Mass Spectrometry of *Rhodopseudomonas palustris* Chromatophores and a Method for Displaying Proteomes.

by

ANTHONY PETER FEJES

B.Sc., University of Waterloo, 2000
B.I.S., University of Waterloo, 2001

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

(Department of Microbiology and Immunology)

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA
FEBRUARY 2004

Library Authorization

Anthony Fejes
Name of Author *(please print)*

23/03/04
Date (dd/mm/yyyy)

Title of Thesis: Mass Spectrometry of Rhodopseudomonas palustris Chromatophores and a Method for Displaying Proteomes

Degree: Master of Science
Year: 2004

Department of Microbiology & Immunology
The University of British Columbia
Vancouver, BC Canada

Abstract

Mass Spectrometry of proteins and biologically relevant molecules is an area in which a growing interest in being expressed. However, the field is still in its infancy with respect to the compilation of proteomes of both sub-cellular fractions and whole cells. A two step approach has been used to evaluate the suitability of the mass spectrometry technique on the purple non-sulphur bacteria, *Rhodopsuedomonas palustris*. I first analyzed the mass spectrometry of isolated chromatophores, vesicles formed by invaginations of the bacterial inner membrane, to evaluate our approach. I searched for proteins expected to be located in these structures, identified proteins that may be associated with the chromatophores and searched for potentially novel photosynthetically related hypothetical proteins. Subsequently, I investigated the complete proteome of the bacteria under a number of different environmental conditions and used a mutant strain of this bacterium. From the preliminary results, I created a new approach for display of the proteomics data obtained by my collaborators. This allows for the rapid examination of qualitative and quantitative aspects of proteins by colour-coded display of grouped peptides.

Table of Contents

List of Tables

List of Figures

Abbreviations and Symbols

2DGE - 2 dimensional gel electrophoresis

ACN - acetonitrile

BChl – bacteriochlorophyll

CID - collision-induced dissociation

ESI - electrospray ionization

COG(s) - clusters of orthologous groups

DNA - deoxyribose nucleic acid

DTT - dithiothreitol

HIPIP - high potential iron protein

KEGG(s) - groupings based on the Kyoto Encyclopedia of Genes and Genomes

LC - liquid chromatography

LC/MS/MS - liquid chromatography tandem mass spectrometry

LH1 - light-harvesting complex 1

LH2 - light-harvesting complex 2

MS - mass spectrometry

orf(s) - open reading frame(s)

PEEK - polyetheretherketone

PM - photosynthetic media

QTOF - quadrupole-time-of-flight

RC - reaction center

SDS - sodium dodecylsulfate

SQL - structured query language

TFA - triflouroacetate

TMHMM - transmembrane hidden Markov model software

VBA - Visual Basic for Applications

Acknowledgements

A great many people were involved in creating the data used in this document and deserve recognition. Without their assistance, it would have been impossible to have accumulated enough information to perform these experiments.

Elaine Humphries and Garnet Martin were instrumental in obtaining the Electron Micrographs used here, and were a great resource throughout the time I worked in the Bio-imaging laboratory.

Eugene Yi and David Goodlett at the Institute for Systems Biology performed the Mass Spectrometry on the chromatophores and were very helpful in teaching us how to understand the MS data we obtained from the ISB.

Nathan Verberkmoes was the source for the Mass Spectrometry data obtained from the complete proteome of *R. palustris* used as the basis of the visualization and graphic representation project, as well as providing frequent and helpful advice.

Although not included in this thesis, I was also involved in development and publication of a project (Andronescu et al. 2004) involving the reverse RNA folding problem with the Bioinformatatics, Empirical and Theoretical Algorithms (BETA) laboratory in the Computer Science department. I thank professors Holger Hoos and Anne Condon and the other two members of my project team, Mirela Andronescu and Frank Hutter, for their patience and conviction that our project was worth pursuing in the end.

Finally, I thank all of the people in my laboratory for their patience, understanding and their ever present sense of humour, which created an excellent environment for learning and exploring new avenues of research. My apologies to Dr.

Ali Tehrani, who became my sounding board for many topics that he never expected to learn about. And, of course, many thanks to Dr. J. T. Beatty, who gave me the freedom to wander off the beaten track.

1. Introduction

1.1. Background on *Rhodopseudomonas palustris*

The purple phototrophic gram-negative bacterium *Rhodopseudomonas palustris (R. palustris)* is capable of growing under a wide variety of conditions by altering its metabolism between photoautotrophic, photoheterotrophic, chemoautotrophic and chemoheterotrophic modes of growth in response to changing environmental conditions (Larimer et al. 2004). Known for this metabolic diversity, *R. palustris* has been found - and is able to thrive - in nearly every environmental condition available at the surface of Earth, as well as microgravity (Yang et al. 1999). It is also well known for its ability to use a wide variety of carbon sources, particularly aromatic compounds, and for being able to generate hydrogen gas as a by product of nitrogen fixation (Meyer et al. 1978). *R. palustris* is thought to be one of the most metabolically versatile bacteria in existence (Larimer et al. 2004).

The sequencing of the *R. palustris* genome has been completed (Larimer et al. 2004) and the annotation of the predicted genes has been refined over the past two years. The genome consists of 5,459,213 base-pairs arranged in a single circular chromosome and a further 8,427 base-pairs contained on a circular plasmid. Excluding the plasmid, there are currently 4815 genes predicted in this organism. The completion of this project has opened the doors to probing the genetic basis of *R. palustris*' metabolic diversity on a number of levels, from the creation of gene arrays to the use of mass spectrometry on the *R. palustris* proteins. Without confident predictions for the proteins encoded in the genome, it is impossible to determine the origins of the peptides identified by mass spectrometry. Thus, the completion of the genome and its annotation are a significant milestone for the study of this bacterium for the use of proteomics approaches.

1.2. Photosynthetic Properties of *R. palustris*

1.2.1. Structural Configuration

Although best known for its ability to degrade relatively simple aromatic compounds and its metabolic diversity, *R. palustris* has also long been studied for its photosynthetic capabilities. Under photosynthetic (anaerobic, illuminated) growth conditions, the intracellular structure changes significantly from that of cells grown under aerobic, dark growth conditions because of the formation of intracytoplasmic membrane structures that are derived from and are continuous with the cytoplasmic membrane (Figure 1). This cellular differentiation is exhibited by many purple phototrophic bacteria and, depending on the species, the intracytoplasmic membrane system may range from simple tubular invaginations of the cytoplasmic membrane to relatively large and elaborate thylakoid-like structures as in *R. palustris* (Drews and Golecki 1995; Varga and Staehelin 1983). Upon disruption of cells, segments of the intracytoplasmic membranes are released, most of which spontaneously form vesicles that are called chromatophores (Prince et al. 1975). Purple phototrophic bacterial chromatophores are readily obtained and, because they are formed from the membrane containing all of the known photosynthetic proteins, have been used in experiments ranging from assays of photosynthetic catalytic activities to the purification and crystallization of photosynthetic complexes such as the RC and LH2 (Cogdell et al. 1999; Lancaster et al. 1995; Lilburn et al. 1995).

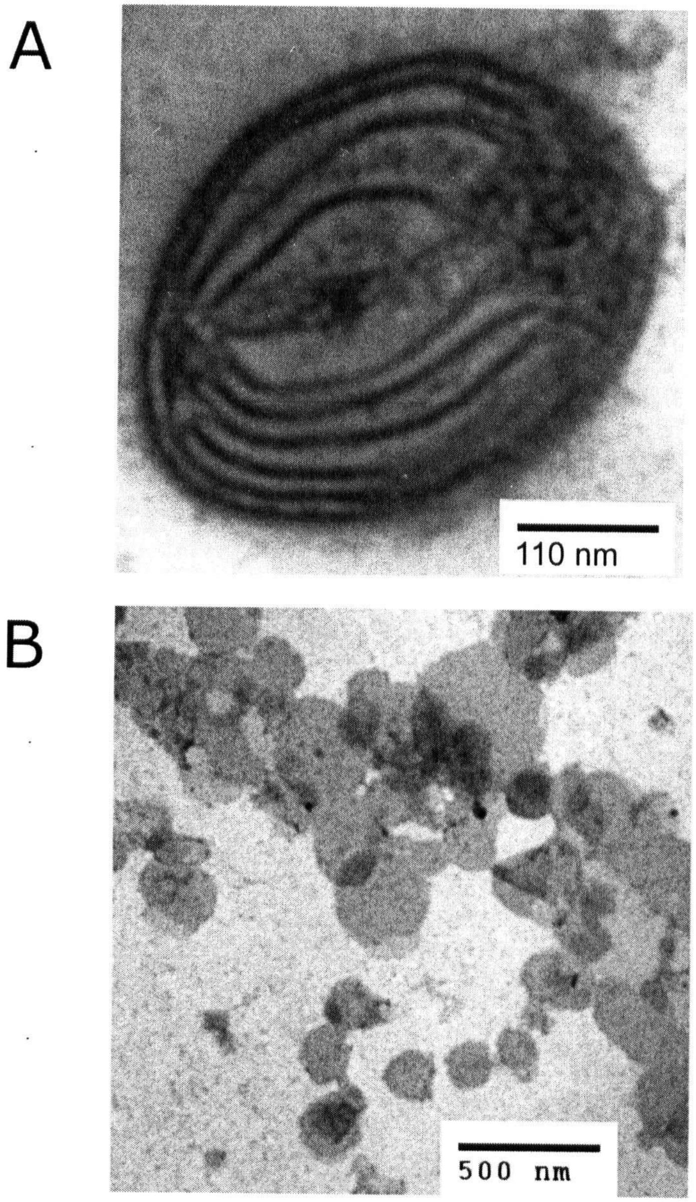


Figure 1: Electron micrographs of an *R. palustris* cell and purified chromatophores. A; Thin section through a cell showing the intracytoplasmic membrane organization. B; Negatively stained chromatophore particles. The bars give dimensions in nanometers (nm).

1.2.2. *R. palustris* Genes Involved in Photosynthesis

From the completed *R. palustris* genome sequence, it is known that most of the predicted photosynthesis genes are located near each other on the single chromosome, along with open reading frames (orfs) of unknown function. Orfs believed to be genes encoding components of the photosynthetic apparatus have been identified, such that the predicted amino acid sequences of all the proteins mentioned below are known; however, there are some uncertainties (see Results and Discussion; and http://genome.ornl.gov/microbial/rpal/).

The photo-active proteins of the purple photosynthetic bacteria occur in three complexes, the reaction center (RC), the light-harvesting complex 1 (LH1) and the light harvesting complex 2 (LH2). The RC consists of three transmembrane proteins, bacteriochlorophyll (BChl) and other cofactors, and operates as a light-driven quinone reductase (Okamura et al. 2000). Immediately surrounding the RC is the LH1 complex, which consists of oligomeric repeats of a transmembrane heterodimeric α/β protein subunit that contains BChl and transfers light energy to the RC (Loach 2000). In purple bacteria such as *R. palustris,* a second type of light-harvesting complex, LH2, transfers light energy to LH1. Although LH2 contains fewer subunits than LH1, it too is built of oligomeric α/β heterodimeric subunits that bind BChl. In general, light energy is thought to follow the pathway LH2 → LH1 → RC (Cogdell et al. 1999). The light energy harvested is transferred through a cyclic series of electron transfer reactions coupled to the translocation of protons across the cytoplasmic membrane (Prince 1990) (Figure 2). Quinones are reduced and protonated at the cytoplasmic side of the RC, and diffuse through the membrane to the periplasmic side of the transmembrane cytochrome *b/c*$_1$ complex which oxidizes quinols with the release of protons. These protons contribute to a transmembrane electrochemical potential that is used by the F_1F_0 ATPase for ATP synthesis. Electrons are transferred from the cytochrome *b/c*$_1$ complex to the RC by one or more types of carriers, typified by the soluble cytochrome c2, to complete the electron transfer cycle (Okamura et al. 2000).

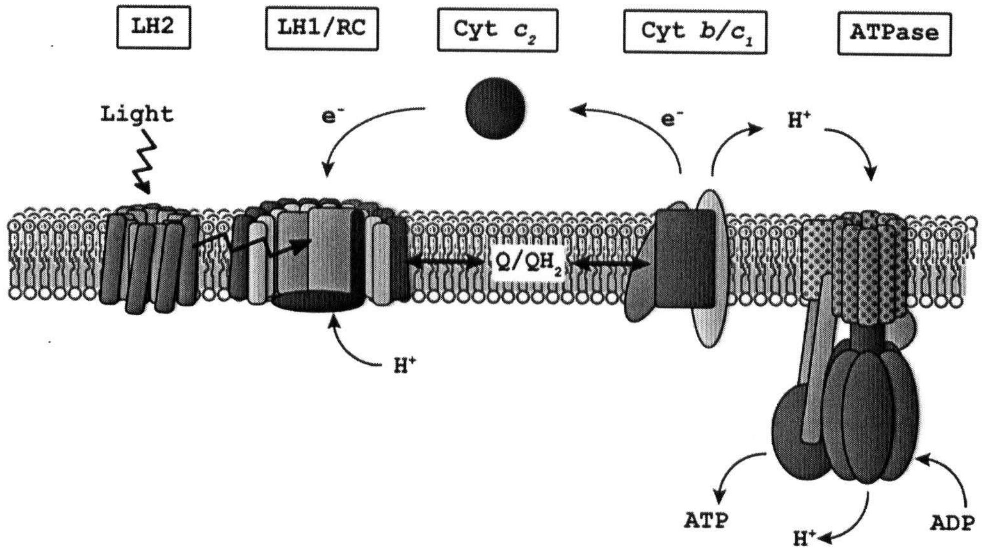


Figure 2: Representation of the intracytoplasmic membrane of *R. palustris*
Illustrating the activities of membrane-integral complexes that catalyze the conversion of light energy to the phosphate ester bond of ATP. LH2, light-harvesting complex 2; LH1, light-harvesting complex 1; RC, reaction center complex; Q/QH$_2$, quinone/quinol pool; cyt *b/c$_1$*, cytochrome *b/c$_1$* complex; *cyt c$_2$*, cytochrome *c$_2$*; ATPase, F$_1$F$_0$ ATP phosphohydrolase complex; H$^+$, protons; e$^-$, electrons; zigzag arrows, light energy. Stippled components of the ATPase indicate the F$_0$ a and c proteins that were not detected in chromatophores (see section 3.3).

Interestingly, *R. palustris* contains five pairs of LH2 structural genes (*pucBA*) that encode variants of LH2 α and β proteins (Tadros et al. 1993), and the genome sequence reveals that these gene pairs are dispersed throughout the chromosome (Larimer et al. 2004).

1.3. *R. Palustris* is a Model Organism for Proteomics

Because genome sequence information drives MS proteomics approaches to identify proteins that are present in restricted spatial or temporal domains (Aebersold and Goodlett 2001), the availability of the complete and annotated genome for *R. palustris* was a significant requirement, only recently achieved. In addition, there are a number of other advantages that suggest the use of *R. palustris* as a model organism for this type of study.

1. The abundance of membrane proteins, many of them extremely hydrophobic, present under photosynthetic conditions provides a model system to evaluate the difficulty in isolating measurable quantities of these proteins for MS analysis.
2. Although chromatophore membranes are widely thought to be photosynthesis-specific regions of the cellular membrane system (Varga and Staehelin 1985), it was not clear whether chromatophores also contain membrane proteins known to be present in undifferentiated cell membranes.
3. It is not clear if the complete set of photosynthetic proteins is currently known, and the evaluation of chromatophores may provide insight leading to the discovery of novel photosynthesis related proteins. As well, the products of hypothetical genes present in the genome near photosynthesis gene clusters may be detected, potentially implicating them in photosynthesis-related processes. (eg. RC or LH complex assembly factors (Aklujkar et al. 2000; Young and Beatty 2003).)
4. The diverse conditions under which *R. palustris* is able to thrive can be exploited, allowing the analysis of a simple genome that gives rise to a diverse set of proteomes, depending on the growth conditions.
5. *R. palustris*' prokaryotic genome does not contain introns, which eliminates the possibility of alternative splicing and greatly simplifies the analysis of the MS based proteomic results in comparison to eukaryotic organisms.

1.4. Mass Spectrometry

1.4.1. Mass Spectrometry of Biological Samples

Mass spectrometry (MS) is a powerful method which allows atoms or molecules to be separated simultaneously by mass and charge. By applying a magnetic field across the path of a molecule with a velocity, a change in the direction of travel can be induced. The size of the alteration in the course of the molecule is proportional both to the magnitude of the charge on the molecule and its mass. Thus, for a grouping of molecules containing a wide distribution of masses and charges, a single pass through a mass spectrometer will lead to a variety of peaks, with a single peak for each molecule, given a unique charge and mass. However, when charges are induced upon each molecule, it is possible to impart multiple charges to each molecule. This is reflected in the ratio of mass (*m*) to charge (*z*), referred to as the *m/z* ratio. This ratio determines the degree of effect the magnetic field will have upon the path of a given molecule. For two identical molecules with differing *m/z* ratios, two separate peaks will appear on the mass spectra (Peng and Gygi, 2001).

In practice, a single run is insufficient to separate and identify the components of a complex mixture. To alleviate this problem, MS is often combined with alternative methods to increase the resolution. For proteins, a common practice is to perform a 2D gel electrophoresis (2DGE) to separate as many of the proteins as possible, and perform MS on each of the resolved spots that can be visualized (Peng and Gygi, 2001). This approach has been a standard part of the MS repertoire and has demonstrated its reliability for abundant proteins. However, as the number of proteins increases and the need for high throughput techniques increases, this technique has become too slow and is unable to locate proteins that are not visualized, or that do not enter the resolving range of the 2DGE. Thus, more efficient methods have become necessary.

One such technique, used in the collection of data for this study, is LC/MS/MS, or liquid chromatography, tandem mass spectrometry. In this technique, the molecules in a sample are first separated over a liquid chromatography column, before being passed through the mass spectrometer (Peng and Gygi, 2001). This step greatly reduces the variety of molecules passing through the detectors at a given point in time, allowing for better detection and identification. The second MS step further reduces the number of compounds in the second detector by sampling individual peaks from the first mass spectra, allowing for a single molecule or small set of molecules to be analyzed at a time, despite the potential complexity of the original sample. Although sometimes used on samples containing full-length proteins, improved results are obtained when the sample proteins are pre-digested with trypsin, other proteases or a chemical treatment, such as cyanogen bromide, which reduces the average size of the molecules being injected (Kasper, 1970).

In particular, this technique has been successfully applied to proteins from other organisms (Kolker et al. 2003) as well as subcellular fractions (Ferro et al. 2003, Fejes et al. 2003). Because the peptide bond between amino acid residues is particularly susceptible to breakage, the addition of a collision chamber between the two mass spectrometry steps, breaks peptides into smaller fragments. The identical peptides collide with a stream of gas (e.g. Ar, N_2) in the collision chamber, shearing peptide bonds, generating smaller peptides that ideally represent all of the possible sequences that can be derived from the original peptide. In particular, for a peptide, p, comprised of n amino acid residues, the set of peptides $\{p_1 \ldots p_n, p_1 \ldots p_{n-1}, p_1 \ldots p_{n-2}, \ldots p_1\}$ and its corresponding set $\{p_1 \ldots p_n, p_{1+n} \ldots p_n, p_{2+n} \ldots p_n, \ldots pn\}$ are generated. Because they differ by a single amino acid removed from one end of the peptides, they generate a sequence of peaks, separated by the difference in mass of the amino acid(s) lost. By investigating the gaps between the peaks in a given spectrum and equating the difference in masses between two adjacent peaks to an amino acid of appropriate mass, a sequence of amino acids can be obtained. This allows for the sequence of the original peptide to be determined with great accuracy, and has been used to investigate many biological questions of importance (Graves and Haystead, 2002).

Despite the broad application of MS to biological samples, the applications to photosynthesis are few - for examples, see (van Wilk 2000; Hippler et al. 2001; Ouellette and Barry 2002; Zabrouskov et al. 2003). In this case, the chromatophores of *R. palustris*, in which the proteins that catalyze photosynthesis are located, are an excellent biological sample upon which the use of MS can be evaluated.

In addition to the MS of sub-cellular fractions and restricted domains, MS can also be used to enumerate the presence of proteins in whole cells and, potentially, their environments (Spoof et al. 2003). Because the composition of such a sample is more complex, the preparation and processing of the sample is also significantly more complicated. In this thesis, the separation of disrupted cell fractions was performed by centrifugation, which separates proteins broadly into soluble and membrane-bound proteins. Subsequently, LC/MS/MS can be used to provide a list of mass spectra of the peptides, and hence the proteins from which they are derived, located in each fraction.

1.4.2. Mass Spectrometry as a Tool for Probing Effects of Genetic Change

Mass spectrometry can not only be used to investigate the effects of environmental conditions upon the proteome of a cell, but also the effects of gene mutations. However, the methodology and visualization of the data are identical to that of changing environmental conditions. In this thesis, all references to investigating environmental changes upon a cell can also be broadly interpreted as being effective for the investigation to changes to gene expression (mutations) inside the cell.

To illustrate this point, the same procedures carried out upon the *R. palustris* cells for changing environmental conditions were also performed on an *R. palustris* strain carrying a deletion of the *lhaA* gene. This gene encodes a protein thought to be an LH1 assembly factor (Young and Beatty, 2003). Under photosynthetic (anaerobic, illuminated) conditions, where the expression of the protein is thought to be required for maximal assembly of LH1, *R. palustris* may display a reduced amount of photosynthetic pigments. The effect of this mutation is less obvious on the growth of cells under aerobic conditions, where photosynthesis gene expression is repressed and should not differ significantly from the wild type strain. The data that were observed for this sample support this hypothesis.

1.4.3. Caveats for the Use of MS on Biological Samples

Despite the tremendous potential of MS for the generation of proteomic information, there remain a number of pitfalls - particularly with the datasets used here. Chief among them are problems associated with post-translational modifications (PTMs), lack of cleavage sites for generating peptides and the intractability of membrane proteins.

Although post-translational modifications are not a problem with the MS itself, they represent a problem that has yet to be overcome. As genomic information is used to generate a list of all possible fragments that may be obtained from the MS itself, the more extensive the list is, the less likely a peptide is to be omitted from the final results. However, the longer the list of potential fragments, the more computational time is required to process the data obtained. In the case of post-translational modifications, although there are a limited number of them that are possible, a list of all of the peptides available from an organism's genome as well as all of the possible PTMs is quite extensive. Thus, for *R. palustris*, the peptide lists have all been generated with the exclusion of possible PTMs. Any peptide that contains a post-translationaly modified amino acid residue would simply not have been correctly identified, thus would be left out of the results.

A second problem with many MS analyses is in the processing of peptides. A frequent practice used is the tryptic digest of the full length proteins to yield shorter peptides. Trypsin is an enzyme catalyst for the degradation of proteins, which cleaves the peptide bond after K and R residues, when not followed by a proline residue (Kasper, 1970). When this approach is used, a lack of trypsin cleavage sites prevents a protein from being broken into sufficiently small fragments, leaving only the intact protein that may not be detected efficiently during the MS experiment. In addition, some trypsin cleavage sites may be inaccessible to the enzyme, which also results in an inability to yield fragments of the desired sizes. Although many alternative enzymes exist for the digestion of proteins, trypsin is one of the more site-specific proteases, and the use of alternative enzymes is uncommon.

Finally, the difficulty of locating membrane proteins in a MS sample is also a significant challenge. In addition to the difficulty in generating tryptic fragments for a protein that is embedded in a membrane (or a detergent after solubilization), the hydrophobicity of the protein makes it difficult to work with and presents numerous opportunities for the protein to be lost from the sample during processing. Fortunately, there are alternative methods under development to circumvent some of these problems, yielding short peptides which are more soluble and are less prone to being excluded from the results. Cyanogen bromide, which cleaves proteins at methionine residues (Quach et al. 2003), can be used as an alternative to trypsin to generate peptides and may provide improved results for membrane proteins.

1.5. Representation of MS Data

One of the greatest challenges to compiling an MS based proteome is to organize the data in a useful fashion that highlights the differences and similarities between samples in an easily recognizable manner. MS data are typically difficult for non-specialists to understand and practitioners have not yet settled on a single standard method by which proteomics data can be compared and displayed. In part, this exists because of the myriad ways of displaying MS data. Each peptide found can be given a number of different scores. Most noticeably, a so called "Xcorr" value can be associated with each peptide (MacMoss et al. 2002), which is a simplified value that takes numerous factors into account to give a rating of how likely a given predicted peptide matches the spectrum from which it was identified. The greater the Xcorr value, the greater the confidence that a given peptide was correctly identified. Derived from the Xcorr value and other MS parameters, a "probability value" can be calculated (Keller et al. 2002).

