

To the Graduate Council:

I am submitting herewith a dissertation written by David R. McWilliams entitled "Bioinformatic and Proteomic Investigation of Chloroplast Transit Peptide Motifs and Genesis". I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Life Sciences.

Barry D. Bruce

Major Professor

We have read this dissertation
and recommend its acceptance:

Robert L. Hettich

Igor B. Jouline

Frank W. Larimer

Steven W. Wilhelm

Accepted for the Council:

Linda Painter

Interim Dean of Graduate Studies

(Original signatures on file with official student records.)

Bioinformatic and Proteomic Investigation of Chloroplast Transit Peptide Motifs and Genesis

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

David R. McWilliams
May 2007

Dedication

I dedicate this dissertation to my incomparable wife

Susanne Carroll McWilliams

With whom any project seems not only possible, but an adventure

And to my children

Jennifer, Andrew, and Anthony

Whose support, encouragement, challenge, and humor helped immeasurably

Acknowledgments

I would like to acknowledge all those who have assisted me in this project. First, I would like to especially thank my advisor, Dr. Barry D. Bruce, for taking a chance with a 'non-traditional' student, and a not very compliant or humble one at that. I would also like to thank my advisory committee, Drs. Robert Hettich, Igor Jouline, Frank Larrimer and Steve Wilhelm for their time and patience demanding hypothesis-driven instead of data-driven research. I thank Dr. Nathan VerBerkmoes for training me on the LCQ, and Dr. Hayes McDonald for helpful discussion on chromatography and other technique matters. Thanks are also due to Dr. Fu-Min Menn for help in trouble-shooting and exorcising LCQ demons. Finally, I would like to thank the past and present members of the Bruce laboratory: Anton Mitsky, Evan Reddick, Michael Vaughn, and Sarah Wright for trying to teach me biochemistry technique, sample isolation, and general biology lore.

Abstract

The eukaryotic mitochondrion was formed by the endosymbiotic association of an α -proteobacterium and a primordial phagocytic eukaryote. A second, and later, endosymbiosis between the eukaryote and a cyanobacterium gave rise to the chloroplast of plants. Following each of these events most of the organellar DNA was exported to the nucleus. A system evolved wherein proteins produced on cytosolic ribosomes are targeted to organelle protein translocators by N-terminal targeting sequences. Protein sorting between the chloroplast and the mitochondrion in the plant cell by the general import pathways shows remarkable fidelity despite a lack of sequence conservation among transit peptides and pre-sequences and despite very little sequence difference between these two targeting peptides. There is evidence for a hydrophobic recognition motif in mitochondrial pre-sequences, and a similar motif has been proposed for the chloroplast transit peptide. We have developed novel motif-finding methods and applied them to our own chloroplast proteome data and to literature mitochondrial data. We fail to find a hydrophobic motif that discriminates the chloroplast and the mitochondrion. Another little understood phenomenon of organelle protein trafficking is how the targeting sequence is acquired after transfer of organelle DNA to the nucleus. It has been hypothesized that the transit peptide is acquired by exon shuffling. We find no correlation of transit peptide lengths with exon boundaries. Furthermore, using highly expressed cyanobacterial proteins conserved in plants, we find that the transit peptide appears as likely to be attached within the primordial sequence as without, indicating a more stochastic process for the origin of the transit peptide.

Contents

Introduction	1
1 Comparison of MS/MS Search Algorithms	16
1.1 Abstract	16
1.2 Introduction	16
1.3 Materials and Methods	18
1.4 Results	18
1.5 Discussion	29
2 Chloroplast Proteomics	31
2.1 Abstract	31
2.2 Introduction	31
2.3 Materials and Methods	34
2.3.1 Plant production	34
2.3.2 Organelle isolation	34
2.3.3 Mass spectrometry	35
2.3.4 Mass spectrometric data analysis	36
2.3.5 Bioinformatics	36
2.4 Results	37
2.4.1 MS Search Strategy Comparison	38
2.4.2 Localization prediction	41
2.4.3 Conservation in Cyanobacteria	46
2.4.4 Functional characterization	49
2.5 Discussion	50
3 Motif Analysis of Chloroplast Transit Peptides	54
3.1 Abstract	54
3.2 Introduction	54
3.3 Materials and Methods	59
3.4 Results	60
3.4.1 Application of MEME to targeting sequences	60
3.4.2 Heuristic analysis of targeting sequences	63
3.5 Discussion	77

4	<i>Synechocystis</i> PCC 6803 Proteomics	79
4.1	Abstract	79
4.2	Introduction	79
4.3	Materials and Methods	81
4.3.1	<i>Synechocystis</i> culture and fractionation	81
4.3.2	Protein digestion	81
4.3.3	Column packing and sample loading	81
4.3.4	High performance liquid chromatography	82
4.3.5	Mass spectrometry conditions	82
4.3.6	Mass spectrometric data analysis	83
4.3.7	Bioinformatic analyses	84
4.4	Results	85
4.4.1	The Experimental Proteome	85
4.4.2	Protein Detection Frequency is Proportional to Codon Adaptation Index	91
4.4.3	Experimental proteins are more likely to be conserved in plants . .	93
4.5	Discussion	97
4.5.1	Proteomics as expression analysis	97
4.5.2	Conservation in plants	98
5	Stochastic Nature of Transit Peptide Acquisition	100
5.1	Abstract	100
5.2	Introduction	100
5.3	Materials and Methods	101
5.3.1	Comparison of exon and targeting sequence length	101
5.3.2	Determination of transit peptide attachment points	102
5.3.3	Calculation of random targeting sequences	104
5.4	Results	104
5.4.1	Correlation of targeting sequence length and exon boundaries . . .	104
5.4.2	Distribution of predicted transit peptide ‘attachment points’	107
5.4.3	Predicted targeting specificity of random sequence	113
5.5	Discussion	116
	Summary	120
	References	124
	Appendix	139
A	Chloroplast Proteome Results - Sequest	140
B	Chloroplast Proteome Results - Three Algorithms	148
C	Top-Hits to Proteobacteria	159

D Top-Hits to Yeast	161
E <i>Synechocystis</i> Proteins	164
Vita	174

List of Tables

1.1	Pearson Correlation Coefficients for Search Scores ^a	19
1.2	Estimate of Single-Algorithm False Rejection Rate ^a	23
2.1	Novel localization prediction algorithms cited in PubMed.	33
2.2	Sequest Two Peptide Protein Identifications ^a	39
2.3	Combined Algorithm Prediction ^a	39
2.4	TMHMM Predictions for the Two Search Strategies	39
2.5	Proteins Identified by ‘Any2’ Not Represented in ‘Seq2’ Results	42
2.6	Localization Prediction for Different MS/MS Search Strategies ^a	46
3.1	Motif Definitions	64
4.1	Summary of Proteins Found	85
4.2	Summary Statistics for the Experimental and Predicted Proteomes	86
4.3	Kegg Gene Classification	89
5.1	Arabidopsis Exon Data	105
5.2	Exon and Targeting Sequence Length	106
5.3	Percentage predicted Localizations of Random Targeting Sequences Attached to Mitochondrial Proteins ^a	115
5.4	Percentage predicted Localizations of Random Targeting Sequences Attached to Chloroplast Proteins ^a	115
5.5	Per Cent TargetP Localization of Arabidopsis Intergenic Sequence	116
5.6	Per Cent Predotar Localization Prediction for Intergenic Targeting Sequence	116
A.1	Chloroplast Two-Peptide Proteins	140
B.1	Chloroplast Proteins - Two Peptides or Two Algorithms	148
C.1	‘Any2’ Proteins with Top-Hit to Proteobacteria	159
D.1	‘Any2’ Proteins with Top-Hit to Yeast	161
E.1	Proteomics Results	164

List of Figures

1	General import pathway for mitochondria and chloroplasts	7
2	Structural features of a plant transit peptide in different environments . . .	11
3	Peptide ion fragmentation nomenclature	13
1.1	Comparison of scores for spectra in which all three algorithms identified the same peptide, z=1	20
1.2	Comparison of scores for spectra in which all three algorithms identified the same peptide, z=2	21
1.3	Comparison of scores for spectra in which all three algorithms identified the same peptide, z=3	22
1.4	Comparison of scores irrespective of peptide assignment, z=1	24
1.5	Comparison of scores irrespective of peptide assignment, z=2	25
1.6	Comparison of scores irrespective of peptide assignment, z=3	26
1.7	Number of proteins identified by the algorithms	28
2.1	Pseudo-gel plot of chloroplast proteins.	40
2.2	Comparison of Localization Predictors	47
2.3	Proportion of top-hits to Cyanobacteria.	48
2.4	Mapman functional characterization of chloroplast proteins.	49
3.1	Motifs in plant transit peptides	57
3.2	Globin fold logo plot	58
3.3	MEME motif for TargetP chloroplast sequences	61
3.4	N-terminal sequence logos	62
3.5	Percentage test sequences matched by rule 21	66
3.6	Percentage test sequences matched by rule 22	67
3.7	Percentage TargetP sequences matched by rule 16	69
3.8	Percentage TargetP sequences matched by rule 35	70
3.9	Percentage TargetP sequences matched by rule 22	71
3.10	Percentage TargetP sequences matched by rule 37	72
3.11	Percentage TargetP sequences matched by rule 38	73
3.12	Percentage proteomic sequences matched by rule 35	74
3.13	Percentage proteomic sequences matched by rule 22	75
3.14	Percentage TargetP sequences matched by rule38	76
4.1	Comparison of <i>Synechocystis</i> proteomic studies	86

4.2	Pseudo-2D Gel Plot of <i>Synechocystis</i> Experimental and Predicted Proteins	87
4.3	Functional categorization of proteins found. Data expressed as a fraction of those in the predicted proteome with the same annotation.	90
4.4	Frequency distribution of codon adaptation index for the experimental and predicted proteomes	92
4.5	Comparison of total peptides per protein and CAI (data binned)	92
4.6	Rank distribution of codon adaptation index for the experimental and predicted proteomes	94
4.7	BLAST results for predicted and experimental proteome with different filtering criteria	95
4.8	Proteins with higher CAI are more likely to have plant homologue	96
5.1	Difference in first exon length and TargetP-predicted targeting sequence length, <i>Arabidopsis</i> predicted proteome	106
5.2	Difference in nearest exon boundary and targeting sequence length for TargetP training set chloroplast sequences	108
5.3	Difference in nearest exon boundary and targeting sequence length for TargetP training set mitochondrial sequences	108
5.4	Distribution of attachment points	109
5.5	Distribution of attachment points, truncated plot	111
5.6	ClustalW alignment of selected ftsZ sequences	112
5.7	Residue frequencies for organelle and random sequence	114

Introduction

Endosymbiosis

The eukaryotic cell might be considered life's smallest ecosystem. The organelles of prokaryotic origin and other compartments, the flow of information, the transport of nutrients, and protein trafficking are a truly amazing system. The presence of the cytoskeleton and flexible plasma membrane permit a larger cell size and more morphological flexibility than the peptidoglycan bound bacteria. Other features of the flexible plasma membrane important to eukaryotic evolution are invagination, which may have led to the formation of the endomembrane system, vacuoles, and other single-membrane bound organelles. Endocytosis, by which external materials may be engulfed and assimilated, is also made possible by the flexible plasma membrane[17]. The eukaryotic ancestor was probably a motile anaerobic heterotroph that fed on bacteria or other organic materials in the environment. About 2 billion years ago an engulfed α -proteobacterium became stably associated with eukaryote and eventually evolved into the mitochondrion[73]. What did the bacterium do for the eukaryote, or more properly, what was the evolutionary advantage of retaining it rather than digesting it? At the time this event is thought to have occurred the oxygen production by cyanobacteria had exceeded the geochemical reductive capacity of the earth and was beginning to increase in the atmosphere. It is likely that the oxygen-tolerant α -proteobacterium provided oxygen detoxification to the anaerobic eukaryote - the so-called 'OxTox' hypothesis[73]. The intracellular parasite *Rickettsia*, which is the genetically closest extant relatives of mitochondria, possesses two other features which may have led to the success of this arrangement - they have the ability to import pyruvate, and they have an ATP importer. Since mitochondria import pyruvate and *export* ATP, perhaps the membrane topology of the ATP transporter was reversed, in a mechanism similar to that proposed for the evolution of the protein import system (*vide infra*). Indeed the development of the ATP export system might have been the definitive event in the transformation of the commensally parasitic bacterium into an organelle.

The second endosymbiotic event occurred about 500 million years after the first when a cyanobacterium was assimilated[74]. The advantages to the host are more obvious in this case, since the photosynthetic endosymbiont could provide carbohydrates and reducing equivalents. Interestingly, the second endosymbiotic event also coincided with an atmospheric signature, a period in which the CO₂ was elevated to a level 10-200 times present levels[64, 74]. Whether this is related somehow to cyanobacterial endosymbiosis is not clear. While modern cyanobacteria are apparently capable of intracellular parasitism (within diatoms for instance), many more examples of more mutualistic associations with sponges, ascidians, and especially fungi are known[116]. This may be an indication that cyanobacteria also possessed the genetic and biochemical ability to adopt the intracellular habit. While the Rickettsia-like α -proteobacteria had the parasitic ability to siphon host nutrients, there seems to be no similar strategy for the autotrophic cyanobacterial endosymbiont. Perhaps protection from predation or a motile habit were the selective forces involved.

Each of these primary endosymbiotic events was apparently monophyletic: the mitochondria of all extant eukaryotes derive from a single association with an α -proteobacterium[73] and the plastids of all plastid-containing organisms derive from a single cyanobacterial association[91]. A possible exception for plastids is the amoeboid *Paulinella chromatophora* which contains two cyanobacterial endosymbionts related to *Synechococcus*[65]. The endosymbionts cannot be cultivated outside the host and divide synchronously with it. A closely related species *P. ovalis* lacks endosymbionts, but preferentially feeds on *Synechococcus*. It appears therefore that this single species may represent at least one other primary endosymbiosis in which *Synechococcus* has passed from “fodder to feeder”[73]. Leaving aside this exception, it seems incredible that only these two endosymbiotic events were successful. Either the bacteria assimilated were particularly well suited for association with the eukaryote or other endosymbiotic associations were lost in one of the many mass extinctions that have occurred in the past billion years. What has occurred many times are secondary endosymbioses, in which plastid-containing algae were engulfed by another eukaryote. The secondary plastids are surrounded by three or four membranes - the two from the original prokaryotic progenitor, plus the plasma membrane of the second host and/or the membrane from vacuole created to contain the engulfed algae. Some of these secondary endosymbionts are very successful contributors to the marine environment. Among these are the haptophyte algae whose blooms may be large enough to be seen from space[65], and the apicomplexa, a large group of secondary endosymbionts which subsequently lost the plastid. They are successful intracellular parasites, a notable member of which is the malaria parasite, *Plasmodium falciparum*.

One of the observations that supports the endosymbiotic theory of organelle origin is the circular genome of the bacterial type retained by both the chloroplast and mitochondrion. Almost all the coding capacity however has been exported to the eukaryotic nucleus. The protist *Reclinomonas americana* has the most mitochondrial genes retained by any organism so far known, 67. The total among all other organisms is 49[73]. Plastid genomes retain slightly more genes, 88 in *Arabidopsis*, and 76 in rice. Primitive plastid containing organisms appear to retain more, with the red alga *Porphyra* retaining 200 and the glaucocystophyte *Cyanophora* 136[89]. In comparison, the free-living α -proteobacterium *Escherichia coli* has 4237 genes, the parasitic *Rickettsia prowazekii* has 835, and the cyanobacterium *Synechocystis* has 3661. One explanation offered for this wholesale export of coding capacity is ‘Muller’s ratchet’ which states that small populations of asexually reproducing organisms will accumulate deleterious mutations[73]. Once they were established in the eukaryotic cell, the endosymbionts were cut-off from the genetic exchange with the parental population and mutational inactivation of genes may have become important. Export to the nucleus may have offset this through sexual reproduction. Doolittle proposed a more functional or process-oriented ‘ratchet’ mechanism[27] in which organelle lysis provides a steady supply of DNA to the nucleus. If the organellar copy became lost, and this loss was fixed in all the organelles, the nuclear copy became essential. If the nuclear copy were lost, evolution would ‘try again’ later. This mechanism assumes that more than one organelle is present or that the endosymbiosis was not obligate when it occurred since it requires organelle destruction. This is not the case in the chlorophyte *Chlamydomonas* which is basal to all the green algae and plants and has only a single chloroplast. Other primitive plastid containing organisms such as *Cyanophora* have two or more plastids however. This would appear to be the primordial situation, since the eukaryotic host would have had to somehow regulate the reproduction of the formerly free-living endosymbionts.

Several remarkable experiments have quantified the rate of export of organellar DNA to the nuclear genome. They furthermore indicate that this process is frequent, extensive, and on-going. Using yeast mitochondria transformed with a selectable marker gene, Thorsness *et al.* found the rate of DNA transport to the nucleus to be 2 in 10^5 per cell per generation[154]. The transfer from the nucleus to the mitochondrion was estimated to be at least 100,000 times lower. These estimates seem to support Doolittle’s functional ratchet for gene flow. More recently Stegemann *et al.* measured a rate of 1 in 5 million events for transfer of a chloroplast marker gene to the nucleus of somatic cells of tobacco[144]. This rate is within an order of magnitude of the previous results for the yeast mitochondrion. Huang *et al.*[51] measured an even higher rate of 1 in 16000 in a paternally transmitted

chloroplast marker gene. Since chloroplasts are maternally inherited in angiosperms, this may indicate a mechanism of DNA incorporation after degradation of the chloroplast during pollen production. It also implies an enhanced rate of gene transfer to the nucleus in the germ line which is necessary for the trait to be heritable. These estimates may be high since the marker genes were provided with promoters (in the chloroplast experiments at least) which would somehow have to be acquired in the natural event. But even so, given the population size and reproduction time of unicellular eukaryotes, or even the population size of some modern plants, these rates seem sufficient to support the facile transfer of DNA to the nucleus.

Two recent plant genome comparisons support facile organelle transport to the nucleus. Noutsos *et al.* examined several insertions of organellar DNA in the *Arabidopsis* and rice genomes[106]. The ages of the insertions could be inferred by comparison of the insertions to the organellar genomes, and many were found to be relatively recent. The insertions were of two types. The first type was linear segments of the organelle chromosome. Rice contains an almost complete copy of the linearized chloroplast genome, and *Arabidopsis* an almost complete copy of the mitochondrial genome. Older sequences took on the appearance of the background nuclear DNA through random mutations. The other type of insertion observed was more confusing. They appeared to be mosaics of random fragments of organellar sequence, but the mutation analysis indicated that the sequence of the fragments was very similar to the organellar source and did not arise from mutation to the nuclear background. The authors proposed that the fragments had been ligated in the cytosol and integrated into the nucleus as a unit. Given this apparent barrage of DNA, why is the nuclear genome not continually expanding? Matsuo *et al.* performed an analysis similar to Noutsos *et al.* for the rice genome only, and found that the nuclear genome appears to eliminate the organellar DNA over time. Larger segments were ejected faster than smaller segments[92]. They additionally dated the large rice chloroplast insertion at 600,000 years. In another comparison of various angiosperm genomes, Kellogg and Bennetzen find a wide variation in genome size, even among closely related species[67]. Some of this was due to polyploidy, but most was due to long terminal repeat (LTR) retrotransposons. The ages of these indicate that plants can eliminate them from their genomes. Thus plant genomes at least appear to have a remarkable ability to both accept and reject extra-nuclear DNA. The observation that the human genome has a large chunk of mitochondrial DNA suggests that similar processes occur in animals[77].

If transfer of DNA to the nucleus is this easy, then transfer of foreign DNA could conceivably occur. One example at least of this is known. *Bigeloviella natans* is a chlo-

rachniophyte that contains a secondary plastid derived from the green algae. *B. natans* is photosynthetic but is also capable of phagocytosis. A study of plastid targeted genes in this organism found that 21% were more closely related to organisms other than the existing green alga *Chlamydomonas* implying that they may have been obtained from prey genomes[6]. Most of these were from the red algae, which is reasonable since genes from these secondary endosymbionts would already possess both the signal peptide and the transit peptide necessary for plastid import. There were also genes of bacterial origin and even one that was encoded in the plastid of a red alga, showing that the necessary targeting information could still be acquired by the foreign DNA.

Why are any genes retained at all in the organelle genomes? The ‘hydrophobic hypothesis’ is one possibility, since some retained proteins tend to be very hydrophobic[48]. Others however, such as the plastid chlorophyll binding proteins, are very hydrophobic and are encoded in the nucleus. A second possibility is special gene processing factors. The mitochondrial genome has some several codons that are not the same as the universal code[4], so these may not be transcribed properly if transferred to the nuclear genome. Finally, kinetic or assembly factors might pertain. Photosystem D1 for example, is rapidly damaged by oxidation in photosynthesis and must be replaced frequently. It is still encoded in the plastid probably so it can be more quickly produced at need. Many ribosomal proteins are also still organelle encoded, perhaps to be available for the stepwise assembly of the ribosome[99].

Lateral gene transfer is a well known phenomenon among bacteria, but the transfer of prokaryotic DNA from organelles to the eukaryotic nucleus must be the largest such event in the history of life. Most of this transfer must have happened early in development of eukaryotes, since surviving primitive algae and protists also have reduced organelle genomes. The experiments cited above however indicate that the mechanisms for DNA transfer are still operative. The evolutionary advantages to the eukaryote are at least two-fold. First, DNA transfer to the nucleus would enable control of the formerly free-living organelles, both to keep their growth in check and to fully integrate them into cellular processes. Second, the continual influx of DNA must have been an engine of diversification for the eukaryotic genome. Mutation rates are capable of supplying significant variation for adaptation of prokaryotic genomes, but not eukaryotes ([105], p.23). The input of organellar DNA might be a hitherto unrecognized form of genomic variation similar to mobile elements (and polyploidy in plants) that is important in eukaryote evolution.

Protein Trafficking

The export of organellar DNA to the nucleus required a system to import the proteins necessary for proper organelle function. Since bacteria apparently do not possess a protein import system, this had to be created *de novo* by the evolving eukaryote. A functionally similar system evolved for both the mitochondrion and the chloroplast. Proteins produced on cytosolic ribosomes possess an N-terminal targeting sequence (termed ‘pre-sequences’ for mitochondria and ‘transit peptides’ for chloroplasts) which directs the protein to large, multi-protein import complexes in the outer membranes of the organelles. These complexes consist of β -barrel porins surrounded by ancillary proteins that serve a recognition or chaperone function. Proteins thread the outer membrane translocon (‘TOC’ or ‘TOM’, translocon of the outer membrane of the chloroplast or mitochondrion) and then a cognate translocon of the inner membrane, which is probably in contact with the outer membrane apparatus. Soluble and bound chaperones in the inner membrane space assist and direct the protein. Chaperones in the matrix or stroma assist transit either actively, in a DNA requiring process, or by prohibiting retrograde motion (the ‘Brownian ratchet’). Proteases then cleave the targeting sequence to generate the mature sequence. This is called the ‘general import pathway’ and is illustrated in simplified form in Figure 1. Proteins may be imported by other pathways, but the majority of proteins in each organelle are imported using this system.

In the chloroplast, TOC34 and TOC159 are GTPases which serve as recognition proteins and may be the product of an ancient gene duplication. They possess an acidic domain which is not well understood, but may be involved in recognition of the positively charged basic residues in transit peptides. TOC64 is a recently discovered outer membrane protein which has a tetratricopeptide repeat (TPR) domain. This domain is involved in protein-protein interaction, implying that TOC 64 may serve some recognition function as well[142]. Translocating proteins are then passed in a poorly understood process to the porin TOC75. This β -barrel protein is the translocation channel and has a homologue, SynToc75 in *Synechocystis*[121]. SynToc75 is apparently related to bacterial outer membrane protein (OMP) 85, members of which serve to *export* proteins to the outer membrane.

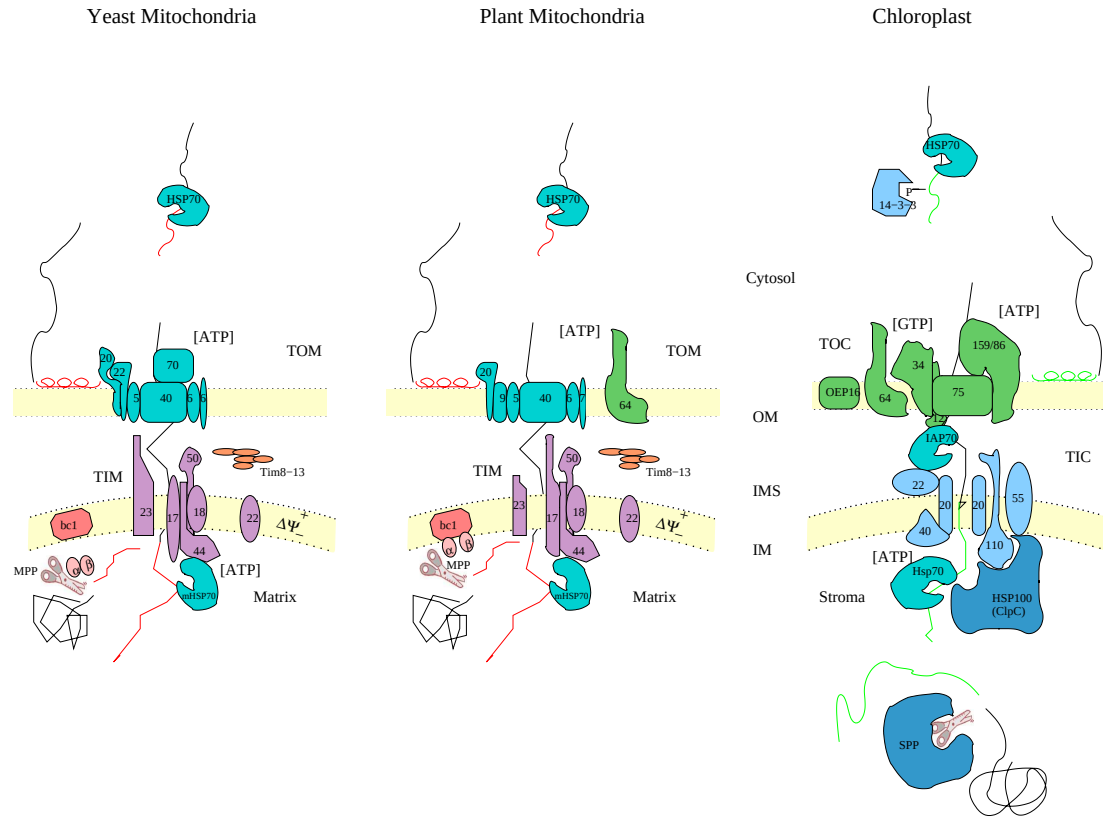


Figure 1: General import pathway for mitochondria and chloroplasts. TOM, translocon of the outer membrane of mitochondria; TIM, translocon of the inner membrane of mitochondria; TOC, translocon of the outer membrane of chloroplasts; TIC, translocon of the inner membrane of mitochondria; IM inter-membrane space; MPP, mitochondrial processing peptidase; SPP, stromal processing peptidase; numbers are approximate molecular weights in kDa and are used to designate the components, thus 'TOC34'. Adapted from [15], [79], [162].

It is thus possible that this protein was adapted for protein import by reversing its membrane ‘polarity’[122]. Another β -barrel protein OEP16 may be involved in the specialized import of photosynthetic proteins [118]. The inner membrane translocator (TIC) components are not understood as well since they are biochemically less accessible. Either TIC20 or TIC110 forms the translocation channel. The former is of cyanobacterial origin, whereas the latter is of eukaryotic origin. TIC110 has a sizable component in the inner membrane space and may serve as a point of contact with the TOC complex or the transiting protein. HSP90 proteins in the stroma may assist with import, but the HSP100 ClpC which is associated with the stromal side of TIC110 appears to be the primary ‘motor’ for protein import. A soluble peptidase cleaves the transit peptide in one or more steps.

In yeast mitochondria, three proteins appear to have recognition functions - TOM20, TOM22, and TOM70. TOM20 has a binding domain with a hydrophobic patch surrounded by acidic residues in which pre-sequences fit as an amphipathic α -helix[100]. TOM22 and TOM70 also have acidic domains. TOM70 has an additional TPR domain. The TOM20, TOM22, TOM70 arrangement is conserved in all fungi and animals. Plants do not possess a TOM70 analogue, and TOM22 is truncated to 9 to 11 kDa depending on the species[79, 84]. It appears that plant TOM20 adapted to assume some of the functions of TOM22, since it has more acidic residues than the yeast TOM20[84]. A homologue of TOC64 has been found in the mitochondrion using proteomic data[18]. No import function has been proved yet, but this component may substitute for the missing plant TOM70. The TOC import channel is TOC40, which is presumed to be of bacterial origin since it is a β -barrel, but no homologue has been identified[37]. In yeast the translocation channel in the inner membrane appears to be formed by TIM23, which possess N-terminal extensions into the inner membrane that may contact the TOC system. A similar system, built around the related TIM22 does not translocate proteins across the inner membrane, but inserts them in the ‘carrier import pathway’ (only TIM22 is shown in Figure 1). In plants on the other hand, the N-terminal extension of TIM23 is missing. Instead, TIM17 has an N-terminal extension[79]. TIM17,22,23 share about 22% similarity with the previously mentioned TOC component OEP16. Analogously to the chloroplast, there is an HSP70 associated with a membrane bound TIM component which assists or drives import in to the matrix. Most matrix targeted proteins also require the presence of the potential across the mitochondrial inner membrane as well. The pre-sequence is cleaved by a mitochondrial processing peptidase (MPP). In yeast this protein is soluble in the matrix. In plants, it is bound to the inner membrane as part of the bc_1 complex.

Thus, the mitochondrial and chloroplast systems have a great deal of functional similar-

ity. In the outer membrane the translocation channel is a β -barrel porin of bacterial origin, mediated by recognition proteins probably of eukaryotic origin. It might be expected that with the appearance of the cyanobacterial endosymbiont, that the evolving plastid would adapt to the problem of discriminating mitochondrion-bound and plastid-bound proteins. But comparison of the plant and animal/fungal mitochondrial systems show that significant adaptation occurred in both the inner and outer translocons.

Since the import systems are similar it must be that the information necessary to target the correct organelle must reside in the targeting sequences. Indeed, for the general import pathways, the pre-sequence or transit peptide will direct either native or heterologously attached proteins to the correct organelle. The nature of this specificity is not clear however. Examination of targeting sequences shows that they have more similarities than differences. While related sequences may show some sequence similarity in the targeting peptide, collections of pre-sequences or transit peptides show very little similarity or regular features that might imply recognition sites. Both pre-sequences and transit peptides are enriched in basic residues and hydroxylic residues and depleted in acidic residues, though mitochondrial sequences appear to have more arginine and chloroplast sequences more serine. Chloroplast transit peptides are generally longer than mitochondrial pre-sequences. Furthermore plant pre-sequences are somewhat longer than animal or fungal pre-sequences[162]. These observations imply that more information was required in the targeting sequences after the arrival of the plastid.

The similarity of both the import apparatus and the targeting sequences of mitochondria and chloroplasts are exemplified by the phenomenon of dual-targeting, examples of which have increased in recent years. Dual-targeted proteins fall into the categories: DNA and RNA maintenance, translation components, and cellular defense[85]. Of 24 organellar targeted tRNA synthetases found in *Arabidopsis* for instance, 15 are shared between the mitochondrion and chloroplast[28]. Mechanistically, dual-targeting can be achieved in several ways. Alternate targeting sequences can be obtained for the same protein via alternative start sites or alternative splicing. These approaches would presumably provide the plant with mechanisms to regulate the flow of the proteins to the different organelles. Of more interest for understanding import recognition are the so-called ‘ambiguous’ targeting sequences, which are recognized by both chloroplast and mitochondrial import systems[107]. Ambiguous sequences appear to have characteristics of both targeting sequences and in addition have an increased content of hydrophobic residues. In a related experiment it was demonstrated that a chloroplast protein with its transit peptide was slowly imported into spinach mitochondria, but could be imported efficiently into yeast mitochondria at about

40% of the rate in spinach chloroplasts. This indicates the discriminating effect of the modifications to the plant import apparatus mentioned above, but with the dual-targeting phenomenon, also illustrates the high degree of similarity in function of the two import apparatus and in the features of targeting sequence they recognize.

Another feature in which mitochondrial and plastid targeting sequences are similar is their secondary structure. Both are largely unstructured in solution and appear to attain at least some α -helical structure upon contact with lipid bilayers[14, 125, 160]. Some experiments for both pre-sequences and targeting-peptides hint that this might be specific for the lipids found in the membranes of their target organelles[20, 141, 160]. Similar behavior is found for membrane active bacterial and other toxins such as the honey bee venom melittin (see [71] for instance). A model illustrating this for a higher plant targeting peptide is shown in Figure 2. This model was obtained from circular dichroism and NMR structural analysis. The peptide exists mostly as a random coil in aqueous solution, but has induced α -helical structure in the more hydrophobic TFE/H₂O solvent. Mixed micelles which included a lipid present in the outer envelope of the chloroplast, monogalactosyldiacylglycerol (MGDG), also showed this behavior, though the helical content was somewhat less. Similar experiments have been conducted for mitochondrial pre-sequences. No experiments have been done yet that confirm whether these events occur *in vivo* however.

The structure of targeting sequences is consistent with theoretical models which show that unstructured proteins have increased binding rates because they can search a given space faster - the ‘fly-casting’ mechanism[137]. Targeting sequences are unstructured in solution and this may facilitate finding the correct organelle. Contact with the outer membrane might induce structure which would present potential binding motifs to the translocon recognition components. Alternatively, since targeting sequences have membrane activity, this interaction might stimulate assembly of the translocation apparatus.

The study of organelle origin and function, their control and coordination by the nucleus, and the mechanisms of protein trafficking to them is critical to understanding eukaryotic cell evolution and function. Much has been learned about protein trafficking in the past twenty years but the molecular basis of translocon recognition and discrimination of two the organelles in plants has remained obscure. One of the goals of this dissertation is to evaluate sequence motifs that may be important for organelle discrimination.

Proteomics

One of the barriers to progress in understanding the molecular basis of organelle protein trafficking has been the relatively small set of examples available for analysis. If the in-

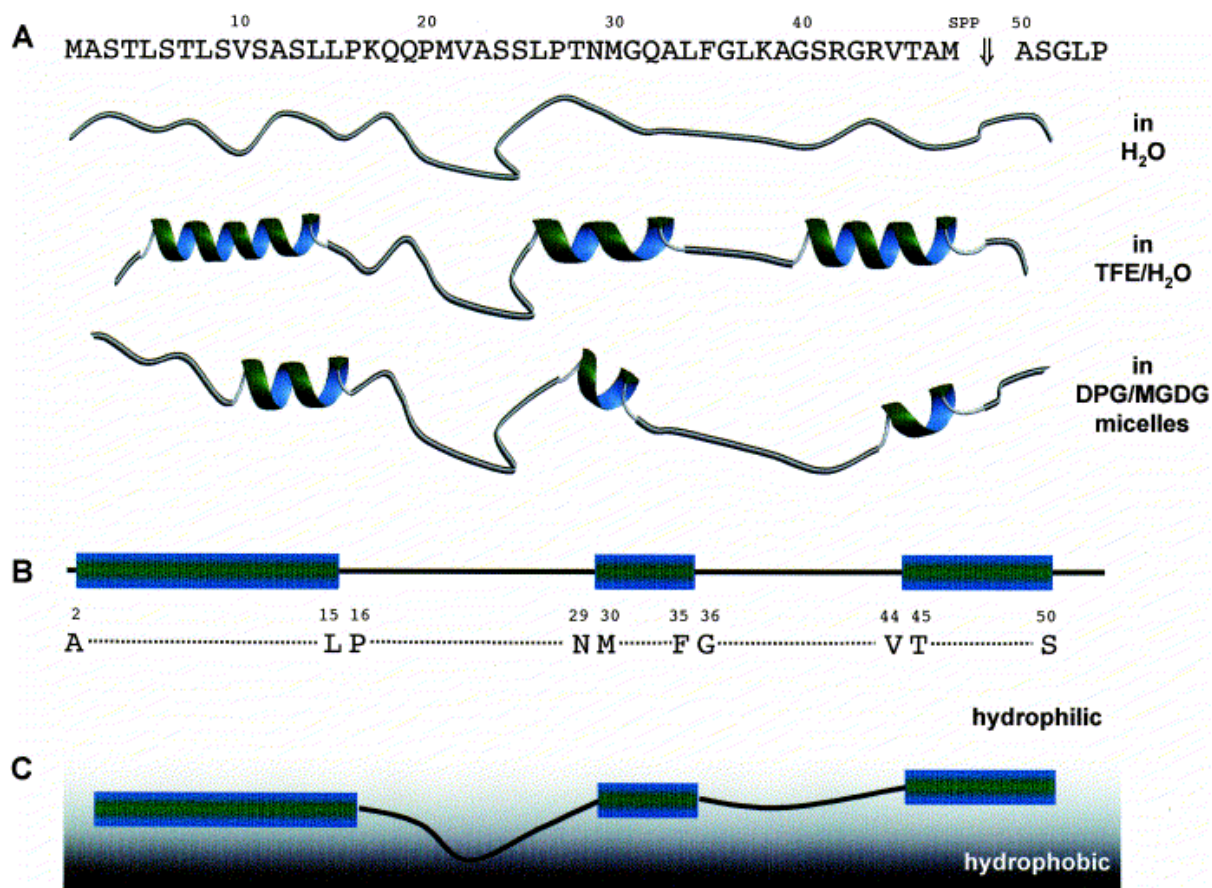


Figure 2: Structural features of a plant transit peptide in different environments. A: The sequence is that of the *Silene* ferredoxin transit peptide. The arrow indicates the cleavage site. The top structure illustrates the random coil found in solution, while the lower two show the helical structures found in TFE or in mixed micelles of dodecyl phosphocholine, dodecyl phosphoglycol, and monogalactosyldiacylglycerol (DPG/MGDG). B: A block diagram showing the three helical regions and intervening random coil delimited by the helix-breaking residues proline and glycine. C: Hypothetical diagram of potential membrane interaction of the helix-coil-helix arrangement. Data of Wienk *et al.*[160]; figure taken from Bruce[15].

formation content of the targeting sequences is low, then more examples might help in its detection. Progress is being made through the use of high throughput, mass spectrometry based protein analysis or ‘proteomics’. Mass spectrometry separates ions in electric or magnetic fields by their mass to charge ratio (m/z). The progress in mass spectrometry parallels that of the great advances in physical science of the twentieth century, with five Nobel prizes awarded for mass spectrometry or its applications. In the 1950’s, one of the first of modern ‘hyphenated’ analytical systems was created when the gas chromatographic separation technique was interfaced with a mass spectrometer (‘GC/MS’) for the separation and analysis of volatile compounds. Analysis of biopolymers and other polymers was hampered by the difficulty of ionizing and transferring these large molecules to the vapor phase without destroying them. Some progress was made with rapid pyrolysis, fast atom bombardment (FAB), and chemical ionization (CI), but these were not efficient enough to provide detailed information about biopolymers. The foundation technologies were provided by the discovery of the ‘soft ionization’ techniques matrix-assisted laser desorption ionization (MALDI) and electrospray ionization (ESI)[33, 60, 150]. In MALDI, the protein is embedded in matrix of a UV-absorbing organic acid. The matrix is irradiated and volatilized with a laser in the process transferring the protein to the gas phase with positive charge. In ESI the protein in an acidic or basic solution is sprayed through a needle which is part of the electrical circuit of the mass spectrometer creating charged droplets which travel toward the opposing electrode. Several pumping stages are employed to remove the solvent and other neutral molecules to increase sensitivity. Since ESI requires the sample to be in solvent, it is readily adaptable for interface to liquid separation technologies such as high performance liquid chromatography (HPLC).

For complex molecules such as proteins, mass alone is not sufficient for identification unless the universe of possibilities can be limited. In parallel with the soft ionization developments were ‘soft-fragmentation’ strategies such as collision activated dissociation[56, 55]. These techniques promote cleavage at the peptide bonds of proteins. While cleavage may occur preferentially at certain peptide bonds due to the particular amino acids or their sequence, in general cleavage can occur randomly on the peptide chain. The mass difference of cleavage at successive bonds gives the mass of the residue and the collection of random fragments is capable of yielding the sequence of the peptide or protein. It is thus possible to sequence peptides in the mass spectrometer, which is a tremendous improvement over wet chemical methods such as Edmann degradation which is limited to relatively short peptides and may be blocked by N-terminal modifications. Fragmentation typically yields two series of fragment ions since the charge may remain on either the N-terminal or C-

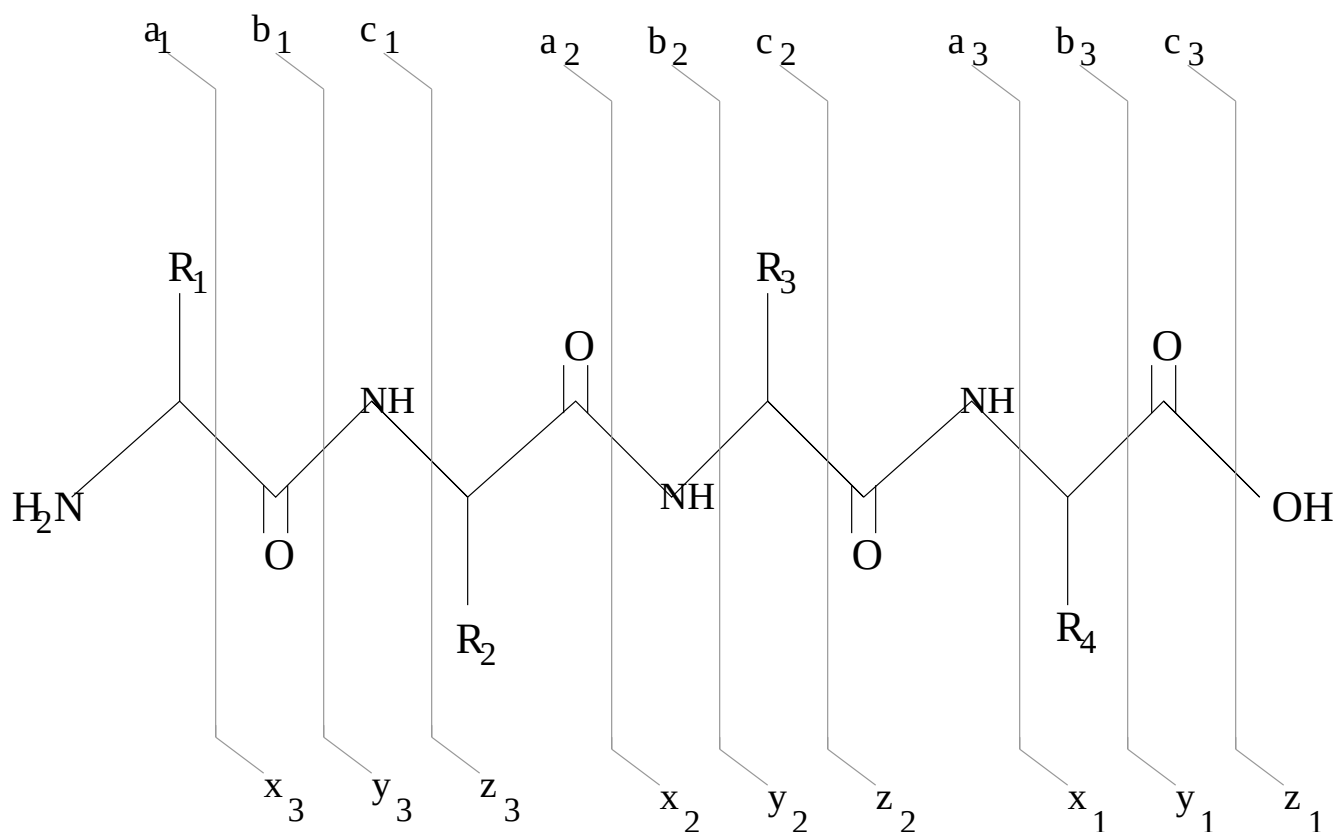


Figure 3: Peptide ion fragmentation nomenclature. Charge residing on the N-terminus gives the a,b,or c series. Charge residing on the C-terminus gives the x,y, or z series. R_n denotes arbitrary amino acid side chains.

terminal fragment. A terminology has been developed to describe these and is illustrated in Figure 3. For low energy CAD as typically employed for peptides, b- and y-ions are obtained.

Proteomics is currently delineated by the high throughput, non-targeted analysis of intact proteins or peptides produced from them by proteolysis. Analysis of intact proteins is usually called ‘top-down’ proteomics[66, 80]. This has the advantage of determination of the actual functional protein with any potential post-translational modifications, but is hampered currently by technical problems. Separation of intact proteins may be difficult due to the range of sizes and abundance. Even in a bacterial proteome, proteins may range from 1000 Daltons to 200,000 Daltons, and abundance may range over five or six orders of magnitude. Eukaryotic abundance ranges are up to ten orders of magnitude. The wide variety of post translational modifications possible for proteins also complicates their separation. In the mass spectrometer, proteins yield multiple charge states after ionization, requiring relatively expensive and high resolution instruments for analysis. Finally,

fragmentation of proteins may be more difficult, yielding shorter ion series. Progress is being made in all these areas however. Electron capture dissociation and electron transfer dissociation for instance show promise for efficient fragmentation of larger peptides and proteins[147].

Mass spectrometric analysis of the mixture of peptides obtained from enzymatic or chemical proteolysis is called ‘bottom up’ or ‘shotgun’ proteomics. Peptides are more tractable in separations, and if the enzyme is trypsin the size range is smaller and relatively constant. Trypsin has the added advantage of high specificity, and since it cleaves C-terminal to basic residues tends to produce doubly charged ions which produce good quality tandem mass spectra in mid-range instruments.

