging to the functional category in that growth state

[c] NSAF values were totaled for proteins in each individual functional category (ΣNSAF) to determine the percentage of the total expressed proteome within each functional category



A) Sp245 Optimal growth control

B) Sp245 N2-fix molybdate

C) Sp245 N2 fix vanadium

Figure 3. 1 Proteome expression profile by relative abundance (NSAF) in Sp245 cells
Pie charts were constructed by first totaling the NSAF values (ΣNSAF) of proteins found within each functional category, then determining what percentage of the total ΣNSAF of all proteins detected was devoted to each individual functional category. From this it is clear that nitrogen fixing cells devote less of the overall expressed proteome energy to translation and ribosomal biosynthesis (J) and more energy to energy production and conversion (C) and amino acid transport and metabolism (E).

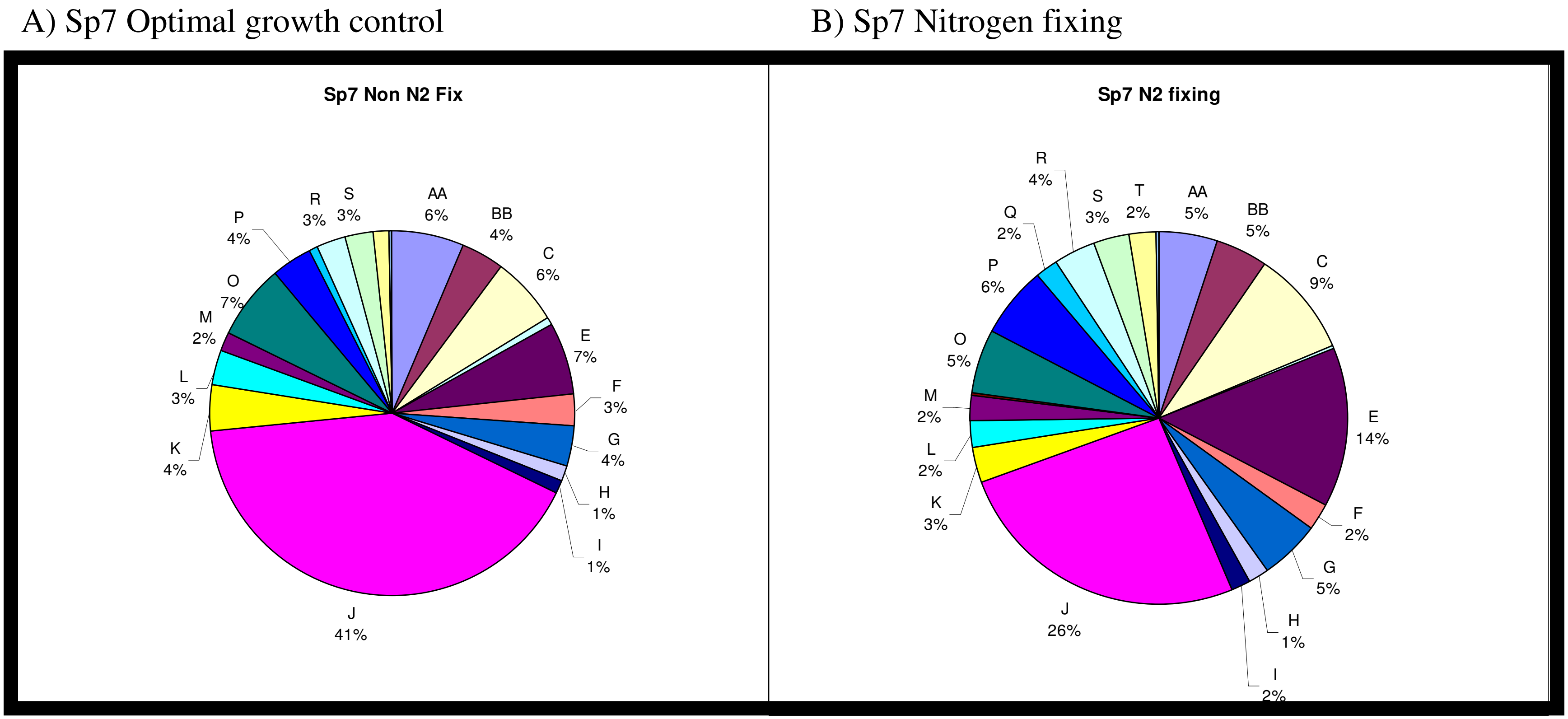
Table 3. 3 Numbers and percentages of relative abundance of proteins (by NSAF) identified in each Sp7 sample by functional category

		Sp7	Sp7		Sp7-N2Fix	
FC	Functional Category Description	No. Common Proteins[a]	No. Proteins[b]	ΣNSAF Percentage[c]	No. Proteins[b]	ΣNSAF Percentage[c]
B	Chromatin structure and dynamics	1	1	NA	1	NA
C	Energy production and Conversion	80	90	6.3%	109	9.0%
D	Cell Division and Chromosome Partitioning	14	18	0.4%	16	0.4%
E	Amino Acid Transport and Metabolism	110	135	6.5%	137	13.9%
F	Nucleotide Transport and Metabolism	33	37	2.6%	42	2.2%
G	Carbohydrate Transport and Metabolism	40	49	3.9%	55	5.3%
H	Coenzyme Metabolism	33	45	1.2%	48	1.5%
I	Lipid Metabolism	26	35	1.2%	34	1.8%
J	Translation, Ribosomal Structure and Biogenesis	113	126	41.1%	119	25.9%
K	Transcription	35	45	4.2%	37	2.8%
L	DNA Replication, Recombination and Repair	25	33	3.0%	34	2.4%
M	Cell Envelope Biogenesis, Outer Membrane	22	37	1.6%	38	2.1%
N	Cell Motility and Secretion	10	17	0.2%	19	0.3%
O	Posttranslational Modification, Protein turnover, Chaperones	49	59	6.6%	57	5.5%
P	Inorganic Ion Transport and Metabolism	32	43	3.7%	49	6.3%
Q	Secondary metabolites biosynthesis, transport and catabolism	21	24	0.6%	33	1.9%
R	General Function Prediction Only	54	75	2.5%	76	3.7%
S	Function Unknown	42	61	2.5%	73	2.9%
T	Signal Transduction Mechanisms	21	31	1.3%	37	2.2%
U	Intracellular trafficking and secretion	7	11	0.3%	7	0.3%
V	Defense mechanisms	2	4	0.05%	4	0.05%
AA	Hypotheticals with similarity or homology to known	41	64	6.3%	61	4.9%
BB	Hypothetical	52	77	3.8%	95	4.6%
	TOTAL	863	1117	99.8%	1181	99.9%

[a] Number of proteins identified in each growth state

[b] Number of proteins identified in each functional category in that growth state

[c] NSAF values were totaled for proteins in each individual functional category (ΣNSAF) to determine the percentage of the total expressed proteome within each functional category



A) Sp7 Optimal growth control

B) Sp7 Nitrogen fixing

Figure 3. 2 Sp7 protein expression by functional category in A) Sp7 optimal growth controls and B) Sp7 nitrogen fixing cultures
Pie charts were constructed by first totaling the NSAF values of proteins found within each functional category (ΣNSAF), then determining what percentage of the total NSAF of all proteins detected was devoted to each individual functional category. From this it is clear that nitrogen fixing cells devote less of the overall expressed proteome energy to translation and ribosomal biosynthesis (J) and more energy to energy production and conversion (C) and amino acid transport and metabolism (E).

Expression of nitrogenase structural components

Examination of the genome structure shows clusters of nitrogen fixation genes in addition to the genes encoding the structural components of nitrogenase. The majority of the structural genes and associated synthesis proteins are found within a 30kb region of DNA, with other associated proteins scattered throughout the genome. All nitrogenase structural components, NifH, NifD, and NifK were detected in all Sp245 nitrogen fixing cultures (Table 3.4), with nitrogenase alpha chain having 40% sequence coverage. None of these components were detected in cultures that were grown with ammonium, consistent with the induction of these proteins only under nitrogen fixation conditions.

The Sp7 genome also shows two nitrogen fixation clusters in a 30kb region of Sp7 DNA [98] with additional nitrogen fixation related proteins being scattered throughout the genome [135]. Table 3.5 contains a list of those proteins involved in nitrogen fixation that were detected in the Sp7 proteomes. Interestingly, only NifH subunit of nitrogenase (abra_1247) was found in Sp7 nitrogen-limited cultures and other structural components could not be detected. This is not surprising since a BLASTx [134] search of the translated coding sequences of the Sp7 strain *nif* operon revealed only a 32% and 27% sequence similarity to the Sp245 molybdate nitrogenase alpha (abra_1246) and beta (abra_1245) chains, respectively. The Sp7 translated coding sequence for the Fe protein, NifH, however, showed a 100% sequence similarity to the Sp245 NifH sequence, while other translated coding sequences in the Sp7 *nif* operon showed between 94% and 99% sequence similarity to those found in the Sp245 genome sequence.

Table 3. 4 Nitrogenase components and related proteins necessary for nitrogen fixation detected in both nitrogen fixing and optimal growth control Sp245 cultures

Gene Name	Description	Sp245-N2Fix- Mo		Sp245-N2Fix-V		Sp245	
		No. Replicates[a]	Average NSAF[b]	No. Replicates[a]	Average NSAF[b]	No. Replicates[a]	Average NSAF[b]
abra_0572	nitrogen regulatory protein P-II	2	9.7E-03	2	8.6E-03	2	1.0E-03
abra_1231	cysteine desulfurase NifS	2	2.5E-04	2	2.4E-04	0	ND[c]
abra_1232	NifU Fe-S cluster assembly protein NifU	2	1.7E-03	2	1.3E-03	0	ND
abra_1236	ferredoxin III, nif-specific	2	5.9E-04	2	7.9E-04	0	ND
abra_1237	protein of unknown function DUF683	2	3.1E-03	2	2.8E-03	0	ND
abra_1238	nitrogen fixation protein	1	3.6E-04	2	1.7E-04	0	ND
abra_1239	nitrogen fixation protein NifX	2	5.8E-04	2	1.6E-03	0	ND
abra_1245	nifK nitrogenase molybdenum-iron protein beta chain	2	1.1E-03	2	1.3E-03	0	ND
abra_1246	nitrogenase molybdenum-iron protein alpha chain	2	5.1E-04	2	6.0E-04	0	ND
abra_1247	nifH nitrogenase iron protein	2	1.0E-02	2	1.1E-02	0	ND
abra_1251	nitrogenase-associated protein	1	4.6E-04	0	ND	0	ND
abra_1254	conserved hypothetical protein	2	2.1E-03	2	3.2E-04	0	ND
abra_1257	nitrogen fixation protein FixT	2	2.1E-03	2	2.2E-03	2	4.1E-04
abra_1350	NifQ family protein	1	1.5E-04	2	2.8E-04	0	ND
abra_3965	ptsN PTS IIA-like nitrogen-regulatory protein PtsN	2	1.0E-03	2	1.1E-03	2	1.4E-03
abra_0535	glnD protein-P-II uridylyltransferase	0	ND	0	ND	1	3.0E-05
abra_1564	GlnA glutamine synthetase, type I	2	7.2E-04	2	4.9E-04	2	9.4E-04
abra_3054	ntrC nitrogen regulation protein NR(I)	1	7.0E-05	0	ND	0	ND

[a] Number indicates how many replicates contained the identified protein

[b] Measure of relative abundance of each component based on spectral abundance

[c] ND = not detected in these cultures

Table 3. 5 Nitrogenase components and related proteins necessary for nitrogen fixation detected in both nitrogen fixing and optimal growth control Sp7 cultures

Gene Name	Description	Sp7- N2 fix		Sp7	
		No. Replicates[a]	Average NSAF[b]	No. Replicates[a]	Average NSAF[b]
abra_0572	nitrogen regulatory protein P-II	2	4.5E-03	2	5.7E-04
abra_1247	nifH nitrogenase iron protein	1	1.9E-04	0	ND
abra_1257	nitrogen fixation protein FixT	2	8.5E-04	0	ND
abra_3965	ptsN PTS IIA-like nitrogen-regulatory protein PtsN	2	6.5E-04	2	1.2E-03
abra_0535	glnD protein-P-II uridylyltransferase	2	4.0E-05	0	ND
abra_1564	GlnA glutamine synthetase, type I	2	2.5E-03	2	8.9E-04
abra_3054	ntrC nitrogen regulation protein NR(I)	0	ND	0	

[a] Number indicates how many replicates contained the identified protein

[b] Measure of relative abundance of each component based on spectral abundance

[c] ND = not detected in these cultures

Differentially expressed proteins

Direct comparisons of relative abundance of a given protein in the proteome were done through a direct comparison of NSAF values of a given protein in the nitrogen fixing proteome with the relative abundance of that same protein within the optimally grown culture (in presence of ammonium and with oxygen and so "non nitrogen fixing") proteome. Up-regulated fold change of individual proteins was determined by the formula $(NSAF)_{N2fix}/(NSAF)_c$ where $(NSAF)_{N2fix}$ is the NSAF value for a given protein under nitrogen fixing conditions, and $(NSAF)_c$ is the NSAF value for the corresponding protein in optimal growth control conditions. Levels of down-regulation were determined by $(NSAF)_c / (NSAF)_{N2fix}$. Levels of up- and down-regulation for proteins detected in Sp245 cultures are given in Tables 3.6 and 3.7, respectively, while those for Sp7 strain are given in Tables 3.8 and 3.9, respectively. Fold-change levels determined as above were tabulated for a direct comparison between Sp245 and Sp7, and results presented Tables 3.10 and 3.9 for up- and down-regulated proteins, respectively.

Discussion

Percentages of proteins detected in other studies of bacterial proteomes range from 24-34% of gene coding sequences detected in proteomic studies [25, 26, 30, 32, 66], slightly higher than the proteome coverage from bacterial cultures in our study. Further, the numbers of detected proteins in *R. palustris* control proteomes is higher in a 5-step MudPIT procedure than the numbers of identified proteins in these studies. This could be due to several different reasons. First, peptides from abundant ribosomal proteins eluted in every chromatographic step, thereby decreasing the apparent dynamic range of the instrument due to the inability to sample peptides from other proteins that are of lower

Table 3. 6 Fold change of proteins found to be up-regulated in Sp245 nitrogen fixing cultures in comparison to optimal growth Sp245 controls

Gene	Product	FuncCat	Avg Fold Change Upreg	Upreg N2Fix-Mo	Upreg N2Fix-V
abra_0588	Rubrerythrin	AA	6.1	6.1	6.1
abra_3154	UspA domain protein	AA	8.1	7.5	8.6
abra_2681	conserved hypothetical protein	BB	15.6	12.2	19.0
abra_4017	conserved hypothetical protein	BB	5.8	5.4	6.3
abra_0347	cytochrome c oxidase, cbb3-type, subunit II	C	4.7	5.1	4.4
abra_1181	molybdopterin oxidoreductase	C	5.7	6.3	5.1
abra_1191	H+transporting two-sector ATPase B/B' subunit	C	4.3	5.8	2.9
abra_4904	Pyridoxal 4-dehydrogenase	C	6.6	6.0	7.2
abra_0572	nitrogen regulatory protein P-II	E	8.5	9.0	8.0
abra_1798	Extracellular ligand-binding receptor	E	5.1	4.5	5.8
abra_1842	Extracellular ligand-binding receptor	E	5.1	ND[a]	5.0
abra_1868	cationic amino acid ABC transporter, periplasmic binding protein	E	8.8	8.5	9.1
abra_4656	putative branched-chain amino acid ABC transport system substrate-binding protein	E	4.9	4.0	5.9
abra_2710	PQQ-dependent dehydrogenase, methanol/ethanol family	G	8.0	9.1	7.0
abra_2756	elongation factor G domain IV	J	10.8	12.3	9.3
abra_0430	periplasmic solute binding protein	P	4.2	3.0	5.2
abra_0617	Cytochrome-c peroxidase	P	8.1	6.9	9.1
abra_2041	3'(2'),5'-bisphosphate nucleotidase	P	4.1	5.2	3.0
abra_2753	Ferritin Dps family protein	P	4.6	5.3	3.9
abra_1257	nitrogen fixation protein FixT	Q	5.3	5.2	5.5
abra_6359	acetoacetyl-CoA reductase	Q	5.3	5.6	5.1
abra_1884	TRAP transporter solute receptor, TAXI family	R	6.9	5.3	8.6
abra_0139	conserved hypothetical protein	S	46.2	49.6	42.8
abra_0141	conserved hypothetical protein	S	29.9	33.6	26.3
abra_1003	protein of unknown function DUF1476	S	7.3	4.4	10.1
abra_1380	protein of unknown function DUF1013	S	5.9	4.5	7.3

[a] Not detected to be upregulated in nitrogen fixing cultures when compared to non nitrogen fixing control cultures

Table 3. 6 (cont) Fold change of proteins found to be upregulated in Sp245 nitrogen fixing cultures in comparison to optimal growth Sp245 controls

Gene	Product	FuncCat	Avg Fold Change Upreg	Upreg N2Fix-Mo	Upreg N2Fix-V
abra_1485	protein of unknown function DUF344	S	7.7	6.2	9.2
abra_3461	conserved hypothetical protein	S	5.8	5.9	5.7
abra_4625	conserved hypothetical protein	S	4.7	5.9	3.6
abra_0303	CBS domain containing protein	T	12.7	11.5	13.9

[a] Not detected to be downregulated in nitrogen fixing cultures when compared to non nitrogen fixing control cultures

Table 3. 7 Fold change of proteins identified in Sp245 nitrogen fixing cultures to be down-regulated over Sp245 optimal growth control cultures

				Downreg	Downreg
			Avg Fold Change	Sp245	Sp245
Gene	Product	FuncCat	Downregulated	N2Fix_Mo	N2Fix_V
abra_0403	helix-turn-helix domain protein	AA	9.7	ND[a]	9.7
abra_2568	helix-turn-helix domain protein	AA	7.3	9.6	5.1
abra_0639	hypothetical protein	BB	6.9	ND	6.9
abra_1792	hypothetical protein	BB	6.0	3.4	6.0
abra_3834	conserved hypothetical protein	BB	12.2	ND	12.2
abra_6669	periplasmic nitrate reductase, large subunit	C	11.2	11.2	ND
abra_4719	Glu/Leu/Phe/Val dehydrogenase	E	4.8	5.3	4.3
abra_4885	Substrate-binding region of ABC-type glycine betaine transport system	E	14.0	ND	14.0
abra_1004	adenylosuccinate lyase	F	6.2	7.1	5.3
abra_1950	glyceraldehyde 3-phosphate dehydrogenase	G	5.9	6.2	5.6
abra_1731	thiamine biosynthesis protein ThiC	H	10.0	ND	10.0
abra_2597	Polyprenyl synthetase	H	4.0	5.9	2.1
abra_3047	lipoic acid synthetase	H	5.3	5.3	ND
abra_3999	adenosylhomocysteinase	H	5.3	3.3	7.3
abra_0152	sun protein	J	4.3	3.5	5.2
abra_0364	ribosomal protein S20	J	5.6	4.9	6.3
abra_0947	ribosomal protein L21	J	5.1	5.5	4.8
abra_4091	translation initiation factor IF-1	J	4.0	2.4	5.7
abra_4317	glutamyl-tRNA(Gln) amidotransferase, C subunit	J	4.6	5.5	3.7
abra_2337	RNA polymerase sigma factor RpoD	K	6.2	3.7	8.5
abra_3468	Cold-shock protein DNA-binding	K	3.8	5.2	2.3
abra_0693	DEAD/DEAH box helicase domain protein	L	9.8	ND	9.8
abra_3041	DEAD/DEAH box helicase domain protein	L	7.5	9.8	5.1
abra_0563	chaperone protein DnaJ	O	8.1	7.6	8.6
abra_2258	Redoxin domain protein	O	6.2	5.7	6.6
abra_3139	FeS assembly protein SufD	O	4.9	4.4	5.5

[a] Not detected to be downregulated in nitrogen fixing cultures when compared to non nitrogen fixing control cultures

Table 3.7 (cont) Fold change of proteins identified in Sp245 nitrogen fixing cultures to be down-regulated over optimal growth control cultures

Gene	Product	FuncCat	Avg Fold Change Downregulated	Downreg Sp245 N2Fix_Mo	Downreg Sp245 N2Fix_V
abra_0350	SAM-dependent methyltransferase	R	5.2	3.6	6.8
abra_3520	cobalamin biosynthesis protein CobW	R	5.0	5.1	5.0
abra_5399	amidohydrolase	R	9.4	9.4	ND[a]
abra_3007	signal recognition particle-docking protein FtsY	U	4.6	5.6	3.6

[a] Not detected to be downregulated in nitrogen fixing cultures when compared to non nitrogen fixing control cultures

Table 3. 8 Fold change of proteins up-regulated in Sp7 nitrogen fixing cultures in comparison to Sp7 optimal growth control cultures

Name	Description	UpregSp7-N2Fix [a]	FuncCat
abra_2702	cytochrome c class I	14.8	C
abra_5376	cytochrome c class I	6.5	C
abra_0362	extracellular solute-binding protein family 1	8.9	E
abra_0572	nitrogen regulatory protein P-II	7.9	E
abra_3314	extracellular solute-binding protein family 3	9.3	E
abra_3517	Extracellular ligand-binding receptor	6.0	E
abra_3555	Extracellular ligand-binding receptor	7.1	E
abra_4115	ABC transporter related	8.9	E
abra_4656	putative branched-chain amino acid ABC transport system substrate-binding protein	6.5	E
abra_6441	4-phytase	26.2	E
abra_3820	extracellular solute-binding protein family 1	6.0	G
abra_4586	Methyltransferase type 11	13.7	H
abra_4652	acyl-CoA dehydrogenase domain protein (lipid metabolism)	12.0	I
abra_1829	Lytic transglycosylase catalytic	5.9	M
abra_0617	Cytochrome-c peroxidase	13.4	P
abra_2519	NMT1/THI5 like domain protein	5.5	P
abra_2753	Ferritin Dps family protein	9.2	P
abra_4287	sulfate ABC transporter, periplasmic sulfate-binding protein	7.7	P
abra_3090	TRAP dicarboxylate transporter- DctP subunit	9.1	Q
abra_4158	TRAP dicarboxylate transporter- DctP subunit	13.1	Q
abra_6186	TRAP dicarboxylate transporter- DctP subunit	16.4	Q
abra_6359	acetoacetyl-CoA reductase	5.6	Q
abra_4265	beta-lactamase domain protein	6.9	R
abra_0139	conserved hypothetical protein	9.2	S
abra_1485	protein of unknown function DUF344	9.9	S
abra_0303	CBS domain containing protein	14.8	T
abra_0588	Rubrerythrin	5.2	AA
abra_2770	serine/threonine protein kinase	6.5	AA
abra_3660	Tetratricopeptide TPR_2 repeat protein	5.3	AA

a Numbers represent fold change of upregulation in nitrogen fixing cultures

Table 3. 9 Fold Change of Proteins down-regulated in Sp7 and Sp245 nitrogen fixing cultures in comparison to optimal growth control cultures

			Fold Change	Downreg	Downreg
Gene	Product	FuncCat	Downreg in N2Fix	Sp7N2Fix	Sp245N2Fix_Mo
abra_2568	helix-turn-helix domain protein	AA	9.5	C[a]	9.5
abra_6669	periplasmic nitrate reductase, large subunit	C	11.2	2.6	11.2
abra_2185	phosphoadenosine phosphosulfate reductase	E	7.2	7.2	C[a]
abra_4719	Glu/Leu/Phe/Val dehydrogenase	E	5.3	C[a]	5.4
abra_1004	adenylosuccinate lyase	F	7.2	C[a]	7.2
abra_1950	glyceraldehyde 3-phosphate dehydrogenase	G	6.2	C[a]	6.2
abra_2597	Polyprenyl synthetase	H	5.9	C[a]	5.9
abra_3047	lipoic acid synthetase	H	5.3	2.6	5.3
abra_0947	ribosomal protein L21	J	5.4	5.2	5.5
abra_4317	glutamyl-tRNA(Gln) amidotransferase, C subunit	J	5.5	ND[b]	5.5
abra_3468	Cold-shock protein DNA-binding	K	5.2	C[a]	5.2
abra_3041	DEAD/DEAH box helicase domain protein	L	9.8	2.0	9.8
abra_0563	chaperone protein DnaJ	O	7.5	4.2	7.5
abra_2258	Redoxin domain protein	O	5.7	3.8	5.7
abra_4879	OsmC family protein	O	7.8	C[a]	7.8
abra_0593	ThiJ/PfpI domain protein	R	9.5	9.5	4.8
abra_3520	cobalamin biosynthesis protein CobW	R	5.1	2.3	5.1
abra_5399	amidohydrolase	R	9.4	ND[b]	9.4
abra_3007	signal recognition particle-docking protein FtsY	U	5.6	2.2	5.6

[a] C= Control cultures only, protein was not detected in nitrogen fixing cultures

[b] ND = Protein was not detected in nitrogen fixing cultures

Table 3. 10 Fold Change of common proteins up-regulated in Sp7 and Sp245 molybdate nitrogen fixing cultures in comparison to each respective optimal growth control culture

Name	Description	FuncCat	Avg Fold Change Upreg N2 Fix	Upreg Sp245N2fix	Upreg Sp7N2Fix
abra_0588	Rubrerythrin	AA	5.7	6.2	5.2
abra_2770	serine/threonine protein kinase	AA	6.5	2.0	6.5
abra_3154	UspA domain protein	AA	7.5	7.5	4.8
abra_2681	conserved hypothetical protein	BB	12.2	12.1	NF[a]
abra_4017	conserved hypothetical protein	BB	5.4	5.4	ND[b]
abra_0347	cytochrome c oxidase, cbb3-type, subunit II	C	5.1	5.1	NF[a]
abra_1181	molybdopterin oxidoreductase	C	6.3	6.3	2.2
abra_1191	H+transporting two-sector ATPase B/B' subunit	C	5.7	5.7	NF[a]
abra_2702	cytochrome c class I	C	14.8	NF[a]	14.8
abra_4904	Pyridoxal 4-dehydrogenase	C	6.0	6.0	4.8
abra_5376	cytochrome c class I	C	6.5	NF[a]	6.5
abra_0362	extracellular solute-binding protein family 1	E	8.9	NF[a]	8.9
abra_0572	nitrogen regulatory protein P-II	E	8.4	9.0	7.9
abra_1868	cationic amino acid ABC transporter, periplasmic binding protein	E	8.5	8.5	2.5
abra_3314	extracellular solute-binding protein family 3	E	9.3	NF[a]	9.4
abra_3517	Extracellular ligand-binding receptor	E	6.1	3.3	6.0
abra_4115	ABC transporter related	E	8.9	4.3	8.9
abra_4656	putative branched-chain amino acid ABC transport system substrate-binding protein	E	6.5	4.0	6.5
abra_6441	4-phytase	E	26.2	NF[a]	26.2
abra_3555	Extracellular ligand-binding receptor	E	7.2	NF[a]	7.1
abra_2710	PQQ-dependent dehydrogenase, methanol/ethanol family	G	9.2	9.2	NF[a]
abra_3820	extracellular solute-binding protein family 1	G	6.0	ND[b]	6.0
abra_4586	Methyltransferase type 11	H	13.7	NF[a]	13.7
abra_4652	acyl-CoA dehydrogenase domain protein	I	12.0	ND[b]	12.0

a NF = Protein was detected only in nitrogen fixing cultures
b ND = Protein was not detected to be upregulated in nitrogen fixing cultures

Table 3.10 (cont) Fold Change of common proteins up-regulated in Sp7 and Sp245 molybdate nitrogen fixing cultures in comparison to each respective optimal growth control culture

Name	Description	FuncCat	Avg Fold Change Upreg N2 Fix	Upreg Sp245N2fix	Upreg Sp7N2Fix
abra_2756	elongation factor G domain IV	J	12.3	12.3	NF[a]
abra_1829	Lytic transglycosylase catalytic	M	5.8	NF[a]	5.9
abra_2519	NMT1/THI5 like domain protein	P	5.5	NF[a]	5.5
abra_2753	Ferritin Dps family protein	P	7.2	5.3	9.2
abra_4287	sulfate ABC transporter, periplasmic sulfate-binding protein	P	7.7	2.4	7.7
abra_1257	nitrogen fixation protein FixT	Q	5.2	5.2	NF[a]
abra_3090	TRAP dicarboxylate transporter- DctP subunit	Q	9.1	NF[a]	9.1
abra_4158	TRAP dicarboxylate transporter- DctP subunit	Q	13.1	NF[a]	13.1
abra_6186	TRAP dicarboxylate transporter- DctP subunit	Q	16.4	ND[b]	16.4
abra_6359	acetoacetyl-CoA reductase	Q	5.6	5.6	5.6
abra_1884	TRAP transporter solute receptor, TAXI family	R	5.3	5.3	4.3
abra_4265	beta-lactamase domain protein	R	6.9	NF[a]	6.9
abra_0139	conserved hypothetical protein	S	29.4	49.6	9.2
abra_0141	conserved hypothetical protein	S	33.6	33.6	NF[a]
abra_1485	protein of unknown function DUF344	S	8.1	6.2	9.9
abra_3461	conserved hypothetical protein	S	5.9	5.9	NF[a]
abra_4625	conserved hypothetical protein	S	5.9	5.9	2.9
abra_0303	CBS domain containing protein	T	13.2	11.5	14.8

a NF = Protein was detected only in nitrogen fixing cultures

b ND = Protein was not detected to be upregulated in nitrogen fixing cultures abundance. Second, a number of phage sequences are known to be integrated into the genome [128], and perhaps these are called as candidate genes but not expressed under these growth conditions. Third, investigation of the genome sequence reveals a number of redundant genes in the genome. Perhaps only one copy needs to be expressed, or one version is expressed in one growth state while a different version is expressed in a different growth state, and perhaps still others are expressed under stress conditions. Further, strain Sp7 identifications may be low due to the use of a genome that is not specific for that strain. However, BLASTx comparisons of the Sp7 strain sequenced genes available in Genbank with the Sp245 genome indicate that the two strains possess genes with a high degree of similarity with more than 95% of the available gene sequences for Sp7 showing greater than 90% similarity to Sp245 genes. This is partly expected since these are two strains of the same species. Only 1 gene (Sp7 *nirS*) did not have a homolog in the Sp245 sequence. Two others (nitrogenase structural components, alpha chain and beta chain of the molybdate nitrogenase) showed 27% and 32% similarity to those found within the Sp245 genome. Therefore, although some proteins may not be detected in the Sp7 strain due to sequence differences in the Sp245 genome, it is likely to be a small fraction of these, as reflected in the slight differences in the numbers of identifications and the percentage of the proteome detected between the two strains. Further, the proteins that are detected and identified in Sp7 in the approach used here are likely to be accurate since there is a high degree of sequence similarity between the Sp245 strain and Sp7 strain sequences.

Sp245 proteome evaluation

Figure 3.1 illustrates basic shifts in a number of functional categories when cultures are grown under nitrogen fixing conditions, most notably a decrease in percentages of proteins detected from functional category J representing those proteins involved in translation, ribosomal structure and biogenesis. Although the actual numbers of proteins detected in functional category J under each growth condition are similar (Table 3.2), the decrease in proteome percentage calculated by ΣNSAF values of proteins in this category (~37% in optimal growth control cells versus ~ 26% in nitrogen fixing cells) suggests a lower abundance of these proteins in the overall protein expression of the cells, which in turn suggests a slower protein synthesis and/or turnover in nitrogen fixing cells. The entire expression profile of the detected proteome as represented by functional categories suggests a basic change in metabolism in cultures that are grown under conditions of nitrogen fixation (Table 3.2 and Figure 3.1). As discussed in Chapter 1, nitrogen fixation requires a substantial amount of energy expenditure, with concomitant adjustments in metabolism. An additional strategy to conserve energy includes transport of molecules in order to glean all available nitrogen from the surroundings as well as to provide additional metals needed for nitrogenase structural components [141]. The necessity of transporting molecules combined with the need for more energy in cells that have no fixed nitrogen available [141] is reflected in the proteome expression profile in the following ways. Under nitrogen fixation conditions, a higher percentage of proteins were detected in functional categories involved with amino acid (category E: ~11% versus ~6% for control cells) and inorganic ion transport and metabolism (category P: ~6% versus 3% for optimal growth control cells), and in energy production and conversion (category C, ~8.5% versus ~6% for cells grown with ammonium) compared to protein expression levels in cells grown with ammonium. These observations taken together support the assumption that cells adjust their physiology and metabolism when grown under different conditions.

Nitrogen fixation proteins

Nitrogen fixation is an energetically costly endeavor for the cell, requiring 16 ATP and 8 NADH for every molecule of nitrogen converted to 2 molecules of ammonia [142]. As mentioned in Chapter 1, conversion of nitrogen to ammonia is accomplished by the nitrogenase enzyme, a complex containing two proteins that are encoded by the *nifHDK* operon [96]. Reduction of nitrogen is accomplished by dinitrogenase, which is encoded by *nifDK*, and contains an iron-molybdenum (Fe-Mo) co-factor. Dinitrogenase reductase protein is encoded by *nifH*, contains an Fe co-factor, and serves to donate electrons to dinitrogenase [96]. Transcription of genes from the *nifHDK* operon is tightly controlled by several layers of regulation, all dependent upon the nitrogen nutritional status of the cell and nitrogen availability in the environment. The promoter site driving transcription of the nitrogenase enzymes is an *rpoN* promoter site, and transcription requires the presence of NtrA (RpoN) and NifA, a transcription factor responsible for *nif* gene transcription [91]. NifA is constitutively expressed in *A. brasilense*, and levels of its activity are regulated by posttranslational modification of the protein and are mediated by both ammonia and oxygen levels: high levels of ammonium and oxygen act together to repress NifA activity [143]. In growth conditions in which nitrogen availability is low and oxygen levels are less than 2%, NifA is active, and nitrogenase structural components are transcribed. All structural components are found to be expressed in nitrogen fixing cultures (Table 3.4) while being absent in nitrogen-replete oxygen-optimized cultures, consistent with the expression of the *nif* operon in environments with high carbon to nitrogen ratios.

A number of bacterial species including *Azotobacter, Anabaena* [144], and *Rhodopseudomonas* [105] can synthesize an alternate nitrogenase which uses vanadium as a co-factor under conditions where molybdate concentrations are low and vanadium is present. Examination of the genome sequence of Sp245 reveals the presence of a *vnf* operon encoding the alpha, beta and delta chains of vanadium nitrogenase. However, the nitrogen fixing cultures supplemented with vanadium did not show evidence of expression of a vanadium nitrogenase under the growth conditions tested in this work. Vanadium nitrogenase contains an iron (Fe) protein with 89% sequence similarity to molybdenum Fe-protein. The iron-vanadium (FeV) protein of vanadium nitrogenase is encoded by *vnfDKG* and consists of 3 different subunits, alpha, beta and delta [144]. BLASTp [52] alignment between the vanadium nitrogenase alpha subunit (abra_6100) and the molybdate nitrogenase alpha subunit (abra_1248) shows 37% sequence identity between the two alpha chains, with gaps and alignments as shown in Figure 3.3. The detected peptide sequences in both molybdate and vanadium nitrogen fixing cultures for the molybdate nitrogenase alpha chain are in red text and underlined in Figure 3.3. An additional peptide N-terminal to the alignment shown was also detected in both vanadium and molybdate nitrogen fixing cultures.