The probability value states the likelihood of the available data having given a correct prediction of the presence of the protein. This gives the likelihood of a given peptide or protein to have actually been present in the sample. Probability values are used in the analysis of chromatophore proteins (Section 3.2 - 3.4).

For larger samples, where greatly overlapping numbers of peptides are found for each protein, these data can be abstracted by one level. Instead of considering the probability value for each peptide, a cut-off can be used, so that only peptides with a high probability value are considered. In this case, those accepted peptides can be sorted by their protein of origin, and used to calculate the percent sequence coverage (Lauber et al. 2001). Percent sequence coverage indicates the fraction of the complete predicted protein that was located by the mass spectrometry analysis, which does not reflect the number of times that any given peptide fragment was found (Figure 3). It is simply the total percentage of amino acid residues identified in any number of peptides found, uniquely identified as originating from a single protein. Although it incorporates a bias based on the size of the peptide and the number of available cleavage sites, when percent coverage is compared between the same proteins in different samples, these issues become minimal. When the same procedure is used on two samples, identical numbers of a protein in the two samples would be subjected to identical treatments, and should give rise to identical percent sequence coverage. Only when the numbers of a protein in the two samples are different would greatly altered percent sequence coverage be observed. After comparing some of the available methods, percent coverage was chosen, in consultation with my collaborators, as the basis of the analysis for the *R. palustris* proteome in the whole-cell proteomics experiments (sections 4.5 - 4.7). However, it has not been settled within the MS community as to whether the percent coverage is the most accurate measure for the population of a given protein.

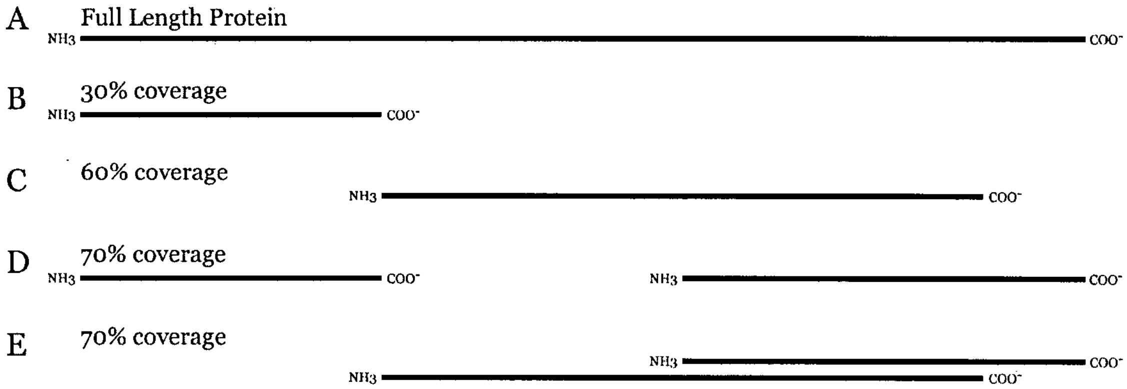


Figure 3: Percent coverage example.
Percent coverage forms the basis by which protein populations are deemed to change. A low percent coverage suggests a low population of the protein within the cell, while a greater percent coverage suggests more protein was present. The example above shows how percent coverage is determined compared to a full length protein (A). The bar shown in (B) indicates a single peptide covering 30% of the original full length protein from which it was derived (A) would yield 30% coverage for that peptide. The same is shown for a peptide of 60% in example (C). Where two peptides of 30% and 40% (D) of the length of the full protein from which they were derived (A), the percent coverage is summed to give 70%. When two peptides are located that overlap (E), the total percent coverage is the sum of the two peptides, subtracting the overlapping area, (40% + 60% - 30% = 70%)

Another approach that can be used to investigate the population of a single protein is to compare the number of unique peptides located in a given sample. In this case, unique refers to a unique sequence identified by mass spectrometry, irrespective of its charge. Because a single protein can give rise to a variety of fragments in the tryptic cleavage (or alternative enzymatic or chemical digestion) process, the population of fragments for similarly treated samples should be representative of the starting population. This allows for an alternative means to identify changes in the relative amounts of proteins under changing genomic or environmental conditions.

1.6. Overview of Thesis

In this thesis, the results of a proteomics experiment on *R. palustris* chromatophore preparations are described, using a shotgun approach that is essentially a simplification of a type of analysis described (Link et al. 1999). This represents the first step in applying proteomics methodology to reveal how the genome sequence of a purple phototrophic bacterium relates to its protein composition. These results also addressed some of the questions raised in section 1.3.

In a subsequent investigation, the proteome of the entire cell of *R. palustris* is considered. Using MS data of cells grown in a variety of environmental conditions, provided by our collaborators, I was able to create a new method by which the proteome information was compiled and displayed with colour-coding to make qualitative and quantitative aspects of the data intuitively interpretable. Ultimately, such techniques can be applied to understand how cells control and integrate all of the biological potential encoded in the genome sequence.

2. Materials and Methods

2.1. Chromatophore Purification

R. palustris strain CGA009 (Harwood and Gibson 1986) was cultivated photoheterotrophically (anaerobically) in a defined succinate medium designated PM (Kim and Harwood 1991), incubated at 30°C in a 20 liter glass fermentation vessel illuminated by four 100 W tungsten filament flood lamps. Cells were harvested by centrifugation and suspended in 2 ml of buffer A (10 mM Tris-HCl pH 8.0, 2 mM $MgCl_2$, 150 mM NaCl) for each gram (wet weight) of cell paste. Samples were maintained on ice or at 4°C in the subsequent steps. After addition of a few crystals of DNAse I, the cells were disrupted by two passages through a French press and the resultant sample was centrifuged at 30000 X g for 10 min. The supernatant fluid was centrifuged at 257000 X g for 120 min to pellet chromatophores, which were suspended in buffer A. The suspended chromatophores were layered over a sucrose step gradient (20/30/40/50% sucrose in buffer A) and centrifuged at 20000 X g for 18 hr, resulting in a band of chromatophores at approximately the 30/40% interface. This band was collected, diluted in two volumes of buffer A and the chromatophores pelleted (435000 X g for 15 min). The pellet was suspended in buffer A and banded a second time in a sucrose gradient as above, and the chromatophores pelleted and stored at -80°C.

2.2. Electron Microscopy

Chromatophores were negatively stained with OsO_4 for transmission electron microscopy using formvar coated copper grids, onto which 10 µl of a chromatophore suspension in buffer A were placed. After 1 min, the excess liquid was wicked off with a piece of filter paper, and 10 µl of a 2% OsO_4 solution were placed on the grid. After allowing the OsO_4 to stain the sample for 1 min, the excess solution was wicked off as above, and the grid was allowed to air-dry. Grids were examined in a Hitachi H-7600 transmission electron microscope.

Throughout the entire procedure for fixing and staining of cells for electron microscopy of thin sections (up to but not including the addition of 100% resin), the sample was agitated on a rotating rack. Unless otherwise noted, all steps were done at room temperature. Pelleted cells were fixed by suspension in a solution of 2.5% glutaraldehyde in 0.1 M sodium cacodylate buffer (pH 7.3) for 30 min, and washed three times in the cacodylate buffer without glutaraldehyde. The fixed cells were treated with OsO_4, for 30 min, followed by a single 10 min rinse in distilled water. Enblock stain was applied for 30 min, followed by a 10 min rinse in distilled water. Ethanol was introduced to replace the water. This was accomplished by successive 5 min washes in each of 30%, 50%, 70%, 85% and 95% ethanol solutions in distilled water, and three washes in 100% ethanol. Epon resin was used to embed the dehydrated sample and was introduced in increasing concentrations to replace the ethanol, beginning with a 3:1 ethanol:resin solution, subsequently increasing to 1:1, 1:3 and 100% resin at hourly intervals. The sample was allowed to sit in the 100% resin for 24 hours, and was then baked at 65 °C overnight. Rough sectioning and facing of the resin-embedded samples were performed on a Reichert OM3 Ultramicrotome. The thin sectioning of resin-embedded samples was performed using a Leica ULTRACUT UCT universal microtome and a Diatome diamond knife. The sections obtained with the diamond knife were placed on copper grids, stained with 2% uranyl acetate for 12 min, and stained with lead citrate for 6 min in the presence of NaOH pellets. After each stain a rinse in distilled water was performed, the excess water was wicked from the grid, and the grid was air dried on filter paper. Grids were examined in a Zeiss EM10C transmission electron microscope.

2.3. Mass Spectrometry of Chromatophores

Chromatophore pellets were sent to my collaborators Eugene Yi and David Goodlett at the Institute for Systems Biology, who resuspended the chromatophores in 0.5 ml of 50 mM NH_4HCO_3 (pH 8.3), 0.5% SDS. Membrane proteins were solubilized by incubating for 30 min at 60 °C with occasional vortexing. The resultant mixture was diluted to 0.05% SDS in 50 mM NH_4HCO_3 (pH 8.3), modified trypsin (Promega, Madison, WI) was added at an enzyme:substrate mass ratio of 1:100, and the sample was incubated overnight at 37 °C. Prior to LC/MS analysis, the resultant peptides were purified by OASIS MCX (mixed-mode cation-exchange reversed-phase; Waters, Beverley, MA) chromatography following the manufacturer's protocol. Peptide mixtures were injected into a C18 trap cartridge (Michrome Bioresources, Auburn, CA) for cleanup using a FAMOS autosampler (DIONEX, Sunnyvale, CA), and then passed through a 10 cm x 100 μm i.d. microcapillary HPLC (μ–LC) column packed with Magic C18 (Michrome Bioresources, Auburn, CA). The effluent from the μ–LC column entered a miniaturized electrospray ionization (μ–ESI) source in which peptides were ionized and passed directly into a quadrupole-time-of-flight (QTOF) mass spectrometer

(Waters, Beverly, MA). The C18 trap cartridge, μ–ESI-emitter/μ–LC column combination, a high voltage line for ESI and the waste line were each connected to separate ports of a four port union (Upchurch Scientific, Oak Harbor, WA) constructed entirely out of polyetheretherketone (PEEK). Ion selection for collision-induced dissociation (CID) was automated using top-down charge state-data-dependent ion selection of only $[M + 2H]^{2+}$ ions from a survey scan of 400-1500 *m/z* (including a 3 min dynamic exclusion period to prevent re-selection of previously selected ions), and repeated continuously throughout the μ–LC-ESI-MS/MS analysis. For each of three analyses, ~ 2 μg of total peptide (estimated from protein concentration prior to digestion) were loaded onto the combined μ–ESI-emitter/μ–LC column. This analysis was done in triplicate to increase coverage (Yi et al. 2002). Peptides were eluted by a linear gradient of acetonitrile from 5-32% over 150 min at a flow rate of ~ 300 nl/min. Proteins were identified by SEQUEST search of peptide tandem mass spectra (Finnigan MAT, San Jose, CA) against the *R. palustris* database (http://genome.ornl.gov/microbial/rpal/) using previously defined criteria (Eng et al. 1994), which excluded trypsin and keratin sequences. The raw data from all three μ–LC runs were combined for further analyses using guidelines as described (Goodlett et al. 2000). The complete dataset can be viewed at http://www.microbiology.ubc.ca/tbeatty/RpalChromMS.pdf.

2.4. Data Sources

2.4.1. P-SORTB

Predictions for protein localization were obtained through the use of the P-SORTB algorithm, available at http://www.psort.org/psortb/ (Gardy et al. 2003). This utility attempts to predict the localization of a given protein, classifying it as cytoplasmic, inner membrane, periplasmic, outer membrane or extracellular; or, where no prediction can be made, as unknown. Although the default threshold score for assigning localization is greater than 80% confidence, for the purposes of my work, if a given localization was greater or equal to 50% confidence, and there was no other localization proposed with greater than 30% confidence, the threshold was lowered to accept the localization with the greatest confidence.

2.4.2. COGS

Clusters of orthologous groups (COGs) (Tatusov et al. 1997) were obtained from the Computational Biology at ORNL website, at http://genome.ornl.gov/microbial/rpal/1/fun.html, and from Loren Hauser at Oak Ridge National Laboratories (ORNL) (personal communication). COG data were imported to the Structured Query Language (SQL) database (see below) through intermediary steps which included either the creation of a Microsoft Excel intermediate file or through a text file, which could be imported into the database.

2.4.3. Genome

The complete genome and its annotation were obtained through the Oak Ridge National Labs Genome Analysis Group (available at http://genome.ornl.gov/microbial/rpal/). All predictions for transmembrane segments generated by the TMHMM software (Sonnhammer et al. 1998) and signal peptides generated by the SignalP algorithms were derived from the feature tables of the annotated genome, viewable through Artemis software (Rutherford et al. 2000).

2.5. Growth of *R. palustris* and Processing of Samples for Whole Cell Proteomics

For the whole cell proteome experiments, all steps prior to the transfer of data (sections 2.5-2.6) were performed by my collaborators in the Harwood (Iowa State University), Tabita (Ohio State University) and Larimer (Oak Ridge National Labs) laboratories.

2.5.1. Anaerobic/Photosynthetic (Photoheterotrophic Growth)

R. palustris strain CGA009 (Harwood and Gibson 1986) was cultured anaerobically, in PM (Kim and Harwood, 1991), a defined mineral medium plus sodium bicarbonate to final concentration of 10 mM. The carbon source was succinate at a final concentration of 10 mm. Cultures were grown anaerobically, illuminated by 40 W incandescent and 32 W fluorescent lamps at 30°C. Starter cultures were grown in 150 ml screw cap bottles filled to the top with PM, at pH 6.8. 15-20 ml of starter culture were used to inoculate 1.1 L anaerobic cultures. Growth was monitored spectrophotometrically and cells were harvested at final OD 600 nm of 0.9. The final pH of the culture was 7.26 and the cells were collected by centrifugation (6000 rpm for 20 min) and washed once with PM and stored at - 80^0C. Yield was approximately 1.4 g wet weight of cell paste/1.1 L culture.

2.5.2. Aerobic (Chemoheterotrophic Growth)

R. palustris strain CGA009 (Harwood and Gibson 1986) was grown on PM media modified to contain 25 mM succinate and 0.5 g/L yeast extract as the carbon sources. Colonies were taken from a fresh agar plate, and used to inoculate small shaker flasks, which served as seed cultures. Six baffled Fernbach flasks, were inoculated with 100 mL each and grown in the dark at 30°C, aerobically and shaken at 150 RPM. Cells were harvested at an OD 600 nm between 1.2 and 1.5. Cells were collected by centrifugation at 4°C and frozen at - 4°C.

2.5.3. LhaA mutant (Chemoheterotrophic Growth)

The LhaA deletion mutant, derived from wild type *R. palustris* CGA009 (Harwood and Gibson 1986), was grown exactly as the aerobic wild type sample (section 2.5.2).

2.5.4. Nitrogen fixation (Photoheterotrophic Growth)

R. palustris strain CGA009 (Harwood and Gibson 1986) was cultured as the anaerobic samples with the following exceptions; ammonium sulfate was replaced with sodium sulfate in the PM medium and nitrogen gas was added to the head space.

2.5.5. Lysis and Fractionation

4.0 grams of *R. palustris* bacterial pellet were washed twice in 15 ml cold buffer (50 mM Tris pH 7.5, 10 mM EDTA). The first wash was done in a 40 ml centrifuge tube and spun at 6000 rpm for 5 minutes. After the second wash, cells were resuspended at 20% cell wt/buffer volume (2.0 g cell / 10 ml buffer) and divided into 2 tubes, kept on ice, for sonication using a Branson 185 sonifer cell disrupter. Each sample was sonicated 10 times for 0.5 minute at ~ 4 power with cooling between each 0.5 minute sonication. Samples were spun twice for 10 minutes at 6500 rpm (5000 g) at 4°C. The supernatant was removed and transferred to 10 ml centrifuge tubes and spun for 1 hour at 33,000 rpm (100,000 g) in a Beckman Ultracentrifuge at 4°C.

Four fractions were collected. The crude soluble fraction was collected as half of the supernatant (~9 ml) from the one hour spin. The pellet was resuspended (by sonication) and washed twice in 9 ml of buffer and spun for 33,000 rpm (100,000 g) for one hour to yield the membrane fraction. The remaining soluble fraction from the 1 hour spin was collected and spun overnight in 10 ml centrifuge tubes at 33,000 rpm at 4°C, and was designated as the cleared soluble fraction. The pellet obtained from the overnight spin was collected and resuspended (by sonication) in 5 ml of buffer and was termed the ribosomal fraction. All four fractions were stored as 1 ml aliquots in 1.5 ml Eppendorf tubes at - 80°C.

2.5.6. Trypsin digestion

5 mg from each fraction obtained in section 2.5.5 (0.5-2.0 ml) were digested with sequencing-grade trypsin from Promega in 15 ml centrifuge tubes: Samples were adjusted to 6 M guanidine and reduced with dithiothreitol (DTT) (10 mM) for 1 hour at 60°C. Samples were cooled to room temperature and the guanidine was diluted with (~5 volumes) 50 mM Tris, 1 mM $CaCl_2$, pH 7.5. Trypsin was added to the samples at a concentration of 20 μg trypsin/mg protein. Tubes were put in a 37°C incubator on a rotator for 6 hours. Trypsin (20 μg/mg protein) was added again to the samples for overnight digestion. 10 mM DTT was subsequently added for further reduction.

2.5.7. Sep-Pak Clean Up

Using a 10 ml syringe, +C18 Sep-Pak (Waters), was pre-wet with 0.1% triflouroacetic acid (TFA) in acetonitrile (ACN) and rinsed with 0.1% TFA in water. The sample was pushed slowly through the cartridge twice and washed with 0.1% TFA in water. Bound protein (peptides) were eluted with 0.1% TFA in ACN. The sample was processed a second time as above, using a new cartridge. Four fractions were collected per Sep-Pak (~12 ml total). Samples were reduced to 500 μl volume using a Savant Speed Vac system. Estimated concentrations were 10 μg of protein/μl.

2.6. MS Analysis of Fractionated Cells

Dried samples from all four fractions were dissolved in 500 μl water containing 0.1 % TFA to an estimated concentration of 10 μg/μl. The samples were analyzed by 1D LC-MS/MS (Electrospray) with 5 m/z ranges (400-700, 690-900, 890-1200, 1190-1500, 1490-2000) and a full mass range for chromatography diagnostics. Samples were injected (60 μl per injection) by the Famos Autosampler on to a 50 μl loop and separated with a 25cm x 300um 18 VYDAC 218MS column into the electrospray source at a flow rate of 4 μl/min, voltage of 4.5 kV with sheath gas 40, multiplier of 900 V, a heated cap at 175°C with an MS/MS isolation width 5 m/z. Solvent A contained 95% H_2O, 5% CAN and 0.5% formic acid. Solvent B contained 30% H_2O, 70% ACN and 0.5% formic acid. The gradient used through the MS procedure was as follows, from 0-20 minutes, 100% solvent A was used. From 20-200 minutes, 100% solvent A was slowly replaced with solvent B until only 50% solvent A was used. From 200-230 minutes, solvent A was slowly reduced until the 50% solvent A - 50% solvent B was replaced with 100%

solvent B. From 230-240 minutes, 100% solvent B was used, and the MS was turned off. Solvent A was used for 25 minute to re-equilibrate the column.

All .Raw files obtained were converted to .dta files with Finnigan .dta generator, all .dta files were searched with SEQUEST against an *R. palustris* database with SEQUEST (Finnigan MAT, San Jose, CA) using autosequest_batch.pl (trypsin as enzyme). Data from fractions (section 2.5.5) were combined to generate whole cell proteomes. All data were processed with DTASelect and Contrast.

2.7. Software

2.7.1 Overview of Software

Data can be obtained from a variety of sources, with the minimum requirement that each row of information be associated with a single gene. If this is condition is met, it can be included in the dataset and correlated with any other source. Once the information has been collected in a single database, it can be processed within the database for on-the-fly queries, or exported for visualization and portability. Contrast files are the first file format produced from the MS data that can be used by a bioinformatics approach to interpret the results. They provide the core information used in the database, but can also be used independently to identify genes of interest, through the use of the "Gene Caller" program. (Figure 5)

2.7.2. Gene Caller

As part of my thesis, I developed a new software program, termed "Gene Caller", which selects genes of interest based on either the percent coverage, or the number of unique peptides (peptides of the same sequence, regardless of their charge) found in the

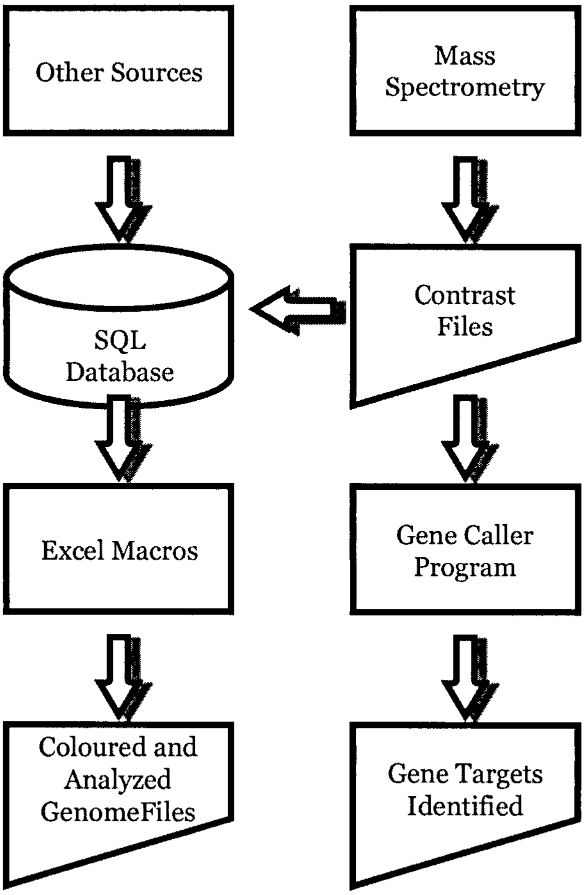


Figure 4: Flow chart of data processing.
Schematic representation of the processing of data in this thesis. Mass spectrometry is processed into Contrast files, which can be integrated into a database, or analyzed with the "Gene Caller" program. The Gene caller provides a text file listing of the genes of interest (section 2.7.1). The database, in addition to processing MS data, can also be used to integrate and correlate other data source. These have included P-SORTB data (section 2.4.1), COGs data (section 2.4.2) and further genomic data (section 2.4.3). Using the SQL driven database, queries can be exported to Microsoft Excel files, where macros or other processing can be done (section 2.7.3). The end product of the processing in this thesis is the Microsoft Excel spreadsheet which contains the colourized and further annotated proteome.

MS contrast file, containing the datasets for two different samples. If the percent sequence coverage is used, the sum of the percent sequence coverage for one sample must exceed the sum of the percent sequence coverage of the other sample by 30%, for each run included in the contrast file.

$$\sum_{i=1}^{n}(X_i) \geq \sum_{i=1}^{n}(Y_i) + (30 \cdot n)$$

or

$$\sum_{i=1}^{n}(X_i) \leq \sum_{i=1}^{n}(Y_i) - (30 \cdot n)$$

Where n is the number of runs used, and X and Y are the percent coverage for runs i to n. If the number of unique peptides is used, a similar equation is used,

$$n_x \geq n_y + (n_{xy} \cdot t)$$

or

$$n_x \leq n_y - (n_{xy} \cdot t)$$

Where t is the threshold value, n_x and n_y are the number of unique peptides for sets x and y respectively, and n_{xy} is the total number of unique peptides for the given protein under both growth conditions. Values for t greater than 0.5 give an acceptable sensitivity, and 0.6 was used as the standard threshold.

As a default value, the program was written to accept two runs for each sample, but can be easily expanded to process larger datasets. For this thesis, the samples used were obtained from whole cells grown as described in section 2.5.

2.7.3. Microsoft Access/SQL Database

Data were incorporated into a single central SQL driven database (Microsoft Access), in which SQL queries are created through either the graphical interface or through the SQL text interface (see Figure 4).