While analytically simpler, this approach introduces a new level of complexity since the peptides are identified in a mixture, outside their original protein context, and must be re-assigned to its parent protein. Bottom-up proteomics became a viable technology with the growth of databases of protein sequences. Algorithms like Sequest[31] and Mascot[110] were developed to match experimental peptide mass spectra to those predicted from *in silico* digestion of proteins from databases. By the late 1990’s these informatics and separations techniques had demonstrated the ability to identify proteins in complex mixtures and proteomes[78, 88, 94, 136]. Another problem with bottom-up proteomics is that post-translational modification information may be lost or difficult to recover. Searches for specific modifications may be made in the information processing stage, but this severely inflates the computational time required. There are techniques available for specific modifications such as phosphorylation, which have the advantage of reducing the complexity of the sample at the cost of added experimental difficulty. The tremendous productivity of bottom-up proteomics outweighs these difficulties and make it the current method of choice - the identification of 1000 to 1500 proteins is now possible with the application of a few days instrument time[26].

Mass spectrometry based proteomics is an exciting new technology capable of identifying hundreds or thousands of proteins to address problems of biological importance. It is the primary experimental technique employed in this work, where we employ it develop chloroplast and cyanobacterial protein data sets for analysis of transit peptide motifs and evolution.

General Outline

The goal of this research is to employ bioinformatics and proteomics to detect features of transit peptides involved in import specificity and to infer their evolutionary origin. The general outline of this work is the following. Since mass spectrometry was used to develop candidate proteins for analysis, this technology and the required informatics was developed in the laboratory. *Chapter 1* compares three widely available tandem mass spectrometric search algorithms. In *Chapter 2* is presented the analysis of the chloroplast proteome. The prediction accuracy of two localization algorithms is evaluated using the data, and the relationship of the proteins to the *Cyanobacteria* is analyzed. The results of Chapter One are applied to the data and shown to yield an increase in detections. *Chapter 3* uses the data of Chapter 2 in the bioinformatic search for transit peptide recognition motifs. The *Synechocystis* PCC 6803 proteome is presented in *Chapter 4*. The proteomics data are analyzed as a form of expression analysis by comparison to the codon adaptation index. The conservation of the proteins in plants is also examined. Finally, in *Chapter 5* the data of Chapter 4 are used to infer the mode of transit peptide acquisition.

Chapter 1

Comparison of MS/MS Search Algorithms

1.1 Abstract

A critical ancillary component of modern high throughput liquid chromatography/mass spectrometry proteomics technology is the computer algorithm used to assign the amino acid sequence to the hundreds or thousands of spectra accumulated in a typical analysis. Different algorithms have different approaches for interpreting spectra, and until recently, there have been few comparisons of their performance. We have compared the proteins identified by the three algorithms Sequest, Mascot, and X!Tandem from a large set of mass spectra from a shotgun proteomics study of the *Arabidopsis thaliana* chloroplast. The algorithms were found to have different sensitivities even after adjusting each to the same false positive rate. The implications and potential applications of the differential responses are discussed.

1.2 Introduction

The invention of the soft ionization techniques for mass spectrometry has permitted the analysis of biological molecules such as lipids, carbohydrates, and in particular, proteins, in a way not previously possible[33, 150]. For proteins and peptides, not only the mass, but the sequence can be determined in a suitable mass spectrometer equipped to fragment them and perform additional stages of mass separation. Electrospray ionization in particular is suitable for connection to a liquid chromatography separation system, permitting the identification of dozens or hundreds of proteins in an automated fashion. Analysis of intact

proteins has been termed ‘top-down proteomics’ and is capable of not only identifying the protein, but defining post-translational modifications or other changes to the primary sequence. Application of top-down analyses has been hampered by the difficulties inherent in separating proteins, which may vary greatly in size and chemical properties, and in fragmenting the large molecules efficiently in the mass spectrometer. Some recent advances, such as electron capture dissociation [147], may address the latter problem in the near future.

The analysis of peptides produced from proteins by chemical or enzymatic digestion is termed ‘bottom-up’ proteomics. Peptides are more easily separated by chromatographic methods and more easily sequenced in mass spectrometers with moderate capabilities. A new level of complexity is introduced however, since the number of spectra generated quickly exceeds the ability to manually interpret them to obtain the sequence. This has stimulated the development of algorithms for the machine interpretation of peptide mass spectra. These may be divided into four general categories. So-called *de novo* algorithms mimic human analysis by interpreting the fragment-ion series directly[151, 83]. Other methods depend on the existence of databases of known protein sequences for comparison. Similar to *de novo* methods, sequence tag methods derive a partial sequence from the spectrum and search sequence databases with FASTA or similar algorithms[87, 149]. Probabilistic algorithms assign a statistical score to the experimental spectrum match to theoretical spectra calculated from sequence databases[23, 24, 110]. Cross-correlation methods such as Sequest correlate theoretical and experimental spectra and apply various heuristics to improve the accuracy[31]. These are rough divisions and as the technology has matured, hybrids have been developed. X!Tandem for instance uses a sequence-tag approach to quickly identify candidate sequences in a database for a subsequent probabilistic match.

Due to limited computational resources, a preliminary comparison of available algorithms to search MS/MS data against predicted proteomes was undertaken, in particular to ascertain whether the open-source algorithm, X!Tandem, returned similar results to the more widely used Sequest and Mascot. Rough search time comparisons of the X!Tandem algorithm showed it to be approximately ten times as fast as Sequest for a typical search. The savings in search time would obviously be considerable if the algorithms produced comparable identifications. The other advantage of open-source algorithms such as X!Tandem is portability and scalability - the program can be compiled on any machine and copied to multiple processors without additional cost.

1.3 Materials and Methods

All programs were written in Perl. Statistical analysis and plots were done using the ‘R’ statistical package[115].

The *Arabidopsis thaliana* predicted proteome was downloaded from The Arabidopsis Information Resource (TAIR, www.arabidopsis.org) in August 2004. A decoy database was constructed by reversing each sequence and appending these to the forward sequences[109]. Mass spectra were taken from the shotgun proteomic analysis of the tryptic digest of the chloroplast proteins described in Chapter 2. The data set represented five distinct biological replicates with two 1D (reverse phase HPLC only) and three 2D (strong cation exchange followed by reverse phase HPLC) LC/MS/MS analyses.

A subset of the proteomic data, consisting of more than 11,000 spectra, was searched with each of the three algorithms, Sequest [31], Mascot[110], and X!Tandem[24] using the combined database. Sequest searches were performed with a peptide mass tolerance of 3.0 Da and a fragment tolerance of 0 Da. No enzyme specificity was required and 4 missed cleavages were allowed. Mascot searches were done with the default settings for an ion trap mass spectrometer. These were peptide mass tolerance of 1.2 and fragment tolerance 0.8. The enzyme specificity was trypsin with 2 missed cleavages. X!Tandem searches were also done with the default for an ion trap mass spectrometer. These were peptide mass tolerance of 1.5 and fragment tolerance 0.4. Trypsin and two missed cleavage sites was used for both Mascot and X!Tandem.

The search results for each algorithm were filtered by adjusting their individual parameters so that each algorithm detected less than 1% false positives. The false positive score was taken as: $100 * 2 * fp / total$, where fp is the number of matches to the reversed sequences and $total$ is the total number of matches.[109].

The three algorithms were then used to search the entire set of mass spectra using the forward database only. Parsers were written to extract the results from each algorithm and the results were stored in a Postgres database (www.postgresql.org). The Sequest XCorr, the Mascot score, and the X!Tandem expectation score were used for comparison of the algorithms’ responses. The parameters that achieved a 1% false positive rate were used for the protein-level comparison.

1.4 Results

The Sequest cross-correlation score (XCorr), the X!Tandem expectation score (represented as $-\log_{10}(Expectation)$) and the Mascot score are compared for those spectra in which

all three algorithms identified the same peptide, about 10% of the total (Figures 1.1-1.3). Scores for doubly-charged ions in Figure 1.2 clearly show correlation for all three algorithms indicating that spectra which score highly with one algorithm are likely to score highly in the others as well. The plots indicate that Sequest and X!Tandem are the most highly correlated, and Mascot and X!Tandem the least. These observations are corroborated by the correlation coefficients in Table 1.1. For the +3 ions the correlations are worse, and poorer still for the +1 ions, with very little linear trend in the scores for the latter (Figures 1.1 and 1.2, Table 1.1). Only 3.3% of the spectra assigned to the same peptide by all three algorithms were for the single charge state. While the trends in the plots and the correlations exhibit a good degree of agreement among the algorithms, they do respond to spectra differently, since the correlations are not high. The correlation coefficient between Sequest and Mascot, 0.57, means that 43% of the variability in the comparison is not accounted for by a linear comparison, for instance.

Table 1.1: Pearson Correlation Coefficients for Search Scores^a

Comparison	<i>z</i>		
	+1	+2	+3
Sequest-Mascot	0.1497	0.4405	0.1887
Sequest-X!Tandem	0.2338	0.5675	0.4039
Mascot-X!Tandem	0.0546	0.3712	0.3367

^aData are the correlation coefficients (r^2) for the spectra in which all three algorithms identified the same spectrum.

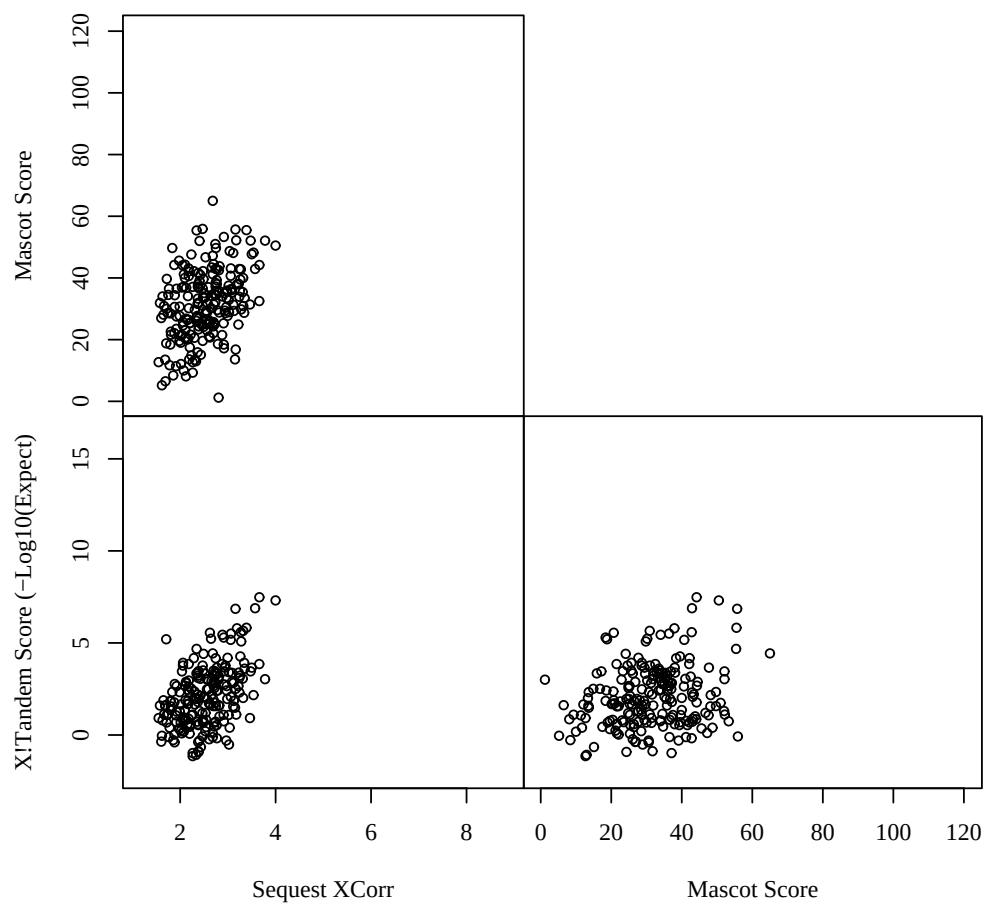


Figure 1.1: Comparison of scores for spectra in which all three algorithms identified the same peptide, $z=1$.

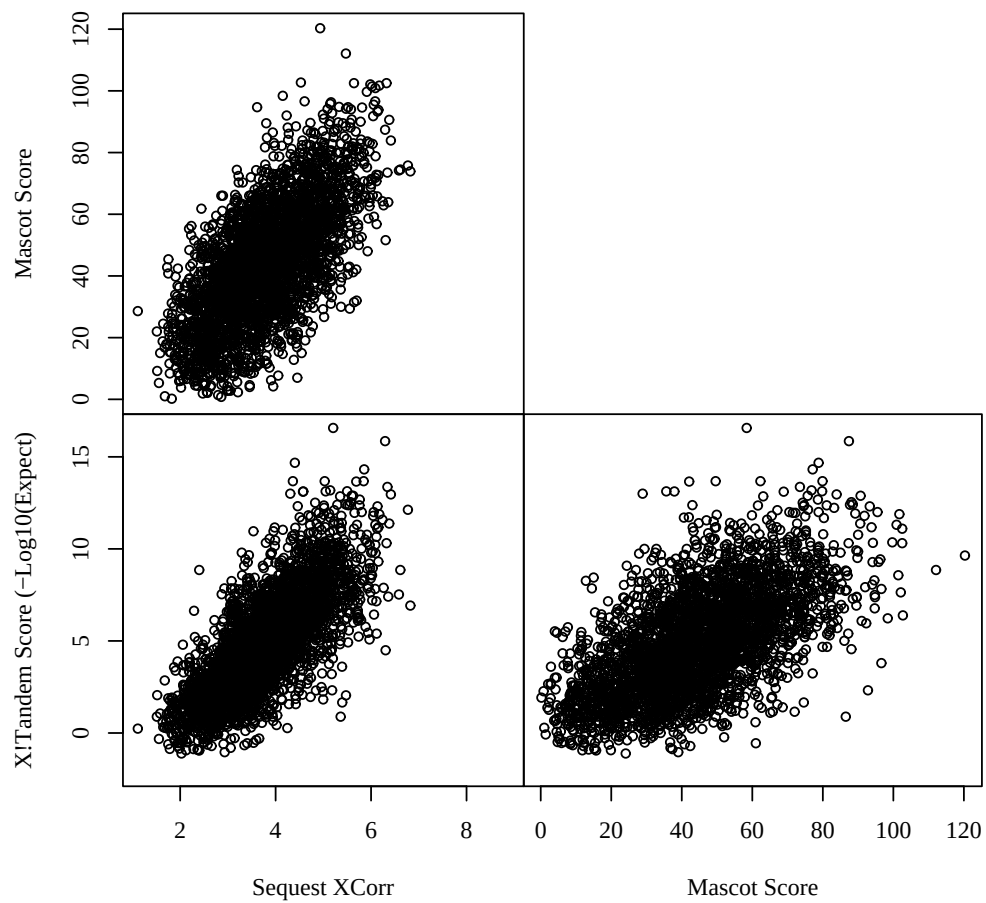


Figure 1.2: Comparison of scores for spectra in which all three algorithms identified the same peptide, $z=2$.

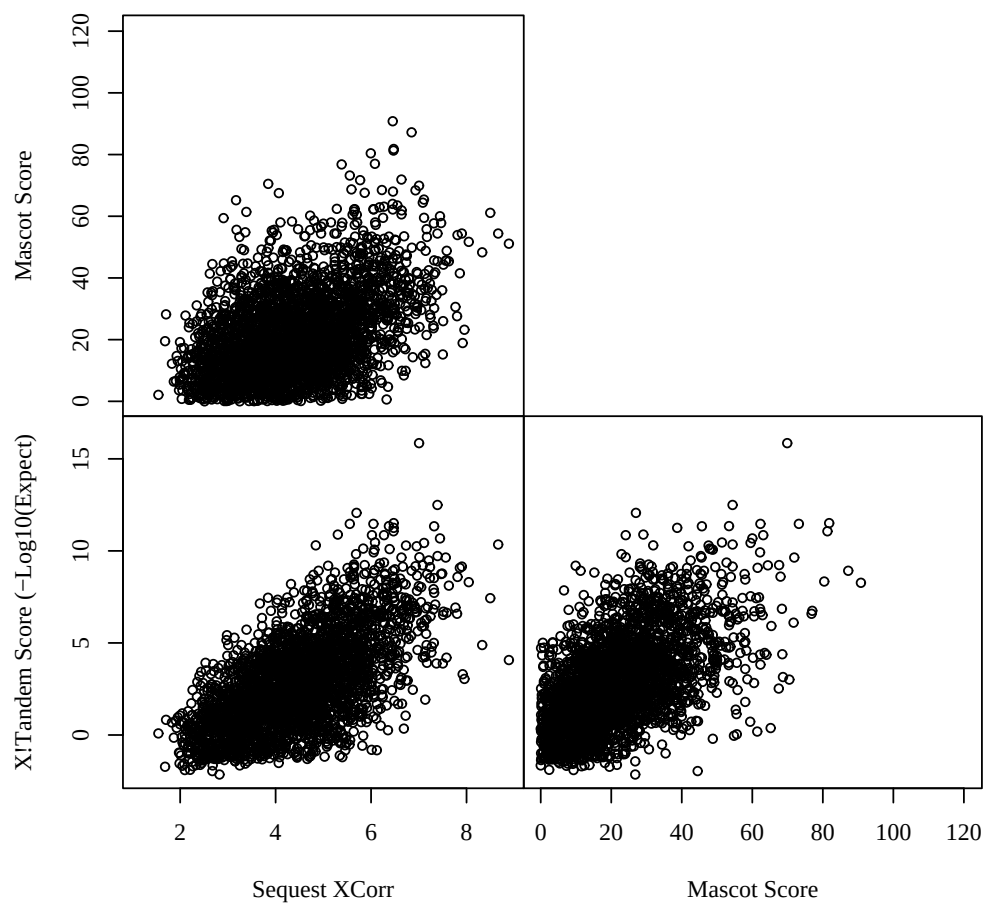


Figure 1.3: Comparison of scores for spectra in which all three algorithms identified the same peptide, $z=3$.

Table 1.2: Estimate of Single-Algorithm False Rejection Rate^a

Algorithm	z		
	+1	+2	+3
Sequest ^b	0.3	1.9	7.5
Mascot ^c	0.5	3.8	23.7
X!Tandem ^d	2.0	10.8	23.0

^aPercentage of identifications rejected by the algorithm tuned to a 1% false positive rate under the assumption that spectra identified by all three algorithms are correct; ^b $\Delta\text{CN} > 0.08$, $\text{XCorr} > 1.8(z=+1)$, $2.2(z=+2)$, $3.2(z=+3.2)$; ^ca constant score of 20 was taken for all ions; ^d $-\log_{10}(\text{expectation value}) > 2.5$.

If it is assumed that it is highly unlikely for all three algorithms to wrongly assign the same peptide to the same spectrum (i.e. these assignments represent true positives), the plots also suggest that a significant fraction of spectra may be wrongly rejected by typical filtering criteria (Table 1.2). At cut-offs which limit false positive identifications to less than 1% (see below), this amounts to 2%, 4% and 11% for Sequest, Mascot, and X!Tandem respectively for ions in the +2 charge state. Mascot and X!Tandem would wrongly reject 23% of +3 ions if this analysis is correct. Sequest would have the lowest rates for all ions under this assumption.

Figures 1.1-1.3 demonstrate a correlation among the algorithms when they assign a spectrum to the same peptide. In Figures 1.4-1.6 are compared the scores for each spectrum irrespective of the peptide assignment. Superimposed on the plots are the score cut-offs for a 1% false-positive rate determined from searching against a database including reversed sequences (since the cut-off for Mascot was variable, i.e. the score exceeding the ‘identity score’, the median value of the score is plotted). Points falling in the upper left and lower right quadrants defined by these cut-offs would be accepted by one of the algorithms in the comparison and rejected by the other. Points in the lower left and upper right quadrants would be rejected or accepted by both, respectively. For each comparison it appears that there are scores, some very high, that would be accepted by one algorithm, but rejected by the others. This is best illustrated for the +2 charge state shown in Figure 1.5. In keeping with the correlations in Table 1.1, Sequest and X!Tandem seem to be the most similar, though there is a noticeable distribution of Sequest scores that would be rejected by X!Tandem. The Mascot-X!Tandem comparison shows the most dissimilarity, with large populations in the upper left and lower right quadrants.

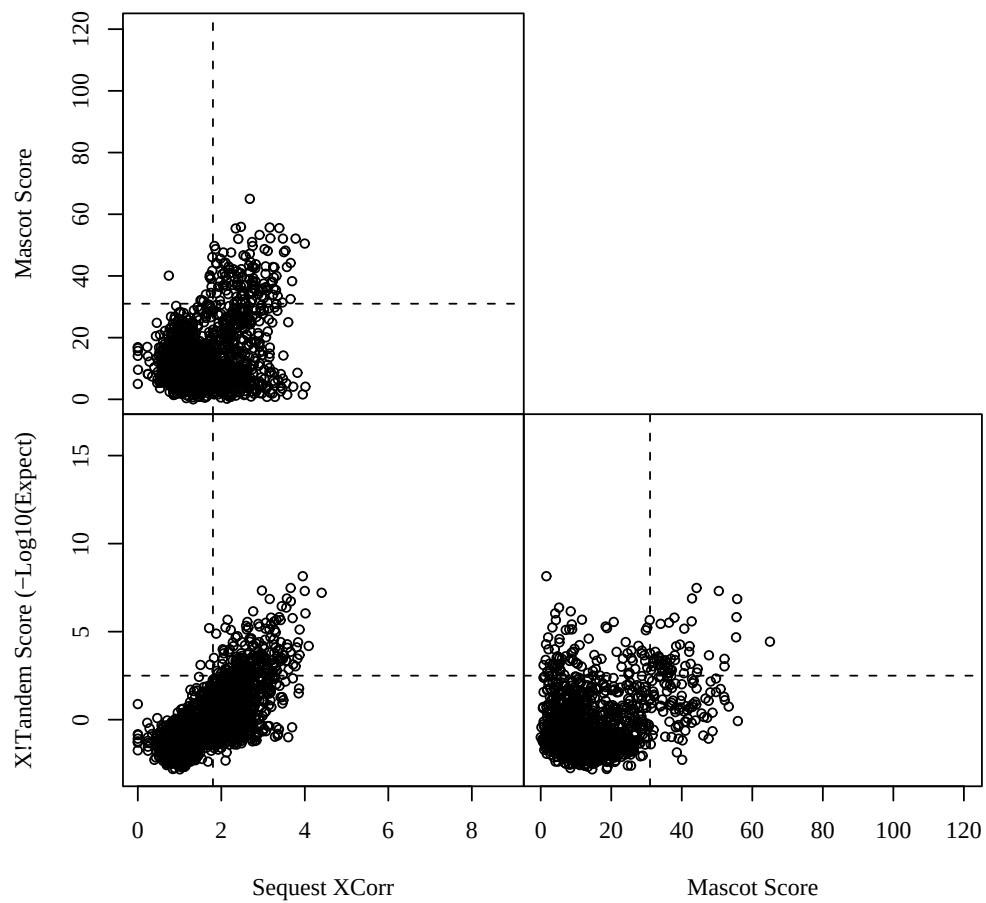


Figure 1.4: Comparison of scores irrespective of peptide assignment, $z=1$. Dotted lines represent 1% false-positive cut-offs.

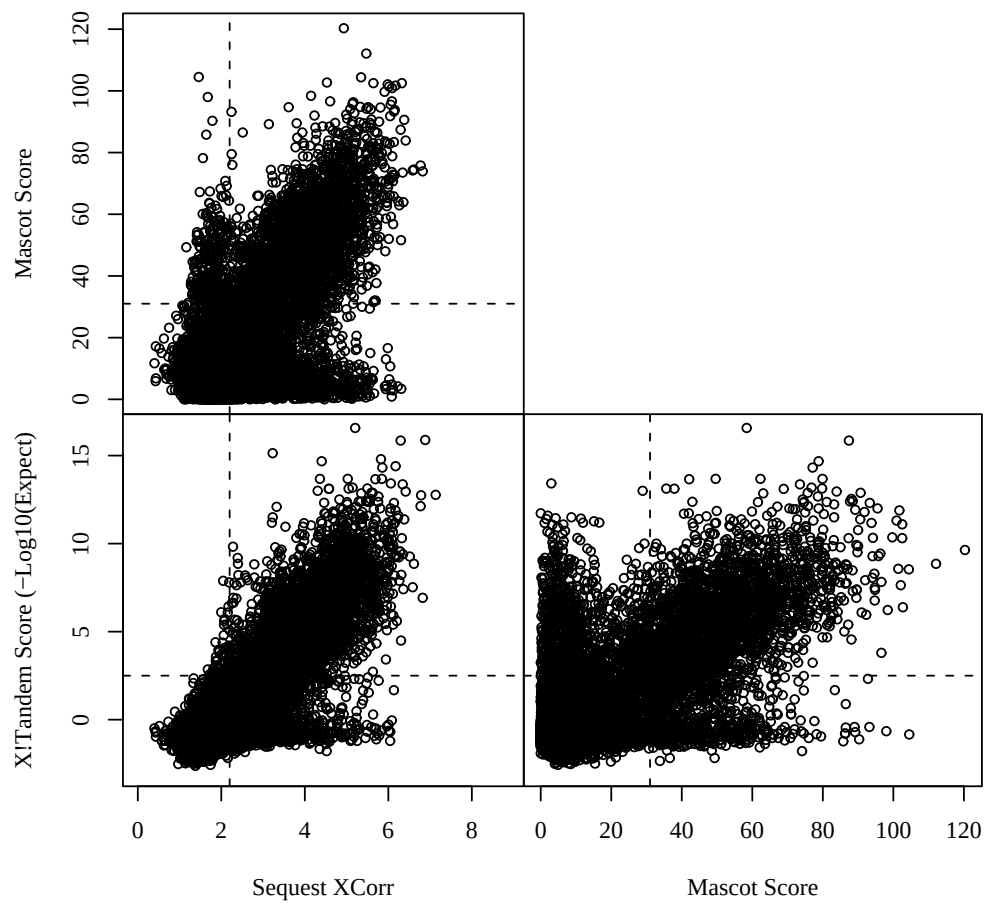


Figure 1.5: Comparison of scores irrespective of peptide assignment, $z=2$. Dotted lines represent 1% false-positive cut-offs.

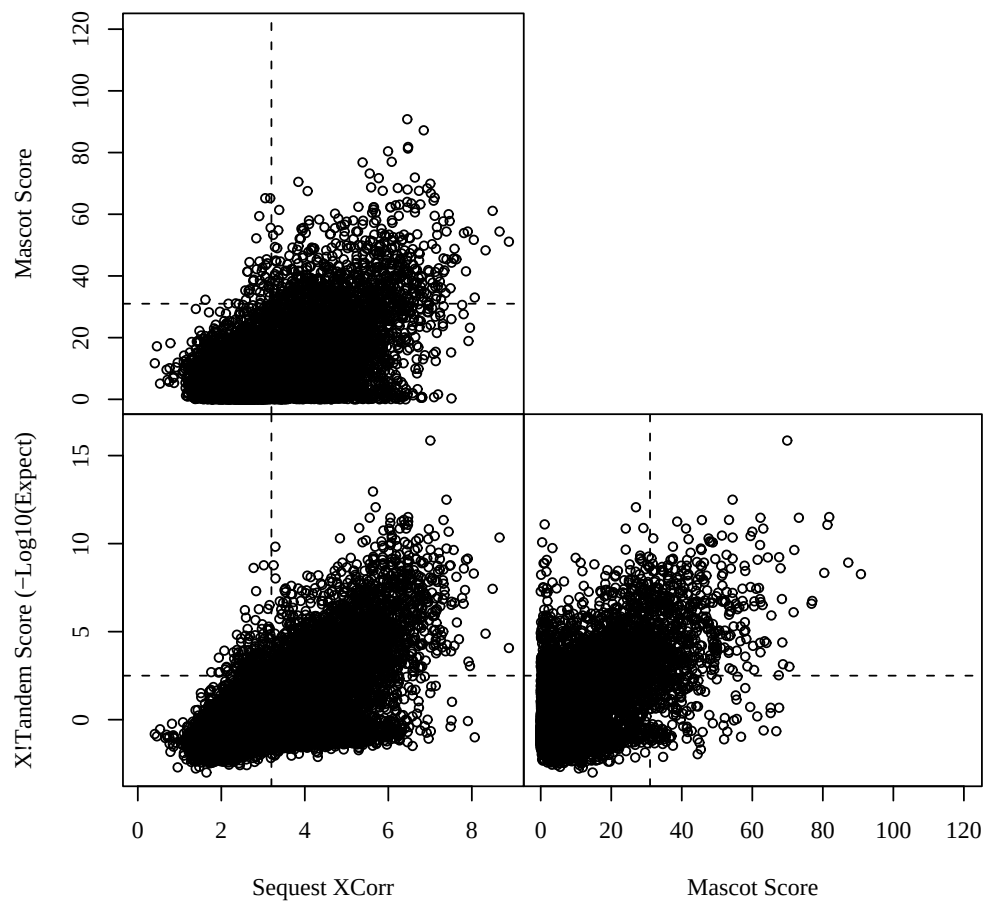


Figure 1.6: Comparison of scores irrespective of peptide assignment, $z=3$. Dotted lines represent 1% false-positive cut-offs.

The existence of these disparities illustrates the possibility of false positives from any of the algorithms and suggests that application of more than one may be a means of eliminating them.

The previous comparisons were on the basis of raw scores only. While they reflect the way each algorithm scores common spectra, it may not be informative of their utility in protein identification since additional parameters are typically applied by each to improve their accuracy. Sequest reports the ΔCN , for example, which is the difference in the top two Xcorr values for a given spectrum, with a larger difference representing a better match.

To compare the results of the algorithms at the protein level, parameters for each were found that yielded a false positive rate of $<1\%$ when searching a subset of the spectra against a database with reversed sequences appended. For all algorithms, only fully tryptic peptides were considered. This was the single-most important factor in reducing false positives. For the Sequest parameters given below that return 1% false positives, the rate jumped to about 35% if non-tryptic peptides were allowed. For Sequest the parameters were $\Delta\text{CN}>0.08$, $\text{XCorr}>1.8$ ($z=+1$), 2.2 ($z=+2$), and 3.2 ($z=+3$). For Mascot a peptide was accepted if the score exceeded the identity score and it was the top hit. For X!Tandem, $-\log_{10}(\text{expectation score})$ exceeding 2.5 yielded less than 1% false positives. These parameters were then used to filter the peptide identifications of the entire set of mass spectra searched against the real sequences only. The algorithms were thus ‘normalized’ to the same false-positive level and the identified proteins sets compared. This comparison is presented in Figure 1.7 for both a one-peptide and a two-peptide identification criteria. The union of these data is 795—416 (one-peptide—two-peptide). Only one-third (267/795) is identified by all three algorithms. This fraction improves only slightly to 42% (178/416) if a more conservative two-peptide criterion is applied. For proteins identified by a single algorithm, the fractions were 43.0% and 28.1%. In keeping with the score correlations in Table 1.1 Sequest and X!Tandem had the most protein identifications in common, and Mascot and X!Tandem the least. Sequest had the most proteins not identified by the others; X!Tandem the least. This is somewhat surprising given that these two algorithms had the highest correlations in scores. On the other hand, it does fit with the observation that X!Tandem appears to have a lower fraction of peptides with passing scores rejected by the other algorithms (upper left quadrant of Figures 1.4-1.6). Thus the protein data comparison is in general agreement with the previous scores comparison: a relatively small fraction of the proteins are identified by all three algorithms and the pairwise agreement follows the correlation in the scores.

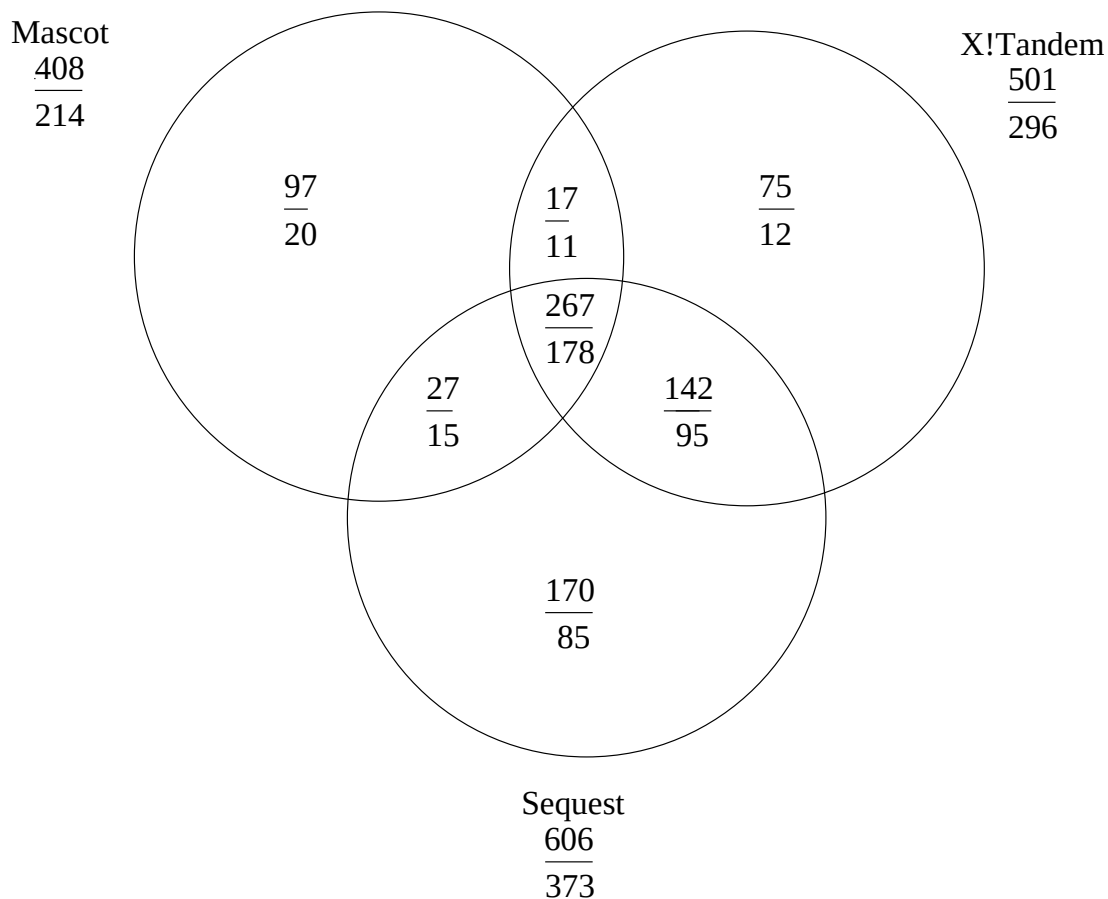


Figure 1.7: Number of proteins identified by the algorithms. Each algorithm was parameterized to yield a 1% false-positive rate. The top number of each pair is a one-peptide criterion; the bottom number is for two or more peptides required for identification.

1.5 Discussion

The goal of this experiment was to ascertain whether the three algorithms, Sequest, Mascot, and X!Tandem produced similar protein identifications from searching the same set of mass spectra. Despite tuning to the same false positive rate, the numbers of proteins identified varied widely and the number identified by all three algorithms was relatively low, 33.6% or 42.8% of the total identified for a one- or two-peptide identification criterion. Sequest identified the most proteins and also had the highest number uniquely identified. X!Tandem had the fewest proteins uniquely identified and the highest number shared with the other two algorithms. Mascot identified the fewest proteins and had an intermediate level of unique identifications. Thus it would appear that Sequest is the most sensitive and X!Tandem the most accurate. This conclusion is weakened however by the largest defect in this experiment, namely the failure to determine the true positives. This makes it impossible to conclude, for instance, whether the high number of unique identifications by Sequest is high specificity (correct detection of true positives) or sensitivity (correct rejection of true negatives).

A recent study of four algorithms, including the same three algorithms as here, did address this[59]. Using 3952 spectra from a single one-dimensional LC/MS/MS run of an immunoglobulin-depleted human plasma sample, these authors assumed that peptides that were identified by all the algorithms were true positives and manually interpreted the others. Their results were presented only as peptide identifications, not protein identifications. They find a higher fraction of peptides identified by all four algorithms (55%), than we did for three (10%). The fraction of peptides uniquely identified was much lower than the similar value we determined for proteins, 12% instead of 43% (single-peptide criterion). The order of these was the same as that for proteins in this study however; Sequest had the highest level of unique identifications and X!Tandem the lowest. In general our comparison agrees with that of Kapp *et al.* with the major disparity being the number of common and unique identifications. This may be due to the difference in the experiments. Their data set was less than 4000 spectra from a single one-dimensional LC/MS/MS analysis, while ours was more than 50,000 spectra from five experiments including three two-dimensional experiments. It is therefore likely that we have a higher variation in spectral quality and in peptide diversity.

A second deficiency in this study was the difference in search parameters employed. The Sequest search was done with 4 missed cleavages allowed, the other two were done with two. This might have contributed to the higher number of unique peptides assigned by Sequest. In practice, tryptic peptides with larger number of missed cleavages result

in +3 or higher charge states and yield poor spectra. These would be removed in the filtering process, which means the protein identifications probably are not affected by this difference. The fragment ion tolerance also differed. The Sequest search was conducted with this parameter set to zero, for Mascot and X!Tandem it was 0.8 and 0.4 respectively. This may have affected the sensitivity of each algorithm.

In conclusion we find that the protein identifications by Sequest, Mascot, and X!Tandem vary considerably and the common identifications are fairly low. This and the correlations of the raw scores indicate that the algorithms have different sensitivities and probably interpret the spectra in different ways. This is both a curse and a blessing. From this experiment we cannot conclude which algorithm is ‘correct’, but differential responses imply that a combination of the algorithms might yield a more accurate interpretation of a collection of spectra. In the analysis of the chloroplast proteome in Chapter 2 we evaluate such an approach. An extension of this experiment might be to optimize the algorithms collectively, rather than individually, and analyze the sensitivity and specificity.

Chapter 2

Chloroplast Proteomics

2.1 Abstract

The shotgun proteomic analysis of agar-grown *Arabidopsis thaliana* chloroplasts is reported. Three hundred twenty-seven proteins are identified, some of which may be indicative of growth in the high sucrose agar. Most of the proteins were found to have homologues in the Cyanobacteria, but there were significant contributions from α -proteobacteria and fungi, reflecting the mosaic nature of the plastid proteome. The proteins were used to evaluate localization algorithms and found the TargetP accuracy to be as claimed, the Predotar somewhat less. Additionally, the proteins were used to evaluate a three-algorithm MS/MS spectral search strategy that appears to give a significant increase in protein identifications.

2.2 Introduction

Plastids are essential organelles of plants that arose through the endosymbiotic association of a primitive eukaryote and a cyanobacterium. The term 'plastid' (Greek $\pi\lambda\alpha\sigma\tau\acute{o}\varsigma$, in the sense moldable, plastic) denotes the remarkable ability of these organelles to assume specialized morphologies and biochemical functions in higher plants. In the green photosynthetic state, the chloroplast sustains autotrophophy by the light-driven production of carbohydrates, ATP, and NADPH. Even the chloroplast may be further differentiated in a tissue-specific or developmental manner. As chloroplasts age for instance, they tend to get larger and decrease in number, and their import characteristics change. In C4 plants, chloroplasts are further differentiated into bundle sheath and mesophyll forms, an adaptation which increases the capture of CO₂ in hot climates. Plastids may exist in other, non-green forms[104]. The etioplast is a colorless or yellowish form that exists in dark-

grown tissues and converts to the chloroplast upon exposure to light. A form specialized for the production of pigments in colored tissues such as petals is known as the chromoplast. The amyloplast is a colorless form specialized for starch storage. The term leucoplast is used for various colorless forms that may be specialized for lipid or protein storage, or have unknown function. While the structures and functions of many of the proteins in the chloroplast photosynthetic and Calvin cycles are well known, those of other forms or developmental states are less well known.

During the conversion of the endosymbiotic prokaryotes to the mitochondria and plastid, most of the organelle DNA and coding capacity was exported to the eukaryotic nucleus[2, 89]. Most organelle proteins are now translated on cytosolic ribosomes and targeted to their destination by an N-terminal peptide sequence which is usually cleaved after import via protein translocons. The proteins are not exclusively returned to their organelle of origin however. A protein may be directed to either or both organelles. The complement of amino-acyl tRNA synthases in particular seem to have been reduced in number due to dual import, perhaps as a means to exert control over translation[28]. Proteins may also be directed to another location altogether, as exemplified by the targeting of plastid rps13 to the mitochondrion in *Arabidopsis*[97]. It has been estimated that about 18% of the *Arabidopsis thaliana* proteome is derived from cyanobacteria, and that most of these proteins are not targeted back to the plastid[91]. Estimates of the size of the chloroplast proteome have ranged from 600 to 6000[1]. Using the availability of the *Arabidopsis* genome, Bruce estimated the proteome size at about 3500 based on the number of proteins with a predicted chloroplast transit peptide[15]. Such estimates however, depend on the reliability of the prediction algorithms and may miss cytosolically produced proteins that are directed to the plastid outer envelope and do not require targeting peptides. There are also proteins that are imported without using the standard TOC/TIC import pathway, a recent example being α -carbonic anhydrase which is apparently imported via vesicle fusion after passing through the endoplasmic reticulum and being N-glycosylated[156].

As the number of eukaryotic genomes increases, the prediction of their localization remains an important area of research. Indeed, the number of algorithms seems to be increasing faster than the number of genomes. A recent review lists 18[130] and several new ones appear every year (Table 2.1). These prediction tools use machine learning methods such as neural nets[29, 139], support vector machines[9, 50], and k-nearest neighbor discrimination[103, 135] trained on proteins with known localization. Many use the N-terminal sequence for localization, others may include evolutionary and structural information [102], motifs[49, 133], dipeptide frequencies[9] or other sequence-based features.

Table 2.1: Novel localization prediction algorithms cited in PubMed.

Year	Number
2006	9
2005	5
2004	9
2003	4

Besides the localization, another desirable feature of these algorithms would be a prediction of the targeting sequence length. TargetP is apparently the only current algorithm which predicts the cleavage site of cleavable targeting signals in addition to localization. The performance of any of these classifiers will depend on the size, quality, and diversity of the known examples used in their training and validation sets. High-throughput technologies such as proteomics can enable better localization prediction by providing large sets of proteins with known localization for algorithm training, testing and refinement. The definition of the cleavage site with a high-throughput method, inferred through determination of the intact mass of the mature protein or by determination of the sequence of the N-terminal peptide, would also be desirable.

Since the publication of the *Arabidopsis* genome, there have been several proteomic analyses of the chloroplast. Most of these have been determinations of membrane fractions or other sub-compartments[40, 44, 69, 126]. While this work was being completed, a whole chloroplast proteome study identifying 690 proteins was reported[70]. This study found a remarkably low fraction of proteins with a TargetP-predicted transit peptide. It was criticized in a later review[155] wherein it was observed that Kleffmann et al. detected proteins that were also detected in other studies had a TargetP predicted chloroplast localization of 70-92%. The Predotar predicted chloroplast localizations were similarly high[155]. For the Kleffmann et al. proteins *not* found in other studies, the TargetP predictions were only 37%. These results imply a high level of false positives or contamination from other cell compartments. A more recent study compared the proteomes of bundle sheath and mesophyll chloroplasts isolated from *Zea mays* using both two-dimensional gel separations followed by MALDI identification and LC/ESI-MS/MS[86]. Isotope-coded affinity tag (ICAT) labeling was employed for a quantitative comparison of the proteins in the differentiated plastids. Four hundred proteins were identified and their relative location and quantity were related to the functional specializations of the bundle sheath and mesophyll chloroplasts.

In this chapter we present the determination of the proteome of chloroplasts isolated from the model plant *Arabidopsis thaliana* using 'bottom-up' proteomics and use the

protein set to evaluate the performance of the localization algorithms TargetP[29] and Predotar[139]. Additionally, we apply the conclusion of Chapter 1 that MS search algorithms have different performance by combining them as a 'meta-search' tool. We also use the data in Chapter 3 to test our heuristic motif analysis.

2.3 Materials and Methods

2.3.1 Plant production

Arabidopsis thaliana, Columbia, was grown on sterile agar in Magenta boxes essentially as in Fitzpatrick and Keegstra[34]. This method of plant production gives a high yield of chloroplasts, but may be plagued by fungal or bacterial growth. Pre-sterilization of the Magenta boxes and additional treatment of contaminated boxes with 1.5% hypochlorite bleach kept the number of failed boxes to about 1%. Agar was prepared by mixing 8 g agar, 10 g sucrose, and 4.41 g of Murashige and Skoog media in one liter of water. This was boiled until mostly clear, and approximately 40 mL poured into pre-sterilized Magenta boxes. The filled boxes were autoclaved and allowed to cool. Approximately 300 mg of *Arabidopsis* seeds were surface sterilized by gentle mixing for 15 minutes in a 15 mL Falcon tube containing 5 mL of 1.5% hypochlorite bleach and 0.02% Triton-X100. All further operations were carried out in a sterile hood. The seeds were washed three times with sterile water and re-suspended in 8 mL of sterile water. Using a sterile P1000 tip trimmed to a diameter large enough to admit the seeds, approximately 300 μ L of the seed suspension was evenly applied to the surface of the agar in each box. Approximately 24 boxes could be planted with 1 L of media and 300 mg of seeds. The boxes were distributed in an environmental growth chamber maintained at 22°C and 100 μ E light with a light/dark cycle of 18/6 hours. Plants were harvested at 21 days (bolting occurs after about 25 days with these conditions). Plants grown under these conditions were about one inch high with leaves approximately 1 mm long.