Although cultures supplemented with vanadium show evidence of nitrogen fixation, only Mo-nitrogenase was detected in cultures supplemented with vanadium

```
AB1246  24   EKSRKRRAKH--LNVLEAEAKDCGVKSNIKSIPGVMTIRGCAYAGSKGVVWGPIKDMIHI  81
             +K    R KH  L   + + +D    SN  +IPG ++ RGCA+ G+K V+ G +KD I +
AB6100  9    DKDIPEREKHIYLKAPDEDTRDYLPLSNAATIPGTLSERGCAFCGAKLVIGGVLKDTIQM  68

AB1246  82   SHGPVGCGYYSWSGRRNYYVGDTGVDSWGTMHF------TSDFQEKDIVFGGDKKLHKVI  135
             HGP+GC Y +W  +R  Y  D G       HF      ++D +E  IVFGG+K+L K +
AB6100  69   IHGPLGCAYDTWHTKR--YPTDNG-------HFNMKYVWSTDMKESHIVFGGEKRLEKSM  119

AB1246  136  EEINELFPLVNGISIQSECPIGLIGDDIEAVARAKSEELGKP---VVPVRCEGFRGVSQS  192
             E  +  P V  + + + CP  LIGDDI+AVA+   +   +P   V  V C GF GVSQS
AB6100  120  HEAFDAMPDVKRMIVYTTCPTALIGDDIKAVAKKVMD--ARPDVDVFTVECPGFSGVSQS  177

AB1246  193  LGHHIANDVIRDWIFEKTEPKEGFVSTPYDVTIIGDYNIGGDAWASRILLEEIGLRVIAQ  252
             GHH+ N     WI EK    E  +++PY +  IGD+NI GD    +   + +G++VIA
AB6100  178  KGHHVLN---IGWINEKVGTLEPEITSPYTMNFIGDFNIQGDTQLLQTYWDRLGIQVIAH  234

AB1246  253  WSGDGTLAELENTPKAKVNLIHCYRSMNYIARHMEEKFGIPWMEYNFFGPSQIAESLRKI  312
             ++G+GT  +L    +A++N+++C RS  YIA  +++++GIP ++ + +G + +AE +RKI
AB6100  235  FTGNGTYDDLRKMHRAQLNVVNCARSSGYIANELKKQYGIPRLDIDSWGFNYMAEGIRKI  294

AB1246  313  AALFDDTIKENAEKVIAKYQPMVDAVIAKFKPRLEGKKVMIYVGGLRPRHVVDAYH-DLG  371
              A F   I+E  E +IA+   +    +  +K RL+G ++ I+ GG R  H   +   DLG
AB6100  295  CAFFG--IEEKGEALIAEEYALWKPKLDWYKERLKGTRMAIWTGGPRLWHWTKSVEDDLG  352

AB1246  372  MEIVGTGYEFAHNDDYQRTQHYVKEGTLIYDDVTAFELEKFVEVMRPDLVASGIKEKYVF  431
             +E+V    +F H +D+++     +EGT   DD    E  + +++++PD++ +G +   +
AB6100  353  VEVVAMSSKFGHEEDFEKVIARGREGTYYIDDGNELEFFEIIDLVKPDVIFTGPRVGELV  412

AB1246  432  QKMGLPFRQMHSWDYSGPYHGYDGFAIFARDMDLAINNPV  471
             +K  +P+   H++ ++GPY G++GF   ARDM  A+NNP+
AB6100  413  KKQHIPYVNGHAY-HNGPYMGFEGFVNLARDMYNAVNNPL  451
```

Figure 3. 3 BLASTp alignment of alpha subunits of molybdate nitrogenase (AB1246) and vanadium nitrogenase (AB6100)

BLAST alignment of molybdenum nitrogenase alpha chain (abra_1246) and vanadium nitrogenase alpha chain (abra_6100). There is 35% sequence identity between the alpha subunits of molybdate and vanadium nitrogenases. The peptide sequences of abra_1246 identified in Sp245 cultures are in red bold underlined letters. An additional peptide upstream of this alignment was also detected in all Sp245 nitrogen fixing cultures.

trichloride. No peptides from vanadium nitrogenase (V-nitrogenase) were identified in vanadium nitrogen-fixing cultures, suggesting that the V-nitrogenase is not expressed or is expressed only at very low levels under the conditions of these experiments. This could be due to several reasons. First, V-nitrogenase is often hierarchically expressed, and will not be transcribed in cultures where as little as 2 nM molybdate is present [103]. If Sp245 V-nitrogenase is hierarchically expressed, then even small amounts of molybdate would affect its expression. Since metals were not chelated prior to growth, it is possible that a high enough concentration of molybdate was present in the water with which the growth media was made to ensure repression of the V-nitrogenase. V-nitrogenase has low efficiency in converting nitrogen, and makes small amounts of hydrazine in addition to ammonia, but is more effective at lower temperatures [144]. Because of this lower efficiency, it is possible that once the culture became molybdate-limited it began to express vanadium nitrogenase but the amount was below the level of detection in these experiments. Alternatively, the V-nitrogenase could be expressed at a low level simultaneously with the Mo-nitrogenase, but not detected due to very low abundance.

Growth and energy production and conversion

Comparison of the most abundant proteins and the degree of up-regulation in nitrogen fixing cultures with those of cultures grown with ammonium suggests a shift in basic metabolism in nitrogen fixing cells. Investigation of the most abundant proteins in each growth condition listed in order by NSAF values (found in attached spreadsheets Sp245N2FixData and Sp7N2FixData) suggests rapid growth in cultures grown with ammonium. In these cultures, the most abundant proteins are heavily populated with a number of different ribosomal proteins, translation elongation factors and chaperonins. Additionally, DNA and RNA binding proteins are prevalent. Together, these data suggest that transcription of DNA and translation of mRNA into proteins is a high priority under these conditions and also that cells are growing rapidly, synthesizing biomass and/or with an active protein turnover. By comparison, cells grown under nitrogen fixing conditions include a wider variety of proteins and a more diverse representation of functional categories among their most abundant proteins. Although ribosomal proteins, translation elongation factors, chaperonin proteins, and DNA and RNA binding proteins are found among the most abundant proteins in nitrogen fixing cultures, proteins in these functional categories are not the clear majority. Instead, down-regulation of greater than 5-fold (Table 3.7) is noted for DNA binding proteins (abra_0403, 2568, 0693, 3041, 3468), ribosomal proteins (abra_0364, 0947), chaperone protein DnaJ (abra_0563), and RNA polymerase sigma factor RpoD (abra_2337). Cell division proteins, including MinC, MinD, FtsA, FtsK and FtsZ, are all down-regulated by 2-fold or greater (found in attached spreadsheet 2fold_DownN2Fix), suggesting slower growth and cell division. Further, a number of ribosomal proteins, chaperones and tRNA synthetases are down-regulated by 2-fold or greater (found in attached spreadsheet 2fold_DownN2Fix), indicating a slower protein synthesis in nitrogen fixing cells. These data suggest that while nitrogen fixing cells are actively growing, their growth is slower, biomass synthesis is reduced and protein turnover less active, consistent with the observation that cells grow slower and to lower cell density under conditions of nitrogen fixation [140].

As discussed in Chapter 1, nitrogen fixation is an energetically expensive process, and nitrogen fixing cells must address this need for more energy [141]. A number of cytochrome c proteins are also found among the most abundant proteins in cells grown under nitrogen fixation conditions, consistent with the need for more energy in order to fix nitrogen. Three cytochrome c proteins (abra_5376, 4904, and 3146) are found to be up-regulated by greater than 2-fold (found in attached table 2fold_UpN2Fix), while cytochrome c oxidase (abra_0347) is up regulated by greater than 5-fold (Table 3.6) with a concomitant increase in sequence coverage to 50%. Additional energy requirements are met through up-regulation of ATP synthase (abra_1191, 5-fold and abra_0929, 2-fold) and PQQ-dependent dehydrogenase methanol/ethanol family (abra_2710). Pyrrolo-quinoline-quinone (PQQ)-dependent dehydrogenase is a periplasmic quinoprotein that oxidizes methanol or ethanol to formaldehyde, passing the electrons derived from this oxidation to soluble cytochrome cL. Additional reducing equivalents are provided by 2-fold up-regulation (found in attached table 2fold_DownN2Fix) of citric acid cycle enzymes, malate dehydrogenase (abra_2984 and 0339), and succinate dehydrogenase (abra_2959).

Sensing energy and nitrogen status of the cell becomes very important when cells are grown under nitrogen fixing conditions [141]. Nitrogen regulatory P-II proteins serve as a sensor for nitrogen status in the cell, and are involved in regulatory control of nitrogen metabolism on several different levels [91, 96, 97]. Abra_0572, annotated as nitrogen regulatory protein P-II, was identified with over 80% sequence coverage and an up-regulation of 8.5-fold in all nitrogen fixing cultures (Table 3.6). Additionally, all nitrogen fixing cultures show substantial up-regulation of abra_0303, a cystathionine-beta-synthase (CBS) domain containing protein, identified with over 80% sequence coverage in all nitrogen fixing cultures. CBS domains are small intracellular modules that pair together to form stable globular domains and are often combined with a variety of other protein domains. They are known to bind ATP, AMP, and S-adenosylmethionine (SAM), and as such can function as energy sensors or intracellular sensors of the metabolic status of the cell [145], although they can also function as sensors for osmolarity as well [146]. A BLASTp [134] search against the NCBI non-redundant protein database revealed a 38% similarity to signal transduction protein from *Rhodospirillum rubrum* ATCC 11170 (YP 428825.1). Since abra_0303 is up-regulated by greater than 11.5-fold in all nitrogen fixing cultures, it is likely to be important in sensing the status of energy or intracellular metabolites within the cell.

Several proteins involved in metal homeostasis or nucleotide metabolism are found to be significantly up-regulated in all Sp245 nitrogen fixing cells (Table 3.6). For instance, all Sp245 cultures grown under nitrogen-fixation conditions show up-regulation of abra_0588 (rubrerythrin), abra_2753 (ferritin-Dps family protein) and abra_1485 (protein of unknown function DUF344). Rubrerythrin, up-regulated by 6-fold under nitrogen fixing conditions, is a protein with a ferritin-like fold proposed to play a role in metal ion binding. It belongs to the same family as Ferritin-Dps (DNA protection during starvation protein), which is also up-regulated in Sp245 nitrogen fixing cultures, and is also involved in non-heam storage of iron. Abra_1485 contains a domain structure similar to nucleoside triphosphate hydrolases containing a P-loop, and members of this family are known to be involved in purine and pyrimidine metabolism [147]. It further shows sequence similarities of greater than 65% to a number of hypothetical proteins and polyphosphate kinases from other bacterial species.

Molecular transport and metabolism shifts under nitrogen fixing conditions

Also among the most abundant proteins in cells grown under nitrogen fixing conditions are a number of extracellular ligand binding domain proteins and molecular transport proteins. These proteins are likely involved in more efficient use of available nutrients such as amino acids or metals, consistent with expected changes in metabolism between these 2 conditions [148, 149]. Of particular interest in molecular transport are two conserved hypothetical proteins, abra_0139 and abra_0141, shown to be up-regulated by greater than 30-fold under conditions of nitrogen fixation in Sp245 strain (Table 3.6), suggesting they play major roles in the physiology of cells under these conditions. Sequence coverage of abra_0139 (379 amino acid length) increases from 23% in ammonium-replete cultures to 67% in nitrogen fixing cultures, while sequence coverage increases from 28% to 78% for abra_0141 (323 amino acids). The ORFS encoding abra_0139 and abra_0141 are not genetically linked and are not contained with an operon. They are located 1944 nucleotides from one another in the genome sequence, and are transcribed in opposite directions (Figure 3.4). Abra_0139 is flanked by a methyl-accepting chemotaxis protein (abra_0140) with a start site downstream 157 nucleotides and an operon containing dimethylmenaquinone methyltransferase (abra_0137) and an auxin efflux carrier (abra_0138) which ends 391 nucleotides upstream, while abra_0141 is preceded by a response regulator receiver modulated metal-dependent phosphohydrolase which ends 248 nucleotides upstream of its start site. The close proximity of an MCP to these hypothetical proteins suggests that both proteins may function in response to the same signaling pathway.

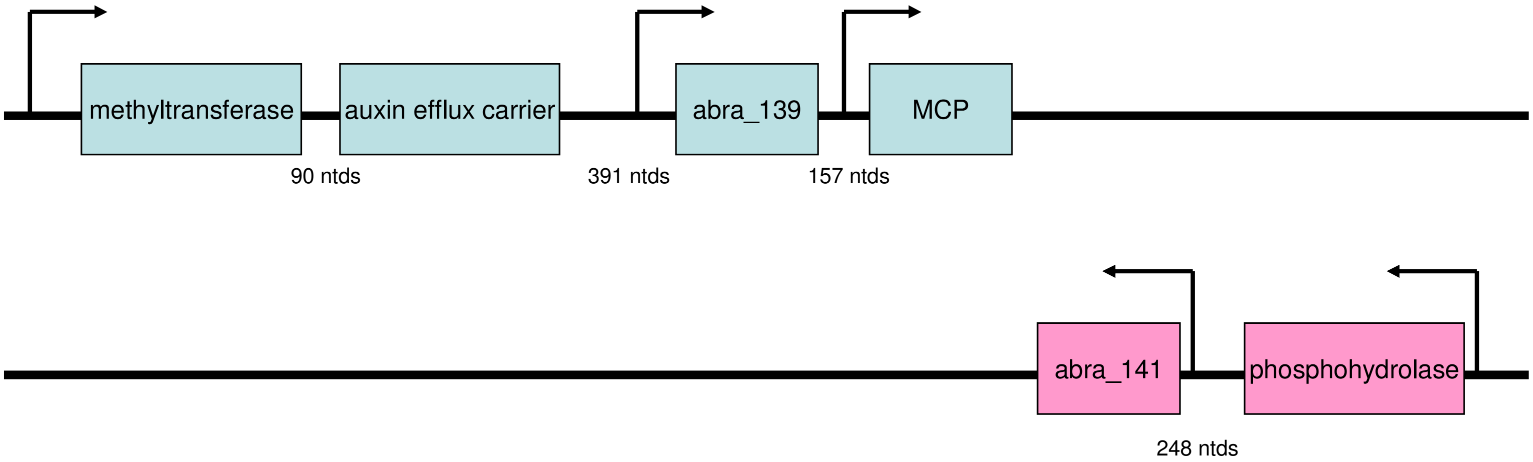
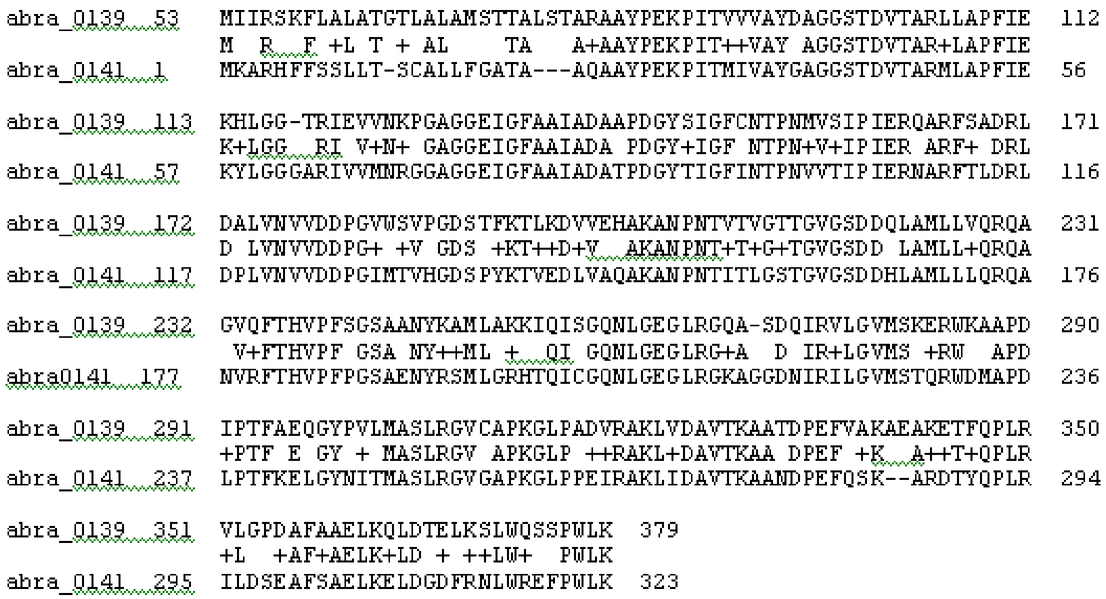


Figure 3. 4 Schematic representation of genomic region surrounding Abra_0139 and Abra_0141
Abra_0139 and abra_0141 are both hypothetical proteins containing BUG (Bordatella Uptake Gene) domains, thought to be important in solute binding and uptake. Sp245 nitrogen fixing cultures show a large degree of upregulation of these proteins. The close proximity of auxin efflux carrier to abra_0139 is intriguiging because of the known secretion of plant hormones in Azospirillum species associated with plants.

Analysis of abra_0139 and abra_0141 by BLASTp against the NCBI non-redundant protein database indicate that both proteins have a 51% and 54% sequence similarity to a putative secreted protein from *Bordetella petrii* and 49% and 50% sequence similarity to a probable extra-cytoplasmic solute receptor from *Ralstonia eutropha* H16, respectively. Aligning the amino-acid sequences of abra_0139 and abra_0141 using BLASTp [52] (Figure 3.5) indicates that they share 66% overall sequence similarity. Interestingly, abra_0141 possesses a periplasmic binding domain at its N-terminal end that is not found in abra_0139 while both proteins possesses a C-terminal BUG (*Bordetella* uptake gene) domain. BUG domains are widespread in bacteria, and seem especially abundant in the genomes of soil bacteria. Although their function has not been characterized, BUG domain proteins are suggested to be involved in nutrient transport [150]. Average up-regulation of 46-fold for abra_0139 and 30-fold for abra_0141 in Sp245 nitrogen fixing cultures suggests that both may be important for nutrient uptake or transport in nitrogen fixing cells.

Interestingly, periplasmic domains of ABC transporter systems and extracellular ligand-binding receptors are also found to be abundant in nitrogen fixing cultures. An extracellular ligand-binding receptor, abra_1798, is the most abundant protein found in all cultures grown under nitrogen fixation conditions, and is up-regulated by 4.5- and 5.8-fold in Sp245 molybdate and vanadium nitrogen fixing cultures, respectively (Table 3.6). This protein is not among the most abundant proteins in cultures grown with ammonium, suggesting an important function under nitrogen fixation conditions. A BLASTp [52, 134] search against the NCBI non-redundant database (http://blast.ncbi.nlm.nih.gov/) using the translated coding sequence revealed that abra_1798 shows 75% similarity to the



```
abra_0139  53    MIIRSKFLALATGTLALAMSTTALSTARAAYPEKPITVVVAYDAGGSTDVTARLLAPFIE  112
                 M  R  F +L T + AL    TA   A+AAYPEKPIT++VAY AGGSTDVTAR+LAPFIE
abra_0141  1     MKARHFFSSLLT-SCALLFGATA---AQAAYPEKPITMIVAYGAGGSTDVTARMLAPFIE  56

abra_0139  113   KHLGG-TRIEVVNKPGAGGEIGFAAIADAAPDGYSIGFCNTPNMVSIPIERQARFSADRL  171
                 K+LGG  RI V+N+ GAGGEIGFAAIADA PDGY+IGF NTPN+V+IPIER ARF+ DRL
abra_0141  57    KYLGGGARIVVMNRGGAGGEIGFAAIADATPDGYTIGFINTPNVVTIPIERNARFTLDRL  116

abra_0139  172   DALVNVVDDPGVWSVPGDSTFKTLKDVVEHAKANPNTVTVGTTGVGSDDQLAMLLVQRQA  231
                 D LVNVVDDPG+ +V GDS +KT++D+V  AKANPNT+T+G+TGVGSDD LAMLL+QRQA
abra_0141  117   DPLVNVVDDPGIMTVHGDSPYKTVEDLVAQAKANPNTITLGSTGVGSDDHLAMLLLQRQA  176

abra_0139  232   GVQFTHVPFSGSAANYKAMLAKKIQISGQNLGEGLRGQA-SDQIRVLGVMSKERWKAAPD  290
                  V+FTHVPF GSA NY++ML +  QI GQNLGEGLRG+A  D IR+LGVMS +RW  APD
abra0141  177    NVRFTHVPFPGSAENYRSMLGRHTQICGQNLGEGLRGKAGGDNIRILGVMSTQRWDMAPD  236

abra_0139  291   IPTFAEQGYPVLMASLRGVCAPKGLPADVRAKLVDAVTKAATDPEFVAKAEAKETFQPLR  350
                 +PTF E GY + MASLRGV APKGLP ++RAKL+DAVTKAA DPEF +K  A++T+QPLR
abra_0141  237   LPTFKELGYNITMASLRGVGAPKGLPPEIRAKLIDAVTKAANDPEFQSK--ARDTYQPLR  294

abra_0139  351   VLGPDAFAAELKQLDTELKSLWQSSPWLK  379
                 +L  +AF+AELK+LD + ++LW+  PWLK
abra_0141  295   ILDSEAFSAELKELDGDFRNLWREFPWLK  323
```

Figure 3. 5 BLASTp alignment of protein sequences for hypothetical proteins abra_0139 and abra_0141
Hypothetical proteins Abra_0139 and Abra_0141 both contain a BUG (Bordatella Uptake Gene) domain and are shown to be highly upregulated in nitrogen fixing conditions, suggesting they are important for nutrient uptake under nitrogen fixing conditions. BLASTp analysis of the two sequences shows 67% identity between the two with 81% positives.

periplasmic binding domain of ABC-type branched-chain amino acid transport systems of *Magnetospirillum magnetotacticum*, a phylogenetically close relative of *A. brasilense.* Searches against the NCBI conserved domain database [151] revealed that abra_1798 possesses conserved sequences matching to protein superfamily of periplasmic ligand-binding domains of ABC-type transporters with a domain structure of the LivK family. LivK is a periplasmic leucine-, isoleucine-, and valine-specific-binding protein involved in leucine transport, consistent with the suggestion that under conditions of nitrogen fixation, cells expend more of their energy in transporting molecules (Table 3.2).

An additional extracellular ligand-binding receptor (abra_5402) is also found within the most abundant proteins in all nitrogen fixing cultures. Analysis of closest homologs of abra_5402 in the NCBI microbial database using BLASTp [52] indicates that abra_5402 has high sequence similarity to the extracellular ligand-binding domain of ABC-type transporters, with the highest sequence similarity (47%) to an extracellular ligand-binding receptor from *Rhodospirillum rubrum* that has an unknown function. Further evidence of the need for increased molecular transport is seen in the abundance and up-regulation of cationic amino acid ABC-transporter (abra_1868) and putative branched-chain amino acid ABC transport system substrate binding protein (abra_4656). Abra_1868 is found among the most abundant proteins in cells grown under nitrogen fixing conditions and is up-regulated by 8.5-fold and 9.1-fold in molybdate and vanadium nitrogen fixing cultures, respectively (Table 3.6). Abra_4656 is up-regulated by 5-fold, and a number of additional ABC transporter binding proteins are up-regulated by greater than 2-fold (found in attached table 2fold_UpN2Fix). The abundance and variety of substrate binding proteins in nitrogen fixing cultures serves to emphasize the need for nitrogen fixing cultures to import nutrients rather than expend the energy to synthesize them. Taken together, these results suggest the need for increased molecular transport in nitrogen fixing growth state.

Sp245 nitrogen fixing cultures show an abundance of cupin 2 barrel domain protein (abra_3559) with a 2-fold increase in expression levels in nitrogen fixing Sp245 cultures (attached table 2fold_UpN2Fix). Cupin proteins are related to sugar storage proteins in plant seeds, and have been shown to have a number of different roles in bacteria including cell wall biosynthesis, secondary metabolite synthesis, stress response and sugar metabolism [152]. Interestingly, cupin domains have been found to be present in auxin binding proteins, where one of the five exons of auxin binding protein encodes a region that is thought to act as a receptor for IAA [152]. Of further interest, bacterial cupin domain proteins have been linked to oxalate metabolism. Oxalate, $C_2O_4^{2-}$, has been linked to either chelation and precipitation or to solubilization of metals in soils such that the metal-oxalate compound is made available [152]. Cupin domain proteins are not abundant in bacterial genomes [152], so the presence of this protein among the most abundantly expressed proteins in Sp245 nitrogen fixing cultures suggests that this protein plays an important role in this growth condition. Sp245 is known to produce more IAA under nitrogen fixing conditions [153], and *A. brasilense* is known to change its cell morphology under nitrogen fixing conditions [154], so the role of this particular protein could be varied, involved in transport of IAA, or cell wall remodeling or chelation of metals. This protein is not detected in Sp7 cultures at all, either because it does not exist within the Sp7 genome, or because it is either not expressed in Sp7 under these conditions or is expressed at a level below our detection limit. Further, the abundance of this protein in Sp245 cultures and non-detection of it in Sp7 cultures suggests that these two strains may employ different strategies for adaptation to nitrogen fixation.

Interestingly, both Sp245 molybdate and vanadium nitrogen fixing conditions show a 7-fold up-regulation of abra_3154, a UspA domain protein (Table 3.6). Sp7 nitrogen fixing cultures also show an up-regulation of abra_3154 of 4.7-fold, suggesting this protein is likely to be important under nitrogen fixing conditions. In *E. coli*, Universal Stress Protein (Usp) A is a small cytoplasmic bacterial protein which lends stress endurance to the cell, enhancing cell survival rate [155]. Increased expression is noted in cells that are under stress, including starvation for carbon, nitrogen or amino acids, where Usp family proteins protect against oxidative stress and mediate iron homeostasis and motility/adhesion in stress conditions [156]. Activity levels of Usp family proteins in *E. coli* have been shown to be dependent upon serine-threonine autophosphorylation [155]. In *A. brasilense*, a mutant lacking an Usp-family protein has been described. This Sp7 Tn5 mutant shows increased susceptibility to carbon deprivation, decreased EPS production and impaired flocculation, but does not exhibit sensitivity to oxidative stress agents [157]. Since flocculation has been proposed as a strategy used by the microaerophilic *Azospirillum* species to control oxygen diffusion through the cells [158], perhaps the up-regulation of UspA in nitrogen-fixing cultures plays a role in both EPS metabolism for nitrogen fixation energy and flocculation to protect nitrogenase integrity.

Differences between Sp245 molybdate versus vanadium nitrogen fixing cultures

Vanadium is a transition metal ubiquitous in soil and water, with particularly high concentrations in the oceans. It exists in a variety of oxidation states, from V(V) to V(II), with V(V) and V(IV) being the most commonly found in nature, and V(V), V(IV) and V(III) all shown to exist within different organisms. [159, 160] The anionic species, vanadate, or V(V), has the chemical formula of $H_2VO_4^-$. It is soluble and is the prevalent oxidation state found within the soil, most often found in oligomeric forms. The cationic species, vanadyl, or V(IV), has a chemical formula VO_2^+. It is insoluble in water although it is the prevalent form found in suspension in rivers. It is often complexed to other compounds, such as the porphyrins in petroleum, so that refinement of petroleum results in creation of vanadium sulfide [159]. The lowest oxidation state that can stay in solution is V(III), or V^{3+}, but must be coordinated to another ligand (such as iron transport protein, transferrin) or in a very acidic environment to do so. V(V) and V(IV) function as a redox couple with an E = 1.31V [160].

The biological effect of vanadium is dependent upon its oxidation state. Vanadate, V(V), can act as a phospho-mimetic, competing with phosphate HPO_4^{2-} while vanadyl, V(IV) can act as a transition metal ion to compete with other divalent metal ions [160]. Vanadate has been shown to inhibit the action of phosphatases, ribonucleases, and ATPases by serving as a transition state analog or by replacing the terminal phosphate in ATP, ADP or AMP and thus preventing further action by the enzymes [159, 160]. *In vitro* studies indicate that it can also inhibit the action of serine and cysteine proteases when in a sulfate salt form [161]. Vanadium has also been shown to simply act as a metal catalyst to catalyze the oxidation of sulfates [159]. *In vitro* studies using

Azotobacter Mo-nitrogenase suggest that vanadium can be used in the more traditionally expressed molybdate nitrogenase. In these studies, the reduced Fe protein was shown to reduce vanadate to vanadyl, and then use this form as a substitute for magnesium in the Mg-ATP bound to the MoFe protein [162]. Growth experiments with a number of species of bacteria indicate that though most bacteria may not be able to reduce vanadium, they can still tolerate more than trace amounts of vanadium, with most species and strains tested growing well in cultures with up to 3 g/L concentration of vanadate [163]. High concentrations, however, are toxic to some bacteria, such as *Azotobacter vinelandii*, which shows decreased growth in media with concentrations greater than1µM vanadium. [164] The effect of vanadium on *Azospirillum* species has not been studied.

Addition of vanadium to nitrogen fixing cultures did not appear to affect growth of the cultures. A similar amount of growth and clumping behavior was observed in both vanadium and molybdate cultures, with vanadium nitrogen fixing cultures growing to the same optical density as the molybdate nitrogen fixing cultures after 48 hours of growth. Microscopic investigation of the vanadium nitrogen fixing cultures showed actively motile cells, with no significant difference in size or shape of individual cells. In comparing patterns of protein expression of molybdate and vanadium nitrogen fixing cultures when compared to a control of optimal growth control cultures, expression profiles are similar across all functional categories (Figure 3.1 and Table 3.2). The list of proteins detected to be up-regulated in both nitrogen fixing cultures is the same (Table 3.6) with similar levels of up-regulation for each protein. This list includes a number of extracellular ligand binding proteins, hypothetical proteins and transporter proteins as well as cytochrome c oxidase and ATPase subunits, all of which were discussed earlier.

Similar levels of up-regulation of each protein within this list suggests similar needs for molecular transport and energy generation under both nitrogen fixation conditions.

One notable exception is extracellular ligand binding receptor abra_1842, which is detected to be up-regulated by 5-fold in vanadium nitrogen fixing cultures but not up-regulated at all in molybdate nitrogen fixing cultures (Table 3.6). Abra_1842 encodes a protein with 61% similarity to a LivK family protein from *Methylobacterium radiotolerans* JCM 2831, suggesting it is involved in branched chain amino acid transport. It is part of a putative operon that encodes components of an ATP-dependent branched chain amino acid transport system but other components of this system are not detected to be up regulated (Figure 3.6). Upstream of this operon by 332 nucleotides is Abra_1839, which encodes a MarR family protein. Proteins within the MarR family of transcriptional regulators share a conserved winged helix structure, and binding to DNA is mediated by the presence of anionic lipophilic ligands [165]. MarR dimers bind to specific palindromic DNA sequences in order to affect transcription from Mar (multiple antibiotic resistance) operons or oxidative stress operons or from operons involved in aromatic catabolism [165]. They function as environmental sensors, and some have been shown to repress transcription of those operons of which they are a part, or of operons that are close to them in the genome structure, while others have been shown to enhance transcription [165]. Interestingly, in vanadium nitrogen fixing cultures abra_1839 is not detected to be down-regulated, while it is down-regulated by 4-fold in molybdate nitrogen fixing cultures (attached spreadsheet 2Fold_DownN2fix). The proximity of this MarR homolog, abra_1839, to transport protein abra_1842, taken together with the 5-fold up-regulation of abra_1842 in vanadium cultures, while no up-regulation is found in

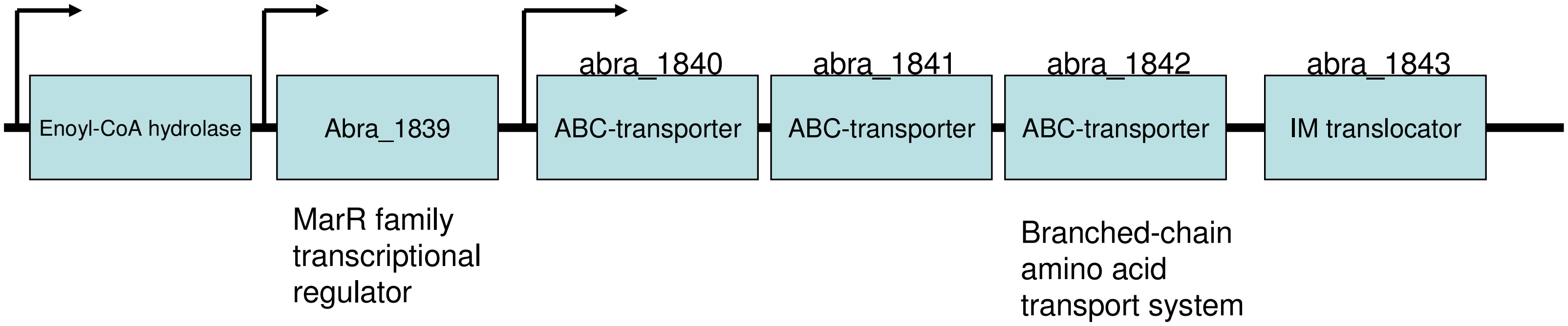


Figure 3. 6 Genomic region encompassing Abra_1839 MarR family transcriptional regulator
Abra_1839 shows similarity to MarR family transcriptional regulator. This protein is downregulated in molybdate nitrogen fixing cultures but not in vanadium nitrogen fixing cultures. Interestingly, this protein is in close proximity to a family of ABC-transporters involved in branched chain amino acid transport. MarR proteins have been shown to respond to environmental changes, and to affect transcription of genes in close proximity to them. Transporter Abra_1842 is upregulated in vanadium nitrogen fixing cultures but not in molybdate nitrogen fixing cultures, suggesting the possibility that abra_1839 may play a role in regulation of this operon of ABC transporters in response to environmental stimuli related to the presence of vanadium in the culture..

molybdate cultures where abra_1839 is down-regulated by 4-fold suggests the intriguing possibility that abra_1839 MarR protein is acting to affect transcription of transport protein abra_1842.

Twenty one common proteins are down-regulated in both molybdate and vanadium Sp245 nitrogen-fixing cultures (Table 3.7), although the levels of down-regulation are variable between cultures. Ribosomal proteins, chaperone proteins, DNA binding proteins, and synthesis proteins show similar levels of down-regulation in both cultures (Table 3.7), supporting the assumption that both molybdate and vanadium nitrogen fixing cultures grow slower with concomitant decreases in DNA and protein synthesis. Notable exceptions include helix-turn-helix binding domain protein encoded by abra_0403, and DEAD/DEAH box helicase domain protein encoded by abra_0693, both of which are down-regulated by more than 9-fold in vanadium cultures but are not detected in molybdate cultures. Additional hypothetical proteins abra_0639 and 3834 are also substantially down-regulated in vanadium cultures but not detected in molybdate cultures. The lack of detection of these proteins in molybdate cultures indicates they were either not expressed in molybdate nitrogen fixing cultures, or that they were expressed at very low levels such that they were below our limit of detection, which would imply a larger level of down-regulation of these proteins than in vanadium cultures.

Interestingly, molybdate nitrogen fixing cultures show substantial down-regulation of periplasmic nitrate reductase (abra_6669), but this protein is not detected in vanadium nitrogen fixing cultures (Table 3.7). Periplasmic nitrate reductase is a molybdenum-containing enzyme involved in the dissimilatory process of the nitrogen cycle by catalyzing the reduction of nitrate (NO^{3-}) to either nitrite (NO^{2-}) or ammonia [166].   Due to the lack of nitrate in the culture combined with the limited amount of molybdenum and the necessity of using molybdenum for Mo-nitrogenase, down-regulation of this protein is consistent with nitrogen fixation needs of the cell. The lack of detection of periplasmic nitrate reductase in vanadium nitrogen fixing cultures suggests that in a situation in which molybdate is severely limited either abra_6669 is not expressed at all or is expressed at levels below those we can detect. This data suggests that nitrogen cycling may be affected by the addition of vanadium to nitrogen fixing cultures, perhaps due to limited amounts of molybdate (which may in turn lead to lower expression levels of Mo-nitrogenase) or alternatively to less abundant or less efficient nitrogenase activity in the vanadium cultures, or due to interference of the vanadium with iron uptake pathways.

In addition, lipoic acid synthetase (abra_3047), which catalyzes the synthesis of lipoic acid, a coenzyme bound to pyruvate dehydrogenase and alpha-ketoglutarate dehydrogenase, is down-regulated by greater than 5-fold in molybdate nitrogen fixing cultures only (Table 3.7). In *Rhizobium etli*, lipoic acid synthetase has been shown to be down-regulated as it enters stationary phase when the cell energy needs are decreased [167]. Recent studies with a mutant of *B. subtilis* lacking a gene for lipoic acid synthetase have indicated that it also plays a role in branched chain fatty acid synthesis, and can affect the composition of the outer membrane lipopolysaccharides (LPS) [168]. The above data suggests that the 5-fold down-regulation of lipoic acid in Sp245 nitrogen fixing cells would result in a decreased output from the TCA cycle due to the reduction in alpha-ketoglutarate dehydrogenase and pyruvate dehydrogenase activity. Further, down-regulation of lipoic acid synthetase in nitrogen fixing cells may also contribute to outer membrane remodeling such that the LPS of the outer membrane contain a greater percentage of straight chain saturated fatty acids.

Sp7 Proteome evaluation

The proteome expression profiles of Sp7 nitrogen fixing and optimal growth control (nitrogen replete) cells are shown in Figure 3.2. As seen in Sp245 nitrogen fixing cultures, Sp7 also shows a shift in metabolism of nitrogen fixing cells. The most notable differences in protein abundances by functional category are a decrease in percentages of proteins detected from functional category J representing those proteins involved in translation, ribosomal structure and biogenesis, and increases in functional categories C and E representing energy production/conversion and amino acid transport/metabolism, respectively. The number of proteins detected in categories J (translation, ribosomal structure and biogenesis) and K (transcription) are lower in nitrogen fixing cells, while the numbers of proteins detected in categories C (energy production and conversion) and Q (secondary metabolite biosynthesis, transport and catabolism) are substantially higher (Table 3.3). The protein expression profile for Sp7 nitrogen fixing cells is indicative of slower growth and increased need for energy and molecular transport, just as in Sp245 nitrogen fixing cells. The decrease in protein abundances of those proteins involved in translation (category J, decrease from 41% to 25.5%) support the assumption that cells are making less protein when growing under nitrogen fixing conditions. Further, the most abundant proteins when listed in order by NSAF values (attached table Sp7N2Fix_Data) have far less ribosomal proteins under nitrogen fixing conditions, and contain a number of extracellular ligand-binding proteins and solute binding proteins. The increased need for molecular transport and energy in nitrogen fixing cells is also illustrated in Sp7 nitrogen fixing cells by the increase in percentages of protein abundance by ΣNSAF in those proteins contributing to energy production and conversion (category C, 9% versus 6% in nitrogen replete cells) and to amino acid transport and metabolism (category E, ~14% versus 6.5% in non nitrogen fixing cells).