MS peptide data obtained from Oak Ridge National Labs were imported with the following column names: *ID*, *Orf*, *Length*, *MolWt*, *pI*, *Peptide Charge*, *SubDir*, *XCorr*, *DeltaCN*, *PrecursorMass*, *TotalIntensity*, *SpRank*, *IonProportion*, *Sequence*, *SequencePos*, *Tryptic* and *UniqeToLocation*. *ID* is an auto-incrementing key field, generated at the time of import, but is not used for any other cross referencing. *Orf*, *Length*, *MolWt* and *pI* are all properties of the protein with which a given peptide has been associated. *Peptide Charge*, *Xcorr*, *DeltaCN*, *PrecursorMass*, *TotalIntensity*, *Sequence* and *SequencePos* are all properties of the located peptide, and are used to filter out peptides which do not pass the minimum confidence threshold. *UniqueToLocation* is used to specify whether a given peptide is unique to that protein, as it is possible for a located sequence to be a subsequence of more than one protein. Approximately 300,000 peptides, with unique *m/z* ratios, are located for each run of single sample, and are stored in a table unique to the run.

For comparisons using percent sequence coverage, contrast files were also imported into the SQL database, with the headings: *ID* and *Locus*, followed by a single heading for each run obtained, containing the percent coverage for each run. *ID* is again an auto-incrementing key field, not used for cross-referencing between tables, whereas the *Locus* column contains the orf number of the protein located. All of the data used in generating tables in this thesis where percent coverage formed the basis for the analysis were obtained from a single contrast file, imported to the database to yield a table with the name *tblAllRunsCoverage*. This table contains a line for each protein found in any run incorporated into the contrast file. A full list of proteins encoded by the genome and their percent coverage, regardless of whether the protein was observed in any run, can be obtained by executing an outer join on the *crossref* table and *tblAllRunsCoverage*.

Psort-B Data was included in the table *tblPSORT-B*, with the column headers *ID*, *SeqID*, *New Number*, *HMMTOP_Localization*, *HMMTOP_Details*, *Motif_Localization*, *Motif_Details*, *OMPMotif_Localization*, *OMPMotif_Details*, *SCL-BLAST_Localization*, *SCL-BLAST_Details*, *Signal_Localization*, *Signal_Details*, *SubLocC_Localization*, *SubLocC_Details*, *Cytoplasmic_Score*, *InnerMembrane_Score*, *Periplasmic_Score*, *OuterMembrane_Score*, *Extracellular_Score*, *Final_Localization*, *Final_Score and Anthony'sLocalization*. *ID* is again an auto-incrementing key field, not used for cross-referencing between tables, whereas the *SeqID* column contains the orf number of the protein to which the PSORT-B prediction belongs, as well as the gene name. *New Number* holds an alternative annotation for the *R. palustris* genome, which can be cross referenced through the *crossref* table. (New annotations are in the form RPAxxxx, where xxxx is the numerical designation of the gene. The previous annotation, for which the numbering method does not correspond directly with the new annotation, was in the form orfzzzz, where zzzz was the numerical designation of the gene.) All column headers ending in _Localization or _Details contain information about a prediction for a given localization except *Final_Localization*, which gives a text label, indicating the predicted localization of the protein. Columns ending in _Score contain a final likelihood of the protein being found in a given cellular location. The column labeled

Anthony'sLocalization contains an alternative prediction to *Final_Localization*, with lowered thresholds. See section 2.4.1 for an explanation of thresholds used.

Information on COGs were kept in a table labeled *tblCogs_summary*, with the column headings *Contig*, *Gene*, *Num Prot*, *Group*, *COG*, *COG Description*, *Gene Name*, *Score*, *E-Value* and *Category*. *Gene* is the key field in this table, and contains the orf number of the gene for which a COG is assigned. The *COG* column contains the group number for the COG to which the gene has been assigned. The *COG Description* field contains a short description of the COG to which the gene has been assigned. *Category* and *E-value* contain further information on the COG and the probability of that gene belonging to the assigned COG, respectively.

Data from the MS of the chromatophore fractions were stored in the table *Mass Spec Data -Chromatophore*, in the format used by the SEQUEST (Finnigan MAT, San Jose, CA) software. A summary of those data was generated through the query *qryMassSpecChromatophore(Summary)*, containing two fields; *Orf* and *MaxOfProbability*. The *Orf* column contains a unique gene number, while the *MaxOfProbability* contains the maximum probability obtained for any peptide identified as belonging to that gene.

To generate the dataset exported to Microsoft Excel, the query *qryCOG-MS11 (By Strand/Gene)* was used, for which the SQL is given in Appendix 1. Sorting was done by DNA strand location and by the new annotation numbers. The final queries were exported to Microsoft Excel spreadsheets using the export feature.

2.7.4. Microsoft Excel Macros

Spreadsheets were opened in Microsoft Excel, where three Visual Basic for Applications (VBA) macros were sequentially used to perform colorization, formatting and further calculations. The first macro demarks where genes on one DNA strand are adjacent, from which some basic information on potential operons can be inferred. The macro also colours the percent coverage based on an array-like colouring scheme, to create a more intuitive and visual means of viewing the data (see Figure 6 and Figure 7).

The second Excel macro uses the percent coverage data to determine a baseline expression level for each protein product. This is done by finding the highest percent coverage for each protein in a single growth condition and comparing these values across the sampled growth conditions to find the lowest highest percent coverage for the whole dataset. This can be expressed by the formula

$$\min_{n=1..j}\left[\left(\max_{m=1..k}\{X_m\}\right)_n\right] > 0$$

where j is the number of environmental conditions used and k is the number of assays performed on each environmental condition. These values are sorted from lowest to highest, and the lowest one greater than 10% is used. If no value is found to meet that criterion, a baseline value of 10% is chosen. All other highest percent coverage values are then divided by the baseline to generate the normalized data (see Figure 8). By using a minimum value of 10% for the normalization denominator, the range over which any protein may be up-regulated or down-regulated is constrained below tenfold increase in expression. Alternate values may be used, depending on the range of normalized expression levels desired.

The third Excel macro uses the normalized data to evaluate and predict the expression levels of the genes based on the expression profile. To accomplish these tasks, predetermined criteria are used (Table 1). For each condition that was defined numerically, an up-regulation value of at least 0.4 or greater than the threshold value was required to be included in a given profile. For example, if nitrogen fixation conditions were being tested (row 1 of Table 1), the nitrogen fixation sample would need to be greater than each of the normalized values for the aerobic, anaerobic environmental conditions and the LhaA mutant by at least 0.4, relative to the baseline;

$$X_i > X_j + 0.4$$

or

$$X_i < X_j - 0.4$$

where X is the value of the normalized data for two different environmental conditions, indicated respectively by i and j.

Table 1: Criteria used to evaluate protein expression patterns

Column	Aerobic	Anaerobic	LhaA mutant	N2 fixation
N2 Fixation criteria:	N2 fixation > Aerobic	N2 Fixation > Anaerobic	N2 Fixation > LhaA mutant	---
Aerobic criteria:	---	Aerobic > Anaerobic	---	Aerobic > N2 fixation
Anaerobic criteria:	Anaerobic > Aerobic	---	Anaerobic > LhaA mutant	---
LhaA mutant criteria 1:	LhaA mutant > Aerobic	LhaA mutant > Anaerobic	---	Lha mutant > N2 fixation
LhaA mutant criteria 2:	LhaA mutant < Aerobic	LhaA mutant < Anaerobic	---	Lha mutant < N2 fixation
Unused criteria:	Aerobic = 0	Anaerobic = 0	LhaA mutant = 0	N2 Fixation = 0
Always on criteria:	Aerobic <> 0	Anaerobic <> 0	LhaA mutant <> 0	N2 Fixation <> 0
Unchanged criteria:	0.4 < Aerobic < 1.5	0.4 < Anaerobic < 1.5	0.4 < LhaA mutant < 1.5	0.4 < N2 Fixation < 1.5

1) A further condition is checked, in which none of the above conditions occur.
2) If all criteria for a row are observed, then the condition is accepted as at least a weak hit.
3) To be accepted as a likely (strong) hit, a minimum distance of 0.4 is required for all conditions in a given row, where inequalities are observed (rows 1-5).
4) "---" indicates that the data from the baseline normalized dataset indicated in the column were not used in determining the expression pattern in the row shown.
5) Criteria are assessed using normalized data (section 2.7.3)

3. Results of the MS of *R. palustris* Chromatophores

3.1. Cell Membrane and Chromatophore Structures

As shown in Figure 1A, electron micrographs of thin sections through cells indicated that these *R. palustris* cells produce an intracytoplasmic membrane system that consists of large, flattened sacs layered in parallel, as previously described (Drews and Golecki 1995; Varga and Staehelin 1983). After cell disruption and sucrose density gradient purification of chromatophores, electron microscopy of negatively stained preparations showed vesicles ranging from approximately 100 to 400 nm in diameter (Figure 1B). Larger structures of irregular shape were often seen. Although the relative amounts of cytoplasmic and intracytoplasmic membranes in these preparations were not determined, the protein composition should consist predominantly of proteins imbedded in intracytoplasmic membrane fragments, proteins bound to the surface of these membranes (such as by interaction with membrane-imbedded proteins), and soluble proteins entrapped within vesicles. Electron microscopy does not reveal whether chromatophore vesicles contain periplasmic or cytoplasmic components (*i.e.*, membrane vesicle topology), but previous experiments on *Rhodobacter* species indicated that essentially all vesicles resulting from French press disruption of cells were formed with the cytoplasmic face on the outside (Prince et al. 1975). Thus, soluble periplasmic proteins such as cytochrome c_2 could be present within these *R. palustris* chromatophore vesicles.

3.2. Photosynthesis-Specific Proteins Detected in Chromatophores

Table 2 gives a list of RC, LH1 and LH2 proteins that were detected, on the basis of the mass spectra of tryptic peptides. These proteins represent controls, in the sense that they are known to be extremely hydrophobic proteins and known to be major components of *R. palustris* chromatophores (Varga and Staehelin 1985). The detection of all three RC proteins as well as both proteins of LH1 indicates that even these hydrophobic proteins contained segments that were accessible to trypsin cleavage, and that the resultant peptides were resolved in the chromatography system used. It is surprising that the maximum probability value of RC L peptides (0.3792) was significantly lower than the maximum value of RC M peptides (0.9999). These two proteins are about 30% identical in sequence alignments, both contain five transmembrane segments and they exist in a 1:1 ratio in the RC. A number of factors enter into the SEQUEST probability value calculation (Keller et al. 2002), but in this case at least part of the reason for the difference may relate to 21 RC M-assigned peptides being obtained as opposed to 10 RC L-assigned peptides (data not shown). (For complete data, see http://www.microbiology.ubc.ca/tbeatty/RpalChromMS.pdf.) In turn, the larger number of RC M peptides detected may be due in part to the larger number of predicted trypsin cleavage sites in RC M (17 vs. 13 in RC L).

There are five potential LH2 structural (*pucBA*) gene operons indicated by the *R. palustris* strain CGA009 genome sequence, although the DNA sequence of the orf RPA3010 (LH2 α protein) appears to contain a frameshift (http://genome.ornl.gov/microbial/rpal/). Although five *pucBA* gene pairs were previously reported in another strain of *R. palustris* (Tadros et al. 1993), several

Table 2: Photosynthetic light energy transduction proteins

Protein	Orf Number	Maximum Probability Value[a]
RC L	RPA1527	0.3792
RC M	RPA1528	0.9999
RC H	RPA1548	0.9999
LH1 β	RPA1525	0.9951
LH1 α	RPA1526	0.8762
LH2 β	RPA4291	0.0091
LH2 α	RPA4292	0.8492
LH2 β	RPA1491 [b]	0.0091
LH2 α	RPA1492	ND[c]
LH2 β	RPA3009	ND
LH2 α	RPA3010 [d]	ND
LH2 β	RPA3013	0.8801
LH2 α	RPA3012	ND
LH2 β	RPA2654	ND
LH2 α	RPA2653	ND

[a] Highest score of all peptides attributed to the designated protein.
[b] Proteins have the same sequence.
[c] Not detected.
[d] Not present in the database because the DNA sequence indicates a frameshift.

sequences differ from the strain CGA009 homologues (including the absence of a frameshift). All five *pucBA* gene pairs were reported by Tadros et al. (1993) to be transcribed as dicistronic *pucBA* messages when cultures were grown under a high light intensity and three were transcribed under low light intensity. The culture used in our experiments was grown at an intermediate light intensity and, as shown in Table 2, peptides of only one LH2 α (orf RPA4292) and one LH2 β (orf RPA3013) protein were detected with high probability values. These proteins are predicted to be encoded by genes that are located in separate transcription units. One or both of the identical LH2 β proteins of orfs RPA4291 and RPA1491 may have been detected, but the probability value was extremely low. All of the LH2 proteins are predicted to contain at least three trypsin cleavage sites, and so either some residues were not accessible by trypsin or the proteins were absent from the chromatophore sample. Because the products of orfs RPA4292 (LH2 α) and RPA3013 (LH2 β) were cleaved by trypsin and peptides detected with high probability values, and the predicted amino acid sequences of the eight other LH2 proteins are (respectively) very similar, it appears that the predominant LH2 complex in these chromatophores consisted of the orf RPA4292 α and orf RPA3013 β proteins. This is surprising given the almost certain existence of separate orf RPA4291/RPA4292 and orf RPA3013/RPA3012 *pucBA* transcripts.

Peptides of Bchl and carotenoid biosynthetic enzyme homologues were also found in this chromatophore sample, as summarized in Table 3. The high probability values indicate that these enzymes genuinely co-purified with the chromatophores.

Table 3: Photosynthetic pigment biosynthetic enzymes

Protein	Orf Number	Maximum Probability Value[a]
BchI	RPA1506	0.9923
CrtI	RPA1512	0.7426
CrtF	RPA1520	0.7804
BchX	RPA1522	0.6865
BchY	RPA1523	0.9999
BchP	RPA1532	1.0000
BchM	RPA1546	0.9998

[a] Highest score of all peptides attributed to the designated protein.

3.3. Photosynthesis-Related Proteins

Table 4 lists *R. palustris* homologues of proteins that have been shown to catalyze both photosynthetic and respiratory electron transfer reactions in other species (Meyer and Donohue 1995; Zannoni 1995). The *R. palustris* genome DNA sequence predicts that the cytochrome *b/c₁* complex cytochrome *b* and *c₁* moieties are present in a single polypeptide sequence, encoded by orf RPA1193. A similar gene arrangement exists in *Bradyrhizobium japonicum*, in which a precursor protein is cleaved to yield the separate cytochromes *b* and *c₁* that are typically found in cytochrome *b/c₁* complexes (Thöny-Meyer et al. 1991). Therefore *R. palustris* may cleave a precursor protein to yield both cytochromes as separate molecules, which gave rise to peptides that were found with high probability values (Table 4).

Table 4: Relevant electron transfer proteins.

Protein	Orf Number	Maximum Probability Value[a]
Cyt b/c_1[b]	RPA1193	1.000
Rieske Fe/S	RPA1192	1.000
Rieske Fe/S	RPA1692	ND[c]
Cyt c_2	RPA1535	0.9959
HIPIP	RPA0744	0.0437
Cyt c_y	RPA3693	0.9885

[a] Highest score of all peptides attributed to the designated protein.
[b] DNA sequence predicts a single polypeptide.
[c] Not detected.

The *R. palustris* genome sequence encodes two putative Rieske iron-sulfur proteins (orfs RPA1192 and RPA1016) that are 52% identical in an alignment. The orf RPA1192 sequence contains an approximately 35 amino acid N-terminal extension relative to orf RPA1016 and *Rhodobacter* homologues, and appears to be co-transcribed with the predicted cytochromes *b*- and c_1-encoding orf RPA1193. The MS data (Table 4) indicate that of these two potential Rieske iron-sulfur proteins only the orf RPA1192 protein was present in the chromatophore preparation, which is designated as the iron-sulfur protein component of the *R. palustris* cytochrome b/c_1 complex.

The periplasmic, soluble cytochrome c_2 was at one time thought to be obligatory for electron transfer from the cytochrome b/c_1 complex to the RC in purple phototrophic bacteria, but several types of electron carriers have been found to perform this function. These include alternative cytochromes such as the membrane-anchored cytochrome c_y of *Rhodobacter* species, and a high potential iron protein called HIPIP (Meyer and Donohue 1995). Homologues of all of these three electron carriers are encoded in the *R. palustris* genome sequence, but only the putative cytochromes c_2 and c_y were found with confidence in the chromatophore preparation (Table 4). The detection of cytochrome c_2 indicates that this predicted soluble protein was entrapped within chromatophore vesicles, as was previously reported for *Rhodobacter* species (Prince et al. 1975). The low probability value of 0.044 for the single HIPIP- assigned peptide indicates that this protein may not have been present.

The final membrane protein complex required for transduction of light energy to the potential energy of the phosphate ester bond in ATP is the F_1F_0 ATPase. Table 5 lists the nine proteins that are predicted by the genome sequence to comprise the *R. palustris* F_1F_0 ATPase, of which seven were detected with high probability values. The two proteins that were not detected (F_0 components a and c) are the most hydrophobic components of this complex (see Figure 2), and so perhaps they were not detected because they were not solubilized or cleaved by trypsin. A recent review noted that these two proteins have been recalcitrant in proteomics experiments on other organisms (Patton et al. 2002).

Table 5: F_1F_0 ATPase complex proteins.

Protein	Orf Number	Maximum Probability Value[a]
a	RPA0846	ND[b]
c	RPA0845	ND
b'	RPA0844	1.0000
b	RPA0843	1.0000
δ	RPA0179	1.0000
α	RPA0178	1.0000
γ	RPA0177	0.9994
β	RPA0176	1.0000
ε	RPA0175	1.0000

[a] Highest score of all peptides attributed to the designated protein.
[b] Not detected.

3.4. Hypothetical Proteins

Genome sequence analyses typically yield a large number of orfs that could potentially encode proteins that either have no homologues in the protein sequence databases, or are homologous to genomic sequences of unknown function, and therefore are annotated as encoding hypothetical or conserved hypothetical proteins. These orfs are of interest as candidates for gene disruption to reveal new biological functions, especially if the predicted proteins can be shown to be temporally produced or spatially located within the cell along with proteins that have a known biological activity. This approach is most likely to provide information of specific phenotypic significance if an orf appears to be co-transcribed or at least clustered with genes known to encode proteins involved in a cellular process. Table 6 lists all the orfs containing the word "hypothetical" in the genome draft annotation (http://genome.ornl.gov/microbial/rpal/) that were identified on the basis of a chromatophore peptide identified with >0.75 probability value, and which were located near a putative photosynthesis-related gene. These data show that these orfs are genuine genes, which encode chromatophore-associated proteins that may have functions related to photosynthesis.

Table 6: Proteins encoded by potentially photosynthesis-related "hypothetical" orfs

Orf Number	Maximum Probability Value[a]	Orf Neighbors and Properties
RPA0257	1.0000	next to homologue of uroporphyrinogen -III synthase orf RPA0256
RPA0258	0.9998	part of cluster with orfs RPA0256, RPA0257, RPA0260 and RPA0259
RPA0261	0.9999	next to homologue of photosynthesis gene regulator, AppA/PpaA orf RPA0260
RPA1494	1.0000	next to *pucC* orf RPA1493, possibly part of LH2 operon
RPA1495	1.0000	next to orf RPA1494, possibly part of LH2 operon
RPA1510	0.9994	in cluster of *bch*/*crt* genes, similar to conserved proteins in photosynthetic organisms
RPA1549	0.7694	overlaps *puhA* (RC H) gene, putative assembly factor
RPA1550	0.9998	possibly co-transcribed with *puhA* and orf RPA1549, putative assembly factor
RPA3011	1.0000	next to LH2 α orf RPA3012
RPA1504	0.9956	next to *bch* genes
RPA0259	1.0000	next to homologue of photosynthesis gene regulator, AppA/PpaA orf RPA0260

[a] Highest score of all peptides attributed to the designated protein.

4. Results of the MS of the *R. palustris* Whole Cell Proteome

4.1 Analysis of Fractions

During preparation of the whole cell proteome, four fractions were obtained and termed as "membrane", "ribosomal", "cleared soluble" and "crude soluble" fractions (see section 2.5.5). Each fraction was injected into the LC-MS-MS, and data were obtained on each fraction individually. Attempts to analyze these fractions separately were unsuccessful because of poor resolution between samples. For example, ribosomal proteins were found in all four fractions and were not significantly enriched in the "ribosomal" fraction (data not shown). This led to the decision to pool the MS data, creating a single proteome for the whole cell. Thus, subsequent computational processing and the creation of Contrast files were performed using combined data, for each MS run (section 2.6).

4.2. Analysis of Contrast Files by Peptide and Coverage

Once contrast files have been obtained for a set of runs (see section 2.6), they can be processed in numerous ways. In addition to their incorporation into the SQL database, they can also be analyzed in their raw format. This allows the raw peptide information, which is lost from the final database version of the table, to be processed. To accomplish this goal, the "Gene Caller" program was written. This short C++ program can be run from a compiler, or as a standalone DOS/Windows program that can be modified to accept command line parameters, including the name of the contrast file to be processed.

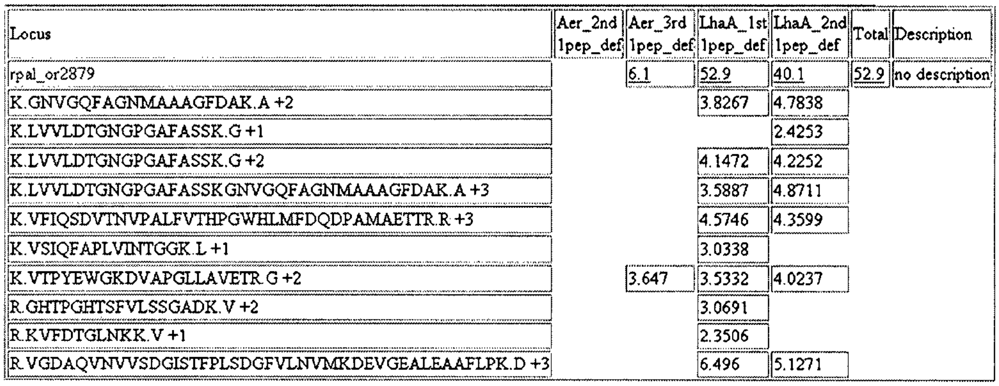
The "Gene Caller" program was designed to identify if any proteins have significantly changed their level of expression between two different environmental conditions. The contrast file, containing two separate runs of each of two environmental conditions, is required as input, and can be used as generated by the Contrast software. The "Gene Caller" moves through this file in a linear sequence, checking both the percent coverage and the number of unique peptides for each protein that has been identified in the MS experiment. As each protein is inspected, the "Gene Caller" will compare the results against threshold criteria set by the program, and if the protein passes those criteria, writes out to a plain text file the name of the orf of the protein, which criteria it passed, and the margin by which it passed.

Figure 5 includes an excerpt from the contrast file, showing the listing for orf 2879, now annotated as RPA2488, a conserved unknown protein, as well as an excerpt from the output of the "Gene Caller" program. (Contrast file data have not yet been generated using the new annotation for the *R. palustris* genome.)

4.3. Compilation of Non-Uniform Data Sources and Database Structure

Much of the information included in the database came from a diverse range of sources including web pages, spread sheets and flat data files. Compiling these data sources into a single accessible database provided a key to generating an interface which provides sufficient information as a base for analysis. The challenge involved in importing these sources into the database is variable for each type of information.