2.3.2 Organelle isolation

All operations were carried out on ice or in the cold room at 4°C. The agar was freed from the box with a spatula and the plant plug popped out. Plants were cut at the level of the agar with scissors, being careful not to include any agar. About 80 g of wet tissue was obtained from 24 Magenta boxes of 21 day old plants. The harvested plants were diced with a new razor blade in a large glass Petri dish with grinding buffer (100 mM HEPES,

660 mM sorbitol, pH 8.0). This brie was transferred with additional grinding buffer (total volume about 3 times that of the brie) and gently ground in a cold mortar. The brie was sluiced onto a filter consisting of one layer of Mira cloth and 2 layers of cheese cloth. The brie was gently squeezed to enhance recovery. The filtrate was spun at 2000 *g* for 6 minutes. The supernatant was discarded and the dark green pellet gently re-suspended in import buffer (100 mM HEPES, 600 mM sorbitol, 3 M MgCl₂, 3 M MnCl₂, 0.5 M EDTA, pH 7.3) using a paint brush until homogeneous. The chloroplast homogenate was layered atop a previously prepared Percoll gradient (35 mL 50% Percoll in import buffer, 37,000 *g*, 30 minutes) and spun at 6000 *g* for 14 minutes. This pellet was re-suspended in fresh import buffer, pelleted, and placed atop a fresh 50% Percoll gradient. The intact band was taken, diluted with import buffer and pelleted at 2000 *g*. The final pellet was re-suspended in import buffer at a concentration of about 1 mg/mL chlorophyll.

2.3.3 Mass spectrometry

Chloroplast suspensions were made 90% in -20°C acetone and kept at -20°C for 30 minutes. The suspension was centrifuged at 5000*g* and the green supernatant removed. The pellet was solutioned with 6 M guanidine-HCl, 50 mM Tris, 5 mM DTT, pH 8.0 and heated at 60°C for 60 minutes. The solution was diluted with at least 5 volumes of 50 mM Tris and sequencing grade trypsin (Promega) was added at a ratio of 1:50 (trypsin:protein). The sample was tumbled overnight at 37°C . Another aliquot of trypsin was added and digestion continued for 4 hours. The digest was then made 10 mM in DTT and heated at 60°C for 60 minutes. The digest was desalted on a C₁₈ cartridge (Waters) with water/0.1% TFA; the peptides were eluted with acetonitrile/0.1% TFA. The acetonitrile extracts were concentrated on a Speed-Vac and passed through a 400 μm spin filter (Millipore). Samples were made up in 0.1% TFA for one dimensional chromatography or 0.1% formic acid for two dimensional chromatography.

Samples were analyzed by both one- and two-dimensional LC/MS/MS using an LC Packings HPLC system connected to an LCQ DECA XP mass spectrometer (Thermo-Finnian). For one-dimensional HPLC, the sample was injected on a 300 μm id by 15 cm, 300 Å (Vydac) column using the autosampler. A gradient of 100% solvent A (95% H₂O, 5% ACN, 0.5% formic acid) to 100% solvent B (95% H₂O, 5% ACN, 0.5% formic acid) was applied at 4 $\mu\text{L}/\text{min}$ for 120 minutes. The column was connected to the electrospray source with 100 μm id fused silica. The electrospray voltage was operated at 4.5kV. Three separate runs employing smaller *m/z* ranges spanning 400 *m/z* to 2000 *m/z* were made. The mass spectrometer was operated in data-dependent mode with dynamic exclusion enabled; the

top 4 peaks for each scan were selected for MS/MS analysis.

For two dimensional HPLC, the sample was loaded onto an LC Packings SCX column (500 μm id x 15 mm) in valve A of the Switchos valve controller. Peptides eluting from the SCX column were trapped on a C_{18} pre-column (300 μm x 5 mm, 300 Å) on valve B. After desalting, the valve was switched to direct the flow to the analytical column (75 μm id x 15 cm, 300 Å, 5 μm , Grace Vydac). After each salt bump a gradient of 100% solvent A to 100% solvent B was applied at 200 nL/min for 2 hours. Eight ammonium acetate salt bumps covering the range 25 mM to 2 M were applied. The nanospray source was operated at 2.0 kV and a single mass range was scanned from 300 to 1700 m/z.

2.3.4 Mass spectrometric data analysis

The *Arabidopsis thaliana* predicted proteome was downloaded from The Arabidopsis Information Resource (TAIR, www.arabidopsis.org) in August 2004. A decoy database was constructed by reversing each sequence and appending these to the forward sequences[109]. A subset of the mass spectra, about 10% of the total, were searched against the combined data base with Sequest. Parameters were adjusted to yield a 1% false positive rate(Chapter 1). For the Sequest-only identifications (denoted 'Seq2') these were: XCorr 1.8 (+1 m/z), 2.2 (+2 m/z), and 3.2 (+3 m/z), $\Delta\text{CN} > 0.08$ and fully tryptic peptides. Two peptides were required for a peptide identification. DTASelect and Contrast were used to collate the data for the Sequest-only analysis[148].

Proteins were also identified by combining the results of the three search algorithms Sequest, Mascot, and X!Tandem. Each algorithm was first parameterized to yield a 1% false positive rate as described in the previous paragraph and in section 1.3, p.18. The results for all the spectra, scored by all three algorithms were stored in a Postgres database and queries conducted with SQL and Perl. A protein was identified if: two or more constituent peptides passing the individual thresholds were identified by any single algorithm; or two or more algorithms assign the same peptide for the same mass spectrum. Thus a single peptide identification by a single algorithm, or peptides for which the algorithms disagree, are excluded. This search strategy is denoted 'Any2' for brevity.

2.3.5 Bioinformatics

BLAST analysis was done with WU-BLAST v.2.0[41] downloaded from <http://blast.wustl.edu/blast/executables> using the default settings. A database for inferring chloroplast protein phylogeny was constructed with representatives of the bacterial orders avail-

able from NCBI in December 2004. The genomes used were: Actinobacteria (*Mycobacterium tuberculosis*, *Streptomyces coelicolor*), Aquificae (*Aquifex aeolicus*), Archaeoglobaceae (*Archaeoglobus fulgidus*), Bacteroidetes (*Bacteroides thetaiotaomicron*, *Chlorobium tepidum*), Chlamydeae (*Chlamydophila pneumoniae*), Crenoaeraeota (*Sulfolobus solfataricus*), Cyanobacteria (*Gloeobacter violaceus*, *Prochlorococcus marinus*, *Synechocystis* sp., *Synechococcus* sp., *Nostoc* sp., *Thermosynechococcus elongatus*), Firmicutes (*Bacillus anthracis*, *Bacillus subtilis*, *Clostridium tetani*, *Mycoplasma pneumoniae*, *Staphylococcus aureus*), Fungi (*Saccharomyces cerevisiae*), Fusobacteria (*Fusobacterium nucleatum*), Halobacteria (*Halobacterium salinarum*), Methanococci (*Methanocaldococcus jannaschii*), Methanomicrobia (*Methanosarcina acetivorans*), Methanopyri (*Methanopyrus kandleri*), Nanoarchaeota (*Nanoarchaeum equitans*), Plantomycetes (*Rhodopirellula baltica*), Proteobacteria (*Bordetella bronchiseptica*, *Bradyrhizobium japonicum*, *Escherichia coli*, *Geobacter sulfurreducens*, *Helicobacter pylori*, *Rhodopseudomonas palustris*, *Rickettsia prowazekii*, *Shewanella oneidensis*), Spirochaetes (*Treponema pallidum*), Thermococci (*Pyrococcus furiosus*), Thermoplasma (*Thermoplasma volcanium*).

Predicted grand average hydropathy (GRAVY) and pI for identified proteins were obtained from the ‘Plastid Proteome Database’ (<http://ppdb.tc.cornell.edu/>)[35]. The ‘MapMan’[152] *Arabidopsis* protein annotations were downloaded from the same site.

Localization predictions of both the Seq2 and Any2 protein sets were made using the two algorithms TargetP[29] and Predotar[139].

2.4 Results

A bottom-up LC/MS/MS analysis of the *Arabidopsis* chloroplast proteome was performed. Chloroplasts were isolated from five separate plant sets grown on agar and purified on two Percoll gradients with intermediate washing steps. The isolations appear to be very pure. Of 51 ribosomal proteins identified, 12 are plastid-encoded and 22 have a plastid annotation. Of the remaining 17, 15 are predicted by both TargetP and Predotar to be plastid localized. One of the remaining pair is predicted to be located in the mitochondrion by both algorithms; the last is predicted to be cytosolic by TargetP and mitochondrial by Predotar. The lack of any cytosolic ribosomal proteins implies a clean preparation largely free of cytosolic contamination. Some proteins with predicted mitochondrial localization were detected, as discussed below, but the absence of the abundant respiratory chain proteins suggests that the preparation is free of major mitochondrial contamination as well.

2.4.1 MS Search Strategy Comparison

In Chapter 1 it was concluded that the three search algorithms Sequest, Mascot, and X!Tandem had a differential response, implying that a combination of them might give better sensitivity and accuracy than any alone. This was tested on the bottom-up proteomics results for chloroplast proteins, consisting of two one-dimensional and three two-dimensional LC/MS/MS analyses. The data were searched using all three algorithms and the results for each filtered using the criteria described in Chapter 1 for a 1% false positive rate. For the Sequest-only results, a minimum of two tryptic peptides was required for a protein identification. For the combination search strategy, a peptide was assigned to a protein if: two or more constituent peptides were identified by any single algorithm; the peptide was assigned to the same spectrum by two or more algorithms.

For the relatively conservative Sequest search criterion (denoted ‘Seq2’), 327 non-redundant proteins were detected. These are listed in Appendix A and their physico-chemical properties summarized in Table 2.2. Since the chloroplast proteome is not known with certainty, it is not possible to calculate the values in Table 2.2 for all proteins (as might be done with a bacterial predicted proteome, for example see Table 4.2) to estimate how well the proteome is sampled. It is evident however that hydrophilic proteins are preferentially sampled, since the GRAVY scores are shifted to negative values. GRAVY values as high as 0.5 are found, but some of the photosystem proteins have GRAVY as high as 1.5 and these are not detected. Likewise proteins with up to 11 transmembrane helices are detected, but most have zero or one (Table 2.4).

Using the combination search (denoted ‘Any2’) 421 non-redundant proteins are identified, an increase of 29%. There were 307 consensus identifications, 114 identified only by the Any2 strategy, and 20 identified only by the Seq2 search. The proteins found by the Any2 search are listed in Appendix B and summarized in Table 2.3. The median data for all the parameters are almost identical to the Seq2 search, with the largest difference being an increase in the number of hydrophobic proteins detected. In all, 10 additional proteins were found with GRAVY greater than zero. There is also a slight increase in the number of membrane proteins found (average TMHMM, Table 2.4).

The pseudo-2D gel plot in Figure 2.1 shows very similar dispersion of the protein sets found by each algorithm. The bimodal pI distribution characteristic of bacteria is also evident in the plastid proteome(cf. Chapter 4, p.85 and ref.[132]). Since most of the proteins detected were soluble (as indicated by the relatively lower numbers with predicted transmembrane helices) this probably reflects the necessity of proteins to have a pI different from the stromal or lumenal pH to maintain solubility.

Table 2.2: Sequest Two Peptide Protein Identifications^a

	GRAVY	pI	MW	TMHMM
mean	-0.178	7.6	39,900	0.24
median	-0.169	7.7	34,940	0
min	-0.989	4.3	7,250	0
max	0.571	12.2	241,900	11

^aData are summaries of predicted values downloaded from ppdb.
tc.cornell.edu[35].

Table 2.3: Combined Algorithm Prediction^a

	GRAVY	pI	MW	TMHMM
mean	-0.185	7.7	40,300	0.25
median	-0.175	7.9	34,560	0
min	-0.989	4.3	7,700	0
max	0.998	12.2	241,900	11

^aData are summaries of predicted values downloaded from ppdb.
tc.cornell.edu[35].

Table 2.4: TMHMM Predictions for the Two Search Strategies

TMHMM	Seq2	Any2
0	314	387
1	12	15
2	7	8
3	0	3
4	1	4
5	2	2
6	1	1
7	1	1
8	0	0
9	2	2
10	0	0
11	1	1

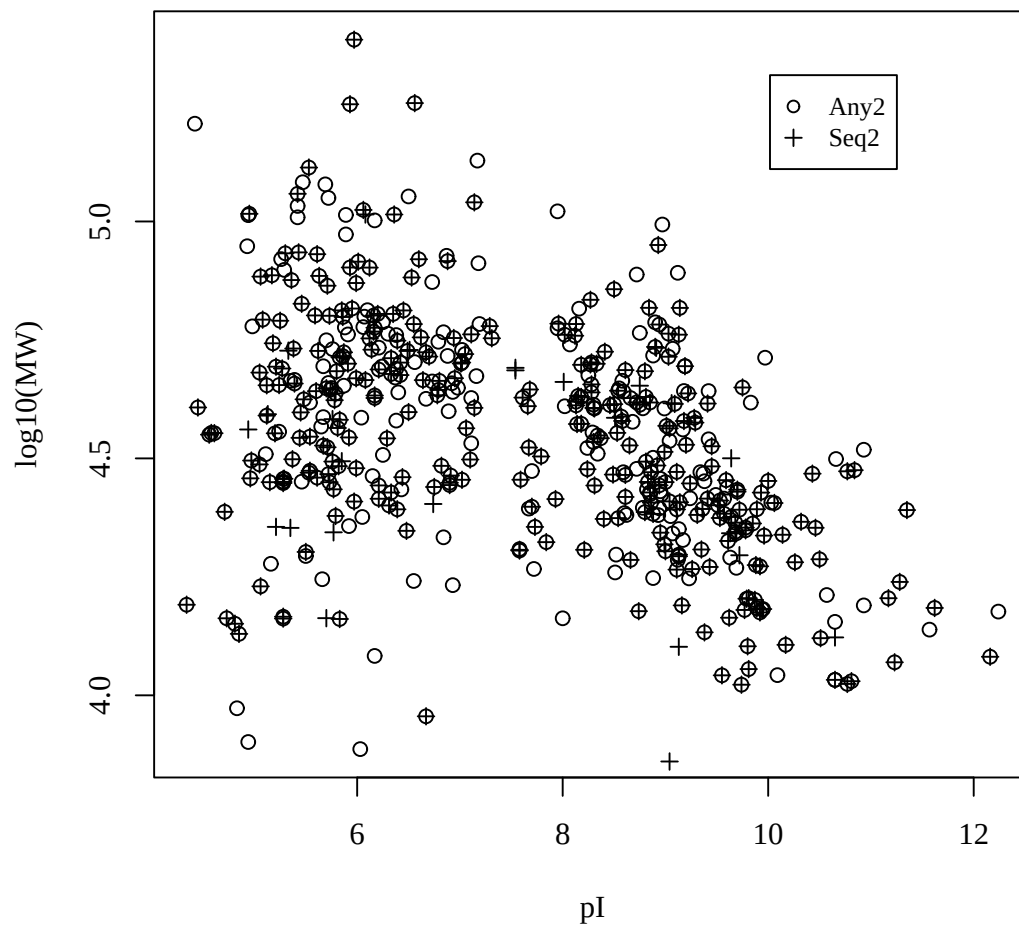


Figure 2.1: Pseudo-gel plot of chloroplast proteins.

The 114 additional proteins identified by the Any2 search are listed in Table 2.5. Most of these have a predicted chloroplast localization and have annotations consistent with a chloroplast function. There are 3 proteins that were predicted by both TargetP and Predotar to have a mitochondrial localization; all are annotated ‘putative’: At1g11860, aminomethyl transferase; At4g33010 glycine dehydrogenase; and At5g08680, ATP synthase β -chain. Another ATP β chain component with mitochondrial localization was detected by both search strategies (At5g08670). Thus, the identification of mitochondrial ATP synthase might be due to either mitochondrial contamination or detection of a protein with a potential dual localization. In this category, the protease At3g19170, identified only by the Any2 search is known to be targeted to both the mitochondrion and the chloroplast[10]; it had a chloroplast localization by TargetP and mitochondrial by Predotar. Similarly, the two tRNA synthases detected only by Any2 had split localization predictions. Many of the plant tRNA synthases are known to have dual localization as well[28].

Cellular localization for each protein set were predicted using the algorithms TargetP and Predotar and are summarized in Table 2.6 (data for plastid-encoded proteins were removed). The predictions by TargetP for the two sets of proteins were identical. Predotar returned a lower fraction of chloroplast predictions for the Any2 proteins; this difference was reflected in a decrease for chloroplast proteins predicted by both algorithms. Some of the predicted mitochondrial proteins were discussed above. Here we note that there are several proteins detected by the Any2 search that are probably outer membrane proteins. These lack an N-terminal targeting signal and thus will not have a plastid localization prediction. In this category are At3g08580, ADP/ATP carrier, At4g02510, chloroplast outer membrane protein, and At5g05000, TOC34, all of which have a predicted cytosolic localization by both TargetP and Predotar.

In conclusion, it appears that the ‘Any2’ search strategy achieved a 30% increase in sensitivity. The slight decrease in chloroplast fractional localization by Predotar may indicate an increase in false positives, but this appears to be slight.

2.4.2 Localization prediction

If the argument that the chloroplast preparation is clean is accepted, the proteomic results may be used to evaluate the performance of the localization predictors TargetP and Predotar. The data in Table 2.6 are in good agreement with the 85% sensitivity originally claimed for TargetP[29]. This result also agrees with an estimate of TargetP sensitivity obtained in a proteomic study of thylakoid lumen proteins[146]. These authors estimated the TargetP sensitivity to be between 85% and 90%. However, Kleffmann, et al. in their

Table 2.5: Proteins Identified by ‘Any2’ Not Represented in ‘Seq2’
Results

Locus	H ^a	T ^b	P ^c	G ^d	MW ^e	pI	Annot
At1g02560.1	0	C	C	-0.197	32.35	8.34	ATP-dependent Clp protease proteolytic subunit
At1g03630.1	0	C	C	-0.351	43.88	9.18	protochlorophyllide reductase C, chloroplast
At1g06950.1	0	C	C	-0.300	112.12	5.72	Encodes a protein thought to be a part of the
At1g08520.1	0	C	C	-0.326	83.28	5.26	magnesium-chelatase subunit chlD, chloroplast,
At1g11750.1	0	C	C	-0.128	29.38	9.37	ATP-dependent Clp protease proteolytic subunit
At1g11860.1	0	M	M	-0.263	44.44	8.55	aminomethyltransferase, putative, similar to
At1g12250.1	0	C	C	-0.333	30.06	8.72	thylakoid lumenal protein-related, weak
At1g17650.1	0	C	M	0.171	37.78	8.68	similar to 6-phosphogluconate dehydrogenase
At1g18170.1	0	C	C	-0.268	26.53	8.96	immunophilin FKBP-type peptidyl-prolyl
At1g21270.1	0	S		-0.300	81.64	7.18	wall-associated kinase 2 (WAK2), identical to
At1g21350.1	0	S		-0.284	17.08	6.93	expressed protein
At1g29930.1	0	C	C	0.020	28.24	5.46	chlorophyll A-B binding protein 2, chloroplast
At1g52670.1	0	C		-0.220	29.57	8.59	biotin/lipoyl attachment domain-containing
At1g56050.1	0	C	C	-0.010	45.62	5.72	GTP-binding protein-related, similar to
At1g56070.1	0	-		-0.227	93.89	5.89	elongation factor 2, putative EF-2, putative,
At1g58290.1	0	C	C	-0.198	59.51	7.95	glutamyl-tRNA reductase 1 GluTR (HEMA1),
At1g63970.1	0	C	C	-0.031	24.81	8.94	2C-methyl-D-erythritol 2,4-cyclodiphosphate
At1g64510.1	0	C	C	-0.305	22.76	5.92	ribosomal protein S6 family protein, similar to
At1g64770.1	0	C		-0.213	38	6.38	expressed protein
At1g65260.1	0	C	C	-0.561	36.39	9.17	PspA/IM30 family protein, contains Pfam PF04012:
At1g66430.1	0	C	C	-0.070	41.47	5.54	pfkB-type carbohydrate kinase family protein,
At1g67700.1	0	M		-0.475	26.01	9.57	expressed protein
At1g68010.1	0	-		-0.044	42.24	6.67	glycerate dehydrogenase NADH-dependent
At1g71500.1	0	C	C	-0.226	31.72	8.88	Rieske (2Fe-2S) domain-containing protein,
At1g74640.1	0	C	C	-0.221	41.05	8.3	expressed protein, contains Pfam profile: PF00561
At1g74880.1	0	C	C	-0.392	17.65	9.23	Encodes subunit NDH-O of NAD(P)H: plas-toquinone
At1g76100.1	1	C	C	0.221	17.58	5.66	plastocyanin, identical to plastocyanin
At1g79850.1	0	C	C	-0.313	16.28	10.57	30S ribosomal protein S17, chloroplast CS17
At1g80480.1	0	C	C	-0.331	49.48	5.67	PRLI-interacting factor L, putative, similar to
At2g04030.1	0	C	C	-0.551	88.66	4.93	heat shock protein, putative, strong similarity

^aTHMM; ^bTargetP; ^cPredotar ^dGRAVY; ^eMW * 10⁻³.

Continued on next page

Table 2.5: Continued

Locus	H ^a	T ^b	P ^c	G ^d	MW ^e	pI	Annot
At2g04400.1	0	C		-0.204	44.57	6.98	indole-3-glycerol phosphate synthase (IGPS),
At2g05620.1	0	-	C	-0.079	14.29	10.65	expressed protein
At2g26140.1	0	M		-0.169	77.27	8.72	encodes an FtsH protease that is localized to the
At2g28820.1	0	M		-0.545	32.98	10.93	similar to alanine aminotransferase, putative
At2g30170.1	0	C		-0.012	32.28	5.11	expressed protein
At2g31810.1	0	C	C	-0.048	53.87	9.07	acetolactate synthase small subunit, putative,
At2g40300.1	0	C		-0.268	29.02	6.15	ferritin, putative, similar to ferritin subunit
At2g41680.1	0	C	C	-0.313	57.95	6.3	thioredoxin reductase, putative NADPH-dependent
At2g44050.1	0	C	M	0.092	24.02	8.62	6,7-dimethyl-8-ribityllumazine synthase DMRL
At2g44920.1	0	C	C	-0.152	22.43	9.13	thylakoid lumenal 15 kDa protein, chloroplast,
At2g46820.1	0	C	C	0.07	18.48	7.72	expressed protein
At3g01440.1	0	C		-0.337	24.78	7.67	oxygen evolving enhancer 3 (PsbQ) family protein
At3g02730.1	0	C	C	-0.141	19.32	9.12	thioredoxin, putative, similar to SP:P29450
At3g03910.1	0	M		-0.200	44.52	5.75	glutamate dehydrogenase, putative, similar to
At3g08580.1	3	-		-0.120	41.47	9.83	ADP, ATP carrier protein 1, mitochondrial/
At3g14415.1	0	-		-0.031	40.3	8.99	(S)-2-hydroxy-acid oxidase, peroxisomal, putative
At3g15520.1	0	C	C	-0.135	50.48	6.56	peptidyl-prolyl cis-trans isomerase TLP38
At3g17810.1	0	C	C	-0.275	46.84	6.37	dihydroorotate dehydrogenase family protein
At3g19170.1	0	C	M	-0.320	121.01	5.47	Zinc metalloprotease pitrilysin subfamily A.
At3g20390.1	0	C	C	0.223	19.81	8.52	endoribonuclease L-PSP family protein, contains
At3g24430.1	0	C	C	-0.009	57.76	5.91	expressed protein, contains Pfam profile PF01883:
At3g27850.1	0	C	C	0.045	19.68	5.5	50S ribosomal protein L12-3, chloroplast
At3g27890.1	0	M		-0.028	21.55	6.84	NADPH-dependent FMN reductase family protein,
At3g32930.1	0	C	C	-0.361	27.43	9.6	expressed protein
At3g44620.1	0	C		-0.428	26.95	8.95	similar to protein tyrosine phosphatase-like
At3g44880.1	0	C	C	-0.407	60.75	7.19	Rieske (2Fe-2S) domain-containing protein,
At3g45140.1	0	C	C	-0.465	102.04	5.42	lipoxygenase (LOX2), identical to SP:P38418
At3g47650.1	0	C	C	-0.300	14.54	8	bundle-sheath defective protein 2 family bsd2
At3g48110.1	0	C		-0.152	119.77	5.69	similar to glycyl-tRNA synthetase alpha chain and
At3g51140.1	4	C		-0.016	31.53	10.66	expressed protein

^aTHMM; ^bTargetP; ^cPredotar ^dGRAVY; ^eMW * 10⁻³.

Continued on next page

Table 2.5: Continued

Locus	H ^a	T ^b	P ^c	G ^d	MW ^e	pI	Annot
At3g52930.1	0	-		-0.230	38.54	6.04	fructose-bisphosphate aldolase, putative, similar
At3g53460.1	0	C	C	-0.488	36	5.24	29 kDa ribonucleoprotein, chloroplast
At3g54090.1	0	C	C	-0.467	53.78	5.76	pfkB-type carbohydrate kinase family protein,
At3g59980.1	0	C	C	-0.252	29.74	7.7	tRNA-binding region domain-containing protein,
At3g61870.1	4	C	C	0.093	29.58	9.34	expressed protein, hypothetical protein
At4g01900.1	0	C	C	-0.136	21.27	9.17	P II nitrogen sensing protein (GLB I), identical
At4g02510.1	0	-		-0.479	160.82	4.42	chloroplast outer membrane protein, putative
At4g09040.1	0	C		-0.602	34.03	7.11	RNA recognition motif (RRM)-containing protein
At4g12060.1	0	C		-0.166	26.56	8.85	Clp amino terminal domain-containing protein,
At4g12830.1	0	M	C	-0.337	43.76	8.59	hydrolase, alpha/beta fold family protein, low
At4g13010.1	0	-		0.034	34.43	9.04	oxidoreductase, zinc-binding dehydrogenase family
At4g13430.1	0	C	C	-0.170	55.01	8.07	aconitase family protein, aconitate hydratase
At4g14210.1	0	C	C	-0.135	62.96	6.07	phytoene dehydrogenase, chloroplast phytoene
At4g18440.1	0	C		-0.149	59.75	5.88	adenylosuccinate lyase, putative /
At4g21210.1	0	C		-0.316	43.68	6.93	expressed protein, contains Pfam domain PF03618:
At4g25100.1	0	-		-0.290	23.79	6.05	superoxide dismutase (Fe), chloroplast (SODB)
At4g25130.1	0	C		-0.525	28.64	8.95	peptide methionine sulfoxide reductase, putative,
At4g25370.1	0	C		-0.273	26.05	9.24	Clp amino terminal domain-containing protein,
At4g25650.1	2	C	C	-0.340	61.26	8.9	Rieske (2Fe-2S) domain-containing protein,
At4g27700.1	0	C	C	-0.316	24.87	8.78	rhodanese-like domain-containing protein,
At4g33010.1	0	M	M	-0.140	112.92	6.5	glycine dehydrogenase (decarboxylating), putative
At4g33760.1	0	C	M	-0.193	74.51	6.73	tRNA synthetase class II (D, K and N) family
At4g34090.1	0	C	C	-0.140	36.91	5.65	expressed protein
At4g35630.1	0	C	C	-0.229	47.36	8.25	phosphoserine aminotransferase, chloroplast
At5g01650.1	0	-		0.203	12.11	6.17	macrophage migration inhibitory factor family
At5g05000.1	0	-		-0.151	34.7	9.42	translocate of chloroplast 34 (TOC34)

^aTHMM; ^bTargetP; ^cPredotar ^dGRAVY; ^eMW * 10⁻³.*Continued on next page*

Table 2.5: Continued

Locus	H^a	T^b	P^c	G^d	MW^e	pI	Annot
At5g05740.1	3	C		0.215	60.06	4.98	S2P-like putative metalloprotease, also contain
At5g08540.1	0	C	C	-0.404	38.65	5.67	expressed protein, similar to unknown protein
At5g08680.1	0	M	M	-0.143	59.86	6.06	ATP synthase beta chain, mitochondrial, putative,
At5g12040.1	0	C		-0.174	40.33	8.78	carbon-nitrogen hydrolase family protein, similar
At5g13120.1	0	C	C	-0.303	28.3	9.38	peptidyl-prolyl cis-trans isomerase
At5g14260.1	0	C	C	-0.273	57.58	6.38	SET domain-containing protein, low similarity to
At5g16660.1	1	C	C	-0.479	18.17	8.51	expressed protein
At5g17560.1	0	C	C	-0.201	19.53	9.63	BolA-like family protein, contains Pfam profile:
At5g19940.1	0	C	C	-0.158	26.48	9.49	plastid-lipid associated protein PAP-related
At5g22800.1	0	C		-0.245	107.83	5.42	aminoacyl-tRNA synthetase family protein
At5g35360.1	0	C	C	-0.135	58.38	6.84	acetyl-CoA carboxylase, biotin carboxylase
At5g35970.1	0	C		-0.156	105.04	7.95	DNA-binding protein, putative, similar to
At5g36700.1	0	C	C	-0.134	39.76	6.89	phosphoglycolate phosphatase, putative, similar
At5g38420.1	0	C	C	-0.211	20.35	7.58	ribulose biphosphate carboxylase small chain 2B
At5g38530.1	0	C	C	-0.229	55.74	6.79	tryptophan synthase-related, low similarity to
At5g42070.1	0	C	C	-0.684	17.68	8.88	expressed protein, similar to unknown protein
At5g42650.1	0	C	C	-0.231	58.19	8.75	allene oxide synthase (AOS) hydroperoxide
At5g43780.1	0	C	C	-0.310	52.12	8.88	sulfate adenylyltransferase 4, ATP-sulfurylase 4
At5g51110.1	0	C	C	-0.155	23.89	9.05	expressed protein
At5g54190.1	0	C	C	-0.166	43.86	9.42	protochlorophyllide reductase A, chloroplast
AtCg00140.1	2	C		0.998	7.97	4.94	ATPase III subunit
AtCg00580.1	1	C		0.046	9.39	4.83	PSII cytochrome b559. There have been many
AtCg00710.1	1	C		0.350	7.7	6.03	Encodes a 8 kD phosphoprotein that is a component
AtCg00730.1	3	C		0.551	17.43	6.55	A chloroplast gene encoding subunit IV of the
AtCg00750.1	0	C		-0.394	15.02	12.24	30S chloroplast ribosomal protein S11
AtCg00770.1	0	C		-0.336	15.48	10.93	chloroplast 30S ribosomal protein S8
AtCg00810.1	0	C		-0.621	18.58	9.69	encodes a chloroplast ribosomal protein L22

^aTHMM; ^bTargetP; ^cPredotar ^dGRAVY; ^eMW * 10⁻³.

Table 2.6: Localization Prediction for Different MS/MS Search Strategies^a

Localization Prediction	%Chloroplast Localization ^b	
	Any2	Sequest
TargetP	88.4	88.2
Predotar	67.0	75.7
Both ^c	72.3	74.1

^aSee text for complete description of the search strategies;

^bplastid-encoded proteins removed; ^cpercentage of loci with a chloroplast localization by both TargetP and Predotar.

reported chloroplast proteome determination, found only a 59% predicted plastid localization using TargetP[70]. The conflict with this and the thylakoid lumen study suggests a possible contaminated isolation. Kleffmann et al. also used a non-specific enzyme cleavage in their MS/MS database search, which may have returned a high level of false positive results. Predotar claims to have 82% sensitivity[139], but we find a lower 72.3% to 75.5% depending on the search strategy(Table 2.6). Our proteomic data therefore confirm the reported TargetP specificity, but find a somewhat lower value than cited for Predotar.

It has been reported that the application of multiple localization prediction algorithms gives more accurate results than any alone[123]. Table 2.6 also gives fraction of the experimental proteins for which both algorithms assigned a chloroplast localization. Seventy-four per cent of the Seq2 predictions have a chloroplast localization by both algorithms (the intersection). The Any2 search is lower at 68%. The results are recast in the Venn diagram of Figure 2.2. Of the Seq2 proteins assigned a chloroplast localization by only one algorithm, approximately half have annotations such as chlorophyll binding protein or porphyrin biosynthesis which suggest a correct chloroplast localization (data not shown). The proportions for the Any2 proteins is lower. Thus, it appears that the algorithms have different specificities, and the intersection, while increasing accuracy, may hurt sensitivity. If the union of the identifications is taken, the the chloroplast localization for the Seq2 proteins is 89% and for the Any2, 86%.

2.4.3 Conservation in Cyanobacteria

Since the chloroplast is derived from a cyanobacterial ancestor, it is expected that proteins detected in the chloroplast would have homologues in these organisms. To test this, the experimental proteomes were compared to a database of 6 cyanobacterial and 31 other

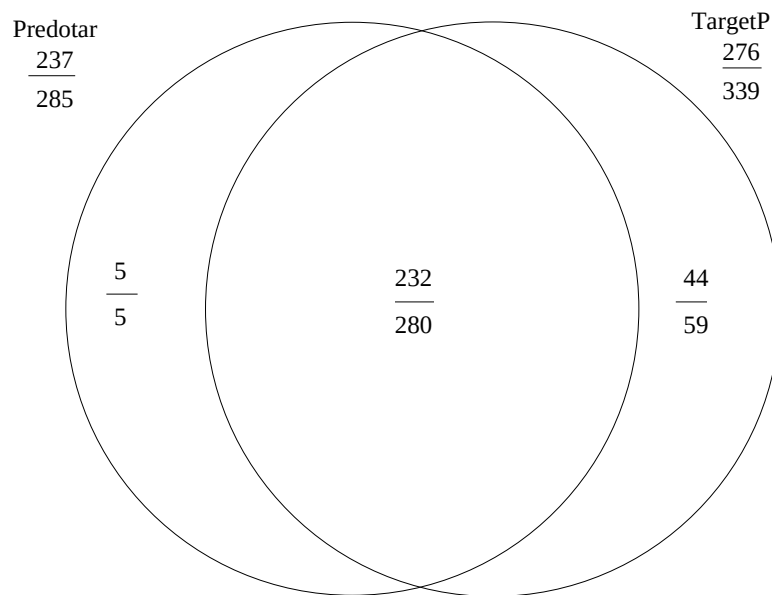


Figure 2.2: Comparison of localization predictors. Data are the number of proteins with a predicted chloroplast localization; the top number is for the Seq2 search, the bottom, Any2. Plastid-encoded proteins are removed.

bacterial proteomes using WU-BLAST. The yeast proteome was included to represent eukaryotic contributions to the plastid proteome. For the Seq2 protein set, 49 queries did not have a significant match at an expectation value of 10^{-6} ; 44 of the Any2 proteins did not have a significant match. This may be due to the limited size of the database searched and might be improved by including newly available genomes in the database. Results of the comparison are presented as a fraction of the number of queries in Figure 2.3. As expected, most of the chloroplast proteins had a top hit to a cyanobacterial protein, but the fraction of these hits was only 65%. The hits to cyanobacteria in the ‘Any2’ proteins was slightly lower, apparently due to an increase in the top-hits to yeast.

Interestingly, the next highest contributions after cyanobacteria are from the Proteobacteria and yeast, both at 11% of the top-hits. The proteobacterial homologues probably represent contributions from the mitochondrial proteome which was derived from an α -proteobacterial progenitor similar to *Rickettsia*, which was represented in the search database. Among these matches were a ribosomal protein (rps30), a heat shock protein, and several hypothetical proteins, but most were synthesis-related enzymes like kinases, aldolases, and lyases (Supplemental Table C). The annotations for the top hits to yeast proteins were similar, but include several that suggest functions for control of the organelle such as At1g07920 (EF-1 α), At1g21270.1 (spindle-pole associated kinase), At1g56070.1 (EF-2), and nucleolar proteins with RNA binding motifs (At2g37220, At3g52150, At3g52380,

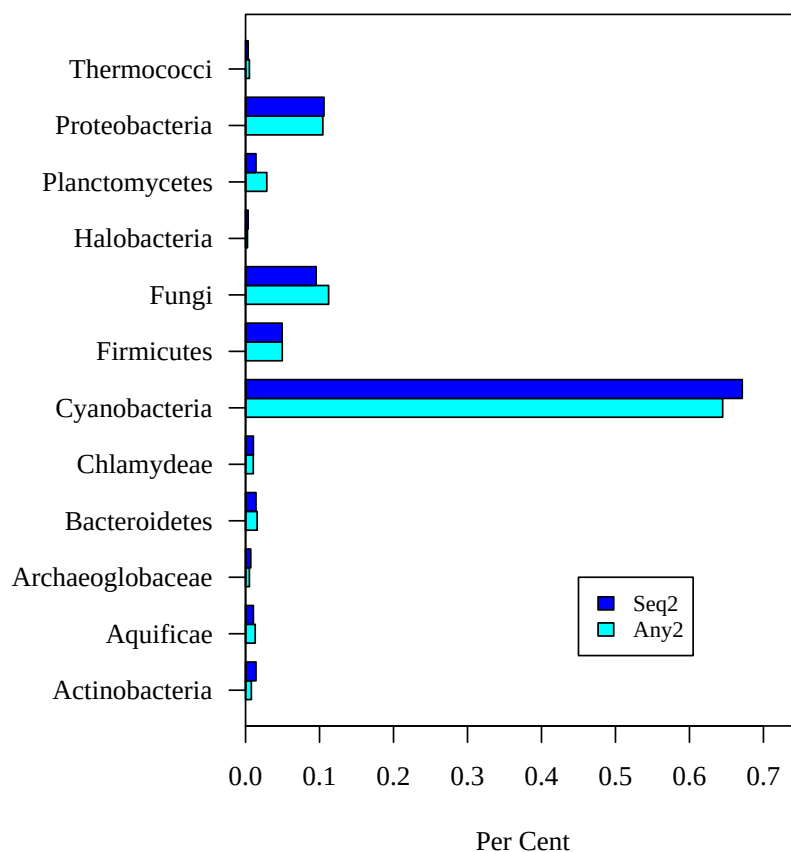


Figure 2.3: Proportion of top-hits to Cyanobacteria for different search strategies. $E < 10^{-6}$

At4g24770). Import machinery-related component components At4g02510 (Toc159, annotated hypothetical in yeast), and the HSP70s At2g30170 and At3g09440 also had top hits to yeast. The triose-phosphate transporter, At5g44110 (also hypothetical in yeast) indicates not so much control, but mutualism by extraction of photosynthate, a phase that probably preceded full symbiosis. The next highest fraction of top-hits is to the Firmicutes, which include the photosynthetic heliobacteria, though none were included in the database used here. This may be indicative of some contribution of the heliobacteria to the chloroplast progenitor. The hits to the other orders are all 3% or less. In conclusion, these data illustrate the composite nature of the chloroplast proteome. Two-thirds of the proteins are of cyanobacterial origin, but there are significant contributions from mitochondrial and eukaryotic proteins. The proteins with top-hits to the Firmicutes suggest a relationship with the heliobacteria which could be followed up now that one of these has been sequenced.

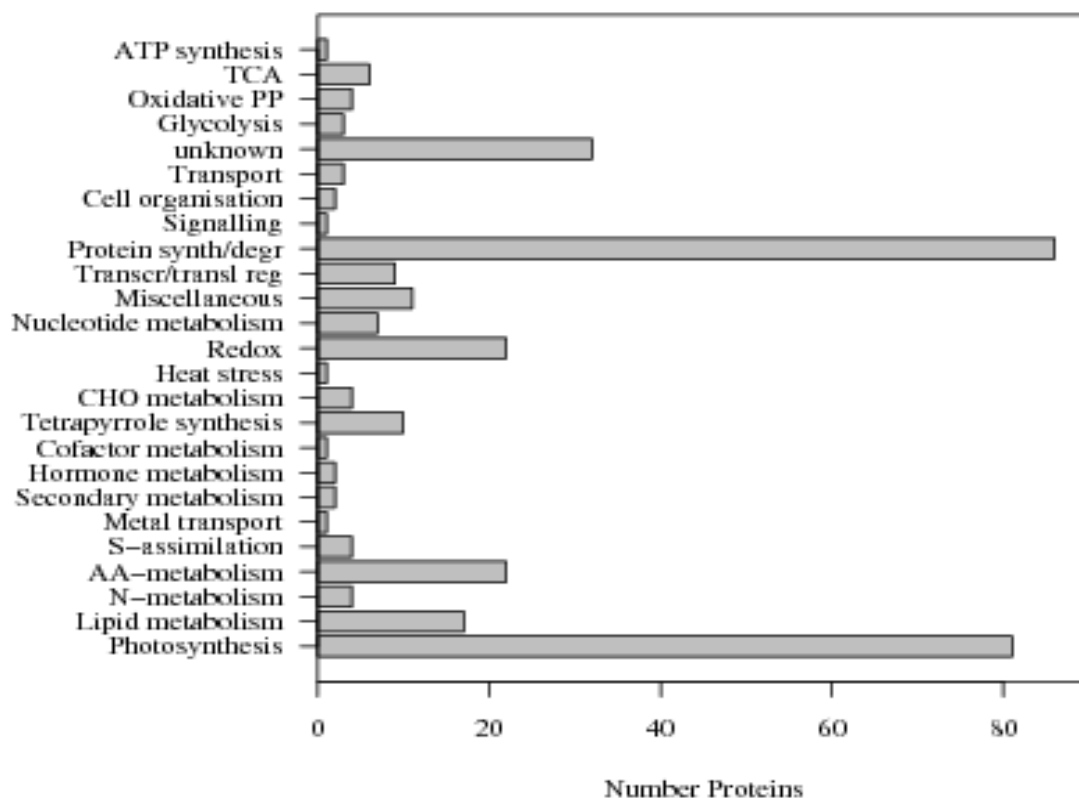


Figure 2.4: Mapman functional characterization of chloroplast proteins.

2.4.4 Functional characterization

The Mapman functional characterization of the Seq2 proteins are summarized in Figure 2.4. Proteins involved in photosynthesis and protein synthesis account for nearly half the total. Proteins with unknown function are the next highest category. Proteins of lipid metabolism, amino acid metabolism and redox function contribute approximately 20 each. The protein distribution therefore reflects the replete photosynthetic growth state of the plants.

A detailed biochemical and pathway analysis of these proteins has not yet been done, but the following observations are offered. Acyl Carrier Protein 2 (ACP2, At1g54580) was detected. The ACP4 isoform, which was not detected, is normally expressed in leaves and is light-regulated, while the ACP2 isoform was expressed in roots and not light regulated[12]. Microarray data archived at TAIR (www.arabidopsis.org/cgi-bin/afgc/atExpressioncgi.pl?su_id=141607, no other citation) indicate that ACP2 is up regulated when grown with sucrose. In a recent microarray experiment of *Arabidopsis* seedlings etiolated on a sucrose-containing medium, 368 genes were found to be up-regulated (ACP2 was not in this set, however). Of the 368 up-regulated genes, we find 18 of the 99 that had a predicted chloroplast localization. Only one of these was not a synthase: OEP16

(At2g28900). This outer membrane porin is a transporter of amino-acids and other amino-containing solutes, but has recently been shown to participate with TOC33 in the import of protochlorophyllide oxidoreductase A (porA)[118, 119]. Protochlorophyllide oxidoreductase A was detected as well (At5g54190). Its up-regulation may be therefore be a response to increased protein synthesis or photosynthesis rate and may correspond to the high numbers of proteins with these annotations detected (Figure 2.4). These plants were grown on a sucrose-containing medium and 18% of chloroplast localized genes shown to respond to sucrose were detected. However since we did not have a control state for comparison, we cannot state absolutely that their detection is due to the growth state.

A surprisingly large fraction of the protein import machinery was detected, considering that these are mostly membrane proteins and no special effort was taken to isolate these. Four of five outer membrane components were detected: Toc159, Toc75, Toc34, and the OEP16 mentioned earlier (At4g02510, At3g46740, At5g05000, and At2g28900, respectively). Only Toc64 was missed. The inner membrane Tic110 (At1g06950) and Tic55 (At2g24820) were found; Tic20, Tic22 and Tic40 were not. Two HSP90's (At5g50920 and At3g48870) and two HSP70's (At5g49910 and At4g24280) were detected. Three of these were detected by the Any2 search only, the others were detected by both the Seq2 and Any2 searches. The multiple HSPs detected may represent inner membrane and stromal forms or forms not involved in import activities.

2.5 Discussion

Modern mass spectrometric techniques identify hundreds of proteins and may describe the major features of the proteome of an organism in a given metabolic or developmental state. In this experiment we report the identification of several hundred proteins of the photosynthesizing chloroplast and use the data to evaluate prediction algorithms and infer the phylogenetic origin of the plastid proteome.

The plants used in this study were grown on sucrose-supplemented agar. While this method has been shown to produce a high yield of functional chloroplasts, the chloroplast metabolism may be distorted by the sucrose, since normally the only time the plant grows with excess sucrose is during etiolation from the seed. This is reflected in the morphology of the plants which were small and upright, with no sign of the appressed rosette of soil grown plants. Our proteomics data may support a metabolic shift. ACP2 which is expressed in roots was detected instead of the light-regulated leaf isoform ACP4. Eighteen of 99 chloroplast-localized proteins up-regulated during exposure of etiolating plants

to sucrose were also detected, including the porin OEP16. The latter has been shown to be involved in the import of a variety of solutes and the nuclear encoded protochlorophyllide A (*porA*, also detected). *PorA* has in addition been shown to be regulated by its substrate, protochlorophyllide[117, 118, 119]. In normal plastid development, etiolating seedlings rapidly lose protochlorophyllide as they green and the import of *porA* drops dramatically[117]. Together these observations suggest that these plants are ‘juvenilized’ by the sucrose media, since components of metabolic systems associated with etioplasts are detected in these 21 day-old green plants. These observations are suggestive only, since the proteomic determination was for a single growth state. Additional experiments with plants grown on soil would provide a useful comparison.

A surprisingly large fraction of the known protein import machinery was detected[58]. If the chloroplasts are indeed in a juvenile and rapidly growing state, this might imply that the import machinery is relatively abundant.