Surprisingly, only one structural component of nitrogenase, abra_1247 nifH nitrogenase Fe-protein was detected in Sp7 nitrogen fixing cultures (Table 3.5). The *nif* operon from Sp7 strain was sequenced earlier [98, 169] and sequences entered into the NCBI database (http://www.ncbi.nlm.nih.gov/). BLASTx [134] comparison of the translated coding sequences of the Sp7 strain *nif* operon against the Sp245 genome structure showed remarkable levels of similarity (94 – 100%) for the majority of Sp7 translated coding sequences with those of the coding sequences of the *nif* operon in Sp245 genome. The notable exceptions to this included nitrogenase molybdate nitrogenase alpha and beta chains (abra_1246 and abra_1245, respectively). The Sp7 translated coding sequences showed only 32% and 27% similarity, respectively, to those same sequences from Sp245 genome. With the low degree of similarity in the alpha and beta nitrogenase subunits, it is therefore not surprising that no peptides were detected from these proteins in the Sp7 nitrogen fixing cultures. The Sp7 translated coding sequence for the NifH protein, however, showed 100% sequence similarity to that from the Sp245 genome, thus explaining the detection of the NifH protein in Sp7 cultures.

Growth and Energy Production

Nitrogen fixing Sp7 cells seem to have a somewhat different strategy for coping with nitrogen fixing conditions than do Sp245 nitrogen fixing cells. Although a large difference in percentage of proteins based on abundance by NSAF values detected that are involved in translation and transcription are noted, only one ribosomal protein L21 (abra_0947), is found to be down-regulated by more than 5-fold (Table 3.9). However, a number of other ribosomal proteins as well as RNA polymerase, sigma factor RpoD, and translation initiation and termination factors are down-regulated by 2 to 3-fold (attached spreadsheet 2fold_DownN2Fix), thus suggesting an overall decrease in protein production or turnover. One protein, abra_0593, containing a ThiJ/PfpI domain, is found to be down-regulated by 9.5-fold under nitrogen fixing conditions (Table 3.9). The ThiJ/PfpI domain is found in a diverse group of proteins involved in regulation of RNA-protein interaction, thiamine biosynthesis, or protease activity, and has also been found in transcriptional regulators. While abra_0593 is shown to be down-regulated by 4-fold in Sp245 nitrogen fixing cultures (attached spread sheet 2fold_DownN2fix), down-regulation of DNA binding proteins, chaperonins and cell division proteins that is prevalent in Sp245 nitrogen fixing cultures is not noticed in Sp7 nitrogen fixing cultures, perhaps suggesting either a different strategy for reducing growth rate or larger levels of down-regulation of these proteins in Sp7 cultures such that they are not detected at all under nitrogen fixing conditions.

Increased energy requirements are met in Sp7 nitrogen fixing cells through up-regulation of a number of respiratory chain components. ATP synthase subunits (abra_0925 and 0928) are among the most abundant proteins in Sp7 nitrogen fixing cultures. Additional ATP synthase subunits (abra_0927 and 0929) are shown to be up-regulated by greater than 2-fold (attached spreadsheet 2fold_UpN2Fix), suggesting Sp7 cells are using this ATP synthase (among the multiple ATP synthase subunits present in the Sp245 genome) to synthesize the large amount of ATP required for nitrogen fixation. The up-regulation of nearly 15-fold of cytochrome c (abra_2702) and 6.5-fold of cytochrome c (abra_5376), shown in Table 3.8, along with greater than 2-fold up-regulation of 2 electron-transferring-flavoprotein dehydrogenases (abra_3654 and 4624) and ubiquinol-cytochrome c reductase (abra_4336) suggests the importance of the electron transport chain in energy production in nitrogen fixing Sp7 cells. As in Sp245 nitrogen fixing cells, reducing equivalents are provided by 2-fold up-regulation of malate dehydrogenase (abra_2984) although no other citric acid cycle enzymes are shown to be up-regulated (attached spreadsheet 2fold_UpN2Fix).

Molecular transport and metabolism changes under nitrogen fixation

The strategy adopted by Sp7 nitrogen fixing cells for molecular transport includes a number of extracellular ligand-binding receptor proteins, extracellular solute binding proteins, ABC-transporters and DctP subunits of TRAP dicarboxylate transporters, all found within the most abundant proteins in Sp7 nitrogen fixing cultures (attached spreadsheet Sp7_N2FixData). Extracellular solute binding protein family 1 (abra_0362 and 3820) and family 3 (abra_3314) proteins are up-regulated by 8.9-, 9.3- and 6-fold, respectively, while extracellular ligand-binding receptor proteins (abra_3517 and 3555) are up-regulated by 6- and 7-fold, respectively (Table 3.8). An additional 7 extracellular solute binding proteins and 7 ligand-binding receptor proteins are up-regulated by more than 2-fold. Both abra_0362 and abra_3820 have LysR substrate binding domains, which are present in proteins involved in regulation of metabolism of carbohydrates and amino acids, often binding co-factors or co-inducers [170]. Abra_3314 has a domain structure similar to those found in penicillin binding proteins. Extracellular ligand-binding receptor, abra_1798, is among the most abundant protein found in Sp7 cultures grown under nitrogen-fixation conditions (attached spreadsheet Sp7_N2FixData), and is up-regulated by 4-fold (attached spreadsheet 2fold_UpN2fix). As discussed earlier, abra_1798 shows similarity to ABC-type transporters with a domain structure of the LivK family. BUG domain protein, abra_0139, (up-regulated by greater than 40-fold in Sp245 nitrogen fixing cultures), is up-regulated by only 9-fold in Sp7 cultures, while abra_0141 (up-regulated by greater than 25-fold in Sp245 nitrogen fixing cultures) is detected only in Sp7 nitrogen fixing cultures, but not in Sp7 non-nitrogen fixing cultures so a level of up-regulation could not be determined. Taken together, the above data suggests that Sp7 cells use a different, more diverse strategy for molecular transport than does Sp245, perhaps reflecting different physiological abilities to cope with the metabolic demands of nitrogen fixation.

Nitrogen fixation conditions (high C:N ratio) are known to induce exopolysaccharides production in *Azospirillum* species and various mutants affected in the composition of the exopolysaccharides have been characterized [171, 172]. Earlier studies in strain Sp7 cells have shown that the composition of exopolysaccharide (EPS) changes dramatically based on the growth phase and energy status of the cell [173]. In exponential phase, the cells produce glucose-rich EPS which is later degraded and used as a carbon source. In stationary phase, arabinose-rich EPS is produced, facilitating cell aggregation and flocculation, a behavior thought to promote resistance from hostile environmental conditions [173]. Precursor components of EPS are synthesized in the cytoplasm, and then must be transported out to the outer membrane. This transport is accomplished through ATP family transporters [174], an observation also consistent with the up-regulation of a number of ATP transporters detected in Sp7 grown under nitrogen fixation (Table 3.8).

Several proteins suggested to be localized to the inner membrane of cells were found to be enriched in the proteomes of cells grown under conditions of nitrogen fixation, suggesting that growth under these conditions is accompanied by major changes in membrane-associated proteins. In Sp7 nitrogen fixing cultures, basic membrane lipoprotein (abra_1538) and inner membrane lipoprotein A (NLPA lipoprotein, abra_4083) are among the most abundant proteins (Attached spreadsheet Sp7_N2FixData), showing a 2.5 and 1.5-fold increase respectively between nitrogen fixing and non-nitrogen fixing Sp7 cultures (attached spreadsheet 2Fold_UpN2Fix). BLASTp searches [52] against the NCBI conserved domain database using the translated coding sequence of abra_1538 show it to have high similarity with the family of periplasmic binding domains of basic membrane lipoproteins. The proteins within this family are localized to the cell surface, and contain a sugar-binding protein-like fold, which could be important in transport of EPS or LPS components. Members of this family include Med, a positive transcriptional regulator from *B. subtilis* [175] and PnrA from *Treponema pallidum*, a purine nucleoside binding receptor belonging to an ABC transport system [176], suggesting that abra_1538 may be involved in nutrient transport or binding. The function of NLPA lipoprotein is unknown, although studies with pathogenic forms of *E. coli* have suggested that it plays a role in vesicle formation [177].

Further, it is suggested to be structurally distantly related to solute binding proteins known to transfer solutes from the inner membrane for concentration within the cytoplasm [178]. Taken together, this data suggests that abra_1538 and abra_4083 play an important role in Sp7 nitrogen fixing cells.

Interestingly, the protein with the largest degree of up-regulation in Sp7 nitrogen fixing cells (26-fold, Table 3.8) is a protein annotated as 4-phytase (abra_6441). Phytase (phytate-3-phosphatase) catalyzes hydrolysis of phosphate groups from phytate, the main storage form of phosphorous in cereal grasses and legumes [179]. Phosphorous stored in the form of phytate is poorly utilized by the plant, making microbial phytase a key enzyme in phosphorous metabolism. Studies have shown that plants are better able to use phosphorous in low phosphorous soils when inoculated with phytase-secreting microbes [179]. Although there are four different classes of phytases, studies with microbial genomes have indicated that beta-propeller phytases are the most common type found in soil microbes [180]. Beta propeller phytases possess six calcium ion binding sites scattered throughout the protein structure but have no conserved binding motif. Further, phytases belonging to this group share very little sequence identity (< 25%) between them [180]. A BLASTp [52] search of the translated coding sequence of abra_6441 against the NCBI non-redundant protein database indicates that it shows greater than 47% similarity to a number of extracellular solute binding proteins from *Roseobacter* sp. and *Bordatella* sp.. The 26-fold up-regulation in Sp7 nitrogen fixing cells indicates that this protein is likely to be important in nitrogen fixation, perhaps in increasing phosphorous availability to the cells or in making phosphorous available for use in phospholipids that contribute to cell membrane remodeling. This protein is also detected with 50% sequence coverage in Sp245 nitrogen fixing cultures, although it is not detected in Sp245 non-nitrogen fixing cultures, making it impossible to determine if it is up-regulated or simply expressed only in nitrogen fixing conditions in Sp245 strain.

Sp7 cells grown under nitrogen fixing conditions are smaller in size and rounder, thus increasing the cell surface-to-volume ratio, indicating active remodeling of morphology. They also accumulate PHB (poly-hydroxybutyrate) granules, indicating a need to store carbon for later use. Earlier studies with both endophytic and rhizospheric *Azospirillum* species showed that different strains exhibited different growth rate responses to the same substrates. All strains produce large amounts of exopolysaccharides (EPS) in early growth phases [158], with amount produced being inversely proportional to the amount of ammonium available to the cell [149]. The stored EPS is then metabolized for energy during nitrogen fixation, while intracellular carbon stores such as PHB are used primarily for cell maintenance activities [158]. The changes observed in morphology of Sp7 nitrogen fixing cells as well the changes observed in carbon utilization during nitrogen fixation are consistent with the changes in proteome expression profiles of these cells.

Conclusions

Nitrogen fixation results in predictable changes in proteome expression profiles reflecting the basic metabolic shifts necessary for adaptation to nitrogen fixing conditions. The slower growth observed in these cells is reflected in the proteome by decreased numbers and amounts of ribosomal proteins, elongation factors, and DNA and RNA binding proteins. The greater energy requirements for nitrogen fixation are found within the proteome in the form of increased expression of electron transport components to provide energy and increased expression of certain TCA cycle enzymes to provide additional reducing power. The most notable change noted in the proteome is in the greater abundance and variety of transport molecules, suggesting a greater need for molecular transport in nitrogen fixing cells. This is especially noted in both number and amount of those proteins expressed that are predicted to be involved in branched chain amino acid transport. These transport components also appear quite diverse and they are apparently not genetically linked, suggesting a global change in translation that affects the entire physiology of the cell.

As stated earlier, Sp7 and Sp245 occupy different ecological niches, and thus must adapt to different situations, which could be distinguished by different physiology and metabolism between the strains when studied *in vitro*. Proteomic investigations have revealed a number of similarities between these two species, with 959 proteins commonly expressed in both Sp7 and Sp245 cultures grown under conditions of molybdate nitrogen fixation. Proteins that are commonly up-regulated in all nitrogen fixing cultures (Table 3.10), are likely to play very important roles in adaptation to nitrogen fixing conditions. This includes proteins such as solute binding proteins abra_0139 and 0141 and UspA domain protein abra_3154 as mentioned earlier, but also includes a number of other proteins involved in molecular transport as well as a number of hypothetical proteins.

Perhaps most interesting to note is the pattern of up- and down-regulation between nitrogen fixing cultures of each species. In Table 3.10, a number of proteins are not detected to be up-regulated in Sp7 cultures that show substantial fold change levels of up-regulation in Sp245 nitrogen fixing cultures. This same pattern is noted with down-regulated proteins as shown in Table 3.9. However, the majority of the up-regulated proteins are detected in nitrogen fixing cultures (indicated by NF in Table 3.10), many with sequence coverage of 50 % or greater, but are not detected in non-nitrogen fixing Sp7 cultures. Therefore, a level of up-regulation could not be determined for these proteins. For down-regulated proteins, the majority of those proteins were noted in Sp7 control (nitrogen replete) cells, most at low levels of sequence coverage, but were not noted in nitrogen fixing cells, so a level of down-regulation could not be determined. This could be due to the large numbers of ribosomal proteins in Sp7 nitrogen replete control cultures which decreased the depth of proteome coverage and resulted in lower numbers of proteins detected as discussed earlier. Overall, the data suggests that Sp7 species has less dramatic changes in protein abundance than does the Sp245 species. This could, however, be an artifact related to the current lack of complete genome sequence for Sp7. It is possible that some of the proteins found to be expressed in the Sp245 proteome are simply not present in the Sp7 genome. Another possibility is that many unknown proteins found in Sp7 could have similar functions as in Sp245. This comparative proteome study has nevertheless revealed differences in the physiology of cells under different growth conditions. Further, a subset of proteins has been identified that may distinguish the unique physiology of Sp7 and Sp245 under different growth conditions.

Chapter 4. Comparison of Proteome Expression between *Azospirillum brasilense* wild-type and mutants in a chemotaxis-like signal transduction pathway

Introduction

As discussed in Chapter 1, a number of bacteria have multiple chemotaxis operons that can mediate response to a variety of environmental conditions [111]. *A. brasilense* species have a large genome size and exhibit extensive ability to adapt to a number of different environmental conditions. The draft sequence of *A. brasilense* Sp245 indicates the presence of four chemotaxis-like operons and 45 sensory transducers. One of these chemotaxis-like operons, named Che1 [117], has been previously characterized [139] and found to contribute to chemotaxis, although that is not its primary function [118]. The Che1 operon comprises components that are typically found in prototypical chemotaxis pathways that regulate motility patterns. As discussed in Chapter 1, this includes a two component signal transduction pathway consisting of a histidine kinase, CheA1 and its cognate response regulator, CheY1, a coupling protein CheW1 and adaptation proteins, chemoreceptor-specific CheB1 methylesterase and CheR1 methyltransferase. Although the operon structure suggests involvement in chemotaxis, experimental evidence indicates that Che1 gene products function to modulate both cell length at division and cell-to-cell aggregation (named clumping) as well as contribute to chemotaxis [117]. In particular, mutants lacking a functional forward signaling pathway due to deletion of genes encoding histidine kinase CheA1, response regulator CheY1 or deleted for the entire operon, Che1, were found to be shorter in length than wild-type, to maintain a smooth swimming direction for longer periods of time, and to flocculate sooner and more than wild-type [117]. In contrast, a mutant lacking a functional adaptation pathway (or "molecular memory") due to deletion of genes encoding both CheB1 methylesterase and CheR1 methyltransferase, here termed Δ*cheB1cheR1*, were found to be longer than wild-type, to divide at longer lengths than wild type, to be defective in chemotaxis and to be impaired (with substantial delays) in flocculation [117]. Both mutants were found to make less exopolysaccharide (EPS) than wild-type, with the colonies of the Δ*cheY1* mutant having a wrinkled, wet appearance distinctly different from wild-type, while the colony morphology of the Δ*cheB1cheR1* mutant was similar to wild-type [117]. From the above experimental results, it can be hypothesized that the Che1 operon has an indirect effect on chemotaxis with its primary function being the control of cell length at division and the propensity for cell-to-cell interaction (clumping).

However, as stated above, the molecular components of the Che1 pathway show extensive similarity to that found in prototypical chemotaxis signal transduction pathways that exclusively function in controlling the swimming motility pattern. The molecular basis of the Che1 control of cell length at division and clumping remains to be determined. In order to identify potential pathways and functions that may be regulated by this operon as well as possible physiological differences that result from the activity of this operon, we here undertake a proteomic comparison of wild type Sp7 cultures versus mutant Δ*cheB1cheR1* and mutant Δ*cheY1* derivative strains cultures.

Materials and methods

Bacterial Strains and Cell growth

Wild–type *A. brasilense* strain Sp7 (ATCC 29145) was used as control throughout this study, and is here referred to as Sp7. The mutant lacking functional CheB1 and

CheR1 [Δ(cheB1-cheR1)::Km] (BS104) was constructed previously [118], and is herein referred to as Δ*cheB1cheR1*. The mutant lacking functional CheY1 [Δ(CheY1)::Km] (AB102) was also constructed earlier [117] and is referred to in this work as Δ*cheY1*.

All cultures were grown at 25°C on an orbital shaker in minimal media (MMAB) [139] with 5 mM malate and 5 mM fructose as carbon sources, and ammonium chloride as nitrogen source. One culture was grown for proteome preparations. No replicate was grown, so all studies include only one culture. Because earlier studies indicated a maximum difference in size at low optical density (below OD_{600} 0.5), cultures were harvested at an OD_{600} 0.3, and cell pellets weighed and frozen at -80°C for later proteome preparation. A crude membrane fraction for the Sp7 control was prepared as follows from a culture grown to an OD_{600} of 0.6. Cells were harvested and lysed by sonication. Cleared lysate was centrifuged at 18000 xg in a Sorvall ultracentrifuge, and the pellet resolubilized in 50 mM Tris-HCl, pH 7.5 to a protein concentration of 1 mg/ml.

Proteome preparation

Proteome preparation was done as discussed earlier. Briefly, cell pellets were re-suspended at a rate of 0.1g wet cell pellet weight per 500 µl buffer in 6M guanidine hydrochloride and lysed by sonication. Lysate was then cleared of debris by centrifugation at 12000 xg in a Sorvall centrifuge, and cleared lysate was diluted for subsequent digestion. After addition of 10 mM DTT and 1 hour incubation at 60°C, sequencing-grade trypsin (Promega) was added, and solution was incubated overnight at 37°C. An additional 20 µg trypsin was added to each sample the following morning, and samples were incubated for an additional 6 hours. Digestion was halted by addition of

0.1% formic acid. Digested samples were then desalted using Sep-Pak Plus C-18 solid phase extraction (Waters, Milford, MA) following manufacturer's recommendations. Eluates were solvent exchanged via vacuum centrifugation into HPLC-grade water (Burdick-Jackson) containing 0.1% formic acid. Samples were aliquoted into 40 µl volumes and frozen at -80°C for later analysis.

LC/LC-MS/MS

Analysis of cleaned, digested samples was accomplished using MudPIT [17, 20] as described in previous chapters, with each sample being analyzed in technical duplicates. Only two technical replicates were run due to limited instrument time available. Briefly, samples were loaded via pressure cell onto a back column consisting of 3 cm reverse-phase (RP) material followed by 3 cm strong cation exchange (SCX). This back column was then coupled to a 15 cm RP resolving column, packed into a fused silica column pulled to a tip with a 5 µm opening. This front column was then placed directly in-line with a linear ion trapping quadrupole mass spectrometer (LTQ, Thermo-Finnagan). Chromatography was accomplished in 12 steps, with each step consisting of an initial salt pulse to elute a subset of peptides from SCX, followed by a gradient elution of these subsets of peptides with increasing concentration of organic solvent, as described more thoroughly in earlier chapters. Peptides were eluted directly into the mass spectrometer via nano-electrospray for analysis.

The mass spectrometer was operated according to parameters established in Chapter 2 of this work, with instrument operation parameters and data collection being under the control of Xcaliber software. Data was collected in data dependent mode, with one parent ion scan being followed by six tandem MS scans.

Data Analysis

Raw files were searched using SEQUEST [3] against the Sp245 database created earlier as described in Chapter 3 since an Sp7 species-specific database is not yet available. Search parameters were as follows: tryptic digestion, peptide mass tolerance of 3 m/z and a fragment ion tolerance of 0.5 m/z. Additionally, search parameters included two dynamic modifications: 1. methylation represented by a mass shift of +14 m/z on glutamate residues, and 2. deamidation followed by methylation represented by a mass shift of +15m/z on glutamine residues in order to facilitate later analysis of receptor methylation patterns which were beyond the scope of this dissertation. Results were then filtered and collated using DTASelect [54] with the following parameters: XCorr filter levels of 1.8 for peptides with a charge state of +1, 2.5 for those with charge state +2 and 3.5 for charge state +3, minimum delta CN of 0.08, non-tryptic status and 2 peptides per protein identification. Normalized spectral abundance factors (NSAF) [72] were determined for each protein in a given run, and were added to the search results. All DTASelect output files, both HTML and text formats, can be found at https://compbio.ornl.gov/mspipeline/ azospirillum / study1/ status.html.

Search results were imported into Microsoft Access for analysis. Two replicates of each sample were used in all data analysis, and comparisons were done using Sp7 wild-type cultures grown to an OD_{600} 0.3 as control. Pie charts were constructed using the total percentages of proteins expressed in each individual functional category based on ΣNSAF values of each individual protein expressed. Levels of up-regulation and down-regulation were determined as described in Chapter 3, with NSAF values of each protein in mutant cultures divided by NSAF values of the same protein in Sp7 control cultures. Although potentially interesting, proteins expressed only in mutant cultures were not considered in this analysis since levels of up- or down-regulation could not be determined. Further, filter levels of 2.5-fold change were applied to these data sets because the Sp7 nitrogen fixing data sets examined in Chapter 3 of this work indicated a lower level of up- or down-regulation in Sp7 strain. Additionally, because work with individual Sp7 genes located on the pRHICO plasmid have suggested a possibility that some genes, and perhaps their gene products, can vary significantly from the sequence found in the Sp245 genome [172], a BLASTx [134] comparison of available Sp7 translated coding sequences against the Sp245 genome was done.

Results

Sixty seven individual gene sequences for *Azospirillum brasilense* strain Sp7 are available in the NCBI database (http://www.ncbi.nlm.nih.gov/). BLASTx [134] searches using translated coding sequences from these Sp7 strain sequences indicate that 95% of these genes show between 90% and 100% similarity to coding sequences in the Sp245 genome. The high degrees of similarity between coding sequences of the Sp7 genes and those of the Sp245 genome lend credibility to those protein identifications made using the Sp245 genome. One Sp7 translated coding sequence, from gene *nirS*, does not have a homolog in the Sp245 genome. Two Sp7 translated coding sequences, molybdenum nitrogenase alpha and beta subunits, show a low degree of similarity (27% and 32%, respectively) to those from the Sp245 genome, as discussed in Chapter 3. While some proteins may not be identified in the Sp7 proteome by using the translated coding sequences from the Sp245 genome, the high degree of similarity of available Sp7 translated coding sequences to the Sp245 genome sequences gives confidence to those identifications that are made.

Overall proteome analysis

The *Azospirillum brasilense* Sp7 control (Sp7) yielded 1366 identifications from combined technical replicate runs, while the Δ*cheB1cheR1* and Δ*cheY1* mutant strains gave 1355 and 1280 identifications, respectively. 921 expressed proteins were common to all cultures, and reproducibility (determined by the number of common proteins detected in both runs (not shown)/ total number of proteins detected in combined datasets) between technical replicates was 67% for Δ*cheY1* samples and 71% for both Δ *cheB1cheR1* and Sp7 cultures (Table 4.1). Δ*cheY1* and Δ*cheB1cheR1* cultures had 1003 common protein identifications, giving a good representation of proteins required for basic metabolic functions in both mutant cultures. Attached spreadsheet Mutant_Proteome_Data gives a list of all proteins detected in each culture.

The assigned protein identifications were grouped according to functional category as described in Chapter 3, and the numbers of proteins identified in each category for each sample are given in Table 4.2. Also as described in Chapter 3, the relative abundance of each protein detected within a sample was determined through calculating a normalized spectral abundance factor (NSAF) for each protein identified [181]. NSAF values were summed for all proteins in each individual functional category (ΣNSAF) and the percentage of totaled ΣNSAF values of identified proteins that belong

Table 4. 1 Total number of protein identifications in each Sp 7 technical replicate run

	Number of Protein Identifications			Reproducibility between runs
Sample	Replicate 1	Replicate 2	Combined Dataset	
Sp7 OD 0.3	1179	1160	1367	71.1%
Sp7 *ΔcheY1* mutant	1077	1074	1281	67.0%
Sp7 *ΔcheB1cheR1* mutant	1087	1237	1356	71.3%

Table 4. 2 Comparison of protein expression in the cultures of the parental strain Sp7 and its Δ*cheB1cheR1* and Δ*cheY1* mutant derivatives by functional category

			Sp7 OD 0.3		Δ*cheB1cheR1*		Δ*cheY1*	
FC	Functional Category Description	No. Common Proteins[a]	No. Proteins[b]	ΣNSAF Percentage[c]	No. Proteins[b]	ΣNSAF Percentage[c]	No. Proteins[b]	ΣNSAF Percentage[c]
B	Chromatin structure and dynamics	0	1	NA	1	NA	1	NA
C	Energy production and Conversion	89	118	11.6%	114	12.5%	104	10.4%
D	Cell Division and Chromosome Partitioning	16	20	0.5%	20	0.6%	17	0.7%
E	Amino Acid Transport and Metabolism	120	157	8.7%	152	8.9%	143	8.3%
F	Nucleotide Transport and Metabolism	37	52	2.9%	47	3.2%	44	2.9%
G	Carbohydrate Transport and Metabolism	41	60	4.8%	58	5.2%	51	4.3%
H	Coenzyme Metabolism	38	71	1.9%	67	2.0%	50	1.8%
I	Lipid Metabolism	28	44	1.5%	39	1.8%	39	1.9%
J	Translation, Ribosomal Structure and Biogenesis	118	133	32.2%	138	29.2%	121	29.3%
K	Transcription	35	52	3.9%	55	3.8%	49	3.7%
L	DNA Replication, Recombination and Repair	26	46	2.3%	39	2.3%	44	2.7%
M	Cell Envelope Biogenesis, Outer Membrane	40	55	3.7%	60	3.5%	61	4.1%
N	Cell Motility and Secretion	12	33	0.9%	40	0.6%	41	0.7%
O	Posttranslational Modification, Protein turnover, Chaperones	56	69	6.9%	63	6.3%	69	7.3%
P	Inorganic Ion Transport and Metabolism	34	42	2.9%	42	3.7%	42	2.6%
Q	Secondary metabolites biosynthesis, transport and catabolism	17	23	0.9%	24	0.9%	22	1.5%
R	General Function Prediction Only	59	98	2.8%	87	2.7%	87	2.7%
S	Function Unknown	42	78	2.7%	76	2.8%	71	3.4%
T	Signal Transduction Mechanisms	26	51	1.2%	59	1.7%	54	1.6%
U	Intracellular trafficking and secretion	10	10	0.3%	13	0.4%	11	0.5%
V	Defense mechanisms	1	4	0.05%	3	0.03%	10	0.1%
AA	Hypotheticals with similarity or homology to known	34	60	3.5%	64	4.2%	56	5.2%
BB	Hypothetical	42	90	3.5%	95	3.8%	94	4.2%
	TOTAL	921	1366	99.6%	1355	100%	1280	100%

[a] Number of proteins identified in every growth state

[b] Number of proteins identified in each functional category in that growth state

[c] NSAF values were totaled for proteins in each individual functional category(ΣNSAF) to determine the percentage of the total expressed proteome within each functional category

to each category were calculated based on their overall abundance as represented by NSAF values (Table 4.2). A graphical representation of proteome expression profiles compiled by percentage of ΣNSAF values totaled for each functional category is given in Figure 4.1. The proteome expression profiles for all cultures look remarkably similar given the differences in length and colony morphology noted in the different mutant cells [117].

Differences between mutant cultures are more apparent in the expression levels of individual proteins. Levels of up-regulation and down-regulation in comparison to control cultures were determined as described in Chapter 3, dividing the calculated NSAF value of the protein in a mutant culture by the calculated NSAF value of the same protein in a control Sp7 wild-type culture to determine levels of up-regulation, and dividing calculated NSAF values of the Sp7 control divided by those of the same protein in a mutant culture to determine levels of down-regulation. Table 4.3 shows the levels of up-regulation (fold-change) of proteins in both mutant cultures when compared to Sp7 controls, while Table 4.4 shows the levels of down-regulation (fold-change) of proteins in both mutant cultures when compared to Sp7 controls. Examination of these tables clearly illustrates a very different pattern of differential protein expression levels in Δ*cheY1* mutants versus Δ*cheB1cheR1* mutants when compared to Sp7 controls, discussed more thoroughly below.

Perhaps most intriguing in the Δ*cheY1* proteome is upregulation of abra_4441 type VI secretion protein. Although this protein is detected in Sp7 wild-type controls, it is not detected in Δ*cheB1cheR1* mutants at all. Type VI secretion systems (T6SS) are

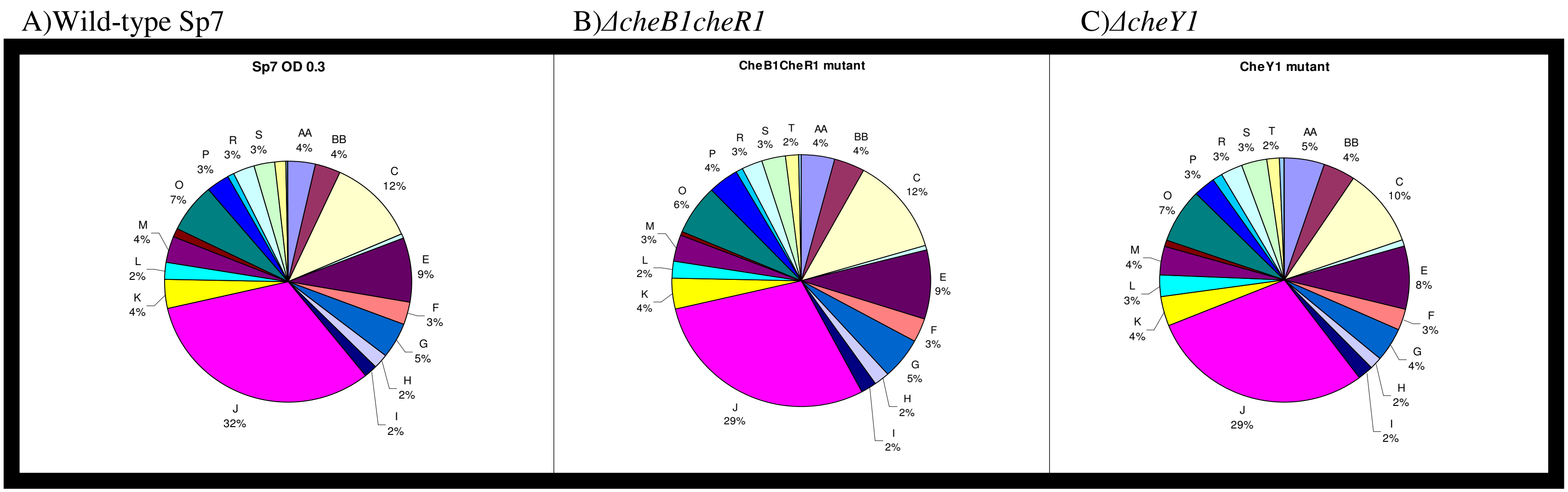


Figure 4. 1 Proteome expression profiles for A) wild-type Sp7 control, B) mutants lacking functional CheB1 and CheR1, and C) mutants lacking functional CheY1

Proteins expressed are grouped by functional category and NSAF values for each individual protein within that category are totaled (ΣNSAF). From the total ΣNSAF of each category, a total percentage of proteins in each functional category based on relative abundance as represented by NSAF values is calculated. Pie chart representations indicate that all cultures are growing rapidly, with only slight differences in the protein expression profile of the mutants versus the wild-type control.