A



Locus	Aer_2nd 1pep_def	Aer_3rd 1pep_def	LhaA_1st 1pep_def	LhaA_2nd 1pep_def	Total	Description
rpal_or2879		6.1	52.9	40.1	52.9	no description
K.GNVGQFAGNMAAAGFDAK.A +2			3.8267	4.7838		
K.LVVLDTGNGPGAFASSK.G +1				2.4253		
K.LVVLDTGNGPGAFASSK.G +2			4.1472	4.2252		
K.LVVLDTGNGPGAFASSKGNVGQFAGNMAAAGFDAK.A +3			3.5887	4.8711		
K.VFIQSDVTNVPALFVTHPGWHLMFDQDPAMAETTR.R +3			4.5746	4.3599		
K.VSIQFAPLVINTGGK.L +1			3.0338			
K.VTPYEWGKDVAPGLLAVETR.G +2		3.647	3.5332	4.0237		
R.GHTPGHTSFVLSSGADK.V +2			3.0691			
R.KVFDTGLNKK.V +1			2.3506			
R.VGDAQVNVVSDGISTFPLSDGFVLNVMKDEVGEALEAAFLPK.D +3			6.496	5.1271		

B

```
orf .      Peptide Criteria   Coverage Criteria   PC1    PC2    PC3    PC4    Pep1   Pep2   Total
or4354     Peptides 1         -no coverage-       41.4   20.1   7.1    11.4   10     2      10
or6102     Peptides 2         -no coverage-       7.8    3.8    19.2   32.7   3      13     13
or6159     Peptides 1         -no coverage-       21.1   9.7    3.8    9.4    10     3      10
or6351     Peptides 1         -no coverage-       54.4   41.9   20     16.3   21     7      21
or7174     Peptides 2         Coverage 2          23.7   17.6   59.8   48.6   13     34     34
or2879     Peptides 2         Coverage 2          0      6.1    52.9   40.1   1      9      9
or3995     -no peptide-       Coverage 2          0      21.9   28.4   57.9   1      4      4
or4331     -no peptide-       Coverage 2          0      6.8    49     43.2   1      6      6
or7052     -no peptide-       Coverage 2          0      11.8   40.3   40.3   1      5      5
or0068     -no peptide-       Coverage 2          10.6   0      56.4   56.4   1      2      2
or7136     -no peptide-       Coverage 2          18.9   0      53.7   53.7   1      2      2
or1159     -no peptide-       Coverage 1          44.6   58.5   0      27.7   3      1      3
or3644     Peptides 1         Coverage 1          36.3   35.6   0      8.6    14     2      14
or0876     -no peptide-       Coverage 1          42     42     13     0      3      1      3
or2605     -no peptide-       Coverage 1          45.5   45.5   13.6   0      5      1      5
or5400     -no peptide-       Coverage 1          57     30.2   17.4   0      3      1      3
or7002     Peptides 1         -no coverage-       31     3.7    2.1    0      9      1      9
or7003     Peptides 1         -no coverage-       40.4   17.6   11.6   0      12     4      12
or0490     Peptides 2         -no coverage-       0      0      24.2   14     0      8      8
or1361     -no peptide-       Coverage 2          0      0      36.7   36.7   0      2      2
or2444     Peptides 2         -no coverage-       0      0      16.8   20.5   0      7      7
or2852     Peptides 2         -no coverage-       0      0      10.1   12.7   0      7      7
```

Figure 5: Input and output of the "Gene Caller" program.
A; shows the HTML formatted representation of the contrast file used to identify differences between cells grown in differing environmental conditions. The gene's orf number is visible in the top left corner in the format rpal_or2879, indicating *R. palustris* orf 2979. Below the orf number is a list of all sequenced peptides found that are associated with that gene. To the right of the orf number are the four percent sequence coverage obtained for each run for the sample indicated in the column headers above, as well as the total percent coverage for all samples combined. For each peptide, the best xcorr value is given below that, if the peptide was observed in that run. Where no xcorr value is given for a peptide and run, the peptide was not observed. B; shows the output of the "Gene Caller" program, with columns indicating the gene name, whether the list of peptides is biased in favor of one environmental condition, whether the percent coverage was biased in favor of one environmental condition, the percent sequence coverage for each of the four samples, the number of peptides found for each environmental condition and the total number of peptides observed associated with that gene. For information on the thresholds used to determine chosen genes, see section 2.7.1

Fortunately, there exist a plethora of tools to assist in these tasks. Depending on the platform and type of database in use, these tools may be as simple as using export functions built into spreadsheet software packages. Microsoft Excel contains a great number of both import and export functions that can convert data from most common formats. As well, because of the ever increasing integration of web interfaces and desktop tools, these same software packages are now able to perform cut and paste operations from web tables into spreadsheets, which can greatly simplify the data gathering.

Although represented as a single box in Figure 4, "other sources" of data includes such diverse sources as the web interface for the P-SORTB algorithm, 15 tables of pre-grouped genes according to proposed functions from the *R. palustris* web pages, spreadsheets of COG assignments, manually generated lists of transmembrane segments and miscellaneous bioinformatics assignments of superfamilies, and gene types. Fortunately, in all of these data sources, each item of information is given a single gene with which it corresponds. This allows each item to be accepted as an individual row of data into an SQL database. Although each item must correspond with a single gene, the converse is not necessarily true. In many instances, information sources may provide more than one piece of information for a single gene. This is often the case for COGs, where a gene may belong to more than one family and two or more rows of information are provided. Thus, the gene assignment is frequently not suitable as a unique key field; however, it is still the preferred means of joining tables and for that reason, should be indexed.

When sources of data generate unique rows for each gene or protein assignment, such as P-SORTB (Gardy et al. 2003), it may significantly improve SQL database performance to use the gene assignment as the key field as the database increases in size, regardless of the inconsistency of its use as a key field in all tables. Such sources are often the preferred type of data for larger queries, where fewer filters/criteria are required to generate a single column in a final generated report. P-SORTB is again a good example, as a putative cellular localization prediction can be added to any query with a single join and by including a single column from the joined table.

Once the tables are created, SQL queries can be performed on an ad-hoc basis to compile lists that reference numerous sources. The flexibility of the SQL database is in the ability to modify and alter these queries as needed, quickly. Microsoft Access was used here to take advantage of a graphical query designer that simplifies the process. In addition, it also provides a number of export options, and so data generated at this step may be obtained and manipulated without being obliged to execute the same query to generate the information over again. The data provided by queries are returned in a format similar to that of a spreadsheet, and can be exported to a variety of formats. By utilizing this function, it was possible to create a single spreadsheet containing the information that was most useful, which could be further processed as a spreadsheet. An alternative approach would have been to use the query information to create a new table, upon which further queries could have been conducted, and from which the SQL database could have been used to identify further targets. Whereas the method used here allows for the easier creation of portable Excel spreadsheets, the alternative approach would provide greater portability from which a web application could be created.

4.4. Analysis of Data – Grouping and Evaluation

The SQL database was used to store large amounts of dissimilar information, which could be used to generate a coherent picture concerning the MS information and any additional information that could be obtained relevant to individual genes. However, not all sources of information proved to be useful in creating a coherent picture. Many different variations of sort orders and groupings were attempted before any significant patterns could be visualized.

Many attempts involved grouping the genes by predicted function. For many *R. palustris* genes, a significant amount of annotation has been done; allowing groupings to be created based on key words, or based on known pathways. However, for any grouping created in this manner, only a fraction of the total genes, those for which a function is predicted, can be included. This eliminates the possibility of performing predictions upon genes of unknown function. In fact, this fundamental problem underlies the use of any predictive grouping as the basis of an organizational system.

While most of the predictive and grouping methods tested, such as COGs, Superfamilies and Kyoto Encyclopedia of Genes and Genomes (KEGGs) groupings, or any other homology based means are unable to be used as a primary means for arranging data, they provide an excellent secondary source of information. Frequently, they can support or raise valid objections to annotations or putative functions, giving a better understanding of the gene and its neighbors. For that reason, many of these secondary sources of information are included in the final colourized representation of the proteome, but have little bearing upon the processing or grouping of the data.

In contrast, the most successful means by which to group the MS data, in terms of ease of use, has been by the order of the genes themselves. This can be done either by traveling along one strand, and then back along the other, or simply by proceeding around the chromosome. In either case, as long as the transcriptional orientation of the genes is known, it is a simple matter to include that information in the queries used, and thus include those data in the final, visualized proteome files.

Because many bacterial genes are naturally grouped into operons or transcription units, displaying the genes based on their position on the chromosome provides much insight into the function and usage of proteins that are present, as well as those that do not appear in the proteome. In this manner, even those proteins which are not found in the MS data are able to contribute to the understanding of the proteome simply by providing information about their location and by inference from the data obtained from their neighbors, where operons are likely to exist. This can be particularly helpful for membrane proteins that form complexes, where only some of the units of the complex are found. For example, the F_1F_0 complex a and c proteins were not detected, while the b' and b proteins that appear to be expressed from the same transcriptional unit were observed (see Table 5 - subunit b (RPA0843) is the last gene of the transcriptional unit). Of course, this type of information still requires confirmation and may be an artifact, and must be interpreted with care.

At the completion of these operations, once the secondary information is included, and the data are sorted appropriately, the information generated can be treated in a number of manners. In the examples given here, the complete dataset used for analysis was exported to Microsoft Excel, a spreadsheet program, where the data could be further processed. Similar operations could have been done using extensions to the SQL database, but would not have been portable outside of the database.

4.5. Analysis of Data

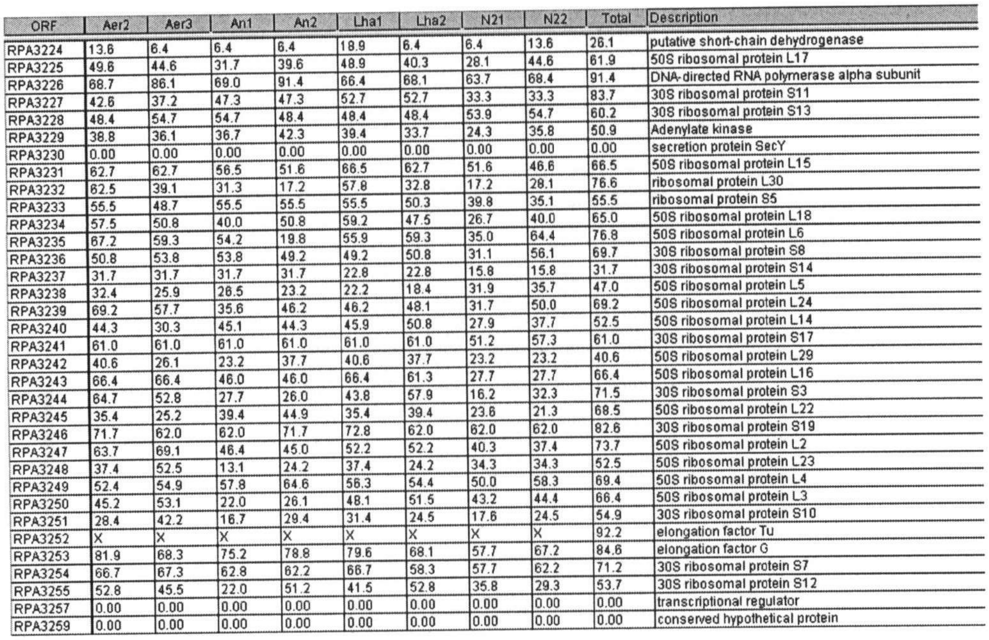
Although it is not difficult to visually inspect the data listed in a tabular format (Figure 6A and Figure 7A), it is difficult to inspect each row individually to identify trends that may be present. Furthermore, it is not a simple matter to utilize the numerical data to see trends between rows, and even more difficult to perceive the boundaries between potential operons without constant scrolling.

In order to utilize the data present in the spreadsheet and create a single readable source of data in as small space as possible, a graphical interpretation of the data was generated. By using a modified colour scheme (Table 7), similar to that used in micro-array data, it was possible to provide a colour interpretation of the data, facilitating visual inspection of the numerical data. By using the colour as a background and ensuring that each colour used would not block the numerical data, this provides both a general overview in the form of the colour interface as well as the raw data in numerical format. Thus, no information is lost, preserving the integrity of the data.

Table 7: Colour scheme for percent coverage

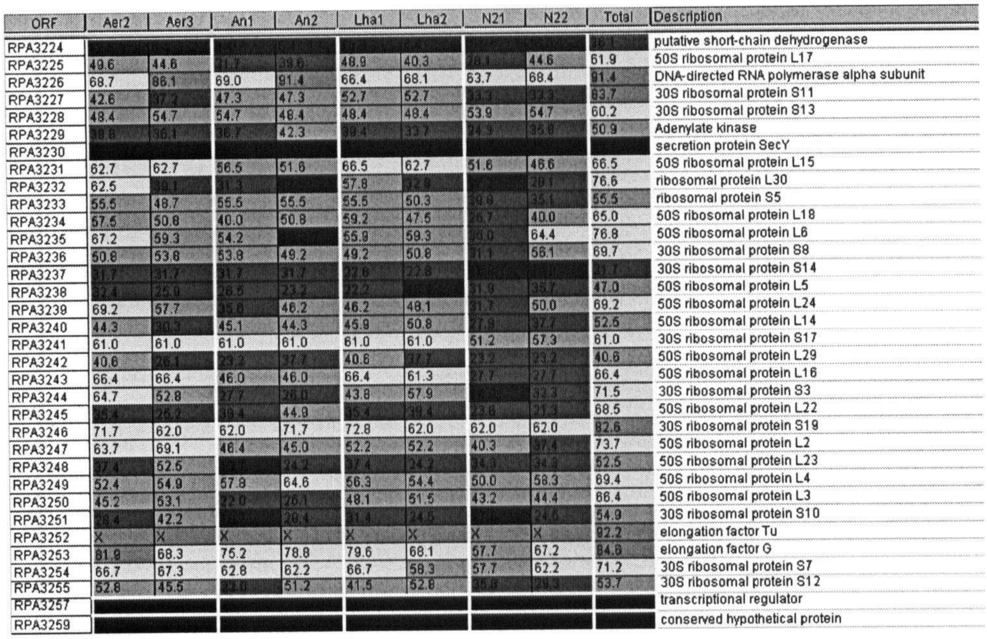
Colour	Percent Coverage
Black	0%
Dark Red	0.1%-19.9%
Red	20.0%-39.9%
Orange	40.0%-59.9%
Yellow	60.0%-79.9%
Green	80.0%-100.0%

A



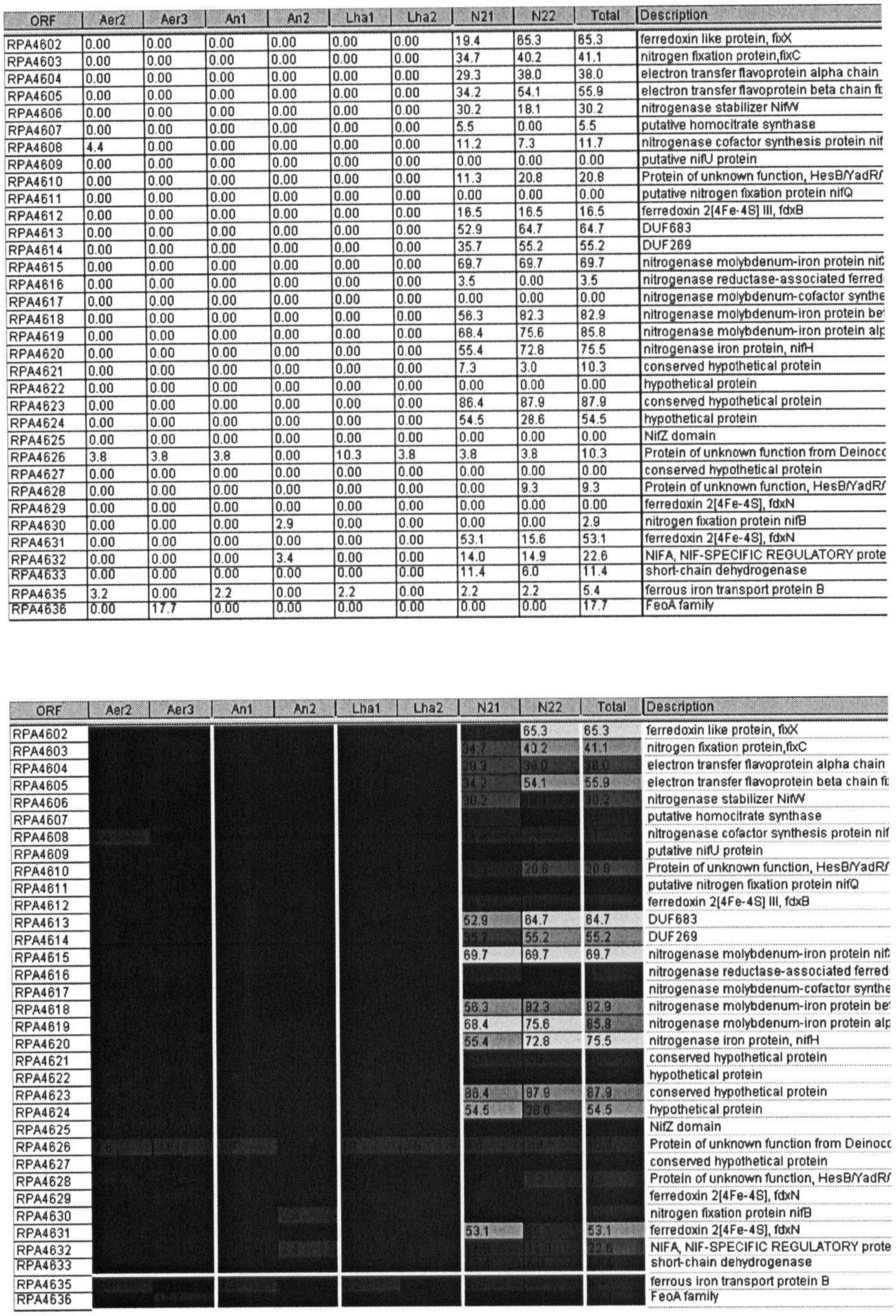
ORF	Aer2	Aer3	An1	An2	Lha1	Lha2	N21	N22	Total	Description
RPA3224	13.6	6.4	6.4	6.4	18.9	6.4	6.4	13.6	26.1	putative short-chain dehydrogenase
RPA3225	49.6	44.6	31.7	39.6	48.9	40.3	28.1	44.6	61.9	50S ribosomal protein L17
RPA3226	68.7	86.1	69.0	91.4	66.4	68.1	63.7	68.4	91.4	DNA-directed RNA polymerase alpha subunit
RPA3227	42.6	37.2	47.3	47.3	52.7	52.7	33.3	33.3	83.7	30S ribosomal protein S11
RPA3228	48.4	54.7	54.7	48.4	48.4	48.4	53.9	54.7	60.2	30S ribosomal protein S13
RPA3229	38.8	36.1	36.7	42.3	39.4	33.7	24.3	35.8	50.9	Adenylate kinase
RPA3230	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	secretion protein SecY
RPA3231	62.7	62.7	56.5	51.6	66.5	62.7	51.6	46.6	66.5	50S ribosomal protein L15
RPA3232	62.5	39.1	31.3	17.2	57.8	32.8	17.2	28.1	76.6	ribosomal protein L30
RPA3233	55.5	48.7	55.5	55.5	55.5	50.3	39.8	35.1	55.5	ribosomal protein S5
RPA3234	57.5	50.8	40.0	50.8	59.2	47.5	26.7	40.0	65.0	50S ribosomal protein L18
RPA3235	67.2	59.3	54.2	19.8	55.9	59.3	35.0	64.4	76.8	50S ribosomal protein L6
RPA3236	50.8	53.8	53.8	49.2	49.2	50.8	31.1	56.1	69.7	30S ribosomal protein S8
RPA3237	31.7	31.7	31.7	31.7	22.8	22.8	15.8	15.8	31.7	30S ribosomal protein S14
RPA3238	32.4	25.9	26.5	23.2	22.2	18.4	31.9	35.7	47.0	50S ribosomal protein L5
RPA3239	69.2	57.7	35.6	46.2	46.2	48.1	31.7	50.0	69.2	50S ribosomal protein L24
RPA3240	44.3	30.3	45.1	44.3	45.9	50.8	27.9	37.7	52.5	50S ribosomal protein L14
RPA3241	61.0	61.0	61.0	61.0	61.0	61.0	51.2	57.3	61.0	30S ribosomal protein S17
RPA3242	40.6	26.1	23.2	37.7	40.6	37.7	23.2	23.2	40.6	50S ribosomal protein L29
RPA3243	66.4	66.4	46.0	46.0	66.4	61.3	27.7	27.7	66.4	50S ribosomal protein L16
RPA3244	64.7	52.8	27.7	26.0	43.8	57.9	16.2	32.3	71.5	30S ribosomal protein S3
RPA3245	35.4	25.2	39.4	44.9	35.4	39.4	23.6	21.3	68.5	50S ribosomal protein L22
RPA3246	71.7	62.0	62.0	71.7	72.8	62.0	62.0	62.0	82.6	30S ribosomal protein S19
RPA3247	63.7	69.1	46.4	45.0	52.2	52.2	40.3	37.4	73.7	50S ribosomal protein L2
RPA3248	37.4	52.5	13.1	24.2	37.4	24.2	34.3	34.3	52.5	50S ribosomal protein L23
RPA3249	52.4	54.9	57.8	64.6	56.3	54.4	50.0	58.3	69.4	50S ribosomal protein L4
RPA3250	45.2	53.1	22.0	26.1	48.1	51.5	43.2	44.4	66.4	50S ribosomal protein L3
RPA3251	28.4	42.2	16.7	29.4	31.4	24.5	17.6	24.5	54.9	30S ribosomal protein S10
RPA3252	X	X	X	X	X	X	X	X	92.2	elongation factor Tu
RPA3253	81.9	68.3	75.2	78.8	79.6	68.1	57.7	67.2	84.6	elongation factor G
RPA3254	66.7	67.3	62.8	62.2	66.7	58.3	57.7	62.2	71.2	30S ribosomal protein S7
RPA3255	52.8	45.5	22.0	51.2	41.5	52.8	35.8	29.3	53.7	30S ribosomal protein S12
RPA3257	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	transcriptional regulator
RPA3259	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	conserved hypothetical protein

B



ORF	Aer2	Aer3	An1	An2	Lha1	Lha2	N21	N22	Total	Description
RPA3224	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	26.1	putative short-chain dehydrogenase
RPA3225	49.6	44.6	31.7	39.6	48.9	40.3	28.1	44.6	61.9	50S ribosomal protein L17
RPA3226	68.7	86.1	69.0	91.4	66.4	68.1	63.7	68.4	91.4	DNA-directed RNA polymerase alpha subunit
RPA3227	42.6	37.2	47.3	47.3	52.7	52.7	33.3	33.3	83.7	30S ribosomal protein S11
RPA3228	48.4	54.7	54.7	48.4	48.4	48.4	53.9	54.7	60.2	30S ribosomal protein S13
RPA3229	38.8	36.1	36.7	42.3	39.4	33.7	24.3	35.8	50.9	Adenylate kinase
RPA3230										secretion protein SecY
RPA3231	62.7	62.7	56.5	51.6	66.5	62.7	51.6	46.6	66.5	50S ribosomal protein L15
RPA3232	62.5	39.1	31.3	[illegible]	57.8	32.8	[illegible]	28.1	76.6	ribosomal protein L30
RPA3233	55.5	48.7	55.5	55.5	55.5	50.3	39.8	35.1	55.5	ribosomal protein S5
RPA3234	57.5	50.8	40.0	50.8	59.2	47.5	26.7	40.0	65.0	50S ribosomal protein L18
RPA3235	67.2	59.3	54.2	[illegible]	55.9	59.3	35.0	64.4	76.8	50S ribosomal protein L6
RPA3236	50.8	53.8	53.8	49.2	49.2	50.8	31.1	56.1	69.7	30S ribosomal protein S8
RPA3237	31.7	31.7	31.7	31.7	22.8	22.8	[illegible]	[illegible]	31.7	30S ribosomal protein S14
RPA3238	32.4	25.9	26.5	23.2	22.2	[illegible]	31.9	35.7	47.0	50S ribosomal protein L5
RPA3239	69.2	57.7	35.6	46.2	46.2	48.1	31.7	50.0	69.2	50S ribosomal protein L24
RPA3240	44.3	30.3	45.1	44.3	45.9	50.8	27.9	37.7	52.5	50S ribosomal protein L14
RPA3241	61.0	61.0	61.0	61.0	61.0	61.0	51.2	57.3	61.0	30S ribosomal protein S17
RPA3242	40.6	26.1	23.2	37.7	40.6	37.7	23.2	23.2	40.6	50S ribosomal protein L29
RPA3243	66.4	66.4	46.0	46.0	66.4	61.3	27.7	27.7	66.4	50S ribosomal protein L16
RPA3244	64.7	52.8	27.7	26.0	43.8	57.9	[illegible]	32.3	71.5	30S ribosomal protein S3
RPA3245	35.4	25.2	39.4	44.9	35.4	39.4	23.6	21.3	68.5	50S ribosomal protein L22
RPA3246	71.7	62.0	62.0	71.7	72.8	62.0	62.0	62.0	82.6	30S ribosomal protein S19
RPA3247	63.7	69.1	46.4	45.0	52.2	52.2	40.3	37.4	73.7	50S ribosomal protein L2
RPA3248	37.4	52.5	[illegible]	24.2	37.4	24.2	34.3	34.3	52.5	50S ribosomal protein L23
RPA3249	52.4	54.9	57.8	64.6	56.3	54.4	50.0	58.3	69.4	50S ribosomal protein L4
RPA3250	45.2	53.1	22.0	26.1	48.1	51.5	43.2	44.4	66.4	50S ribosomal protein L3
RPA3251	28.4	42.2	[illegible]	29.4	31.4	24.5	[illegible]	24.5	54.9	30S ribosomal protein S10
RPA3252	X	X	X	X	X	X	X	X	92.2	elongation factor Tu
RPA3253	81.9	68.3	75.2	78.8	79.6	68.1	57.7	67.2	84.6	elongation factor G
RPA3254	66.7	67.3	62.8	62.2	66.7	58.3	57.7	62.2	71.2	30S ribosomal protein S7
RPA3255	52.8	45.5	22.0	51.2	41.5	52.8	35.8	29.3	53.7	30S ribosomal protein S12
RPA3257										transcriptional regulator
RPA3259										conserved hypothetical protein

Figure 6: Colourization of data - ribosomal proteins.
Demonstration of the colourization of data, performed on a cluster of ribosomal proteins, which are normally abundant in the proteome. The same dataset is shown without colour in (A), and after being coloured in (B). Vertical bars divide the datasets performed under different environmental conditions, and horizontal bars indicated where there are genes on the opposite strand. The colour scheme from Table 7 is used here. Column are used as follows: Orf indicates the gene annotation, Aer2 and Aer3 are two sets of data from the same sample of aerobically grown cells (section 2.5.2.), An1 and An2 are two sets of data from the same sample of anaerobically grown cells (section 2.5.1.), Lha1 and Lha2 are two sets of data from the same sample of LhaA mutant cells (section 2.5.3.), N21 and N22 are two sets of data from the same sample of cells grown under nitrogen fixation conditions (section 2.5.4.), Total displays the total percent sequence coverage for the protein indicated over all sets of data shown in the table, and Description gives a brief annotation for the protein indicated.