Since Mereschkowsky[96], the plant chloroplast has been thought to be derived from cyanobacteria based on structural and biochemical observations. More recent data supports this by demonstrating that plastids group with cyanobacteria in phylogenetic trees[91]. It is known however that not all cyanobacterially derived proteins in *Arabidopsis* are targeted back to the chloroplast[91] and conversely that some organelle proteins are derived from sources other than their progenitor prokaryote. This experimental protein data was compared to a database of bacterial and yeast sequences using BLAST to infer the origin the plastid proteome. A relatively low 65% of proteins had a top-hit to the Cyanobacteria. The next highest contributions were 10% each from the Proteobacteria and yeast. The former include the α -proteobacteria which are the presumed ancestors of the mitochondrion. Therefore this fraction must represent the replacement or addition of proteins from the other plant organelle. The 10% contribution from yeast represents the contribution from the eukaryotic host. While most of the proteins from either source were enzymes, several of the yeast related proteins had annotations suggestive of regulatory functions. Much of the protein import machinery is also of eukaryotic origin. These proteins may therefore represent part of the systems added by the host to control the bacterial endosymbiont. There was also a eukaryotic triose-phosphate transporter, which may be indicative of stage in the progress toward full endosymbiosis in which the host extracted photosynthate from the proto-organelle. There were also some hits to the Firmicutes, which may indicate a contribution of the photosynthetic heliobacteria to the cyanobacterial progenitor since these bacteria belong to this order. For both the protein data sets examined, approximately 14% of the chloroplast proteins had no homologue in the target database of 37 proteomes. The

smaller database was used instead of non-redundant (nr) database to exclude hits to other plants. A better strategy might have been to use the nr with plants excluded. In addition there have been more genomes sequenced, particularly cyanobacteria, since this analysis was done. The larger and newer database might assign homologues to the ‘orphan’ proteins and confirm the linkage to the Firmicutes. Finally, it must be noted that proteomics is biased towards more highly expressed proteins. The phylogenetic distribution noted here may reflect the particular growth state examined.

One of the goals of this analysis was to test the performance of localization prediction algorithms with a large set of experimental proteins. Our data agree with the published TargetP sensitivity of 85% for TargetP[29]. For Predotar we find a lower sensitivity than that published, 75% instead of 82%[139]. A relatively high proportion of this dataset, 74%, is predicted by both algorithms to be plastid localized. Analysis of the excluded proteins however reveals that each algorithm falsely rejects proteins accepted by the other, indicating that there is room for improvement of both. Additionally, the two algorithms split on one known dual-targeted protein and several other predictions split between these two locations. As more of the sequences become known, provision could be made for dual-targeting in algorithm upgrades. Neither do these algorithms account for outer-membrane proteins which have no N-terminal targeting sequence, yet are *bona fide* components of the plastid proteome. Improvements might try to incorporate these as well, or include other algorithms in a multiple sensor approach. There has been some effort in prediction of plastid outer membrane proteins[129]. Finally, it must be noted that many of the proteins in the TargetP and Predotar training set represent chloroplasts in the photosynthetic state as do the data here. The good agreement may be due to a ‘representation bias’. More experiments, such as this one devised for other plastid developmental states, might both test and improve prediction algorithms.

The MS/MS data for this experiment were analyzed both by a standard Sequest two-peptide criterion and by a multiple search algorithm criterion based on the results of Chapter 1. A subset of MS/MS spectra, consisting of about 10% of the total were searched by each of Sequest, Mascot, and the open source algorithm X!Tandem using a database of *Arabidopsis* sequences with reversed sequences appended. Parameters for each algorithm were adjusted to return a false positive rate of 1% as measured by hits to the reversed sequences. The entire set of mass spectra were then searched with each algorithm and the results filtered with the chosen parameters. The algorithms were combined as a ‘meta-algorithm’ by using as minimal criteria for protein acceptance: two or more unique peptides identified by any single algorithm or two or more algorithms matching the same spectrum

to the same peptide. Thus a single peptide identification is only accepted if identified by two or three algorithms. The meta-algorithm produced about 30% more identifications. The phylogenetic data and the localization prediction results both suggest that the increase in identifications may have come at the expense of false positives. The fraction of hits to Cyanobacteria dropped slightly and the fraction to yeast increased slightly. Likewise the fraction with chloroplast localization dropped by about 5% for the Predotar predictions only. This implies more random hits to other proteins since the plastid proteome is about one-fifth the total *Arabidopsis* proteome. However, many of the new proteins detected had annotations suggesting valid chloroplast localizations. And as mentioned previously, some of them were outer membrane components which would not be correctly detected by the localization algorithms used and might be of eukaryotic origin. Even if the increase in false positives is 5%, the much larger increase in identifications, most of which appear to be valid, argues that the overall approach of combining algorithms which have different sensitivities is valid. The combination employed here was an attempt to apply a ‘two-peptide’ rule using all three algorithms, it might be that a more sophisticated approach, perhaps using artificial intelligence tools, could optimize both the sensitivity and specificity.

Chapter 3

Motif Analysis of Chloroplast Transit Peptides

3.1 Abstract

The motif ‘FPLK’ has been demonstrated to be an important feature of transit peptide in chloroplast protein import. Other studies have shown the importance of the residue phenylalanine for protein import into the primitive plastids of the glaucocystophyte *Cyanophora paradoxa*, implying a relict preference of the plastid import system for this residue. The motif-finding tool MEME and a heuristic approach were applied to curated TargetP targeting sequences and to proteins from our proteomic analysis of the *Arabidopsis thaliana* chloroplast. The ‘FPLK’ motif was detected, but equally well in chloroplast and mitochondrial sequences and appears to have no discriminatory power.

3.2 Introduction

As discussed in the general introduction (Chapter , the import apparatus of the chloroplast and mitochondrion share a great deal of functional similarity. Likewise, the targeting sequences which direct proteins to the organelles have similar residue composition, domain organization and secondary structure behavior. The phenomenon of dual import directed by ambiguous targeting sequences also argues for a high degree of correspondence in the two systems. So there is truly a ‘paradox of plastid transit peptides’([15]) or more generally of organelle targeting: how is targeting fidelity attained when there is little apparent difference within or between chloroplast and mitochondrial targeting sequences?

The lengths of mitochondrial pre-sequences are on average, longer than the signal pep-

tides which direct proteins to the secretory system. Chloroplast transit peptides are longer yet, implying that as each new targeting situation arose, the cell had to impart more information to the targeting sequence. (Although there are transit peptides that are shorter than pre-sequences and *vice versa*.) The nature of the information and how it contributes to organelle discrimination is still obscure however. The information or recognition requirements of the targeting sequences might be important in the following import steps:

- chaperone binding
- organelle discrimination
- translocon binding
- translocation
- suborganellar targeting
- targeting sequence processing by peptidases

Recognition motifs have been found or implied in several of these steps. Cytosolic HSP70 chaperones have been shown to be involved in both mitochondrial and chloroplast protein import[162], serving perhaps to keep proteins in an unfolded state to minimize the energy requirements for translocation. HSP chaperones exist in the matrix and in the stroma, and may help pull the protein through the translocon in an ATP dependent manner[21, 162]. A 14-3-3 protein has also been shown to interact with a phosphorylated serine on a chloroplast targeting sequence[93]. Motifs for both the HSP70 and the 14-3-3 recognition site have been detected bioinformatically[57, 93]. Since import occurs specifically *in vitro*, in the absence of cytosolic factors, their recognition sites must not be involved in organelle discrimination. Other motifs that have been found are for thylakoid protein import by the twin-arginine system and for peptidase recognition in both organelles. These likewise have no bearing on import specificity, since the proteins must first be imported via the general import pathway.

Hydrophobic motifs have been proposed for both mitochondrial and chloroplast translocon recognition. An NMR study of rat TOM20 complexed with pre-sequences showed that the pre-sequence adopts an amphipathic α -helical conformation in the binding groove, with the hydrophobic face of the helix contacting the hydrophobic residues in the groove[100]. Using results from five pre-sequences, these authors proposed the TOM20 recognition motif: $\phi\chi\chi\phi\phi$, where ϕ is an aromatic or hydrophobic residue, and χ is any residue. For chloroplasts, a motif 'FPLK' has been investigated for many years as a chloroplast recognition element. This was first proposed in 1986 using a few orthologous sequences known at

the time[61]. While unrelated targeting sequences show little or no similarity in sequence, orthologues or paralogues may retain significant similarity, especially at the C-terminus of the transit peptide ([16] and Figure 5.6). Later, Pilon et al. using more sequences from unrelated proteins provided further support for the motif, generalizing it to F|W-G|P-K|R[113]. The motif was predominantly located in the middle or C-terminal third of the transit peptide (Figure 3.1). It was demonstrated that deletions in regions of the ferredoxin transit peptide containing the motif reduced import *in vitro*. These authors deleted relatively large stretches of the transit peptide however. In a more recent study, single residues or pairs of neighboring residues of the transit peptide of the small subunit of rubisco (prSSU) were replaced with alanines throughout the length of the peptide[76]. Import of fusion proteins of the transit peptides with green fluorescent protein in *Arabidopsis* protoplasts, showed very little effect of most of the mutations on import. Those substitutions that did show a reduction in import were in a region containing the FPATRK sequence. In our own laboratories, mutations of components of the FTGLK and FPVSR sequences in tobacco prSSU severely reduced *in vitro* import into pea chloroplasts (Sarah Wright, personal communication). The studies cited previously imply a motif that might be generalized as an aromatic residue (F or W), a helix-breaking residue (P or G), an aliphatic residue (A,I,L,V), and a basic residue (R or K). Thus the mutagenesis studies indicate that the motif is important for protein import into the chloroplast.

A recent report has shown the importance of the single residue phenylalanine for plastid import in the glaucocystophyte *Cyanophora paradoxa* which is thought to be similar to the first phototrophic eukaryotes since it still possesses a peptidoglycan layer and a relatively large complement of plastid-encoded genes[145]. *Cyanophora* plastid proteins can be imported into higher plant plastids, but not *vice versa*. Analysis of these proteins showed a conserved phenylalanine in the third or fourth position of the transit peptide. If this motif were engineered into higher plant transit peptides, they were capable of effecting import into *Cyanophora* plastids. Phenylalanine is also involved in directing bacterial proteins destined for export to the outer membrane through OEP85. The authors propose a model for the evolution of plastid import in which OEP85 is inverted, becoming the importer TOC75 rather than an exporter, and retaining the requirement for an N-terminal phenylalanine. While there is no apparent conserved phenylalanine in the N-termini of higher plant transit peptides, it may be that some affinity for phenylalanine is involved in translocon recognition and organelle discrimination as implied by the evidence for the ‘FPLK’ motif cited previously.

```

ara MAS---T-ALSSAIVGTSPFIIRSPAPISLRSLPSANTQSLFGLKSGTARGGRTAM 52
spi MAA---T-TTIMGMAITVFPKQPAFMMAALPSNTGSLFGLKKTGS-AGGRTM-M 50
sil MAS---T-LSTLSVSASLLPK-QCPMVMSALTNTMGSLFGLKAGSR-GRVTM 48
pea MAT---TPALYGTAVSTSFRLTQPMMSVTT-TRAFSNGFLGLETSLRKGGDLAVAM 52
may MATVLGSPRAPAFFSSSSSLRAAPAPTAV-ALPAAKV-GIMGRSASSRR--RLRAQ 52

```

acyl	MASIAAASASIS-----	<u>LQARPRQLVSLFPR</u>	<u>ISHLVLLVLLWQ</u>	<u>AIASQVKSPFNGR</u>	<u>SSLSFNLRLQL</u>	<u>PTRLTVSC</u>	70
tufa	MAISA PAAC SSSSRILCSYSS PS LC PA ISTSGKIKLTGLSSSP LP SYSLTTSASQSTARRSPTVR-----						68
ract	MAAAVSTVGAINRA PL SLNGSGSGAVAPAS-----	<u>TELKKVVT</u>	<u>VSRFAQSN</u>	<u>ASNG</u>	<u>-----</u>	<u>SFKVL</u>	58
epsp	MAQVSRICNGVQN PS LISNLKSSQQRK SP LS-----	<u>VSL</u>	<u>KTQQHPRAY</u>	<u>PISSWG</u>	<u>LKSGMT</u>	<u>LIGSELRLPKVMSSV</u>	72
ferr	MASTA-LSSAIVGTSTFI RR SP AP ISLRSIP SA NTQ-----	<u>SIF</u>	<u>-----</u>	<u>GLKSGT</u>	<u>ARGGRV</u>	<u>TAM</u>	52
ssur	MAASM-LSSAAV TS -----	<u>PAQATMVAP</u>	<u>TGLKSS</u>	<u>-----</u>	<u>ASFVT</u>	<u>RKANNDITS</u>	55
plas	MAA IT -----SATV TI PSF-----	<u>TGLK</u>	<u>LA</u>	<u>VSSKPK</u>	<u>-----</u>	<u>TLSTIS</u>	47
lhcp	MAASTM-----ALSS PAE -----	<u>AGKAVNLS</u>	<u>PAASEV</u>	<u>-----</u>	<u>-----</u>	<u>LGSGRVT</u>	34
	**						

elip	MAVSSCQSIM--SNSMTNISRRSRVNF-----TNIPSVYIPTLRNVSLKVR--M	48
fnre	MAAAVTAAVSLPYSNSTSLPIRTSIYAP-----ERLVFKKVLNIVSISGRVGTIRA	52
carb	MSTSSINGFSLSLSLPAKTSKTRTLLPFVSAISLNTSSSSSSSTPSLIDQEPVFASSSPIITPVLAERM	70
ri18	MALLCFNSFTTTTPTVSSSSSLPHPTANPISRVRIGLPTNCLKG-FRLITP	49
plas	MAT-----VTSTTVAIPSFSGGLKTN-AAKTVSMAKIPSTSTSQSPRLCV-----RASLK-----	48
ssur	MAS-----MISSVAVTVSRASTVQSAAVAPEGGGLKSMTGFPVKKVNTDITSITSNQGVKVC	57
ferr	MAT-----TPALYG-TAVST-SFLRTQPMFM-SVTITKAFSNQFLGLK-TSLKRGDLAVAM-----	52
sodm	MA-----QTLVSP-SPLSSHLRSTFSGV-SVKLAPQFST-LA-T-SNFPK-LTUVV-----	48
hsp7	MAS-----QSVSLSTI-ASPI-----LSQKPGSSVKSTPPCM-----ASFPLRQLPRLGLR-N	46
lhcp	MA-----A-----SSSSSMALSSPTLAGEQLK--L-----NPSSQELGAAR--F-T	36
ri24	MV-----AMAMASL-QSSMSSSLSLSNSFLGQPLSPITL-----SPFLQKGPTEKKCLI-V	49
glns	MA-----QILAPSTOWMRITKTSPCATPITSKMSLSLVM-----EQTKVAHSAK--FR-V	49
ri09	MASSTLSSLSSTPL-----QHSFADNLKTCSSQFPNKSSG-----	34
ri25	MASVSSIFPGCGVSMAPNSSLRNKAIRT-----ERR	30

57

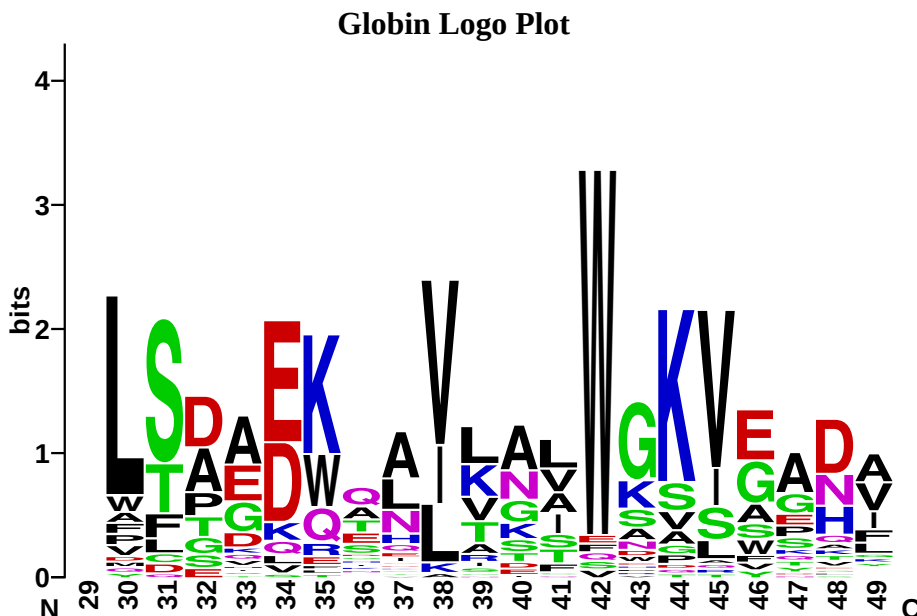


Figure 3.2: Globin fold logo plot. The plot was produced from 58 aligned globin-binding sequences using WebLogo[25].

Here we test using bioinformatic tools the hypothesis that the ‘FPLK’ motif is furthermore important not only for plastid import but is also involved in discrimination of the chloroplast and mitochondrion.

Motifs in biological sequences are generally considered to be fixed width, though their position in the larger sequence might be variable. Presumably this is because they must be geometrically fixed to interact with cognate sites in another sequence. Conventional motif-finding tools such as MEME[7] and HMMER <http://hmmer.janelia.org/> have been developed to find signals such as these. Visualization tools such as sequence logos are also available to conveniently visualize the frequency of residues at each site in a motif[131]. The paradigmatic TATA box and the globin fold are examples (Figure 3.2). Figure 3.1 shows that if the motif exists, both its position and its width are variable. To detect this type of motif we have developed a more ‘heuristic’ approach in which a sliding window is applied to the test sequence and the required residues sought. A similar approach was recently reported in a method to find motifs specific for methyltransferase activity[63].

3.3 Materials and Methods

Sequences of proteins localized to the mitochondrion, chloroplast, secretory system, and cytosol were downloaded from <http://www.cbs.dtu.dk/services/TargetP/datasets/>. The localization of these sequences is known experimentally and the examples were further curated by the authors of TargetP to be used in the training of that algorithm[29]. To validate the discrimination rules developed using the TargetP sequences, the chloroplast proteins detected by shotgun proteomics described in Chapter 2 and filtered using the ‘Sequest 2 peptide rule’ were used. A similar shotgun proteomic data set of 385 proteins detected in the *Arabidopsis* mitochondrion was also used[46].

For the heuristic analysis, rules defining variations of motifs were formulated as regular expressions and applied in a sliding window to each protein sequence in the set being analyzed using the Perl programming language. For all rules, glutamate and aspartate were excluded. Thus if D or E were encountered, the window is advanced and the rule applied again. The rule as presented in the last section would be `[F|W] AND [P|G] AND [A|I|L|V] AND [R|K] NOT [D|E]`. In some rules, the window was allowed to be variable, that is the match continued until it failed and the length recorded. In other rules, fixed windows of 5 to 9 residues were used. Matches could be saved for logo plot generation. A variety of match statistics were developed, but here we employ only the percentage of sequences in the training set matched by a given rule.

Sets of synthetic sequences, each 1000 in number and 50 residues long, were created to test the rules. The set ‘fixed forward’ had FPLK residue inserted in random sequence at a random position; ‘fixed both’ had 500 sequences each with FPLK or KLPF; ‘ordered’ had FPLK in this order, but with randomly inserted residues, to fill out a random length of 5 to 9 residues; ‘shuffled’ was created by inserting the FPLK residues randomly in a random window of 5 to 9 residues; finally, ‘random’ had 1000 sequences of 50 random residues with no inserted motif.

The MEME server (meme.sdsc.edu/, [7]) was used with default settings for motif searching of real and synthetic protein sequences. Sequence logos were created with WebLogo[25] which was downloaded from weblogo.berkeley.edu/.

3.4 Results

3.4.1 Application of MEME to targeting sequences

The motif finding tool MEME[7] was applied to the synthetically generated sequences. Due to limits on the size of submissions to the server, only subsets of the 1000 sequences were analyzed. For 100 sequences from the ‘fixed forward’ sequences, each of which contained the FPLK motif, MEME detected the motif in each sequence with an overall expectation of $2.5\text{e-}282$. For 100 sequences of the ‘ordered’ motifs, each of which contained FPLK but in a variable window of 5 to 9 residues, MEME detected the motif ‘FPLK’ with expectation $2.0\text{e-}39$, but in only 49 sequences. A second significant motif, ‘FPxLK’ was detected in an additional 16 sequences with an expectation value of $1.0\text{e-}7$. These probably represent motifs wherein the 4 residues fell at the same position by chance. The combined results means that 65% of these motifs were correctly detected. For the ‘shuffled’ motifs in 100 sequences, MEME detected the motif ‘KPLF’ in only 5 sequences with a relatively poor expectation value of $6.8\text{e-}1$. If the analysis was increased to cover 250 of these sequences, no significant motif was detected. Thus, even though every sequence in these test data possessed a motif, the ability of MEME to detect them declined as the positional specificity was relaxed.

When applied to the TargetP training chloroplast proteins, MEME finds a single motif for all of the 141 sequences with expectation value $5.4\text{e-}50$ (Figure 3.3). For 96 sequences, this motif is N-terminal. This ‘MAS’ motif has been noticed for some time and may be picked out by inspection of sequences[30]. More surprisingly, it is found in the interior of 19 more sequences and for six of these, it is coincident with the experimental cleavage site of the transit peptide. For several of the proteins with the motif in the interior, it appears that another could have been assigned at the N-terminus as well. For example, sequences with Swiss-Prot numbers P323062, P32063, P32061, and P42910 had interior motif assignments while their N-termini were MAL, MAL, MAM, and MAS, respectively. The motifs in the remaining 26 sequences (18% of the total) lie well within the mature sequence of the protein. MEME found no motifs within the transit peptides of our experimental chloroplast protein data, despite the fact that the N-terminal ‘MAS’ motif is nearly as strong as that in the TargetP training set (c.f Figure 3.4, A and B).

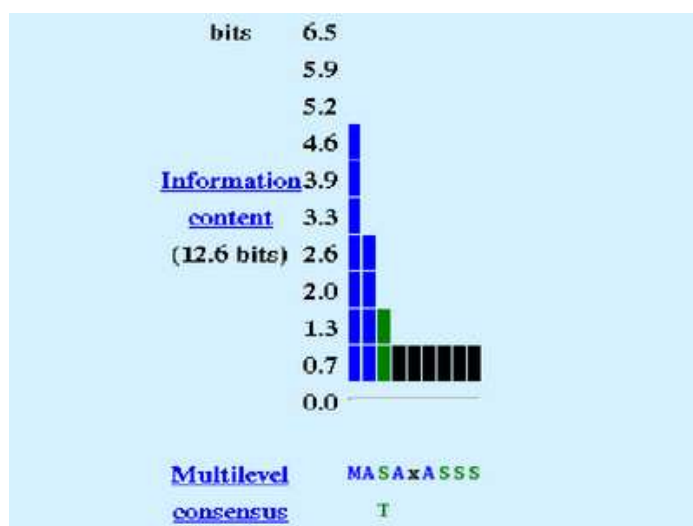


Figure 3.3: MEME motif for TargetP chloroplast sequences.

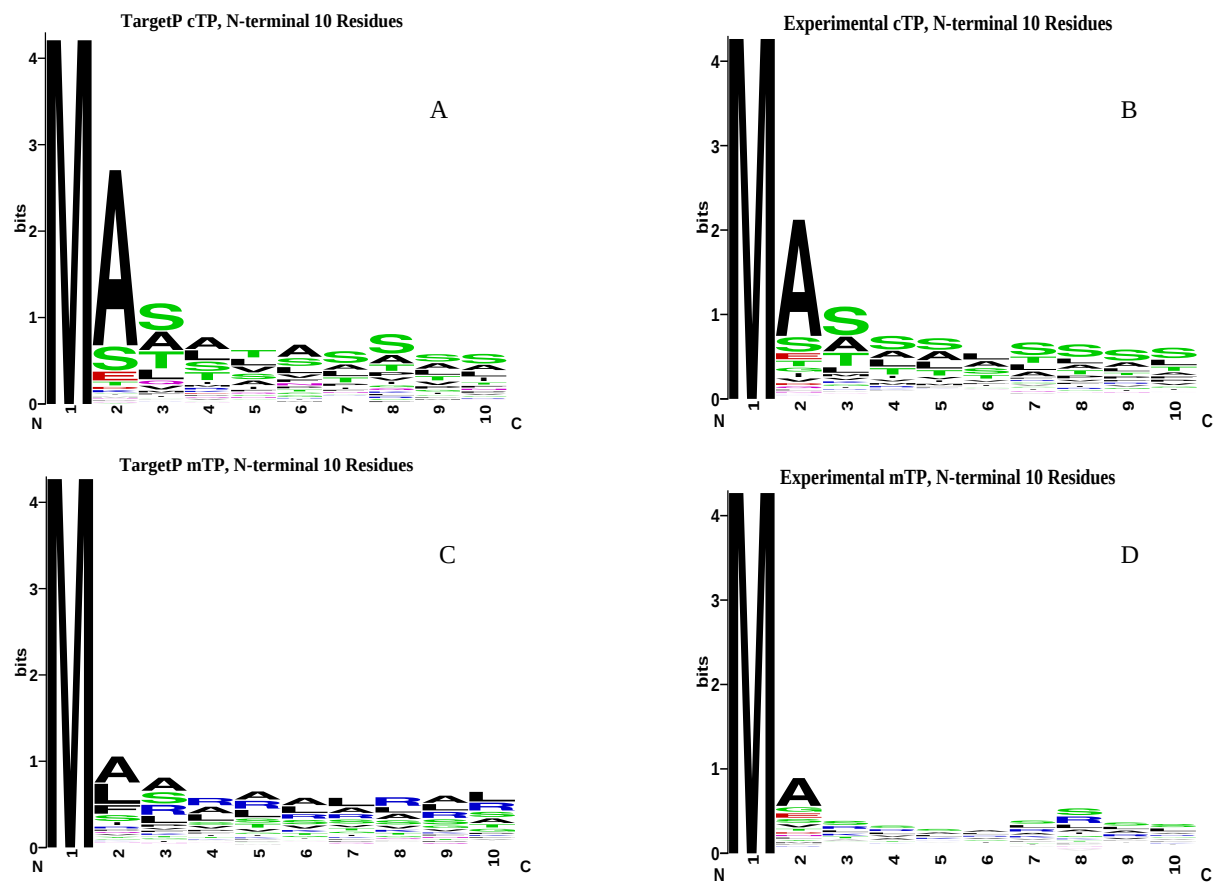


Figure 3.4: N-terminal sequence logos. A: TargetP chloroplast proteins; B: experimental chloroplast proteins, this work; C: TargetP mitochondrial proteins; D: experimental mitochondrial proteins, Heazlewood et al.[46]

When applied to the TargetP mitochondrial training set, MEME found no motif within the pre-sequences. This is surprising, since there is an N-terminal motif similar to the chloroplastic one, though less informative (cf. Figure 3.4 A and C). This may reflect a sensitivity limitation of MEME. For the 385 *Arabidopsis* mitochondrial sequences of Heazlewood, et al. ([46]), MEME found several significant motifs. One, 'TWHHTFYNY' ($E=3.3e-16$), was found in only five sequences, all of which were actins. Likewise eight proteins were found with the motif 'EQGTHFLIP' ($E=6.1e-10$); seven were prohibitins, the other was a stomatin. Prohibitins and stomatins belong to the same superfamily, so these proteins appear to be related as well[101]. Finally, four proteins with motif 'FETTKYYC' ($E=7.4e-16$) were detected, all were annotated EF1- α . Thus is likely that the Heazlewood, et al. data set is not non-redundant or that MEME has otherwise detected conserved regions in paralogues that are not motifs definitive for pre-sequences. Even so, the three motifs are unrelated and occur in only 17 of the 385 proteins.

The sequence logo comparison of the N-termini in Figure 3.4 show some difference in the mitochondrial and chloroplast targeting sequences. Both have alanine at the second position, but this position has more than twice the information content in chloroplast transit peptides. The remaining residues show a preponderance of serine and aliphatic residues for both of the chloroplast protein sets. For the TargetP mitochondrial proteins, arginine and aliphatic residues are frequent. Somewhat surprisingly the Heazlewood, et al. proteins show very little information content after the second position. In a recent mutagenesis study, it was shown that deletion of the first ten residues of prSSU severely reduced import and replacement of one or two of the initial residues seem to indicate that the hydrophobicity balance of the sequence was important for import[76, 75]. This conserved motif appears therefore to be important at least for chloroplast import.

3.4.2 Heuristic analysis of targeting sequences

In the previous section, the analysis of synthetic data sets demonstrated that MEME failed to detect motifs as the positional specificity was relaxed. A second, rule-based or heuristic approach was taken therefore to look for the 'FPLK' motif. The various rules employed are listed in Table 3.1. Most of the rules require the residues FPLK, and all rules exclude the acidic residues D and E. The variations on the rules generalize the simple motif to attempt a better match frequency. Thus 'F|W|Y' would represent any aromatic residue and 'A|L|I|V' hydrophobic residues. For some rules, the pattern was allowed to match until it failed, but for most the rule was applied iteratively using a sliding window (or 'word length') of 5 to 9 residues.

Table 3.1: Motif Definitions

Rule	Definition ^a
1	F W AND P G AND K R AND A L AND S T
2	F W AND P G AND K R AND A L
3	Any 4 of (F W, P G, K R, A L, T)
4	F W required, any 3 of (P G, K R, A L, T)
5	F W required, any 3 of (P G, K R, A L, S)
6	F W Y, P G
7	[FWPGKRALST]+, variable window, minimum 4
8	F W AND P G AND K R AND A L AND S
9	[FWYPGKRALST]+, variable window, minimum 4
10	F W AND P G AND K R N AND A L AND S T
11	F W AND P G AND K R N AND A L V AND S variable window
12	F W AND P G AND K R N AND A L AND S variable window
13	F W AND P G AND K R N AND A L V AND S T
14	F W AND P G AND K R N AND A L V
15	F W L V AND P G AND K R N AND optional S T variable window
16	F W L V AND P G AND K R N AND optional S T
17	F W Y AND P G variable window
18	F (to generate logoplot for context)
19	F W Y AND P G
20	F W Y AND P G
21	F W AND P G AND K R N Q AND A L I V
22	F AND P G AND K R AND A L V
23	F W AND P G AND K R AND A L
24	F W AND P G AND K R AND A L V AND S
25	F W AND P G AND K R AND A L V AND S T
26	F AND P G AND K R AND A L
27	F AND P G AND K R AND A L V AND S
28	F AND P G AND K R AND A L V AND S T
29	F AND P G AND K R AND A L AND S T
30	F AND A L AND P G AND K R AND S T W Y I M
31	F AND A L AND P G AND K R W Y I M

Continued next page

Table 3.1: Continued

Rule	Definition ^a
32	F AND L AND P G AND K R AND S T W Y I M
33	F AND P G AND K AND A L AND S T
34	F AND P AND K R AND A L AND S T
35	F AND P AND K AND L AND S T
36	F AND A L AND P G AND K R AND S T
37 ^b	pxxpp, where p=[LMWCAYFI] and x is any residue
38 ^c	[AILMFWYV] _n , where n = 0.6 * window length

^aAll rules exclude D and E; vertical bar '|' indicates alternation, any single residue; brackets indicate 'any of'; '+' means one or more; ^bthis rule formulates the hydrophobic mitochondrial motif given by Muto[100]; ^crule 38 was formulated as a generalization of rule 37 - require 60% of the residues in a window to be hydrophobic

Rule 37 encodes the mitochondrial Tom20 recognition motif proposed by Muto et al.[100]. Rule 38 is a generalization of this rule over the length of the windows examined. The two criteria used in evaluating the rules was the percentage of sequences matched, and how well each rule discriminated the chloroplast and mitochondrion.

The sensitivity of the rules was tested on the synthetic data sets. For the 'fixed' motif data sets, both forward, backward, and both, most of the rules detected the motif in 100% of the sequences with a window size of four. Detection fell off slightly at larger window sizes, probably because of the D|E exclusion rule. Those rules that did not detect the fixed motif were the ones that required other residues such as S or T. The performance of the rules on the more positionally random motifs was variable. Since there are so many rules and window sizes, this will be illustrated with the extreme responses. Plotted in Figure 3.5 are results for the relatively permissive rule 21. The rule matches fixed motifs well at all window sizes. For the ordered and shuffled motifs, approximately 70% match at a window size of 4 and the match percentage progressively increases with window size. The matches to random sequence is also significant however (last column of each cluster). This might be taken as an indication of false positive matching. For the more restrictive rule 22, which closely approximates the 'FPLK' motif, the fixed motifs again are detected easily (Figure 3.6). The randomized ones match well, but larger window sizes are required. For this more restrictive rule, the matches to random sequence are about half those to rule 21. Thus the heuristic approach is capable of detecting the motif even when fully randomized.

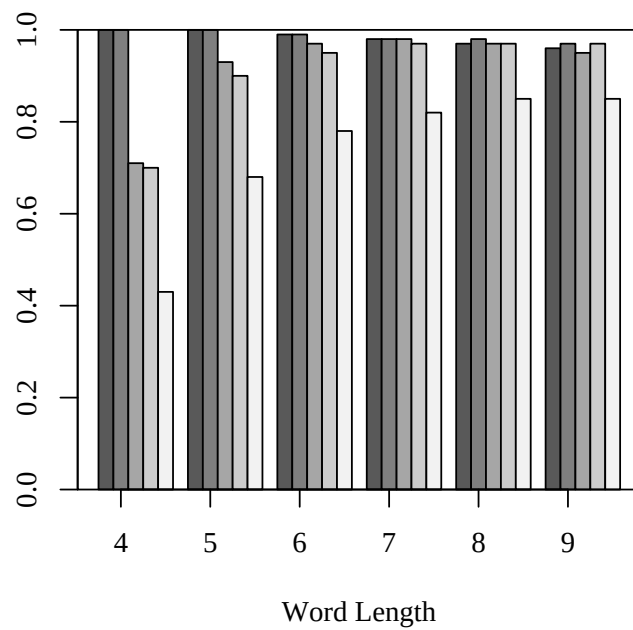


Figure 3.5: Percentage test sequences matched by rule 21: F|W AND P|G AND K|R|N|Q AND A|L|I|V. At each window size are the results for the data sets fixed, both, ordered, shuffled, and random.

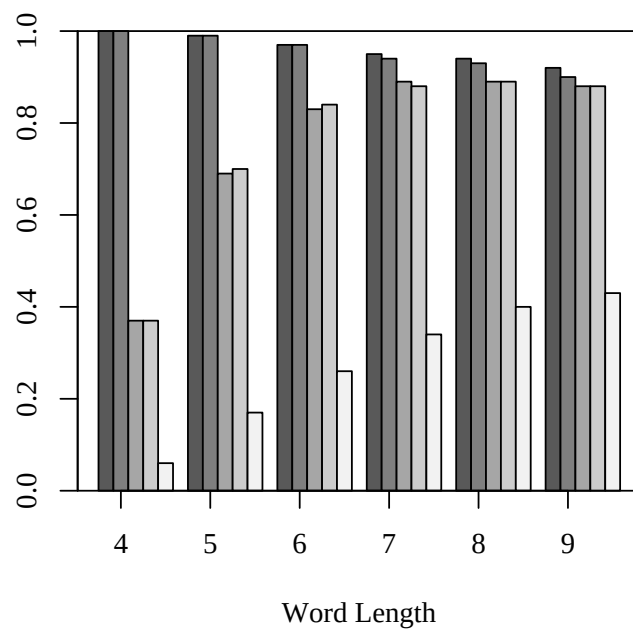


Figure 3.6: Percentage test sequences matched by rule 22: F AND P|G AND K|R AND A|L|V . At each window size are the results for the data sets fixed, both, ordered, shuffled, and random.

The loss of information caused by relaxing positional specificity shows up in this method as a requirement for larger window sizes to detect the motif, and an increase in detection in random sequence.

The suite of rules was applied to the TargetP training set data to find that variation of the FPLK motif that best discriminated the chloroplast and mitochondrion. It must be noted that 9 of the TargetP chloroplast sequences do not have phenylalanine in the transit peptide, so 93.6% (132/141) is the best any rule can match if this residue is required. As was seen for the synthetic test sequences, the sensitivity and specificity of the rules varied. Rule 16 matched the most chloroplast sequences, but matched mitochondrial, nuclear and cytosolic sequences nearly as well (Figure 3.7). It appears mainly to select against secretory proteins. Rule 35 was the most specific for chloroplast proteins, matching them five-fold better than mitochondrial proteins, but matched only about 30% of the chloroplast proteins in total (Figure 3.8). Finally, the rule 22 which closely approximates the FPLK motif, is intermediate. Eighty per cent of the chloroplast proteins are matched, but so are 50% of the mitochondrial for a ratio of 1.6:1. These results parallel those found for the test sets, a range of sensitivities and specificities are obtained based on the particular heuristic applied. The rule 22, “F AND P|G AND K|R AND A|L|V NOT D|E” does match the chloroplast proteins fairly well but does not discriminate well (Figure 3.9).

Interestingly, the TOM20 recognition motif proposed by Muto et al., $\phi\chi\chi\phi\phi$, does match mitochondrial proteins slightly better than chloroplast proteins, but matches secretory proteins best of all (Figure 3.10). Since this motif was for a fixed window size, it was generalized somewhat in rule 38, where 60% of sequences in the window were required to be hydrophobic (3/5, 4/6, 4/7, etc.). This rule matches secretory proteins very well, and discriminates well against other cellular compartments at large window sizes (Figure 3.11).

Finally, the list of rules was applied to additional experimental proteins identified by shotgun proteomics. The chloroplast data are reported in this work, Chapter 2, and are believed to be essentially free of contamination based on the lack of abundant proteins from other cell compartments. The mitochondrial data used was from a study of *Arabidopsis* grown in cell culture ([46]). Based on marker enzyme assays, the authors believed their preparation to have less than 1.5% plastid contamination and less than 0.5% cytosolic contamination. The first 100 residues of the proteins in each set were analyzed by all the rules as for the TargetP training set. Rule 35 was again the most discriminating, with a ratio of chloroplast to mitochondrial proteins matched of 2.6, about half the ratio of the TargetP training set (Figure 3.12). Rule 22 did match up to 90% of chloroplast sequences, but matched nearly as well for mitochondrial sequences (Figure 3.13).

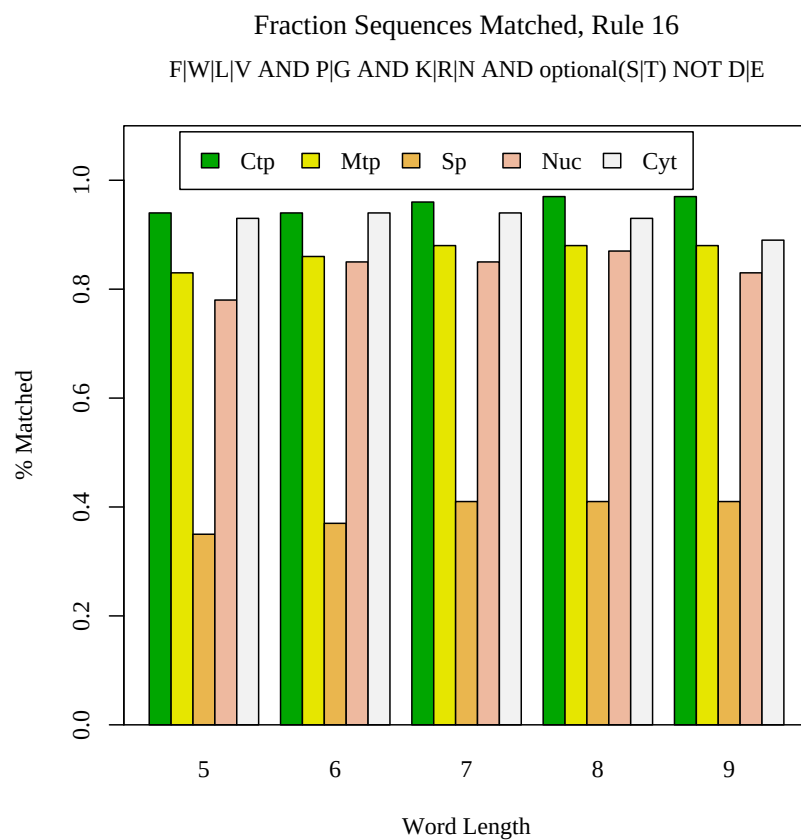


Figure 3.7: Percentage TargetP sequences matched by rule 16.

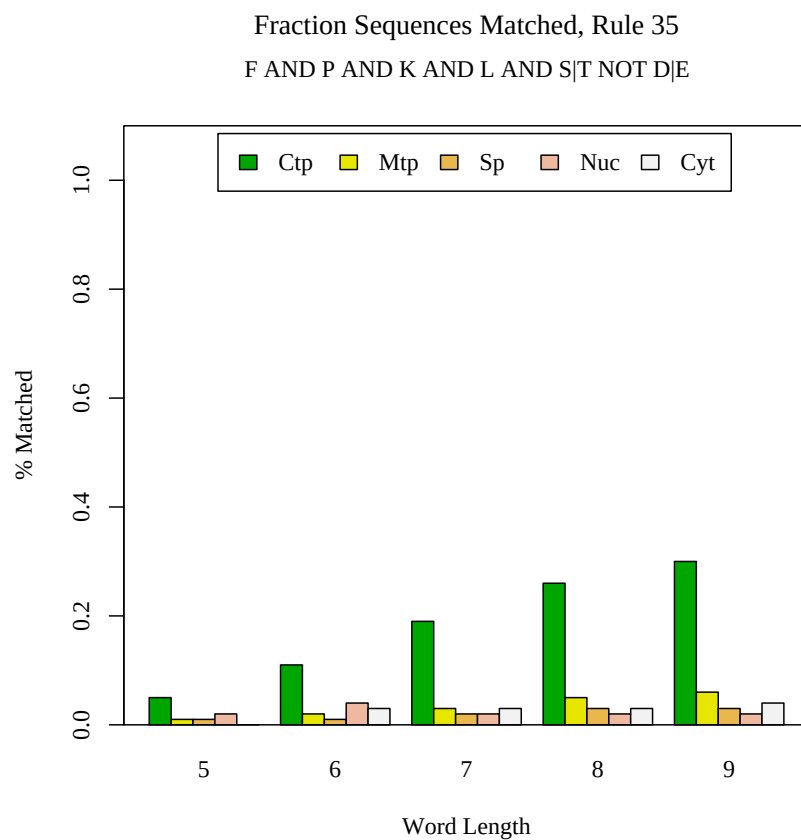


Figure 3.8: Percentage TargetP sequences matched by rule 35.

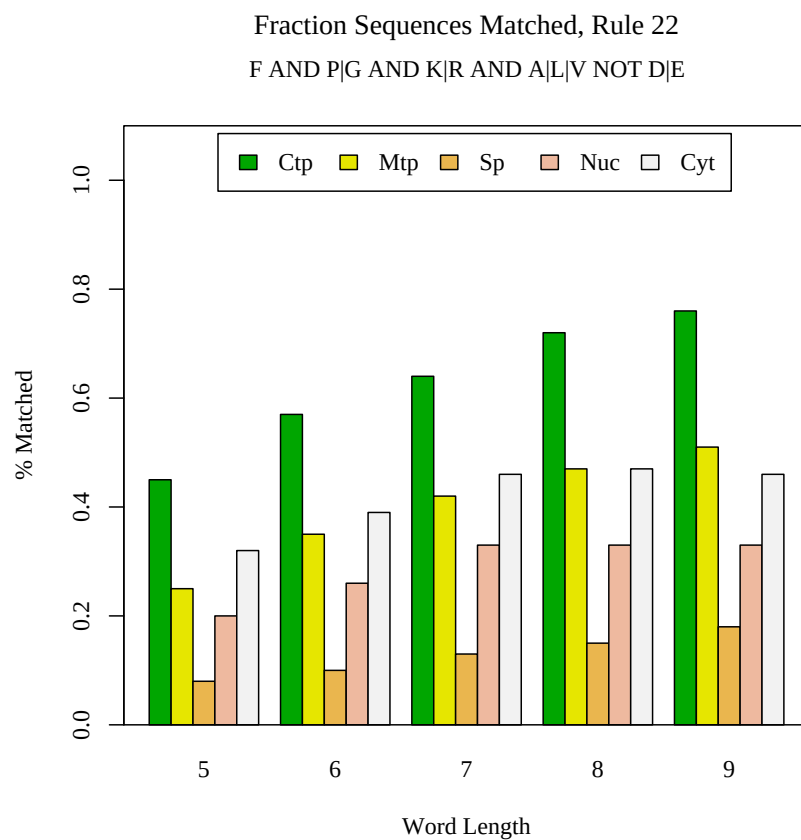


Figure 3.9: Percentage TargetP sequences matched by rule 22.