Table 4. 3 Fold change of proteins up-regulated in both *ΔcheB1cheR1* and *ΔcheY1* mutants versus Sp7 controls

Gene	Product	FuncCat	Avg fold-change upreg	*ΔcheB1cheR1*	*ΔcheY1*
abra_0714	NUDIX hydrolase	AA	2.7		2.7
abra_1169	phasin family protein	AA	5.0	2.9	7.1
abra_1276	ribosomal protein L36	AA	2.8		2.8
abra_2407	flagellar assembly regulator FliX	AA	2.6		2.6
abra_3170	phasin family protein	AA	2.9	2.6	3.3
abra_6288	histidine kinase HAMP region domain protein	AA	5.2	5.2	
abra_1987	conserved hypothetical protein	BB	4.3		4.3
abra_2391	conserved hypothetical protein	BB	3.5		3.5
abra_2655	hypothetical protein	BB	5.2		5.2
abra_3281	Sporulation domain protein	BB	3.0	3.0	
abra_3348	hypothetical protein	BB	4.1		4.1
abra_3990	hypothetical protein	BB	13.3	8.5	18.0
abra_4002	hypothetical protein	BB	3.8		3.8
abra_4507	flagellin domain protein	BB	2.6		2.6
abra_6429	conserved hypothetical protein	BB	4.0		4.0
abra_0437	iron-sulfur cluster binding protein	C	3.7		3.7
abra_2275	cytochrome c oxidase, subunit II	C	3.5		3.5
abra_2759	cytochrome c prime	C	2.9		2.9
abra_4337	Cytochrome b/b6 domain	C	2.6	2.6	
abra_4829	ubiquinol oxidase, subunit II	C	2.7	2.7	
abra_4904	Pyridoxal 4-dehydrogenase	C	2.8	2.8	
abra_6420	Phosphoenolpyruvate carboxylase	C	2.9		2.9
abra_0574	cell divisionFtsK/SpoIIIE	D	3.5		3.5
abra_1289	Cobyrinic acid ac-diamide synthase	D	3.0		3.0
abra_1445	TonB box domain protein	D	2.8		2.8
abra_4240	chromosome segregation protein SMC	D	2.7		2.7
abra_5418	Cobyrinic acid ac-diamide synthase	D	4.3	3.6	5.0
abra_0573	aminotransferase class I and II	E	2.6		2.6
abra_0897	Prephenate dehydratase	E	2.5		2.5
abra_1479	Extracellular ligand-binding receptor	E	4.1		4.1
abra_1565	nitrogen regulatory protein P-II	E	5.3		5.3

Table 4.3 (cont) Fold change of proteins upregulated in both *ΔcheB1cheR1* and *ΔcheY1* mutants versus Sp7 controls

Gene	Product	FuncCat	Avg fold-change upreg	*ΔcheB1cheR1*	*ΔcheY1*
abra_1858	Pyridoxal-5'-phosphate-dependent protein beta subunit	E	3.5	2.8	4.2
abra_1864	aminotransferase class I and II	E	3.0		3.0
abra_1935	spermidine/putrescine ABC transporter ATPase subunit	E	2.8		2.8
abra_2184	phosphoadenosine phosphosulfate reductase	E	2.7	2.7	
abra_2247	Glycine dehydrogenase (decarboxylating)	E	3.2		3.2
abra_3996	ornithine carbamoyltransferase	E	4.4		4.4
abra_4110	4-aminobutyrate aminotransferase	E	2.9		2.9
abra_4286	Prephenate dehydrogenase	E	3.1	2.9	3.4
abra_2870	Dihydroorotate oxidase	F	4.0		4.0
abra_0068	Phosphomannomutase	G	2.6		2.6
abra_3118	glycogen/starch/alpha-glucan phosphorylase	G	3.5		3.5
abra_4739	KDPG and KHG aldolase	G	2.7		2.7
abra_6513	glucose sorbosone dehydrogenase	G	3.2	3.2	
abra_6523	TRAP dicarboxylate transporter, DctP subunit	G	3.8	4.1	3.5
abra_0548	glutathione synthetase	H	3.5		3.5
abra_0553	oxygen-independent coproporphyrinogen III oxidase	H	2.7		2.7
abra_1554	phosphomethylpyrimidine kinase	H	4.2		4.2
abra_1954	5-formyltetrahydrofolate cyclo-ligase	H	2.6		2.6
abra_2162	riboflavin synthase, alpha subunit	H	3.5		3.5
abra_3240	pantetheine-phosphate adenylyltransferase	H	2.6		2.6
abra_2604	6-phosphogluconate dehydrogenase, NAD-binding	I	4.0	4.0	
abra_3050	2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase	I	2.8		2.8
abra_3442	phosphatidyl-N-methylethanolamine N-methyltransferase	I	3.6		3.6
abra_5859	acetyl-CoA acetyltransferase	I	3.0		3.0
abra_0826	tRNA (guanine-N1)-methyltransferase	J	3.2		3.2
abra_2439	ribosomal protein L24	J	2.7		2.7
abra_2440	ribosomal protein L14	J	4.2		4.2
abra_2443	ribosomal protein L16	J	3.2		3.2
abra_6528	GntR domain protein	K	3.2		3.2
abra_1522	AAA ATPase central domain protein	L	2.7		2.7
abra_1817	replicative DNA helicase	L	2.5		2.5

Table 4.3 (cont) Fold change of proteins upregulated in both *ΔcheB1cheR1* and *ΔcheY1* mutants versus Sp7 controls

Gene	Product	FuncCat	Avg fold-change upreg	*ΔcheB1cheR1*	*ΔcheY1*
abra_2111	transcription-repair coupling factor	L	4.2		4.2
abra_1829	Lytic transglycosylase catalytic	M	2.9		2.9
abra_1965	peptidoglycan-associated lipoprotein	M	3.0	2.7	3.4
abra_1966	polysaccharide biosynthesis protein CapD	M	2.6		2.6
abra_2091	glucosamine/fructose-6-phosphate aminotransferase, isomerizing	M	2.8		2.8
abra_6505	efflux transporter, RND family, MFP subunit	M	11.3		11.3
abra_2408	flagellar P-ring protein	N	3.1		3.1
abra_2567	chemotaxis sensory transducer	N	5.4	2.8	8.0
abra_3025	chemotaxis sensory transducer	N	3.1		3.1
abra_5341	chemotaxis sensory transducer	N	3.2		3.2
abra_5409	PAS sensor protein	N	2.6	2.6	
abra_0257	cytochrome c-type biogenesis protein CcmI	O	2.7	2.6	2.8
abra_3997	Hsp33 protein	O	2.9		2.9
abra_4742	ATP-dependent chaperone ClpB	O	4.2		4.2
abra_1378	Carbonate dehydratase	P	3.4		3.4
abra_2177	potassium efflux system protein	P	3.4	2.6	4.2
abra_2189	ABC transporter related	P	2.8	2.8	
abra_0652	Carboxymethylenebutenolidase	Q	3.8		3.8
abra_3360	5-oxopent-3-ene-1,2,5-tricarboxylate decarboxylase	Q	2.8	2.9	2.7
abra_5423	acetoacetyl-CoA reductase	Q	2.9		2.9
abra_6359	acetoacetyl-CoA reductase	Q	3.9		3.9
abra_0487	alanine racemase domain protein	R	2.6		2.6
abra_1766	enoyl-(acyl carrier protein) reductase	R	4.1	4.1	
abra_1884	TRAP transporter solute receptor, TAXI family	R	3.5		3.5
abra_2263	Mitochondrial processing peptidase	R	2.5		2.5
abra_3266	protein of unknown function DUF815	R	3.3		3.3
abra_4621	Serine/threonine protein kinase-related	R	3.0	3.0	
abra_6620	basic membrane lipoprotein	R	2.6	2.5	2.7
abra_0139	conserved hypothetical protein	S	30.3	34.2	26.5
abra_0151	Heparinase II/III family protein	S	2.9		2.9

Table 4.3 (cont) Fold change of proteins upregulated in both *ΔcheB1cheR1*and *ΔcheY1* mutants versus Sp7 controls

Gene	Product	FuncCat	Avg fold-change upreg	*ΔcheB1cheR1*	*ΔcheY1*
abra_0162	HemY domain protein	S	2.9		2.9
abra_1380	protein of unknown function DUF1013	S	2.9		2.9
abra_1388	protein of unknown function DUF526	S	3.9		3.9
abra_2142	protein of unknown function DUF490	S	6.5		6.5
abra_2717	40-residue YVTN family beta-propeller repeat protein	S	2.7	2.7	
abra_2767	virulence protein SrfB	S	2.7	2.7	2.7
abra_2768	conserved virulence factor protein	S	3.9	3.0	4.8
abra_3284	iron-sulfur cluster assembly accessory protein	S	3.9		3.9
abra_4441	type VI secretion protein, VC_A0107 family	S	2.8		2.8
abra_4455	type VI secretion-associated protein, ImpA family	S	2.6		2.6
abra_4533	putative periplasmic ligand-binding sensor protein	S	2.9		2.9
abra_0407	PhoH family protein	T	3.2	2.7	3.7
abra_0805	response regulator receiver	T	3.5		3.5
abra_2692	response regulator receiver	T	3.0		3.0
abra_3691	PAS sensor protein	T	3.6		3.6
abra_4542	Stage II sporulation E family protein	T	3.5	2.8	4.2
abra_5331	PAS sensor protein	T	3.3		3.3
abra_3264	preprotein translocase, YajC subunit	U	2.7	2.8	2.6

Table 4. 4 Fold change of proteins down-regulated in both *ΔcheB1cheR1* and *ΔcheY1* versus Sp7 controls

			Avg fold-change		
Gene	Product	FuncCat	Downreg	*ΔcheB1cheR1*	*ΔcheY1*
abra_0403	helix-turn-helix domain protein	AA	6.5	3.5	9.5
abra_3900	NLP/P60 protein	AA	2.5	2.5	
abra_3938	acyl-CoA reductase	AA	2.5	2.5	
abra_0735	conserved hypothetical protein	BB	4.6		4.6
abra_0844	Hpt domain protein	BB	2.6		2.6
abra_2512	conserved hypothetical protein	BB	2.6		2.6
abra_2736	hypothetical protein	BB	2.7		2.7
abra_2750	conserved hypothetical protein	BB	2.6	2.6	
abra_0918	dihydrolipoamide dehydrogenase	C	2.7		2.7
abra_1181	molybdopterin oxidoreductase	C	10.4	17.8	3.0
abra_1305	ferredoxin	C	4.0	4.0	
abra_1578	NADH-quinone oxidoreductase, F subunit	C	3.6		3.6
abra_3380	D-isomer specific 2-hydroxyacid dehydrogenase NAD-binding	C	2.6		2.6
abra_0517	spermidine synthase	E	3.5	3.5	
abra_0979	2-isopropylmalate synthase/homocitrate synthase family protein	E	3.1		3.1
abra_1757	urease, gamma subunit	E	3.0		3.0
abra_1865	ABC transporter related	E	3.1	3.1	
abra_1868	cationic amino acid ABC transporter, periplasmic binding protein	E	3.4	3.4	
abra_2642	peptidase M24	E	2.7	2.7	
abra_2966	aspartate-semialdehyde dehydrogenase	E	3.6	3.6	
abra_3967	aspartate kinase	E	2.8		2.8
abra_4885	Substrate-binding region of ABC-type glycine betaine transport system	E	3.2		3.2
abra_4887	glycine betaine/L-proline ABC transporter, ATPase subunit	E	3.4		3.4
abra_1146	Orotate phosphoribosyltransferase	F	4.3		4.3
abra_6433	thiazole biosynthesis family protein	F	2.7		2.7
abra_0423	PfkB domain protein	G	2.8		2.8
abra_0921	transaldolase	G	3.7		3.7
abra_1553	phosphoglucosamine mutase	G	2.8	2.7	2.9
abra_1642	Triose-phosphate isomerase	G	2.7		2.7
abra_1950	glyceraldehyde 3-phosphate dehydrogenase	G	3.3		3.3
abra_2097	pantoate/beta-alanine ligase	H	2.6		2.6

Table 4.4 (cont) Fold change of proteins downregulated in both *ΔcheB1cheR1* and *ΔcheY1*

			Avg fold-change		
Gene	Product	FuncCat	Downreg	*ΔcheB1cheR1*	*ΔcheY1*
abra_2597	Polyprenyl synthetase	H	2.7		2.7
abra_4347	deoxyxylulose-5-phosphate synthase	H	2.9	2.9	
abra_0946	ribosomal protein L27	J	2.8	2.7	3.0
abra_0980	RNA methyltransferase, TrmH family, group 1	J	4.3	4.3	
abra_2435	ribosomal protein L6	J	3.1		3.1
abra_3254	gid protein	J	2.6	2.6	
abra_3690	tyrosyl-tRNA synthetase	J	2.9		2.9
abra_2463	transcription termination/antitermination factor NusG	K	3.1		3.1
abra_6192	LysR substrate-binding	K	3.9	3.9	
abra_4247	DNA methylase N-4/N-6 domain protein	L	2.5		2.5
abra_3497	Serine-type D-Ala-D-Ala carboxypeptidase	M	3.2		3.2
abra_2635	chemotaxis sensory transducer	N	3.3		3.3
abra_6213	flagellin domain protein	N	15.0	7.0	23.0
abra_0172	2-alkenal reductase	O	2.7	2.7	
abra_0444	20S proteasome A and B subunits	O	4.3	4.3	
abra_0471	protein of unknown function DUF255	O	4.8	4.8	
abra_4609	Endopeptidase Clp	O	2.9		2.9
abra_3260	Superoxide dismutase	P	2.9		2.9
abra_6242	protein of unknown function DUF47	P	2.5		2.5
abra_6590	conserved hypothetical signal peptide protein	P	2.6		2.6
abra_0566	ABC transporter related	Q	3.0	3.0	
abra_0593	ThiJ/PfpI domain protein	R	6.6		6.6
abra_0945	GTP-binding protein Obg/CgtA	R	3.0		3.0
abra_1538	basic membrane lipoprotein	R	2.6		2.6
abra_3520	cobalamin biosynthesis protein CobW	R	5.6	5.6	
abra_0131	protein of unknown function DUF971	S	2.9	2.9	
abra_0464	protein of unknown function DUF299	S	2.5		2.5
abra_1433	rpsU-divergently transcribed protein	S	3.0		3.0
abra_0303	CBS domain containing protein	T	3.6	3.6	

Table 4. 5 NSAF values of Type 6 Secretion System components detected in mutant strains and in Sp7 control cultures

Type 4 Secretion System IcmF, DotU region

		ΔcheB1cheR1	*ΔcheY1*	Sp7
abra_0639	hypothetical protein	8E-4	7E-4	9E-4
abra_0643	dctP TRAP dicarboxylate transporter, DctP subunit	3E-4	3E-4	3E-4
abra_0646	type VI secretion protein IcmF	5E-5	6E-5	ND[a]
abra_0647	type IV / VI secretion system protein, DotU family	ND[a]	ND[a]	5E-5

Type 6 Secretion System region

		ΔcheB1cheR1	*ΔcheY1*	Sp7
abra_4428	sigma-54 factor interaction domain-containing protein	6E-5	ND[a]	6E-5
abra_4435	type VI secretion ATPase, ClpV1 family	ND[a]	2E-4	8E-5
abra_4439	protein of unknown function DUF796	ND[a]	5E-4	ND[a]
abra_4440	type VI secretion protein, EvpB/VC_A0108 family	ND[a]	7E-4	5E-4
abra_4441	type VI secretion protein, VC_A0107 family	ND[a]	2E-3	8E-4
abra_4442	conserved hypothetical protein	ND[a]	2E-4	1E-4
abra_4452	type VI secretion protein, VC_A0114 family	5E-5	ND[a]	1E-4
abra_4455	type VI secretion-associated protein, ImpA family	ND[a]	2E-3	1E-5
abra_4458	hypothetical protein	ND[a]	1E-4	ND[a]

[a] ND = not detected

multi-component transport systems, thought to be involved in bacterial interaction with a eukaryotic host. Table 4.5 catalogs all components of the T6SS expressed in all cultures with their concomitant NSAF values. Additionally, the NSAF values of Type 4 secretion system that are homologs to the T6SS components are given in Table 4.5. While components of the T4SS are detected in all cultures, several components of the T6SS are detected in both Δ*cheY1* mutant proteomes and in Sp7 wild-type control cultures, but are not seen in the Δ*cheB1cheR1* mutant at all, suggesting the intriguing possibility that this system is responding, either directly or indirectly, to the Che1 signaling system.

Discussion

Proteome expression profiles are similar for control and mutants with only slight differences noted. Since all cultures were harvested at similar points in their growth (early log phase) and all were growing rapidly and dividing at the same rate, this result is not unexpected and it is consistent with previous observations that indicated that differences between the mutants and the parental strain were subtle [117]. The slight differences noticed in percentage of abundances as represented by total NSAF values for proteins in individual functional categories could be attributed to general variation in culture growth.

As expected for early mid-log phase cultures, the most abundant proteins found within all cultures grown (attached spreadsheet Mutant_Proteome_Data) indicate the cultures were growing rapidly and were metabolically active. Active protein synthesis is suggested by the abundance and variety of ribosomal proteins, chaperonins and translation elongation factors, a common set of which is found in both mutant cultures as well as in the Sp7 wild-type control culture. DNA manipulation and RNA synthesis is implied by the presence of a number of DNA binding proteins. All cultures have an abundance of ATP synthase subunits, cytochrome c proteins and electron transferring flavoproteins, suggesting they are actively respiring. Several malate dehydrogenase proteins and glyceraldehyde 3-phosphate dehydrogenase proteins are present in all cultures, suggesting active glycolysis and TCA cycles as expected in early mid-log phase cultures.

Differential expression in proteins related to general metabolism

Patterns of up- and down-regulation (Tables 4.3 and 4.4, respectively) are very different for Δ*cheB1cheR1* and Δ*cheY1* mutants, indicating that each mutant has a unique physiological response to absence of different components of the Che1 pathway. As noted in other bacterial species, the lack of a forward signaling response regulator in a two component signaling system consisting of histidine kinase-response regulator pair elicits a more dramatic response from the mutant cells than those lacking an adaptation pathway [182]. In keeping with this observation, Δ*cheY1* mutants exhibit a greater response to the loss of the forward signaling pathway than do Δ*cheB1cheR1* mutants to the loss of the adaptation pathway.

Protein metabolism

Protein metabolism appears to be very different in the two mutants. In Δ*cheY1* mutants, a number of proteins involved in amino acid and cofactor biosynthesis as well as in amino acid modification are found to be up-regulated, while only a few of these same proteins show up-regulation in the Δ*cheB1cheR1* mutant, suggesting an increased need for amino acid metabolism in the Δ*cheY1* mutant. Of specific interest are two proteins involved in tyrosine metabolism, prephenate dehydratase (abra_0897) and prephenate dehydrogenase (abra_4286), up-regulated in Δ*cheY1* mutants by 2.6- and 3.4-fold, respectively. Tyrosine residues have been found to be important in transport of molecules through outer membrane protein OMP85 [174]. In *E coli*, OMP85 is the outer membrane protein responsible for transport of a number of outer membrane proteins [174]. OMP85 exists in a large complex in the outer membrane anchored to the peptidoglycan by lipoprotein and recognizes a motif containing tyrosine residues followed by three hydrophobic residues in those proteins it is to transport [174, 183].

Additionally, abra_4742 ClpB protease, an ATP-ase that functions to disaggregate proteins and restore their function [184], is also up-regulated in Δ*cheY1*. Further, glycine/betaine transporters, abra_4885 and 4887 are down-regulated by 3-fold in Δ*cheY1* mutants. Betaines are zwitterionic compounds containing cationic functional groups such as ammonium on one end of the molecule and anionic functional groups such as carboxylate at the other end [185]. Glycine betaines are amino acid derivatives that often serve as methyl donors, or can act as an osmolyte that both stabilizes proteins and adjusts solute concentration in the cytoplasm to restore hydration and protect a cell from dehydration [185]. Transporters act as osmolarity sensors in the surrounding environment [185]. Down-regulation of these transporters may therefore simply be due to the smaller size of the cells, but may also lead to a reduced ability to sense changes in osmolarity of the surrounding media. Further, since glycine betaines serve to stabilize proteins within the cell [185] the down-regulation of this protein in *ΔcheY1* mutants could lead to more aggregated proteins due to unfolding, thus leading to a higher expression of

ClpB protease as mentioned above. In contrast, Δ*cheB1cheR1* mutants show few proteins participating in amino acid metabolism to be up-regulated, instead exhibiting the greatest degree of up-regulation in hypothetical proteins.

Molecular transport

Molecular transport appears to be different between the two mutants as well. Interestingly, although Δ*cheB1cheR1* mutants exhibit a larger cell size, ABC transporters abra_1865 and abra_1868, as well as ABC transporter related protein abra_0566 are down-regulated while only one ABC transporter, abra_2189, is up-regulated. In contrast, Δ*cheY1* mutants do not show changes in expression levels of ABC transporter proteins. However, both mutants show up-regulation of abra_3264 YajC preprotein translocase, a component of Type III protein secretion systems which serves to translocate proteins from the cytoplasm to the periplasmic space, suggesting the need for increased protein translocation in both mutants. While this is understandable in the Δ*cheB1cheR1* mutant due to its larger size which presumably has more proteins on its outer membrane due to the greater surface area of the membrane, it is an interesting phenomenon in the smaller Δ*cheY1* cells.

Intriguingly, a protein found to be abundantly up-regulated in Sp245 cells grown under nitrogen fixation, abra_0139, is also found to be among the most abundant proteins of both Δ*cheB1cheR1* and Δ*cheY1* mutant proteomes. As discussed in Chapter 3, abra_0139 is a small extracellular solute binding protein with a *Bordatella* Uptake Gene (BUG) domain, thought to be involved in nutrient uptake. Interestingly, Sp7 cells grown under nitrogen fixing conditions show a 9-fold up-regulation of abra_0139 (Chapter 3), while the Δ*cheB1cheR1* and Δ*cheY1* mutants show a 34- and 26-fold up-regulation of this protein, respectively (Table 4.3). An additional solute binding protein containing a BUG domain, abra_0141, is also present in both mutant proteomes, but was not detected in control cultures so levels of up-regulation could not be determined (attached spreadsheet Mutant_Proteome_Data). This was also the case for Sp7 nitrogen fixing cultures where abra_0141 was detected only in nitrogen fixing cells. The large degree of up-regulation of solute binding receptor protein abra_0139 and the presence of solute binding receptor protein abra_0141 in both mutant proteomes once again suggests that it is important in nutrient uptake or transport, further suggesting that the Δ*cheB1cheR1* and Δ*cheY1* mutants are sensing their environment in a different way than are wild-type cells.

Carbon metabolism

Carbon metabolism also seems to be affected in the Δ*cheY1* mutant. Evidence of an altered carbon metabolism is given in the up-regulation of several proteins involved in carbon metabolism. Abra_4739 KDPG and KHG aldolase, an Entner-Doudoroff pathway enzyme which catalyzes creation of pyruvate and glyceraldehyde-3-phosphate from 2-keto-3-deoxy-6-phosphogluconate (KDPG) is up-regulated by ~3-fold in the Δ*cheY1* mutant, but not at all in the Δ*cheB1cheR1* mutant. Additionally, abra_6420 PEP carboxylase, which converts phosphoenolpyruvate to oxaloacetic acid, is up-regulated by 3-fold in the Δ*cheY1* mutant only while abra_1950 glyceraldehyde 3-phosphate dehydrogenase, involved in the glycolytic pathway, is down-regulated by 3-fold.

Carbon storage in the form of PHB granules may also be affected in both mutants. One phasin family protein, abra_3170, is among the most abundant proteins in all proteomes (attached spreadsheet Mutant_Proteome_Data), although it is of higher abundance in both mutant proteomes. Examination of the up-regulated proteins in Δ*cheB1cheR1* (Table 4.3) shows other phasin family proteins, abra_1168 and 1169, to be up-regulated by 2.6-fold and 2.9-fold respectively, while in Δ*cheY1* cultures abra_3170 is upregulated by 3.2-fold, abra_1169 is upregulated by 7-fold and abra_1168 does not show any fold-change (Table 4.3). Phasins are structural proteins embedded within the lipid layer that surrounds poly-3-hydroxybutyrate (PHB) granules [186]. PHB granules consist of a polyester core surrounded by a barrier layer thought to be composed of a lipid monolayer with embedded and associated proteins that function to synthesize, store, and access the PHB carbon when needed [186]. *Azospirillum* species are known to produce small PHB granules under optimal conditions, with the amount of PHB increasing up to 30-fold when grown under conditions of high C:N ratio and high amounts of oxygen [154]. During times of nutrient limitation, specifically when carbon sources are abundant while nitrogen is limited, a bacterium can accumulate up to 80% of its dry weight in PHB granules [186, 187]. It is thought that the presence of phasin proteins prevents individual PHB granules from fusing, with phasin synthesis and abundance being highly correlated to PHB synthesis and abundance [186]. Microscopic examination of cells shows the presence of PHB granules in all cultures, both mutants and Sp7 control. However, the amount of PHB in each was not quantified.

PHB synthesis is accomplished in a very simple pathway consisting of three steps (Figure 4.2). Condensation of acetyl-CoA subunits to acetoacetyl-CoA is catalyzed by beta-ketothiolase, while acetoacetyl-CoA-reductase catalyzes the subsequent formation of

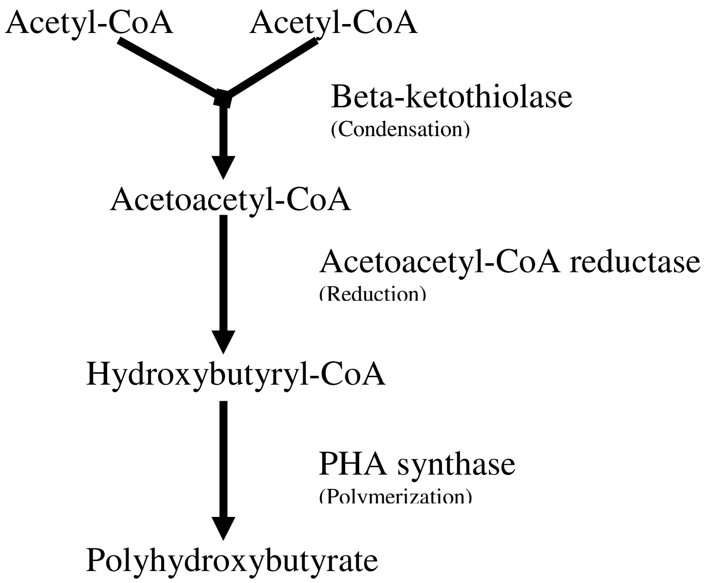


Figure 4. 2 Pathway of PHB biosynthesis
Acetyl-CoA subunits are condensed to form acetoacetyl-CoA (catalyzed by beta-ketothiolase). Acetoacetyl-CoA-reductase catalyzes the subsequent formation of hydroxybutyryl-CoA, and PHA synthase then acts to polymerize individual units of hydroxybutyryl-CoA to form poly-hydroxybutyrate. As discussed in the text, mutants lacking a functional CheY1 show upregulation of a number of proteins annotated as acetoacetyl CoA reductase, although no proteins from this pathway were detected in any cultures. In addition a number of phasin proteins, which surround PHB granules, are upregulated in this mutant, suggesting an altered carbon metabolism.

hydroxybutyryl-CoA. PHA synthase then acts to polymerize individual units of hydroxybutyryl-CoA to form poly-hydroxybutyrate [186]. Gene sequences for those Sp7 genes involved in PHB synthesis are present in the NCBI database. BLASTx [134] comparison of these translated coding sequences to the Sp245 genome coding sequences showed beta-ketothiolase (phaA) to correspond to abra_4231 with 100% similarity, acetoacetyl-CoA reductase (phbB) to abra_4232 with 97% similarity, and poly-beta-hydroxybutyrate synthase (phbC) to abra_4230 with 98% similarity. An additional gene involved in PHB manipulation is PHB depolymerase (phaZ), corresponding to abra_0608 with 97% similarity. None of these genes were detected in any culture, either control or mutant. However, 3 different proteins annotated as acetoacetyl-CoA reductase (abra_5859, 5432, and 6359) were detected to be up-regulated by 3.0-, 2.9-, and 3.9-fold, respectively, in Δ*cheY1* mutants (Table 4.3).

It is perhaps expected for the Δ*cheB1cheR1* mutant to have a greater amount of PHB granules because of the larger size, especially if carbon storage in the form of PHB is relative to the size of the cell, but the increase in phasin family proteins for the Δ*cheY1* mutant is somewhat unexpected. Since carbon storage in the form of PHB granules is known to increase in a high C:N ratio medium, which was not the case for these mutants, the differential carbon storage in the form of PHB granules provides another clue that these mutants may be sensing the environment differently than the wild type control.

Nitrogen and metal metabolism

Nitrogen regulatory protein P-II (abra_1565) is up-regulated by 5-fold in Δ*cheY1* mutants (Table 4.3), just as it is in nitrogen fixing Sp7 cells, suggesting an increased need for nitrogen sensing in these mutants. PhoH family protein (abra_0407) is upregulated in both mutant cultures by 3-fold (Table 4.3). In enteric bacteria, PhoH is a phosphate starvation inducible protein that possesses nucleoside triphosphate hydrolase activity. *E. coli* PhoH is proposed to be part of the Pho regulon, whose expression is controlled by two component system PhoR-PhoB [188]. The response to phosphate starvation and the role of PhoH in *A. brasilense* or any other *Azospirillum* species has not been investigated. However, the up-regulation of this protein in both mutants could suggest a role for this PhoH family protein related to growth or clumping behavior in *Azospirillum* cells.

Cofactor biosynthesis

Both mutant cultures show a down-regulation of molybdopterin oxidoreductase (abra_1181), although Δ*cheB1cheR1* shows a far greater down-regulation (17-fold) than does Δ*cheY1* (3-fold, Table 4.4). Molybdopterin oxidoreductase complexes molybdenum to pterin, thus activating it for use in more than 50 molybdate-containing enzymes [189]. Since the molybdenum cofactor is used in a diverse set of oxidation and reduction reactions involved in nitrogen, sulfur and carbon metabolism [190], the down-regulation of this one enzyme could potentially affect the function of many additional enzymes. Δ*cheB1cheR1* mutants show a 5-fold down-regulation of cobalamin biosynthesis protein CobW, one enzyme within the pathway responsible for synthesis of co-factor cobalamin (vitamin B-12). Cobalamin is a co-factor for a number of other enzymes [191], and its down-regulation suggests a related effect on the function of those enzymes dependent on it for their function, which also suggests differences in the metabolic activity of Δ*cheB1cheR1* versus Δ*cheY1* and Sp7.

Respiratory chain

Differences in the requirement for (or expression of) respiratory chain components is implied in the up-regulated proteins for Δ*cheY1* and Δ*cheB1cheR1* (Table 4.3). In Δ*cheY1* mutants, abra_0437, an FeS binding protein, is up-regulated, just as it is in nitrogen fixing Sp7 cells. Additionally, cytochrome c oxidase (abra_2759) and cytochrome c' (abra_2759) is up-regulated in Δ*cheY1* cells. In contrast, Δ*cheB1cheR1* mutants show up-regulation of abra_4337 cytochrome b/b6 and abra_4829 ubiquinol oxidase, while ferredoxin abra_1305 is down-regulated, suggesting the use of a different or additional respiratory pathway in Δ*cheB1cheR1* mutants. Taken together, these data suggest that each mutant is utilizing a different respiratory metabolism with the notable similarity of Δ*cheY1* to Sp7 nitrogen fixation. This is interesting as nitrogen fixing Sp7 cells are also usually shorter and rounder relative to cells grown under nitrogen-replete conditions.

Signaling proteins

Interestingly, a number of signaling proteins are differentially expressed in the two mutant cultures. *ΔcheY1* shows up-regulation of chemotaxis sensory transducers abra_3025 and abra_5341 while abra_2635 is down-regulated. One chemotaxis sensory transducer, abra_2567, is up-regulated in both mutant cultures. Response regulator receivers abra_0805, and abra_2692 are up-regulated in Δ*cheY1,* but no response regulator receivers are found to be up- or down-regulated in Δ*cheB1cheR1.* PAS sensor proteins abra_5331 and abra_3691 are up-regulated in Δ*cheY1* mutants, while PAS sensor protein abra_5409 is up-regulated in Δ*cheB1cheR1*. PAS domains are found in cytosolic signaling transduction proteins that function to sense the intracellular energy status of the cell [192]. They bind small co-factors such as heme or flavin or adenine, with the small molecule bound allowing for specificity in sensory input [192, 193]. In contrast to Δ*cheY1* mutants, Δ*cheB1cheR1* mutants show a greater than 5-fold up-regulation of only one signaling protein, abra_6288. A BLASTp [52] search of the abra_6288 translated sequence against the NCBI non-redundant protein database revealed that it has 46% identity to a chemotaxis sensory transducer from *Rhodopseudomonas palustris* BisB5, and contains a C-terminal methyl-accepting chemotaxis-like domain with a HAMP region directly upstream of that domain, which indicates that this is a chemoreceptor or MCP (Figure 4.3). HAMP (histidine kinase, adenylyl cyclase and methyl carrier protein) regions are common signaling domains occurring in a wide variety of signal transduction proteins in bacteria, and function both as extracellular sensors and as dimerization domains [193]. The up-regulation of only this MCP signaling protein in *ΔcheB1cheR1* mutants suggests that this MCP has a special role that is relevant or specific to the physiology of cells such as Δ*cheB1cheR1*. Notably, these differences are also consistent with changes in the physiology of these cells that require different sensing and signaling systems.

Proteins related to changes in cell morphology/clumping

As discussed earlier, Δ*cheB1cheR1* mutant cells lack the methylesterase and methyltransferase that functions in the adaptation pathway of the Che1 operon. Although these mutants grow well and at a rate indistinguishable from that of the wild type cells, they also produce EPS of different properties relative to the wild type cells [117] which may correspond to less EPS or different types of EPS. This change in EPS production

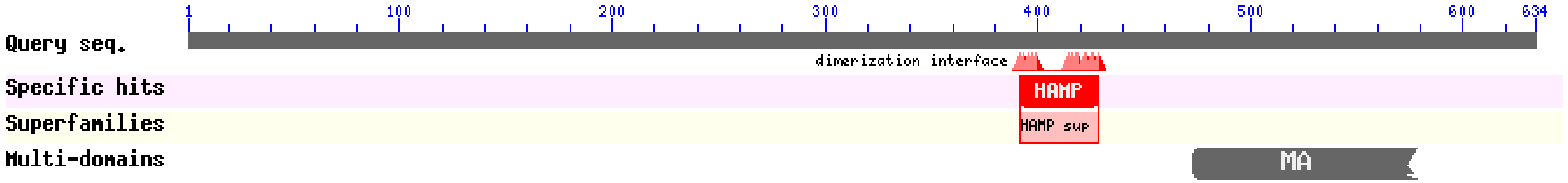


Figure 4. 3 BLASTp depiction of conserved domains in abra_6288
The protein sequence indicates that this protein contains a C-terminal methyl-accepting chemotaxis-like domain with a Histidine kinase, Adenylyl cyclase, Methyl-accepting protein, and Phosphatase (HAMP) domain. The presence of these two domains suggests that this protein may be involved in sensing environmental changes and signaling signaling those changes. 5-fold upregulation of this signaling protein in only ΔcheB1cheR1 cells suggests that it may play an important role in cells exhibiting a phenotype such as these.

also leads to the observed defect of Δ*cheB1cheR1* mutant in clumping by cell-to-cell aggregation, an event that is a prerequisite to flocculation, which is not observed in cultures of the Δ*cheB1cheR1* mutant. Further, they grow to longer length prior to cell division, and divide at longer lengths than wild-type [117]. In contrast, Δ*cheY1* mutant cells lack the signaling output of the pathway, since CheY1 is homologous to the response regulator CheY from *E.coli* that interacts with the flagellar motor to modulate the direction of rotation of the flagellum. Whether CheY1 interacts with *Azospirillum*'s flagellar motor is unknown but mutations in CheY1 affect the swimming motility bias of cells [117] suggesting that it does interact with the flagellar motor. In addition, mutations in CheY1 increase the propensity of cells for clumping, perhaps by modulating EPS production (amount and/or composition). Individual *A. brasilense* Δ*cheY1* mutant cells also divide at shorter cell length relative to the wild type [117], as discussed earlier. All the phenotypes displayed by the Δ*cheY1* mutant strain are similar to that of a strain lacking the entire Che1 pathway, suggesting that CheY1 functions as the signaling output of the pathway to modulate clumping, cell length at division and chemotaxis, either through direct interaction or through an indirect effect on other pathways.

With that in mind, one goal of this proteomics experiment was to identify changes that could be common to both mutants, thus pointing to a single output of the Che1 pathway or unique to each mutant and thus pointing to unique proteins involved in clumping or in increasing cell length. For instance, if a protein is found to be up-regulated in Δ*cheY1* mutant strain while remaining unchanged, down-regulated, or not expressed in the Δ*cheB1cheR1* mutant, then the possibility exists that it is related to clumping behavior. Likewise, if a protein is found to be up-regulated in the Δ*cheB1cheR1* mutant strain while at the same time remaining unchanged, unexpressed or down-regulated in the Δ*cheY1* mutant strain, the possibility exists that it is related to cell growth or elongation. Those that are unique to each mutant whether up- or down-regulated are probably related to unique cellular responses that result from mutation effect on cell length and clumping and so are likely specific to the physiology of the mutant.

Cell wall biosynthesis

Phenotypic characterization of Δ*cheY1* mutant cells indicate a shorter cell with an increased tendency to flocculate and an altered colony morphology from wild-type, suggesting differences in the outer cell wall composition and attached exopolysaccharide (EPS) [117].   Up-regulation of a number of proteins in the Δ*cheY1* mutant proteome supports the hypothesis that Δ*cheY1* mutants are undergoing extensive morphological remodeling. Altered peptidoglycan synthesis in Δ*cheY1* mutants is suggested by up-regulation of abra_1829 lytic glycosylase while abra_3497 Serine-type D-alanyl-D-alanine carboxypeptidase is down-regulated. Three-fold up-regulation of two glycosyltransferases, one with phosphorylase activity (abra_3118), and another without (abra_1786) suggests that either new lipid A or EPS components are being synthesized or different sugars are being added in Δ*cheY1* cells.   Outer polysaccharide layer changes are further suggested in Δ*cheY1* mutants by the 2.6-fold up-regulation of polysaccharide biosynthesis protein CapD (abra_1966), and 3.5-fold upregulation of abra_3118 glycogen/starch/alpha-glucan phosphorylase.

The cell envelope of gram-negative bacteria is composed of both an inner membrane and an outer membrane separated by a periplasmic space containing peptidoglycan [194]. Peptidoglycan is composed of long chains of alternating N-acetylglucosamine and N-acetylmuramic acid subunits. These long glycan chains are cross-linked together with short peptide bridges, forming a strong and flexible protective structure surrounding the inner membrane. Lytic transglycosylases, such as abra_1829 which is up-regulated in *ΔcheY1* mutants (Table 4.3), are enzymes that cleave peptidoglycan in order to allow biosynthesis and turnover of peptidoglycan layers [195], while D-Ala-D-Ala carboxypeptidases such as abra_3497, which is up-regulated in *ΔcheY1* mutants, facilitate formation of peptide cross-links. Differential regulation of these proteins in in *ΔcheY1* mutants suggests these cells may be remodeling peptidoglycan layer.

The inner membrane of the cell envelope is a phospholipid bilayer with embedded transmembrane proteins and associated lipoproteins, while the outer membrane is an asymmetric bilayer with only the inner leaflet composed of phospholipid [174]. The outer leaflet of the outer membrane is a selectively permeable membrane composed of lipopolysaccharides (LPS) with associated lipoproteins and imbedded proteins that act as transport channels for small molecules less than 600 Da [183]. LPS is composed of three elements: a complex glycolipid called lipid A which anchors the LPS to the outermembrane, a core oligosaccharide that provides an attachment site for O-antigen, and O-antigens which are composed of repeating units of oligosaccharides [174, 196]. Each component of LPS is synthesized individually via a pathway beginning with activation of cytosolic sugars through the addition of UDP. Lipid A is synthesized in the cytosol via the action of UDP-N-acetylglucosamine acyltransferase, deactylase and N-acyltransferase. The synthesis pathway for the LPS core oligosaccharide inner region is unknown, although the outer core region of hexose sugars is added to the inner core by glycosyl-transferases which are specific for the sugar molecule added. *ΔcheY1* mutants show 3-fold up-regulation of two glycosyltransferases, abra_3118 and abra_1786, while *ΔcheB1cheR1* do not show up-regulation of any glycosyltransferases, suggesting that either new lipid A or EPS components are being synthesized or different sugars are being added in Δ*cheY1* cells. This observation is consistent with the phenotypic observation of morphology differences between the two mutants.

Exopolysaccharides (EPS) are polymers of a wide variety of high molecular weight polysaccharides and/or proteins that are loosely attached to the outer cell membrane [197]. Capsular polysaccharides (CPS) appear as fibrous material at the cell surface, anchored to the outer membrane via covalent linkages to phospholipid or to lipid A. Synthesis and transport of EPS and CPS utilizes many of the same pathways and proteins as those involved in LPS synthesis [197]. EPS is constructed from intracellular nucleoside diphosphate sugar precursors. Some are synthesized through undecaprenol intermediates, attached to a fatty acyl carrier and polymerized in the same manner as LPS [183]. Others are not associated with lipids, but are instead polymerized by synthetases in the cytosol and transported to the outer membrane via ABC-type transporters. *Azospirillum* species synthesize both exopolysaccharides (EPS) that are loosely attached and capsular polysaccharides (CPS) that are anchored to the outer leaflet of the outer membrane [88]. Differences in synthesis or export of EPS and CPS in *ΔcheY1* mutants is suggested by the up-regulation of polysaccharide biosynthesis protein CapD (abra_1966) and glycogen/starch/alpha-glucan phosphorylase (abra_3118) in Δ*cheY1* mutants. CapD is a gamma- glutamyltranspeptidase in gram positive *Bacillus anthracis* cells and functions in anchoring polysaccharide layers to the peptidoglycan [198, 199]. In gram negative *Rickettsia prowazekii* cells, CapD catalyzes the epimerization of UDP-glucose to UDP-galactose, and is important in polysaccharide biosynthesis [200]. Glycogen phosphorylase is involved in the degradation of glycogen, although the role of glycogen storage in *Azospirillum* is still not well understood. Recently, an *A. brasilense* Sp7 mutant was created that lacked one glycogen phosphorylase gene termed *glgP* that had 97% similarity at a protein sequence level to abra_3118. Although the morphology of this mutant was similar to wild-type, the mutant showed impaired biofilm formation, decreased amounts and different composition of EPS, but greater capability for survival under stress conditions such as starvation and osmotic pressure [201]. Up-regulation of abra_3118 glycogen phosphorylase protein in Δ*cheY1* mutants suggests that this protein may play a role in Δ*cheY1* mutant cells, either in using glycogen stores as an additional carbon source, or in modifying EPS. Taken together, the above data suggests that Δ*cheY1* mutants are engaged in active remodeling of their cell wall and attached EPS, perhaps synthesizing more EPS and LPS than wild-type cells.