A

ORF	Aer2	Aer3	An1	An2	Lha1	Lha2	N21	N22	Total	Description
RPA4602	0.00	0.00	0.00	0.00	0.00	0.00	19.4	65.3	65.3	ferredoxin like protein, fixX
RPA4603	0.00	0.00	0.00	0.00	0.00	0.00	34.7	40.2	41.1	nitrogen fixation protein,fixC
RPA4604	0.00	0.00	0.00	0.00	0.00	0.00	29.3	38.0	38.0	electron transfer flavoprotein alpha chain
RPA4605	0.00	0.00	0.00	0.00	0.00	0.00	34.2	54.1	55.9	electron transfer flavoprotein beta chain fi
RPA4606	0.00	0.00	0.00	0.00	0.00	0.00	30.2	18.1	30.2	nitrogenase stabilizer NifW
RPA4607	0.00	0.00	0.00	0.00	0.00	0.00	5.5	0.00	5.5	putative homocitrate synthase
RPA4608	4.4	0.00	0.00	0.00	0.00	0.00	11.2	7.3	11.7	nitrogenase cofactor synthesis protein nif
RPA4609	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	putative nifU protein
RPA4610	0.00	0.00	0.00	0.00	0.00	0.00	11.3	20.8	20.8	Protein of unknown function, HesB/YadR/
RPA4611	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	putative nitrogen fixation protein nifQ
RPA4612	0.00	0.00	0.00	0.00	0.00	0.00	16.5	16.5	16.5	ferredoxin 2[4Fe-4S] III, fdxB
RPA4613	0.00	0.00	0.00	0.00	0.00	0.00	52.9	64.7	64.7	DUF683
RPA4614	0.00	0.00	0.00	0.00	0.00	0.00	35.7	55.2	55.2	DUF269
RPA4615	0.00	0.00	0.00	0.00	0.00	0.00	69.7	69.7	69.7	nitrogenase molybdenum-iron protein nif
RPA4616	0.00	0.00	0.00	0.00	0.00	0.00	3.5	0.00	3.5	nitrogenase reductase-associated ferred
RPA4617	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	nitrogenase molybdenum-cofactor synthe
RPA4618	0.00	0.00	0.00	0.00	0.00	0.00	56.3	82.3	82.9	nitrogenase molybdenum-iron protein be
RPA4619	0.00	0.00	0.00	0.00	0.00	0.00	68.4	75.6	85.8	nitrogenase molybdenum-iron protein alp
RPA4620	0.00	0.00	0.00	0.00	0.00	0.00	55.4	72.8	75.5	nitrogenase iron protein, nifH
RPA4621	0.00	0.00	0.00	0.00	0.00	0.00	7.3	3.0	10.3	conserved hypothetical protein
RPA4622	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	hypothetical protein
RPA4623	0.00	0.00	0.00	0.00	0.00	0.00	86.4	87.9	87.9	conserved hypothetical protein
RPA4624	0.00	0.00	0.00	0.00	0.00	0.00	54.5	28.6	54.5	hypothetical protein
RPA4625	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	NifZ domain
RPA4626	3.8	3.8	3.8	0.00	10.3	3.8	3.8	3.8	10.3	Protein of unknown function from Deinocc
RPA4627	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	conserved hypothetical protein
RPA4628	0.00	0.00	0.00	0.00	0.00	0.00	0.00	9.3	9.3	Protein of unknown function, HesB/YadR/
RPA4629	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	ferredoxin 2[4Fe-4S], fdxN
RPA4630	0.00	0.00	0.00	2.9	0.00	0.00	0.00	0.00	2.9	nitrogen fixation protein nifB
RPA4631	0.00	0.00	0.00	0.00	0.00	0.00	53.1	15.6	53.1	ferredoxin 2[4Fe-4S], fdxN
RPA4632	0.00	0.00	0.00	3.4	0.00	0.00	14.0	14.9	22.6	NIFA, NIF-SPECIFIC REGULATORY prote
RPA4633	0.00	0.00	0.00	0.00	0.00	0.00	11.4	6.0	11.4	short-chain dehydrogenase
RPA4635	3.2	0.00	2.2	0.00	2.2	0.00	2.2	2.2	5.4	ferrous iron transport protein B
RPA4636	0.00	17.7	0.00	0.00	0.00	0.00	0.00	0.00	17.7	FeoA family

B

ORF	Aer2	Aer3	An1	An2	Lha1	Lha2	N21	N22	Total	Description
RPA4602								65.3	65.3	ferredoxin like protein, fixX
RPA4603							34.7	40.2	41.1	nitrogen fixation protein,fixC
RPA4604							29.3	38.0	38.0	electron transfer flavoprotein alpha chain
RPA4605							34.2	54.1	55.9	electron transfer flavoprotein beta chain fi
RPA4606							30.2		30.2	nitrogenase stabilizer NifW
RPA4607										putative homocitrate synthase
RPA4608										nitrogenase cofactor synthesis protein nif
RPA4609										putative nifU protein
RPA4610								20.8	20.8	Protein of unknown function, HesB/YadR/
RPA4611										putative nitrogen fixation protein nifQ
RPA4612										ferredoxin 2[4Fe-4S] III, fdxB
RPA4613							52.9	64.7	64.7	DUF683
RPA4614							35.7	55.2	55.2	DUF269
RPA4615							69.7	69.7	69.7	nitrogenase molybdenum-iron protein nif
RPA4616										nitrogenase reductase-associated ferred
RPA4617										nitrogenase molybdenum-cofactor synthe
RPA4618							56.3	82.3	82.9	nitrogenase molybdenum-iron protein be
RPA4619							68.4	75.6	85.8	nitrogenase molybdenum-iron protein alp
RPA4620							55.4	72.8	75.5	nitrogenase iron protein, nifH
RPA4621										conserved hypothetical protein
RPA4622										hypothetical protein
RPA4623							86.4	87.9	87.9	conserved hypothetical protein
RPA4624							54.5	28.6	54.5	hypothetical protein
RPA4625										NifZ domain
RPA4626										Protein of unknown function from Deinocc
RPA4627										conserved hypothetical protein
RPA4628										Protein of unknown function, HesB/YadR/
RPA4629										ferredoxin 2[4Fe-4S], fdxN
RPA4630										nitrogen fixation protein nifB
RPA4631							53.1		53.1	ferredoxin 2[4Fe-4S], fdxN
RPA4632									22.6	NIFA, NIF-SPECIFIC REGULATORY prote
RPA4633										short-chain dehydrogenase
RPA4635										ferrous iron transport protein B
RPA4636										FeoA family

Figure 7: Colourization of data - nitrogen fixation genes.
Demonstration of the colourization of data, performed on a cluster of Nitrogen fixation genes, which are normally only present under nitrogen fixation conditions. The same dataset is shown without colour in (A), and after being coloured in (B). Vertical bars divide the datasets performed under different environmental conditions. Column headings given in Figure 6. The colour scheme from Table 7 is used here.

As well, because the proteins present are often expressed from genes present in transcription units, it was useful to use a simple dividing line between proteins where the genes adjacent on the chromosome are on the opposite strand. This allows for putative transcriptional units to be visualized, which can be tentatively confirmed by looking for clues in the annotation of the genome.

The coloured versions of the Figures 6A and 7A can be seen in Figures 6B and 7B respectively. In Figure 6B, a thicker white line separates the genes RPA3255 and RPA3257, indicating that gene RPA3256 is encoded by the opposite strand, although located between these two genes on the chromosome. For the nitrogen fixation gene cluster (Figure 7B), the genes which encode all of the proteins shown are sequentially arranged on the same strand of the chromosome, thus the figure represents what could be a single transcription unit. However, because the protein expression appears to change significantly with gene RPA4626, it is unlikely that all of these proteins are from a single mRNA transcript. While it is possible that the regulation of the protein occurs post-translationally, it is less likely to be the case when the pattern of the presence of the protein differs significantly from those in the same putative transcriptional unit.

Despite the improvement that the colourization of the data creates, it is not sufficient for locating subtle trends among the protein expression profiles. Unlike array data, where the genes investigated are up- and down-regulated with respect to some baseline condition, the mass spectrometry data give a simple percent coverage. While the percent coverage data may give an indication of the level of expression of the gene, it is not explicitly stated. Thus, determining a baseline expression level for each gene provides a reference point by which the up-regulation of each gene can be assessed.

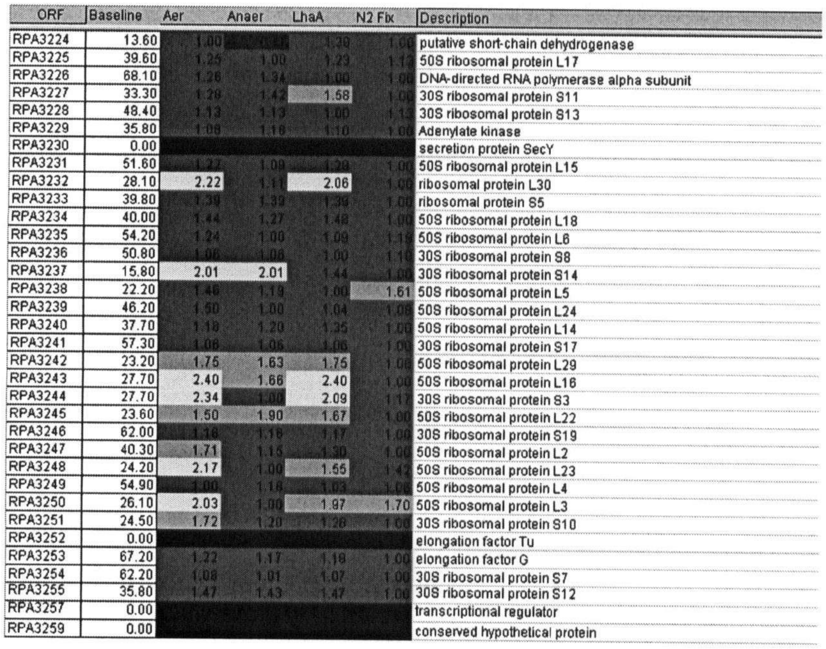
For the purposes here, where only two replicate mass spectrometry experiments were performed on each environmental condition, it is a challenging task to determine an accurate value for each condition. Any method of obtaining a baseline value for such a limited dataset caries significant element of uncertainty, and the methods used here should be given a critical review before implementation on a more robust dataset. In this case, a naïve approach was utilized, wherein the highest value for each protein under each condition was accepted as the true value. This allowed for an overall simplification of the dataset. These values were then divided by the baseline to generate a reduced and normalized version of the full dataset. Again, as with the raw percent coverage data, the normalized data were colourized using a similar colour scheme, to render the data more accessible (Table 8).

Table 8: Colour scheme for normalized data

Colour	Ratio Relative to Baseline	Interpreted as:
Black	0.00	not found
Dark Red	0.01-0.69	down regulated
Red	0.70-1.49	similar to baseline
Orange	1.50-1.99	up regulated
Yellow	2.00-4.99	up regulated
Green	>5.00	up regulated

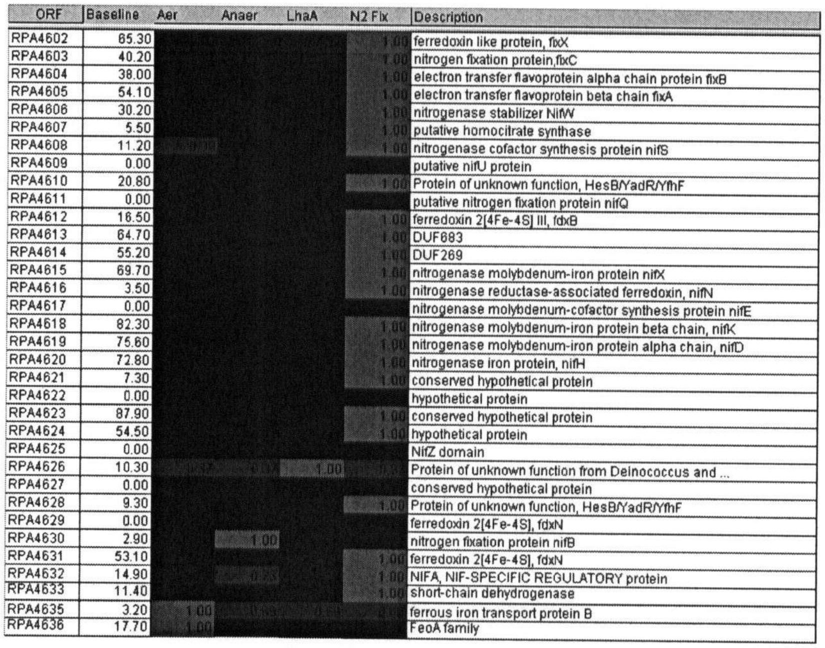
Two examples of the baseline data are shown, using ribosomal proteins (Figure 8A) and a cluster of nitrogen fixation genes (Figure 8B). For the ribosomal cluster, the baseline data clearly indicate that the majority of these genes are expressed in a constitutive manner and that under aerobic conditions, some of the ribosomal proteins

A



ORF	Baseline	Aer	Anaer	LhaA	N2 Fix	Description
RPA3224	13.60	1.00	[illegible]	1.30	1.00	putative short-chain dehydrogenase
RPA3225	39.60	1.25	1.00	1.23	1.12	50S ribosomal protein L17
RPA3226	68.10	1.26	1.34	1.00	1.00	DNA-directed RNA polymerase alpha subunit
RPA3227	33.30	1.29	1.42	1.58	1.00	30S ribosomal protein S11
RPA3228	48.40	1.13	1.13	1.00	1.13	30S ribosomal protein S13
RPA3229	35.80	1.09	1.16	1.10	1.00	Adenylate kinase
RPA3230	0.00					secretion protein SecY
RPA3231	51.60	1.27	1.09	1.29	1.00	50S ribosomal protein L15
RPA3232	28.10	2.22	1.11	2.06	1.00	ribosomal protein L30
RPA3233	39.80	1.39	1.39	1.39	1.00	ribosomal protein S5
RPA3234	40.00	1.44	1.27	1.48	1.00	50S ribosomal protein L18
RPA3235	54.20	1.24	1.00	1.09	1.16	50S ribosomal protein L6
RPA3236	50.80	1.06	1.06	1.00	1.10	30S ribosomal protein S8
RPA3237	15.80	2.01	2.01	1.44	1.00	30S ribosomal protein S14
RPA3238	22.20	1.46	1.19	1.00	1.61	50S ribosomal protein L5
RPA3239	46.20	1.50	1.00	1.04	1.08	50S ribosomal protein L24
RPA3240	37.70	1.18	1.20	1.35	1.00	50S ribosomal protein L14
RPA3241	57.30	1.08	1.06	1.06	1.00	30S ribosomal protein S17
RPA3242	23.20	1.75	1.63	1.75	1.00	50S ribosomal protein L29
RPA3243	27.70	2.40	1.66	2.40	1.00	50S ribosomal protein L16
RPA3244	27.70	2.34	1.00	2.09	1.17	30S ribosomal protein S3
RPA3245	23.60	1.50	1.90	1.67	1.00	50S ribosomal protein L22
RPA3246	62.00	1.18	1.18	1.17	1.00	30S ribosomal protein S19
RPA3247	40.30	1.71	1.15	1.30	1.00	50S ribosomal protein L2
RPA3248	24.20	2.17	1.00	1.55	1.42	50S ribosomal protein L23
RPA3249	54.90	1.00	1.16	1.03	1.06	50S ribosomal protein L4
RPA3250	26.10	2.03	1.00	1.97	1.70	50S ribosomal protein L3
RPA3251	24.50	1.72	1.20	1.28	1.00	30S ribosomal protein S10
RPA3252	0.00					elongation factor Tu
RPA3253	67.20	1.22	1.17	1.19	1.00	elongation factor G
RPA3254	62.20	1.08	1.01	1.07	1.00	30S ribosomal protein S7
RPA3255	35.80	1.47	1.43	1.47	1.00	30S ribosomal protein S12
RPA3257	0.00					transcriptional regulator
RPA3259	0.00					conserved hypothetical protein

B



ORF	Baseline	Aer	Anaer	LhaA	N2 Fix	Description
RPA4602	65.30				1.00	ferredoxin like protein, fixX
RPA4603	40.20				1.00	nitrogen fixation protein,fixC
RPA4604	38.00				1.00	electron transfer flavoprotein alpha chain protein fixB
RPA4605	54.10				1.00	electron transfer flavoprotein beta chain fixA
RPA4606	30.20				1.00	nitrogenase stabilizer NifW
RPA4607	5.50				1.00	putative homocitrate synthase
RPA4608	11.20	[illegible]			1.00	nitrogenase cofactor synthesis protein nifS
RPA4609	0.00					putative nifU protein
RPA4610	20.80				1.00	Protein of unknown function, HesB/YadR/YfhF
RPA4611	0.00					putative nitrogen fixation protein nifQ
RPA4612	16.50				1.00	ferredoxin 2[4Fe-4S] III, fdxB
RPA4613	64.70				1.00	DUF683
RPA4614	55.20				1.00	DUF269
RPA4615	69.70				1.00	nitrogenase molybdenum-iron protein nifX
RPA4616	3.50				1.00	nitrogenase reductase-associated ferredoxin, nifN
RPA4617	0.00					nitrogenase molybdenum-cofactor synthesis protein nifE
RPA4618	82.30				1.00	nitrogenase molybdenum-iron protein beta chain, nifK
RPA4619	75.60				1.00	nitrogenase molybdenum-iron protein alpha chain, nifD
RPA4620	72.80				1.00	nitrogenase iron protein, nifH
RPA4621	7.30				1.00	conserved hypothetical protein
RPA4622	0.00					hypothetical protein
RPA4623	87.90				1.00	conserved hypothetical protein
RPA4624	54.50				1.00	hypothetical protein
RPA4625	0.00					NifZ domain
RPA4626	10.30	[illegible]	[illegible]	1.00	[illegible]	Protein of unknown function from Deinococcus and ...
RPA4627	0.00					conserved hypothetical protein
RPA4628	9.30				1.00	Protein of unknown function, HesB/YadR/YfhF
RPA4629	0.00					ferredoxin 2[4Fe-4S], fdxN
RPA4630	2.90		1.00			nitrogen fixation protein nifB
RPA4631	53.10				1.00	ferredoxin 2[4Fe-4S], fdxN
RPA4632	14.90		[illegible]		1.00	NIFA, NIF-SPECIFIC REGULATORY protein
RPA4633	11.40				1.00	short-chain dehydrogenase
RPA4635	3.20	1.00	[illegible]	[illegible]	[illegible]	ferrous iron transport protein B
RPA4636	17.70	1.00				FeoA family

Figure 8: Baseline representation of ribosomal and nitrogen fixation proteins. Colourized representation of normalized data for the genes shown in Figures 6 and 7. Column headers are as follows, ORF indicates the gene which gives rise to the protein, Baseline indicates the percent sequence coverage shown as 1.00 in the normalized data, Description gives the annotation associated with the protein and Aer, Anaer, LhaA and N2 Fix indicate the cells used in collecting the data, respectively aerobically (section 2.5.2), anaerobically (section 2.5.1), LhaA mutant (section 2.5.3) and Nitrogen fixation conditions (section 2.5.4) (A) Demonstrates the constitutive nature of the ribosomal proteins. *R. palustris* wild type grown under aerobic and the LhaA mutant (both grown aerobically) shows a slightly increased protein expression, in comparison to the samples grown under anaerobic and N_2 fixation conditions. (B) The proteins encoded by genes known to be involved in nitrogen fixation are consistently present under nitrogen fixation conditions but generally not present under any other condition. This becomes clear when the data is normalized, as opposed to percent sequence coverage data (Figure 6 and Figure 7).

become up-regulated. In contrast, the nitrogen fixation cluster shows a set of genes which are expressed only under nitrogen fixation conditions.

It is worth mentioning the LhaA data, seen in the third set of data (columns 6 and 7) in Figure 7 and Figure 8. As discussed in section 1.4.2, the *lhaA* gene is thought to be a photosynthetic-specific gene, and so the *lhaA* mutation would not be expected to make a significant difference in the proteomic profile of the aerobically grown *R. palustris* cells used in this mass spectrometry experiment. Instead, its profile would be expected to match the profile seen in the aerobic grown wild-type cells, in the second and third columns of Figures 6 and 7. Indeed, this is seen in both examples, as well as throughout the majority of the complete dataset.

4.6. Predictions

Genes were grouped by expression profiles using relatively conservative criteria (Table 1) to which a simplified colour scheme was applied (Table 9). Both Boolean (true or false) or analog methods, in which the distances are recorded, can be used to categorize data. While the Boolean method provides the greatest simplicity and can answer such basic questions as whether a given protein is observed or not, it fails to yield more interesting answers. A more useful question to ask of this dataset is whether a given protein is more highly expressed under a given environmental condition, and by how much. However, the answer to that question comes in two parts - a Boolean answer to whether it is more highly expressed, and then the analog part - the difference between the two scalars. Quantifying the difference is a simple matter, but determining a threshold over which the answer is useful is not intuitively obvious. In this case, an arbitrarily chosen distance of 0.4 was used, which appeared to give satisfactory results. (Section 4.7)

Table 9: Colour scheme for gene profile predictions

Colour	Gene evaluation
Black	No criteria or zero
Dark Green	Weak criteria met[a]
Green	Criteria met

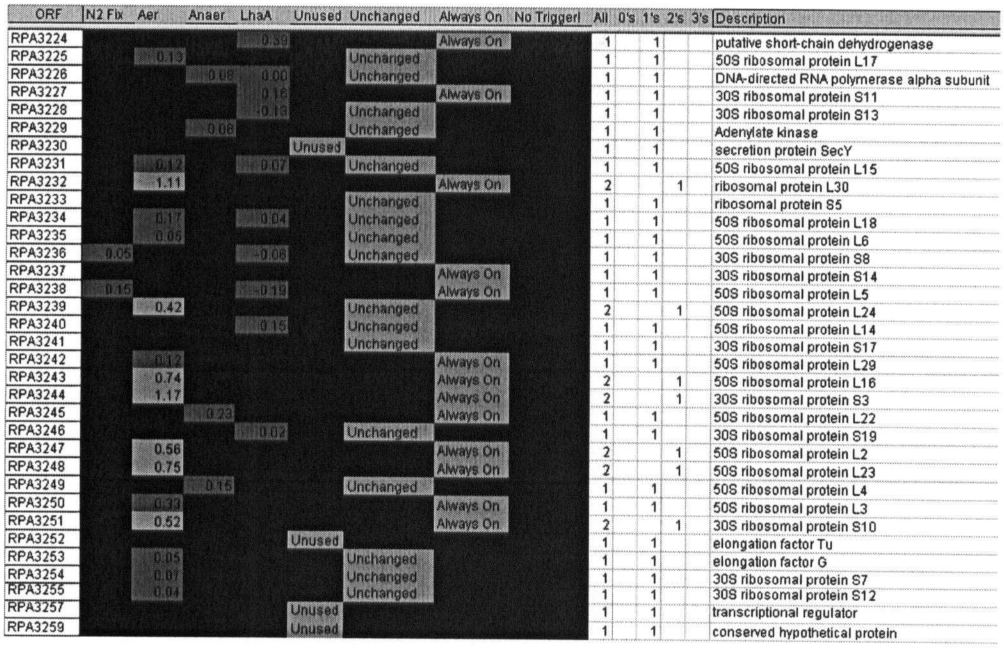
[a] not counted as a function call.