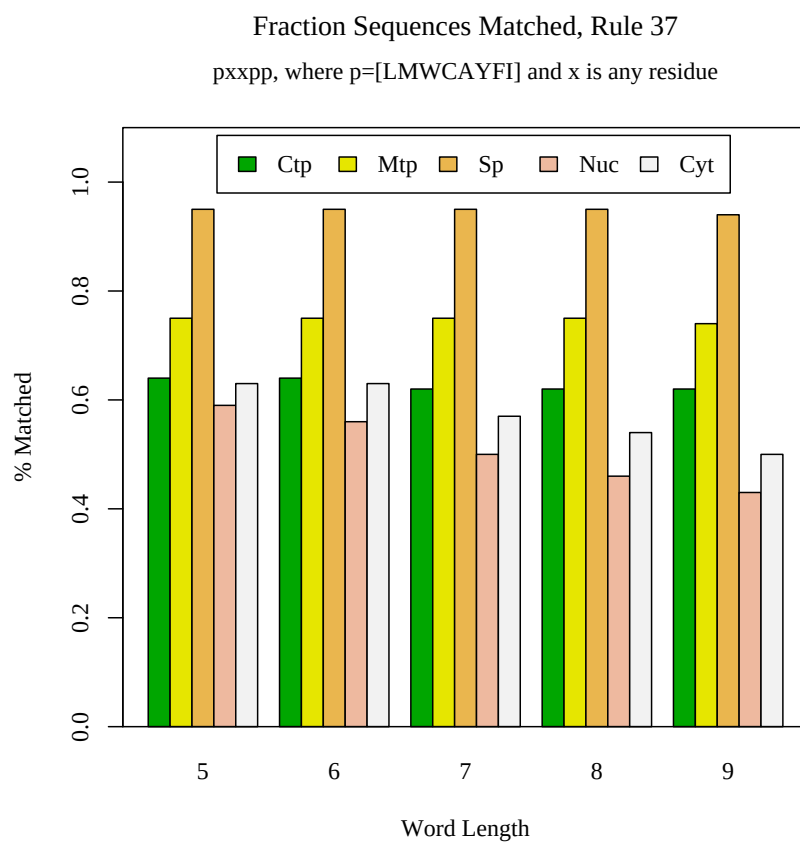


Figure 3.10: Percentage TargetP sequences matched by rule 37 (Muto et al. TOM20 recognition motif[100])

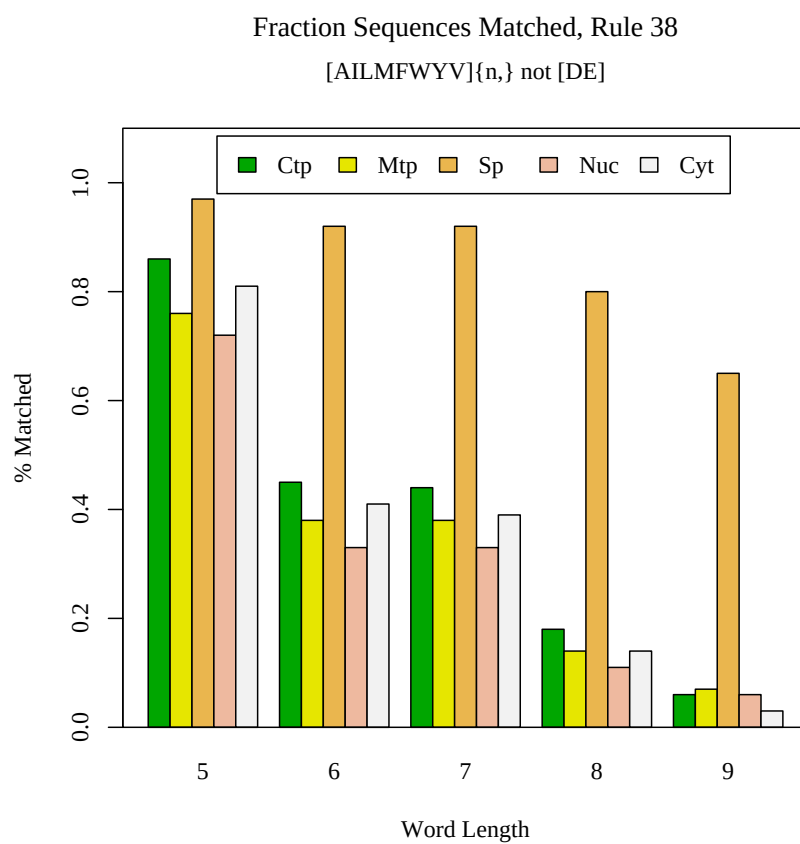


Figure 3.11: Percentage TargetP sequences matched by rule 38. The number of hydrophobic residues, ‘n’, was chosen to be approximately 60% of the window size.

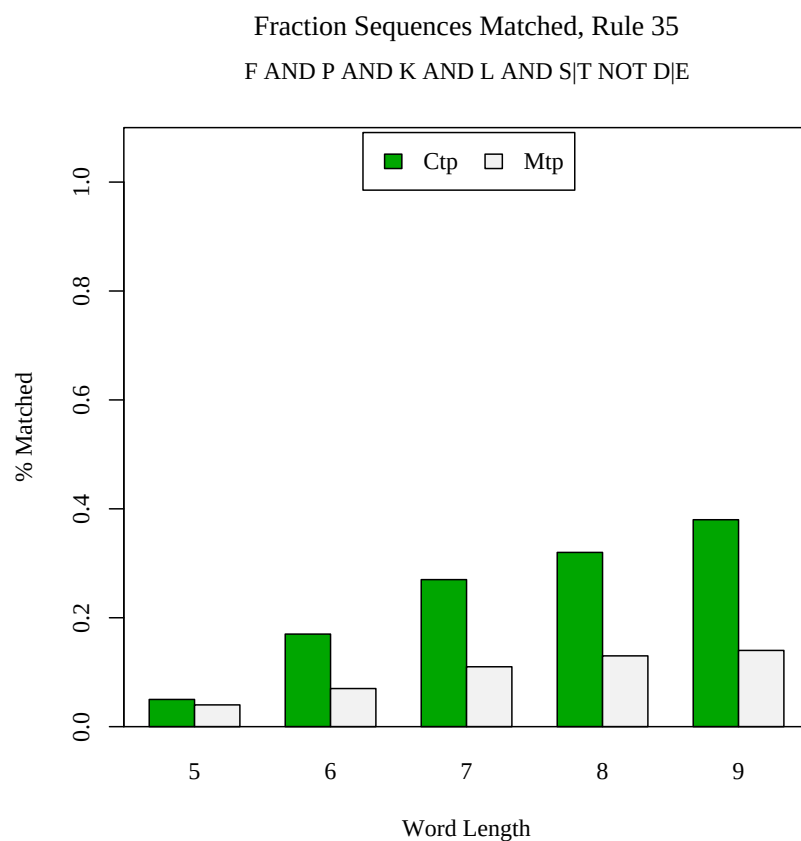


Figure 3.12: Percentage proteomic sequences matched by rule 35. Chloroplasts sequences are from this work. Mitochondrial sequences are from Heazlewood et al. [46].

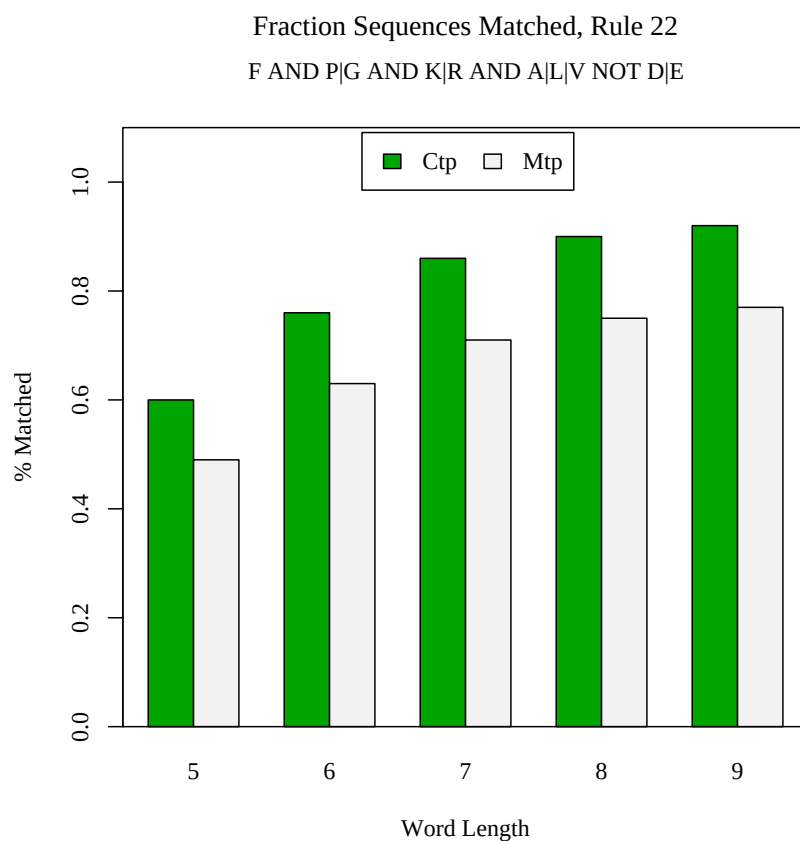


Figure 3.13: Percentage proteomic sequences matched by rule 22. Chloroplasts sequences are from this work. Mitochondrial sequences are from Heazlewood et al. [46].

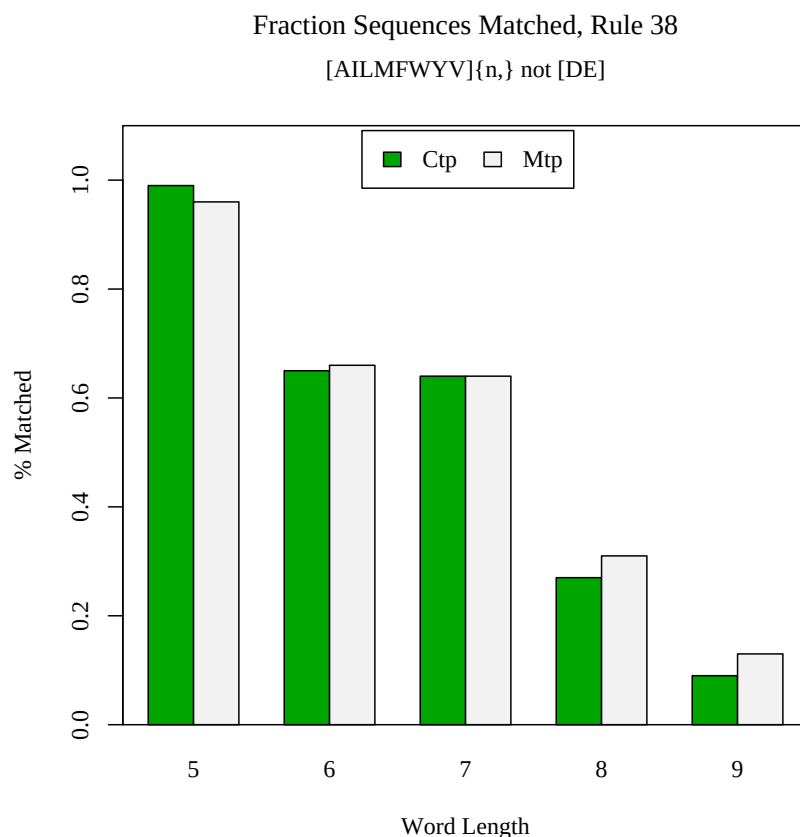


Figure 3.14: Percentage proteomic sequences matched by rule 38. The number of residues, ‘n’, was chosen to be approximately 60% of the window size.

The more inclusive rule 21 matched about 91% of sequences with no discrimination (not shown). Likewise, the rule proposed by Muto et al. matched nearly 100% of proteins from both organelles. The more generalized form of this rule, rule 38, also failed to discriminate the organelles, and discriminated against both as the window size increased (Figure 3.14; the Muto et al. TOM20 motif would correspond to the window size 5 in this plot). The analysis reveals that the proteomic data responded remarkably similarly to the TargetP training set. The more inclusive rules matched the most sequences, and the rules that differentiated best matched a minority of the data set. The percentage matches for the proteomic data were slightly higher for than for TargetP data. The rule 22 which best approximates the FPLK motif matches up to 90% of the chloroplast proteins, but does not discriminate it from the mitochondrion very well.

3.5 Discussion

The danger of the heuristic approach is that it encodes our expectations and we do indeed find what we seek. We constrained the arbitrary aspects of this approach by the analysis of synthetic sequence and showed that it is capable of discriminating sequences with the motif from random sequence, even as positional specificity was relaxed. The conventional motif-finding tool MEME quickly lost sensitivity as the positional requirements were relaxed. MEME was run from the web server with its default settings, but is available in a command line version with numerous options. It is possible that optimization of MEME's parameters for this particular problem might have improved its performance. The heuristic comparisons also showed that rules with increased sensitivity for the motif lost specificity by increased matches to random sequence. A refinement to the method might be to quantify this trade off by using a receiver-operating-curve (ROC). Katz et al. employed a heuristic method similar to ours in a search for methyl transferases in genomic databases[63]. They reported a high level of false positives for their approach, as measured by results on a test database of curated sequences. A problem with using random sequence as a control as we did is that chloroplast and mitochondrial targeting sequences are themselves very similar to random sequence, as we shall see in Chapter 5. A second constraint on the heuristic approach as employed here is that we used it to discriminate two classes of sequences of similar size and not as a tool for finding a motif in a much larger database.

The heuristic approach did detect the motif, as shown by the results for rule 22 ("F AND P|G AND K|R AND A|L|V NOT D|E") which matched 80% of TargetP chloroplast sequences and 50% of the mitochondrial sequences (1.6:1). For the proteomic data, the rule matched 90% of chloroplast sequences and 80% of mitochondrial sequences (1.1:1). A ROC analysis might indicate whether these differences are significant, but the motif clearly occurs almost as frequently in the mitochondrion as the chloroplast, which negates our hypothesis that this motif distinguishes the two organelle targeting sequences. Since the motif contains hydrophobic, hydrophilic, and helix-breaking residues, its occurrence in both the targeting sequences may reflect the requirement that they be random coils in solution, but capable of adopting an amphipathic α -helix upon contact with the membrane or translocon. A corollary is that the residue phenylalanine required in the motif does not serve any discriminatory purpose either. The overall phenylalanine content of the targeting sequences is similar, about 4%, and actually doubles to 8% in known ambiguous dual-targeting signals[107]. If phenylalanine is involved, as proposed by Steiner et al.[145], it must have a similar effect in both organelles, which is reasonable since the translocation channels of both organelles are β -barrels of bacterial origin. The most discriminating rule,

rule 35, was similar to rule 22, except that serine or threonine are required. This is significant, since that one of the few differences known between chloroplast and mitochondrial targeting sequences is a higher content of serine in the former.

We also detected the TOM20 binding motif proposed by Muto et al. [100] in both chloroplast and mitochondrial sequences. This supports the importance of the motif to import, since the motif was originally proposed based on the analysis of only five sequences. It may also indicate that a similar adoption of an α -helical conformation in binding may occur in the chloroplast as well.

MEME detected in the N-terminal ‘MAS’ motif in the TargetP chloroplast proteins, but not our proteomically detected proteins, even though sequence logos show a the same motif with slightly lower information content. Interestingly, the motif was detected in the interior of 19 sequences. This suggests a model for transit peptide evolution by translation initiation at an upstream start site. This idea will be discussed further in Chapter 5. Sequence logos also suggest that the MAS motif is a feature of mitochondrial pre-sequences, though the information content is about half the chloroplast version.

Thus the bioinformatic experiments reported do not support a discriminating role for the ‘FPLK’ motif, instead they add to the list of features in which chloroplast and mitochondrial targeting sequences are similar. A possible exception might be if some of the features which appear to differ only in degree might serve to discriminate if combined in a sequential recognition process or served to increase the kinetics of import in one organelle in the other.

Chapter 4

Synechocystis PCC 6803 Proteomics

4.1 Abstract

The higher plant plastid is derived from a cyanobacterium that was assimilated by a primordial eukaryote. Most of the prokaryotic genes were exported to the nucleus and a system evolved to target proteins translated in the cytosol back to the plastid. Not all proteins were returned to the plastid however, some being targeted to the mitochondrion or cytosol, and some lost altogether. Which part of the cyanobacterial proteome was retained in plants is thus of interest. Here we present the identification of more than 400 proteins from *Synechocystis* sp. PCC 6803 using liquid chromatography followed by ESI mass spectrometry. This number includes 91 proteins annotated as hypothetical or putative. We find that more highly expressed proteins as indicated by the codon adaptation index (CAI) are preferentially detected in our proteomic analysis and that there is a correlation between the number of peptides detected and the CAI. A BLAST analysis using the experimental proteome found that these proteins are almost three times as likely to be conserved in plants than the *Synechocystis* predicted proteome.

4.2 Introduction

Cyanobacteria gave rise to the chloroplasts of higher plants and algae through an endosymbiotic association with a primordial eukaryote. The formation of the eukaryotic plastid was apparently a singular event leading to the monophyletic association of cyanobacteria with algal and plant plastids[95]. Following the endosymbiotic event, there was wholesale export of plastid DNA to the nucleus leaving approximately 100 (depending on plant species) proteins encoded in the plastid. This required the development of a targeting mechanism

to direct proteins produced in the cytosol back to the plastid and in the plastid, a machinery to import them. In a recent study the fraction of the model plant *Arabidopsis thaliana* proteome that is derived from cyanobacteria has been estimated at 18% based on phylogenies constructed by BLAST comparison of every *A.thaliana* protein to various yeast, archaeobacterial, cyanobacterial and other eubacterial predicted proteomes.[91] This estimate was thought to be conservative, since proteins without highly conserved sequence are frequently rejected in the phylogenetic analysis. It is also likely that some proteins of cyanobacterial origin have been supplanted by proteins of mitochondrial or eukaryotic origin as indicated in Chapter 2. Other parts of the cyanobacterial proteome, such as those necessary for a free-living existence, were probably lost altogether.

Most studies of the protein targeting system to organelles has been directed towards discovering the components and functions of the import apparatus and those parts of the targeting sequences that interact with them, since this is an active and essential process in the contemporary cell. Less attention has been paid to the problems of how the genes were exported and integrated into the nuclear genome or how targeting sequences arose in the first place or how they co-evolved with the import system. Since the identity of the cyanobacterial progenitor is unknown, a study of expressed cyanobacterial protein conservation in plants might illuminate these mechanisms.

We wished to do a whole-proteome investigation of the model cyanobacterium *Synechocystis* and to use the results to address the issues raised above. First, what fraction of the cyanobacterial proteome was used and retained by the evolving plant? Secondly, can the cyanobacterial proteome be used to infer the origin of the transit peptide? We address the first question in this chapter and the second in the next.

Cyanobacteria have been the subject of several proteomic studies in recent years. There have been several analyses of isolated outer and plasma membranes [36, 53, 52, 3], as well as membrane complexes [47]. The peripheral thylakoid proteome has been investigated [159] and the composition of *Synechocystis* photosystem II[62]. The latter study was notable for detecting 5 novel components of PSII, two of which were later demonstrated to be necessary for the assembly and maintenance of active PSII complexes. [153]. Soluble proteins from *Synechocystis* [138] and *Nostoc* [54] have also been studied. All of these studies have utilized gel separations followed by in-gel digestion and MALDI or LC/ESI/MS analysis. An early whole-proteome study used N-terminal microsequencing of 2D gel separated soluble and membrane fractions to identify 143 proteins[128]. Recently, 776 *Synechocystis* proteins were identified in a comparison of protein and peptide pre-fractionation methods, but no analysis of the data beyond counts was provided.[38]. In the next chapter we further use

this data set to make inferences about the origin of the transit peptide.

4.3 Materials and Methods

4.3.1 *Synechocystis* culture and fractionation

Synechocystis PCC 6803, (ATCC), were grown to late log phase in Bethesda Research Labs air lift fermentors under 70 μ E light in BG11 medium. Cells were harvested by centrifuging at 3000 *g*. Fractionation was essentially as described [42] except that detergents were omitted. In brief, cells were lysed in 750 mM potassium phosphate, pH 7, using a French press and centrifuged at 25,000 *g*. This pellet was taken as the thylakoid layer. The supernatant was layered on a sucrose step gradient and spun at 136,000 *g*. Intact phycobilisomes were removed from the 1.0 M sucrose layer. The colorless layer was taken as the cytosolic fraction.

4.3.2 Protein digestion

The cell fractions were dialyzed at 4°C in 50 mM ammonium bicarbonate using a 3500 MWCO membrane to remove sucrose and phosphate. The dialyzate was then lyophilized with repeated additions of 0.1% formic acid. The lyophilized protein was denatured at 60°C and reduced with 5 mM DTT before digestion. For the cytosolic sample, the chaotropic agent was 6 M guanidine in 50 mM Tris. For the phycobilisome and thylakoid samples, it was 100% dimethyl sulfoxide (DMSO)[108]. The samples were diluted to 1% (guanidine) or 10% (DMSO) denaturant using 50 mM ammonium bicarbonate/1 mM CaCl₂ before digesting with trypsin (TRL3, Worthington Biochemical, Lakewood NJ) at 37°C for 18 hours. The trypsin was added in 3 portions at a total ratio of 1:30[120]. Peptides were desalted using C₁₈ SPE cartridges (# 20928, Alltech, State College, PA) with 0.1% TFA and recovered with 0.1% TFA in acetonitrile. The eluates concentrated using the Speed-Vac.

4.3.3 Column packing and sample loading

Columns for LC/MS/MS were 75 μ m internal diameter (ID) by 12 cm fused silica (Polymicro, Phoenix, Arizona) with flame drawn tips slurry packed with 5 μ m, 300 Å C₁₈ (Adsorboshere, Alltech #16088) in MeOH using a pressure cell. Columns were connected to the pump and subjected to a manual step gradient (from 100% solvent A to 95% solvent B

and back to 100% A in four steps each, where solvent A is 95:5:0.1, H₂O:acetonitrile:formic acid, and solvent B is 100:0.1 acetonitrile:formic acid) before the first use. Pre-columns, 150 μ m, ID were packed with 5 cm of the same C₁₈ using the pressure cell. The absorbent was retained in the pre-column using a filter (M-520, Upchurch Scientific, Oak Harbor, Washington). Pre-columns were washed 5 minutes each with 100% solvent B and 100% solvent A before loading. For two-dimensional chromatography, the pre-column was packed with 5 cm of strong cation exchange resin 5 μ m, 300 Å (Phenomenex, Torrance, California) behind 1 cm of C₁₈.

Since the phycobiliprotein absorbance interfered with both the BCA test and the Bradford test, quantitation of sample loading was difficult, but approximately 10 μ g of protein were loaded for a one dimensional chromatographic run, and approximately 40 μ g for a two dimensional run. The protein dissolved in solvent A was centrifuged at 14,000 rpm for 2 minutes or filtered through a 0.45 μ m filter. The requisite volume was placed in a 300 μ L tapered vial and loaded onto the pre-column in the pressure cell. The effluent volume was monitored to prohibit the pre-column from going dry.

4.3.4 High performance liquid chromatography

The pre-column was placed in line with the column using a union and a linear gradient of 100% solvent A to 70% B in 120 minutes was applied at a flow rate of 200-300 nL/min. A five minute stabilization time using 100% solvent A was applied before and after the gradient. A 60 minute triangular gradient (100% solvent A, 95% solvent B, 100% solvent A) was applied after every sample run (with spectral acquisition). For two dimensional chromatography, salt bumps were applied as increasing percentages of solvent C (1N ammonium formate in solvent A) for 3 minutes after the 5 minute initial stabilization time. A 5 minute desalting wash with solvent A was applied after the salt bump before starting the 120 minute gradient.

4.3.5 Mass spectrometry conditions

Mass spectrometry was performed on a LCQ Deca XP+ (ThermoFinnigan, San Jose, California) operating in data dependent mode with dynamic exclusion. The 4 most intense masses in the full mass range scan were selected for MS/MS. The exclusion limit was one minute. The source voltage was 2.5 kV. Initially the voltage was applied directly to the solvent stream to the column, but after repeated plugging it was moved to the split line (see Results, p. 85). For the 1D analyses, a run was made with scanning over the mass range 400-1700 m/z, followed by runs with scanning over the limited mass ranges 400-800

m/z, 750-1200 m/z, and 1150-1800 m/z (gas phase fractionation, GPF). For 2D runs, the mass spectrometer was scanned over the range 400-1700 m/z for each salt bump.

4.3.6 Mass spectrometric data analysis

The *Synechocystis* PCC 6803 predicted proteome, including plasmids, was downloaded from Cyanobase (<ftp://ftp.kazusa.or.jp/pub/cyanobase/Synechocystis/>, file dated 07/09/2003). A decoy database was created by reversing all the protein sequences and appending them to the true sequences. All MS/MS data were searched against this database using Sequest[31]. Search parameters were chosen to yield a 5% false positive rate which was defined as $100 * 2 * fp/total$, where fp is the number of matches to the reversed sequences and $total$ is the total number of matches[109]. The Sequest parameters yielding a 5% false positive rate in these experiments were: XCorr 1.6(+1), 2.2(+2), 3.0(+3) and $\Delta CN > 0.08$. For those loci identified by a single peptide, one or more of the spectra for the peptide were manually verified using the criteria: the base peak and major ions assigned and a continuous series of y-ions accounting for 50% of the peptide sequence; in the case of a spectrum dominated by a y-ion formed from cleavage N-terminal to proline, the corresponding b-ion was required to be present. Approximately two-thirds of single peptide identifications were rejected using these criteria.

Spectra were visualized with an in-house program, 'plotdta'. The program is called with a peptide sequence and a file of mass spectral data in the dta format as arguments. Plotdta calls the program 'theospec' [11] to calculate a table of ions. These ions are then assigned to the peaks in the dta file within a user-selectable tolerance. The labeled peaks are then plotted by calling the open source plotting program 'gnuplot' (www.gnuplot.info).

Sequest '.out' files were parsed and stored in a Postgres (www.postgresql.org) database. Scripts were written in Perl to query the database. Query results were summarized using the formalism of Yang, *et al.* in their DBParser program[161]. Briefly, proteins are classified as distinct (identified by only discrete peptides), differentiable (identified by at least one discrete peptide) or equivalent (identified by the same set of degenerate peptides), where distinct peptides identify one protein and degenerate peptides identify more than one. Subset proteins are identified by a set of peptides which may be part of a larger set from another protein. Subsumable proteins are identified by degenerate peptides that may be distributed among two or more proteins. The summary categories non-redundant equivalent (equivalent proteins counted once), maximum sum (sum of distinct, differentiable, subsumable, subset and equivalent) and parsimonious sum (sum of distinct, differentiable, subsumable and non-redundant equivalent) are also defined.

4.3.7 Bioinformatic analyses

NCBI BLAST v. 2.2.12 [5] was used locally for similarity searches of the experimentally detected proteins. BLAST results were filtered at a number of expectation values. A parameter for the length of a match was developed by summing the lengths of all high-scoring sequence pairs from the BLAST result that pass the expectation value cut-off. This sum was divided by the length of the *Synechocystis* query sequence. The query sequence was chosen since the plant target proteins may possess a transit peptide or may consist of multiple domains and so reduce the apparent match length. The non-redundant (nr) and taxonomy databases were downloaded from <ftp.ncbi.nlm.nih.gov> 22-Nov-2005. Perl routines were written to select the top hits for each query and to assign taxa. Hits to *Synechocystis* are excluded. Note that matches to *Synechocystis* paralogues are excluded by this approach as well. For detection of plant homologues, a database of plant sequences was constructed with the *Arabidopsis thaliana* and *Oryza sativa* databases from TIGR. The plant database was queried with the entire *Synechocystis* predicted proteome and with the experimental protein data set. A subset of the NCBI nr database containing no sequences from the phylum *Cyanobacteria* was also created.

TargetP [29] was used with default settings to predict plant protein localization. THMM [143] was used to predict the number of transmembrane helices. Codon adaptation index (CAI) was calculated according to Sharp and Li [134]. The reference gene set was constructed from *Synechocystis* orthologues of the genes used by Sharp and Li in their calculations. Orthologues were not found for the *Escherichia coli* proteins ompA, ompC, ompF, tufB, or lpp (BLAST E value cut-off 10^{-5}). GroEL was added to the *Synechocystis* reference protein set as an additional highly expressed protein. The Kyte-Doolittle parameter was calculated by averaging the parameters over the entire length of the protein or peptide. Hydrophathy values were taken from Mount [98]. pI was calculated with the algorithm of Tabb, described in <http://fields.scripps.edu/DTASelect/20010710-pI-Algorithm.pdf>. To validate the Kyte-Doolittle and pI calculations, the results were compared to those generated by the ExPasy server for test sequences <http://us.expasy.org/tools/>. Results were pre-computed for the *Synechocystis* proteome and stored in a database for look up.

Table 4.1: Summary of Proteins Found

Category ^a	Count
Distinct	409
Differentiable	13
Non-Redundant Equivalent	4
Equivalent	2
Subset	6
Subsumable	0
Maximum sum:	434
Parsimonious sum:	424

^aSee 4.3, p.83 for explanation of categories.

4.4 Results

4.4.1 The Experimental Proteome

The number of proteins found is summarized in Table 4.1 (see 4.3, p.83 for an explanation of the categories). A maximum of 434 proteins were found, or 424 after deducting the equivalent and subsumable proteins. This number is compared to the Gan *et al.* [38] and Norling [36, 53, 52] groups' results in Figure 4.1. Fifty-nine per cent of the proteins in this study were found by Gan *et al.* in their methods comparison, which was applied to the cytosolic fraction only. The count given is the combined result of six different multi-dimensional separations. Our count exceeds any single method of theirs, excepting one (protein IEF - peptide IEF, 467 proteins). We also found about half the proteins detected in the combined Norling group membrane protein studies.

In Table 4.2 are listed summary statistics for the proteins identified in this analysis. The median protein molecular weight and pI detected are close to those for the proteome as a whole. These parameters are further compared in the pseudo-2D gel plot Figure 4.2. The distribution of the experimental proteins mirrors that of the predicted proteome. Both the predicted and experimental *Synechocystis* proteins follow the bimodal distribution of pI seen in other bacteria. It is thought that the low pI proteins represent cytosolic proteins while the higher pI population represents membrane proteins[132]. Excepting the very low molecular weight proteins, it appears that both these compartments were sampled in a representative way. The median hydropathy results, as measured by the Kyte-Doolittle parameter, show the greatest failure to detect extremes. Those detected ranged about -1 to +1 for this parameter, while the predicted proteome ranges from -1.79 to +1.79.

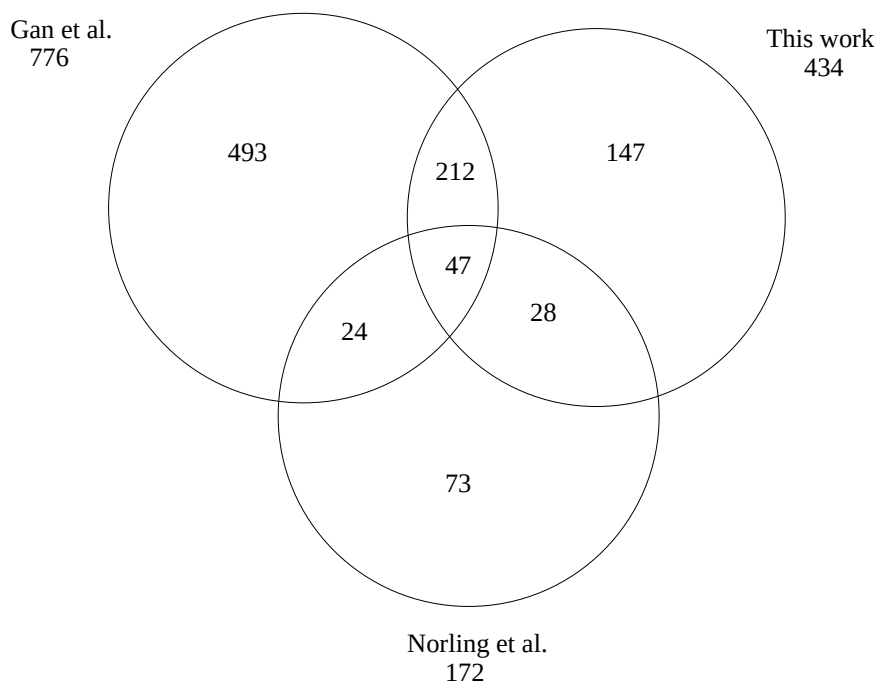


Figure 4.1: Comparison of *Synechocystis* proteomic studies.

Table 4.2: Summary Statistics for the Experimental and Predicted Proteomes

	Len ^a	MW ^b	pI ^c	K-D ^d	CAI ^e	TMH ^f
Experimental						
median:	311.5	34483.5	5.5	-0.16	0.707	0.0
avg:	352.0	38681.2	6.1	-0.16	0.706	0.4
max:	1741.0	178260.0	12.0	0.99	0.830	12.0
min:	40.0	4532.0	3.6	-0.97	0.531	0.0
Predicted						
median:	256.0	28472.6	6.1	-0.17	0.664	0
avg:	311.5	34623.1	6.7	-0.17	0.656	1
max:	4199.0	442444.2	13.0	1.71	0.853	17
min:	29.0	3397.9	3.3	-1.79	0.414	0

^aLength; ^bmolecular weight; ^cisoelectric point; ^dKyte-Doolittle parameter; ^ecodon adaptation index; ^fnumber of transmembrane helices.

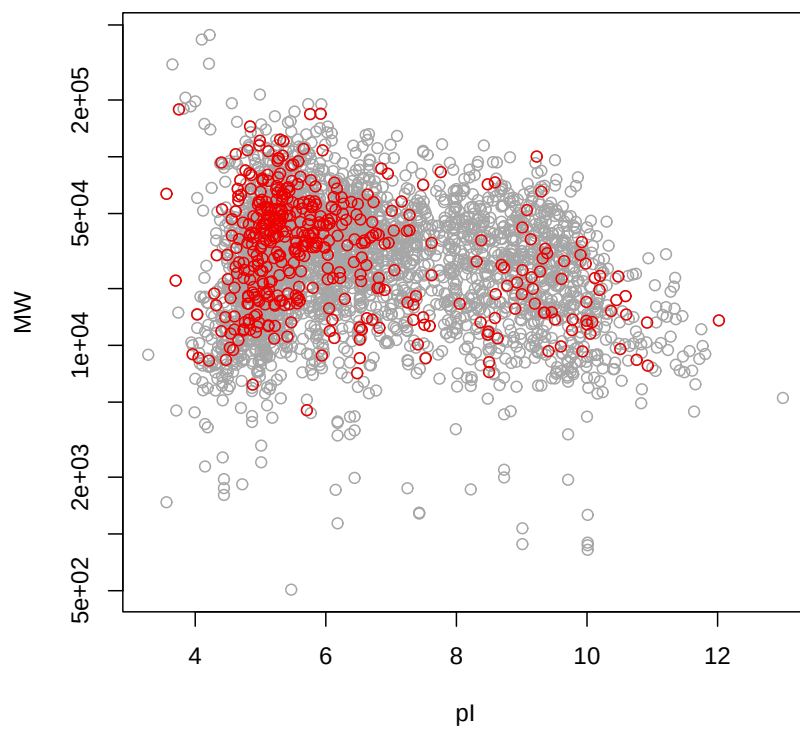


Figure 4.2: Pseudo-2D Gel Plot of *Synechocystis* Experimental and Predicted Proteins

There were 18 proteins in the predicted proteome with hydropathy less than -0.97. Six were small, very basic ribosomal proteins of 100 residues or less; the remainder were annotated as hypothetical, and most of these were very basic as well. Ribosomal proteins are often extensively modified by acetylation and methylation and this may have prevented their detection by shotgun proteomics. There were 47 hydrophobic proteins with hydropathy greater than 0.99. Most were membrane-bound photosystem proteins or porins; 25 were annotated as hypothetical or unknown. Since proteins with as many as 12 transmembrane helices (TMH) were detected, failure to detect these predominantly photosynthetic membrane proteins may indicate an incomplete isolation or digestion of the thylakoid membrane. In summary, the experimentally detected proteins seem to be representative of the proteome in regards to their chemical properties. The extremes of the measured properties are not sampled as well, but this may be due either to the isolation methods used or to a lower number of representatives for proteins with these properties.

The functional classifications of the experimentally detected proteins are summarized in Table 4.3. Ninety-one proteins, or 20% of the total found, are annotated as 'hypothetical' or 'unknown'. This is the largest category in the predicted proteome as well, with over 50% of the proteins in *Synechocystis* having these annotations. The next most frequently detected categories are proteins involved in photosynthesis, translation, and energy metabolism, consistent with the log-phase growth of these bacteria in the light. The functional classifications are illustrated as a fraction of the predicted proteome in Figure 4.3. On a fractional basis, the hypothetical proteins are detected less frequently, even though they make up more than 50% of the predicted proteome. These proteins tend to have lower codon adaptation index (CAI) and presumably lower expression (see below), which would account for their lower detection. Conversely, nearly 50% of the annotated photosynthesis proteins are detected. Similarly, high fractions of the translation and transcription machinery are found. All of the phycobiliproteins were detected, with average coverage of 59%. All of the phycobilisome linker proteins were detected, with average coverage of 38%, excepting *cpcG2* (*sll1471*). This includes the 100 kDa core-membrane linker *apcE*, which was detected with 48% coverage. This protein is located in the thylakoid membrane and anchors the phycobilisome to its surface. Since this protein is in the thylakoid membrane, the failure to detect the photosystem proteins mentioned previously must be due to their extreme hydrophobicity, rather than some problem with the thylakoid preparation.

Table 4.3: Kegg Gene Classification

Found	Proteome	%	Description
33	97	34.0	Amino acid biosynthesis
20	125	16.0	Biosynthesis of cofactors, prosthetic groups, and carriers
10	67	14.9	Cell envelope
21	79	26.6	Cellular processes
7	31	22.6	Central intermediary metabolism
5	75	6.7	DNA replication, restriction, modification, recombination, and repair
36	93	38.7	Energy metabolism
6	39	15.4	Fatty acid, phospholipid and sterol metabolism
69	1269	5.4	Hypothetical
20	366	5.5	Other categories
60	127	47.2	Photosynthesis and respiration
12	43	27.9	Purines, pyrimidines, nucleosides, and nucleotides
12	156	7.7	Regulatory functions
11	30	36.7	Transcription
65	166	39.2	Translation
20	200	10.0	Transport and binding proteins
22	679	3.2	Unknown

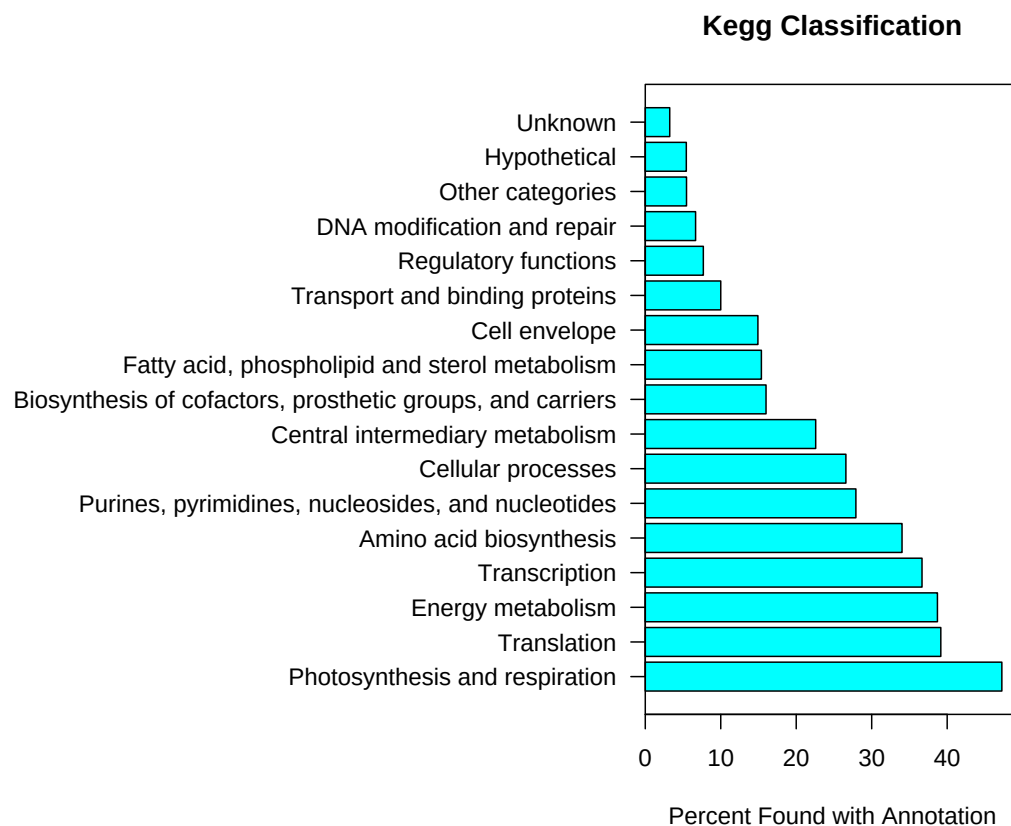


Figure 4.3: Functional categorization of proteins found. Data expressed as a fraction of those in the predicted proteome with the same annotation.

No detailed biochemical characterization of the detected proteins has been done yet, but a few observations are in order. One protein, sls5080, “non-heme chloroperoxidase”, detected with 3 peptides and 15% coverage, is encoded on the *Synechocystis* plasmid. Two other plasmid-encode proteins were detected with one peptide each (sls7085 and sls7089). These are both annotated unknown. Another protein, slr1251, “peptidyl-prolyl cis-trans isomerase”, also detected with 15% coverage, has its top BLAST hits to *Caenorhabditis*, plants, and other eukaryotes. An unrooted neighbor-joining tree shows that this protein groups in a diffuse cluster with animal, fungal, and protist homologues, while other cyanobacterial homologues group with plants (data not shown). A more careful phylogenetic analysis might reveal whether this protein is a lateral gene transfer (LGT) from *Synechocystis* to animals and fungi, or an LGT from other bacteria to *Synechocystis* with subsequent loss of the cyanobacterial gene. While most of the hypothetical proteins were detected with low frequency and coverage (see below 4.4.2), two, slr0923 and slr8847, were detected with 56% and 67% coverage. The former has a COG domain suggesting a ribosomal protein, the latter had no recognizable domain. Both were conserved in cyanobacteria, algae, and plants and may be essential proteins with unrecognized function.

4.4.2 Protein Detection Frequency is Proportional to Codon Adaptation Index

Quantitation of proteins in mass spectrometric proteomics experiments is difficult without some sort of labeling technique, but in general, more abundant proteins are detected more frequently as has been seen in this study wherein the phycobiliproteins dominate results unless they are depleted. Codon adaptation index (CAI) is a parameter that predicts the expression of a gene based on its codon usage frequency relative to a reference set of highly expressed genes. CAI was calculated for *Synechocystis* proteins using the orthologues of the original Sharp and Li reference proteins set [134]. CAI for the predicted proteome ranged from 0.414 to 0.853. The phycobiliproteins had average CAI of 0.778, higher than the 0.737 found for ribosomal proteins. The phycobilisome linker proteins had CAI almost as high, with average 0.720. Proteins with CAI as low as 0.531 were detected, but from the plot of CAI distribution in Figure 4.4, it is evident that in this experiment the proteins detected are biased towards high expression. We find that the proteomic and codon-based expression measures are related. Comparing peptide count per protein to CAI was very noisy (not shown), but binning the data in both dimensions illustrates the overall trend of increasing protein detection frequency with increasing CAI (Figure 4.5). The regression shown has $r^2 = 0.98, p < 0.0009$.

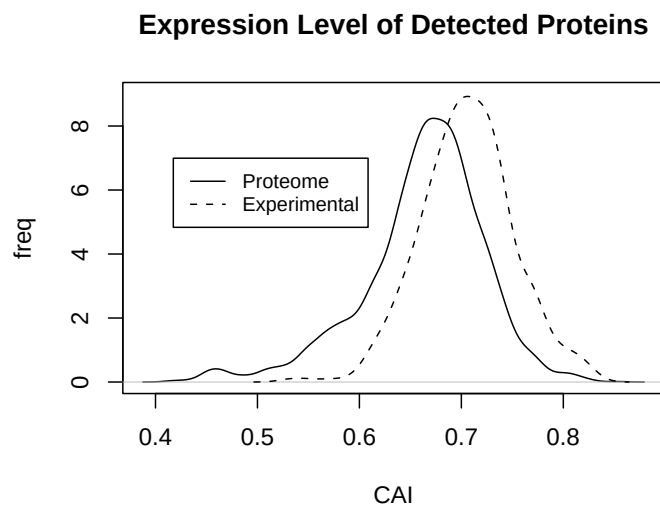


Figure 4.4: Frequency distribution of codon adaptation index for the experimental and predicted proteomes. Areas are not proportional to the number of proteins.

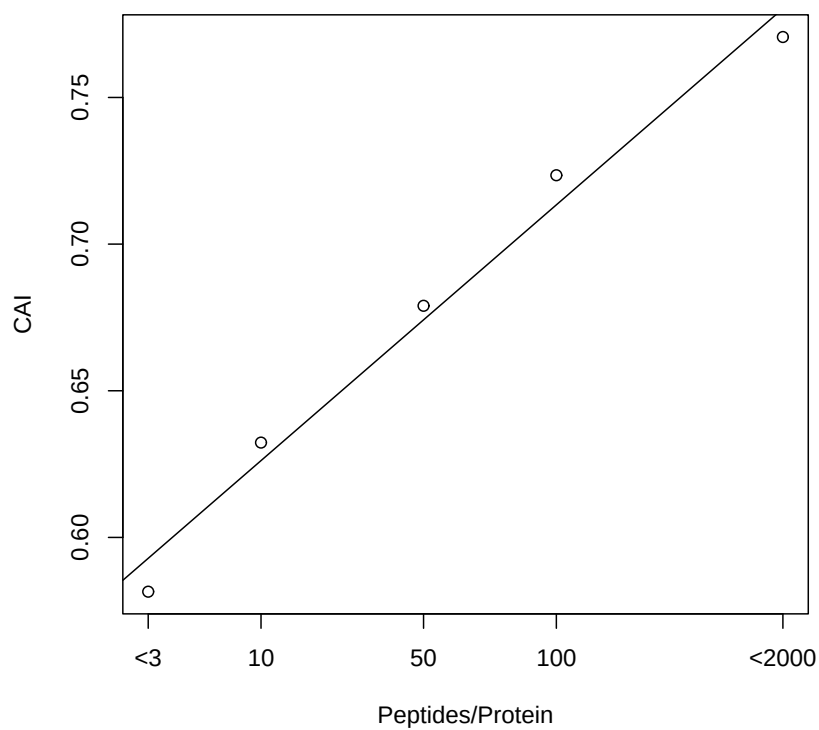


Figure 4.5: Comparison of total peptides per protein and CAI (data binned)

This plot illustrates the relationship but hides the information for individual loci. In Figure 4.6, the peptide count data are plotted as the rank (where that protein with the most peptides is rank 1 and so on). Analyzing data as ranks is beneficial in cases like this where the data are not normally distributed [140]. The linear relationship is statistically significant ($p < 2.2 * 10^{-16}$) and explains 16% of the variability. From the plot we can see that the hypothetical loci tend to be distributed more to the low CAI, low frequency of detection end of the plot, though some of the hypothetical proteins have do high expression levels in both measures.