Under conditions of nutrient limitation when carbon sources are abundant (high C:N ratio), *Azospirillum* species will aggregate, with accompanying loss of motility, and formation of PHB granules and a thick polysaccharide-rich coat or capsule [202]. Extent of flocculation and/or aggregation is dependent upon the composition of the EPS surrounding the cells [87, 154]. Although both Δ*cheB1cheR1* and *ΔcheY1* cultures were grown under high aeration conditions in minimal media containing both carbon and nitrogen sources (not conditions to induce flocculation and aggregation), Δ*cheY1* still exhibits an increased tendency to clump, while Δ*cheB1cheR1* cells show the opposite tendency and are in fact impaired in flocculation. The above mentioned results, taken together with up-regulation in Δ*cheY1* mutants and no change in Δ*cheB1cheR1* mutants, suggest a number of possible proteins that could contribute to a morphological change leading to clumping in Δ*cheY1* mutant cells.

Type VI secretion systems

An interesting set of proteins shown to be up-regulated in the Δ*cheY1* mutant but not detected in the Δ*cheB1cheR1* mutant includes components of the Type VI secretion system (T6SS), expressed as shown in Table 4.5. Type VI secretion systems were only recently recognized and named as a separate secretion system for export of proteins either with or without leader peptides into the extracellular media [203]. The primary characterization of components of this system has been done in pathogenic bacteria due to the requirement of the T6SS for virulence in pathogenic strains of *E. coli, Vibrio cholerae, Vibrio anguillarum, Salmonella typhimurium* and *Pseudomonas aeruginosa*, to name a few [204]. Nevertheless, only a few of the proteins found within large genomic regions encoding the T6SS components have been characterized, and identity of proteins secreted by this system remains elusive [204].

The *Azospirillum brasilense* Sp245 genome shows 2 regions containing genes annotated as Type VI secretion components. One small region contains genes encoding IcmF (Intracellular multiplication F, abra_0646) and a DotU family protein (Defect in organelle trafficking, abra_0647), both involved in Type 4 secretion systems as well, arranged as shown in Figure 4.4. Another large genomic region contains conserved T6SS components (discussed below) Vgr family protein (Valine-lysine repeat protein, abra_4434), ClpV1 Atp-ase (abra_4435), EvpB (abra_4440) and ImpA (abra_4455), arranged as depicted in Figure 4.5. No Hcp (Hemolysin coregulated protein) protein, a conserved secreted T6SS component, is found within the Sp245 genome structure, although a number of proteins annotated as "hemolysin" are present within the genome. In both Sp7 and in Δ*cheY1* mutants, seven known components of the latter T6SS are expressed, including those components known to be necessary for T6SS function (ImpA, ClpV1, and EvpB) with abra_4441 VCA0107 and abra_4455 ImpA being up-regulated in Δ*cheY1* mutants over Sp7 control cells (Table 4.5). Additional hypothetical proteins and a protein of unknown function DUF796 are also detected in Δ*cheY1* cultures. In contrast, only abra_4428 σ^{54} interaction domain containing protein and abra_4452 Type VI secretion protein VCA0014 family are detected in Δ*cheB1cheR1* cultures, while no other components are detected (Table 4.5). In Sp7 control cultures as well as in each mutant culture, one protein from the former region containing genes for IcmF and DotU are detected. Nearby hypothetical protein abra_0639 and dctP TRAP dicarboxylate transporter abra_0643 are also detected in all cultures, suggesting this operon is expressed at similar levels in all cultures.

Type VI secretion systems were first discovered in *Rhizobium leguminosarium* in 1996 in mutants that were impaired in nodulation to pea plants [205] but not recognized as a distinct and conserved secretion system until 2006 [206]. This protein excretion system has now been discovered to be conserved among a wide range of bacteria and is present in about 100 bacterial genomes, primarily in pathogenic bacteria and non-

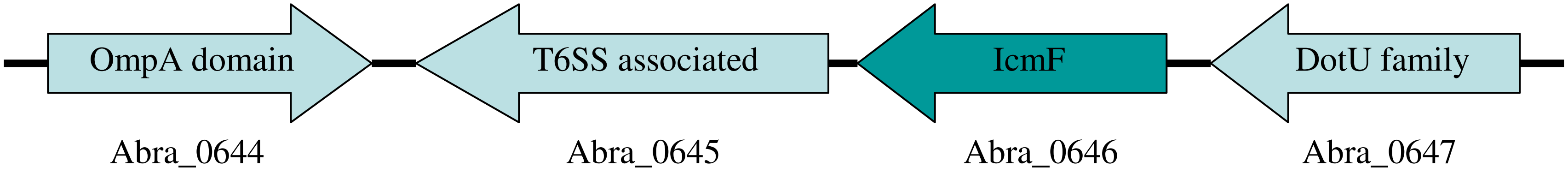


Figure 4. 4 Schematic depiction of genomic region containing type VI secretion system homologs to IcmF (abra_0646) and and DotU (abra_0647) Abra0645 is a type VI secretion associated protein of the BMA_A0400 family, while abra_0646 is an IcmF homolog, and abra_0647 is a DotU family homolog. Abra_0646 IcmF homolog is upregulated in CheY1 mutants, but not in Sp7 or in CheB1CheR1 mutants. IcmF homologs contain a transmembrane domain and a Walker A nucleotide binding motif, while DotU-like proteins contain a transmembrane segment in the C-terminal end, often with C-terminal extensions that show similarity to OmpA proteins. T6SS secretory apparatus are thought to function through association of IcmF and DotU family homologs.

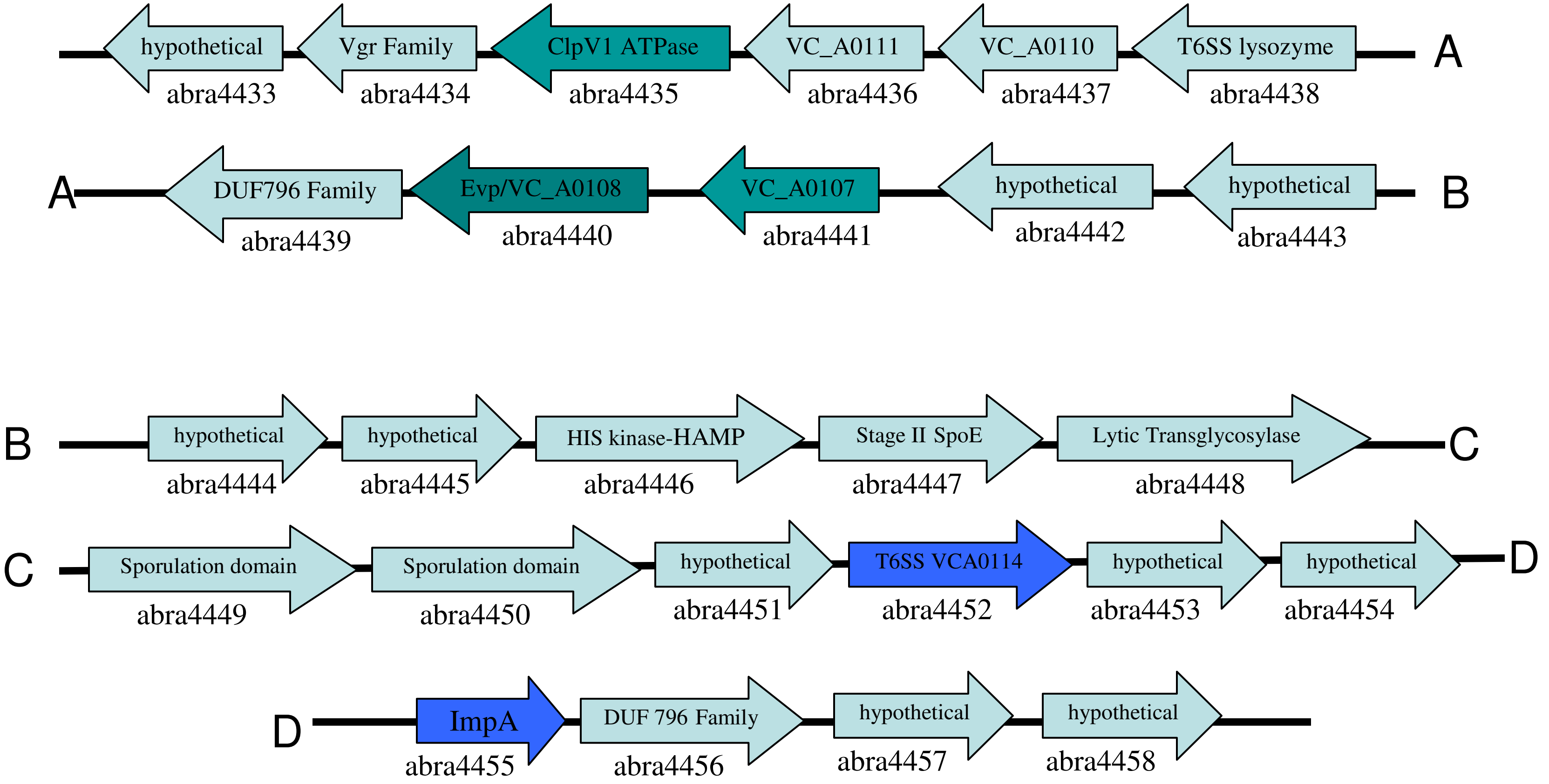


Figure 4. 5 Schematic representation of extended genomic region surrounding the type VI secretion system components in the Sp245 genome Black line represents genomic DNA, while arrows are representative of the direction/strand on which individual genes are located. Each individual line represents a possible grouping of genes in an operon structure based on distance between the genes. Dark aqua represents those components upregulated and CheY, while dark blue indicates those gene products that were detected in Sp7 only. The number of sporulation domain proteins and hypothetical proteins in this region combined with the upregulation of proteins only in smaller CheY cells present intriguing possibilities for the secretion system playing a role in morphological changes related to clumping and cell size.

pathogenic soil bacteria of the proteobacteria subclass [207]. Found in loci of 12 – 30 genes, with components located both within operon structures and also in close proximity to those operons, the T6SS is thought to mediate interaction with eukaryotic hosts [203] and to be involved in processes such as adherence, cytotoxicity and survival and persistence within a host cell [207].

Commonly recognized genes within T6SS encode proteins similar to those of *Legionella pneumophila* Type IV secretion proteins, IcmF (intracellular multiplication F) and DotU (defect in organelle trafficking) [203]. IcmF-like proteins, often called ImpA like the IcmF homolog originally found in *Rhizobium*, contain a transmembrane domain and a Walker A nucleotide binding motif, while DotU-like proteins contain a transmembrane segment in the C-terminal end, often with C-terminal extensions that show similarity to OmpA proteins [203]. Secretion function is dependent upon association of IcmF and DotU homologs in most known systems, although precisely how these two proteins interact and function to mediate secretion is not known [203]. As stated above, the Sp245 genomic region encompassing abra_0644 to abra_0647 contains genes encoding DotU and IcmF homologs, and all cultures show expression of genes in this region (Table 4.5). An additional conserved gene found within T6SS loci encodes a ClpB-family protein, ClpV, an ATP-ase whose function is unknown, although it is thought to unfold proteins for transport through the secretory apparatus and to provide energy for protein translocation [204]. Also common within T6SS loci are genes encoding putative outer membrane lipoproteins, forkhead associated (FHA) domain proteins, membrane-associated proteins and ATP-ases that may be contributing to regulation and/or secretion function [208]. A second extended region within the Sp245 genome encompassing genes abra_4433 to abra_4458 contains ClpV ATP-ase (Figure 4.5), which is detected only in Sp7 and *ΔcheY1* cells, but not in *ΔcheB1cheR1* cells.

The only known proteins to be secreted from the T6SS are Hcp (Hemolysin co-regulated protein), and VgrG (valine-lysine repeat protein G), which together are thought to compose part of the secretion apparatus of T6SS systems. Hcp forms hexameric rings that polymerize *in vitro* to form long tubes, and thus it is thought to function as a channel for protein export to extracellular medium [204]. VgrG proteins have a conserved region with similarity to T4 bacteriophage tail spike proteins, gp27 and gp5, which associate as heterotrimers to form a needle or spike that facilitates injection of DNA into a target bacterium [208, 209]. An additional conserved gene within T6SS clusters, VCA0109, shows similarity to T4 bacteriophage gp25, a component of the baseplate for the bacteriophage tail spike that serves to anchor it into the cell wall [208], making it likely that this protein also participates in the structural secretory apparatus. Additional effector proteins, RsbB ribose-binding protein in *Rhizobium* [203], and EvpB virulence factor protein in *Edwardsiella tarda* [206], are also thought to be secreted from T6SS. Of these, Vgr family protein (abra_4434), and EvpB (abra_4440) are found within the T6SS extended region in the Sp245 genome, while Hcp protein is not found within the genome at all. EvpB is expressed in both Sp7 and *ΔcheY1* cultures, but is not detected in *ΔcheB1cheR1* cultures (Table 4.5).

In *Vibrio anguillarum*, a pathogenic marine bacteria which causes hemorrhagic septicemia in fish, the T6SS operon contains 8 genes, 4 that encode known components of the T6SS and 4 that encode solute binding proteins, transport proteins and a D-ala-D-ala ligase [210]. The four atypical T6SS proteins in this species function as either positive or negative regulators of secretory function and control expression of extracellular proteases via affecting the expression levels of RpoS involved in stress response, and thus of VanT involved in quorum sensing [210], suggesting that it is possible that proteins encoded by genes in close proximity to or within the T6SS operon can serve a widely variant regulatory function. In the larger T6SS genomic region of *Azospirillum brasilense* Sp245 a number of hypothetical proteins, proteins of unknown function containing DUF domains, sporulation proteins, lytic transglycosylase and a His-kinase HAMP domain protein are present (Figure 4.5). Proteins such as sporulation domain proteins, lytic transglycosylase and HAMP domain proteins can be involved in cell remodeling on a number of levels, and their position in close proximity of the genes encoding theT6SS system suggest the intriguing possibility that this secretion system may be involved, either directly or indirectly, in morphological change of the cell.

In other bacterial species, expression of T6SS genes are known to be tightly controlled at a transcriptional level [203, 206, 207]. Transcription is regulated by members of the AraC family of transcriptional regulators and by RpoN (σ^{54}) [207, 211], and seems to be inversely related to transcriptional regulation of Type III secretion system components [206, 207]. Interestingly, Δ*cheY1* mutants show up-regulation of abra_1953 spermidine/putrescine ABC transporter, a polyamine transport system (Table 4.3). Spermidine transport pathways have been shown to regulate expression of Type III secretion systems in *P. aeruginosa* [212], and up-regulation of this transporter in Δ*cheY1* mutants but not in Δ*cheB1cheR1* mutants suggest it could play a role in the coordinated regulation of the Type III and Type VI secretion systems. A fold-change level or even detection of Type III secretion system components could not be determined because no proteins in the database are annotated as such.

In microarray studies, expression of T6SS components in *Vibrio cholerae* have been shown to be controlled through the action of the flagellar regulatory network, with motility and virulence being oppositely regulated [211]. Expression of T6SS components in vitro in *Pseudomonas aeruginosa* cultures can be induced by deleting the RetS gene, part of a two component system important in establishing chronic infection in the host cells [203]. Taken together, the above data suggests that T6SS expression can be controlled by two-component systems, and may be related to a variety of cell morphology changes such as those observed in the Δ*cheY1* mutants. Further, the abundant presence of T6SS components in Δ*cheY1* mutant cells in comparison to Δ*cheB1cheR1* mutant cells suggests that the Che1 operon is functioning, either directly or indirectly, to modulate expression of T6SS components, although further studies are needed to investigate this possibility.

Conclusions

Proteomics investigations give a snapshot picture of the proteins present at a given moment in time in the dynamic life of a cell. In this study we have examined the proteomes of mutants in the Che1 chemotaxis pathway grown to early mid log phase, and compared the detected proteins to those detected in a wild type Sp7 culture. Although all cultures were grown under the same conditions, some significant differences emerge between the detected proteome derived from the *cheY1* mutant lacking a functional forward signaling pathway, and that lacking an adaptation pathway (Δ*cheB1cheR1*).

Earlier studies revealed significant physiological differences between the two mutants, and the proteome further identifies differences in metabolic function between the two.

Patterns of up-regulation and down-regulation reveal very different metabolic function, especially in carbon and amino acid metabolism. Of particular interest is the increased abundance of phasin proteins, suggesting a tendency for the smaller Δ*cheY1* mutant to store carbon, and perhaps indirectly suggesting an altered sensing of the environment in Δ*cheY1* cells. An *Azospirillum* Sp7 PHB synthase mutant strain was constructed and characterized [213]. This mutant exhibited longer generation times and a reduced ability to deal with stress conditions and also had a reduced chemotactic response. Further, it produced greater amounts of exopolysaccharides (EPS) and exhibited increased aggregation [213]. This same phenomenon is noticed in Δ*cheY1* cells, which show an increased tendency to clump and altered EPS composition and/or amount.

The detection of type VI secretion system components in both Sp7 and *ΔcheY1* mutants, with up-regulation noted in *ΔcheY1* cells, is very interesting. Further, *ΔcheB1cheR1* cells show only a few components of the same system, suggesting that the Che1 operon two component signaling system has an effect, either direct or indirect, on the expression or operation of the Type VI secretion system. Further, four components of the same system are detected in Sp7 controls grown to late mid log phase, but none are detected in nitrogen fixing cells. Further investigation is needed to help clarify the role of this secretion system in *Azospirillum* cells.

Chapter 5. Development of a set of gateway-compatible destination vectors containing C-terminal tags

Introduction

In previous chapters of this work, proteomic surveys of *Azospirillum* cell cultures were conducted. These surveys yield a snapshot of proteins present at a single point in time, but give no information about the functionality of these proteins. Proteins rarely work alone within a cell, but instead function in association with other proteins. Thus, the putative function of an unknown protein can possibly be inferred from its known interacting partners. A number of hypothetical proteins and proteins of unknown function are detected in *Azospirillum* proteomes and are further seen to be up-regulated under different growth conditions or in chemotaxis mutants; however, as discussed, the function of most of these proteins remains unknown. Expression of proteins fused to an affinity tag provides an easy way to purify a variety of proteins using a uniform purification scheme [80]. A number of vector systems have been developed and optimized for heterologous expression and purification of fusion proteins from bacterial systems such as *E coli* [214-217] and *B. subtilis* [218], and for investigation of protein complexes in model systems such as *E. coli* [83] or *S. cerevisiae* [76, 79, 219]. However, although *E. coli* can be used for heterologous expression of proteins from other bacteria, exploration of protein interacting partners needs to be done in the system of interest (*in vivo*) to avoid false positives and facilitate interpretations. Organism-specific interacting partners may be uniquely present in a system of interest but may not have any homolog in *E. coli*. Ideally and to be widely applicable, vector sets with broad host range expression capabilities which allow ease of cloning and also provide flexibility in the choice of tag employed are needed for investigation of protein complexes in systems other than *E. coli* [85]. In this chapter, a set of vectors has been developed and tested to facilitate investigation of protein-protein interactions on a genome wide scale in a wide variety of non-model bacteria, for which dedicated vectors have not been developed.

Coupling affinity tag purification with mass spectrometry has been shown to be a viable method for purifying protein complexes on a large scale and subsequently identifying interacting partners using the tool of mass spectrometry [73, 74, 77]. A single affinity tag fused to a protein allows for a homogeneous purification method but can lead to a high degree of background contamination, making it difficult to determine legitimate interacting partners [80, 220]. Tandem affinity tags allow for a much "cleaner" purification than single step purification, and can also facilitate elution under native conditions when a protease cleavage site is included [220]. Further, affinity tags can influence the level of expression and the degree of solubility of a protein of interest in an individual protein-dependent manner [85]. Therefore, in considering a large-scale investigation of proteins and their interacting partners, it is desirable to have a set of affinity tags for flexibility and increased coverage of protein expression and investigation of interacting partners. Fodor et al [221] have created a modular broad host range vector system allowing for insertion of different promoters and containing either single or tandem (StrepII-FLAG) affinity tags for expression of proteins and investigation of protein complexes in native bacterial systems other than *E. coli* [221]. However, most of the vectors created have only a single affinity tag, and expression in bacterial systems other than *E. coli* relies on change of promoter through cloning into restriction sites. Cloning of a gene of choice into this set of vectors is done through restriction digest of genomic DNA or of PCR products and subsequent ligation into complementary restriction sites. Ligation of restriction-digested DNA into complementary vector restriction sites can be a time-consuming and ineffective process, so vectors containing a ligation-free cassette for cloning a gene of choice would definitely be an improvement.

Earlier work at Oak Ridge National Laboratory (ORNL) resulted in creation of a vector based on the pBBRMCS5 [222, 223] parental vector that could be used for high-throughput analysis of protein complexes in a bacterial system. These plasmids will stably replicate in a number of gram-negative bacteria [224], making them a versatile vector for expression and interaction studies in a wide range of bacterial species. To facilitate ease of cloning for high-throughput applications, a ligation-free cloning cassette used in Invitrogen Gateway® system [225] was included, along with a gentamicin resistance marker. In this work, the numbers of vectors in this newly designed tool box have been increased by inclusion of a variety of C-terminal tags for tandem affinity purifications under native elution conditions, allowing for complementary protein expression to give the greatest coverage of proteins and their interacting partners in a bacterial system of interest.

Earlier work in investigating protein complexes in *E. coli* [74, 76] and in *S. cerevisiae* [16, 77-79] utilized the technique of integration of a tagged version of the protein into the genome in order to reduce the likelihood of non-physiological interactions as a result of over-expression. In order to address this issue, we also developed a set of vectors based on a pJQ200KS [226] parental vector. The p15A origin of replication of pJQ200KS makes this parental vector non-replicative (and hence a “suicide” vector) in non-enterobacterial hosts [227]. A gentamicin resistance cassette is also included in this backbone vector for positive selection of those colonies possessing the plasmid containing the gene of interest in enterobacteria, or for positive selection of single crossover events in non-enterobacteria. To this backbone, we added the gateway cloning cassette with one of three different C-terminal tags. In this chapter, development of the above outlined set of vectors is chronicled, and characterization of the application of this tool set from immunoprecipitation experiments conducted with both integrating and exogenous expression vectors in *Rhodopseudomonas palustris*, an alpha-proteobacterium, are presented.

Materials and Methods

Bacterial strains and growth

Three strains of *E. coli* were used. Library Efficiency® DB3.1™ Competent Cells (Invitrogen, Carlsbad, CA) were used for vector backbone maintenance and propagation. Subcloning Efficiency™ *E. coli* DH5α™ Competent Cells (Invitrogen, Carlsbad, CA) were used for maintenance and propagation of entry and destination clones. *E. coli* S17-1 [228] chemically competent cells were used for conjugative transfer of pJQ200SK-based plasmids into *Rhodopseudomonas palustris*. DB3.1 cells were grown on Luria-Bertani (LB) media containing 10 µg/mL gentamicin and 100 µg/mL chloramphenicol for positive selection of vector backbone with correct tag inserts. DH5-α cells were grown on LB with either 10 µg/mL gentamicin for selection of positive transformants of destination clones, or with 50 µg/mL kanamycin for selection of positive transformants of entry clones. Plasmids are maintained in DH5-α cells, with glycerol stocks maintained at –80°C.

S17-1 chemically competent cells were prepared by standard protocols [229]. Briefly, cells were grown in LB broth overnight. Two ml of overnight culture was used to inoculate a 250 ml culture in LB containing 20 mM magnesium sulfate, which was then grown to an OD_{600nm} of 0.5. Cells were pelleted, and suspended in 20 mL of TFB-1 (30 mM potassium acetate, 50 mM manganous chloride tetrahydrate, 100 mM potassium chloride, 10 mM calcium chloride dihydrate, 15% w/v glycerol, pH 5.8). After a 20 minute incubation, cells were gently pelleted and resuspended in 1ml TFB-2 buffer (10 mM MOPS, 75 mM calcium chloride, 10 mM potassium chloride, 15% w/v glycerol). Aliquots were frozen with liquid nitrogen and stored at -80°C.

Transformation of plasmids into all *E coli* strains was done following standard protocols [229]. Briefly, chemically competent cells were incubated on ice for 1 hour with plasmid DNA, then heat shocked at 42°C for 45 seconds. After being placed on ice for 2 minutes, LB media was added and cells were allowed to recover through incubation with shaking at 37°C for 1 hour. Transformations were then plated out and incubated overnight at 37°C. When selecting for parent vectors with individual tags added, *E. coli* DB3.1 cells were grown on LB medium containing both 10 µg/ml gentamicin and 30 µg/ml chloramphenicol. When selecting for positive transformants containing vectors with genes of interest, all strains were grown on LB medium containing 10 µg/ml gentamicin.

R. palustris strain CGA010 was grown under photoheterotrophic conditions [30], in defined minimal media with 10 mM succinate as a carbon source, with 100 µg/mL gentamicin for selection of positive transformants or without antibiotics. Cells were grown anaerobically in the light at 25°C, with stirring for large scale cultures. Growth for small scale expression studies was done in 25 mL culture tubes, while 1.5 L cultures were grown for immunoprecipitation experiments.

Transformation of pJQ200KS-based plasmids into *R. palustris* strain CGA010 was accomplished through mating with *E. coli* strain S17-1 using standard protocols. Briefly, *R. palustris* cells were grown as described without antibiotics. S17-1 cells containing the pJQ-based plasmid with a gene of interest were grown overnight in LB media with gentamicin. After centrifugation and washing, S17-1 cells were mixed with *R. palustris* cells in a 1:5 ratio and the mixture spread on LB plates overnight. Selection for *R. palustris* clones containing the gene of interest within the genome structure due to a single crossover event was accomplished by anaerobic growth on minimal media plates containing 100 µg/mL gentamicin. Transformation of pBBR-Dest42-based plasmids into *R. palustris* strain CGA010 was accomplished through electroporation, as described elsewhere [11, 81].

Construction of pBBR-based destination vectors

The parent backbone vector was based on pBBR1MCS [223] with gentamicin resistance, which was modified to include a Gateway cloning cassette followed by tandem C-terminal V5 and 6xHIS epitopes and named pBBR5-Dest42 [81]. The pBBR5-Dest42 vector backbone was restriction digested with BstB1 restriction endonuclease (New England Biolabs, Ipswich, MA), to linearize the vector and remove the HIS-V5 epitope. The resulting digest was run out on an agarose gel, and the 6000 bp band representing the vector backbone was then gel-purified using a QIAquick Gel Extraction Kit (Qiagen, Valencia, CA) following manufacturer's protocols.

Amino acid sequences for the components of the TAP tag (ProteinA-TEV protease site-Calmodulin Binding Protein (CBP)) were obtained from Rigaut et al. [80] and those for the SPA (3xFLAG-TEV protease site- CBP) tag were obtained from Zeghouf et al [75]. For the STH (StrepII-TEV protease cleavage site-6xHIS) tag, the strepII sequence was obtained from Schmidt et al. [230]. A 4-cysteine motif (CCPGCC) was added to facilitate detection of recombinant protein expression [84] and an additional TEV protease site [231] was included in the tag constructs for more efficient proteolytic cleavage of the outer tag. The modified affinity tags are termed TAP2, SPA2, and STH. Amino acid sequences and molecular weights for each tag are given in Table 5.1. Amino acid sequence for each tag construct was back-translated using VectorNTI (Invitrogen, Carlsbad,CA) and codons optimized for expression in *R. palustris*, which possesses a high-GC content genome. Codon optimization tables can be obtained at http://genome.ornl.gov/microbial/rpal/. DNA sequences for the TAP2 tag, the SPA2 tag and the STH tag were synthesized by GenScript (Piscataway, NJ) and placed in pUC vectors [232]. Insertion of the newly designed C-terminal tags into parent backbone vectors is illustrated by the flow chart presented in Figure 5.1. Tag sequences were PCR amplified from each vector using M13 universal primers (M13 forward 5' CGC CAG GGT TTT CCC AGT CAC GAC 3'; M13 reverse 5' GAG CGG ATA ACA ATT TCA CAC AGG 3'). PCR product was then restriction digested using Cla1 restriction endonuclease (New England Biolabs, Ipswich, MA), creating single stranded DNA ends with complementary sequence to those created in the parent pBBR5-Dest42 after BstB1

Table 5. 1 C-terminal tags and sequences included in pBBR-based and pJQ200KS-based vectors

Tag Name	Tag Motif	Amino Acid Sequence*	length	Molecular weight (kDa)
STH	4Cys – 2x StrepII tag – 2x TEV protease site – 6x HIS	CCPGCCASAWSHPQFEKSGWSHPQFEKGGTGSENLYFQGGRGGSENLYFQGEGTGSHHHHHH	239bp	6.7
SPA2	4Cys- Calmodulin Binding Protein (CBP) -2xTEV-3xFLAG	CCPGCCASKRRWKKNFIAVSAANRFFKKISSSGALDYDIPTTASENLYFQGGRGGSENLYFQGELDYKDHDGDYKDHDIDYKDDD	314bp	9.6
TAP2	4Cys- CBP- 2xTEV - 2xProteinA	CCPGCCASKRRWKKNFIAVSAANRFKKISSSGALDYDIPTTASENLYFQGGRGGSENLYFQGELKTAALAQHDEAVDNKFNKEQQNAFAEILHLPNLNEEQRNAFIQSLKDDPSQSANLLAEAKKLNGAQAPKGVDNKFNKEQQNAFYEILHLPNLNEEQRNAFIQSLKDDPSQSANLLAEAKKLNGAQAPK	632bp	21.2

* Underlined portions indicate the amino acid sequence of the motifs separated by lines in the motif area

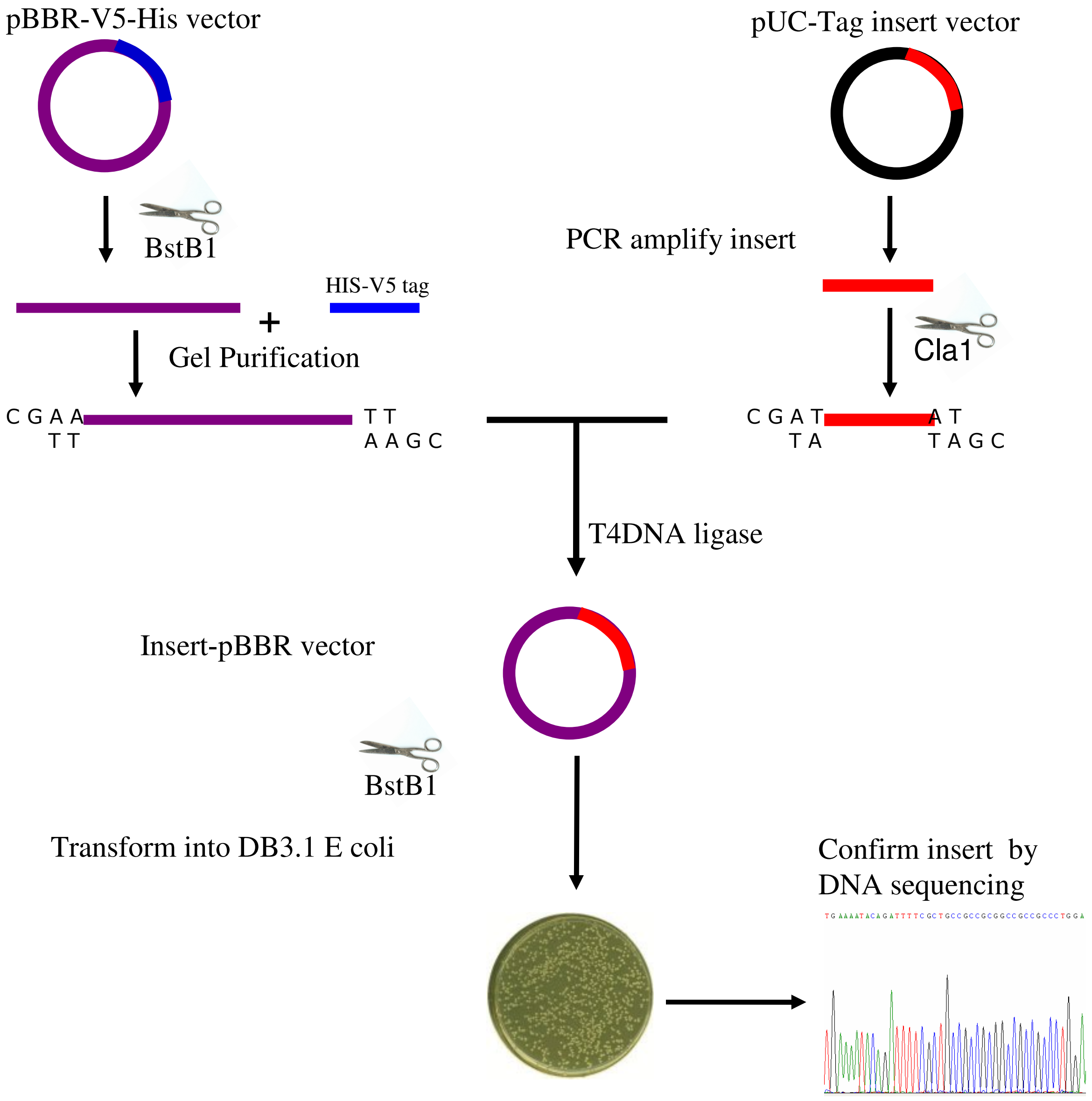


Figure 5. 1 Flow chart of vector construction steps
The parent vector pBBR-Dest42 with HIS-V5 tandem tag was restriction digested with BstB1, while the PCR product amplifying the tag construct was digested with restriction endonuclease Cla1, leaving complementary overhanging DNA sequence for easier ligation. Tag sequence was then ligated into gel-purifed parent backbone vector, with resulting ligation mixture transformed into DB3.1 *E. coli* cells. Positive clones were confirmed by diagnostic restriction digest and DNA sequencing.

restriction digest. The resulting fragment was ligated into the BstB1 site of the gel-purified pBBR5-Dest42 vector using T4 DNA ligase (Promega, Madison WI) following manufacturer's instructions with the following exception. In order to increase yield, the ligation reaction was allowed to proceed overnight at 17°C before transformation into DH5α cells. Ligated parent vectors were transformed into DH5α *E. coli* cells following standard transformation protocols [229] , as described above. Five colonies of each type of backbone vector clone were chosen for overnight growth in liquid media and subsequent DNA plasmid mini-prep (Qiagen, Valencia, CA). Positive clones were confirmed by diagnostic restriction digest and by DNA sequencing using universal M13 primers.

Construction of pJQ200KS-based destination vectors

The pJQ200KS (5370bp) vector backbone was linearized by digestion with Xba1 restriction endonuclease (New England Biolabs, Ipswich, MA). DNA sequence including the Gateway® cassette (attR sites, chloramphenicol resistance cassette, and ccdB gene) with the added C-terminal tag (~2100bp) was PCR amplified using T7 Universal primer (5' TAA TAC GAC TCA CTA TAG GG 3') and TagRevSeq (5' CGA CCG GGT CGA ATT TGC 3'). The resulting PCR product was digested with Xba1 restriction endonuclease (New England Biolabs, Ipswich, MA), and then ligated into the vector backbone using T4 DNA ligase (Promega, Madison WI) following manufacturer's instructions with the following exception. The ligation mixture was incubated overnight at 17°C in order to ensure a higher number of correct ligations. Ligated pJQ200KS-tag vectors were transformed into DB3.1 *E. coli* cells following standard transformation protocols [229]. Transformants were plated on LB-agar plates containing both 10 µg/ml gentamicin and 30 μg/ml chloramphenicol, and five individual colonies were chosen for plasmid DNA mini-prep as above. Positive clones for the base parent vectors containing C-terminal tags were confirmed by restriction digest and sequencing.