The above approaches have been built into the colour scheme used to display the results. Figure 9 displays the results of these criteria on the ribosomal proteins and the nitrogen fixation proteins. For the "unused", "unchanged" and "always on" columns, the results are displayed either in black or bright green, reflecting the Boolean nature of the criteria used. In the other columns, the values of the minimum distance found using inequality criteria are shown: where the distance appears under the threshold value of 0.4, the background of the result is shown in dark green, indicating that the results are not significant; where the result is greater than the threshold, the background for that value has been coloured bright green, indicating that the difference is above the threshold, and may be of significance (that is, the amount of the protein in cells is likely to be genuinely regulated in response to different growth conditions).

4.7. Overall Results

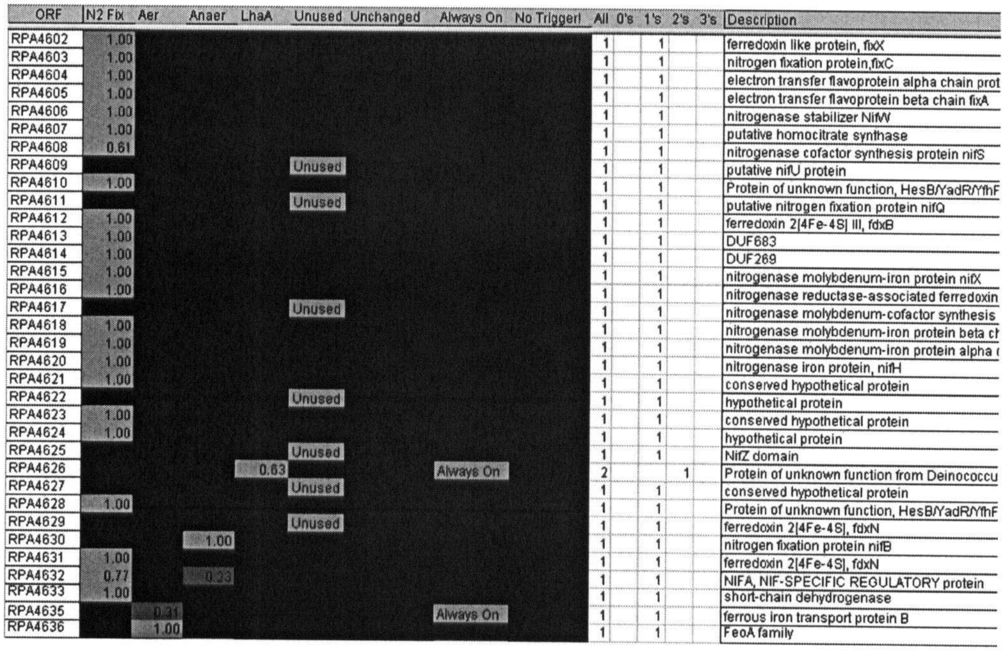
To demonstrate the outcomes of the use of this method to process MS data, two tables have been included that show the overall results obtained. The first (Table 10) demonstrates that the number of functional predictions per gene are consistently low and that few genes are not classified. (Classification of a gene as being unused, where the

A



ORF	N2 Fix	Aer	Anaer	LhaA	Unused	Unchanged	Always On	No Triggerl	All	0's	1's	2's	3's	Description
RPA3224				0.39			Always On		1		1			putative short-chain dehydrogenase
RPA3225		0.13				Unchanged			1		1			50S ribosomal protein L17
RPA3226			0.08	0.00		Unchanged			1		1			DNA-directed RNA polymerase alpha subunit
RPA3227				0.16			Always On		1		1			30S ribosomal protein S11
RPA3228				-0.13		Unchanged			1		1			30S ribosomal protein S13
RPA3229			0.08			Unchanged			1		1			Adenylate kinase
RPA3230					Unused				1		1			secretion protein SecY
RPA3231		0.12		0.07		Unchanged			1		1			50S ribosomal protein L15
RPA3232		1.11					Always On		2			1		ribosomal protein L30
RPA3233						Unchanged			1		1			ribosomal protein S5
RPA3234		0.17		0.04		Unchanged			1		1			50S ribosomal protein L18
RPA3235		0.05				Unchanged			1		1			50S ribosomal protein L6
RPA3236	0.05			-0.06		Unchanged			1		1			30S ribosomal protein S8
RPA3237							Always On		1		1			30S ribosomal protein S14
RPA3238	0.15			-0.19			Always On		1		1			50S ribosomal protein L5
RPA3239		0.42				Unchanged			2			1		50S ribosomal protein L24
RPA3240				0.15		Unchanged			1		1			50S ribosomal protein L14
RPA3241						Unchanged			1		1			30S ribosomal protein S17
RPA3242		0.12					Always On		1		1			50S ribosomal protein L29
RPA3243		0.74					Always On		2			1		50S ribosomal protein L16
RPA3244		1.17					Always On		2			1		30S ribosomal protein S3
RPA3245			0.23				Always On		1		1			50S ribosomal protein L22
RPA3246				0.02		Unchanged			1		1			30S ribosomal protein S19
RPA3247		0.56					Always On		2			1		50S ribosomal protein L2
RPA3248		0.75					Always On		2			1		50S ribosomal protein L23
RPA3249			0.15			Unchanged			1		1			50S ribosomal protein L4
RPA3250		0.33					Always On		1		1			50S ribosomal protein L3
RPA3251		0.52					Always On		2			1		30S ribosomal protein S10
RPA3252					Unused				1		1			elongation factor Tu
RPA3253		0.05				Unchanged			1		1			elongation factor G
RPA3254		0.07				Unchanged			1		1			30S ribosomal protein S7
RPA3255		0.04				Unchanged			1		1			30S ribosomal protein S12
RPA3257					Unused				1		1			transcriptional regulator
RPA3259					Unused				1		1			conserved hypothetical protein

B



ORF	N2 Fix	Aer	Anaer	LhaA	Unused	Unchanged	Always On	No Triggerl	All	0's	1's	2's	3's	Description
RPA4602	1.00								1		1			ferredoxin like protein, fixX
RPA4603	1.00								1		1			nitrogen fixation protein,fixC
RPA4604	1.00								1		1			electron transfer flavoprotein alpha chain prot
RPA4605	1.00								1		1			electron transfer flavoprotein beta chain fixA
RPA4606	1.00								1		1			nitrogenase stabilizer NifW
RPA4607	1.00								1		1			putative homocitrate synthase
RPA4608	0.61								1		1			nitrogenase cofactor synthesis protein nifS
RPA4609					Unused				1		1			putative nifU protein
RPA4610	1.00								1		1			Protein of unknown function, HesB/YadR/YfhF
RPA4611					Unused				1		1			putative nitrogen fixation protein nifQ
RPA4612	1.00								1		1			ferredoxin 2[4Fe-4S] III, fdxB
RPA4613	1.00								1		1			DUF683
RPA4614	1.00								1		1			DUF269
RPA4615	1.00								1		1			nitrogenase molybdenum-iron protein nifX
RPA4616	1.00								1		1			nitrogenase reductase-associated ferredoxin
RPA4617					Unused				1		1			nitrogenase molybdenum-cofactor synthesis
RPA4618	1.00								1		1			nitrogenase molybdenum-iron protein beta ch
RPA4619	1.00								1		1			nitrogenase molybdenum-iron protein alpha
RPA4620	1.00								1		1			nitrogenase iron protein, nifH
RPA4621	1.00								1		1			conserved hypothetical protein
RPA4622					Unused				1		1			hypothetical protein
RPA4623	1.00								1		1			conserved hypothetical protein
RPA4624	1.00								1		1			hypothetical protein
RPA4625					Unused				1		1			NifZ domain
RPA4626				0.63			Always On		2			1		Protein of unknown function from Deinococcu
RPA4627					Unused				1		1			conserved hypothetical protein
RPA4628	1.00								1		1			Protein of unknown function, HesB/YadR/YfhF
RPA4629					Unused				1		1			ferredoxin 2[4Fe-4S], fdxN
RPA4630			1.00						1		1			nitrogen fixation protein nifB
RPA4631	1.00								1		1			ferredoxin 2[4Fe-4S], fdxN
RPA4632	0.77		0.23						1		1			NIFA, NIF-SPECIFIC REGULATORY protein
RPA4633	1.00								1		1			short-chain dehydrogenase
RPA4635		0.31					Always On		1		1			ferrous iron transport protein B
RPA4636		1.00							1		1			FeoA family

Figure 9: Protein expression predictions for ribosomal and nitrogen fixation proteins. Using the same genes shown through Figures 6-8, putative environmental conditions are assigned for proteins. For the ribosomal proteins (A), with the exception of two that are not detected in the mass spectrometry (labeled as Unused), the others are all classified as unchanged or always on (See Table 1). For the nitrogen fixation proteins, although some were not detected by mass spectrometry (labeled as unused), most of them are correctly identified as nitrogen fixation related. To the right of the coloured area of both (A) and (B), the number of potential environmental conditions assigned to a given protein is shown. The column labeled "All" displays the number of conditions to which the protein was assigned, and the four following columns give an alternate representation of the data.

protein product is not detected, is included as a single prediction for that gene.) The majority of genes (86.0%) are clearly given only a single prediction, less than 10% of genes are predicted to have 2 or more classifications and while less than 5% cannot be classified.

Table 10: Function calls per gene based on criteria used to evaluate protein profile expressions

	0 functions called per gene (no trigger)	1 function called per gene	2 function called per gene	3 function called per gene	4 or more function called per gene
Number of genes	229	4142	401	43	0
% of genome	4.8%	86.0%	8.3%	0.9%	0.0%

The second table (Table 11) shows the groupings of the predictions by the classifications used. The majority of genes are classified as "unused", while 13.52% of genes are considered "always on". All other classifications include between seven and 9% of the total number of genes. It is worthy of note that these classifications are not necessarily unique. A gene that is up regulated significantly for use under nitrogen fixation conditions may also be expressed under other growth conditions, and thus also be classified as "always on" (See Materials and Methods, Table 1)

Table 11: Overall statistics of protein expression pattern assignments

	Nitrogen Fixation	Aerobic	Anaerobic	Unused	Unchanged	Always on	LhaA Mutant
Genes:	400	410	378	2534	352	651	348
Percentage:	8.3%	8.5%	7.9%	52.6%	7.3%	13.5%	7.2%

Notes: Groupings are not exclusive.

5. Discussion

5.1. Shotgun Proteomic Approach to Chromatophore Protein Content

The shotgun proteomic approach used to identify proteins in the chromatophore fraction from the purple phototrophic bacterium *R. palustris* was chosen primarily because, unlike 2 D electrophoresis – MS methods (Patton et al. 2002), it allows direct analysis of hydrophobic membrane proteins, and also because it serves as a relatively rapid screen of expressed genes.

Briefly, the method involved solubilization of membrane proteins using the detergent SDS, proteolysis of the denatured proteins to peptides with trypsin, μ–LC-ESI-MS/MS analysis using a QTOF mass spectrometer to generate peptide tandem mass spectra, and identification of parent proteins by searching the *R. palustris* genome sequence database using the SEQUEST search engine (see Materials and Methods for details). Although this method is rapid and direct there are several points that are worthy of mention:

1) Peptide ion selection for CID during μ–LC introduction is "top-down" and to some degree random (Yi et al. 2002), meaning that peptides which ionize well and that are from the most abundant proteins in the original mixture are the most likely to be selected;

2) For a protein to be identified the peptide tandem mass spectrum used in the database search must be of sufficient "quality" (which in part is related to the abundance of the peptide) to match a sequence in the database;

3) the absence of a sequence in the database for which a high quality peptide tandem mass spectrum is generated may lead to a false-positive because the software can generate a best-fit to a highly similar sequence that is present, although the probability scoring routine used minimizes this;

4) High versus low percent sequence coverage lends more weight to a protein identification and may be an indication of its relative abundance amongst proteins present in the original mixture;

5) Our search results were based solely on matching predicted genome sequences (*i.e.*, post-translational modifications of amino acids were not considered).

The most important point to be garnered from the above caveats is that a failure to identify a protein is not necessarily an indication of its absence from the sample.

5.2. Whole Cell Proteomics and Visualization

Interpretation of a dataset as encompassing as a complete proteome is a difficult undertaking because of the large amount of data available. Not only is it challenging to understand the significance of a single datum, but to comprehend the significance of the datum in relation to the entire dataset. In this case, where multiple proteomes are being compared, the issue becomes not only a challenge, but a significant obstacle to overcome.

The first hurdle is the need for the basic data to be readily interpretable. The development of a simple method for visualizing and categorizing a genome by expression profiles is a significant step in understanding how an organism manages its proteome. This is especially important for organisms like *R. palustris*, which are able to adapt to a wide variety of growth conditions. In addition to having a complete proteome and to be able to identify the conditions under which each protein is present, it is important to have the ability to display these data in a visually clear manner. This allows the reader to investigate and understand the proteome in a manner that has previously been inaccessible.

5.3. Integration of Diverse Datasets

One of the major strengths of the approach used in this thesis has been the easy integration of a number of different datasets By creating a database in which one can correlate different sources and types of data (i.e. homology, localization and functional predictions), it becomes easier to accumulate more information on any given target gene of interest. Thus, any single bioinformatics prediction can be interpreted in relation to the collection of information available for a given gene. In the collection of data available for *R. palustris* proteome, while some independently generated information may not be consistent, the majority of the data sources can be used together to build a stronger argument for the function or location of a protein in the proteome. The more independent sources that agree, the greater are the chances of the prediction being correct (Tong et al, 2002).

The integration of datasets has also provided the ability to perform genome/proteome wide searches for unknown proteins. A recent publication (Roszak et al. 2003) on the crystal structure of the photosynthetic reaction centre of *R. palustris* showed the presence of a protein with a mass of 10,707 D, which was termed protein "W" of unknown sequence. Based on its location in the crystal structure, "W" is likely to be a homologue of *PufX*, from *Rhodobacter* species (Frese et al. 2000), however, there is no known *R. palustris* homologue of *PufX* in the genome (http://genome.ornl.gov/microbial/rpal/). By using the mass and single known transmembrane helix of the "W" protein, it was possible to generate a list of candidate genes which may include the "W" protein, which is N-terminally blocked, preventing sequencing (Roszak et al. 2003). Unfortunately, because the post-translational modification(s) yielding the "W" protein in the crystal are unknown, the identity of the protein could not be determined (data not shown).

5.4. Predictions

The normalized dataset (sections 2.7.3. and 4.5) is simple to interpret and utilize, providing a quick evaluation of each gene. However, when the number of genes present in the genome is taken into consideration, it becomes clear that this large dataset requires at least one further step of processing. In the case where multiple datasets are used, representing differing environmental conditions or genetic modifications (i.e. *lhaA* deletion), resulting in differential gene expression, the change in observed protein abundance can be utilized to predict the expression profile of a given gene or family of genes. In this case, the normalized percent sequence coverage can be expanded to fill this role.

One caveat for the predictions obtained made is that the predictions themselves are only as accurate as the data input. For the portion of genes which were never observed in the proteome, the "unused" label may not necessarily apply to all of those cases. As discussed in section 1.4.3, a number of factors may influence the ability to detect some proteins or the trypic peptides they yield. Thus, the use of predictive groupings must be evaluated in the context of each individual protein.

For those cases where information about the relative abundance of a protein is available *a priori*, for instance, the ribosomal proteins and the nitrogen fixation proteins, the predictive groupings obtained from my algorithm appear to confirm what is already known about these proteins. A much more useful function of the groupings, which must be evaluated by further testing, is the predictions for proteins of unknown function or classification. While many hypothetical genes have now been confirmed by mass spectrometry, much work now remains to be done to determine their functions.

5.5. Justification of the Use of the Dataset

Although the datasets used in these experiments may not be complete and are likely to have missed proteins in low abundance, it is interesting to note that just over half of the proteins in the proteome have not been found under any of the growth conditions assayed (see Table 11). While this may seem to be a low percentage of the genome to have been expressed, many of the proteins that were not detected can reasonably be expected to have functions involving the processing of alternative carbon or nitrogen sources, stress responses and the like. *R. palustris*' ability to grow under a wide variety of growth conditions is consistent with the large numbers of genes that are present in the genome for the transport and breakdown of amino acids and compounds such as benzoate (Larimer et al. 2003). For example, *R. palustris* encodes approximately 325 transport systems, comprising at least 700 genes, of which 20 systems appear to be related to branched chain amino acid transporters (Larimer et al, 2003). In the defined media used for our experiments, few of these transporters should be needed. Similar conclusions were reached by Wassinger et al. (2000).

5.6. Weaknesses of the MS Approach

A number of additional concerns became apparent upon completion of the analysis of the dataset used here. As mentioned above, the limited number of genes being expressed as proteins is not unreasonable; however, the distribution of expressed genes (indicated by proteins detected) in the genome appears to contain artificial gaps. One example of this is the *nifE* gene (RPA4617), which appears to be centered in the *nif* operon. The *nifE* and *nifN* genes encode two subunits of the NifNE protein complex and would be expected to be co-transcribed (Fani et al. 2000). However, despite the detection of proteins encoded by both *nifN* and the genes present on the other side of *nifE* (i.e. *nifK*) the *nifE* gene product was not detected. Similarly, in the shotgun proteomics experiments on the *R. palustris* chromatophores, the a and c ATPase proteins were not detected (Fejes et al. 2003).

The gaps that appear in the proteome may occur for a variety of reasons. While many of them are likely to be because the protein was absent from the cell, some of them may exist because the proteins are highly hydrophobic and may not be solubilized during sample preparation. Poor detection may also occur for proteins which are present in small quantities or contain a low number of trypsin cleavage sites. Furthermore, if tryptic cleavage sites are inaccessible, the protein will not be cleaved or the peptides may not be of an appropriate size to be detected. While most proteins in this proteome are expected to yield tryptic fragments, there are likely to be a few proteins which do not generate peptides of the appropriate size for analysis. In these cases, such proteins will appear to be absent.

Loss of proteins can also occur during the processing of the mass spectrometry data. Because each protein fragment is compared back to the predicted spectrum for each of the peptides available, any fragment for which the predicted spectrum is missing will not appear in the results. This would certainly be the case for any fragment containing a post-translational modification such as phosphorylation, as is the *Rhodobacter sphaeroides* type I RubisCO protein (Wang and Tabita, 1992). In such cases, while the peptide may be present, the MS peaks do not correspond to the predicted peptides for that gene, and the association between the peptide and the protein from which it was derived cannot be made.

It is very clear that a protein that is missing from the final dataset does not necessarily indicate that the protein was not present in the cell (see section 1.4.3 and 3.3). Further testing has been done to demonstrate that two runs is insufficient, (unpublished data from VerBerkmoes et al, performed on *Shewanella*). Thus, the rate of false negatives is probably significant in the data shown in this thesis. However, the peptides that were detected were filtered strongly, to ensure that those proteins which do appear can be accepted with >95% confidence. This allows the data to be evaluated with a minimal likelihood of false positives.

Another weakness in this approach is the outstanding question of whether the percent coverage of a protein by the identified fragments can be accepted as an indication of the relative amount of the protein in the organism. It may not be obvious why the percent sequence coverage should be correlated to the amount of a protein. In any given sample, a more or less random selection of peptides is passed through to the detector, which leads to the eventual call of a "hit" for any given protein. Each spectrum for a given peptide is in fact derived from of a number of identical peptides that have together passed through the LC and first MS detector. Thus, a single spectrum is derived from of a population of identical peptides. The larger the pool of a given protein is in the cells, the more of a given peptide can be generated from it. When less of a protein is present in the cell, fewer peptides will be obtained to give rise to a single spectrum. If the amount of a peptide is sufficiently low that it becomes undetectable, then the percent sequence coverate of the protein from which the peptide originate will decrease. Although a number of other factors are involved (from the size and sequence of the peptides to the detector's ability to identify the peptides) that may determine the percent coverage for each gene, this trend appears to exist throughout the proteome.

Finally, it is worth mentioning the major issue of post-translational modifications (PTMs), which were not investigated for either the chromatophore dataset or the whole cell proteomes. One example of a common PTM in bacterial cells is the phosphorylation of residues in signal relay proteins. Because phosphorylation is used in the activation and inactivation of regulatory proteins, it plays a key role in the composition of the proteome. For example, R. palustris is predicted to encode 451 potential regulatory and signaling genes, of which 225 are predicted to be signal transduction proteins (Larimer et al. 2004), frequent targets of PTMs. However, peptides containing PTMs such as phosphorylation are not identified by MS unless they are specifically targeted in the search algorithms. This is because the mass of the added PTM alters the spectrum of the peptide, preventing it from matching the predicted peptide spectrum.

Unfortunately, despite the significant role of PTMs in the regulation of proteins, the computationally expensive task of searching for peptides affected by PTMs has not been undertaken on these datasets. Thus, for proteins regulated in this manner, percent sequence coverage may actually decrease when a peptide is activated or deactivated, because peptides containing any form of PTM will become invisible to the software used to match the peptide with its protein of origin. However, other peptides from proteins that are post-translationally modified are not affected by the PTMs and will be detected whether or not PTMs are taken into consideration. Thus, PTM could result in the loss of one peptide from a protein and reduce the percent sequence coverage observed.

5.7. Processing of Data

Assuming that percent coverage is a reasonably accurate means of estimating the population of a protein in the proteome, the problem of using these data still exists. It is impossible to directly compare the percent coverage between two different proteins to obtain meaningful information, and so, to alleviate this problem, an independent baseline for each gene was used. Each protein was evaluated individually to determine its minimum level of expression, which was used as the normalizing or "baseline" condition. In this case, a minimum value for the baseline was set at 10% coverage. Although this value was a compromise between the reliability of the data and a desire to keep the baseline low for samples where higher baselines could not be obtained, it appears to have given reasonable results. Percent coverage data for two proteins below 10% (i.e. 3% and 6%) may not be significantly different, given the overall quality of the data used in this thesis, but would significantly alter the ratios of highly expressed genes when used as a denominator for normalizing data (see section 2.7.3).

For each protein with a baseline level of expression, it is possible to – at a glance – determine how the proteome changes as a result of altering culture conditions, relative to the baseline. It is equally important to be able to visualize these data in a meaningful manner, and for that reason, colours have been included to allow these changes to become intuitively obvious. There are few examples of colour schemes for scientific data that are intuitively obvious as well as widely accepted, and as such, there was no clearly apparent method by which it should be done in this case. For both the baseline and the percent coverage data, the colour scheme used was chosen to be similar, but not identical, to that used in the visualization of genome array data (Shalon et al. 1996). Because the data do not occur in a format conducive to the treatment that would be performed for data generated by array experiments, the same colour scheme could not be used in an identical fashion. Instead, by using a perceived change in "brightness" of the same colours, it was possible to generate an easily readable chart in which the relative brightness replaces the need for using only the underlying data that generated the graphical display. By using the colour as a background to the data itself, it was also possible to keep the numerical data visible, allowing both types of information to be presented, while maintaining a compact format for their display. Although not a novel concept in itself, this colourization allows for the proteomic data to be quickly scanned, and scrutinized at relevant locations as needed. This can be done for both the percent coverage as well as the normalized "baseline" data.

The final step in using the MS data is to use the accumulated information to perform a predictive analysis of the data. Although not all of the information derived from the proteome is of a predictive nature, this allows for hypotheses-based approaches for further studies. In this case, it was done by searching each protein's profile for a given set of patterns. This can be seen in Figure 9, where the various groupings are shown. Although the set of criteria (Table 1) used are not the only ones that are potentially available, they provide a simple displays through which the dataset can be evaluated.

5.8. SQL Based Web Application

Despite the simplicity of the tools created in this thesis, there are a number of limitations placed upon their use. Currently, the most daunting aspect of these tools is the lack of user-friendly interface. In order to alleviate this problem, the creation of an SQL based web application would be a logical step, for which the algorithms used here could be adapted without significant modification.

With the data in a spreadsheet, it was possible to perform simple line-by-line or column-by-column transformations of the data. Despite the column-based approach used here, it is important to note that all of the formatting and processing algorithms discussed in this thesis could be done on a row-by-row basis. The ability to perform these operations in either direction is important for the future usability of the algorithms presented, particularly if an SQL based approach were used in the creation of a web interface, particularly in HTML, ASP or other web-enabled interface which generates web pages on the fly in a row-by-row manner. Thus, the independence of each row of data from those preceding or following it is important. This also implies that individual rows can be calculated and regenerated independently of the entire dataset, suggesting the possibility of the development of a web-based interface through which queries can be performed.