4.4.3 Experimental proteins are more likely to be conserved in plants

Since a cyanobacterium or its ancestor is thought to be the progenitor of the higher plant chloroplast, the conservation of our protein data set in plants is of evolutionary interest. Since the actual progenitor is unknown, *Synechocystis* conceivably might be an unsuitable probe of plant proteomes. Ninety-six percent (417/434) of the experimental protein set have a top hit to other cyanobacteria when BLAST is performed against the nr database. Of the hits to Cyanobacteria, most (321/417, 77%) are to the to the parent order of *Synechocystis*, *Chroococcales*. Most of these hits in turn are to the single species, *Crocospaera watsonii* (272/321, 85%, or 62% of the total experimental data set). While the identity of the progenitor is likely never to be known, this protein data set seems to be conserved amongst extant (and sequenced) cyanobacteria.

To infer homology to plant proteins, a database was constructed which contained the predicted nuclear, chloroplastic, and mitochondrially encoded proteins for *Arabidopsis thaliana* and *Oryza sativa*. Both the entire *Synechocystis* predicted proteome and the experimental proteome were queried against this database and the results (top hit) filtered by expectation value and the length of the matching sequence. The latter was taken as the length of the summed matching high scoring sequence pairs (HSP) as a fraction of the *Synechocystis* query protein. Results are presented in Figure 4.7. The experimental proteins are more than twice as likely to be conserved in plants as the *Synechocystis* proteome as a whole, and the ratio increases at higher levels of matching stringency. The matches were evenly distributed between the *Arabidopsis thaliana* and *Oryza sativa* (data not shown). Twelve per cent of the top hits were still encoded in the plastid genome ($E = 10^{-5}$, $HSP\% = 40$). This experiment was repeated with a target database consisting of the NCBI nr database having all cyanobacterial sequences removed. The total number of matches to plants was reduced by matches to other bacteria, but the difference between the experimental and

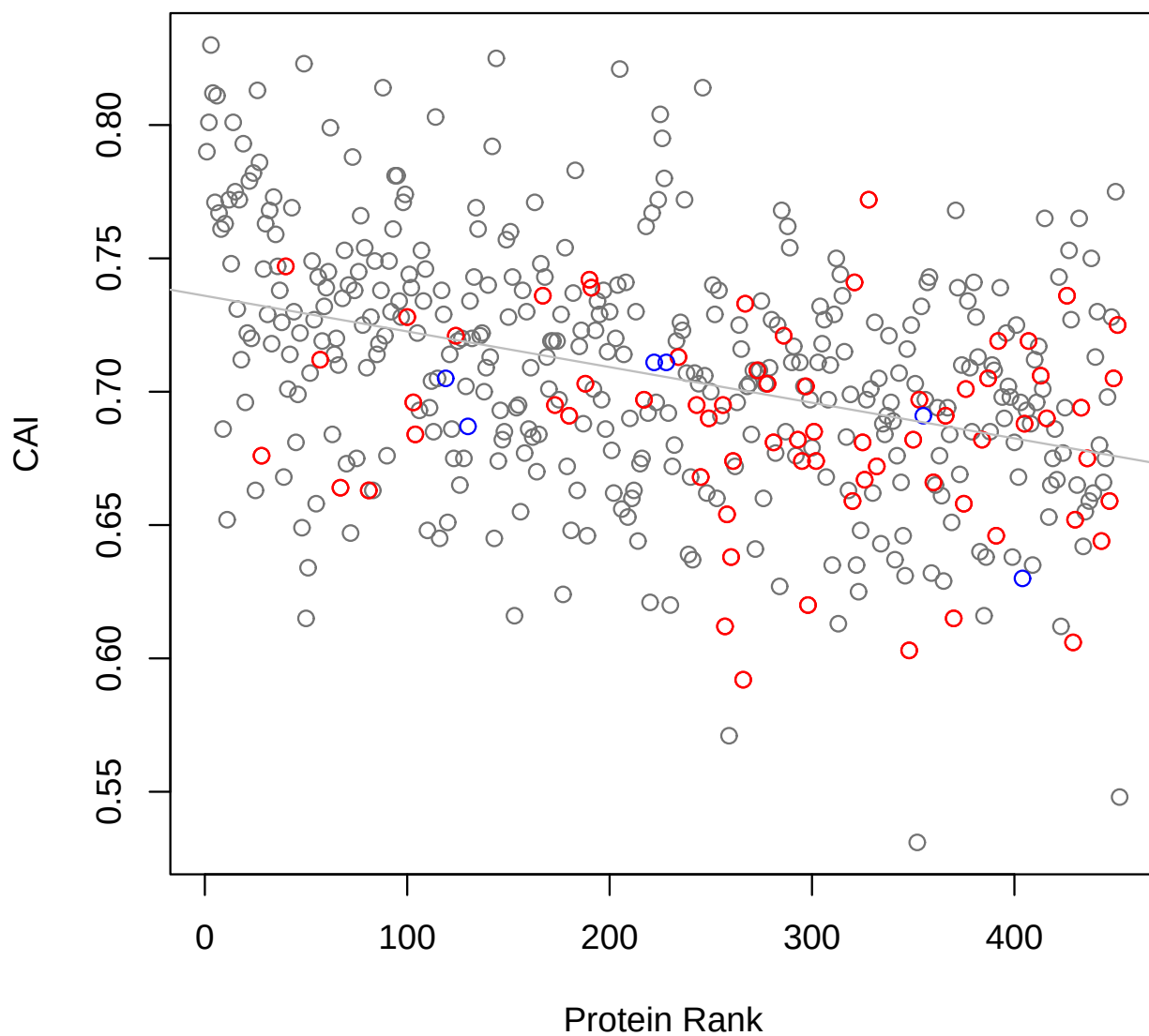


Figure 4.6: Rank distribution of codon adaptation index for the experimental and predicted proteomes. Hypothetical proteins are in red, putative in blue.

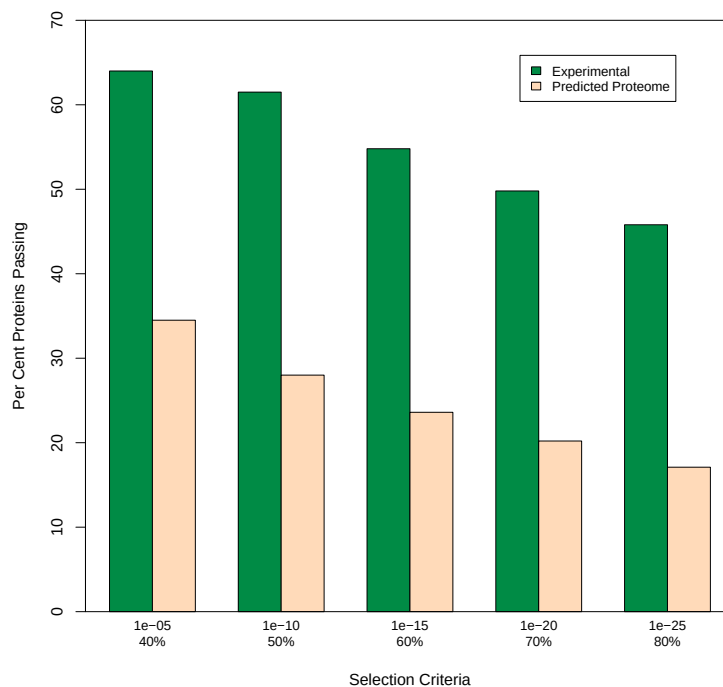


Figure 4.7: BLAST results for predicted and experimental proteome with different filtering criteria. Top label of the ordinate is the expectation value; the bottom label is the length of the matching sequence (HSP) as a percentage of the query protein length.

predicted proteomes was the same (data not shown). It thus appears that the more highly expressed proteins detected by proteomics are preferentially conserved in plants. One final experiment supports this conclusion. In Figure 4.8 the percent conservation in plants is plotted for the lowest third and the highest third of the experimental proteins sorted by CAI. The proteins in the higher CAI fraction, which also had the higher frequency of detection, are more likely to have a plant homologue. The top third is enriched in proteins involved in photosynthesis, energy metabolism, and translation, while the bottom third is has a much higher proportion of hypothetical and unknown proteins. Among those proteins with relatively high expression and no plant homologues are the entire phycobilisome apparatus and the carbon dioxide concentrating mechanism proteins. No homologues were detected for the iron transport proteins bacterioferritins (slr1890, sll1341), iron transport system substrate-binding protein (slr0513), and futA1 (slr1295). The photosystem components PsbU and PsbV (sll1194, sll0258), and the water-soluble carotenoid protein (slr1963) also had no plant homologues detected.

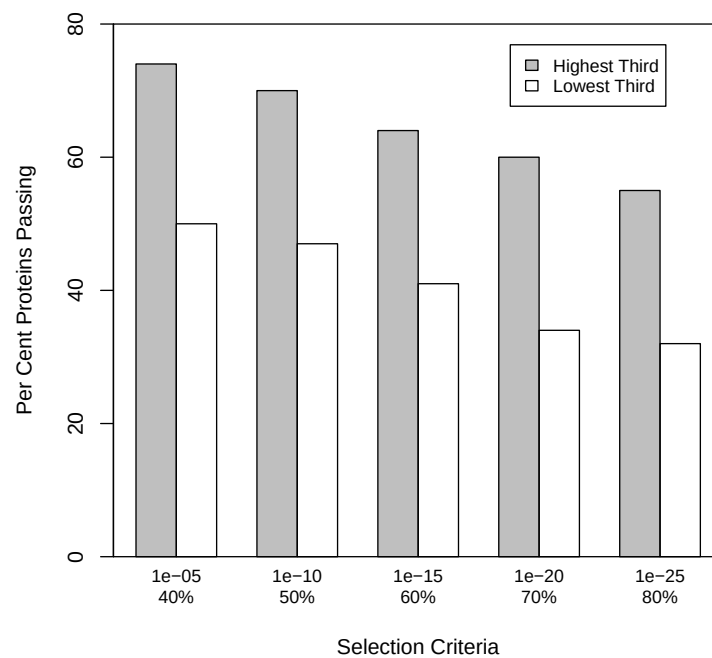


Figure 4.8: Proteins with higher CAI are more likely to have plant homologue. The experimental proteome was ordered by CAI. The fraction of the top and bottom one-third having a plant homologue at the given filtering criteria are plotted.

4.5 Discussion

The major problem for this analysis was depleting the abundant phycobilin proteins. These proteins compose the light-harvesting phycobilisome apparatus on the surface of the thylakoid membranes and may constitute 30-50% of the protein mass of cyanobacteria. Analysis of the unfractionated *Synechocystis* proteome by one-dimensional LC/MS yielded at most two dozen proteins¹. Even after removal of intact phycobilisomes by Na₂PO₄/sucrose density gradient centrifugation, peptides from phycobilisomes proteins were often the most frequently identified in an experiment. The procedure used here for depleting the phycobilisome proteins is designed to isolate pure, intact phycobilisomes, not necessarily to remove them completely. The remaining phycobiliproteins are probably due to phycobilins that were in a dissociated state *in vivo* or were disaggregated during fractionation. Application of both 2D HPLC and gas-phase fractionation (GPF) were also applied to detect lower abundance proteins and to reduce the influence of the highly abundant phycobilins. In addition to problems caused by phycobiliprotein abundance, initial runs were plagued by repeated plugging at the upstream end of the column (columns could be restarted by trimming the butt end) or in the tee where the electrode was inserted, despite the samples having been both centrifuged and filtered through 0.45 μ filters before loading into the pre-column. Since the plugged area was frequently blue, it was presumed that oxidation of phycobilisomes was occurring at the electrode. The plugging problems were completely eliminated by moving the electrode into the split line.

4.5.1 Proteomics as expression analysis

Several factors conspire to make quantitation of proteomic results difficult. The ionization efficiency of a given peptide may depend in unpredictable ways on its sequence or on the chemical environment during the electrospray process. Some proteins may not be represented because of extreme chemical properties as was noted above for the very hydrophobic photosystem proteins or very low molecular weight hydrophilic proteins. The biochemical isolations or the digestion process may likewise discriminate. However it is axiomatic in proteomics that abundant proteins dominate the analysis. These proteins are usually already well known, and so the field is driven to produce higher identification rates and lower detection limits to describe the proteome in finer detail. Liu *et al.* have shown that there is a linear relationship between total peptides detected by mass spectrometry and the amount of standard proteins added to yeast proteome digests over two orders of

¹Gan *et al.* [38] reported 10 proteins for a 1D LC/MS run of a *Synechocystis* cytosolic sample.

magnitude [81]. In a similar experiment, Ishihama *et al.* demonstrated an exponential relationship between protein concentration and the number of unique peptides per protein (normalized by the number of observable proteins). So in controlled conditions it is possible to extract quantitative results from proteomic experiments. Proteomic data has also been related to gene expression or mRNA transcript abundance using a variety of parameters [124, 157, 22, 45]. Since the codon adaptation index (CAI) has been related to protein and RNA expression [43], we sought a relationship between this genomic parameter and our proteomic data. We find a strong logarithmic relationship between total peptides per protein and CAI with binned data. We find a weaker, though statistically significant, linear correlation between the protein rank, when ranked by the number of peptides, and the CAI. While this work was in progress, another report described a very similar relationship [158] for a proteomic analysis of *Mycobacterium smegmatis*. This group also found a linear relationship between rank and CAI, though in this case, the rank was taken as the number of experiments (out of 25 different growth states) in which the protein was detected. The regression coefficient in the Wang *et al.* experiment was 0.74 for data binned over the 25 experiments, whereas that for our binned count-CAI relationship was 0.97. These are probably comparable results given the disparity in the experimental details. Our rank-CAI relationship while not as statistically satisfying was useful in understanding some features in the proteomic data, illustrated above with the superposition of the hypothetical proteins. CAI as a simple estimate of gene expression has a questionable provenance. The organisms for which CAI predicts expression are fast-growing organisms that are probably under selection pressure for translation efficiency to rapidly exploit their environment. CAI is also known to be related to other gene properties such as length, where presumably codon optimization prevents ribosomal detachment during translation of a long mRNA [32]. Other codon preference dependencies such as this would obfuscate the correlation with translation efficiency. *Synechocystis* is a fairly fast-growing bacterium and this experiment exhibits a correlation between protein abundance and CAI.

4.5.2 Conservation in plants

A goal of this experiment was to develop a set of expressed proteins with which to examine the evolutionary relationship of cyanobacteria and the higher plant plastid. In the next chapter, we use the protein data set presented here to elucidate the process of transit peptide acquisition in *Viridiplantae* homologues. Here we have made a preliminary analysis of the conservation of the data set in plants. Most analyses of plant-cyanobacteria endosymbiosis have been plant-centric. It is of interest to take a 'cyano-centric' approach to

find those metabolic activities of free-living cyanobacteria that were retained or rejected in plant evolution. The more highly expressed proteins detected in this experiment are more than twice as likely to have a plant homologue as the *Synechocystis* predicted proteome as a whole, implying that the gene complement employed in the log-phase autotrophic growth state were more likely to be useful to the evolving plant. Furthermore, within the experimental data set, those proteins with higher expression, as measured by CAI or protein count, were more likely to be conserved in plants as those with lower expression. The high expression proteins are dominated by photosynthesis, energy expression, and translation proteins. Most of the photosystem was retained in plants, except for PsbU and PsbV, for which no higher plant homologues were found. The most highly expressed proteins by either measure, those of the phycobilisome, were totally abandoned by higher plants. The carbon-dioxide concentrating mechanism was also abandoned. It is interesting that the bacterioferritins, which are critical to proper photosystem function[68], and two other proteins involved in iron transport were abandoned, and replaced by the eukaryotic ferritins which are apparently targeted back to the chloroplast[112]. While more highly expressed proteins may be preferentially conserved on an absolute basis, there are still approximately 700 genes in the *Synechocystis* genome with plant homologues for which no protein was detected. These may have merely been missed because of the limitations of proteomics pointed out above, or may be expressed in other growth states. It would be interesting to repeat this experiment with another growth state, such as a dark-acclimatized *Synechocystis*, and ascertain whether the expressed proteins are preferentially conserved as well.

Chapter 5

Stochastic Nature of Transit Peptide Acquisition

5.1 Abstract

A critical part of the transformation of endocytosed prokaryotes to eukaryotic organelles was the evolution of an import apparatus for proteins translated in the cytosol and targeting sequences to direct the protein to the import apparatus. The most frequently advanced hypothesis for the acquisition of the targeting sequence is acquisition as a module via exon swapping. Using bioinformatic methods applied to genomic and proteomic data, we find no correspondence between exon boundaries and targeting sequence cleavage sites. In addition, randomly generated sequence is found to be very similar to targeting sequence and a large proportion is predicted to be imported in the mitochondrion and chloroplast.

5.2 Introduction

Most of the studies of organelle functioning in the eukaryotic cell have dealt with the mechanisms of protein trafficking and the composition of the import machinery. In contrast, much less work has been directed toward the origin of the targeting sequence itself. This is probably due to the fact that rugged systems for studying import were worked out before high throughput sequencing methods provided enough data to make inferences about the latter phenomenon. In another sense, protein import is contemporaneous; the organelle is continually importing the proteins necessary for function, whereas the genetic events involved in organelle DNA export to the eukaryotic nucleus were presumed to belong to paleohistory. Because of the increasing availability of genomic data, recent studies have

shown to the contrary that the export of organellar DNA to the nucleus still occurs, and frequently[51, 90, 144, 154]. The following general steps have been outlined for gene transfer to the nucleus[13]:

- Export of the gene to the nucleus
- Coexistence of the nuclear and organellar copies
- Acquisition of a targeting sequence and regulatory elements
- Loss of the organellar copy

Here we focus on the third step, acquisition of the targeting sequence.

The most frequently advanced hypothesis for transit peptide origin is gain by exon swapping with another locus[89]. Examples are known for both mitochondria and chloroplasts[39, 82]. In this theory, the newly arrived gene in the nuclear DNA would receive the targeting sequence by attaching an exon containing it to the 5' end as a module. This implies that transit peptide cleavage sites should be at exon boundaries or that the exon contain the transit peptide and hence be longer. It would also imply that the transit peptides attachment of plant homologues of the contributing cyanobacterial sequences would fall outside the primordial gene start. In this chapter, we test each of these predictions with bioinformatic experiments. Since the results of these experiments imply a random process of targeting sequence attachment, a final *in silico* experiment was performed to test the import of randomly generated sequence.

5.3 Materials and Methods

By 'filter' or 'parse' is meant an *ad hoc* Perl script to transform data from one format to another or to extract the desired information. All data visualization and statistical calculations were done in the 'R' statistical package (www.r-project.org/) unless otherwise noted.

5.3.1 Comparison of exon and targeting sequence length

A database of information on introns and exons has been created by Walter Gilbert to support his research on the origin of introns[127]. While not experimentally verified, the data are filtered from the GenBank non-redundant database by a number of methods including the requirement for a mRNA transcript[127]. This 'gb140.pEID' database was downloaded

from <http://hsc.utoledo.edu/bioinfo/eid> in June 2005. Filters were written to extract *Arabidopsis* protein sequences from the database and to extract the length of the first exon from the information added to the fasta annotation line. Since several variants of a protein may exist in GenBank, each *Arabidopsis* sequence from the exon-intron database was searched for the best BLAST match to the annotated sequences in the database from The Arabidopsis Genome Initiative (TAIR, <http://www.arabidopsis.org/>). TargetP localization predictions for the entire *Arabidopsis* proteome were downloaded from the Munich Information Center for Protein Sequences (MIPS, <http://mips.gsf.de/>). Finally, the exon data keyed to the correct sequence were joined to the TargetP predictions for comparison of the TargetP-predicted cleavage site and the first exon.

The TargetP training sets `planta.cTP.141.rr.fasta` and `planta.mTP.368.rr.fasta` were downloaded from <http://www.cbs.dtu.dk/services/TargetP/datasets/>. The sequences in the training set have been manually evaluated to verify subcellular location and targeting peptide length[29]. The targeting peptide cleavage locations were parsed from the fasta header. The gb140.pEID database was used to find the intron boundaries for the intron closest to the targeting sequence.

5.3.2 Determination of transit peptide attachment points

The presumed ‘attachment points’ of plant targeting peptides to the cyanobacterial homologues were calculated in the following way. The data set of expressed cyanobacterial proteins described in Chapter 4 was searched against the nr database with cyanobacterial sequences removed (p. 84). A hit was retained if it had an initial expectation value less than $1 * e^{-5}$ and it belonged to a plastid-containing organism. A similarity measure was constructed by summing the length of the high scoring sequence pairs (HSPs) and dividing by the length of the query sequence. The BLAST results were then tabled for more stringent filtering later. TargetP was used to predict the localization and transit peptide length for each hit at this stage.

Each hit that had a predicted targeting sequence was then pairwise-aligned with the cyanobacterial query sequence using ClustalW with alignment parameters: `type=protein`, `pwmatrix=blosum`, `pwgapopen=35`, `pwgapext=0.72`. The results were saved in a separate file for each query-hit pair. The TargetP-predicted transit peptide length was then mapped onto the cyanobacterial query sequence. If this point is within the cyanobacterial sequence, the number of residues was counted and the sign taken as negative. Thus, TargetP predicts a transit peptide length of 5 for Os02g19450:

```

                Last residue of transit peptide
                |
Os02g19450  ----MPHRTSKMRRLPRTSRRR-----TALHRRPPP
s110020     MFERFTEKAIKVIMLAQEEARRLGHNFGTEQILLGLIGEGTGVAAKVLKSMGVNLKDAR
                ...: *: *: ::**                ...: : .
                |
                Attachment point

```

The attachment point is taken as the residue to which the transit peptide would have been attached in the aligned sequences. This is numbered -10, the length of MFERFTEKAI. If this sequence contains gaps, they are ignored, i.e. only residues are counted.

If the point fell outside the query sequence, the intervening residues are counted and the sign taken as positive. For example, TargetP predicts a transit peptide length of 76 for Os03g31300:

```

Os03g31300  MAAAPPLAAGLRPAMAAAQAPVVAAGVGARRGAALSSSARCRALRLSRGGGGGRDGWV
s110020     -----

Os03g31300  PPPVVGRRMPRTL SVRCAASNGRITQEFTEMAWQSIVSSPEVAKESKHQIVETEHLMKS
s110020     -----MFERFTEKAIKVIMLAQEEARRLGHNFGTEQILLG
                                ..*** * : *: : * *:. *:* **::: .
                                |         |
                                Attachment point      Query start

```

The attachment point is numbered up to but not including the 'M', so the sequence CAAS-NGRI in this case with length 8 and attachment point = +8. Gaps are treated as above. Note that for this numbering system there is no 'zero' attachment. For the final data set considered here, only those sequences with the plant homologue having a TargetP reliability class prediction of 1 or 2 and a BLAST match with expectation value less than 1×10^{-15} with similarity greater than 60% were considered. The results were not further reduced by homology, so some redundancy exists in the final data set because of similar sequences in the nr for a given locus.

The domain organization of multi-domain proteins was determined from the NCBI conserved domain database <http://www.ncbi.nlm.nih.gov/Structure/cdd/>.

5.3.3 Calculation of random targeting sequences

A script was written to generate random peptide sequences by indexing into the standard codon table with a random number. Either fixed lengths or random lengths between a minimum and maximum could be generated. The TargetP training set *planta.cTP.141.rr.fasta* was used as a data set of known chloroplast sequences. The transit peptide length was obtained from the annotation and used to select the mature sequence. These mature sequences were randomly selected (with replacement) and joined to the randomly generated targeting sequences. These pseudo proteins were then submitted to the TargetP server for localization prediction.

Additional random sequences for targeting prediction were selected from *Arabidopsis* intergenic sequence as follows. Intergenic sequence in the file *TAIR6_intergenic_20060907* was download from TAIR (ftp://ftp.arabidopsis.org/home/tair/home/tair/Sequences/blast_datasets). A random line was selected in the file and codons selected in the upstream direction until either a stop or a start codon was encountered. If a stop codon was encountered, the sequence was discarded; if a start, the sequence was retained if it met length criteria. This process was repeated until the required number of sequences were obtained. The sequences were then joined to randomly selected TargetP chloroplast sequences as above for localization prediction by TargetP and Predotar[139].

5.4 Results

5.4.1 Correlation of targeting sequence length and exon boundaries

Targeting sequence lengths were compared to the lengths of the first exon for proteins predicted to be localized to chloroplasts and mitochondria in *Arabidopsis thaliana*. TargetP predictions for the entire cress proteome were merged with the curated exon-intron database gb140.pEID[127]. The resulting data set is summarized in Table 5.1. The total loci are less than the annotated *Arabidopsis* proteome due to the filtering process to generate the EID database. There were also 350 entries in the MIPS localization database that did not have corresponding entries in the EID database. This was probably due to the inability to link the GenBank names in the EID database to the AGI names used by MIPS and TAIR. The fraction of organelle loci with one exon (no introns) is about the same as or a little larger than the genome as a whole. If exon swapping is necessary for gain of a transit peptide, it must do so without insertion of an intron, since this category is not depleted

Table 5.1: Arabidopsis Exon Data

Location ^b	Total	Number of Exons ^a					
		1		2		>2	
Mitochondrion	2423	516	(21.3%)	360	(14.9%)	1547	(63.8%)
Chloroplast	3420	658	(19.2%)	391	(11.4%)	2371	(69.3%)
Secretory	4081	948	(23.2%)	773	(18.9%)	2360	(57.8%)
Other	16283	2989	(22.5%)	2457	(18.2%)	10837	(66.5%)
Total	26207	5111	(19.5%)	3981	(15.2%)	17115	(65.3%)

^aFrom the exon-intron database[127]; ^blocalization prediction by TargetP.

relative to the entire genome. It might be expected that if an exon were attached to a newly inserted organelle gene, very few of which have introns, there would be an excess of two-exon (one intron) genes. The fraction of loci with two exons is found to be less than that for the genome as a whole (Table 5.1, columns 5 and 6). It is interesting to note that the intron distribution of the organelle genes does not appear to differ very much from the *Arabidopsis* genome as whole for all categories. Apparently, in the time since endosymbiosis, the formerly intronless prokaryotic proteins have been subject to the same intron generating mechanisms as the eukaryotic host.

Summary data for the first exon length and the predicted transit peptide length are given in Table 5.2. Neither of these parameters has a normal distribution, so the medians are a better measure of the center of the distribution. The median targeting sequence lengths are near the typical values for each cell compartment. The median exon lengths are longer in each instance than the corresponding targeting sequence lengths, 46 residues for the mitochondrion and 55 residues for the chloroplast. The organellar exon lengths are also in the order of their the typical transit peptide lengths: chloroplast > mitochondrial > secretory. The summary data indicate therefore that the first exon is long enough to contain the targeting sequence, but the boundaries of the two elements do not coincide as might be expected from the modular transit peptide concept.

For every *Arabidopsis* gene that had both an exon structure prediction and a transit peptide prediction, the difference in the first exon length and the targeting sequence length was taken. The distribution of these differences for chloroplast and mitochondrial predicted localization is presented in the histogram of Figure 5.1. The maxima for each distribution corresponds to the differences in the median lengths mentioned earlier.

Table 5.2: Exon and Targeting Sequence Length

Location	Exon 1 Length ^a		TP Length ^b	
	Mean	Median	Mean	Median
Mitochondrion	113.5	82	45.9	36
Chloroplast	131.0	103	47.3	48
Secretory	129.9	88	62.2	25
Other	107.5	74	–	–

^aFrom the exon-intron database[127]; ^btargeting peptide length as predicted by TargetP.

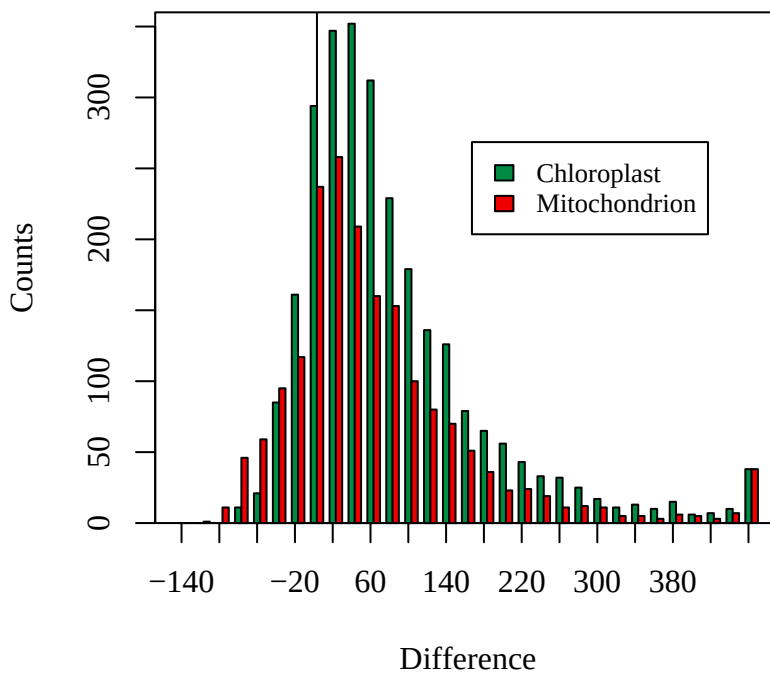


Figure 5.1: Difference in first exon length and TargetP-predicted targeting sequence length, *Arabidopsis* predicted proteome.

Both distributions are log-normal, tailing to large differences and indicating that the first exon is usually longer than the targeting sequence. There appears to be no particular excess of differences near zero as might be expected from the exon-swapping hypothesis.

The experiments described so far have used predictive data for both the exon length as well as the transit peptide length and localization, and thus may be subject to the biases and errors of the particular prediction tool. The last experiment was repeated using the proteins in the TargetP training set. The localizations and lengths of the targeting sequences are based on experimental data and were further validated by the authors of TargetP[29]. In these data, instead of the first exon boundary, the difference was taken between the nearest exon boundary and the targeting sequence cleavage site. For the chloroplast data, there were 34 sequences with matching exon data, for five of these the second exon boundary was closest. For the mitochondrial data, 90 had exon data and for 10 the second exon boundary was closest. The distributions of these data are presented in Figures 5.2 and 5.3. The center of each distribution has moved closer to zero, but they are still shifted to the right. The median difference in exon and targeting sequence lengths are 13 residues for the mitochondrial sequences and 20 for the chloroplast sequences. These are much smaller than the 46 and 55 found for the predicted values above, perhaps because the nearest exon boundary was taken instead of the just the first. For only one of the genes in this data set, At1g42970 (P25857, glyceraldehyde-3-phosphate dehydrogenase) did the exon boundary (exon 1 in this case) correspond with transit peptide cleavage site. These data are therefore consistent with the previous experiments; there does not appear to be any evidence for a correspondence of the exon boundaries with the targeting sequence cleavage site.

5.4.2 Distribution of predicted transit peptide ‘attachment points’

As another test of the modular transit peptide hypothesis, we used the expressed cyanobacterial protein set conserved in plants discussed in Chapter 4 as surrogates for the endosymbiont proteins that were transferred to the eukaryotic nucleus. These were aligned with their plant homologues and the TargetP-predicted cleavage site mapped to the cyanobacterial sequence. If this site fell within the cyanobacterial sequence, i.e. C-terminal to the initial methionine, the attachment is taken as negative. If the location of the cleavage site mapped outside the initial methionine, the attachment is taken as positive. Only sequence pairs in which the TargetP chloroplast localization prediction had a reliability class of 1 or 2 and the BLAST expectation value was less than $1e^{-15}$ were considered. The distribution of attachment points for the resulting 1341 proteins are plotted in Figure 5.4.

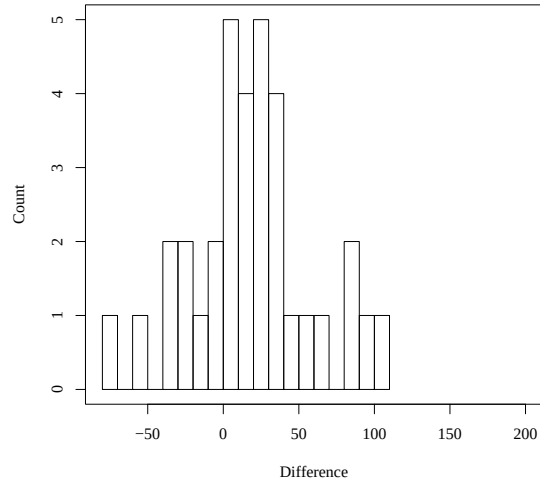


Figure 5.2: Difference in nearest exon boundary and targeting sequence length for TargetP training set chloroplast sequences.

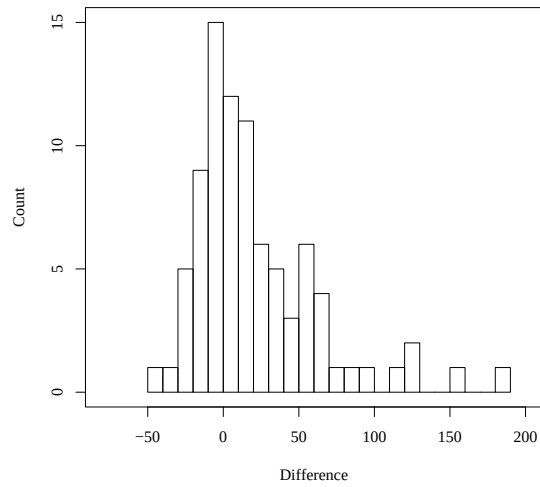


Figure 5.3: Difference in nearest exon boundary and targeting sequence length for TargetP training set mitochondrial sequences. Four mitochondrial differences were greater than 200 residues and are not included.

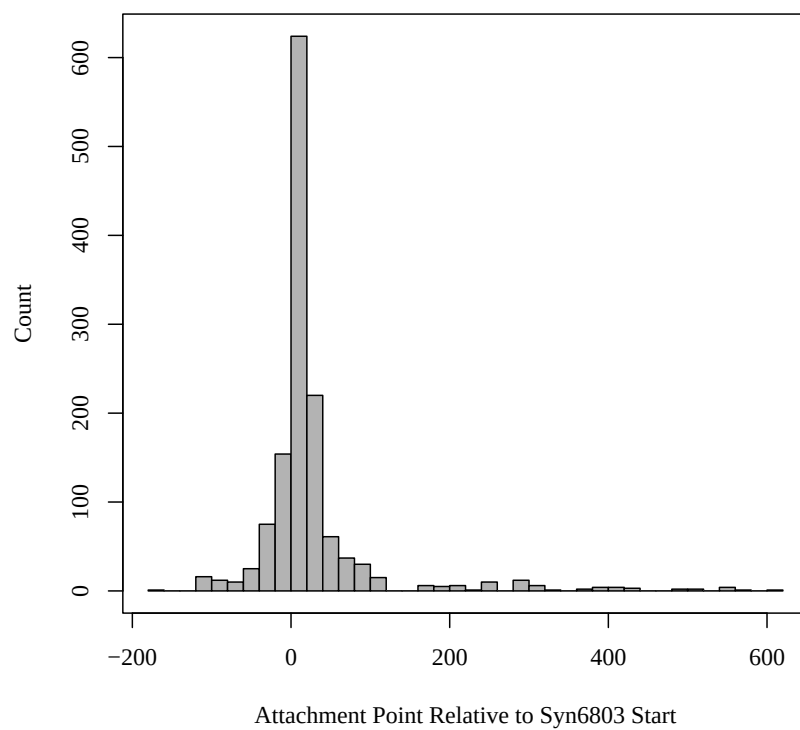


Figure 5.4: Distribution of attachment points. See Materials and methods for determination. Negative values are N-terminal to the cyanobacterial start; positive values are outside the cyanobacterial start.

The alignments of proteins in the long tail of Figure 5.4 with attachment points greater than about 125 residues were confirmed manually and consist of large multidomain proteins within which the cyanobacterial protein aligned. Several starch synthase III (SSIII) genes are found in this group, for instance, *Synechocystis* sll1261 glycogen transferase aligns with the C-terminal transferase domain *Arabidopsis* of At1g11720 with 90% similarity. This domain is preceded by three SSIII domains, which pushes the attachment point of the predicted 80 residue transit peptide out 436 residues. Other proteins in this group are a homologue of the *Synechocystis* elongation factor EF-TS, in which the EF-TS domain in the chloroplast is preceded by two ribosomal S1 domains[8]; the acetolactate synthase homologue of sll0065 has a duplicated synthase domain; finally, the *Oryza sativa* β -glycyl-tRNA synthase region which aligns with slr0220 glyS is preceded by an α glyS region. This subset of proteins with very large attachment points is therefore due to the alignment of the cyanobacteria protein with particular domains of multi-domain eukaryotic homologue.

The central part of the distribution is presented in Figure 5.5. It appears to be roughly Gaussian, but has both skew and kurtosis. The flattened region with negative attachment contains 290 proteins. *Synechocystis* slr1643, petH had several homologues in this region, all with predicted transit peptide lengths of 45 to 65 and attachments -93 to -116. The petH sequence for pea is experimentally known, and the predicted transit peptide length of 51 was correctly detected in this experiment. Several plant ftsZ homologues of sll1633 were found with predicted transit peptide lengths of 28 to 96 and attachments from -39 to -106, but no experimentally verified transit peptides could be found for corroboration. An alignment of several of these sequences is presented in Figure 5.6. For *Medicago trunculata* and *Pisum sativa*, the predicted transit peptide cleavages fall within one residue on the aligned positions. For *Solanum tuberosum* and *Nicotiana tabacum* the cleavage sites correspond, but much further into the highly conserved region. *A.thaliana* is intermediate. Since more closely aligning sequences have similarly predicted cleavage sites, it gives some reassurance that these purely bioinformatic predictions are at least consistent. All the predicted attachment points for this protein were negative, indicating either that the module containing the transit peptide was swapped into the ftsZ gene upon arrival in the nuclear DNA, or the N-terminal sequence mutated into sequence suitable for targeting. The negative side of the distribution is flattened relative to the positive side, since modification in this direction must be limited by the necessity of maintaining the integrity of the functional domains of the protein. The distribution is skewed towards positive attachments (981 proteins), since there is no such limitation in this direction.

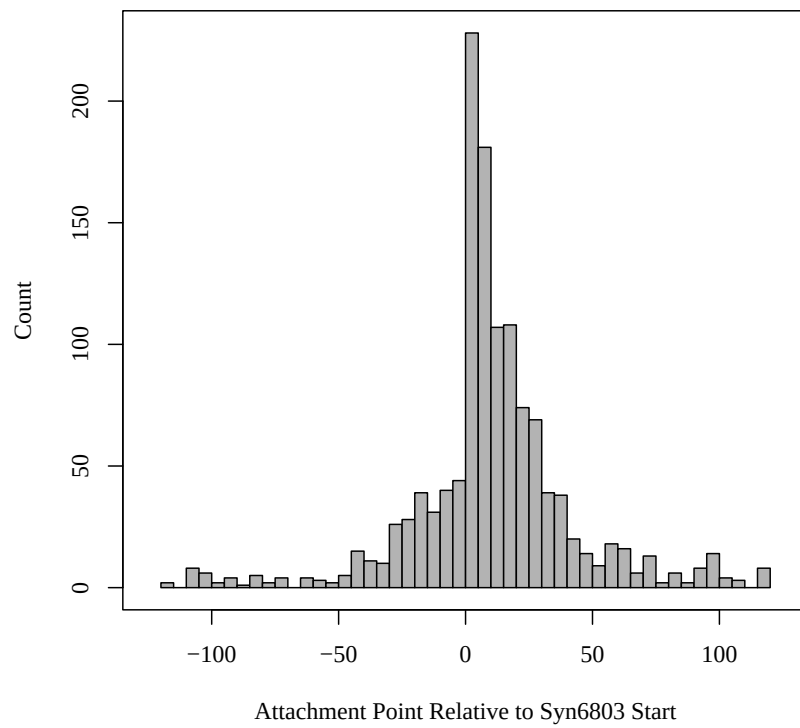


Figure 5.5: Distribution of attachment points, truncated plot. See caption for Figure 5.4.

Mtrunc	-----mvttllpssltnpnkllshsslfhnsslsts-----hsvslypktqrftrr	46
Psativ	-----matllpstisnnpnkltssylfhnaslstspssltttsvsiypktqrfgr	51
Stuber	-----mailglspaelasspsssltfshrlhtsfipkqcfftgvrrksfcrpq	49
Ntabac	-----matisnpaeiaasspsfafyhssfipkqcctkarrkslckpq	43
Athal	maiiplaqlneltisssssssfltksisshslhsscicassrisqfrggfskrrsdstrsk	60
sll1633	-----MTLNNDLPLNNIGFTGSGLNDGTEGLDDLFSSSIPVDNEPLEALVETPTFASPS	53
	:	
Mtrunc	fgsvkcsLAYVDNAKIKVVGIGGGGNNAVNRMIGSGLQGVD FYAINTDAQALLHSA AENP	106
Psativ	fgsvrcSLAYVDNAKIKVVGIGGGGNNAVNRMIGSGLQGVD FYAINTDAQALLHSA AENP	111
Stuber	rfsisssftpm dsakikvvvgvgggggnavnrmigsglqgvdfyaintdaqalvQSA AENP	109
Ntabac	rfsisssftpf dsakikvigvgggggnavnrmigsglqgvdfyaintdaqallQSA AENP	103
Athal	smrlrcsfspmesarikvigvgggggnavnRMISSGLQSVDFY AINTDSQALLQSSAENP	120
sll1633	PNLKR DQIVPSNIAKIKVIGVGGGGCNAVNRMIASGVTGIDFWAINTDSQALTNTNAPDC	113
	: : *:***:~:**** *****. **: .:~*:~****:~*** :~: * :	

Figure 5.6: ClustalW alignment of selected ftsZ sequences. The TargetP predicted transit peptide is shown in lower case.

The peak of the distribution is around +10 or +12 residues . In sum, the distribution of attachment points appears to be approximately normal, but suppressed in the negative direction. The center, or mode, is about 10 residues longer than the origin of the cyanobacterial ancestor proteins.

5.4.3 Predicted targeting specificity of random sequence

The exon-transit peptide boundary difference distribution and the attachment point distributions both appeared to be normal or log-normal, suggesting a stochastic process for acquisition of the transit peptide. Here we test a model that supposes a gene inserted into nuclear DNA and acquires a targeting sequence from upstream sequence¹. Random DNA sequences were produced by indexing into a codon table with a random number. Since the base composition of genomes is known to be biased and not totally random, it might be expected that plant intergenic sequence would improve targeting to the chloroplast in a random insertion model. A second set of sequences was created by selecting a random point in a file of *Arabidopsis* intergenic sequence and selecting codons in the upstream direction until a start codon is found. If a stop codon is encountered, the sequence is abandoned and selection continues. Surprisingly, only 1021 attempts were required to get 500 sequences between 10 and 100 residues (maximum length obtained was 52). The residue frequencies obtained are compared to those for the TargetP organelle frequencies in Figure 5.7. The overall trends in the frequencies are remarkably similar. Tryptophan is the least frequent residue in targeting sequences and in both sets of random sequence. The acidic residues aspartate and glutamate are the next lowest, even though the random frequencies are about twice the organelle frequencies. Phenylalanine, which was investigated for its importance to import in Chapter 3, has a frequency identical to the random sequence and somewhat less than the intergenic sequence. The basic residues lysine and arginine, which are a distinctive component of targeting sequences, are modeled well by random sequence, as is leucine. Another distinctive residue, serine is one of the highest frequency residues in the random and intergenic proteins, though the value is somewhat lower than that found in the organelles. The biggest discrepancy seems to be for alanine, which occurred at lower frequencies in the random and intergenic sequences. Overall, the residue frequencies of purely random sequence are closer to the organelle frequencies than those for the intergenic sequence.

In Table 5.3 are presented the TargetP localizations for 500 random sequences of length 10-100, attached to mature mitochondrial protein sequences selected from the TargetP training set. About 40% of the sequences are predicted to be localized to each of the

¹After these experiments were initiated a similar approach was published by Christensen *et al.*[19].