Protein expression

Entry clones for individual genes had been created earlier [81], and a small set of genes within these entry clones was chosen for expression tests using the newly created vectors. Recombination reactions were performed between the entry clones and the newly created parent vectors using LR clonase enzyme mix (Invitrogen, Carlsbad CA) following manufacturer's directions. To increase the yield of recombinants, the LR reaction was allowed to proceed overnight at 17°C. Resulting destination clones (a set of 5 genes tagged with each of the three individual tags in pBBR-DEST42 parent vectors) were transformed into DH5α *E. coli* cells, and plated on LB media plates containing 10 μg/ml gentamicin. The presence of a wild-type DNA gyrase in these cells allows for negative selection of nonrecombinant plasmids. Expression of the *ccdB* gene found in the Gateway® recombination cassette in the parent vector interferes with the function of wild-type DNA gyrase, thus preventing replication in cells with wild-type DNA gyrase such as DH5α *E. coli* (Invitrogen). Five resulting colonies containing destination clones were chosen for further expression studies. 15 mL cultures were inoculated from these colonies, and grown overnight at 37°C with shaking. Plasmid mini-preps were done using QiaPrep spin mini-prep columns (Qiagen, Valencia CA) and DNA concentration was determined using the NanoDrop spectrophotometer (Thermo Scientific, Wilmington, DE). Additionally, glycerol stocks were made using 1 ml of the overnight cultures.

Destination clones created from the LR recombination reaction were transformed into *R. palustris* CGA010 cells as described above, either by mating for pJQ200KS-based clones or by electroporation for pBBR-Dest42-based clones. Transformants were initially plated on PM-10 plates [30] containing 100 µg/ml gentamicin for positive selection of clones containing plasmid and grown anaerobically with light. Individual colonies were then used to inoculate 25 mL cultures of PM-10 media [30] containing 100 µg/ml gentamicin. Small cultures were harvested by centrifugation at 2000 xg for 5 minutes in a Sorvall tabletop centrifuge, washed, and divided into 2 equal aliquots. One pellet was frozen at -80°C, while the other was lysed using 50 µl of PBS containing Bugbuster® reagent (Novagen), 1 µl/ml benzonase nuclease (Novagen), lysozyme, 1 mg/ml leupeptin and 1% PMSF. An aliquot of lysate was run out on a Precise protein 10-20 % gradient SDS poly-acrylamide gel (Pierce Biotechnology, a division of Thermo Scientific, Rockford, IL) and used for characterization of expression. Western blot analysis was performed using mouse anti-FLAG antibody (Sigma-Aldrich, St Louis MO) for the primary incubation for SPA2 tagged proteins, rabbit anti-CBP antibody (Upstate, a division of Millipore, Billerica MA) for TAP2 tagged proteins, and mouse anti-HIS antibody (Sigma-Aldrich, St Louis MO) for STH tagged proteins. Secondary incubation was done either with goat anti-mouse or goat anti-rabbit IgG antibody complexed to horseradish peroxidase as appropriate depending on the initial antibody incubation. Fusion protein expression was visualized after DAB (3,3´-diaminobenzidine tetrahydrochloride) reaction with horseradish peroxidase following manufacturers recommendations (Pierce Biotechnology).

Once protein expression was confirmed for tagged proteins, colonies from fresh PM-10 plates containing 100 µg/ml gentamicin were used to inoculate 25 mL *R palustris* PM-10 starter cultures. Starter culture tubes contained 100 µg/mL gentamicin, while large cultures did not contain antibiotics. Starter cultures were used to inoculate 1.5 L production cultures, which were then grown to mid-log phase under anaerobic conditions in the light with stirring as described above. Cultures were harvested at an OD_{660} of 0.5 – 0.7. Cells were pelleted by centrifugation at 6500 xg, then washed 3x with phosphate-buffered saline (PBS). After resuspension in PBS, cells were split in two equal aliquots and re-pelleted. Pellet from one aliquot was frozen at -80°C, while the other pellet was resuspended in appropriate lysis buffer at a rate of 0.5 ml lysis buffer/ 0.1g cell pellet weight. Lysis buffers contained 1x Bugbuster® reagent for chemical lysis of cells diluted in binding buffers compatible with initial purification steps (M2 buffer for SPA2 and TAP2 tagged proteins, BF3-FT buffer for STH-tagged proteins, Table 5.2). Benzonase nuclease (1 µl/ml, Qiagen) was included in lysis buffers in order to digest genomic DNA released on cell lysis. PMSF (10 mg/ml) and leupeptin (10 mg/ml) were included for inhibiting protease activity, and lysozyme was included to help degrade the cell wall. Additionally, beta-mercaptoethanol (0.2 µl/ml) was added to the lysis buffer in order to break up disulfide bonds. Cell suspension was incubated on a rotating wheel for 1 hour at room temperature. Lysate was cleared through centrifugation in a Sorvall centrifuge at 18000 xg for 20 mintues at 4°C, and supernatant transferred to a clean 15mL falcon tube. Cleared lysates were frozen at -80°C for later immunoprecipitation experiments.

Table 5. 2 Buffer composition for each affinity purification step

Purification step	SPA and TAP tag purification buffer	STH tag purification buffers
Lysis	**Lysis Buffer:** 10X Bugbuster reagent M2 Binding Buffer Benzonase Nuclease (25U/ul stock) 50uL Lysozyme 100ug/ml PMSF 10 ug/ml Leupeptin	**Lysis Buffer:** 10X Bugbuster reagent M2 Binding Buffer 25U Benzonase Nuclease Lysozyme 100ug/ml PMSF 10 ug/ml Leupeptin
Initial affinity purification step	**Flag or IgG binding -M2 buffer:** 10 mM Tris-HCl @ pH 8.0 100 mM NaCl 10% glycerol 0.1% Triton-X 100	**HIS binding buffer – BF3-FT** 50 mM Tris-HCl @ pH 8.0 150 mM NaCl 50 mM NaH2PO4 10 mM imidazole 10% glycerol
Tev protease cleavage	**TEV cleavage buffer** 50 mM Tris-HCl @ pH 7.9 100 mM NaCl 0.2 mM EDTA 1mM DTT 0.1% Triton X-100	**TEV cleavage buffer** 50 mM Tris-HCl @ pH 7.9 100 mM NaCl 0.2 mM EDTA 1mM DTT 0.1% Triton X-100
Second affinity purification step	**Calmodulin Binding Buffer:** 10 mM Tris-HCl @ pH 7.9 100 mM NaCl 2 mM $CaCl_2$ 10 mM 2-mercaptoethanol	**StrepII-tag Binding Buffer** 50 mM Tris-HCl @ pH 7.9 100 mM NaCl 0.2 mM EDTA 1mM DTT 0.1% Triton X-100
Wash after 2nd affinity purification	**Calmodulin Wash Buffer** 10 mM Tris-HCl @ pH 7.9 100 mM NaCl 0.2 mM $CaCl_2$ 10 mM 2-mercaptoethanol	**StrepII-tag wash Buffer** 100 mM Tris-HCl @ pH 7.9 150 mM NaCl 1 mM EDTA 1 mM DTT
Elution	**Calmodulin Elution Buffer:** 10 mM Tris-HCl @ pH 7.9 100 mM NH_4HCO_3 3 mM EGTA 10 mM 2-mercaptoethanol	**StrepII-tag Elution buffer** 100 mM Tris-HCl @ pH 8 150 mM NaCl 1 mM EDTA 20 mM desthiobiotin

Immunoprecipitation

SPA2 and TAP2 affinity purification was done according to protocols established earlier [11]. STH-tagged proteins were affinity purified following protocols established earlier [84], modified for use in bacterial cells and for batch processing mode. Aliquots (200 µl) of resins for initial tag capture (Table 5.3) were washed three times in appropriate binding buffer, and resuspended in appropriate binding buffer to create a 50% bead slurry. Cleared lysates were incubated with bead slurry for 2-4 hours with 200 µl of washed beads for affinity capture of proteins via the outer tags. For TAP-tagged proteins, IgG Sepharose™ 6 Fast Flow resin (Amersham Biosciences) was used for Protein A capture, while Anti-FLAG® M2 Affinity Gel beads (Sigma-Aldrich) were used for 3xFLAG capture in SPA tagged proteins. Nickel-NTA agarose resin (Qiagen) was used for initial purification of STH-tagged proteins. Resins used for each step of affinity capture for each tag are presented in Table 5.3. After incubation, resins with associated proteins were washed 3 times with appropriate binding buffer, and a final wash with 1 mL of TEV protease buffer. After the final wash, the beads were centrifuged at 1500rpm for 30 seconds, all of the wash buffer removed, and beads resuspended in 200 µl TEV protease buffer. Ac-TEV protease (10 µl, 50U, Invitrogen, Carlsbad, CA) was added to the bead suspension, and the suspension was then incubated on a rotating wheel at room temperature for 30 minutes. Incubation was continued overnight on a rotating wheel at 4°C due to the dramatic reduction in efficiency of TEV protease cleavage at low temperatures. After this incubation, proteins and associated binding partners are present in the supernatant.

Table 5. 3 Resins used for capture in each affinity purification step

Affinity purification	Initial (outer tag) capture	Second (inner tag) capture
SPA affinity purification	FLAG tag capture: Anti-FLAG M2 agarose beads (Sigma-Aldrich)	CBP tag capture: Calmodulin Sepharose 4B (Amersham Biosciences)
TAP affinity purification	Protein A tag capture: IgG sepharose beads (Amersham Biosciences)	CBP tag capture: Calmodulin Sepharose 4B (Amersham Biosciences)
STH affinity purification	6xHIS tag capture: NiNTA beads (Qiagen, Valencia, CA)	StrepII tag capture: Strep-tactin beads (IBA, St Louis)

The second step of affinity purification began with washing a 200 µl aliquot (per sample) of resins required for capture of the only remaining inner tag (CBP for both TAP2 and SPA2 fusion proteins, StrepII for STH fusion proteins). After 4 washes with appropriate binding buffer, bead slurries were centrifuged at 1500rpm for 30 seconds and final wash buffer removed. For CBP capture, the beads were resuspended in 200 µl of appropriate binding buffer, and 1.2 µl of $CaCl_2$ (1M) was added. For StrepII capture, beads were resuspended in StrepII binding buffer. Bead slurries from the TEV protease cleavage step were centrifuged at 1500rpm for 30 seconds to pellet the beads. Supernatant from the TEV protease cleavage step was added to the appropriate bead slurry for secondary capture. Suspension was incubated at 4°C for 3 hours.

Elution was accomplished in three steps. The bead slurries were washed 4 times in appropriate wash buffer. Elution for TAP2 and SPA2 tagged proteins included chelation of calcium from the beads as calmodulin binding protein requires calcium for binding to affinity resins. So, elution buffers included 3 mM EGTA, a chelating agent that is specific for calcium. Elution for StrepII tagged proteins is based on replacing the tagged protein bound to the streptavidin resin with biotin in the elution buffers. Although the manufacturer's protocol suggested 2 mM biotin for elution, we found that concentration to be ineffective for elution. Therefore, we used elution buffers containing 20 mM biotin, which proved to be more effective for elution of tagged protein from the resin. Elution buffer (100 µl) was added to washed beads with protein bound, and incubated at room temperature for 10 minutes. Beads were then pelleted by centrifugation at 1000 xg for 30 seconds, and supernatant was removed to a clean centrifuge tube. This step was repeated twice more, and the three eluates were pooled for a total of 300 µl, and frozen at -20°C until digestion.

Proteolytic digestion with sequencing-grade trypsin (Promega) was accomplished through the addition of 1 µg of trypsin directly to the eluates. The pH was checked to ensure the eluates were at a proper pH between 7 and 8. Calcium chloride (13 mM) was added to the SPA2 and TAP2 eluates due to the presence of EGTA in the eluates. Although trypsin will digest protein without calcium present, specificity is improved in the presence of calcium. For this reason calcium was replenished in those samples. Digestion was allowed to proceed overnight at 37°C. Proteolytic peptide samples were dried down in a vacuum centrifuge and buffer exchanged into Buffer A (95% HPLC-grade water, 5% acetonitrile, 0.1% formic acid). Samples were frozen at -80°C for later analysis by mass spectrometry.

LC-MS/MS

Analysis of proteolytic peptides was accomplished via an automated one-dimensional chromatographic separation followed by mass spectrometry of eluted peptides. Chromatographic setup consisted of a FAMOS autosampler for injection of individual samples to a C-18 reverse phase trapping column, followed by elution to a resolving column via a Switchos and UltiMate HPLC (LC Packings, Sunnyvale, CA) which was coupled directly to a nanoelectrospray source (Proxeon Biosystems, Odense, Denmark) in line with an LTQ linear ion trap mass spectrometer (Thermo-Finnigan, San Jose, CA). The resolving column consisted of 18 cm of 3 µm Jupiter C18 reverse phase material (Phenomenex, Torrance, CA) packed via pressure cell into 100µ ID fused silica, with a tip pulled to an opening of 5 µm (PicoTip, New Objective). This column was directly coupled to the LTQ, and peptides were directly eluted from this column into the mass spectrometer. Samples were run in triplicate in random order, with a single blank sample of buffer A (95% water, 5% acetonitrile, 0.1% formic acid) run between each immunoprecipitation sample in order to minimize carryover between samples. Peptides were eluted from the resolving column in a gradient elution of 0-50% buffer B (30% HPLC-grade water, 70% acetonitrile, 0.1% formic acid). The LTQ was operated under control of Xcaliber software. Data was collected in data dependent mode, with 1 full scan followed by 5 dependent scans in which 2 microscans were used to create each spectrum. Dynamic exclusion was employed, with a repeat count of 1, exclusion list size of 300 and a repeat duration of 180.

Data analysis

Experimentally derived mass spectra were searched against a database of amino acid sequences from the CDS of *R. palustris (*located at http://compbio.ornl.gov/rpal_proteome/databases) using SEQUEST search algorithm [3]. The database was created by first appending a list of common contaminants and then reversing the amino acid sequences and concatenating the reverse sequences to the forward sequences. One tryptic cleavage site was specified for the searches. DTASelect [54] was used to filter and sort the data, with default values (Xcorr of 1.8 for +1 peptides, 2.5 for +2 peptides, and 3.5 for +3 peptides, and deltCN of 0.08) being employed. Tables were compiled in Microsoft Access.

Results

A set of vectors containing three different C-terminal tandem affinity tags with the backbone based on the pBBR-Dest42 parent vector [81] has been created. As shown by the plasmid map in Figure 5.2, pBBR-Dest42 parent vector offers several advantages, including a pBBR replicon for expression in a wide variety of bacterial species, a Gateway® compatible cloning cassette (Invitrogen, Carlsbad, CA), gentamicin resistance for positive selection of clones, and a mobilization region for conjugative transfer to other bacterial species. C-terminal affinity tags included modified Tandem Affinity Purification (TAP –2xCalmodulin Binding Protein(CBP) -Tobacco Etch Virus (TEV) protease site –2xProtein A) tag [80], a modified Sequential Peptide Affinity (SPA-2xCBP - TEV protease site - 3xFLAG epitope) tag [75] or a StrepII-TEV protease site-6xHIS (STH) tag [233], as shown in Figure 5.3. A further advantage to the newly created tag systems is the inclusion of a tetra-cysteine epitope that allows for easy detection of protein expression in lysates, eliminating the necessity of performing a time-consuming western blot procedure.

A second set of vectors was created based on the parent vector, pJQ-200KS [226], which contains a p15A origin of replication, making it non-replicative in non-enterobacteria. Thus it is a “suicide plasmid” which integrates into the genome of non-enterobacteria, allowing for expression of tagged protein from a native promoter sequence. The entire cloning cassette including the Gateway cloning cassette combined with the C-terminal tandem affinity tags described above was placed in the pJQ-200KS backbone vector. Tag sequences were confirmed for both pBBR-Dest42 based vectors

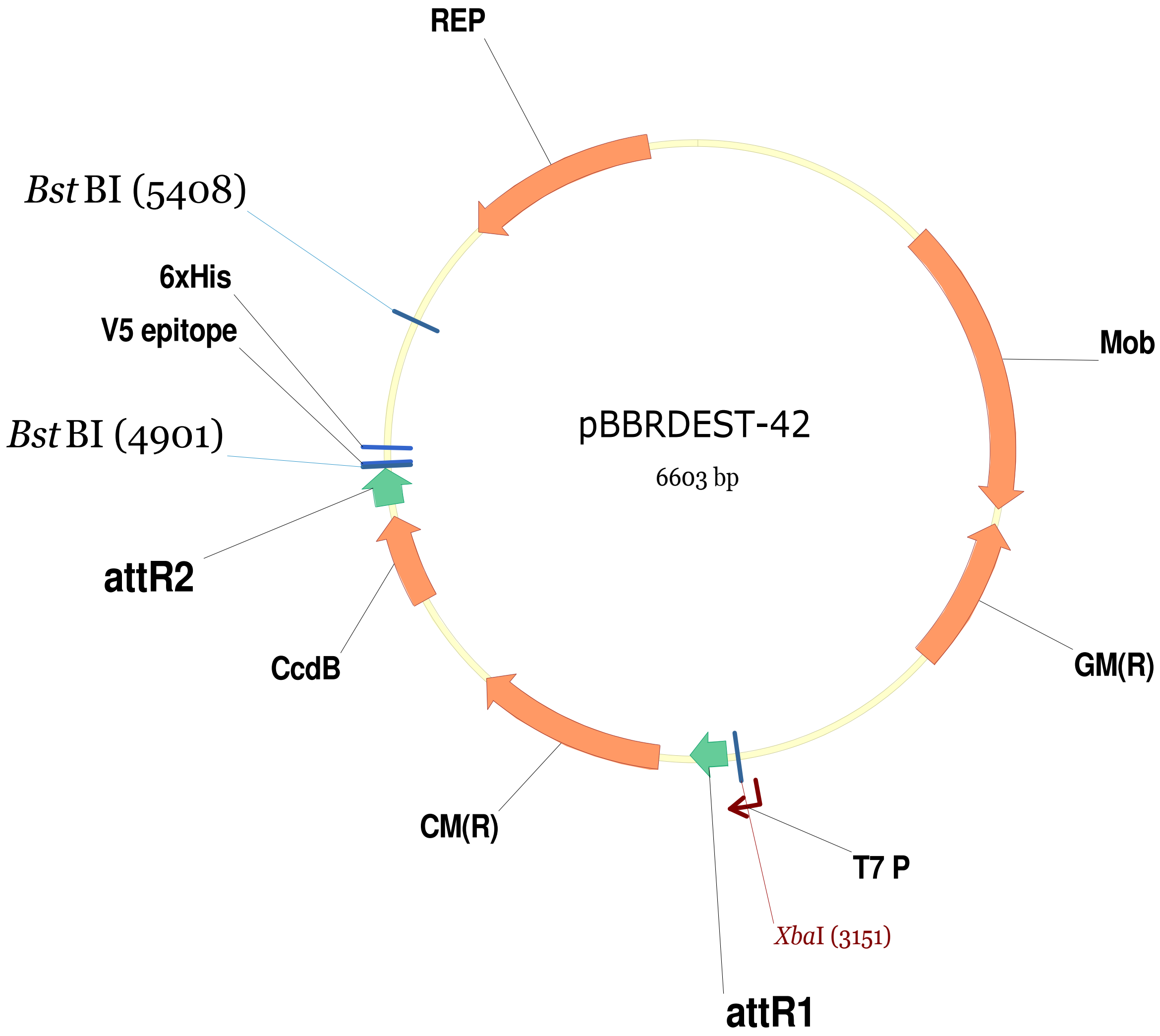


Figure 5. 2 Schematic drawing of the initial pBBR-Dest42 parent vector
The parent vector backbone used for this work was derived from the pBBR1MCS5 vector. Modifications to the initial pBBR1-MCS5 vector included addition of a gateway cloning cassette containing a chloramphenical resistance gene and a ccdB gene flanked by lambda bacteriophage attR attachment sites. A tandem affinity tag consisting of 6xHIS residues followed by a V5 epitope tag was added C-terminal to the cloning cassette. The tandem affinity HIS-V5 tag was flanked by BstB1 restriction sites such that the tag could be removed and a repertoire of different tags inserted.

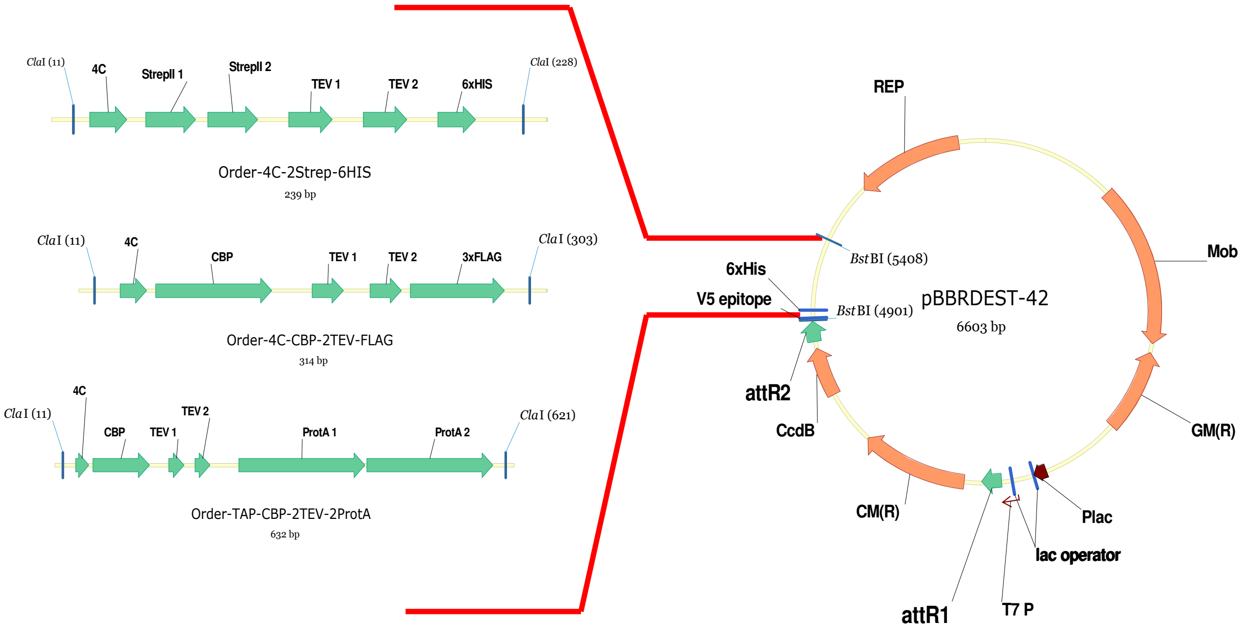


Figure 5. 3 New set of pBBR-Dest42-based vectors showing expansions of the three new C-terminal tag constructs
pBBRDest42 parent vector contains all aspects of the pBBRMCS5 parent vector (genes involved in mobilization of the vector (mob) allowing for conjugative transfer of the vector from a strain containing tra ????? genes in trans, gentamicin reisistance genes and rep genes involved in the replication of the plasmid) in addition to the attR sites for Gateway compatible recombination flanking the cassettes for chloramphenical resistance and ccdB gene for positive selection of transformants. The C-terminal 6xHis-V5 combinatorial tag was removed by restriction digest with BstB1, and the tags shown at left were digested with CLa1 restricion endonuclease and subsequently ligated into the complementary BstB1 site of the digested parent vector. Correct orientation of the tag into the parent vector was confirmed by restriction digest and by sequencing.

and for pJQ200KS-based vectors by both restriction digest and DNA sequencing as described above.

Expression of RpoA, the 36.5 kDa alpha subunit of RNA polymerase, was accomplished in all pBBR-Dest42 based vectors with all tags. Immunoprecipitation experiments were also successful with all RpoA fusion proteins with all tags, with both bait proteins and interacting partners detected via mass spectrometry (Table 5.4). The higher molecular weight beta prime subunit (155.2 kDa) of RNA polymerase (RpoC) was successfully expressed only in combination with the lower molecular weight tags, both SPA2 and STH. Expression of an RpoC-TAP2 fusion from pBBR-Dest42 based plasmids was unsuccessful. RpoC fusion proteins with lower molecular weight tags were successfully purified and bait proteins with associated interacting partners were detected via mass spectrometry. Sequence coverage of detected proteins and interacting partners are presented in Table 5.5. Interestingly, although bait protein was detected in all replicates of all tags, the SPA2 and TAP2 tags resulted in greater sequence coverage for both bait proteins and interacting partners than did the STH tag. TAP and SPA tags were originally tested and optimized for use in bacteria and/or yeast, while the STH tag was optimized for use in mammalian systems.

Integrating vectors were tested only with an *rpoA* gene (rpa_3226) fused to a SPA2 tag. Expression of fusion protein RpoA-SPA2 was confirmed via western blot and mass spectrometry. Both bait protein and interacting partners were identified in mass spectrometry experiments, although no discernible difference in the numbers of non-specifically bound proteins was detected between the integrated version of RpoA-SPA2

Table 5. 4 Sequence coverage of interacting partners for RpoA (rpa3226) detected using mass spectrometry

3226 Interacting partners (http://string.embl.de/newstring_cgi/)		RpoA 3226-			
Gene	Description	-SPA	-STH	-TAP	-pJQ-SPA
rpa3268	rpoB, DNA-Directed RNA polymerase subunit beta	21	2	29	1.5
rpa3267	rpoC, DNA-directed RNA polymerase subunit beta prime	10	2	18	1
rpa3227	rpsK, 30S ribosomal protein S11	ND	ND	ND	ND
rpa3225	rplQ, 50S ribosomal protein L17	ND	ND	14.4	ND
rpa3230	secY,preprotein translocase SecY	ND	ND	ND	ND
rpa1288	rpoD, RNA polymerase sigma factor	ND	ND	ND	ND
rpa3228	rpsM 30S ribosomal protein S13	ND	10.2	ND	ND

Table 5. 5 Sequence coverage of interacting partners for RpoC (rpa3267) detected via mass spectrometry

3267 Interacting partners (http://string.embl.de/newstring_cgi/)		RpoC 3267-	
Gene	Description	-SPA	-STH
rpa3268	rpoB, DNA-Directed RNA polymerase subunit beta	38	2
rpa3226	rpoA, DNA-directed RNA polymerase subunit alpha	45	6
rpa2693	relA, GTP pyrophosphokinase	ND	ND
rpa3269	rplL, 50S ribosomal protein L7/L12	ND	ND
rpa2886	pyrG, CTP synthase	ND	ND
rpa1288	rpoD, RNA poymerase sigma factor	14.5	ND
rpa3270	rplJ, 50S ribosomal protein L10	ND	8.6
rpa3056	ndk, nucleoside diphosphate kinase	ND	ND
rpa0438	nusA, putative NusA protein transcriptional terminator	ND	2.3
rpa2692	rpoZ, DNA-directed RNA polymerase subunit omega	23	ND

and the exogenously expressed RpoA-SPA2. Interestingly, much lower sequence coverage was noted for RNA polymerase subunits when using an integrating vector than when using a plasmid based strategy (Table 5.4), although lysis and immunoprecipitation protocols were identical.

Single crossover events are the most common occurrence when using pJQ200KS in non-enterobacteria [226]. As a result of a single crossover event, the pJQ-SPA2 plasmid would be integrated into the genome sequence in its entirety, with the tagged version (including a stop codon) being translated while the untagged version is not. After a single crossover event, PCR amplification of the genomic region flanked by upstream genomic sequence and by common DNA sequence from the tag region should provide a good diagnostic tool for detection of plasmid integration into the genome. Band size obtained from this diagnostic PCR should reflect the size of the integrated gene plus the size of a partial tag sequence. For instance, the expected size of the rpa3266 (*rpoA*) gene (1019 base pairs) plus a partial SPA tag sequence including the 4C epitope plus the Calmodulin Binding Protein epitope (261 base pairs) is 1280 base pairs. In order to confirm integration of the tagged version of the gene into the genome, colony PCR of *R. palustris* RpoA-SPA2 clones was attempted using a 5' forward primer corresponding to genomic sequence upstream of the *rpoA* gene and a 3' reverse primer corresponding to DNA sequence in the TEV protease site. A control consisting of colony PCR of *E. coli* cells containing the pJQ200KS-SPA2 plasmid with an *rpoA* gene was performed in tandem with the samples, and the results run out on an agarose gel. Unfortunately, only control bands were present on the agarose gels, so establishment of integration of the SPA2 tagged version of the *rpoA* gene into the genome was unsuccessful.

Lastly, a test of detection of protein expression was attempted using the 4C epitope in combination with FlAsh reagent (Invitrogen). Lysates were combined with FlAsh reagent and proteins were separated on an SDS-PAGE gel, following manufacturer's directions. After unsuccessful attempts to visualize protein presence when using samples that had given positive results using a western blot protocol, a test was done to determine whether the Bugbuster® reagent (Novagen) used for chemical lysis of cells was interfering with the fluorescence of the dye. To test for interference of lysis reagent with fluorescence, M2 lysis buffer containing either Bugbuster® or water was used to dilute fluorescein dye. Molar concentrations of fluorescein varied from picomolar (10^{-12}) to millimolar (10^{-3}) concentration, representing concentrations of protein possibly found within affinity purified samples. Dye solutions were spotted onto glass microscope slides and visualized with a Bio-Rad fluorescent imager using Quantity One software. Results indicated that fluorescence was dramatically reduced by the Bugbuster® reagent, with molar amounts less than micromolar (10^{-6}) not being detected in the Bugbuster® solution, while nanomolar (10^{-9}) concentrations were detected in the water solution. Mechanical lysis methods (3 different sonication protocols and freeze-thaw methods) in lysis buffers which did not contain detergent had been attempted earlier and were unsuccessful, with little or no protein detected for any fusion protein. Due to a lack of time, this methodology was not attempted again, so the detection of protein expression in cell lysates via the 4C epitope remains untested for these tags.

Discussion

Proteins have a wide variety of sizes, sequences and structures. Thus, uniform purification strategies such as those used in isolation of DNA are likely to be unsuccessful in isolating a large number of proteins. Instead, the wide variety of protein sequences and structures require optimization of purification strategies for each individual protein. Addition of affinity tags to a protein allows for simplified generic purification strategies that target the tag motif and can facilitate isolation of proteins and their interacting partners on a genome-wide scale. The parental plasmid used in this chapter, pBBR-Dest42 [81] offers a number of attractive advantages for protein expression. First, pBBR-Dest42 (shown in Figure 5.2) is based on the pBBR1MCS5 plasmid, a medium copy number plasmid containing a pBBR1 replicon which allows for replication in a number of different bacterial hosts [223]. It also includes a gentamicin resistance cassette for selection of transformants, and a mobilization region for conjugative transfer of plasmids into those species that are not amenable to transformation through mechanical or chemical means.

pBBR-Dest42 was created through modification of the pBBR1MCS5 backbone to include a Gateway cloning cassette so that pBBR-Dest42 can serve as a destination vector for the Gateway cloning system (Invitrogen, Carlsbad, CA). The Gateway system of recombination is based upon highly specific and efficient site-specific recombination that integrates the bacteriophage lambda sequence into the genome of *E. coli* [225]. Vectors within the Gateway system contain bacteriophage lambda recombination sites that have been modified to improve efficiency and specificity and to facilitate directional cloning. Recombinase enzyme mixes are used for each step in the creation of both entry and destination clones in order to take advantage of the recombination specificity of each step. In the gateway system of cloning, an entry clone is created in two steps. First the gene of interest is amplified through the polymerase chain reaction (PCR) using primers that contain modified lambda phage *attB* attachment site sequences. This PCR product is then cloned into lambda phage *attP* recombination sites of the entry vector through a BP clonase reaction, creating an *attL* recombination site. The entry vector contains a kanamycin resistance cassette for selection purposes. The gene of interest in the entry vector is flanked by lambda phage *attL* attachment sites, and can be subcloned into a destination vector containing *attR* recombination sites through an LR recombinase reaction. The *attR* recombination sites flank a cloning cassette in the destination vector that contains a chloramphenicol resistance gene and a *ccdB* lethality gene from the F plasmid [225]. The *ccdB* gene product interferes with DNA gyrase, and thus destination plasmids must be maintained in a strain of bacteria such as *E. coli* species DB3.1 (Invitrogen, Carlsbad CA) with a mutated DNA gyrase that is impervious to the effects of CcdB [225].   Once an LR recombination reaction is done, this cassette is replaced with the gene of interest, and the destination clone is then used for expression of fusion protein. Inclusion of this gateway cloning cassette in the parental pBBR-Dest42 vectors provides the advantages of a quick and easy ligation-independent cloning method [214, 234-238] that facilitates a high-throughput approach to both surveying protein-protein interactions on a genome-wide scale and to rapidly expressing single fusion proteins of interest.

The pBBR-Dest42 parent vector contains a C-terminal tag consisting of 6xHIS in tandem with a V5 epitope tag and has been used to express tagged “bait” proteins in soil bacterium *R. palustris* on a genome-wide scale [81]. Over 800 tagged “bait” proteins have been expressed and purified in a multi-step purification strategy compatible with the HIS-V5 tandem tag. The eluates were analyzed via mass spectrometry to identify bait proteins and their associated interacting partners. While the 6xHIS tag is a commonly used tag with a simple IMAC (Immobilized Metal Affinity Column) purification strategy, elution of bound bait protein to Nickel-NTA via the 6xHIS and to the anti-V5 resin via the V5 epitope tag requires harsh elution conditions, denaturing the bait protein and possibly disrupting interactions of partner proteins or losing the interacting proteins altogether.

Design considerations for inclusion of new tags in pBBR-Dest42-based vectors

It is sometimes desirable to purify proteins under non-denaturing native biological conditions. Non-denaturing conditions used during purification can improve yields and facilitate maintenance of non-covalent interactions between protein binding partners [239, 240]. Further, the ability to purify a protein and associated interacting partners can be dependent upon the tag or the combination of tags used for the purification, and upon the position of the tag within the fusion protein [234]. Overlap of protein interacting partners identified with different purification strategies can be as low as 4-7% [73]. Therefore, having a set of vectors containing three different C-terminal tandem affinity tags with the backbone based on the pBBR-Dest42 parent vector can be useful in isolating different complexes that may not be amenable to purification using a HIS-V5 tag, and can provide complementary strategies to increase the numbers of proteins amenable to purification on a genome-wide scale. To address these issues, the C-terminal HIS-V5 tag in the pBBR-Dest42 parent vector was replaced with either a modified Tandem Affinity Purification (TAP –2xCalmodulin Binding Protein(CBP) -Tobacco Etch Virus (TEV) protease site –2xProtein A) tag [80], a modified Sequential Peptide Affinity (SPA- 2xCBP - TEV protease site - 3xFLAG epitope) tag [75] or a

StrepII-TEV protease site-6xHIS (STH) tag [233], as shown in Figure 5.3. Use of only C-terminal fusion tags alleviated the need for including ribosomal binding sites that would be required for an N-terminal tag (and thus would likely further limit the application of the vectors) so for this initial work only C-terminal tags were included.

Use of tag systems which combine two different affinity tags, such as the TAP tag developed by Rigaut et al. [80] or the SPA tag developed by Zeghouf et al [75] can increase the applicability and flexibility of a purification scheme, resulting in a higher number of proteins expressed due to increased solubility or increased yield resulting from the combination of tags [241]. The TAP tag is a large tag of ~27kDa which has been used in successful purifications of bait proteins and elucidation of protein interaction networks in both yeast [76-79] and *E. coli* [73, 74], and for this reason was chosen for placement in the pBBR-Dest42 backbone vectors. Larger tags have been shown to increase the solubility and stability of proteins when expressed as fusion proteins, and often can improve the ability to purify these proteins [239]. The SPA tag, which at ~9 kDa is smaller than the TAP tag, has been used for affinity purification of large numbers of protein complexes in *E. coli*, with results comparable to those obtained using bait proteins tagged with a TAP tag [74, 75, 83]. Since not all proteins will express with large tags, the SPA tag was also chosen for inclusion in the pBBR-Dest42-based vectors. The STH tag was chosen for inclusion due to its small size, inexpensive purification reagents, and for its demonstrated success in isolating human protein complexes [233]. Since a smaller tag has less likelihood of interference with protein folding, or with association of interacting partners, the SPA and STH tags may allow purification of bait proteins and associated binding partners that will not express with the larger TAP tag.

As described earlier, TAP tag sequences were obtained from Rigaut et al [80], SPA tag sequences from Zeghouf et al [75] and STH tag sequences from Rich Gianonne (personal communication). STH tag sequence was used as designed. Modifications to the original TAP and SPA sequences included an additional consensus sequence for Tobbaco Etch Virus (TEV) protease between the two affinity epitopes, and inclusion of a tetracysteine (4C) tag with the sequence CCPGCC upstream of the tag sequence. Protease cleavage is a commonly used strategy in purification schemes [75, 80, 231, 238, 241], allowing for removal of part or all of the tag, which is a huge advantage for crystallization of proteins to determine structure [237] or for purifying proteins for medical uses [242]. In a tandem affinity purification scheme, removal by proteolytic cleavage of an outside tag that has very tight binding to affinity resins, such as Protein A to IgG resins, allows for native elution of the protein in all steps [73-75, 80, 243]. As explained in Chapter 1, the outer tag, tightly bound to the affinity resin, is used for initial capture of the "bait" protein. This tag is then cleaved off by the protease, leaving "bait" protein with only an inner tag floating free in solution. A second affinity capture is performed using an inner tag with lower affinity binding that can then be eluted from the affinity resin using gentler elution conditions. TEV protease, originally isolated from the RNA genome of Tobacco Etch Virus (TEV), exhibits a high degree of cleavage specificity [231, 244]. Recent improvements in the ability to purify it from a bacterial host have made it an attractive choice for use in tandem affinity tags [217]. An additional TEV cleavage site was added to the tag constructs inserted in the pBBR-Dest42 backbone in order to improve likelihood and efficiency of cleavage [84, 233]. Improved efficiency results in 95% complete cleavage of inner tag from the protein within 1 hour at room temperature (Rich Giannone, personal communication). Tetracysteine (4C) tags were also included in new tag constructs in the innermost position closest to the gene of interest. Tetracysteine (4C) tags have been used to visualize protein localization in mammalian cells and to detect presence of expressed protein within cell lysates [84]. Protein presence within cell lysates is visualized through interaction of the 4C motif with TC-FlAsH (Fluorescein Arsenical Hairpin) reagent (Molecular Probes, a division of Invitrogen, Carlsbad, CA). Reaction of the tag with the FlAsH reagent creates a hairpin structure with the tag binding to arsenic residues in the reagent. The dye then fluoresces with an emission wavelength of 550nm, allowing easy detection in a fluorescent imager or transilluminator equipped with a 500-600nm filter, or alternatively using an ethidium bromide channel. The dye is sensitive enough to detect nanogram quantities of protein. Presence of the 4C motif in the tags allows detection of expressed protein in cell lysates in the time it takes to run an SDS-PAGE gel, thus eliminating the need for more time-consuming western blot procedures to determine protein expression.