5.9. Future Directions

There is one aspect which has not been explored in this thesis, with respect to pattern searching. One would expect that using the baseline data, similarly controlled genes would give rise to proteins with similar normalized patterns. It should be a relatively simple matter to compare normalized proteins to locate any groups of proteins that share a similar expression profile. However, to obtain optimal results, a larger dataset with greater statistical confidence (i.e. replicates of a given growth condition) is probably required to ensure that the patterns observed are reliable indicators of genuine change in the proteome.

To complement this approach, there are a number of further uses for which these sorts of data may be used. One example that stands out is the comparison of proteome data to mRNA genome array data. By comparing the results obtained with these two independent methods, the uncertainty would be reduced. Comparison of multiple datasets from such independent experiments has been shown to significantly improve the reliability of the combined dataset, which in turn makes predictive statements more accurate by removing false positives (Tong et al. 2001).

6. Summary

6.1. Summary of Chromatophore Experiments

1. Hydrophobic proteins such as of the RC, LH1, LH2 and cytochrome *b/c₁* complexes were detected with high probabilities. However, the absence of some predicted tryptic peptides that would be expected to be present (*e.g.*, of the F_0 a and c proteins, and perhaps some LH2 proteins) indicates that some membrane-imbedded proteins were not detected using this approach.
2. The periplasmic, soluble cytochrome c_2 and membrane-bound cytochrome c_y homologues were detected with a high probability, whereas the HPIP homologue was not. These results indicate that *R. palustris* chromatophore vesicles encircle periplasmic components, and that *R. palustris* cells grown under the conditions employed may utilize either cytochrome c_2 or cytochrome c_y as the electron carrier between the cytochrome b/c_1 and RC complexes. However, as noted above, failure to detect proteins by this method does not prove their absence from the sample.
3. The proteins designated as hypothetical on the basis of genome sequence analysis that were detected with >0.75 probability values are newly revealed as genuine cellular components. Because these proteins co-purified with chromatophores, these proteins are likely to be either membrane-bound or located within the periplasm.

4. Our data pave the way for future research that will test the validity of our protein assignments, and gene disruption experiments to evaluate the possible functions of relevant proteins detected in the chromatophore fraction.
5. The results that we obtained indicate that this shotgun proteomics technique is a powerful, although not a perfect approach that can be expanded to whole cell analyses of cultures grown under different conditions, and for comparisons of mutant to wild type cells with success.

6.2. Summary of Whole Cell Proteome Experiments

1. Mass spectrometry data of a both subcelluar fractions as well as complete proteomes is inherently non-intuitive. I developed algorithms to impose a structured order on the data and display the data in a more intuitive and easily understandable interface. These algorithms provide significantly improved accessibility and could lead to better developed tools for the study of proteomics.
2. Mass spectrometry provides a census of most of the proteins involved in aspects of cellular function, and differences in the proteome of cells grown under differing environmental conditions indicate genome-wide changes in gene expression.
3. The identification of proteins in any sample is a strong indication of the presence of the protein in the cell at the time of disruption, whereas the inability to locate a protein may not be indicative of an absence of the protein within the cell. Improved methods of resolving proteins in low abundance, few tryptic cleavage sites and post-translational modifications, as well as membrane proteins, will be required to create a more accurate map of the proteome.

7. References

Aebersold R. and Goodlett D.R. 2001. Mass spectrometry in proteomics. Chem Rev **101**:69-95

Aklujkar M., Harmer A.L., Prince R.C. and Beatty J.T. 2000. The orf162b sequence of Rhodobacter capsulatus encodes a protein required for optimal levels of photosynthetic pigment-protein complexes. J Bacteriol. **182**:5440-7.

Andronescu M., Fejes A.P., Hutter F., Condon A. and Hoos H.H. 2004. A new algorithm for secondary structure design. J Mol Biol, in press

Cogdell R.J., Isaacs N.W., Howard T.D., McLuskey K., Fraser N.J. and Prince S.M. 1999. How photosynthetic bacteria harvest solar energy. J Bacteriol **181**:3869-3879

Drews G. and Golecki J.R. 1995. Structure, molecular organization, and biosynthesis of membranes of purple bacteria. *In* **Blankenship R.E., Madigan M.T. and Bauer C.E.** (eds), Anoxygenic photosynthetic bacteria p. 231-257. Kluwer Academic Publishers, Dordrecht, The Netherlands.

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

Fani, R., Gallo, R. and Lio, P. 2000. Molecular evolution of nitrogen fixation: the evolutionary history of the *nifD*, *nifK*, *nifE* and *nifN* genes. J Mol Evol **51**:1-11

Fejes A.P., Yi E.C., Goodlet D.R. and Beatty J.T. 2003. Shotgun proteomic analysis of a chromatophore-enriched preparation from the purple photosynthetic bacterium *Rhodopseudomonas palustris*. Photosyn Res **78**:195-203

Ferro M., Salvi D., Brugiere S., Miras S., Kowalski S., Louwagie M., Garin J., Joyard J. and Rolland N. 2003. Proteomics of the chloroplast envelope membranes from *Arabidopsis thaliana.* Mol Cell Proteomics **2**:325-45.

Frese R.N., Olsen J.D., Branvall R., Westerhuis W.H., Hunter C.N. and van Grondelle R. 2000. The long-range supraorganization of the bacterial photosynthetic unit: A key role for PufX. Proc Natl Acad Sci U S A. **97**:5197-202.

Gardy J.L. Spencer C., Wang K., Ester M., Tusnády G.E., Simon I., Hua S., deFays K., Lambert C., Nakai K. and Brinkman F.S.L. 2003. PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria. Nucleic Acids Research **31:**3613-3617

Graves P.R. and Haystead T.A. 2002. Molecular biologist's guide to proteomics. Microbiol Mol Biol Rev. **66**:39-63

Goodlett D.R., Bruce J.E., Anderson G.A., Rist B., Pasa-Tolic L., Fiehn O., Smith R.D. and Aebersold R. 2000. Protein identification with a single accurate mass of a cysteine-containing peptide and constrained database searching. Anal Chem **72:**1112-1122

Harwood C.S. and Gibson J. 1986. Uptake of benzoate by *Rhodopseudomonas palustris* grown anaerobically in light. J Bacteriol **165:**504-509

Hippler M., Klein J., Fink A., Allinger T. and Hoerth P. 2001. Towards functional proteomics of membrane protein complexes: analysis of thylakoid membranes from *Chlamydomonas reinhardtii.* Plant J **28:**595-606

Kasper C.B. 1970. Fragmentation of proteins for sequence studies and separation of peptide mixtures. Mol Biol Biochem Biophys **8**:137-84

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

Kim M-K. and Harwood C.S. 1991. Regulation of benzoate-CoA ligase in *Rhodopseudomonas palustris*. FEMS Microbiol Lett 83:199-204

Kolker E., Purvine S., Galperin M.Y., Stolyar S., Goodlett D.R., Nesvizhskii A.I., Keller A., Xie T., Eng J.K., Yi E., Hood L., Picone A.F., Cherny T., Tjaden B.C., Siegel A.F., Reilly T.J., Makarova K.S., Palsson B.O. and Smith A.L. 2003. Initial proteome analysis of model microorganism *Haemophilus influenzae* strain Rd KW20. J Bacteriol **185**:4593-602.

Lancaster C.R.D., Ermler U. and Michel H. 1995. The structures of photosynthetic reaction centers from purple bacteria as revealed by X-ray crystallography. *In* **Blankenship R.E., Madigan M.T. and Bauer C.E.** (eds), Anoxygenic photosynthetic bacteria p. 503-526. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Lauber W.M., Carroll J.A., Dufield D.R., Kiesel J.R., Radabaugh M.R. and Malone J.P. 2001. Mass spectrometry compatibility of two-dimensional gel protein stains. Electrophoresis **22**:906-18.

Larimer F., Chain P., Lamerdin J., Malfatti S.S., Do L., Land M., Hauser L., Pelletier D.A., Beatty T., Lang A., Tabita F.R., Gibson J., Hanson T., Torres Y Torres J., Peres C., Harrison F., Gibson J. and Harwood C.S. 2004.

Genome sequence of the metabolically versatile photosynthetic bacterium *Rhodopseudomonas palustris*. Nature Biotechnology, **22**: 55-61

Lilburn T.G., Prince R.C. and Beatty J.T. 1995. Mutation of the Ser2 codon of the light-harvesting B870 α polypeptide of *Rhodobacter capsulatus* partially suppresses the *pufX* phenotype. J Bacteriol **177:**4593-4600

Link A.J., Eng J., Schieltz D.M., Carmack E., Mize G.J., Morris D.R., Garvik B.M. and Yates J.R. 1999. Direct analysis of protein complexes using mass spectrometry. Nature Biotech **17:**676-682

Loach P.A. 2000. Supramolecular complexes in photosynthetic bacteria. Proc Natl Acad Sci USA **97:**5016–5018

MacCoss M.J., Wu C.C. and Yates J.R. 3rd. 2002. Probability-based validation of protein identifications using a modified SEQUEST algorithm. Anal Chem. **74**:5593-9.

Meyer J., Kelley B.C. and Vignais P.M. 1978. Nitrogen fixation and hydrogen metabolism in photosynthetic bacteria. Biochimie **60**:245-60

Meyer T.E. and Donohue T.J. 1995 Cytochromes, iron-sulfur, and copper proteins mediating electron transfer from the cyt bc_1 complex to photosynthetic reaction center complexes. *In* **Blankenship R.E., Madigan M.T. and Bauer C.E.** (eds), Anoxygenic photosynthetic bacteria p. 725-745. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Okamura M.Y., Paddock M.L., Graige M.S. and Feher G. 2000. Proton and electron transfer in bacterial reaction centers. Biochim Biophys Acta **1458:**148-163

Ouellette A.J.A. and Barry B.A. 2002. Tandem mass spectrometric identification of spinach Photosystem II light-harvesting components. Photosyn Res **72:**159-173

Patton W.F., Schulenberg B. and Steinberg T.H. 2002. Two-dimensional gel electrophoresis; better than a poke in the ICAT? Curr Opin Biotech **13:**321-328

Peng J., Gygi S.P. 2001. Proteomics, the move to mixtures. J Mass Spectrom **36**:1083-91.

Prince R.C., Baccarini-Melandri A., Hauska G.A., Melandri B.A. and Crofts A.R. 1975. Asymmetry of an energy transducing membrane: the location of cytochrome c_2 in *Rhodopseudomonas sphaeroides* and *Rhodopseudomonas capsulata*. Biochim Biophys Acta **387:**212-227

Prince R.C. 1990. Bacterial photosynthesis: from photons to *Δ*p. The Bacteria **12:**111-149

Quach T.T., Li N., Richards D.P., Zheng J., Keller B.O. and Li L. 2003. Development and applications of in-gel CNBr/tryptic digestion combined with mass spectrometry for the analysis of membrane proteins. J Proteome Res **2**:543-52.

Roszak A.W., Howard T.D., Southall J., Gardiner A.T., Law C.J., Isaacs N.W. and Cogdell R.J. 2003. Crystal structure of the RC-LH1 core complex from *Rhodopseudomonas palustris*. Science **302**:1969-72

Rutherford K., Parkhill J., Crook J., Horsnell T., Rice P., Rajandream M-A. and Barrell B. 2000. Artemis: sequence visualisation and annotation. Bioinformatics **16**:944-945.

Shalon D., Smith S.J. and Brown P.O. 1996. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res. **6**:639-45

Sonnhammer E.L., von Heijne G. and Krogh A. 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol. **6**:175-82

Spoof L., Vesterkvist P., Lindholm T. and Meriluoto J. 2003. Screening for cyanobacterial hepatotoxins, microcystins and nodularin in environmental water samples by reversed-phase liquid chromatography-electrospray ionisation mass spectrometry. J Chromatogr **1020**:105-19

Tadros M.H., Katsiou E., Hoon M.A., Yurkova N. and Ramji D.P. 1993. Cloning of a new antenna gene cluster and expression analysis of the antenna gene family of *Rhodopseudomonas palustris*. Eur J Biochem **217:**867-875

Tatusov R.L., Koonin E.V. and Lipman D.J. 1997. A genomic perspective on protein families. Science **278**:631-637

Thöny-Meyer L., James P. and Hennecke H. 1991. From one gene to two proteins: the biogenesis of cytochromes *b* and c_1 in *Bradyrhizobium japonicum*. Proc Natl Acad Sci USA **88:**5001-5005

Tong A.H.Y., Drees B., Nardelli G., Bader G.D., Brannetti B., Castagnoli L., Evangelista M., Ferracuti S., Nelson B., Paoluzi S., Quondam M., Zucconi A., Hogue C.W.V., Fields S., Boone C. and Cesareni G. 2001. A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. Science **295**:321-324.

van Wilk K.J. 2000 Proteomics of the chloroplast: experimentation and prediction. Trends in Plant Science **5**:420-425

Varga A.R. and Staehelin L.A. 1983. Spatial differences in photosynthetic and non-photosynthetic membranes of *Rhodopseudomonas palustris*. J Bacteriol **154:**1414-1430

Varga A.R. and Staehelin L.A. 1985. Pigment-protein complexes from *Rhodopseudomonas palustris*: isolation, characterization and reconstitution into liposomes. J Bacteriol **161:**921-927

Wang X. and Tabita F.R. 1992. Interaction of inactivated and active ribulose 1,5-bisphosphate carboxylase/oxygenase of *Rhodobacter sphaeroides* with nucleotides and the chaperonin 60 (GroEL) protein, J Bacteriol, **174**: 3607-3611

Wassinger, V.C., Pollack J.D., Humphery-Smith, I. 2000. The proteome of *Mycoplasma genitalium*, Eur. J. Biochem. FEBS **267**:1571-1582

Yang H.F., Huang Z.Y., Zhang D.M., Liu Z.H. 1999. Biological effects of space flight on purple non-sulfur photosynthetic bacteria, Space Med. Med. Eng. (Beijing) **12**:46-50.

Yi E.C., Lee H., Purvine S.O., Aebersold R. and Goodlett D.R. 2002. Approaching complete peroxisome charcterization by gas-phase fractionation. Electrophoresis **23:**3205-3216

Young C. 1997. The Role of the *Rhodobacter capsulatus* Membrane Protein, orf1696, in Light-Harvesting I Complex Assembly, PhD thesis, University of British Columbia

Young C.S. and Beatty J.T. 1998 Topological model of the Rhodobacter capsulatus light-harvesting complex I assembly protein LhaA (previously known as ORF1696). J Bacteriol. **180**:4742-5.

Young C.S. and Beatty J.T. 2003 Multi-level regulation of purple bacterial light-harvesting complexes *In* **Green B.R. and Parson W.W.** (eds) Light-harvesting antennas in photosynthesis, Kluwer Academic Publishers, Dordrecht, Netherlands

Zabrouskov V., Giacomelli L., van Wijk K.J., McLafferty F.W. 2003, New approach for plant proteomics - Characterization of chloroplast proteins of *Arabidopsis thaliana* by top-down mass spectrometry. Mol. Cel. Prot. **2**:1253-1260

Zannoni D. 1995. Aerobic and anaerobic electron transport chains in anoxygenic phototrophic bacteria. *In* **Blankenship R.E., Madigan M.T. and Bauer C.E.** (eds), Anoxygenic photosynthetic bacteria p. 949-971. Kluwer Academic Publishers, Dordrecht, The Netherlands.

8. Appendices

8.1. SQL from Microsoft Access Database query *qryCOG-MS11 (By Strand/Gene)*

```
SELECT [tblAllProteins-July2003].Strand, [tblAllProteins-July2003].[New Number], [tblCogs_summary-
fromLoren].COG, [tblCogs_summary-fromLoren].Gene, [tblPSORT-b].[Anthony'sLocalization],
[qryCOGS-Summary (Good data)].CountOfGene, [tblCogs_summary-fromLoren].[COG Description],
tblAllRunsCoverage081103.[Aer_1st-1pep_def] AS Aer1, tblAllRunsCoverage081103.[Aer_2nd-
1pep_def] AS Aer2, tblAllRunsCoverage081103.[Aer_3rd-1pep_def] AS Aer3,
tblAllRunsCoverage081103.[Anaer_1st-1pep_def] AS An1, tblAllRunsCoverage081103.[Anaer_2nd-
1pep_def] AS An2, tblAllRunsCoverage081103.[LhaA_1st-1pep_def] AS Lha1,
tblAllRunsCoverage081103.[LhaA_2nd-1pep_def] AS Lha2, tblAllRunsCoverage081103.[N2_1st-
1pep_def] AS N21, tblAllRunsCoverage081103.[N2_2nd-1pep_def] AS N22,
tblAllRunsCoverage081103.Total, [qryMassSpecChromatophore(Summary)].MaxOfProbability AS
Chromatophore, [tblAllProteins-July2003].Description
FROM (([qryMassSpecChromatophore(Summary)] RIGHT JOIN ((Crossref LEFT JOIN
([tblCogs_summary-fromLoren] LEFT JOIN [qryCOGS-Summary (Good data)] ON [tblCogs_summary-
fromLoren].COG = [qryCOGS-Summary (Good data)].COG) ON Crossref.[Old Number] =
[tblCogs_summary-fromLoren].Gene) INNER JOIN [tblAllProteins-July2003] ON Crossref.[New
Number] = [tblAllProteins-July2003].[New Number]) ON [qryMassSpecChromatophore(Summary)].Orf =
[tblCogs_summary-fromLoren].Gene) LEFT JOIN [tblPSORT-b] ON Crossref.[New Number] =
[tblPSORT-b].[New Number]) LEFT JOIN tblAllRunsCoverage081103 ON [tblCogs_summary-
fromLoren].Gene = tblAllRunsCoverage081103.Locus
ORDER BY [tblAllProteins-July2003].Strand, [tblAllProteins-July2003].[New Number];
```

8.2. Gene Caller Program

```
#include <iostream> //instead of iostream.h
#include <fstream>  //instead of fstream.h
#include <cstdlib> //for srand and atof()
#include <iomanip>  // to use the setprecision manipulator
using namespace std;

const char tab = '\t';
const char enter = '\n';
const char nulls = '\0';


void writeout(ofstream & fout, char OrfName[8],
        int criteria1,
        int criteria2,
        double SC1,
        double SC2,
        double SC3,
        double SC4,
       int hits1,
       int hits2,
       int peptides) {
 int count=0;
 while (OrfName[count] != tab) {
  fout << OrfName[count];
  count++;
 }

 fout << tab;
 switch (criteria1) {
  case 0:
   fout << "-no peptide-" << tab;
   break;
  case 1:
   fout << "Peptides 1" << tab;
   break;
  case 2:
   fout << "Peptides 2" << tab;
   break;
  default:
   fout << "    Criteria Other: ";
 }
 switch (criteria2) {
  case 0:
   fout << "-no coverage-";
   break;
  case 1:
   fout << "Coverage 1";
   break;
  case 2:
   fout << "Coverage 2";
   break;
  default:
   fout << "    Criteria Other: ";
 }
 fout << tab << SC1 << tab << SC2 << tab << SC3 << tab << SC4 << tab << hits1 << tab << hits2 << tab
<< peptides << enter;
}


void openinput() {
 char t;
 ofstream fout;
 fout.open("C:\\Palustris\\NathansData\\Lh2Mutant\\Contrast_result_s.tab", ios::out);
 ifstream fin;
 fin.open("C:\\Palustris\\NathansData\\Lh2Mutant\\Contrast.txt", ios::in);
 if (!fin) {
  cout << "file input could not be opened." << endl;
  cin >> t;
  exit(1);
 }

 fin.seekg(0,ios::beg);
 streampos pos=fin.tellg();
```

```
//variables used throughout
char c;
double sensitivity = .6;  // works very well at .5

//  Header row variables
char orf[8];
char coverage1[8],coverage2[8],coverage3[8],coverage4[8],coverage5[8],coverage6[8],coverage7[8];
char totalcoverage[8];
char Xcorr1[8],Xcorr2[8],Xcorr3[8],Xcorr4[8],Xcorr5[8],Xcorr6[8],Xcorr7[8];

// Data Row variables
char charge;
char Sequence[100],lastSequence[100];
int sample1hits,sample2hits,lastsample1hits,lastsample2hits;
char Towriteout;
double
SumCoverage1,SumCoverage2,SumCoverage3,SumCoverage4,SumCoverage5,SumCoverage6,SumCover
age7;
int fLocus;
char Locus[6];
int inc, i, Same;
int proteins, peptidePerProtein;
int printoutwhy1,printoutwhy2;
double threshold;
char readin[6];

Towriteout = 'n';
proteins = 0;
c = fin.get();
c = fin.get();
orf[0] = 'q';
printoutwhy1 = 0;
printoutwhy2 = 0;
fLocus = 0;
sample1hits =0;
sample2hits =0;

//read to <enter>rpal

while (fLocus  ==0 ) {
  Locus[0]=Locus[1];         Locus[1]=Locus[2];        Locus[2]=Locus[3];
  Locus[3]=Locus[4];         Locus[4]=Locus[5];        Locus[5]=c;
  c = fin.get();

  // cout << Locus[0] <<Locus[1] <<Locus[2] <<Locus[3] <<Locus[4] <<Locus[5]<< enter;
  if (Locus[0]==enter && Locus[1]=='L' && Locus[2]=='o' && Locus[3]=='c' && Locus[4]=='u' &&
Locus[5]=='s') {
    fLocus++;
  }
}

while (c!= enter) {
  c=fin.get();
}

//*************** start here!***************
```

```
 while ( !fin.eof() ) {  // get one character at a time
  if (c == enter) {
   c = fin.get();
   if (fin.eof()) {
   // don't do anything!  Most especially, don't read another character!
    }
   else if (c == enter) {
   // don't do anything, just go to the next line
   }
   else if (c =='r') {
    cout << proteins << enter;
    proteins++;
    //set sensitivity threshold
    threshold = sensitivity * peptidePerProtein;
    if (threshold < 6) {threshold = 6;}
    // Check if printout criteria are met.
    if (sample1hits > sample2hits + threshold) {
     Towriteout = 'y';
     printoutwhy1=1;
    }
    else if (sample2hits > sample1hits + threshold) {
     Towriteout = 'y';
     printoutwhy1=2;
    }
    else {
     printoutwhy1=0;
    }
    if (Towriteout == 'y' )
{writeout(fout,orf,printoutwhy1,printoutwhy2,SumCoverage1,SumCoverage2,SumCoverage3,SumCoverag
e4,sample1hits,sample2hits,peptidePerProtein);}  // write out the stuff from the orf before
    Towriteout='n';       //reset for next protein.
    peptidePerProtein = 0;
    while (c!= tab) {     //Get ORF Name
        orf[0]=orf[1];
        orf[1]=orf[2];
        orf[2]=orf[3];
        orf[3]=orf[4];
        orf[4]=orf[5];
        orf[5]=c;
        orf[6]=tab;
        c = fin.get();
    }
    c = fin.get();
    inc = 0;
    // Get the coverage for each run
    while (c != tab) {coverage1[inc] = c;  inc++;  c = fin.get();}
    coverage1[inc] = nulls;  c = fin.get();  inc=0;
    while (c != tab) {coverage2[inc] = c;  inc++;  c = fin.get();}
    coverage2[inc] = nulls;  c = fin.get();  inc=0;
    while (c != tab) {coverage3[inc] = c;  inc++;  c = fin.get();}
    coverage3[inc] = nulls;  c = fin.get();  inc=0;
    while (c != tab) {coverage4[inc] = c;  inc++;  c = fin.get();}
    coverage4[inc] = nulls;  c = fin.get();  inc=0;
    // Get the total coverage
    while (c != tab) {totalcoverage[inc] = c;  inc++;  c = fin.get();}
  totalcoverage[inc] = nulls;            inc=0;
  while (c!= enter) {  c = fin.get();}  //read to end of line
  //in some case, genes are "duplicated", so ignore the coverage
  if (coverage1[0] == 'X') {coverage1[0]=nulls;}
  if (coverage2[0] == 'X') {coverage2[0]=nulls;}
  if (coverage3[0] == 'X') {coverage3[0]=nulls;}
  if (coverage4[0] == 'X') {coverage4[0]=nulls;}

  //convert char to Numbers for coverage
  SumCoverage1 = atof(coverage1);
  SumCoverage2 = atof(coverage2);
  SumCoverage3 = atof(coverage3);
  SumCoverage4 = atof(coverage4);

  if ((SumCoverage1 + SumCoverage2) > (SumCoverage3 + SumCoverage4 + 60)) {
    Towriteout = 'y';
    printoutwhy2=1;
  }
  else if ((SumCoverage3 + SumCoverage4) > (SumCoverage1 + SumCoverage2 + 60)) {
    Towriteout = 'y';
    printoutwhy2=2;
  }
  else {
    printoutwhy2=0;
  }
  //reset all variables for peptides - new peptides for a new orf.
  lastsample1hits=0;      lastsample2hits=0;
  sample1hits=0;          sample2hits=0;
  Xcorr1[0] = nulls;      Xcorr2[0] = nulls;      Xcorr3[0] = nulls;
  Xcorr4[0] = nulls;
  lastSequence[0] = enter;
  Sequence[0] = tab;
}
else {
  inc=0;
  while (Sequence[inc]!= tab) {
    lastSequence[inc] = Sequence[inc];
    inc++;
  }
  inc =0;
  while (c != tab) {  Sequence[inc] = c;  inc++;  c = fin.get();}
  Sequence[inc] = tab;  c = fin.get();

  //fix sequence + charge.
  charge=Sequence[inc-1];
  Sequence[inc-2] = tab;

  //Check to make sure that the Peptide is unique.
  Same = 0;
  inc = 0;
  while (Sequence[inc] != tab ) {
    if (Sequence[inc] == lastSequence[inc]) {
      Same=1; inc++;
    }
    else  {
      Same=0;
```

```
          peptidePerProtein++;
          while (Sequence[inc] != tab) {inc++;}
        }
      }
      inc =0;
      if (Same ==0) {
       //If they're not the same, then it doesn't matter what the last sample was.
        lastsample1hits =0;        lastsample2hits =0;
      }
      inc=0;  //reset variables
      Xcorr1[0] = enter;
      Xcorr2[0] = enter;
      Xcorr3[0] = enter;
      Xcorr4[0] = enter;

      //Get the Real Xcorr values for this peptide
      while (c != tab) {  Xcorr1[inc] = c;  c = fin.get(); inc++;}
      Xcorr1[inc] = tab;  c = fin.get();  inc=0;
      while (c != tab) {  Xcorr2[inc] = c;  c = fin.get(); inc++;}
      Xcorr2[inc] = tab;  c = fin.get();  inc=0;
      while (c != tab) {  Xcorr3[inc] = c;  c = fin.get(); inc++;}
      Xcorr3[inc] = tab;  c = fin.get();  inc=0;
      while (c != enter) {Xcorr4[inc] = c;  c = fin.get(); inc++;}
      Xcorr4[inc] = tab;                inc=0;

      if (lastsample1hits !=1) {//if not already counted
        if ((Xcorr1[0]!= tab) || (Xcorr2[0] != tab)) { //and is a new hit
          sample1hits++;  //add it as a hit.
        }
      }
      if (lastsample2hits !=1) {//if not already counted
        if ((Xcorr3[0]!= tab) || (Xcorr4[0] != tab)) {
          sample2hits++;  //add it as a hit.
        }
      }
      // work out last sample
      if ((Xcorr1[0] !=tab) || (Xcorr2[0] != tab)) {lastsample1hits++;}
      if ((Xcorr3[0] !=tab) || (Xcorr4[0] != tab)) {lastsample2hits++;}
    }
  }
  else { c=fin.get();} //if not at the start of line, just keep reading.
 }
 if (Towriteout == 'y')
{writeout(fout,orf,printoutwhy1,printoutwhy2,SumCoverage1,SumCoverage2,SumCoverage3,SumCoverag
e4,sample1hits,sample2hits,peptidePerProtein);} //Print out last protein of file.
 fin.close(); //close files
 fout.close();
}

int main() {
 int t;
 cout << "Anthony's Mass Spec filter - GENE CALLER" << enter;
 openinput(); //main processing procedure
 cout << "Done! ";
 cin >> t;
 return 0;
```

```
}
```