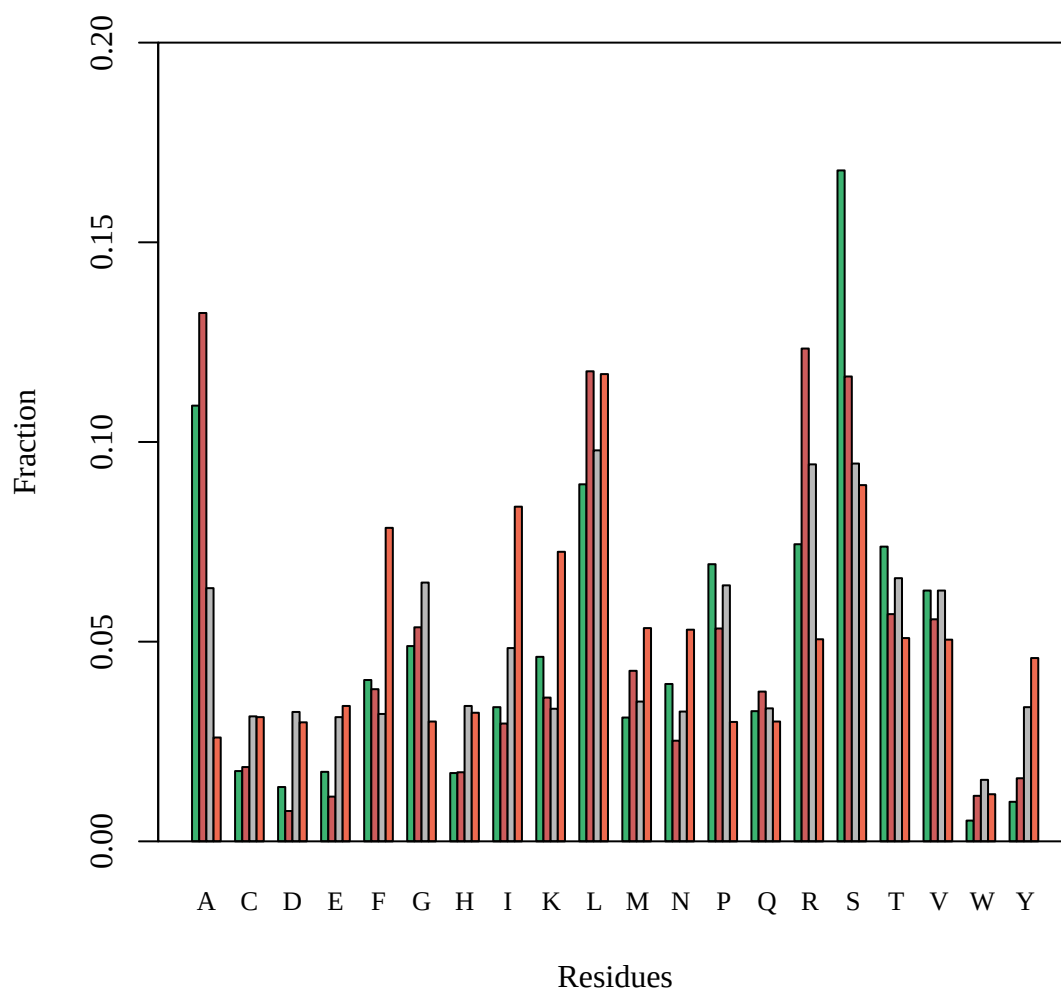


Figure 5.7: Residue frequencies for organelle and random sequence. Residue frequencies for the organelles are taken from the TargetP training set.

Table 5.3: Percentage predicted Localizations of Random Targeting Sequences Attached to Mitochondrial Proteins^a

Location	Reliability Class					Row Sum
	1	2	3	4	5	
Mitochondrion	3.0	6.8	8.2	13.0	9.6	40.6
Chloroplast	0.2	1.2	1.4	3.0	5.6	11.4
Secretory	0.2	1.0	0.2	1.6	2.4	5.4
Other	0.2	5.8	12.4	12.8	11.4	42.6
Total						100

^aLocation prediction by TargetP.

Table 5.4: Percentage predicted Localizations of Random Targeting Sequences Attached to Chloroplast Proteins^a

Location	Reliability Class					Row Sum
	1	2	3	4	5	
Mitochondrion	2.8	7.6	8.6	7.8	11.4	38.2
Chloroplast	0.4	0.2	2.6	3.2	6.4	12.8
Secretory	0.2	1.0	0.2	1.0	2.0	4.4
Other	0.6	5.4	11.8	13.0	13.8	44.6
Total						100

^aLocation prediction by TargetP.

cytosol and the mitochondrion and about 12% to the chloroplast. The same set of random sequences were then joined to chloroplast mature sequence and the localization predicted. These results are given in Table 5.4. The relative percentage of chloroplast predictions increased only slightly from the change in mature sequence (the absolute numbers were 57 and 64 total). A variety of sequence lengths were considered, and similar results were obtained whether the sequences were attached to actual proteins or not (data not shown). For instance, for 250 random sequences all of length 50, 51.2% were predicted to have mitochondrial and 9.2% chloroplastic localization. In all these experiments, mitochondrial localization exceeded chloroplastic by a factor of four or five.

Predictions for the intergenic sequence attached to mature chloroplast sequence is presented in Tables 5.5 and 5.6. The TargetP results show a decrease in targeting to both the chloroplast and the mitochondrion for the intergenic sequence as compared to the ran-

Table 5.5: Per Cent TargetP Localization of Arabidopsis Intergenic Sequence

Location	Reliability Class					Row Sum
	1	2	3	4	5	
Mitochondrion	0.4	2.4	3.2	4.0	6.6	16.6
Chloroplast	0.0	0.2	1.8	0.8	2.0	4.8
Secretory	4.8	4.8	4.2	4.6	8.8	27.2
Other	0.6	12.2	13.0	14.8	10.8	51.4
Total						100

Table 5.6: Per Cent Predotar Localization Prediction for Intergenic Targeting Sequence

Location	Call ^a		Row Sum
	1	2	
Mitochondrion	5.0	10.4	15.4
Plastid	1.4	2.6	4.0
ER	9.6	5.2	14.8
Other	65.8		65.8
Total			100

^aColumn “2” is the Predotar call ‘Possibly Plastid’, etc.

dom, though their relative localization is still about 4:1. The Predotar results were similar (Table 5.6), except this algorithm does not predict as high a localization to the secretory system.

5.5 Discussion

Targeting sequence acquisition by domain swapping is certainly not invalidated by these experiments. Well known examples exist for the correspondence of targeting sequences with exon boundaries[39, 82, 114]. Instead we fail to find a signal strong enough to indicate that it is the necessary or only mode of targeting sequence acquisition. Examination of the exon distribution data for the organelles shows little or no difference when compared with the *Arabidopsis* genome or with proteins predicted to have no organelle localization (Table 5.1). This indicates that organelle proteins have no extra exons. For all genes with 2 or more exons, the comparison of exon 1 boundaries and the targeting sequence

predicted cleavage site does not coincide (Figure 5.1). Instead, the differences in these two boundaries form a distribution that is shifted towards positive values, i.e. the first exon most frequently longer than the targeting sequence. There are significant numbers with negative differences, however, indicating that the targeting sequence cleavage site is within the first exon. The most frequent difference indicated that the first exon was about 55 residues longer than the chloroplast target peptide and about 46 residues longer than the mitochondrial pre-sequence. The previous two experiments are based solely on predicted data. While the exon-intron database[127] has been curated to improve reliability, most of the exon-intron boundaries are not experimentally verified. TargetP was used here because it predicts a cleavage site in addition to the localization prediction. Our proteomic data in Chapter 2 indicate that the TargetP prediction was accurate for chloroplast proteins, but a similar proteomic experiment showed that the accuracy for mitochondrial proteins was not as good[46]. The errors in the prediction of the cleavage site have not been independently challenged by proteomic data and are thus unknown. A high throughput method to sequence the N-terminus of proteins would be valuable to address this issue, as well as other cellular processes in which N-terminal processing is involved in protein functionalization. Therefore it may be that the coincidence in exon and targeting sequence boundaries was obscured by errors in the cleavage site prediction or the exon boundary predictions. However the analysis of the TargetP experimental data in Figures 5.2 and 5.3 show a similar picture. The exon boundaries still are more frequently longer than the cleavage site, but the difference is smaller in the experiment, perhaps because the distance to the nearest boundary was taken, not just the first. For about 10% of these sequences in the training set, the second intron boundary was the nearest. There was only one sequence in the training set for which the boundaries actually coincided, *Arabidopsis* GAP3.

The previous analysis assumes that the targeting sequence module arrives and is attached with an intervening intron, which might not be the case. In the second approach, expressed cyanobacterial proteins were aligned with their plant plastid homologues, and the predicted cleavage site (or ‘attachment point’ from the standpoint of acquisition) mapped to the progenitor protein(Figure 5.5). The distribution of these boundaries was similar to the last experiment - the likelihood of an attachment outside the original protein start is about twice as likely as placement within the original start. The modal value for this distribution was about the same as that from the comparison of targeting sequence and exon boundaries for the TargetP training set.

These experiments neither refute nor support acquisition of the targeting sequence as a module. The module might just as well recombine with the new gene within the sequence

as without and give results consistent with these presented here. Recombination within is of limited extent, either in this model or a random acquisition model (see below) because of the detrimental effect of interfering with gene function. Whether or not recombination or exon swapping is the predominant mode of targeting sequence acquisition, it defers the question of where the targeting sequence came from in the first place. The Gaussian or log-normal distribution for the exon-targeting sequence boundary differences and attachment points hints at a more stochastic process for signal peptide acquisition. Several remarkable experiments in recent years have indicated that the export of chloroplast DNA to the nuclear genome is frequent and extensive. Stegemann *et al.* measured a rate of 1 in 5 million events for transfer of a chloroplast marker gene to the nucleus of somatic cells[144]. This was very similar to an earlier rate determined for yeast mitochondria[154]. Huang *et al.*[51] measured an even higher rate of 1 in 16000 in a paternally transmitted chloroplast marker gene. Since chloroplasts are maternally inherited in angiosperms, this may indicate a mechanism of DNA incorporation after degradation of the chloroplast during pollen production. It also implies an enhanced rate of gene transfer to the nucleus in the germ line. Additionally, a recent analysis of the rice and cress genomes shows the presence of large (and small) segments of organellar DNA in these genomes[106]. Rice contains an almost complete copy of the linearized chloroplast genome, and *Arabidopsis* an almost complete copy of the mitochondrial genome. This analysis also inferred the ages of the insertions and noted that the older fragments were beginning to assume the base frequency of the host DNA background through mutational processes[106]. The frequent insertion of organellar DNA suggests a mechanism for targeting sequence origin via use of random sequence, either intergenic or via mutation towards the the background of the genome².

Random sequences of bases were generated, with fixed length or variable, and attached to actual protein sequences or not. The results were similar in all cases. Around 40% of the random sequences are predicted to be mitochondrially localized and around 10% chloroplast localized. The results for random *Arabidopsis* intergenic sequence gave a similar relative proportion of localization, with a lower overall fraction predicting organelle localization. This might be because the average length of the targeting sequences generated in this experiment was probably lower than for the purely random sequence. The results for TargetP and Predotar were similar for this experiment. In all cases, the proportion of sequences with mitochondrial localization exceeded the chloroplastic by a factor of four or more. This experiment is interesting from several viewpoints. Firstly, the localization preference of random sequence recapitulates the order of organelle endosymbiosis. Current specula-

²After these experiments were done, Christensen *et al.* published a similar result, based on the observation that targeting sequence composition is similar to that one would get from random codon translation[19].

tion is that organelle protein import channels are of prokaryotic origin, perhaps protein or peptide exporters that reversed their membrane polarity(for example, [145]). Some of these exported proteins would have been signal peptides or defensive proteins which have many of the characteristics of targeting sequences - particularly membrane activity and highly variable composition. The primitive importer may have had an ability to recognize these variable sequences. Secondly, the chloroplast targeting signals appear to be more difficult to attain by chance. The advent of the chloroplast apparently led to several changes in the import apparatus. Components of the plant TOM complex were adapted, notably Tom 20 and Tom22 which differ from their animal and fungal functional homologues[84, 111]. In addition, the plant mitochondrial pre-sequences tend to be longer, and the chloroplast targeting peptide longer yet, implying the need to contain more information to ensure correct protein targeting. Finally, these *in silico* random sequence results have some experimental support. In 1987, Baker and Schatz demonstrated that 2.5% of random clones from *E.coli* were capable of targeting proteins to yeast mitochondria, a fraction that is the same order of magnitude of these predictions. A similar experiment has not been done for plants to measure the relative importability of random sequence into the chloroplast, probably due to the difficulty in transforming plants on a large scale. Perhaps the recent development of such a system will facilitate this experiment in the future[72].

Summary

A major goal of this research was to determine the chloroplast proteome and use it test hypotheses for transit peptide motifs. This was achieved. Plant culture and chloroplast isolation procedures were developed and proteomic analysis identified several hundred proteins. The preparation appears to be very pure as indicated by the lack of abundant proteins from other cell compartments and by the high predicted plastid localization by TargetP and Predotar. This is in contrast to another published chloroplast proteomics study[70]. Since this study made somewhat controversial claims about the accuracy of TargetP and Predotar, future studies of the chloroplast proteome should employ marker enzymes or antibodies to verify the purity of the preparation. While we believe our preparation is free of contamination, as noted above, verifying the purity would add increased confidence. The preparation also appeared to cover a wide range of molecular weight and pI, but there was indication that hydrophobic proteins were not detected as well. Since many of the photosystem proteins are very hydrophobic, special efforts might be made to detect these. The bacterial and eukaryotic homologues of the proteins detected reflect the mosaic nature of the chloroplast proteome with probable contribution of proteins originating from the mitochondrion and cytosol. Some of the proteins detected also seem to reflect the metabolic state of the plants, which were grown on agar containing sucrose. This might be followed up with comparison to another growth state using the semi-quantitative analysis developed in Chapter 4 or by labeling studies.

Another goal was to evaluate the accuracy of transit peptide localization programs TargetP and Predotar. This was achieved as well. Unlike a prior publication[70], we found the two algorithms to perform largely to the published accuracy. Since most of the information available for plastid proteins is from the photosynthetic chloroplast state, we had also hoped to perform a similar analysis on the root plastids of *Arabidopsis*, but did not. Besides providing useful information about plastids physiology in these states, the information would be an additional challenge to prediction algorithms since few such proteins are in their training sets. A working hydroponic plant culture system was developed for

Arabidopsis roots, but a successful root plastid procedure has not yet been developed.

Besides the identity or localization of plastid proteins, an important characteristic is the location of the cleavage site. Knowledge of this would provide both the actual sequence of the mature protein and lead discovery of the peptidase cleavage site motifs. We had hoped to develop a method for this but did not. This would require either a very sensitive mass spectrometer to determine the mass of the intact protein with enough accuracy to infer the cleavage site by difference from the predicted sequence of pre-proteins or it would require a method for the isolation of the N-terminal peptide. The latter approach might be preferred, since it would be capable of application on more modest mass spectrometers. Such a method would be a worthy goal not only for organelle proteomics but in general since N-terminal modifications are common.

The proteomic data were used to search for the potential organelle import apparatus motif 'FPLK'. A program which could apply various heuristics was written to search for motifs with relaxed positional restraints. This approach has been applied at least once before[63] and abandoned because of high false positives. These authors had a hand curated test set to make this observation, whereas we did not. However, we did validate the technique with extensive testing of synthetic data sets. While we did not do so, the synthetic data sets could be used to tune the heuristics to a given false positive rate using a statistical approach like a receiver operating curve. We also employed two sets of proteins with experimental localization in the motif search. No evidence for the ability of the motif 'FPLK' to distinguish the chloroplast and mitochondrion was found, instead the motif seemed to exist as frequently in the mitochondrion as the chloroplast. This may indicate that it is a hydrophobic motif involved in translocon recognition for both organelles.

Cyanobacterial proteomics were investigated to provide a ready source of protein to implement proteomics at the University of Tennessee and to identify proteins to make evolutionary inferences about plastid protein conservation and transit peptide origin. For the first goal, cyanobacteria may have been an educational, but not a productive choice. The abundance of proteins like the phycobiliproteins is an issue that is common to many proteomic analyses, blood for instance, and was finally conquered by depletion using biochemical methods. The presence of the covalently bound bilin molecule caused other problems. We suspected oxidation at the electrode of the mass spectrometer which was solved by moving the electrode to the split line. In addition, the bilin tetrapyrrole has cation exchange and chelating characteristics which may have depressed the number of identification in our 2D analyses. An attempt at phosphopeptide isolation by immobilized metal affinity chromatography (not reported here) was probably inhibited by the His-tag like tetrapyr-

role. Several hundred proteins were identified, many of them unknown. The proteomic data were used as a form of expression analysis and correlated to the codon adaptation index. The data were linearized by plotting the abundance versus rank. This was a useful form of presentation since it highlights where proteins fall in the abundance-CAI space. Hypothetical proteins tended to locate in the low CAI, low expression area of the plot, but some did not. These unknown high expression, high CAI proteins might be functionally important and would be candidates for further investigation.

It was also found that the more highly expressed proteins were more likely to be conserved in plants than the *Synechocystis* proteome as a whole. Many of these were involved in photosynthesis, but proteins with other functions were involved as well. Further investigation of these might give insight into other conserved cyanobacterial functions in plants. Since this analysis was conducted with *Synechocystis* grown in the replete photosynthetic state, it would be interesting to conduct a similar analysis with the bacterium cultivated in another growth state and ascertain the expression and conservation in plants. Another part of any additional analysis would examination of those proteins in the *Synechocystis* proteome *not* conserved in the plant.

The proteins from the *Synechocystis* proteome determination were used as 'probes' of plant genomes to infer how transit peptides were obtained. The data indicate a random attachment of targeting sequences with a distribution skewed to the N-terminal, i.e. away from the functional core of the cyanobacterial proteins. This is suggestive and is consistent with the facile and continuing transfer of DNA to the eukaryotic nucleus summarized in the Introduction. The conclusions are weakened however, since the genome scale experiment depends on the accuracy of the cleavage site prediction of TargetP, which is unknown. The availability of a cleavage site determination method mentioned previously would permit such an evaluation to be made.

The randomly generated sequence experiment produced several interesting results. The amino acid distribution obtained was similar to those of chloroplast and mitochondrial targeting sequences. A fairly high fraction was predicted to be imported by TargetP and Predotar. Our proteomic data indicate these localization predictors to be accurate, at least for chloroplast proteins, which provides some confidence for random sequence import. It was also found that randomly selected *Arabidopsis* intergenic sequence could serve as a targeting sequence. In this experiment, it only required four tries on average to obtain a targeting sequence. These results combined with the high rates of DNA export to the nucleus[51, 144, 154] indicate that integration of a gene and acquisition of a targeting sequence could occur readily. In all the random sequence experiments mitochondrial local-

ization exceeded chloroplast localization, implying that the mitochondrion is the 'default' organelle for random sequence. This suggests that the mitochondrial translocon evolved to import this type of sequence. Chloroplast sequences can also be obtained by chance, but more information seems to be required for import. Further modeling work should include a comparison of sequences with predicted import to ascertain the differences in chloroplast and mitochondrial features. The most convincing experiment would be to test in an *in vivo* or *in vitro* import assay whether one of the sequences is capable of targeting a protein to the organelle.

References

References

- [1] F. Abdallah, F. Salamini, and D. Leister. A prediction of the size and evolutionary origin of the proteome of chloroplasts of *Arabidopsis*. *Trends Plant Sci*, 5(4):141–2, Apr 2000.
- [2] Keith L. Adams, Yin-Long Qiu, Mark Stoutemyer, and Jeffrey D. Palmer. Punctuated evolution of mitochondrial gene content: high and variable rates of mitochondrial gene loss and transfer to the nucleus during angiosperm evolution. *Proc Natl Acad Sci U S A*, 99(15):9905–12, Jul 2002.
- [3] M. Aivaliotis, E. Neofotistou, H.W. Remigy, G. Tsimpinos, A. Lustig, F. Lottspeich, and G. Tsiotis. Isolation and characterization of an outer membrane protein of *Clorobium tepidum*. *Photosynth Res*, 79(2):161–6, 2004.
- [4] B. Alberts, J. Lewis, M. Raff, K. Roberts, and James D. Watson. *Molecular Biology of the Cell*. Garland Publishing, 3rd edition, 1994.
- [5] S.F. Altschul, T. L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–402, Sep 1997.
- [6] John M. Archibald, Matthew B. Rogers, Michael Toop, Ken-Ichiro Ishida, and Patrick J. Keeling. Lateral gene transfer and the evolution of plastid-targeted proteins in the secondary plastid-containing alga *Bigelowiella natans*. *Proc Natl Acad Sci U S A*, 100(13):7678–83, Jun 2003.
- [7] T.L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, 2:28–36, 1994.
- [8] Maria Veronica Beligni, Kenichi Yamaguchi, and Stephen P. Mayfield. Chloroplast elongation factor TS pro-protein is an evolutionarily conserved fusion with the S1 domain-containing plastid-specific ribosomal protein-7. *Plant Cell*, 16(12):3357–69, Dec 2004.
- [9] Manoj Bhasin and G.P.S. Raghava. ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res*, 32(Web Server issue):W414–9, Jul 2004.

- [10] Shashi Bhushan, Annelie Stahl, Stefan Nilsson, Benoit Lefebvre, Motoaki Seki, Christian Roth, David McWilliam, Sarah J. Wright, David A. Liberles, Kazuo Shinozaki, Barry D. Bruce, Marc Boutry, and Elzbieta Glaser. Catalysis, subcellular localization, expression and evolution of the targeting peptides degrading protease, AtPreP2. *Plant Cell Physiol*, 46(6):985–96, Jun 2005.
- [11] Andreas M. Boehm, Florian Grosse-Coosmann, and Albert Sickmann. Command line tool for calculating theoretical MS spectra for given sequences. *Bioinformatics*, 20(16):2889–91, Nov 2004.
- [12] Gustavo Bonaventure and John B. Ohlrogge. Differential regulation of mRNA levels of acyl carrier protein isoforms in *Arabidopsis*. *Plant Physiol*, 128(1):223–35, Jan 2002.
- [13] A. Brennicke, L. Grohmann, R. Hiesel, V. Knoop, and W. Schuster. The mitochondrial genome on its way to the nucleus: different stages of gene transfer in higher plants. *FEBS Lett*, 325(1-2):140–5, Jun 1993.
- [14] B. D. Bruce. The role of lipids in plastid protein transport. *Plant Mol Biol*, 38(1-2):223–46, Sep 1998.
- [15] B. D. Bruce. The paradox of plastid transit peptides: conservation of function despite divergence in primary structure. *Biochim Biophys Acta*, 1541(1-2):2–21, Dec 2001.
- [16] Barry D. Bruce and Kenneth Keegstra. *Translocation of proteins across chloroplast membranes*, volume 10, pages 389–430. JAI Press Inc., 1994.
- [17] Thomas Cavalier-Smith. Cell evolution and earth history: stasis and revolution. *Philos Trans R Soc Lond B Biol Sci*, 361(1470):969–1006, Jun 2006.
- [18] Orinda Chew, Ryan Lister, Soumya Qbadou, Joshua L. Heazlewood, Jurgen Soll, Enrico Schleiff, A. Harvey Millar, and James Whelan. A plant outer mitochondrial membrane protein with high amino acid sequence identity to a chloroplast protein import receptor. *FEBS Lett*, 557(1-3):109–14, Jan 2004.
- [19] Alan C. Christensen, Anna Lyznik, Saleem Mohammed, Christian G. Elowsky, Annakaisa Elo, Ryan Yule, and Sally A. Mackenzie. Dual-domain, dual-targeting organellar protein presequences in *Arabidopsis* can use non-AUG start codons. *Plant Cell*, 17(10):2805–16, Oct 2005.
- [20] V. Chupin, R. van ’t Hof, and B. de Kruijff. The transit sequence of a chloroplast precursor protein reorients the lipids in monogalactosyl diglyceride containing bilayers. *FEBS Lett*, 350(1):104–8, Aug 1994.
- [21] Diane Constan, John E. Froehlich, Sowkya Rangarajan, and Kenneth Keegstra. A stromal Hsp100 protein is required for normal chloroplast development and function in *Arabidopsis*. *Plant Physiol*, 136(3):3605–15, Nov 2004.

- [22] Rebecca W. Corbin, Oleg Paliy, Feng Yang, Jeffrey Shabanowitz, Mark Platt, Charles E. Lyons Jr, Karen Root, Jon McAuliffe, Michael I. Jordan, Sydney Kustu, Eric Soupene, and Donald F. Hunt. Toward a protein profile of *Escherichia coli*: comparison to its transcription profile. *Proc Natl Acad Sci U S A*, 100(16):9232–7, Aug 2003.
- [23] Robertson Craig and Ronald C. Beavis. A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun Mass Spectrom*, 17(20):2310–6, 2003.
- [24] Robertson Craig and Ronald C. Beavis. Tandem: matching proteins with tandem mass spectra. *Bioinformatics*, 20(9):1466–7, Jun 2004.
- [25] Gavin E. Crooks, Gary Hon, John-Marc Chandonia, and Steven E. Brenner. Web-Logo: a sequence logo generator. *Genome Res*, 14(6):1188–90, Jun 2004.
- [26] Claire Delahunty and John R. Yates 3rd. Protein identification using 2D-LC-MS/MS. *Methods*, 35(3):248–55, Mar 2005.
- [27] W. F. Doolittle. You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends Genet*, 14(8):307–11, Aug 1998.
- [28] Anne-Marie Duchene, Anatoli Giritch, Beate Hoffmann, Valerie Cognat, Dominique Lancelin, Nemo M. Peeters, Marlyse Zaepfel, Laurence Marechal-Drouard, and Ian D. Small. Dual targeting is the rule for organellar aminoacyl-tRNA synthetases in *Arabidopsis thaliana*. *Proc Natl Acad Sci U S A*, 102(45):16484–9, Nov 2005.
- [29] O. Emanuelsson, H. Nielsen, S. Brunak, and G. von Heijne. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol*, 300(4):1005–16, Jul 2000.
- [30] Olof Emanuelsson. Predicting protein subcellular localisation from amino acid sequence information. *Brief Bioinform*, 3(4):361–76, Dec 2002.
- [31] Jimmy K. Eng, Ashley L. McCormack, and III John R. Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 5(11):976–989, November 1994.
- [32] A. Eyre-Walker. Synonymous codon bias is related to gene length in *Escherichia coli*: selection for translational accuracy? *Mol Biol Evol*, 13(6):864–72, Jul 1996.
- [33] J. B. Fenn, M Mann, C. K. Meng, S. F. Wong, and C. M. Whitehouse. Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246(4926):64–71, Oct 1989.
- [34] L. M. Fitzpatrick and K. Keegstra. A method for isolating a high yield of *Arabidopsis* chloroplasts capable of efficient import of precursor proteins. *Plant J*, 27(1):59–65, Jul 2001.

- [35] Giulia Friso, Lisa Giacomelli, A. Jimmy Ytterberg, Jean-Benoit Peltier, Andrea Rudella, Qi Sun, and Klaas J. van Wijk. In-depth analysis of the thylakoid membrane proteome of *Arabidopsis thaliana* chloroplasts: new proteins, new functions, and a plastid proteome database. *Plant Cell*, 16(2):478–99, Feb 2004.
- [36] S. Fulda, F. Huang, F. Nilsson, M. Hagemann, and B. Norling. Proteomics of *Synechocystis* sp. strain PCC 6803. identification of periplasmic proteins in cells grown at low and high salt concentrations. *Eur J Biochem*, 267(19):5900–7, Oct 2000.
- [37] K. Gabriel, S. K. Buchanan, and T. Lithgow. The alpha and the beta: protein translocation across mitochondrial and plastid outer membranes. *Trends Biochem Sci*, 26(1):36–40, Jan 2001.
- [38] Chee Sian Gan, Kenneth F. Reardon, and Phillip C. Wright. Comparison of protein and peptide prefractionation methods for the shotgun proteomic analysis of *Synechocystis* sp. PCC 6803. *Proteomics*, 5(9):2468–78, Jun 2005.
- [39] J. S. Gantt, S. L. Baldauf, P. J. Calie, N. F. Weeden, and J. D. Palmer. Transfer of rpl22 to the nucleus greatly preceded its loss from the chloroplast and involved the gain of an intron. *EMBO J*, 10(10):3073–8, Oct 1991.
- [40] Lisa Giacomelli, Andrea Rudella, and Klaas Jan van Wijk. High light response of the thylakoid proteome in *Arabidopsis* wild type and the ascorbate-deficient mutant vtc2-2. a comparative proteomics study. *Plant Physiol*, 141(2):685–701, Jun 2006.
- [41] W. Gish. WU-BLAST 2.0. <http://blast.wustl.edu>, 1996-2003.
- [42] Alexander N. Glazer. *Cyanobacteria*, volume 167 of *Methods in Enzymology*, chapter Phycobilisomes. Academic Press, 1988.
- [43] Regina M. Goetz and Anders Fuglsang. Correlation of codon bias measures with mRNA levels: analysis of transcriptome data from *Escherichia coli*. *Biochem Biophys Res Commun*, 327(1):4–7, Feb 2005.
- [44] Stephen M. Gomez, Karl Y. Bil’, Rodrigo Aguilera, John N. Nishio, Kym F. Faull, and Julian P. Whitelegge. Transit peptide cleavage sites of integral thylakoid membrane proteins. *Mol Cell Proteomics*, 2(10):1068–85, Oct 2003.
- [45] Timothy J. Griffin, Steven P. Gygi, Trey Ideker, Beate Rist, Jimmy Eng, Leroy Hood, and Ruedi Aebersold. Complementary profiling of gene expression at the transcriptome and proteome levels in *Saccharomyces cerevisiae*. *Mol Cell Proteomics*, 1(4):323–33, Apr 2002.
- [46] Joshua L. Heazlewood, Julian S. Tonti-Filippini, Alexander M. Gout, David A. Day, James Whelan, and A. Harvey Millar. Experimental analysis of the *Arabidopsis* mitochondrial proteome highlights signaling and regulatory components, provides assessment of targeting prediction programs, and indicates plant-specific mitochondrial proteins. *Plant Cell*, 16(1):241–56, Jan 2004.

- [47] Mirkka Herranen, Natalia Battchikova, Pengpeng Zhang, Alexander Graf, Sari Sirpio, Virpi Paakkarinen, and Eva-Mari Aro. Towards functional proteomics of membrane protein complexes in *Synechocystis* sp. PCC 6803. *Plant Physiol*, 134(1):470–81, Jan 2004.
- [48] Johannes M. Herrmann. Converting bacteria to organelles: evolution of mitochondrial protein sorting. *Trends Microbiol*, 11(2):74–9, Feb 2003.
- [49] Annette Hoglund, Pierre Donnes, Torsten Blum, Hans-Werner Adolph, and Oliver Kohlbacher. MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. *Bioinformatics*, 22(10):1158–65, May 2006.
- [50] S. Hua and Z. Sun. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, 17(8):721–8, Aug 2001.
- [51] Chun Y. Huang, Michael A. Ayliffe, and Jeremy N. Timmis. Direct measurement of the transfer rate of chloroplast DNA into the nucleus. *Nature*, 422(6927):72–6, Mar 2003.
- [52] Fang Huang, Erik Hedman, Christiane Funk, Thomas Kieselbach, Wolfgang P. Schroder, and Birgitta Norling. Isolation of outer membrane of *Synechocystis* sp. PCC 6803 and its proteomic characterization. *Mol Cell Proteomics*, 3(6):586–95, Jun 2004.
- [53] Fang Huang, Ingela Parmryd, Fredrik Nilsson, Annika L. Persson, Himadri B. Pakrasi, Bertil Andersson, and Birgitta Norling. Proteomics of *Synechocystis* sp. strain PCC 6803: identification of plasma membrane proteins. *Mol Cell Proteomics*, 1(12):956–66, Dec 2002.
- [54] Stephen W. Hunsucker, Karsten Klage, Stephen M. Slaughter, Malcolm Potts, and Richard F. Helm. A preliminary investigation of the *Nostoc punctiforme* proteome. *Biochem Biophys Res Commun*, 317(4):1121–7, May 2004.
- [55] D. F. Hunt, J. R. Yates 3rd, J. Shabanowitz, S. Winston, and C. R. Hauer. Protein sequencing by tandem mass spectrometry. *Proc Natl Acad Sci U S A*, 83(17):6233–7, Sep 1986.
- [56] D. F. Hunt, A. M. Buko, J. M. Ballard, J. Shabanowitz, and A. B. Giordani. Sequence analysis of polypeptides by collision activated dissociation on a triple quadrupole mass spectrometer. *Biomed Mass Spectrom*, 8(9):397–408, Sep 1981.
- [57] R.A. Ivey 3rd, C. Subramanian, and B. D. Bruce. Identification of a Hsp70 recognition domain within the rubisco small subunit transit peptide. *Plant Physiol*, 122(4):1289–99, Apr 2000.
- [58] D. Jackson-Constan and K. Keegstra. *Arabidopsis* genes encoding components of the chloroplastic protein import apparatus. *Plant Physiol*, 125(4):1567–76, Apr 2001.

- [59] Eugene A. Kapp, Frederic Schutz, Lisa M. Connolly, John A. Chakel, Jose E. Meza, Christine A. Miller, David Fenyo, Jimmy K. Eng, Joshua N. Adkins, Gilbert S. Omenn, and Richard J. Simpson. An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis. *Proteomics*, 5(13):3475–90, Aug 2005.
- [60] M. Karas and F. Hillenkamp. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal Chem*, 60(20):2299–301, Oct 1988.
- [61] G.A. Karlin-Neumann and E. M. Tobin. Transit peptides of nuclear-encoded chloroplast proteins share a common amino acid framework. *EMBO J*, 5(1):9–13, Jan 1986.
- [62] Yasuhiro Kashino, Wendy M. Lauber, James A. Carroll, Qingjun Wang, John Whitmarsh, Kazuhiko Satoh, and Himadri B. Pakrasi. Proteomic analysis of a highly active photosystem II preparation from the cyanobacterium *Synechocystis* sp. PCC 6803 reveals the presence of novel polypeptides. *Biochemistry*, 41(25):8004–12, Jun 2002.
- [63] Jonathan E. Katz, Mensur Dlakic, and Steven Clarke. Automated identification of putative methyltransferases from genomic open reading frames. *Mol Cell Proteomics*, 2(8):525–40, Aug 2003.
- [64] Alan J. Kaufman and Shuhai Xiao. High CO₂ levels in the proterozoic atmosphere estimated from analyses of individual microfossils. *Nature*, 425(6955):279–82, Sep 2003.
- [65] Patrick J. Keeling. Diversity and evolutionary history of plastids and their hosts. *American Journal of Botany*, 91(10):1481–1493, 2004.
- [66] N. L. Kelleher, S. V. Taylor, D. Grannis, C. Kinsland, H. J. Chiu, T. P. Begley, and F. W. McLafferty. Efficient sequence analysis of the six gene products (7-74 kda) from the *Escherichia coli* thiamin biosynthetic operon by tandem high-resolution mass spectrometry. *Protein Sci*, 7(8):1796–801, Aug 1998.
- [67] Elizabeth A. Kellogg and Jeffrey L. Bennetzen. The evolution of nuclear genome structure in seed plants. *American Journal of Botany*, 91(10):1709–1725, 2004.
- [68] Nir Keren, Rajeev Aurora, and Himadri B. Pakrasi. Critical roles of bacterioferritins in iron storage and proliferation of cyanobacteria. *Plant Physiol*, 135(3):1666–73, Jul 2004.
- [69] T. Kieselbach and W.P. Schroder. The proteome of the chloroplast lumen of higher plants. *Photosynth Res*, 78(3):249–64, 2003.
- [70] Torsten Kleffmann, Doris Russenberger, Anne von Zychlinski, Wayne Christopher, Kimmen Sjolander, Wilhelm Gruissem, and Sacha Baginsky. The *Arabidopsis*

thaliana chloroplast proteome reveals pathway abundance and novel protein functions. *Curr Biol*, 14(5):354–62, Mar 2004.

- [71] J. H. Kleinschmidt, J. E. Mahaney, D. D. Thomas, and D. Marsh. Interaction of bee venom melittin with zwitterionic and negatively charged phospholipid bilayers: a spin-label electron spin resonance study. *Biophys J*, 72(2 Pt 1):767–78, Feb 1997.
- [72] Olga A. Koroleva, Matthew L. Tomlinson, David Leader, Peter Shaw, and John H. Doonan. High-throughput protein localization in *Arabidopsis* using *Agrobacterium*-mediated transient expression of gfp-orf fusions. *Plant J*, 41(1):162–74, Jan 2005.
- [73] C. G. Kurland and S. G. Andersson. Origin and evolution of the mitochondrial proteome. *Microbiol Mol Biol Rev*, 64(4):786–820, 2000. Journal Article Review United states Mmbr.
- [74] U. Kutschera and K. J. Niklas. Endosymbiosis, cell evolution, and speciation. *Theory Biosci*, 124(1):1–24, Aug 2005.
- [75] Dong Wook Lee, Sookjin Lee, Gil-Je Lee, Kwang Hee Lee, Sanguk Kim, Gang-Won Cheong, and Inhwan Hwang. Functional characterization of sequence motifs in the transit peptide of *Arabidopsis* small subunit of rubisco. *Plant Physiol*, 140(2):466–83, Feb 2006.
- [76] Kwang Hee Lee, Dae Heon Kim, Sang Woo Lee, Zhoo Hyeon Kim, and Inhwan Hwang. *In vivo* import experiments in protoplasts reveal the importance of the overall context but not specific amino acid residues of the transit peptide during import into chloroplasts. *Mol Cells*, 14(3):388–97, Dec 2002.
- [77] Dario Leister. Origin, evolution and genetic effects of nuclear insertions of organelle dna. *Trends Genet*, 21(12):655–63, Dec 2005.
- [78] L. Li, C. D. Masselon, G. A. Anderson, L. Pasa-Tolic, S. W. Lee, Y. Shen, R. Zhao, M. S. Lipton, T. P. Conrads, N. Tolic, and R. D. Smith. High-throughput peptide identification from protein digests using data-dependent multiplexed tandem fticr mass spectrometry coupled with capillary liquid chromatography. *Anal Chem*, 73(14):3312–22, Jul 2001.
- [79] Ryan Lister, Orinda Chew, May-Nee Lee, Joshua L. Heazlewood, Rachel Clifton, Karen L. Parker, A. Harvey Millar, and James Whelan. A transcriptomic and proteomic characterization of the arabidopsis mitochondrial protein import apparatus and its response to mitochondrial dysfunction. *Plant Physiol*, 134(2):777–89, Feb 2004.
- [80] D. P. Little, J. P. Speir, M. W. Senko, P.B. O’Connor, and F. W. McLafferty. Infrared multiphoton dissociation of large multiply charged ions for biomolecule sequencing. *Anal Chem*, 66(18):2809–15, Sep 1994.

- [81] Hongbin Liu, Rovshan G. Sadygov, and John R. Yates 3rd. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem*, 76(14):4193–201, Jul 2004.
- [82] M. Long, S.J. de Souza, C. Rosenberg, and W. Gilbert. Exon shuffling and the origin of the mitochondrial targeting function in plant cytochrome c1 precursor. *Proc Natl Acad Sci U S A*, 93(15):7727–31, Jul 1996.
- [83] Bin Ma, Kaizhong Zhang, Christopher Hendrie, Chengzhi Liang, Ming Li, Amanda Doherty-Kirby, and Gilles Lajoie. PEAKS: powerful software for peptide *de novo* sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom*, 17(20):2337–42, 2003.
- [84] D. Macasev, E. Newbiggin, J. Whelan, and T. Lithgow. How do plant mitochondria avoid importing chloroplast proteins? Components of the import apparatus Tom20 and Tom22 from *Arabidopsis* differ from their fungal counterparts. *Plant Physiol*, 123(3):811–6, Jul 2000.
- [85] Sally A. Mackenzie. Plant organellar protein targeting: a traffic plan still under construction. *Trends Cell Biol*, 15(10):548–54, Oct 2005.
- [86] Wojciech Majeran, Yang Cai, Qi Sun, and Klaas J. van Wijk. Functional differentiation of bundle sheath and mesophyll maize chloroplasts determined by comparative proteomics. *Plant Cell*, 17(11):3111–40, Nov 2005.
- [87] M. Mann and M. Wilm. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal Chem*, 66(24):4390–9, Dec 1994.
- [88] S. E. Martin, J. Shabanowitz, D. F. Hunt, and J. A. Marto. Subfemtomole ms and ms/ms peptide sequence analysis using nano-hplc micro-esi fourier transform ion cyclotron resonance mass spectrometry. *Anal Chem*, 72(18):4266–74, Sep 2000.
- [89] W. Martin and R.G. Herrmann. Gene transfer from organelles to the nucleus: how much, what happens, and why? *Plant Physiol*, 118(1):9–17, Sep 1998.
- [90] William Martin. Gene transfer from organelles to the nucleus: frequent and in big chunks. *Proc Natl Acad Sci U S A*, 100(15):8612–4, Jul 2003.
- [91] William Martin, Tamas Rujan, Erik Richly, Andrea Hansen, Sabine Cornelsen, Thomas Lins, Dario Leister, Bettina Stoebe, Masami Hasegawa, and David Penny. Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc Natl Acad Sci U S A*, 99(19):12246–51, Sep 2002.
- [92] Mitsuhiro Matsuo, Yuki Ito, Ryo Yamauchi, and Junichi Obokata. The rice nuclear genome continuously integrates, shuffles, and eliminates the chloroplast genome to cause chloroplast-nuclear DNA flux. *Plant Cell*, 17(3):665–75, Mar 2005.

- [93] T. May and J. Soll. 14-3-3 proteins form a guidance complex with chloroplast precursor proteins in plants. *Plant Cell*, 12(1):53–64, Jan 2000.
- [94] A. L. McCormack, D. M. Schieltz, B. Goode, S. Yang, G. Barnes, D. Drubin, and J. R. Yates 3rd. Direct analysis and identification of proteins in mixtures by lc/ms/ms and database searching at the low-femtomole level. *Anal Chem*, 69(4):767–76, Feb 1997.
- [95] Geoffrey I. McFadden and Giel G. van Dooren. Evolution: red algal genome affirms a common origin of all plastids. *Curr Biol*, 14(13):R514–6, Jul 2004.
- [96] C. Mereschkowsky. Über nature und Ursprung der Chromatophoren im Pflanzenreiche. *Biol. Centralbl*, 25:593–604,689–691, 1905.
- [97] Pascale Mollier, Beate Hoffmann, Cedrig Debast, and Ian Small. The gene encoding *Arabidopsis thaliana* mitochondrial ribosomal protein s13 is a recent duplication of the gene encoding plastid s13. *Curr Genet*, 40(6):405–9, Mar 2002.
- [98] David W. Mount. *Bioinformatics*, chapter 9, page 443. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 2001.
- [99] M .W. Murcha, C. Rudhe, D. Elhafez, K.L. Adams, D.O. Daley, and J. Whelan. Adaptations required for mitochondrial import following mitochondrial to nucleus gene transfer of ribosomal protein s10. *Plant Physiol*, 138(4):2134–44, 2005. Journal Article United States.
- [100] T. Muto, T. Obita, Y. Abe, T. Shodai, T. Endo, and D. Kohda. NMR identification of the Tom20 binding segment in mitochondrial presequences. *J Mol Biol*, 306(2):137–43, Feb 2001.
- [101] R. Nadimpalli, N. Yalpani, G. S. Johal, and C. R. Simmons. Prohibitins, stomatins, and plant disease response genes compose a protein superfamily that controls cell proliferation, ion channel regulation, and death. *J Biol Chem*, 275(38):29579–86, Sep 2000.
- [102] Rajesh Nair and Burkhard Rost. Better prediction of sub-cellular localization by combining evolutionary and structural information. *Proteins*, 53(4):917–30, Dec 2003.
- [103] K. Nakai and P. Horton. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci*, 24(1):34–6, Jan 1999.
- [104] H.E. Neuhaus and M.J. Emes. Nonphotosynthetic metabolism in plastids. *Annu Rev Plant Physiol Plant Mol Biol*, 51:111–140, Jun 2000.
- [105] Karl J. Niklas. *The Evolutionary Biology of Plants*. The University of Chicago Press, 1997.

- [106] Christos Noutsos, Erik Richly, and Dario Leister. Generation and evolutionary fate of insertions of organelle DNA in the nuclear genomes of flowering plants. *Genome Res*, 15(5):616–28, May 2005.
- [107] N. Peeters and I. Small. Dual targeting to mitochondria and chloroplasts. *Biochim Biophys Acta*, 1541(1-2):54–63, Dec 2001.
- [108] Jean-Benoit Peltier, A. Jimmy Ytterberg, Qi Sun, and Klaas J. van Wijk. New functions of the thylakoid membrane proteome of *Arabidopsis thaliana* revealed by a simple, fast, and versatile fractionation strategy. *J Biol Chem*, 279(47):49367–83, Nov 2004.
- [109] Junmin Peng, Joshua E. Elias, Carson C. Thoreen, Larry J. Licklider, and Steven P. Gygi. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *J Proteome Res*, 2(1):43–50, Jan-Feb 2003.
- [110] D. N. Perkins, D. J. Pappin, D. M. Creasy, and J. S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–67, Dec 1999.
- [111] Andrew J. Perry, Joanne M. Hulett, Vladimir A. Likicj, Trevor Lithgow, and Paul R. Gooley. Convergent evolution of receptors for protein import into mitochondria. *Curr Biol*, 16(3):221–9, Feb 2006.
- [112] J. M. Petit, J. F. Briat, and S. Lobreaux. Structure and differential expression of the four members of the *Arabidopsis thaliana* ferritin gene family. *Biochem J*, 359(Pt 3):575–82, Nov 2001.
- [113] M. Pilon, H. Wienk, W. Sips, M. de Swaaf, I. Talboom, R. van ’t Hof, G. de Korte-Kool, R. Demel, P. Weisbeek, and B. de Kruijff. Functional domains of the ferredoxin transit sequence involved in chloroplast import. *J Biol Chem*, 270(8):3882–93, Feb 1995.
- [114] F. Quigley, W. F. Martin, and R. Cerff. Intron conservation across the prokaryote-eukaryote boundary: structure of the nuclear gene for chloroplast glyceraldehyde-3-phosphate dehydrogenase from maize. *Proc Natl Acad Sci U S A*, 85(8):2672–6, Apr 1988.
- [115] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. ISBN 3-900051-07-0.
- [116] Amar N. Rai, Birgitta Bergman, and Ulla Rasmussen. *Cyanobacteria in Symbiosis*. Kluwer Academic Publishers, 2002.