When isolating proteins and their interacting partners via an affinity purification scheme with subsequent identification of those proteins via mass spectrometry (AP-MS), the presence of non-specific interacting proteins becomes an issue [9]. Proteins such as ribosomal proteins and elongation factors are simply “sticky,” adhering to resins and attached protein complexes no matter what purification scheme is used. Tandem affinity purification schemes serve to decrease the numbers of non-specifically interacting proteins [80]. A number of methods have been developed to increase the probability that detected proteins are true interacting partners [9, 12] but a survey of such methods is beyond the scope of this work. Instead, the focus here is on expression of proteins.

Proteins expressed from a plasmid can vary greatly in expression level, depending on a number of factors such as promoters used for protein expression and copy number of the plasmid. Earlier studies in *R. palustris* have shown expression levels of RNA polymerase subunits expressed from the *lac* promoter on a plasmid to vary between 0.7 and 1.7 times that of wild-type expression [11]. Over-expression of a protein can lead to artifactual interactions that are not biologically relevant. Further, expression of tagged versions of protein off of a plasmid increases the pool of the expressed protein in the cell, with both a wild-type and a tagged version of the same protein being present in the cell. This increases the competition for binding partners, and may lead to an inability to detect interacting partners or an increase in the numbers of non-specific binding to the tagged protein.

For this reason, a set of vectors based on the pJQ200KS plasmid was created that would allow endogenous expression of tagged versions of proteins. pJQ200KS is a small plasmid derived from the P15A plasmid first isolated from *E. coli* [226]. It contains a gentamicin resistance cassette for selection of positive transformants, a mobilization region for conjugative transfer, and can be maintained in *E. coli*. However, the P15A origin of replication (ORI) ensures nonreplication in non-enterobacteria, and will integrate into the genome of non-enterobacteria through homologous recombination [226, 227]. Gene replacement through use of pJQ200KS was first tested in *Rhizobium leguminosarum* species and applied to many bacterial species since then, so success in the use of this plasmid for gene integration in diverse bacteria has already been established [226]. Usefulness of this vector in protein-protein interaction studies in non-enterobacteria could include expression of tagged versions of proteins that are lethal, either disrupting cellular function or killing the cell if over-expressed, or expressing versions of proteins that will not express off of an exogenous plasmid, or expression of low expression level proteins such as DNA polymerase subunits which are difficult to purify with associated interacting partners due to competition with endogenously expressed subunits for binding partners. Since this vector integrates into the genome of non-enterobacteria, it is possible that integration will disrupt the expression of an essential protein and thus be lethal to the cell. An additional disadvantage to the use of an integrating vector containing C-terminal tags include limiting the genes expressed to those that are either orphans within the genome or at the end of an operon structure. Since no promoter sequence is included in the vector, C-terminally tagged proteins will express from a native promoter in the genome sequence, but expression of those genes downstream of the tagged protein may be disrupted. Endogenous expression under native promoter is less likely to yield false positives because it limits somewhat the recombinant protein expression levels and thus possible artifacts associated with over-expression. Integration of tagged protein in genome disrupts native copy, and thus expression of a tagged version under a native promoter could allow exploration of protein complexes that will not purify with their interacting partners due to competition with non-tagged protein subunits.

Construction of vectors with new tag sequences

Initial strategies for construction of new vectors included amplification of TAP2, SPA2 and STH tag sequences designed and constructed earlier for inclusion in mammalian expression vectors [233]. PCR product obtained from these vectors was restriction digested and attempts were made to ligate the digested PCR product into linearized pBBR-Dest42 as described above. After many attempts, this approach proved unsuccessful. New tag sequence was designed from back translations of tag amino acid sequences, with codons being optimized for expression in *R. palustris* as described above. New tag sequences were constructed by GenScript (Piscatawny, NJ), and constructs were amplified through PCR. A PCR amplification strategy was found to be necessary due to the methylation sensitivity of the Cla1 restriction enzyme. Vectors propagated through growth in *E. coli* strains contain methylated DNA which could not be restriction digested by Cla1 endonuclease. A PCR strategy amplified the tag sequence without methylation at the restriction enzyme consensus sequence. The pBBR-Dest42 parent vector was designed with restriction digest sites in two places: 1. surrounding the original HIS-V5 tag for easy tag removal and replacement and 2. placed at the 5' end of the Gateway cloning cassette for inclusion of cassettes with N-terminal tags and ribosomal binding sites. Parent vector was linearized and HIS-V5 tags removed through digestion with BstB1 restriction endonuclease, which leaves a complementary overhanging sequence to Cla1 restriction digest. PCR product for each individual tag (STH, SPA2 or TAP2) was restriction digested with Cla1 restriction endonuclease and ligated into BstB1 restriction sites of the linearized parent vector. Ligation of Cla1 digested tags into the parent vector BstB1 restriction sites destroyed the BstB1 restriction site. Therefore, restriction digest of the ligated vectors with BstB1 restriction endonuclease provided a convenient way to eliminate vectors that had re-ligated without a tag insert or had re-ligated with the original HIS-V5 vector. All constructs were confirmed both with diagnostic restriction digest and with DNA sequencing. This same strategy can easily be used to create vectors with additional C-terminal or with N-terminal tags with the pBBR-Dest42 backbone .

The above strategy was also used to insert C-terminal tags into pJQ200KS parent vector [226]. The gateway cassettes including either TAP2, SPA2 or STH C-terminal tags were PCR amplified using universal primers. PCR product was then ligated into a linearized pJQ200KS parent vector, as described above.

Protein expression

Rhodopseudomonas palustris was used as a bacterial host to show protein expression and to illustrate the usefulness of both integrating and exogenous vectors in investigation of protein complexes in a native host. These bacteria were chosen for the proof-of concept set of experiments because of the availability at ORNL of a large amount of data compiled on both a proteome expression level and a functional proteomics level in these bacteria, thereby providing a suite of reference data [30, 81]. The pBBR-Dest42 parent vector from which the new set of vectors was created has been tested extensively in this species, thus facilitating comparison of expression of proteins with new tags to those already characterized. Additionally, the availability of a wide variety of entry clones for native bacterial proteins in this species made it an attractive host for the protein-protein interaction studies in this chapter. *R. palustris* is also an attractive host for testing protein expression from the integrating vector set based on the pJQ200KS parent vector due to previous use of this parent vector to create gene knock-outs in this species [245, 246]. Perhaps the largest disadvantage to using *R. palustris* as a bacterial host for protein expression lay in the amount of time required for growth, with approximately 6 weeks being required from transformation to initial expression tests.

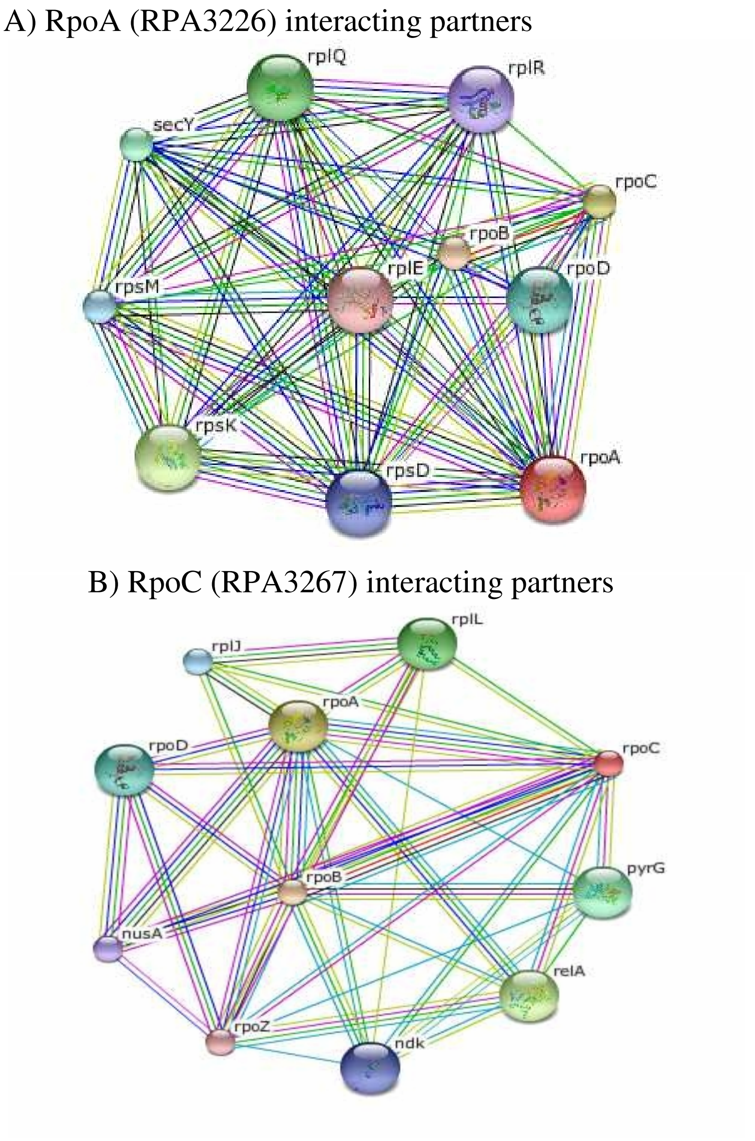
The original goals of this project included not only creation of vectors with different C-terminal tags such that coverage of protein-protein interactions could be increased, but also comparative tests of protein expression across tags. The questions that we hoped to address initially included examining the effects of endogenous expression versus exogenous expression on detection of proteins and their interacting partners. Further, we hoped to characterize the effects of different tags on expression and immunoprecipitation of proteins of low versus high expression level, and the effect of different tag sizes on expression and detection of proteins of high versus low molecular weight. We also hoped to examine the usefulness of any of the new C-terminal tags in the detection of binding partners of transiently interacting proteins. We further hoped to confirm the use of the 4C tag for detection of expression within cell lysates.

Original candidate test gene choices for pBBR-Dest42-based vectors included 2 subunits of DNA polymerase, rpa_0301 DNA polymerase epsilon chain, dnaQ, and rpa_0615 DNA polymerase tau subunit, dnaX, (both with low expression levels), rpa_1175 CheY-like protein (transient interactor) and 2 subunits of RNA polymerase, rpa_3226 RNA polymerase alpha subunit, rpoA (high expression level, low molecular weight) and rpa_3267 RNA polymerase beta prime subunit, rpoC (high expression level, high molecular weight). Unfortunately, due to the time required to create the new vectors combined with the long growth time for *R. palustris*, RNA polymerase subunits rpa_3226 (RpoA) and rpa_3267 (RpoC) were the only proteins successfully expressed, affinity purified and detected via mass spectrometry. Although all other proteins showed some level of expression in cell lysates analyzed by western blot, purification and subsequent detection of the tagged bait proteins and interacting partners in mass spectrometry experiments proved unsuccessful. Original gene choices for expression in pJQ200-KS-SPA2 vector included rpa_3226 rpoA (high expression level), rpa_3059 HolC subunit of DNA polymerase (low expression level), rpa_3522 FtsZ cell division protein, rpa_1631 CheW scaffold protein (no expression off of an exogenous plasmid), and rpa_1175 CheY-like signal protein. Each of these genes was found either at the end of an operon or as an orphan gene not found in an operon structure. Again, although expression in cell lysates was confirmed via western blot, only the RpoA subunit of RNA polymerase was isolated and purified, with bait protein and binding partners detected through mass spectrometry. Perhaps with more time available, immunoprecipitation parameters, such as salt and detergent content of buffers and amount of resins used, and incubation times could have been optimized such that low expression level proteins could have been detected.

RNA polymerase is a multi-subunit enzyme that performs the transcription of DNA template into an RNA polymer. It is well-characterized and abundant in the cell , and has been extensively studied in *E. coli* and *R. palustris* [11, 74, 81]. The core of bacterial RNA polymerase is composed of 5 subunits, consisting of 2 alpha subunits (RpoA, 36.5 kDa), 1 beta subunit (RpoB, 150.6 kDa), 1 beta prime subunit (RpoC, 155.2 kDa), and 1 omega subunit (RpoZ, 10.2 kDa) [11, 247]. Specificity of RNA polymerase is conferred through association of the RNA polymerase core with various sigma factors, which recognize and bind tightly to gene or operon promoter sequences. Different sigma factors are employed at different times, especially during stress conditions, thus allowing for adaptability in response to environmental stress [141]. In earlier genome-wide protein interaction studies conducted in R palustris, RPA1288 RpoD sigma factor was found to interact with both RpoA and RpoC. Additional binding partners for RNA polymerase alpha subunit (RpoA) and for beta prime subunit (RpoC) are cataloged in the STRING database (http://string.embl.de/ newstring_cgi/), and representative figures of interacting partners are shown in Figure 5.4. Because earlier experiments resulted in a thorough catalog of interacting partners for this model system, RNA polymerase alpha (RPA3226, RpoA) and beta prime (RPA3267, RpoC) subunits served as useful bait proteins for verifying expression of low and high molecular weight fusion protein, respectively, from newly created vectors. Fusion proteins of RpoA with either STH tag or SPA2 tag or TAP2 tag were expressed, immunoprecipitation experiments performed and results tabulated in Table 5.4. RpoC was expressed as a fusion protein with either STH tag or with SPA2 tag, but attempts to express a RpoC-TAP2 fusion protein were unsuccessful. Results of immunoprecipitation experiments with RpoC are tabulated in Table 5.5. Thus, the efficacy of this vector set has been shown in initial tests.

Conclusions

The creation of a set of Gateway® compatible destination vectors based on the parent vector pBBR-Dest42, gives an advantage of protein expression in a broad range of bacterial hosts. However, as demonstrated above in a set of preliminary experiments, while the constructs are validated, their application should be tested further, perhaps by using other host strains (such as *Azospirillum*) and/or other bait proteins. These vectors can potentially be used for exploration of protein complexes expressed in native hosts,



A) RpoA (RPA3226) interacting partners

B) RpoC (RPA3267) interacting partners

Figure 5. 4 Schematic depiction of RNA polymerase subunit interacting partners (http://string.embl.de)

Figures were obtained from the STRING interaction database (http://string.embl.de/ newstring_cgi/)A) Interacting partners for RpoA (RPA3226), B) Interacting partners for RpoC (RPA3267). Each “ball” represents a different protein, while each line between individual proteins represents a different form of evidence of interaction, either experimental or in silico. Stronger evidence of interactions between two proteins is represented in this view by more lines between the two proteins.

and are amenable to high-throughput but their application needs to be confirmed. Further characterization of these new vectors could include comparison tests of high versus low expression level proteins with each tag after optimization of affinity purification protocols. Additional tests could include a test of detection of protein presence through using the 4C epitope, which could prove much easier in a bacterial species that does not require chemical lysis procedures.

Chapter 6. Conclusion

Mass spectrometry based proteomics is a powerful tool to use in the characterization of protein expression of newly sequenced microbes. Although genome sequence is an important indication of the physiological capability of a microbe, the expressed proteome gives a more complete picture of the actual physiology of the cells under given conditions. Many microbial genomes contain multiple copies of a number of genes, and genome sequence alone can not tell you which one is expressed in a given situation. Proteomics of bacterial cells grown under different conditions can provide a picture, a snapshot in time, of those proteins expressed and those that are important to existence within a given environmental condition or a given growth state. Thus proteomic analysis can at the very least complement an informed genome sequence annotation.

In Chapter 3 of this work, the tool of mass spectrometry-based proteomics was applied to investigate the expression proteome of single isolate cultures of newly sequenced soil diazotrophic bacteria, *Azospirillum brasilense.* The proteomes of two different strains of *A. brasilense* were examined after growth under optimal conditions in minimalmedia either containing nitrogen in the form of ammonium sulfate, or not containing a nitrogen source and grown under limited oxygen conditions to promote nitrogen fixation (nitrogen fixing conditions). These two cultures were grown under two well-characterized optimum growth conditions, corresponding to distinct physiological abilities for this species. Since no genome sequence is yet available for strain Sp7, the genome of closely related Sp245 strain was used in this work. BlastX comparison of the available Sp7 translated coding sequences with those of the Sp245 strain gene coding sequences indiciated a high degree of similarity (90-100%) between individual coding sequences of each strain, although one gene (Sp7 *nirS*) did not have a homolog in the Sp245 genome, and two other coding sequences (nitrogenase alpha and beta chains) exhibited only ~30% similarity to those in the Sp245 genome. Therefore, inferences regarding protein identification in strain Sp7 using the annotated genome sequence of strain Sp245, while informative, is likely to have some flaws. A recent study indicated that the *wzm* gene found on the pRHICO plasmid of Sp7 strain has 98% identity to the same gene found within the Sp245 genome [172]. Although some genes may have DNA sequence differences, BLASTx comparison of available Sp7 genes suggests that the differences reflected at a protein level seem to be minimal. Nevertheless, some artifactual identifications could be made, or some proteins not detected due to sequence differences. Therefore, the genome sequence for the Sp245 strain is likely the best representation of genome sequence (and therefore translated protein sequence) available with the majority of proteins detected being accurate. However, due to the possible sequence differences between strains, direct comparisons between expression proteomes of each strain could not be done with sufficient confidence to be insightful and accurate.

The high throughput methodology applied to *Azospirillum brasilense* has yielded the first proteome results of this metabolically diverse bacterial species, providing an initial foundation for further in-depth analyses. Combination of proteomic data with microarray studies currently being conducted will further elucidate the overall changes in *Azospirillum* cells due to nitrogen fixation, giving a more complete picture of overall physiological changes. Combination of these two "omics" technologies could further provide a set of useful markers for physiological changes occurring as the cells fix nitrogen or as they prepare to fix nitrogen. This could in turn lead to development of reporter genes useful for determining physiological changes related to nitrogen fixation, in addition to providing insight into the physiology of cells grown under these conditions. Additionally, the proteome data can be useful in developing a more complete annotation of genome sequences. Searching proteome data using a a six frame translation of the genome sequence using a methodology such as that used on data from the plant pathogens *Phytopthera sojae* and *Phytopthera ramorum* can yield more accurate start and stop sites for some genes and can lead to a more accurate annotation of putative genes. This method was used on, and successfully used to provide more complete annotation to the genome sequence of these plant pathogens [248]. Future directions could include a six-frame translation of the Sp245 proteome data such that genome annotation could be more complete.

An additional expression proteome study was done to investigate the effect of a chemotaxis-like signaling pathway on the overall physiology of *Azospirillum brasilense* species Sp7, with results presented in Chapter 4. Expression proteomes of cultures of mutant cells were done in order to elucidate the overall proteome changes when forward signaling output was eliminated versus when only the adaptation pathway was eliminated. The broader view of the physiology of the mutant cells provided by the proteomic study told a story of a dramatic difference in physiology between the two isogenic mutants. There was not a substantial difference between the wild-type cells and the cells which lacked an adaptation pathway, but the differences noted between wild-type and those cells lacking a forward signaling pathway was much different than was previously known. Although a common set of proteins was expressed, representing a number of proteins needed for basic cellular function, the patterns of expression levels of these proteins was very different in the two mutants. Further analysis of proteomic data could provide a discovery of useful markers for physiological studies, perhaps providing a set of useful reporter genes that could indicate certain physiological features such as those occurring in preparation for nitrogen fixation. In addition, given that the molecular mechanisms underlying the different behaviors of the two mutant strains compared is yet to be determined, this comparison lays the foundation for identifying candidate functions that are directly or indirectly modulated by the chemotaxis-like pathway.

Perhaps the most interesting finding was the expression of type 6 secretion system (T6SS) structural components, which mediates interaction with a eukaryotic host, found to be important in nodulation in *Rhizobium*, and in long-term survival in chronic infections of pathogenic bacteria. Found to be expressed and upregulated in forward signaling mutants, many components were not detected in adaptation mutants, suggesting some level of importance in the response to the Che1 signaling pathway. Since little is known about T6SS in non-pathogenic bacteria, these results are intriguing, and could lead to a number of further investigations. For instance, whether expression of T6SS components is regulated directly or indirectly by the Che1 operon or whether it results from the difference in perception of the environment (as evidenced by the distinct physiologies of the cells brought to light by the proteomic analysis) is yet to be determined. Further, the signal input to which this signaling system is responding is yet to be elucidated. All data was searched using a +14 Da modification on glutamate residues, indicative of addition of a single methyl group, although this analysis was not included in Chapter 4. Sensory receptors involved in signaling environmental conditions through chemotaxis pathways are often constitutively modified by methylation, with adaptive response including removal of methylation. Further investigation of methylation patterns within this data set could provide a clue as to what receptors are methylated in response to the Che1 chemotaxis-like pathway. Little is known about the structure of the T6SS in non-pathogenic bacteria, but structural components in pathogenic bacteria are assumed to consist of both an Hcp protein and a Vgr-family protein that combine to form the secretory apparatus. Since there is no Hcp protein within the genome of Sp245 species, an investigation of protein binding partners for the T6SS components detected in the proteome study could provide clues as to the construction of the secretory apparatus for these cells. Further, the function of the T6SS in mediating cellular changes related to aggregation, and possible secreted effectors that may mediate interaction of the *Azospirillum* cells with plant roots could be investigated.

Finally, in Chapter 5, a set of broad host range expression vectors, and an additional set of integrating vectors with a cassette conferring Gateway cloning compatibility was constructed. These vectors contain DNA sequences for different C-terminal tandem affinity tags. Since proteins are not homogeneous molecules, affinity tags lend the capability to do generic purifications, such that proteins can be purified in a high-throughput manner. A set of vectors allows experimentation such that a wider variety of proteins can possibly be expressed. Gateway compatibility further simplifies cloning such that a large percentage of the genome could be expressed and purified and interacting partners determined, thus entering the realm of functional proteomics. Expression of tagged RNA polymerase subunits from these vectors was tested in *R. palustris*, a soil bacterium in which the base parent vector had been extensively tested and for which a thorough set of entry clones were already constructed. Future directions could include a more thorough characterization of these vectors as outlined in Chapter 5. Further, expression of proteins for these vectors could be tested in *Azospirillum* species with a model complex such as RNA polymerase. After successful tests in *Azospirillum*, these vectors could be used to express and purify the components of the T6SS and interacting partners identified by mass spectrometry, perhaps facilitating a model of secretory apparatus.

In summary, the expression proteomics presented in this work yields the first set of proteomics data for *Azospirillum brasilense*. This data set expands the base of knowledge available on this versatile bacterium, and provides a set of proteins for further exploration. Further, the set of broad host range vectors developed provide a set of tools to use for exploration of protein interactions on a genome-wide scale that could lead to a functional understanding of individual proteins.

References

1. Kocher, T. and G. Superti-Furga, *Mass spectrometry-based functional proteomics: from molecular machines to protein networks.* Nature Methods, 2007. **4**(10): p. 807-815.
2. McDonald, W.H. and J.R. Yates III, *Shotgun proteomics and biomarker discovery.* Disease Markers, 2002. **18**(2): p. 99.
3. Eng, J.K., A.L. McCormack, and I. John R. Yates, *An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database.* Journal of the American Society of Mass Spectrometry, 1994. **5**: p. 976-989.
4. Han, X., A. Aslanian, and J.R. Yates III, *Mass spectrometry for proteomics.* Current Opinion in Chemical Biology, 2008. **12**(5): p. 483.
5. MacCoss, M.J. and John R. Yates III, *Proteomics: analytical tools and techniques.* Current Opinion in Clinical Nutrition and Metabolic Care, 2001. **4**: p. 369-375.
6. Tabb, D.L., C.G. Fernando, and M.C. Chambers, *MyriMatch: Highly Accurate Tandem Mass Spectral Peptide Identification by Multivariate Hypergeometric Analysis.* Journal of Proteome Research, 2007. **6**(2): p. 654.
7. Tabb, D.L., et al., *DBDigger: Reorganized Proteomic Database Identification That Improves Flexibility and Speed.* Analytical Chemistry, 2005. **77**(8): p. 2464-2474.
8. Tabb, D.L., A. Saraf, and J.R. Yates, *GutenTag: High-Throughput Sequence Tagging via an Empirically Derived Fragmentation Model.* Analytical Chemistry, 2003. **75**(23): p. 6415.
9. Tackett, A.J., et al., *I-DIRT, A General Method for Distinguishing between Specific and Nonspecific Protein Interactions.* Journal of Proteome Research, 2005. **4**: p. 1752-1756.
10. Ong, S.-E., et al., *Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics.* Mol Cell Proteomics, 2002. **1**(5): p. 376-386.
11. Hervey, W.J., et al., *Evaluation of Affinity-Tagged Protein Expression Strategies Using Local and Global Isotope Ratio Measurements.* Journal of Proteome Research, 2009. **8**(7): p. 3675.
12. Lee, D.J., et al., *Affinity Isolation and I-DIRT Mass Spectrometric Analysis of the Escherichia coli O157:H7 Sakai RNA Polymerase Complex.* Journal of Bacteriology, 2008. **190**(4): p. 1284-1289.
13. Thompson, M.R., et al., *Experimental Approach for Deep Proteome Measurements from Small-Scale Microbial Biomass Samples.* Analytical Chemistry, 2008. **80**(24): p. 9517.
14. McCormack, A.L., et al., *Direct Analysis and Identification of Proteins in Mixtures by LC/MS/MS and Database Searching at the Low-Femtomole Level.* Analytical Chemistry, 1997. **69**(4): p. 767.

15. Liu, K., et al., *Relationship between Sample Loading Amount and Peptide Identification and Its Effects on Quantitative Proteomics.* Analytical Chemistry, 2009. **81**(4): p. 1307.

16. Link, A.J., et al., *Direct analysis of protein complexes using mass spectrometry.* Nature Biotechnology, 1999. **17**.

17. McDonald, W.H., et al., *Comparison of three directly coupled HPLC MS/MS strategies for identification of proteins from complex mixtures: single-dimension LC-MS/MS, 2-phase MudPIT, and 3-phase MudPIT.* International Journal of Mass Spectrometry, 2002. **219**: p. 245-251.

18. Peng, J., et al., *Evaluation of Multidimensional Chromatography Coupled with Tandem Mass Spectrometry (LC/LC-MS/MS) for Large-Scale Protein Analysis: The Yeast Proteome.* Journal of Proteome Research, 2003. **2**(1): p. 43-50.

19. Washburn, M.P., D. Wolters, and J.R.Yates III, *Large-scale analysis of the yeast proteome by multidimensional protein identification technology.* Nature Biotechnology, 2001. **19**(March 2001): p. 242-247.

20. Wolters, D.A., M.P. Washburn, and John R Yates III, *An Automated Multidimensional Protein Identification Technology for Shotgun Proteomics.* Analytical Chemistry, 2001. **73**: p. 5683-5690.

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

22. Prenni, J.E., A.C. Avery, and C.S. Olver, *Proteomics: a review and an example using the reticulocyte membrane proteome.* Veterinary Clinical Pathology, 2007. **36**(1): p. 13-24.

23. Wu, C. and I. John R. Yates, *The Application of Mass Spectrometry to Membrane Proteins.* Nature Biotechnology, 2003. **21**: p. 262-267.

24. Zybailov, B., et al., *Statistical Analysis of Membrane Proteome Expression Changes in Saccharomyces cerevisiae.* Journal of Proteome Research, 2006. **5**(9): p. 2339-2347.

25. Tebbe, A., et al., *Analysis of the cytosolic proteome of Halobacterium salinarum and its implication for genome annotation.* Proteomics, 2005. **5**: p. 168-179.

26. Djordjevic, M.A., ***Sinorhizobium meliloti*** *metabolism in the root nodule: A proteomic perspective.* Proteomics, 2004. **4**(7): p. 1859-1872.

27. Koksharova, O.A., J. Klint, and U. Rasmussen, *Comparative proteomics of cell division mutants and wild-type of Synechococcus sp. strain PCC 7942.* Microbiology, 2007. **153**(8): p. 2505-2517.

28. Malmstrom, J., et al., *Proteome-wide cellular protein concentrations of the human pathogen Leptospira interrogans.* Nature, 2009. **460**(7256): p. 762.

29. Ow, S.Y., et al., *Quantitative Overview of N2 Fixation in Nostoc punctiforme ATCC 29133 through Cellular Enrichments and iTRAQ Shotgun Proteomics.* Journal of Proteome Research, 2009. **8**(1): p. 187-198.

30. VerBerkmoes, N.C., et al., *Determination and Comparison of the Baseline Proteomes of the Versatile Microbe Rhodopseudomonas palustris under Its Major Metabolic States.* J. Proteome Res., 2006. **5**(2): p. 287-298.

31. Victor, J.N. and J.F.S. Marc, *Analysis of environmental stress response on the proteome level.* Mass Spectrometry Reviews, 2008. **27**(6): p. 556-574.

32. Wagner, M.A., et al., *Global analysis of the Brucella melitensis proteome: Identification of proteins expressed in laboratory-grown culture.* Proteomics, 2002. **2**(8): p. 1047-1060.
33. Yang-Hoon, K., et al., *Proteome response of Escherichia coli fed-batch culture to temperature downshift.* Applied Microbiology & Biotechnology, 2005. **68**(6): p. 786.
34. Moberg, M., J. Bergquist, and D. Bylund, *A generic stepwise optimization strategy for liquid chromatography electrospray ionization tandem mass spectrometry methods.* Journal of Mass Spectrometry, 2006. **41**(10): p. 1334-1345.
35. Zybailov, B., et al., *Correlation of Relative Abundance Ratios Derived from Peptide Ion Chromatograms and Spectrum Counting for Quantitative Proteomic Analysis Using Stable Isotope Labeling.* Analytical Chemistry, 2005. **77**(19): p. 6218-6224.
36. Timms, J.F. and R. Cramer, *Difference gel electrophoresis.* Proteomics, 2008. **8**(23-24): p. 4886-4897.
37. Mottaz-Brewer, H.M., et al., *Optimization of Proteomic Sample Preparation Procedures for Comprehansive Protein Characterization of Pathogenic Systems.* Journal of Biomolecular Techniques, 2008. **19**: p. 285-295.
38. Wang, H. and S. Hanash, *Intact-protein based sample preparation strategies for proteome analysis in combination with mass spectrometry.* Mass Spectrometry Reviews, 2005. **24**(3): p. 413-426.
39. Tang, J., et al., *Recent development of multi-dimensional chromatography strategies in proteome research.* Journal of Chromatography B, 2008. **866**(1-2): p. 123.
40. Domon, B. and R. Aebersold, *Mass Spectrometry and Protein Analysis.* Science, 2006. **312**(5771): p. 212-217.
41. John R Yates III, *Mass spectrometry and the age of the proteome.* Journal of Mass Spectrometry, 1998. **33**(1): p. 1-19.
42. Khalsa-Moyers, G. and W. Hayes McDonald, *Developments in mass spectrometry for the analysis of complex protein mixtures.* Briefings in Functional Genomics & Proteomics, 2006. **5**(2): p. 98.
43. McDonald, W. and J.R.Y. III, *Proteomic Tools for Cell Biology.* Traffic, 2000. **1**: p. 747-754.
44. Karas, M. and F. Hillenkamp, *Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons.* Analytical Chemistry, 1988. **60**(20): p. 2299.
45. Fenn, J.B., et al., *Electrospray Ionization for Mass Spectrometry of Large Biomolecules.* Science, 1989. **246**(4926): p. 64.
46. March, R.E., *An Introduction to Quadrupole Ion Trap Mass Spectrometry.* Journal of Mass Spectrometry, 1997. **32**: p. 351-369.
47. March, R.E., *Quadrupole ion traps.* Mass Spectrometry Reviews, 2009. **March 2009**.

48. Schwartz, J.C., M.W. Senko, and J.E.P. Syka, *A two-dimensional quadrupole ion trap mass spectrometer.* Journal of the American Society for Mass Spectrometry, 2002. **13**(6): p. 659.

49. Shukla, A.K. and J.H. Futrell, *Tandem mass spectrometry: dissociation of ions by collisional activation.* Journal of Mass Spectrometry, 2000. **35**: p. 1069-1090.

50. Tabb, D.L., et al., *DirecTag: Accurate Sequence Tags from Peptide MS/MS through Statistical Scoring.* Journal of Proteome Research, 2008. **7**(9): p. 3838.

51. Taylor, J.A. and R.S. Johnson, *Implementation and Uses of Automated de Novo Peptide Sequencing by Tandem Mass Spectrometry.* Analytical Chemistry, 2001. **73**(11): p. 2594.

52. Altschul, S.F., et al., *Protein database searches using compositionally adjusted substitution matrices.* FEBS Journal, 2005. **272**(20): p. 5101-5109.

53. Perkins, D.N., et al., *Probability-based protein identification by searching sequence databases using mass spectrometry data.* Electrophoresis, 1999. **20**(18): p. 3551-3567.

54. Tabb, D.L., W.H. McDonald, and I. John R. Yates, *DTASelect and Contrast: Tools for Assembling and Comparing Protein Identifications from Shotgun Proteomics.* Journal of Proteome Research, 2002. **1**: p. 21-26.

55. Albrethsen, J., *Reproducibility in Protein Profiling by MALDI-TOF Mass Spectrometry.* Clin Chem, 2007. **53**(5): p. 852-858.

56. Mayya, V., et al., *Systematic Comparison of a Two-dimensional Ion Trap and a Three-dimensional Ion Trap Mass Spectrometer in Proteomics.* Molecular and Cellular Proteomics, 2005. **4**: p. 214-223.

57. Moberg, M., K.E. Markides, and D. Bylund, *Multi-parameter investigation of tandem mass spectrometry in a linear ion trap using response surface modelling.* Journal of Mass Spectrometry, 2005. **40**: p. 317-324.

58. Riter, L.S., et al., *Statistical design of experiments as a tool in mass spectrometry.* Journal of Mass Spectrometry, 2005. **40**: p. 565-579.

59. Venable, J.D. and I. John R Yates, *Impact of Ion Trap Tandem Mass Spectra Variability on the Identification of Peptides.* Analytical Chemistry, 2004. **76**: p. 2928-2937.

60. Wenner, B.R. and B.C. Lynn, *Factors that Affect Ion Trap Data-Dependent MS/MS in Proteomics.* Journal of the American Society of Mass Spectrometry, 2004. **15**: p. 150-157.

61. Xie, H. and T.J. Griffin, *Trade-Off between High Sensitivity and Increased Potential for False Positive Peptide Sequence Matches Using a Two-Dimensional Linear Ion Trap for Tandem Mass Spectrometry-Based Proteomics.* Journal of Proteome Research, 2006. **5**: p. 1003-1009.

62. Jiang, X., et al., *Optimization of filtering criterion for SEQUEST database searching to improve proteome coverage in shotgun proteomics.* BMC Bioinformatics, 2007. **8**: p. 323-334.

63. Gade, D., et al., *Towards the proteome of the marine bacterium Rhodopirellula baltica: Mapping the soluble proteins.* Proteomics, 2005. **5**(14): p. 3654-3671.