8.3. Excel Macro - Formatting/Colour Data

```
Sub Macro2()
' Macro2 Macro - colour
' Macro created 23/07/2003 by Anthony Peter Fejes
' Last modified 11/09/2003 by Anthony Peter Fejes - improved colours
' University of British Columbia - Beatty Lab
' Genes by Position

  Columns("A:Z").Select
  FinalRow = 4816     'Range("E372").End(xlDown).Row
  Range("A" & 2, "R" & FinalRow).Select
  With Selection.Borders(xlTop)
   .LineStyle = xlContinuous
   .Weight = xlThin
   .ColorIndex = 1
  End With
  With Selection.Borders(xlRight)
   .LineStyle = xlContinuous
   .Weight = xlThin
   .ColorIndex = 1
  End With
  variable2 = "F"


  For y = 72 To 82 'H to R
   variable2 = Chr(y)
   For x = 2 To FinalRow
    variable1 = Range(variable2 & x).Value
    If IsNull(variable1) Or variable1 = 0 Or variable1 = "" Then
     Range(variable2 & x).Interior.Color = RGB(0, 0, 0)
    ElseIf variable1 < 20 Then
     Range(variable2 & x).Interior.Color = RGB(175, 0, 0)
    ElseIf variable1 < 40 Then
     Range(variable2 & x).Interior.Color = RGB(255, 0, 0)
    ElseIf variable1 < 60 Then
     Range(variable2 & x).Interior.Color = RGB(255, 175, 75)
    ElseIf variable1 < 80 Then
     Range(variable2 & x).Interior.Color = RGB(255, 255, 0)
    Else
     Range(variable2 & x).Interior.Color = RGB(0, 255, 0)
    End If
   Next x
  Next y
  variable2 = "R"  'Chromatophore column
  For x = 2 To FinalRow

   variable1 = Range(variable2 & x).Value
   If IsNull(variable1) Or variable1 = 0 Or variable1 = "" Then
    Range(variable2 & x).Interior.Color = RGB(0, 0, 0)
   ElseIf variable1 < 0.2 Then
    Range(variable2 & x).Interior.Color = RGB(175, 0, 0)
   ElseIf variable1 < 0.4 Then
   Range(variable2 & x).Interior.Color = RGB(255, 0, 0)
  ElseIf variable1 < 0.6 Then
   Range(variable2 & x).Interior.Color = RGB(255, 175, 75)
  ElseIf variable1 < 0.8 Then
   Range(variable2 & x).Interior.Color = RGB(255, 255, 0)
  Else
   Range(variable2 & x).Interior.Color = RGB(0, 255, 0)
  End If
Next x

lastNumber = "RPA0000"
variable2 = "E"  'Number Column
For x = 2 To FinalRow
  variable1 = Range(variable2 & x).Value
  If ((Right(variable1, 4) + 0) <> Right(lastNumber, 4) + 1) Then
   Range("H" & x, "R" & x).Select
   With Selection.Borders(xlEdgeTop)
    .LineStyle = xlContinuous
    .Weight = xlThick
    .ColorIndex = 2
   End With
  End If
  lastNumber = variable1
Next x

Range("H" & 2, "H" & FinalRow).Select  'demark the Messed aerobic
With Selection.Borders(xlEdgeRight)
  .LineStyle = xlContinuous
  .Weight = xlThick
  .ColorIndex = 2
End With
Range("J" & 2, "J" & FinalRow).Select 'demark the Aerobic
With Selection.Borders(xlEdgeRight)
  .LineStyle = xlContinuous
  .Weight = xlThick
  .ColorIndex = 2
End With
Range("L" & 2, "L" & FinalRow).Select  'demark the Anaerobic
With Selection.Borders(xlEdgeRight)
  .LineStyle = xlContinuous
  .Weight = xlThick
  .ColorIndex = 2
End With
Range("N" & 2, "N" & FinalRow).Select  'demark the LHA mutant
With Selection.Borders(xlEdgeRight)
  .LineStyle = xlContinuous
  .Weight = xlThick
  .ColorIndex = 2
End With
Range("P" & 2, "P" & FinalRow).Select  'demark the Chromatophores
With Selection.Borders(xlEdgeRight)
  .LineStyle = xlContinuous
  .Weight = xlThick
  .ColorIndex = 2
End With
Range("Q" & 2, "Q" & FinalRow).Select  'demark the Total Coverage
```

```
  With Selection.Borders(xlEdgeRight)
   .LineStyle = xlContinuous
   .Weight = xlThick
   .ColorIndex = 2
  End With
  Range("A1").Select
End Sub
```

8.4. Excel Macro - Normalizing Data

```
Option Base 1 'set default array base value to 1 (array's first index = 1)

Sub FindBaseline()
' Macro FindBaseline
' Finds the baseline (non-zero) for each protein.
' Last Modified Dec 2/2003
' University of British Columbia - Beatty Lab
' Genes by position

myRange = Columns("A:ZZ").Select
FinalRow = 4816 'myRange.Rows(myRange.Rows.Count).Rows

FirstColumn = 73 'I - don't use semi-aerobic data
Dim Values(5)
Dim Sort(5)
Dim Temp1 As Single, Temp2 As Single

For x = 2 To FinalRow
 For y = 1 To 4 '4 now, used to be 5 with semi-aerobic
  Temp1 = 0
  Temp2 = 0
   If Range(Chr(FirstColumn + (2 * (y - 1))) & x).Value = "" Or _
    Range(Chr(FirstColumn + (2 * (y - 1))) & x).Value = 0 Or _
    Range(Chr(FirstColumn + (2 * (y - 1))) & x).Value = "X" Or _
    IsNull(Range(Chr(FirstColumn + (2 * (y - 1))) & x).Value) Then
    Temp1 = 0
   Else
    Temp1 = CSng(Range(Chr(FirstColumn + (2 * (y - 1))) & x).Value)
   End If
   If Range(Chr(FirstColumn + (2 * (y - 1)) + 1) & x).Value = "" Or _
    Range(Chr(FirstColumn + (2 * (y - 1)) + 1) & x).Value = 0 Or _
    Range(Chr(FirstColumn + (2 * (y - 1)) + 1) & x).Value = "X" Or _
    IsNull(Range(Chr(FirstColumn + (2 * (y - 1)) + 1) & x).Value) Then
    Temp2 = 0
   Else
    Temp2 = CSng(Range(Chr(FirstColumn + (2 * (y - 1)) + 1) & x).Value)
   End If
   If Temp1 > Temp2 Then
    Values(y) = Temp1
   ElseIf Temp2 > Temp1 Then
    Values(y) = Temp2
   ElseIf Temp1 = 0 Then
    Values(y) = 0
   Else
```

```
      Values(y) = Temp1
    End If
Next y
For zb = 1 To 4
  Sort(zb) = Values(zb)
Next zb
For za = 1 To 3
  For z = za To 4
    If Sort(za) > Sort(z) Then
      Temp3 = Sort(z)
      Sort(z) = Sort(za)
      Sort(za) = Temp3
    End If
  Next z
Next za
For zc = 1 To 4 'X-AA
  If Sort(1) < 10 Then
    If Sort(2) < 10 Then
      If Sort(3) < 10 Then
        Denominator = Sort(4)
      Else
        Denominator = Sort(3)
      End If
    Else
      Denominator = Sort(2)
    End If
  Else
    Denominator = Sort(1)
  End If
  Range("U" & x).Value = Denominator
  If Denominator = 0 Then
    Denominator = 10
  End If
  col = 87 + zc
  If col > 90 Then
    col = "A" & Chr(64 + (col - 90))
  Else
    col = Chr(col)
  End If
  Temp4 = CSng(Values(zc) / Denominator)
' If Temp4 > 0 And Temp4 < 1 Then
'   Temp4 = 1
' End If
  Range(col & x).Value = Temp4
  If Temp4 = 0 Then
    Range(col & x).Interior.Color = RGB(0, 0, 0)
  ElseIf Temp4 < 0.7 Then
    Range(col & x).Interior.Color = RGB(175, 0, 0)
  ElseIf Temp4 < 1.5 Then
    Range(col & x).Interior.Color = RGB(255, 0, 0)
  ElseIf Temp4 < 2 Then
    Range(col & x).Interior.Color = RGB(255, 175, 75)
  ElseIf Temp4 < 5 Then
    Range(col & x).Interior.Color = RGB(255, 255, 0)
  Else
    Range(col & x).Interior.Color = RGB(0, 255, 0)
```

```
   End If
  Next zc
Next x
End Sub
```

8.5. Excel Macro - Grouping Protein Expression

```
Option Base 1 'set default array base value to 1 (array's first index = 1)

Sub evaluateData()
' Macro evaluateData
' Evaluates and collates the proteome data
' Last Modified November 14/03
' University of British Columbia - Beatty Lab
' Genes by position

'Version without Column W (semi-aerobic) - Commented out.

myRange = Columns("A:ZZ").Select
FinalRow = 4816 'myRange.Rows(myRange.Rows.Count).Rows

FirstWriteColumn = 68 'D - for use as ("A" & FirstWriteColumn)
FirstReadColumn = 87 'W
Col = 0

'first test - Nitrogen Fixation
For x = 2 To FinalRow
 If CSng(Range("X" & x).Value) < CSng(Range("AA" & x).Value) And _
   CSng(Range("Y" & x).Value) < CSng(Range("AA" & x).Value) And _
   CSng(Range("Z" & x).Value) < CSng(Range("AA" & x).Value) Then
    Max = Range("X" & x).Value
    If Max < Range("Y" & x).Value Then
      Max = Range("Y" & x).Value
    End If
    If Max < Range("Z" & x).Value Then
      Max = Range("Z" & x).Value
    End If
   Range("AD" & x).Value = Range("AA" & x).Value - Max ' difference between the N2 fix and next
largest
 Else
   Range("AD" & x).Value = CSng(0)
 End If
Next x

'Second test - Aerobic
For x = 2 To FinalRow
 If CSng(Range("X" & x).Value) > CSng(Range("Y" & x).Value) And _
   CSng(Range("X" & x).Value) > CSng(Range("AA" & x).Value) Then
    Max = Range("Y" & x).Value
    If Max < Range("AA" & x).Value Then
      Max = Range("AA" & x).Value
    End If
   Range("AE" & x).Value = Range("X" & x).Value - Max ' difference between the N2 fix and next largest
 Else
```

```
  Range("AE" & x).Value = CSng(0)
 End If
Next x

'Third test - Anaerobic
For x = 2 To FinalRow
 If CSng(Range("X" & x).Value) < CSng(Range("Y" & x).Value) And _
   CSng(Range("Z" & x).Value) < CSng(Range("Y" & x).Value) Then
    Max1 = Range("Z" & x).Value
    If Max1 < Range("X" & x).Value Then
     Max1 = Range("X" & x).Value
    End If
   Range("AF" & x).Value = Range("Y" & x).Value - Max1 ' difference between the N2 fix and next
largest
 Else
  Range("AF" & x).Value = CSng(0)
 End If
Next x

'fourth test - LhaA mutant
For x = 2 To FinalRow
 If (CSng(Range("X" & x).Value) < CSng(Range("Z" & x).Value) And _
   CSng(Range("Y" & x).Value) < CSng(Range("Z" & x).Value) And _
   CSng(Range("AA" & x).Value) < CSng(Range("Z" & x).Value)) Then
    Max = Range("X" & x).Value
    If Max < Range("Y" & x).Value Then
     Max = Range("Y" & x).Value
    End If
    If Max < Range("AA" & x).Value Then
     Max = Range("AA" & x).Value
    End If
    Range("AG" & x).Value = Range("Z" & x).Value - Max
 ElseIf (CSng(Range("X" & x).Value) > CSng(Range("Z" & x).Value) And _
   CSng(Range("Y" & x).Value) > CSng(Range("Z" & x).Value) And _
   CSng(Range("AA" & x).Value) > CSng(Range("Z" & x).Value)) Then
    Max = Range("X" & x).Value
    If Max > Range("Y" & x).Value Then
     Max = Range("Y" & x).Value
    End If
    If Max > Range("AA" & x).Value Then
     Max = Range("AA" & x).Value
    End If
    Range("AG" & x).Value = Range("Z" & x).Value - Max
 Else
  Range("AG" & x).Value = CSng(0)
 End If
Next x

'fifth test - SemiAerobic1
'For x = 2 To FinalRow
' If CSng(Range("W" & x).Value) > CSng(Range("Y" & x).Value) And _
'   CSng(Range("W" & x).Value) < CSng(Range("X" & x).Value) Then
'    Range("AH" & x).Value = "Midvalue"
' ElseIf CSng(Range("W" & x).Value) < CSng(Range("Y" & x).Value) And _
'   CSng(Range("W" & x).Value) > CSng(Range("X" & x).Value) Then
'    Range("AH" & x).Value = "Midvalue"
' Else
'   Range("AH" & x).Value = CSng(0)
' End If
'Next x

'sixth test - Semiaerobic2
'For x = 2 To FinalRow
' If (CSng(Range("W" & x).Value) < CSng(Range("X" & x).Value) And _
'    CSng(Range("W" & x).Value) < CSng(Range("Y" & x).Value) And _
'    CSng(Range("W" & x).Value) > 0) Or _
'    (CSng(Range("W" & x).Value) > CSng(Range("X" & x).Value) And _
'    CSng(Range("W" & x).Value) > CSng(Range("Y" & x).Value)) Then
'    Max = Range("Y" & x).Value
'     If Max < Range("X" & x).Value Then
'       Max = Range("X" & x).Value
'     End If
'     Range("AI" & x).Value = Range("W" & x).Value - Max
' Else
'   Range("AI" & x).Value = CSng(0)
' End If
'Next x

'Seventh test - Unused
For x = 2 To FinalRow
  If CSng(Range("X" & x).Value) = 0 And _
    CSng(Range("Y" & x).Value) = 0 And _
    CSng(Range("Z" & x).Value) = 0 And _
    CSng(Range("AA" & x).Value = 0) Then
     Range("AH" & x).Value = "Unused"
  Else
    Range("AH" & x).Value = CSng(0)
  End If
Next x

'eighth test - Constitutive
For x = 2 To FinalRow
      Max1 = Range("X" & x).Value
      If Max1 > Range("Y" & x).Value Then
        Max1 = Range("Y" & x).Value
      End If
      If Max1 > Range("Z" & x).Value Then
        Max1 = Range("Z" & x).Value
      End If
      If Max1 > Range("AA" & x).Value Then
        Max1 = Range("AA" & x).Value
      End If
      Max2 = Range("X" & x).Value
      If Max2 < Range("Y" & x).Value Then
        Max2 = Range("Y" & x).Value
      End If
      If Max2 < Range("Z" & x).Value Then
        Max2 = Range("Z" & x).Value
      End If
      If Max2 < Range("AA" & x).Value Then
        Max2 = Range("AA" & x).Value
      End If
  If Max1 > 0.7 And Max2 < 1.5 Then
      Range("AI" & x).Value = "Unchanged"
  Else
    Range("AI" & x).Value = CSng(0)
  End If
Next x

'Ninth Test - Always On.
For x = 2 To FinalRow
  If CSng(Range("X" & x).Value) > 0 And _
     CSng(Range("Y" & x).Value) > 0 And _
     CSng(Range("Z" & x).Value) > 0 And _
     Range("AI" & x).Value <> "Unchanged" And _
     CSng(Range("AA" & x).Value) > 0 Then
     Range("AJ" & x).Value = "Always On"
  Else
    Range("AJ" & x).Value = CSng(0)
  End If
Next x

For x = 2 To FinalRow
    countingcells = 0
  'Ninth test - Never triggered?
  ' Section moved to the end to use count results - cut duplication

  'Count of relevent results
  If CSng(Range("AD" & x).Value) > 0.4 Then
    countingcells = countingcells + 1
  End If
  If CSng(Range("AE" & x).Value) > 0.4 Or _
     CSng(Range("AE" & x).Value) < -0.4 Then
    countingcells = countingcells + 1
  End If
  If CSng(Range("AF" & x).Value) > 0.4 Then
    countingcells = countingcells + 1
  End If
  If CSng(Range("AG" & x).Value) > 0.4 Or _
     CSng(Range("AG" & x).Value) < -0.4 Then
    countingcells = countingcells + 1
  End If
  If Range("AH" & x).Value <> 0 Then
    countingcells = countingcells + 1
  End If
  If Range("AI" & x).Value <> 0 Then
    countingcells = countingcells + 1
  End If
  If Range("AJ" & x).Value <> 0 Then
    countingcells = countingcells + 1
  End If

  Range("AL" & x).Value = countingcells
  If countingcells = 0 Then
    Range("AM" & x).Value = 1
  End If
  If countingcells = 1 Then
    Range("AN" & x).Value = 1
  End If
  If countingcells = 2 Then
    Range("AO" & x).Value = 1
  End If
  If countingcells = 3 Then
    Range("AP" & x).Value = 1
  End If
Next x

'time to sum it all up
Count = 0
For x = 2 To FinalRow  ' N2Fix
  If CSng(Range("AD" & x).Value) > 0.4 Then
    Range("AD" & x).Interior.Color = RGB(0, 255, 0)
    Count = Count + 1
  ElseIf CSng(Range("AD" & x).Value) = 0 Then
    Range("AD" & x).Interior.Color = RGB(0, 0, 0)
  Else
    Range("AD" & x).Interior.Color = RGB(0, 175, 0)
  End If
Next x
Range("AD" & (FinalRow + 3)).Value = Count

Count = 0
For x = 2 To FinalRow  ' Aerobic
  If CSng(Range("AE" & x).Value) > 0.4 Then
    Count = Count + 1
    Range("AE" & x).Interior.Color = RGB(0, 255, 0)
  ElseIf CSng(Range("AE" & x).Value) = 0 Then
    Range("AE" & x).Interior.Color = RGB(0, 0, 0)
  Else
    Range("AE" & x).Interior.Color = RGB(0, 175, 0)
  End If
Next x
Range("AE" & (FinalRow + 3)).Value = Count

Count = 0
For x = 2 To FinalRow  ' Anaerobic
  If CSng(Range("AF" & x).Value) > 0.4 Then
    Count = Count + 1
    Range("AF" & x).Interior.Color = RGB(0, 255, 0)
  ElseIf CSng(Range("AF" & x).Value) = 0 Then
    Range("AF" & x).Interior.Color = RGB(0, 0, 0)
  Else
    Range("AF" & x).Interior.Color = RGB(0, 175, 0)
  End If
Next x
Range("AF" & (FinalRow + 3)).Value = Count

Count = 0
For x = 2 To FinalRow  ' LhaA mutant
  If CSng(Range("AG" & x).Value) > 0.4 Or _
     CSng(Range("AG" & x).Value) < -0.4 Then
     Range("AG" & x).Interior.Color = RGB(0, 255, 0)
    Count = Count + 1
  ElseIf CSng(Range("AG" & x).Value) = 0 Then
  Range("AG" & x).Interior.Color = RGB(0, 0, 0)
 Else
  Range("AG" & x).Interior.Color = RGB(0, 175, 0)
 End If
Next x
Range("AG" & (FinalRow + 3)).Value = Count

Count = 0
For x = 2 To FinalRow  ' unused
 If Range("AH" & x).Value = "Unused" Then
  Count = Count + 1
  Range("AH" & x).Interior.Color = RGB(0, 255, 0)
 Else
  Range("AH" & x).Interior.Color = RGB(0, 0, 0)
 End If
Next x
Range("AH" & (FinalRow + 3)).Value = Count

Count = 0
For x = 2 To FinalRow  ' Constitutive
 If Range("AI" & x).Value = "Unchanged" Then
  Count = Count + 1
  Range("AI" & x).Interior.Color = RGB(0, 255, 0)
 Else
  Range("AI" & x).Interior.Color = RGB(0, 0, 0)
 End If
Next x
Range("AI" & (FinalRow + 3)).Value = Count

Count = 0
For x = 2 To FinalRow  'Always on
 If Range("AJ" & x).Value = "Always On" Then
  Count = Count + 1
  Range("AJ" & x).Interior.Color = RGB(0, 255, 0)
 Else
  Range("AJ" & x).Interior.Color = RGB(0, 0, 0)
 End If
Next x
Range("AJ" & (FinalRow + 3)).Value = Count

Count = 0
For x = 2 To FinalRow  ' Other
 If Range("AM" & x).Value = 1 Then
  Range("AK" & x).Value = "No Trigger"
  Count = Count + 1
  Range("AK" & x).Interior.Color = RGB(0, 255, 0)
 Else
  Range("AK" & x).Value = CSng(0)
  Range("AK" & x).Interior.Color = RGB(0, 0, 0)
 End If
Next x
Range("AK" & (FinalRow + 3)).Value = Count
End Sub
```