64. Maillet, I., et al., *From the genome sequence to the proteome and back: Evaluation of E. coli genome annotation with a 2-D gel-based proteomics approach.* Proteomics, 2007. **7**(7): p. 1097-1106.
65. Zhang, M.-J., et al., *Comparative proteomic analysis of passaged Helicobacter pylori.* Journal of Basic Microbiology, 2009. **9999**(9999): p. NA.
66. Ha, D.-J., et al., *Proteome analysis of Halobacterium salinarum and characterization of proteins related to the degradation of isopropyl alcohol.* Biochimica et Biophysica Acta, 2007. **1774**: p. 44-50.
67. Gevaert, K., et al., *Stable isotopic labeling in proteomics.* Proteomics, 2008. **8**(23-24): p. 4873-4885.
68. Wiese, S., et al., *Protein labeling by iTRAQ: A new tool for quantitative mass spectrometry in proteome research.* PROTEOMICS, 2007. **7**(3): p. 340-350.
69. Stevenson, S.E., et al., *Validation of gel-free, label-free quantitative proteomics approaches: Applications for seed allergen profiling.* Journal of Proteomics,2009. **72**(3): p 555-566.
70. Old, W.M., et al., *Comparison of Label-free Methods for Quantifying Human Proteins by Shotgun Proteomics.* Mol Cell Proteomics, 2005. **4**(10): p. 1487-1502.
71. Girard, M., et al., *Non-stoichiometric Relationship between Clathrin Heavy and Light Chains Revealed by Quantitative Comparative Proteomics of Clathrin-coated Vesicles from Brain and Liver.* Mol Cell Proteomics, 2005. **4**(8): p. 1145-1154.
72. Florens, L., et al., *Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors.* Methods, 2006. **40**: p. 303-311.
73. Arifuzzaman, M., et al., *Large-scale identification of protein-protein interaction of Escherichia coli K-12.* Genome Res., 2006. **16**(5): p. 686-691.
74. Butland, G., et al., *Interaction network containing conserved and essential protein complexes in Escherichia coli.* Nature, 2005. **433**: p. 531-537.
75. Zeghouf, M., et al., *Sequential Peptide Affinity (SPA) System for the Identification of Mammalian and Bacterial Protein Complexes.* Journal of Proteome Research, 2004. **3**: p. 463-468.
76. Babu, M., et al., *Systematic Characterization of the Protein Interaction Network and Protein Complexes in Saccharomyces cerevisiae using Tandem Affinity Purification and Mass Spectrometry*, in *Methods in Molecular Biology*, I. Staglijar, Editor. 2009, Humana Press. p. 187-207.
77. Gavin, A.-C., et al., *Functional organization of the yeast proteome by systematic analysis of protein complexes.* Nature, 2002. **415**(6868): p. 141.
78. Graumann, J., et al., *Applicability of Tandem Affinity Purification MudPIT to Pathway Proteomics in Yeast.* Molecular and Cellular Proteomics, 2004. **3**: p. 226-237.
79. Ho, Y., et al., *Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry.* Nature, 2002. **415**(6868): p. 180.

80. Rigaut, G., et al., *A generic protein purification method for protein complex characterization and proteome exploration.* Nature Biotechnology, 1999. **17**: p. 1030-1032.

81. Pelletier, D.A., et al., *A General System for Studying Protein-Protein Interactions in Gram-Negative Bacteria.* Journal of Proteome Research, 2008. **7**(8): p. 3319-3328.

82. Collins, M.O. and J.S. Choudhary, *Mapping multiprotein complexes by affinity purification and mass spectrometry.* Current Opinion in Biotechnology, 2008. **19**(4): p. 324.

83. Babu, M., et al., *Sequential Peptide Affinity Purification System for the Systematic Isolation and Identification of Protein Complexes from Escherichia Coli*, in *Methods in Molecular Biology*, J. Reinders and A. Sickmann, Editors. 2009, Humana Press. p. 373-400.

84. Giannone, R.J., Y. Liu, and Y. Wang, *The Monitoring and Affinity Purification of Proteins Using Dual Tags with Tetracysteine Motifs*, in *Micro and Nano Technologies in Bioanalysis*. 2009. p. 421.

85. Lichty, J.J., et al., *Comparison of affnity tags for protein purifcation.* Protein Expression and Purification, 2005. **41**: p. 98-105.

86. Somers, E., J. Vanderleyden, and M. Srinivasan, *Rhizosphere Bacterial Signaling: A Love Parade Beneath Our Feet.* Critical Reviews in Microbiology, 2004. **30**: p. 205-240.

87. Borisov, I.V., et al., *Changes in Azospirillum brasilense motility and the effect of wheat seedling exudates.* Microbiological Research, 2007.

88. Burdman, S., et al., *Extracellular polysaccharide composition of Azospirillum brasilense and its relation with cell aggregation.* FEMS Microbiology Letters, 2000. **189**(2): p. 259.

89. Baldani, V.L.D., et al., *Establishment of inoculated Azospirillum spp. in the rhizosphere and in roots of field grown wheat and sorghum.* Plant Soil, 1986. **90**: p. 37-40.

90. Alexandre, G., S.E. Greer, and I.B. Zhulin, *Energy Taxis is the dominant behavior in Azospirillum brasilense.* Journal of Bacteriology, 2000. **182**(21): p. 6042-6048.

91. Steenhoudt, O. and J. Vanderleyden, *Azospirillum, a free-living nitrogen-fixing bacterium closely associated with grasses: genetic, biochemical and ecological aspects.* FEMS Microbiology Reviews, 2000. **24**(4): p. 487-506.

92. Kamnev, A.A., et al., *Effects of heavy metals on plant-associated rhizobacteria: Comparison of endophytic and non-endophytic strains of Azospirillum brasilense.* Journal of Trace Elements in Medicine and Biology, 2005. **19**(1): p. 91.

93. Assmus, B., et al., *In Situ Localization of Azospirillum brasilense in the Rhizosphere of Wheat with Fluorescently Labeled, rRNA-Targeted Oligonucleotide Probes and Scanning Confocal Laser Microscopy.* Appl. Environ. Microbiol., 1995. **61**: p. 1013-1019.

94. Tarrand, J.J., N.R. Krieg, and J. Do¨bereiner., *A taxonomic study of the Spirillum lipoferum group, with descriptions of a new genus, Azospirillum gen. nov., and two species, Azospirillum lipoferum (Beijerinck) comb. nov. and Azospirillum brasilense sp.nov.* Can. J. Microbiol., 1978. **24**: p. 967-980.

95. Martinez-Argudo, I., et al., *Nitrogen fixation: key genetic regulatory mechanisms.* Biochem. Soc. Trans., 2005. **33**(Pt 1): p. 152-156.

96. Leigh, J.A. and J.A. Dodsworth, *Nitrogen Regulation in Bacteria and Archaea.* Annual Review of Microbiology, 2007. **61**(1): p. 349-377.

97. Klipp, W., et al., eds. *Genetics and Regulation of Nitrogen Fixation in Free-Living Bacteria*. 2004, Kluwer Academic Publishers.

98. Elmerich, C., et al., *Regulation of NIF Gene Expression and Nitrogen Metabolism in Azospirillum.* Soil Biology and Biochemistry, 1997. **29**(516): p. 847-852.

99. Forchhammer, K., *Glutamine Signalling in Bacteria.* Frontiers in Bioscience, 2007. **12**: p. 358-370.

100. Zhulin, I.B., et al., *Oxygen Taxis and Proton Motive Force in Azospirillum brasilense.* Journal of Bacteriology, 1996. **178**(17): p. 5199-5204.

101. Srivastava, A. and A.K. Tripathi, *Adenosine Diphosphate Ribosylation of Dinitrogenase Reductase and Adenylylation of Glutamine Synthetase Control Ammonia Excretion in Ethylenediamine-Resistant Mutants of Azospirillum brasilense Sp7.* Current Microbiology, 2006. **53**: p. 317–323.

102. Chakraborty, B. and K.R. Samaddar, *Evidence for the occurrence of an alternative nitrogenase system in Azospirillum brasilense.* FEMS Microbiology Letters, 1995. **127**: p. 127-131.

103. Eady, R.R. and R.L. Robson, *Characteristics of N2 fixation in Mo-limited batch and continuous cultures of Azotobacter vinelandii.* Biochemisty Journal, 1984. **224**: p. 853-862.

104. Fallik, E., Y.-K. Chan, and R.L. Robson, *Detection of Alternative Nitrogenases in Aerobic Gram-Negative Nitrogen-Fixing Bacteria.* Journal of Bacteriology, 1991. **Vol. 173**(1): p. 365-371.

105. Oda, Y., et al., *Functional Genomic Analysis of Three Nitrogenase Isozymes in the Photosynthetic Bacterium Rhodopseudomonas palustris.* Journal of Bacteriology, 2005. **187**(22): p. 7784-7794.

106. Wadhams, G.H. and J.P. Armitage, *Making Sense of it all: Bacterial Chemotaxis.* Nature Reviews Molecular Cell Biology, 2004. **5**: p. 1024-1037.

107. Baker, M.D., P.M. Wolanin, and J.B. Stock, *Systems biology of bacterial chemotaxis.* Current opinion in Microbiology, 2006. **9**: p. 187-192.

108. Bren, A. and M. Eisenbach, *How Signals Are Heard during Bacterial Chemotaxis: Protein-Protein Interactions in Sensory Signal Propagation.* Journal of Bacteriology, 2000. **182**(24): p. 6865-6873.

109. Baker, M.D., P.M. Wolanin, and J.B. Stock, *Signal transduction in bacterial chemotaxis.* BioEssays, 2006. **28**(1): p. 9-22.

110. Sourjik, V., *Receptor clustering and signal processing in E. coli chemotaxis.* Trends in Microbiology, 2004. **12**(12): p. 569-576.

111. Porter, S.L., G.H. Wadhams, and J.P. Armitage, *Rhodobacter sphaeroides: complexity in chemotactic signalling.* Trends in Microbiology, 2008. **16**(6): p. 251.

112. Larimer, F.W., et al., *Complete genome sequence of the metabolically versatile photosynthetic bacterium Rhodopseudomonas palustris.* Nature Biotechnology, 2004. **22**(1): p. 55.

113. Ferrandez, A., et al., *Cluster II che Genes from Pseudomonas aeruginosa Are Required for an Optimal Chemotactic Response.* J. Bacteriol., 2002. **184**(16): p. 4374-4383.

114. Hickman, J.W., D.F. Tifrea, and C.S. Harwood, *A chemosensory system that regulates biofilm formation through modulation of cyclic diguanylate levels.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(40): p. 14422-14427.

115. Galperin, M.Y., *Structural Classification of Bacterial Response Regulators: Diversity of Output Domains and Domain Combinations.* Journal of Bacteriology, 2006. **188**: p. 4169-4182.

116. Szurmant, H. and G.W. Ordal, *Diversity in Chemotaxis Mechanisms among the Bacteria and Archaea.* Microbiol. Mol. Biol. Rev., 2004. **68**(2): p. 301-319.

117. Bible, A.N., et al., *Function of a Chemotaxis-Like Signal Transduction Pathway in Modulating Motility, Cell Clumping, and Cell Length in the Alphaproteobacterium Azospirillum brasilense.* J. Bacteriol., 2008. **190**(19): p. 6365-6375.

118. Stephens, B.B., S.N. Loar, and G. Alexandre, *Role of CheB and CheR in the Complex Chemotactic and Aerotactic Pathway of Azospirillum brasilense.* J. Bacteriol., 2006. **188**(13): p. 4759-4768.

119. Elias, J.E., et al., *Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations.* Nature Methods, 2005. **2**(9): p. 667-675.

120. Gingras, A.-C., et al., *Analysis of protein complexes using mass spectrometry.* Nat Rev Mol Cell Biol, 2007. **8**(8): p. 645.

121. Narasimhan, C., et al., *MASPIC: Intensity-Based Tandem Mass Spectrometry Scoring Scheme That Improves Peptide Identification at High Confidence.* Analytical Chemistry, 2005. **77**(23): p. 7581.

122. Tanner, S., et al., *InsPecT: Identification of Posttranslationally Modified Peptides from Tandem Mass Spectra.* Analytical Chemistry, 2005. **77**(14): p. 4626.

123. Winnik, W.M. and P.A. Ortiz, *Proteomic analysis optimization: Selective protein sample on-column retention in reverse-phase liquid chromatographystar, open.* Journal of Chromatography B, 2008. **875**(2): p. 478-486.

124. Busch, K.L., *Space Charge in Mass Spectrometry.* Spectroscopy, 2004. **19**(6): p. 35-38.

125. Washburn, M.P., R.R. Ulaszek, and I. John R Yates, *Reproducibility of Quantitative Proteomic Analyses of Complex Biological Mixtures by Multidimensional Protein Identification Technology.* Analytical Chemistry, 2003. **75**(19): p. 5054-5061.

126. Tabb, D.L., et al., *MS2Grouper: Group Assessment and Synthetic Replacement of Duplicate Proteomic Tandem Mass Spectra.* Journal of the American Society for Mass Spectrometry, 2005. **16**(8): p. 1250.

127. Eng, J., *Receiver Operating Characteristic Analysis: A Primer.* Academic Radiology, 2005. **12**(7).

128. Boyer, M., et al., *Bacteriophage Prevalence in the Genus Azospirillum and Analysis of the First Genome Sequence of an Azospirillum brasilense Integrative Phage.* Appl. Environ. Microbiol., 2008. **74**(3): p. 861-874.

129. Caballero-Melladoa, J., L. López-Reyes, and R. Bustillos-Cristales, *Presence of 16S rRNA genes in multiple replicons in Azospirillum brasilense.* FEMS Microbiology Letters, 1999. **178** (2): p. 283-288.

130. Martin-Didonet, C.C.G., et al., *Genome Structure of the Genus Azospirillum.* J. Bacteriol., 2000. **182**(14): p. 4113-4116.

131. Katsy, E.I. and A.G. Prilipov, *Mobile elements of an Azospirillum brasilense Sp245 85-MDa plasmid involved in replicon fusions.* Plasmid, 2009.

132. Onyeocha, I., et al., *Physical map and properties of a 90-MDa plasmid of Azospirillum brasilense Sp7.* Plasmid, 1990. **23**(3): p. 169.

133. Vanbleu, E., et al., *Annotation of the pRhico plasmid of Azospirillum brasilense reveals its role in determining the outer surface composition.* FEMS Microbiology Letters, 2004. **232**(2): p. 165.

134. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucl. Acids Res., 1997. **25**(17): p. 3389-3402.

135. Frazzon, J. and I.S. Schrank, *Sequencing and complementation analysis of the nifUSV genes from Azospirillum brasilense.* FEMS Microbiology Letters, 1998. **159**(2): p. 151.

136. Edgren, T. and S. Nordlund, *Two pathways of electron transport to nitrogenase in Rhodospirillum rubrum: the major pathway is dependent on the fix gene products.* FEMS Microbiology Letters, 2006. **260**(1): p. 30-35.

137. Sperotto, R.A., et al., *The Electron Transfer Flavoprotein fixABCX Gene Products from Azospirillum brasilense Show a NifA-Dependent Promoter Regulation.* Current Microbiology 2004. **49**: p. 267-273.

138. Edgren, T. and S. Nordlund, *The fixABCX Genes in Rhodospirillum rubrum Encode a Putative Membrane Complex Participating in Electron Transfer to Nitrogenase.* J. Bacteriol., 2004. **186**(7): p. 2052-2060.

139. Hauwaerts, D., et al., *A major chemotaxis gene cluster in Azospirillum brasilense and relationships between chemotaxis operons in alpha-proteobacteria.* FEMS Microbiology Letters, 2002. **208**: p. 61-67.

140. Xie, Z., et al., *PAS domain containing chemoreceptor couples dynamic changes in metabolism with chemotaxis.* Proceedings of the National Academy of Sciences. **107**(5): p. 2235-2240.

141. White, D., *The Physiology and Biochemistry of Prokaryotes*. 3 ed. 2007: Oxford University Press, Inc. 628.

142. Chen, G., H. Zhu, and Y. Zhang, *Soil microbial activities and carbon and nitrogen fixation.* Research in Microbiology, 2003. **154**(6): p. 393.

143. Fadel-Picheth, C.M.T., et al., *Regulation of Azospirillum brasilense nifA gene expression by ammonium and oxygen.* FEMS Microbiology Letters, 1999. **179**: p. 281-288.

144. Rehder, D., *Vanadium nitrogenase.* Journal of Inorganic Biochemistry, 2000. **80**(1-2): p. 133.

145. Ignoul, S. and J. Eggermont, *CBS domains: structure, function, and pathology in human proteins.* Am J Physiol Cell Physiol, 2005. **289**(6): p. C1369-1378.

146. Mahmood, N.A.B.N., E. Biemans-Oldehinkel, and B. Poolman, *Engineering of ion sensing by the CBS module of the ABC transporter OpuA.* J. Biol. Chem., 2009: p. M901238200.

147. Ma, B.-G., et al., *Characters of very ancient proteins.* Biochemical and Biophysical Research Communications, 2008. **366**(3): p. 607.

148. I. Kefalogianni, G.A., *Modeling growth and biochemical activities of Azospirillum spp.* Applied Microbiology and Biotechnology, 2002. **58**(3): p. 352.

149. Tsagou, V. and G. Aggelis, *Growth dynamics of Azospirillum lipoferum at steady and transitory states in the presence of NH.* Journal of Applied Microbiology, 2006. **100**(2): p. 286-295.

150. Antoine, R., et al., *Overrepresentation of a Gene Family Encoding Extracytoplasmic Solute Receptors in Bordetella.* J. Bacteriol., 2003. **185**(4): p. 1470-1474.

151. Marchler-Bauer, A., et al., *CDD: specific functional annotation with the Conserved Domain Database.* Nucleic Acids Research, 2009(37(Database issue)): p. D205-10.

152. Dunwell, J.M., S. Khuri, and P.J. Gane, *Microbial Relatives of the Seed Storage Proteins of Higher Plants: Conservation of Structure and Diversification of Function during Evolution of the Cupin Superfamily.* Microbiol. Mol. Biol. Rev., 2000. **64**(1): p. 153-179.

153. Ona, O., et al., *Growth and indole-3-acetic acid biosynthesis of Azospirillum brasilense Sp245 is environmentally controlled.* FEMS Microbiology Letters, 2005. **246**(1): p. 125.

154. Kamnev, A.A., et al., *Responses of Azospirillum brasilense to Nitrogen Deficiency and to Wheat Lectin: A Diffuse Reflectance Infrared Fourier Transform (DRIFT) Spectroscopic Study.* Microbial Ecology, 2008. **56**: p. 615-624.

155. Freestone, P., et al., *The universal stress protein, UspA, of Escherichia coli is phosphorylated in response to stasis.* Journal of Molecular Biology, 1997. **274**(3): p. 318.

156. Nachin, L., U. Nannmark, and T. Nystro¨m, *Differential Roles of the Universal Stress Proteins of Escherichia coli in Oxidative Stress Resistance, Adhesion, and Motility.* Journal of Bacteriology, 2005. **187**(18): p. 6265–6272.

157. Blaha, G., Carlos, and I. Schrank, *An Azospirillum brasilense Tn5 mutant with modified stress response and impaired in flocculation.* Antonie van Leeuwenhoek, 2003. **83**(1): p. 35.

158. Kefalogianni, I. and G. Aggelis, *Modeling growth and biochemical activities of Azospirillum spp.* Applied Microbiology and Biotechnology, 2002. **58**(3): p. 352.

159. Crans, D.C., et al., *The Chemistry and Biochemistry of Vanadium and the Biological Activities Exerted by Vanadium Compounds.* Chem. Rev., 2004. **104**(2): p. 849-902.

160. Rehder, D., *Structure and function of vanadium compounds in living organisms.* BioMetals, 1992. **5**: p. 3-12.

161. Guerrieri, N., et al., *Vanadium inhibition of serine and cysteine proteases.* Comparative Biochemistry and Physiology - Part A: Molecular & Integrative Physiology, 1999. **122**(3): p. 331.

162. Fisher, K., D.J. Lowe, and J. Petersen, *Vanadium(V) is reduced by the 'as isolated' nitrogenase Fe-protein at neutral pH.* Chemical Communications, 2006: p. 2807-2809.

163. Lyalikova, N.N. and N.A. Yurkova, *Role of Microorganisms in Vanadium Concentration and Dispersion.* Geomicrobiology Journal, 1992. **10**: p. 15-26.

164. Bellenger, J.P., T. Wichard, and A.M.L. Kraepiel, *Vanadium Requirements and Uptake Kinetics in the Dinitrogen-Fixing Bacterium Azotobacter vinelandii.* Appl. Environ. Microbiol., 2008. **74**(5): p. 1478-1484.

165. Wilkinson, S.P. and A. Grove, *Ligand-responsive Transcriptional Regulation by Members of the MarR Family of Winged Helix Proteins.* Current Issues in Molecular Biology, 2006. **8**: p. 51-62.

166. González, P.J., et al., *Bacterial nitrate reductases: Molecular and biological aspects of nitrate reduction.* Journal of Inorganic Biochemistry, 2006. **100**(5-6): p. 1015.

167. Taté, R., et al., *Cloning and transcriptional analysis of the lipA (lipoic acid synthetase) gene from Rhizobium etli.* FEMS Microbiology Letters, 1997. **149**(2): p. 165.

168. Martin, N., et al., *A Mutant in lipA (yutB), encoding Lipoic Acid Synthase, Provides Insight into the Interplay Between Branched-Chain and Unsaturated Fatty Acids Biosynthesis in Bacillus subtilis.* J. Bacteriol., 2009: p. JB.01160-09.

169. Galimand, M., et al., *Identification of DNA regions homologous to nitrogen fixation genes nifE, nifUS and fixABC in Azospirillum brasilense Sp7.* Journal of General Microbiology, 1989. **135**(5): p. 1047-1059.

170. Maddocks, S.E. and P.C.F. Oyston, *Structure and function of the LysR-type transcriptional regulator (LTTR) family proteins.* Microbiology, 2008. **154**(12): p. 3609-3623.

171. Lerner, A., et al., *The Azospirillum brasilense Sp7 noeJ and noeL genes are involved in extracellular polysaccharide biosynthesis.* Microbiology, 2009: p. mic.0.031807-0.

172. Lerner, A., Y. Okon, and S. Burdman, *The wzm gene located on the pRhico plasmid of Azospirillum brasilense Sp7 is involved in lipopolysaccharide synthesis.* Microbiology, 2009. **155**(3): p. 791-804.

173. Bahat-Samet, E., S. Castro-Sowinski, and Y. Okon, *Arabinose content of extracellular polysaccharide plays a role in cell aggregation of Azospirillum brasilense.* FEMS Microbiology Letters, 2004. **237**(2): p. 195.

174. Bos, M.P., V. Robert, and J. Tommassen, *Biogenesis of the Gram-Negative Bacterial Outer Membrane.* Annual Review of Microbiology, 2007. **61**(1): p. 191-214.

175. Ogura, M., et al., *A new Bacillus subtilis gene, med, encodes a positive regulator of comK.* J. Bacteriol., 1997. **179**(20): p. 6244-6253.

176. Deka, R.K., et al., *The PnrA (Tp0319; TmpC) Lipoprotein Represents a New Family of Bacterial Purine Nucleoside Receptor Encoded within an ATP-binding*

Cassette (ABC)-like Operon in Treponema pallidum. J. Biol. Chem., 2006. **281**(12): p. 8072-8081.

177. McBroom, A.J., et al., *Outer Membrane Vesicle Production by Escherichia coli Is Independent of Membrane Instability.* J. Bacteriol., 2006. **188**(15): p. 5385-5392.
178. Yu, F., S. Inouye, and I. M, *Lipoprotein-28, a cytoplasmic membrane lipoprotein from Escherichia coli. Cloning, DNA sequence, and expression of its gene.* Journal of Biological Chemistry, 1986. **261**: p. 2284-2288.
179. Rao, D.E.C.S., et al., *Molecular characterization, physicochemical properties, known and potential applications of phytases: An overview.* Critical Reviews in Biotechnology, 2009. **29**(2): p. 182 - 198.
180. Lim, B.L., et al., *Distribution and diversity of phytate-mineralizing bacteria.* ISME J, 2007. **1**(4): p. 321.
181. Paoletti, A.C., et al., *Quantitative proteomic analysis of distinct mammalian Mediator complexes using normalized spectral abundance factors.* Proceedings of the National Academy of Sciences, 2006. **103**(50): p. 18928-18933.
182. Lamarche, M.G., et al., *The phosphate regulon and bacterial virulence: a regulatory network connecting phosphate homeostasis and pathogenesis.* FEMS Microbiol Rev, 2008. **32**(3): p. 461 - 473.
183. Sperandeo, P., G. Dehò, and A. Polissi, *The lipopolysaccharide transport system of Gram-negative bacteria.* Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids, 2009. **1791**(7): p. 594.
184. Doyle, S.M. and S. Wickner, *Hsp104 and ClpB: protein disaggregating machines.* Trends in Biochemical Sciences, 2009. **34**(1): p. 40.
185. Wood, J., *Osmosensing by bateria.* Science STKE, 2006. **357**: p. pe43.
186. Grage, K., et al., *Bacterial Polyhydroxyalkanoate Granules: Biogenesis, Structure, and Potential Use as Nano-/Micro-Beads in Biotechnological and Biomedical Applications.* Biomacromolecules, 2009. **10**(4): p. 660.
187. Potter, M. and A. Steinbuchel, *Poly(3-hydroxybutyrate) Granule-Associated Proteins: Impacts on Poly(3-hydroxybutyrate) Synthesis and Degradationâ€.* Biomacromolecules, 2005. **6**(2): p. 552.
188. VanBogelen, R.A., et al., *Global analysis of proteins synthesized during phosphorus restriction in Escherichia coli.* J. Bacteriol., 1996. **178**(15): p. 4344-4366.
189. Mendel, R.R., *Biology of the molybdenum cofactor.* J. Exp. Bot., 2007. **58**(9): p. 2289-2296.
190. Kisker, C., a.H. Schindelin, and D.C. Rees, *Molybdenum-Cofactor containing Enzymes:Structure and Mechanism.* Annual Review of Biochemistry, 1997. **66**(1): p. 233-267.
191. Raux, E., H.L. Schubert, and M.J. Warren , *Biosynthesis of cobalamin (vitamin B12): a bacterial conundrum.* Cellular and Molecular Life Sciences, 2000. **57**(13): p. 1880.
192. Taylor, B.L. and I.B. Zhulin, *PAS Domains: Internal Sensors of Oxygen, Redox Potential, and Light.* Microbiol. Mol. Biol. Rev., 1999. **63**(2): p. 479-506.
193. Galperin, M.Y., *Bacterial signal transduction network in a genomic perspective.* Environmental Microbiology, 2004. **6**(6): p. 552-567.

194. Scheffers, D.-J. and M.G. Pinho, *Bacterial Cell Wall Synthesis: New Insights from Localization Studies.* Microbiology and Molecular Biology Reviews, 2005. **Vol. 69** (No. 4): p. 585–607.
195. Blackburn, N.T. and A.J. Clarke, *Identification of Four Families of Peptidoglycan Lytic Transglycosylases.* Journal of Molecular Evolution, 2001. **52**(1): p. 78.
196. Garcıa-Calderon, C.B., J. Casadesus, and F. Ramos-Morales, *Rcs and PhoPQ Regulatory Overlap in the Control of Salmonella enterica Virulence.* Journal of Bacteriology, 2007. **189**(18): p. 6635–6644.
197. Kumar, A.S., K. Mody, and B. Jha, *Bacterial exopolysaccharides - a perception.* Journal of Basic Microbiology, 2007. **47**(2): p. 103-117.
198. Wu, R., et al., *Crystal Structure of Bacillus anthracis Transpeptidase Enzyme CapD.* Journal of Biological Chemistry, 2009. **284**(36): p. 24406-24414.
199. Candela, T. and A. Fouet, *Bacillus anthracis CapD, belonging to the gamma-glutamyltranspeptidase family, is required for the covalent anchoring of capsule to peptidoglycan.* Molecular Microbiology, 2005. **57**(3): p. 717-726.
200. Santhanagopalan, V., C. Coker, and S. Radulovic, *Characterization of RP 333, a gene encoding CapD of Rickettsia prowazekii with UDP-glucose 4-epimerase activity.* Gene, 2006. **369**: p. 119-125.
201. Lerner, A., et al., *Glycogen phosphorylase is involved in stress endurance and biofilm formation in Azospirillum brasilense Sp7.* FEMS Microbiology Letters, 2009. **300**(1): p. 75-82.
202. Burdman, S., Y. Okon, and E. Jurkevitch, *Surface Characteristics of Azospirillum brasilense in Relation to Cell Aggregation and Attachment to Plant Roots.* Critical Reviews in Microbiology, 2000. **26**(2): p. 91 - 110.
203. Filloux, A., A. Hachani, and S. Bleves, *The bacterial type VI secretion machine: yet another player for protein transport across membranes.* Microbiology, 2008. **154**(6): p. 1570-1583.
204. Pukatzki, S., S.B. McAuley, and S.T. Miyata, *The type VI secretion system: translocation of effectors and effector-domains.* Current Opinion in Microbiology, 2009. **12**(1): p. 11.
205. Roest, H.P., et al., *A Rhizobium leguminosarum Biovar trifolii Locus Not Localized on the Sym Plasmid Hinders Effective Nodulation on Plants of the Pea Cross-Inoculation Group.* Molecular Plant-Microbe Interactions, 1997. **10**(7): p. 938-941.
206. Bingle, L.E.H., C.M. Bailey, and M.J. Pallen, *Type VI secretion: a beginner's guide.* Current Opinion in Microbiology, 2008. **11**(1): p. 3.
207. Cascales, E., *The type VI secretion toolkit.* EMBO reports, 2008. **9**(8): p. 735-741.
208. Leiman, P.G., et al., *Type VI secretion apparatus and phage tail-associated protein complexes share a common evolutionary origin.* Proceedings of the National Academy of Sciences, 2009. **106**(11): p. 4154-4159.
209. Filloux, A., *The type VI secretion system: a tubular story.* EMBO J, 2009. **28**(4): p. 309.
210. Weber, B., et al., *Type VI secretion modulates quorum sensing and stress response in Vibrio anguillarum.* Environmental Microbiology, 2009. **11**(12): p. 3018-3028.

211. Syed, K.A., et al., *The Vibrio cholerae Flagellar Regulatory Hierarchy Controls Expression of Virulence Factors.* J. Bacteriol., 2009. **191**(21): p. 6555-6570.

212. Zhou, L., J. Wang, and L.-H. Zhang, *Modulation of Bacterial Type III Secretion System by a Spermidine Transporter Dependent Signaling Pathway.* PLoS One, 2007. **12**: p. e1291.

213. Kadouri, D., E. Jurkevitch, and Y. Okon, *Involvement of the Reserve Material Poly-{beta}-Hydroxybutyrate in Azospirillum brasilense Stress Endurance and Root Colonization.* Appl. Environ. Microbiol., 2003. **69**(6): p. 3244-3250.

214. Busso, D., B. Delagoutte-Busso, and D. Moras, *Construction of a set Gateway-based destination vectors for high-throughput cloning and expression screening in Escherichia coli.* Analytical Biochemistry, 2005. **343**(2): p. 313.

215. Melcher, K., *A Modular Set of Prokaryotic and Eukaryotic Expression Vectors.* Analytical Biochemistry, 2000. **277**(1): p. 109.

216. Freuler, F., et al., *Development of a novel Gateway-based vector system for efficient, multiparallel protein expression in Escherichia coli.* Protein Expression and Purification. 2008 **59**(2):p.232

217. Rocco, C.J., et al., *Construction and use of new cloning vectors for the rapid isolation of recombinant proteins from Escherichia coli.* Plasmid, 2008. **59**(3): p. 231.

218. Nagarajan, V., et al., *Modular expression and secretion vectors for Bacillus subtilis.* Gene, 1992. **114**(1): p. 121.

219. Krogan, N.J., et al., *Global landscape of protein complexes in the yeast Saccharomyces cerevisiae.* Nature, 2006. **440**(7084): p. 637.

220. Séraphin, B., et al., *Tandem Affinity Purification to Enhance Interacting Protein Identification*, in *Protein–Protein Interactions: A Molecular Cloning Manual, Chapter 17.* 2002, Cold Spring Harbor Laboratory Press. p. 313-328.

221. Fodor, B.D., et al., *Modular Broad-Host-Range Expression Vectors for Single-Protein and Protein Complex Purification.* Applied and Environmental Microbiology, 2004. **70**(2): p. 712-721.

222. Antoine, R. and C. Locht, *Isolation and molecular characterization of a novel broad-host-range plasmid from Bordetella bronchiseptica with sequence similarities to plasmids from gram-positive organisms.* Molecular Microbiology, 1992. **6**(13): p. 1785-1799.

223. Kovach, M.E., et al., *Four new derivatives of the broad-host-range cloning vector pBBR1MCS, carrying different antibiotic-resistance cassettes.* Gene, 1995. **166**(1): p. 175.

224. Lynch, M.D. and R.T. Gill, *Broad host range vectors for stable genomic library construction.* Biotechnology and Bioengineering, 2006. **94**(1): p. 151-158.

225. Hartley, J.L., G.F. Temple, and M.A. Brasch, *DNA Cloning Using In Vitro Site-Specific Recombination.* Genome Res., 2000. **10**(11): p. 1788-1795.

226. Quandt, J. and M.F. Hynes, *Versatile suicide vectors which allow direct selection for gene replacement in Gram-negative bacteria.* Gene, 1993. **127**: p. 15-21.

227. Cozzarelli, N.R., R.B. Kelly, and A. Kornberg, *A Minute Circular DNA from Escherichia Coli 105*.* BIochemistry, 1968. **60**: p. 992-999.

228. Simon, R., U. Priefer, and A. Puhler, *A Broad Host Range Mobilization System for In Vivo Genetic Engineering: Transposon Mutagenesis in Gram Negative Bacteria.* Nat Biotech, 1983. **1**(9): p. 784.

229. Sambrook, J. and D.W. Russell, *Molecular cloning: a laboratory manual*. 2001, Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory Press.

230. Schmidt, T.G.M., et al., *Molecular Interaction Between the Strep-tag Affinity Peptide and its Cognate Target, Streptavidin.* Journal of Molecular Biology, 1996. **255**(5): p. 753-766.

231. Dougherty, W.G., S.M. Cary, and T.D. Parks, *Molecular Genetic Analysis of a Plant Virus Polyprotein Cleavage Site: A Model.* Virology, 1989. **171**: p. 356-364.

232. Vieira, J. and J. Messing, *The pUC plasmids, an M13mp7-derived system for insertion mutagenesis and sequencing with synthetic universal primers.* Gene, 1982. **19**(3): p. 259.

233. Giannone, R.J., et al., *Dual-tagging system for the affinity purification of mammalian protein complexes.* Biotechniques, 2007. **43(3)**(3): p. 296-300.

234. Berrow, N.S., et al., *A versatile ligation-independent cloning method suitable for high-throughput expression screening applications.* Nucl. Acids Res., 2007. **35**(6): p. e45-.

235. Cabrita, L.D., W. Dai, and S.P. Bottomley, *A family of E. coli expression vectors for laboratory scale and high throughput soluble protein production.* BMC Biotechnology, 2006. **6**(12).

236. Chanda, P.K., W.A. Edris, and J.D. Kennedy, *A set of ligation-independent expression vectors for co-expression of proteins in Escherichia coli.* Protein Expression and Purification, 2006. **47**(1): p. 217.

237. Donnelly, M.I., et al., *An expression vector tailored for large-scale, high-throughput purification of recombinant proteins.* Protein Expression and Purification, 2006. **47**(2): p. 446.

238. Stols, L., et al., *A New Vector for High-Throughput, Ligation-Independent Cloning Encoding a Tobacco Etch Virus Protease Cleavage Site.* Protein Expression and Purification, 2002. **25**(1): p. 8.

239. Braun, P., et al., *Proteome-scale purification of human proteins from bacteria.* Proceedings of the National Academy of Sciences, 2002. **99**(5): p. 2654-2659.

240. Abu-Farha, M., F. Elisma, and D. Figeys, *Identification of Protein–Protein Interactions by Mass Spectrometry Coupled Techniques*. 2008.

241. Waugh, D.S., *Making the most of affinity tags.* Trends in Biotechnology, 2005. **23**(6): p. 316.

242. Arnau, J., et al., *Current strategies for the use of affinity tags and tag removal for the purification of recombinant proteins.* Protein Expression and Purification, 2006. **48**(1): p. 1.

243. Gloeckner, C.J., et al., *A novel tandem affinity purification strategy for the efficient isolation and characterisation of native protein complexes.* Proteomics, 2007. **7**(23): p. 4228-4234.

244. Parks, T.D., et al., *Release of Proteins and Peptides from Fusion Proteins Using a Recombinant Plant Virus Proteinase.* Analytical Biochemistry, 1994. **216**(2): p. 413.

245. Jiao, Y. and D.K. Newman, *The pio Operon Is Essential for Phototrophic Fe(II) Oxidation in Rhodopseudomonas palustris TIE-1.* J. Bacteriol., 2007. **189**(5): p. 1765-1773.

246. Braatsch, S., et al., *Rhodopseudomonas palustris CGA009 Has Two Functional ppsR Genes, Each of Which Encodes a Repressor of Photosynthesis Gene Expression.* Biochemistry, 2006. **45**(48): p. 14441.

247. Alberts, B., et al., *Molecular Biology of the Cell*. 4 ed. 2002: Garland Science. 1463.

248. Savidor, A., et al., *Expressed Peptide Tags: An Additional Layer of Data for Genome Annotation.* Journal of Proteome Research, 2006. **5**(11): p. 3048.

Vita

Gurusahai Khalsa-Moyers grew up in Clinton, Tennessee, and still resides in East Tennessee. She attended the University of Tennessee where she obtained her degree in Electrical Engineering. When her children arrived, she elected to stay home with them, homeschooling them until they reached the eighth grade. When her youngest child entered eighth grade, she returned to the University of Tennessee to further her education in a subject that had always been of interest to her, biochemistry. Because of the 16-year time period away from an academic setting, it was necessary for her to take some background courses before beginning graduate school. In January 2004, Gurusahai entered the Genome science and Technology program at the University of Tennessee. During the process of obtaining her PhD, she re-discovered her initial love for technology, and became fascinated with the application of analytical technology to microbial systems.