- [117] S. Reinbothe, S. Runge, C. Reinbothe, B. van Cleve, and K. Apel. Substrate-dependent transport of the NADPH:protochlorophyllide oxidoreductase into isolated plastids. *Plant Cell*, 7(2):161–72, Feb 1995.
- [118] Steffen Reinbothe, Francoise Quigley, John Gray, Andreas Schemenewitz, and Christiane Reinbothe. Identification of plastid envelope proteins required for import of protochlorophyllide oxidoreductase A into the chloroplast of barley. *Proc Natl Acad Sci U S A*, 101(7):2197–202, Feb 2004.
- [119] Steffen Reinbothe, Francoise Quigley, Armin Springer, Andreas Schemenewitz, and Christiane Reinbothe. The outer plastid envelope protein Oep16: role as precursor translocase in import of protochlorophyllide oxidoreductase A. *Proc Natl Acad Sci U S A*, 101(7):2203–8, Feb 2004.
- [120] Katheryn A. Resing, Karen Meyer-Arendt, Alex M. Mendoza, Lauren D. Aveline-Wolf, Karen R. Jonscher, Kevin G. Pierce, William M. Old, Hiu T. Cheung, Steven Russell, Joy L. Wattawa, Geoff R. Goehle, Robin D. Knight, and Natalie G. Ahn. Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Anal Chem*, 76(13):3556–68, Jul 2004.
- [121] S. Reumann, J. Davila-Aponte, and K. Keegstra. The evolutionary origin of the protein-translocating channel of chloroplastic envelope membranes: identification of a cyanobacterial homolog. *Proc Natl Acad Sci U S A*, 96(2):784–9, 1999. Journal Article United states.
- [122] Sigrun Reumann, Kentaro Inoue, and Kenneth Keegstra. Evolution of the general protein import pathway of plastids (review). *Mol Membr Biol*, 22(1-2):73–86, Jan-Apr 2005.
- [123] E. Richly and D. Leister. An improved prediction of chloroplast proteins reveals diversities and commonalities in the chloroplast proteomes of *Arabidopsis* and rice. *Gene*, 329:11–6, 2004. Journal Article Netherlands.
- [124] Karine G. Le Roch, Jeffrey R. Johnson, Laurence Florens, Yingyao Zhou, Andrey Santosyan, Munira Grainger, S. Frank Yan, Kim C. Williamson, Anthony A. Holder, Daniel J. Carucci, John R. Yates 3rd, and Elizabeth A. Winzeler. Global analysis of transcript and protein levels across the *Plasmodium falciparum* life cycle. *Genome Res*, 14(11):2308–18, Nov 2004.
- [125] D. Roise, S. J. Horvath, J. M. Tomich, J. H. Richards, and G. Schatz. A chemically synthesized pre-sequence of an imported mitochondrial protein can form an amphiphilic helix and perturb natural and artificial phospholipid bilayers. *EMBO J*, 5(6):1327–34, Jun 1986.
- [126] N. Rolland, M. Ferro, D. Seigneurin-Berny, J. Garin, R. Douce, and J. Joyard. Proteomics of chloroplast envelope membranes. *Photosynth Res*, 78(3):205–30, 2003.

- [127] S. Saxonov, I. Daizadeh, A. Fedorov, and W. Gilbert. EID: the Exon-Intron Database—an exhaustive database of protein-coding intron-containing genes. *Nucleic Acids Res*, 28(1):185–90, Jan 2000.
- [128] T. Sazuka, M. Yamaguchi, and O. Ohara. Cyano2Dbase updated: linkage of 234 protein spots to corresponding genes through N-terminal microsequencing. *Electrophoresis*, 20(11):2160–71, Aug 1999.
- [129] Enrico Schleiff, Lutz Andreas Eichacker, Kerstin Eckart, Thomas Becker, Oliver Mirus, Tanja Stahl, and Jurgen Soll. Prediction of the plant beta-barrel proteome: a case study of the chloroplast outer envelope. *Protein Sci*, 12(4):748–59, Apr 2003.
- [130] Gisbert Schneider and Uli Fechner. Advances in the prediction of protein targeting signals. *Proteomics*, 4(6):1571–80, Jun 2004.
- [131] T. D. Schneider and R. M. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, 18(20):6097–100, Oct 1990.
- [132] R. Schwartz, C. S. Ting, and J. King. Whole proteome pI values correlate with subcellular localizations of proteins for organisms within the three domains of life. *Genome Res*, 11(5):703–9, May 2001.
- [133] Michelle S. Scott, David Y. Thomas, and Michael T. Hallett. Predicting subcellular localization via protein motif co-occurrence. *Genome Res*, 14(10A):1957–66, Oct 2004.
- [134] P. M. Sharp and W. H. Li. An evolutionary perspective on synonymous codon usage in unicellular organisms. *J Mol Evol*, 24(1-2):28–38, 1986.
- [135] Hong-Bin Shen and Kuo-Chen Chou. Predicting protein subnuclear location with optimized evidence-theoretic k-nearest classifier and pseudo amino acid composition. *Biochem Biophys Res Commun*, 337(3):752–6, Nov 2005.
- [136] A. Shevchenko, O. N. Jensen, A. V. Podtelejnikov, F. Sagliocco, M. Wilm, O. Vorm, P. Mortensen, A. Shevchenko, H. Boucherie, and M. Mann. Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. *Proc Natl Acad Sci U S A*, 93(25):14440–5, Dec 1996.
- [137] B. A. Shoemaker, J. J. Portman, and P. G. Wolynes. Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. *Proc Natl Acad Sci U S A*, 97(16):8868–73, Aug 2000.
- [138] William J. Simon, John J. Hall, Iwane Suzuki, Norio Murata, and Antoni R. Slabas. Proteomic study of the soluble proteins from the unicellular cyanobacterium *Synechocystis* sp. PCC6803 using automated matrix-assisted laser desorption/ionization-time of flight peptide mass fingerprinting. *Proteomics*, 2(12):1735–42, Dec 2002.

- [139] Ian Small, Nemo Peeters, Fabrice Legeai, and Claire Lurin. Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics*, 4(6):1581–90, Jun 2004.
- [140] George W. Snedecor and William G. Cochran. *Statistical Methods*, chapter 10, pages 191–193. Iowa State University Press, 1980.
- [141] M. M. Snel, A.I. de Kroon, and D. Marsh. Mitochondrial presequence inserts differently into membranes containing cardiolipin and phosphatidylglycerol. *Biochemistry*, 34(11):3605–13, Mar 1995.
- [142] K. Sohrt and J. Soll. Toc64, a new component of the protein translocon of chloroplasts. *J Cell Biol*, 148(6):1213–21, Mar 2000.
- [143] E. L. Sonnhammer, G. von Heijne, and A. Krogh. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol*, 6:175–82, 1998.
- [144] Sandra Stegemann, Stefanie Hartmann, Stephanie Ruf, and Ralph Bock. High-frequency gene transfer from the chloroplast genome to the nucleus. *Proc Natl Acad Sci U S A*, 100(15):8828–33, Jul 2003.
- [145] Jurgen M. Steiner, Fumie Yusa, Johannes A. Pompe, and Wolfgang Löffelhardt. Homologous protein import machineries in chloroplasts and cyanelles. *Plant J*, 44(4):646–52, Nov 2005.
- [146] Qi Sun, Olof Emanuelsson, and Klaas J. van Wijk. Analysis of curated and predicted plastid subproteomes of *Arabidopsis*. subcellular compartmentalization leads to distinctive proteome properties. *Plant Physiol*, 135(2):723–34, Jun 2004.
- [147] John E. P. Syka, Joshua J. Coon, Melanie J. Schroeder, Jeffrey Shabanowitz, and Donald F. Hunt. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc Natl Acad Sci U S A*, 101(26):9528–33, Jun 2004.
- [148] David L. Tabb, W. Hayes McDonald, and John R. Yates 3rd. DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J Proteome Res*, 1(1):21–6, Jan-Feb 2002.
- [149] David L. Tabb, Anita Saraf, and John R. Yates 3rd. GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Anal Chem*, 75(23):6415–21, Dec 2003.
- [150] K. Tanaka, H. Waki, Y. Ido, S. Akita, Y. Yoshida, and T. Yoshida. Protein and polymer analyses up to m/z 100 000 by laser ionization time-of flight mass spectrometry. *Rapid Commun Mass Spectrom*, 2:151–153, 1988.
- [151] J. A. Taylor and R. S. Johnson. Sequence database searches via *de novo* peptide sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom*, 11(9):1067–75, 1997.

- [152] Oliver Thimm, Oliver Blasing, Yves Gibon, Axel Nagel, Svenja Meyer, Peter Kruger, Joachim Selbig, Lukas A. Muller, Seung Y. Rhee, and Mark Stitt. MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J*, 37(6):914–39, Mar 2004.
- [153] Leeann E. Thornton, Hiroshi Ohkawa, Johnna L. Roose, Yasuhiro Kashino, Nir Keren, and Himadri B. Pakrasi. Homologs of plant PsbP and PsbQ proteins are necessary for regulation of photosystem II activity in the cyanobacterium *Synechocystis* 6803. *Plant Cell*, 16(8):2164–75, Aug 2004.
- [154] P. E. Thorsness and T. D. Fox. Escape of DNA from mitochondria to the nucleus in *Saccharomyces cerevisiae*. *Nature*, 346(6282):376–9, Jul 1990.
- [155] Klaas J. van Wijk. Plastid proteomics. *Plant Physiol Biochem*, 42(12):963–77, Dec 2004.
- [156] Arsenio Villarejo, Stefan Buren, Susanne Larsson, Annabelle Dejardin, Magnus Monne, Charlotta Rudhe, Jan Karlsson, Stefan Jansson, Patrice Lerouge, Norbert Rolland, Gunnar von Heijne, Markus Grebe, Laszlo Bako, and Goran Samuelsson. Evidence for a protein transported through the secretory pathway *en route* to the higher plant chloroplast. *Nat Cell Biol*, 7(12):1224–31, Dec 2005.
- [157] Xiu-Feng Wan, Nathan C. Verberkmoes, Lee Ann McCue, Dawn Stanek, Heather Connelly, Loren J. Hauser, Liyou Wu, Xueduan Liu, Tingfen Yan, Adam Leaphart, Robert L. Hettich, Jizhong Zhou, and Dorothea K. Thompson. Transcriptomic and proteomic characterization of the Fur modulon in the metal-reducing bacterium *Shewanella oneidensis*. *J Bacteriol*, 186(24):8385–400, Dec 2004.
- [158] Rong Wang, John T. Prince, and Edward M. Marcotte. Mass spectrometry of the *M. smegmatis* proteome: protein expression levels correlate with function, operons, and codon bias. *Genome Res*, 15(8):1118–26, Aug 2005.
- [159] Y. Wang, J Sun, and P. R. Chitnis. Proteomic study of the peripheral proteins from thylakoid membranes of the cyanobacterium *Synechocystis* sp. PCC 6803. *Electrophoresis*, 21(9):1746–54, May 2000.
- [160] H. L. Wienk, R. W. Wechselberger, M. Czisch, and B. de Kruijff. Structure, dynamics, and insertion of a chloroplast targeting peptide in mixed micelles. *Biochemistry*, 39(28):8219–27, Jul 2000.
- [161] Xiaoyu Yang, Vijay Dondeti, Rebecca Dezube, Dawn M. Maynard, Lewis Y Geer, Jonathan Epstein, Xiongfong Chen, Sanford P. Markey, and Jeffrey A. Kowalak. DBParser: web-based software for shotgun proteomic data analyses. *J Proteome Res*, 3(5):1002–8, Sep-Oct 2004.
- [162] Xiao-Ping Zhang and Elzbieta Glaser. Interaction of plant mitochondrial and chloroplast signal peptides with the Hsp70 molecular chaperone. *Trends Plant Sci*, 7(1):14–21, Jan 2002.

Appendix

Appendix A

Chloroplast Proteome Results - Sequest

The following table contains the non-redundant list of proteins identified in the chloroplast proteomics study using the Sequest two-peptide rule. The annotations were edited to fit.

Table A.1: Chloroplast Two-Peptide Proteins

Locus	Covg. ^a	N ^b	Annotation ^c
AtCg00480	89.4	72	ATPase beta subunit [Arabidopsis thaliana]
AtCg00470	78.8	13	ATPase epsilon subunit [Arabidopsis thaliana]
At2g39730.1	76.8	56	auxin-regulated protein
At3g12780.1	76.1	59	phosphoglycerate kinase -related similar to phosphoglycerate
At4g02770.1	74.5	28	photosystem I reaction center subunit II precursor -related
AtCg00490	72.2	106	large subunit of ribulose-1,5-bisphosphate carboxylase/oxygen
At1g67090.1	71.1	39	ribulose-bisphosphate carboxylase small unit -related simila
At1g42970.1	69.6	48	glyceraldehyde-3-phosphate dehydrogenase identical to glycer
At4g20360.1	69.1	44	elongation factor Tu (EF-Tu) identical to SWISS-PROT:P17745
At1g09340.1	69.0	45	RNA-binding protein -related Identical to gb—Y10557 g5bf gen
At1g03130.1	67.6	27	photosystem I reaction center subunit II precursor -related
At5g66570.1	67.5	47	photosystem II oxygen-evolving complex 33 (OEC33) identical
At3g01500.1	67.0	31	carbonic anhydrase, chloroplast precursor identical to carbo
At3g26650.1	65.7	50	glyceraldehyde 3-phosphate dehydrogenase A subunit (GapA) id
At4g04640.1	65.7	20	ATP synthase gamma chain 1, chloroplast (H(+)-transporting t
At3g50820.1	64.7	33	photosystem II oxygen-evolving complex 33 (OEC33) identical
At5g38430.1	64.6	35	ribulose bisphosphate carboxylase small chain 1b precursor
At5g38410.1	64.6	39	ribulose bisphosphate carboxylase small chain 3b precursor
AtCg01060	64.2	7	PSI 9KDa protein [Arabidopsis thaliana]
At4g28750.1	63.6	12	photosystem I subunit PSI-E - like protein photosystem I cha
At3g60750.1	63.2	48	transketolase - like protein transketolase, Solanum tuberosu
At2g28000.1	62.1	33	RuBisCO subunit binding-protein alpha subunit/60 kDa chapero
AtCg00820	62.0	6	ribosomal protein S19 [Arabidopsis thaliana]
At2g34420.1	61.1	25	photosystem II type I chlorophyll a /b binding protein iden
At1g06680.1	60.8	26	photosystem II oxygen-evolving complex 23 (OEC23) JBC 14:211
At5g35630.1	60.7	30	glutamate-ammonia ligase (EC 6.3.1.2) precursor, chloroplast
At2g34430.1	60.5	20	photosystem II type I chlorophyll a /b binding protein

Continued on next page

Table A.1: Continued

Locus	Covg. ^a	N ^b	Annotation ^c
At4g38970.1	60.3	32	fructose-bisphosphate aldolase, putative strong similarity
At2g05070.1	60.0	16	light-harvesting chlorophyll a/b binding protein
At4g09650.1	59.8	16	H ⁺ -transporting ATP synthase-related protein H ⁺ -transporting
At3g27690.1	59.8	16	light harvesting chlorophyll A/B binding protein, putative s
At2g47400.1	59.7	6	chloroplast protein CP12 -related
At4g21280.1	59.6	40	oxygen-evolving complex protein 16, chloroplast precursor
At1g52230.1	56.6	12	photosystem I subunit VI precursor identical to photosystem
At3g16140.1	56.6	12	photosystem I subunit VI precursor identical to GB:CAB52749
At1g32060.1	56.2	23	phosphoribulokinase precursor identical to phosphoribulokina
At2g20260.1	55.9	11	photosystem I reaction center subunit IV -related
At1g12900.1	55.9	44	glyceraldehyde 3-phosphate dehydrogenase, chloroplast, putative
At4g05180.1	55.7	25	oxygen-evolving complex protein 16, chloroplast precursor
At5g20720.1	55.3	14	chloroplast Cpn21 protein identical to chloroplast 20 kDa ch
At2g37660.1	55.1	16	expressed protein
At1g55490.1	54.3	31	RuBisCo subunit binding-protein beta subunit/60 kDa chaperon
At1g60950.1	54.1	2	ferredoxin, chloroplast identical to FERREDOXIN PRECURSOR GB
At4g10340.1	53.2	21	light-harvesting chlorophyll a/b binding protein chlorophyll
AtCg00540	50.9	22	cytochrome f [Arabidopsis thaliana]
AtCg00130	50.5	12	ATPase I subunit [Arabidopsis thaliana]
At1g03680.1	49.7	8	thioredoxin M-type 1, chloroplast precursor (TRX-M1) nearly
AtCg00790	49.6	6	ribosomal protein L16 [Arabidopsis thaliana]
At2g05100.1	49.1	15	light-harvesting chlorophyll a/b binding protein
At3g47070.1	49.0	6	expressed protein
At4g34620.1	48.7	6	ribosomal protein S16p family ribosomal protein S16, Neurosp
At5g45390.1	48.3	8	ATP-dependent Clp protease proteolytic subunit (ClpP4) ident
At3g62030.1	47.3	13	peptidylprolyl isomerase ROC4
At1g29910.1	47.2	24	photosystem II type I chlorophyll a /b binding protein, put
At3g26060.1	47.2	14	peroxiredoxin -related similar to peroxiredoxin Q GB:BAA9052
At3g63140.1	46.6	12	mRNA binding protein precursor - like mRNA binding protein p
At5g09650.1	46.3	14	inorganic pyrophosphatase - like protein cytosolic inorganic
At1g31330.1	46.2	15	photosystem I subunit III precursor -related similar to phot
AtCg00120	45.6	30	ATPase alpha subunit [Arabidopsis thaliana]
At4g03280.1	45.4	16	Rieske FeS protein (component of cytochrome B6-F complex) id
At1g79040.1	45.0	8	photosystem II polypeptide -related similar to photosystem I
At1g05190.1	44.8	10	ribosomal protein L6p family Similar to Mycobacterium RlpF
AtCg00830	43.8	12	ribosomal protein L2 [Arabidopsis thaliana]
At1g61520.1	43.6	12	light-harvesting chlorophyll a/b binding protein similar to
At1g20340.1	43.1	6	plastocyanin similar to plastocyanin GI:1865683 from [Arabid
At4g25050.1	42.3	6	acyl carrier protein family similar to Acyl carrier protein,
At1g20020.1	42.3	13	ferredoxin-NADP(+) reductase (adrenodoxin reductase), putat
At1g56190.1	42.3	30	phosphoglycerate kinase -related similar to phosphoglycerate
At3g13470.1	41.6	25	chaperonin, putative similar SWISS-PROT:P21240- RuBisCO subu
At3g15360.1	41.5	10	thioredoxin M-type 4, chloroplast precursor (TRX-M4) nearly
At4g24770.1	41.3	11	31 kDa ribonucleoprotein, chloroplast (RNA-binding protein R
At2g28190.1	40.7	5	copper/zinc superoxide dismutase (CSD2) identical to GP:3273
At4g32260.1	40.2	12	H ⁺ -transporting ATP synthase chain 9 - like protein H ⁺ -trans
At2g21330.1	40.1	19	fructose-bisphosphate aldolase, putative strong similarity t
At5g13510.1	40.0	6	ribosomal protein L10p family ribosomal protein L10- Nicotia

Continued on next page

Table A.1: Continued

Locus	Covg. ^a	N ^b	Annotation ^c
At3g01480.1	39.6	9	thylakoid lumen rotamase similar to thylakoid lumen rotamase
At1g30380.1	39.2	4	photosystem I subunit X precursor identical to photosystem I
At3g46780.1	39.0	14	expressed protein
At1g54780.1	38.9	8	thylakoid lumen 18.3 kDa protein SP:Q9ZVL6
At3g08940.1	38.8	8	chlorophyll a/b-binding protein -related similar to chloroph
At3g47470.1	38.6	8	light-harvesting chlorophyll a/b binding protein CHLOROPHYLL
At2g25080.1	38.6	10	glutathione peroxidase, putative Contains Glutathione peroxi
At3g52960.1	38.5	9	peroxiredoxin - like protein peroxiredoxin TPx2, Arabidopsis
At2g37220.1	38.1	7	29 kDa ribonucleoprotein, chloroplast (RNA-binding protein c
At2g05990.1	37.9	9	enoyl-[acyl-carrier protein] reductase [NADH] (enr-A), putat
At4g12800.1	37.9	6	probable photosystem I chain XI precursor photosystem I chai
At1g78630.1	37.8	12	ribosomal protein L13p family similar to ribosomal protein L
AtCg00380	37.8	6	ribosomal protein S4 [Arabidopsis thaliana]
At1g15820.1	37.6	11	chlorophyll a/b-binding protein Lhcb6 nearly identical to Lh
At1g74470.1	37.5	14	geranylgeranyl reductase identical to geranylgeranyl reducta
At3g54890.1	36.9	10	light-harvesting chlorophyll a/b binding protein
At2g43750.1	36.5	9	cysteine synthase, chloroplast (O-acetylserine (thiol)-lyase
At3g63490.1	35.0	10	ribosomal protein L1p family ribosomal protein L1, S.olerace
At5g58330.1	34.8	11	malate dehydrogenase [NADP], chloroplast, putative strong si
At3g27830.1	34.6	11	50S ribosomal protein L12-1, chloroplast precursor (CL12-A)
At3g47520.1	34.5	10	malate dehydrogenase [NAD], chloroplast, putative strong sim
At5g23120.1	34.5	10	photosystem II stability/assembly factor HCF136 sp—O82660
AtCg00780	34.4	3	ribosomal protein L14 [Arabidopsis thaliana]
At5g50920.1	34.3	28	ATP-dependent Clp protease ATP-binding subunit (ClpC1) almos
At4g01310.1	34.0	9	ribosomal protein L5p family
At5g54270.1	34.0	5	light-harvesting chlorophyll a/b binding protein, putative
At2g21530.1	34.0	5	expressed protein
At5g66190.1	33.9	14	ferredoxin-NADP(+) reductase (adrenodoxin reductase), putat
At3g63190.1	33.8	5	expressed protein chloroplast ribosome recycling factor prot
AtCg00280	33.8	22	PSII 43 KDa protein [Arabidopsis thaliana]
At4g03520.1	33.3	5	thioredoxin M-type 2, chloroplast precursor (TRX-M2) nearly
At2g43030.1	33.2	6	ribosomal protein L3p family
At3g11630.1	33.1	7	2-cys peroxiredoxin BAS1 precursor (thiol-specific antioxida
At3g63160.1	31.9	2	expressed protein outer envelope membrane protein E 6.7 - ch
At2g15620.1	31.9	10	ferredoxin-nitrite reductase
At2g28900.1	31.8	5	membrane channel protein -related
At5g61410.1	31.7	5	ribulose-5-phosphate-3-epimerase
At1g63940.1	31.3	11	monodehydroascorbate reductase, putative similar to monodehy
At3g04790.1	31.2	10	ribose 5-phosphate isomerase -related identical to putative
AtCg00840	31.2	2	ribosomal protein L23 [Arabidopsis thaliana]
At1g03600.1	31.0	10	photosystem II protein family similar to SP:P74367 {Synechoc
At2g38140.1	30.5	5	chloroplast 30S ribosomal protein S31
At1g53240.1	30.5	4	malate dehydrogenase [NAD], mitochondrial, putative nearly i
At2g20890.1	29.7	9	expressed protein
At5g14740.1	29.6	18	CARBONIC ANHYDRASE 2
At3g44890.1	29.4	4	50S ribosomal protein L9, chloroplast precursor (CL9)
At5g45680.1	29.3	3	immunophilin / FKBP-type peptidyl-prolyl cis-trans isomerase
At3g52150.1	29.2	2	RNA recognition motif (RRM) - containing protein similar to

Continued on next page

Table A.1: Continued

Locus	Covg. ^a	N ^b	Annotation ^c
At3g61470.1	29.2	6	light-harvesting chlorophyll a/b binding protein
AtCg00680	29.1	33	PSII 47KDa protein [Arabidopsis thaliana]
At5g42270.1	28.8	16	FtsH protease, putative similar to FtsH protease GI:13183728
At5g01530.1	28.6	11	light-harvesting chlorophyll a/b binding protein
At1g75350.1	28.5	4	chloroplast 50S ribosomal protein L31 -related similar to SP
At2g14880.1	28.4	3	expressed protein
At1g74970.1	28.4	5	ribosomal protein S9 -related similar to ribosomal protein S
At4g09010.1	28.4	7	ascorbate peroxidase, putative identical to SP—P82281—TL29_A
At4g34290.1	27.8	2	expressed protein various predicted proteins
At5g01600.1	27.8	9	ferritin 1 precursor
At3g23400.1	27.8	6	plastid-lipid associated protein PAP/fibrillin family contai
At3g53900.1	27.7	6	uracil phosphoribosyltransferase-related protein uracil phos
At5g24490.1	27.6	4	expressed protein contains similarity to ribosomal protein 3
At3g18890.1	27.6	13	expressed protein similar to UV-B and ozone similarly regula
At2g31670.1	27.4	4	expressed protein
At2g44650.1	27.3	3	chloroplast chaperonin 10 identical to chloroplast chaperoni
AtCg00330	27.0	2	ribosomal protein S14 [Arabidopsis thaliana]
At4g18480.1	26.9	5	magnesium-chelatase, subunit chlI, chloroplast (Mg-protoporp
At2g19940.1	26.7	6	N-acetyl-gamma-glutamyl-phosphate reductase -related
At1g13270.1	26.6	3	methionine aminopeptidase I (MAP1) -related similar to methi
AtMg00280	26.4	4	hypothetical protein [Arabidopsis thaliana]
At5g06290.1	26.0	5	2-cys peroxiredoxin-related protein
At1g23740.1	25.9	8	oxidoreductase, zinc-binding dehydrogenase family contains P
At1g31180.1	25.7	7	3-isopropylmalate dehydrogenase, chloroplast, putative stron
AtCg00800	25.7	5	ribosomal protein S3 [Arabidopsis thaliana]
At3g26070.1	25.6	3	plastid-lipid associated protein PAP/fibrillin family contai
At3g25860.1	25.4	7	dihydrolipoamide S-acetyltransferase identical to GB:AAD5513
At3g48870.1	25.3	20	ATP-dependent Clp protease ATP-binding subunit (ClpC2) ident
At4g04020.1	25.2	2	plastid-lipid associated protein PAP/fibrillin, putative str
At5g64040.1	25.1	9	photosystem I reaction center subunit PSI-N precursor (PSI-N
At1g50250.1	25.1	16	chloroplast FtsH protease almost identical to chloroplast Ft
At4g02530.1	25.0	8	chloroplast thylakoid lumen protein SP:022773 ;TL16.iARATH
At5g54770.1	24.9	4	thiazole biosynthetic enzyme precursor (ARA6) (sp Q38814)
At5g40950.1	24.7	7	50S ribosomal protein L27, chloroplast precursor (CL27)
At2g43560.1	24.7	4	immunophilin / FKBP-type peptidyl-prolyl cis-trans isomerase
At2g17265.1	24.6	3	homoserine kinase identical to homoserine kinase [Arabidopsi
At5g23060.1	24.3	6	expressed protein similar to unknown protein (emb—CAB75797.1
At5g53490.1	24.2	7	thylakoid lumenal 17.4 kD pentapeptide repeat family protein
At3g55800.1	23.9	11	sedoheptulose-bisphosphatase precursor
AtCg01120	23.9	3	ribosomal protein S15 [Arabidopsis thaliana]
At2g30950.1	23.6	18	FtsH protease (VAR2) identical to zinc dependent protease VA
At2g47730.1	23.2	2	glutathione transferase, putative (GST6) identical to GB:X95
At4g22240.1	23.2	3	plastid-lipid associated protein (PAP), putative similar to
At5g30510.1	22.8	10	ribosomal protein S1
At1g34000.1	22.7	2	expressed protein contains similarity to photosystem II 22 k
At4g24930.1	22.7	2	thylakoid lumenal 17.9 kDa protein, chloroplast precursor SP
At3g58610.1	22.2	11	ketol-acid reductoisomerase
At5g54600.1	22.2	5	50S ribosomal protein L24, chloroplast precursor (CL24)

Continued on next page

Table A.1: Continued

Locus	Covg. ^a	N ^b	Annotation ^c
At4g01050.1	21.9	7	proline-rich protein family
AtCg00270	21.8	10	PSII D2 protein [Arabidopsis thaliana]
At3g56130.1	21.7	3	expressed protein acetyl-CoA carboxylase, biotin carboxyl ca
AtCg00160	21.6	5	ribosomal protein S2 [Arabidopsis thaliana]
At5g04140.1	21.6	25	glutamate synthase [ferredoxin] (ferredoxin-dependent glutam
At5g08280.1	21.5	7	hydroxymethylbilane synthase (porphobilinogen deaminase, chl
At3g48420.1	21.3	5	haloacid dehalogenase-like hydrolase family low similarity t
At5g46290.1	20.7	5	3-oxoacyl-[acyl-carrier-protein] synthase I precursor (beta-
At5g16390.1	20.7	3	biotin carboxyl carrier protein 1 (BCCP1) identical to bioti
At2g21170.1	20.3	8	triosephosphate isomerase, chloroplast, putative similar to
At5g09660.1	20.3	2	malate dehydrogenase, glyoxysomal identical to SP—Q9ZP05
At1g77490.1	20.2	4	thylakoid-bound ascorbate peroxidase, putative (tAPX) identi
At4g24280.1	20.1	11	heat shock protein cpHsc70-1 heat shock 70 protein - Spinaci
At2g40490.1	20.1	7	uroporphyrinogen decarboxylase (UPD), putative similar to ur
At2g33800.1	20.1	5	ribosomal protein S5 family
At3g54900.1	19.7	2	CAXIP1 protein (CAXIP1) annotation temporarily based on supp
At1g44575.1	19.6	5	photosystem II 22kDa protein -related similar to photosystem
At4g08390.1	19.4	3	stromal ascorbate peroxidase, putative (sAPX) identical to s
At2g35490.1	19.4	5	plastid-lipid associated protein (PAP), putative similar to
At1g16880.1	19.3	7	expressed protein
At1g77090.1	19.2	2	thylakoid lumenal 29.8 kDa protein identical to SP—O49292—TL
At4g23100.1	19.2	5	gamma-glutamylcysteine synthetase
At1g32990.1	18.9	3	ribosomal protein L11p family similar to chloroplast ribosom
At3g54050.1	18.9	8	fructose-bisphosphatase precursor
AtCg00350	18.7	14	PSI P700 apoprotein A1 [Arabidopsis thaliana]
At5g63310.1	18.6	6	nucleotide diphosphate kinase Ia (emb—CAB58230.1)
AtCg01240	18.1	4	ribosomal protein S7 [Arabidopsis thaliana]
At1g06430.1	18.0	12	FtsH protease, putative similar to zinc dependent protease G
At1g80560.1	18.0	4	3-isopropylmalate dehydrogenase, chloroplast, putative stron
At1g08380.1	17.9	3	expressed protein
AtCg00340	17.6	11	PSI P700 apoprotein A2 [Arabidopsis thaliana]
At3g48730.1	17.4	4	glutamate-1-semialdehyde 2,1-aminomutase 2 (GSA 2) (glutamat
At5g66550.1	17.4	3	expressed protein similar to unknown protein (gb—AAD20709.1)
At1g76080.1	16.9	4	thioredoxin family low similarity to thioredoxin (TRX) [Fasc
At1g21500.1	16.7	1	expressed protein
At1g07320.1	16.7	6	50S ribosomal protein L4, chloroplast precursor (CL4) identi
At1g68590.1	16.3	3	expressed protein
At1g69740.1	16.3	2	porphobilinogen synthase (delta-aminolevulinic acid dehydrat
At5g65220.1	16.2	6	ribosomal protein L29p family
At5g56500.1	16.1	9	chaperonin, putative similar to SWISS-PROT:P08927- RuBisCO s
At1g29700.1	16.0	2	expressed protein
At1g03475.1	16.0	3	coproporphyrinogen III oxidase (coproporphyrinogenase) (copr
At5g17170.1	15.9	2	expressed protein
At1g20620.1	15.9	5	catalase 3 almost identical to catalase 3 SP:Q42547, GI:3123
At4g35090.1	15.9	5	catalase 2 identical to catalase 2 SP:P25819, GI:17865693 fr
AtCg00650	15.8	2	ribosomal protein S18 [Arabidopsis thaliana]
At2g30790.1	15.7	9	photosystem II oxygen-evolving complex 23 (OEC23) expression
At3g25920.1	15.5	3	50S ribosomal protein L15, chloroplast precursor (CL15) iden

Continued on next page

Table A.1: Continued

Locus	Covg. ^a	N ^b	Annotation ^c
At5g49910.1	15.5	8	heat shock protein cpHsc70-2 (hsc70-7)
At1g54630.1	15.4	2	acyl carrier protein 3 (ACP), chloroplast nearly identical t
At1g54580.1	15.4	2	acyl carrier protein (ACP), chloroplast, putative strong sim
At4g00620.1	15.3	3	tetrahydrofolate dehydrogenase/cyclohydrolase, putative simi
At4g01690.1	15.3	5	protoporphyrinogen oxidase (PPO) (PPOX) identical to SP—P558
At3g63410.1	15.1	2	chloroplast inner envelope membrane protein, putative simila
AtCg00020	15.0	5	PSII 32 KDa protein [Arabidopsis thaliana]
At2g24020.1	14.8	2	expressed protein
At2g43090.1	14.7	2	aconitase family contains Pfam profile PF00694: Aconitase C-
At2g42220.1	14.5	4	rhodanese-like domain protein contains rhodanese-like domain
At1g35680.1	14.5	2	50S ribosomal protein L21 chloroplast precursor (CL21) ident
At5g16710.1	14.3	2	dehydroascorbate reductase, putative Strong similarity to de
At1g48350.1	14.1	2	ribosomal protein L18p family similar to ribosomal protein L
At4g33680.1	13.7	3	aminotransferase family low similarity to Aromatic Aminotran
At1g02780.1	13.6	2	60S ribosomal protein L19 (RPL19A) similar to ribosomal prot
At3g14930.1	13.6	3	uroporphyrinogen decarboxylase (UPD), putative similar to ur
At3g21055.1	13.6	4	photosystem II 5 kD protein precursor annotation temporarily
At3g15190.1	13.4	2	chloroplast 30S ribosomal protein S20, putative contains Pfa
At1g51400.1	13.2	3	photosystem II 5 KD protein 100% identical to GI:4836947 (F5
At4g23890.1	13.2	3	expressed protein hypothetical protein, Synechocystis sp., P
At5g03880.1	13.0	2	auxin-regulated protein predicted protein, Arabidopsis thali
At3g13120.1	12.6	2	chloroplast 30S ribosomal protein S10, putative similar to 3
At1g55480.1	12.5	4	expressed protein
At4g27440.1	12.5	3	protochlorophyllide reductase B (PCR B/POR B) identical to p
At5g47190.1	12.2	3	ribosomal protein L19, putative similar to plastid ribosomal
At3g14930.1	12.0	2	uroporphyrinogen decarboxylase (UPD), putative similar to ur
At5g07020.1	11.9	3	proline-rich protein family
At4g24620.1	11.9	3	glucose-6-phosphate isomerase glucose-6-phosphate isomerase
At2g33150.1	11.7	2	acetyl-CoA C-acyltransferase (3-ketoacyl-CoA thiolase), puta
At3g16950.1	11.6	2	dihydrolipoamide dehydrogenase 1, plastidic (lipoamide dehyd
At2g25840.1	11.5	3	tryptophanyl-tRNA synthetase -related
At1g62780.1	11.4	2	expressed protein
At5g63890.1	11.3	2	histidinol dehydrogenase
At2g26340.1	11.3	2	expressed protein
At1g16720.1	11.2	4	expressed protein
At3g52380.1	11.2	2	33 kDa ribonucleoprotein, chloroplast (RNA-binding protein c
At5g47840.1	11.0	2	expressed protein contains similarity to adenylate kinase
At4g31990.1	11.0	6	aspartate aminotransferase, chloroplast (transaminase A/AAT1
At5g60600.1	10.9	5	GcpE protein; nuclear gene for chloroplast product annotatio
At5g35170.1	10.4	3	adenylate kinase -related protein adenylate kinase (EC 2.7.4
At5g19220.1	10.3	2	glucose-1-phosphate adenylyltransferase, large subunit 1, ch
At1g34430.1	10.3	6	dihydrolipoamide S-acetyltransferase -related similar to GI:
At3g62410.1	9.9	2	CP12 protein precursor-related protein CP12 protein precurs
At1g07920.1	9.8	3	elongation factor 1-alpha (EF-1-alpha) identical to GB:CAA34
AtCg00720	9.8	2	cytochrome B6 [Arabidopsis thaliana]
At5g63570.1	9.7	2	glutamate-1-semialdehyde 2,1-aminomutase 1 (GSA 1) (glutamat
At1g19920.1	9.5	2	sulfate adenylyltransferase identical to sulfate adenylyltra
At4g16155.1	9.5	2	dihydrolipoamide dehydrogenase 2, plastidic (lipoamide dehyd

Continued on next page

Table A.1: Continued

Locus	Covg. ^a	N ^b	Annotation ^c
At4g31870.1	9.4	3	glutathione peroxidase, putative glutathione peroxidase, Ara
At4g29840.1	9.3	3	threonine synthase, chloroplast identical to SP—Q9S7B5 Threo
At4g37930.1	9.3	2	glycine hydroxymethyltransferase like protein glycine hydrox
At3g27925.1	9.3	4	DegP protease SP:022609; almost identical to DegP protease p
AtCg01110	9.2	4	NADH dehydrogenase 49KDa protein [Arabidopsis thaliana]
At5g41670.1	9.2	2	6-phosphogluconate dehydrogenase
At5g04590.1	8.9	2	sulphite reductase
At4g32520.1	8.9	2	glycine hydroxymethyltransferase (EC 2.1.2.1) - like protein
At2g45290.1	8.8	9	transketolase precursor -related
At5g48300.1	8.7	2	glucose-1-phosphate adenylyltransferase, small subunit, chlo
At1g62750.1	8.4	4	elongation factor Tu family protein similar to elongation fa
At5g14320.1	8.3	2	30S ribosomal protein S13, chloroplast precursor (CS13) ribo
At5g35100.1	8.3	2	expressed protein
At2g30200.1	8.1	2	malonyl-CoA:Acyl carrier protein transacylase -related
At4g34120.1	8.0	2	CBS domain containing protein contains Pfam profile PF00571:
At3g23940.1	7.9	2	dihydroxyacid dehydratase -related similar to dihydroxyacid
At2g33450.1	7.7	2	50S ribosomal protein L28, chloroplast precursor (CL28)
At4g34200.1	7.6	2	D-3-phosphoglycerate dehydrogenase (3-PGDH), putative simila
At5g08410.1	7.6	3	lipoic acid synthase -related annotation temporarily based o
At1g15140.1	7.5	2	oxidoreductase NAD-binding domain-containing protein Contain
At5g17710.1	7.4	2	chloroplast GrpE protein
At3g08740.1	7.2	2	elongation factor P (EF-P) -related similar to elongation fa
At1g24360.1	7.2	2	3-oxoacyl-[acyl-carrier protein] reductase identical to 3-ox
At4g29060.1	6.8	4	expressed protein
At4g22890.1	6.8	2	expressed protein
At1g17220.1	6.4	3	translation initiation factor IF-2, chloroplast precursor Ex
At5g46110.1	6.3	2	phosphate/triose-phosphate translocator, putative identical
At3g54660.1	6.2	2	glutathione reductase, chloroplast nearly identical to SP—P4
At3g57560.1	6.1	2	aspartate/glutamate/uridylate kinase family similar to acety
At1g29900.1	6.1	3	carbamoylphosphate synthetase -related similar to carbamoylp
At1g06690.1	6.1	2	aldo/keto reductase family contains Pfam profile PF00248: ox
At1g64190.1	6.0	2	6-phosphogluconate dehydrogenase -related similar to 6-phosp
At2g38040.1	6.0	2	alpha-carboxyltransferase -related
At5g17530.1	5.9	2	phosphoglucomutase-related protein phosphoglucomutase, chlor
At1g56500.1	5.8	2	haloacid dehalogenase-like hydrolase family low similarity t
At5g52520.1	5.7	2	prolyl tRNA synthetase
At5g08670.1	5.6	3	H+-transporting ATP synthase beta chain (mitochondrial) -rel
At3g57610.1	5.3	3	adenylosuccinate synthetase
At1g66520.1	5.1	2	formyltransferase, putative similar to methionyl-tRNA formyl
At3g46740.1	5.0	2	chloroplast import-associated channel protein homolog chloro
At1g63770.1	4.7	3	aminopeptidase -related similar to aminopeptidase N (alpha-a
At1g23310.1	4.6	2	alanine aminotransferase -related similar to alanine aminotr
At2g41220.1	4.5	4	glutamate synthase [ferredoxin], chloroplast (ferredoxin-dep
At3g22890.1	4.5	3	ATP sulfurylase -related similar to ATP sulfurylase GB:AAF19
At5g48220.1	4.5	2	indole-3-glycerol phosphate synthase (IGPS), putative simila
At4g33030.1	4.4	2	UDP-sulfoquinovose synthase (sulfite:UDP-glucose sulfotransf
At5g55220.1	4.4	2	trigger factor-related protein
At2g24820.1	4.3	2	Rieske [2Fe-2S] domain-containing protein similar to Rieske

Continued on next page

Table A.1: Continued

Locus	Covg. ^a	N ^b	Annotation ^c
At2g13360.1	4.2	2	alanine-glyoxylate aminotransferase
At1g79550.1	4.2	2	phosphoglycerate kinase -related similar to phosphoglycerate
At1g49970.1	4.1	2	ATP-dependent Clp protease proteolytic subunit (ClpR1) ident
At1g66200.1	3.9	1	glutamine synthetase -related similar to glutamine synthetas
At5g14200.1	3.7	2	3-isopropylmalate dehydrogenase, chloroplast, putative stron
At2g37500.1	3.6	2	glutamate/ornithine acetyltransferase -related
At1g20630.1	3.3	3	catalase 1 identical to catalase 1 GI:2511725 from [Arabidop
At1g77590.1	2.9	2	long-chain-fatty-acid-CoA ligase (acyl-CoA synthetase) fami
At3g14390.1	2.7	2	diaminopimelate decarboxylase -related similar to diaminopim
At5g53460.1	2.6	2	glutamate synthase [NADH], chloroplast, putative similar to
At2g35040.1	2.5	2	phosphoribosylaminoimidazolecarboxamide formyltransferase -r
At1g12800.1	2.3	2	heat shock factor protein hsf8 -related

^a%Coverage; ^bnumber of unique peptides; ^cArabidopsis Genome Initiative (AGI) annotation (www.tair.org)

Appendix B

Chloroplast Proteome Results - Three Algorithms

These are the chloroplast proteins resulting from applying the filter: two or more peptides identified by any single algorithm; or two or more algorithms assign the same peptide to the same spectrum. The annotation was edited to fit.

Table B.1: Chloroplast Proteins - Two Peptides or Two Algorithms

Locus	Annotation ^a
At3g12780	phosphoglycerate kinase, putative similar to
At5g01530	chlorophyll A-B binding protein CP29 (LHCB4)
At5g38410	ribulose biphosphate carboxylase small chain 3B
AtCg00350	psaA PSI P700 apoprotein A1
AtCg00490	rbcL large subunit of riblose-1,5-bisphosphate
At2g34420	chlorophyll A-B binding protein / LHCII type I
At1g06680	photosystem II oxygen-evolving complex 23 (OEC23)
At1g15820	chlorophyll A-B binding protein, chloroplast
At1g20340	plastocyanin similar to plastocyanin GI:1865683
At1g31330	photosystem I reaction center subunit III family
At1g61520	chlorophyll A-B binding protein / LHCI type III
At2g20260	photosystem I reaction center subunit IV,
At2g28000	RuBisCO subunit binding-protein alpha subunit,
At2g34430	chlorophyll A-B binding protein / LHCII type I
At2g39730	ribulose biphosphate carboxylase/oxygenase
At3g26650	glyceraldehyde 3-phosphate dehydrogenase A,
At3g48870	ATP-dependent Clp protease ATP-binding subunit
At3g60750	transketolase, putative strong similarity to
At4g02770	photosystem I reaction center subunit II,
At4g05180	oxygen-evolving enhancer protein 3, chloroplast,
At4g10340	chlorophyll A-B binding protein CP26, chloroplast
At4g21280	oxygen-evolving enhancer protein 3, chloroplast,
At4g28750	photosystem I reaction center subunit IV,
At4g38970	fructose-bisphosphate aldolase, putative strong
At5g38430	ribulose biphosphate carboxylase small chain 1B
AtCg00120	atpA ATPase alpha subunit

Continued on next page

Table B.1: Continued

Locus	Annotation ^a
AtCg00130	atpF ATPase I subunit
AtCg00280	psbC PSII 43 KDa protein
AtCg00470	atpE ATPase epsilon subunit
AtCg00480	atpB ATPase beta subunit
AtCg00680	psbB PSII 47KDa protein
At3g50820	oxygen-evolving enhancer protein, chloroplast,
At4g03280	cytochrome B6-F complex iron-sulfur subunit,
At5g50920	ATP-dependent Clp protease ATP-binding subunit /
AtCg01060	psaC PSI 9KDa protein
At2g30790	photosystem II oxygen-evolving complex 23,
At3g01500	carbonic anhydrase 1, chloroplast / carbonate
At3g27690	chlorophyll A-B binding protein (LHCB2:4) nearly
At4g24770	31 kDa ribonucleoprotein, chloroplast, putative /
At3g62030	peptidyl-prolyl cis-trans isomerase, chloroplast
At1g79040	photosystem II 10 kDa polypeptide identical to
At5g54270	chlorophyll A-B binding protein / LHCI type III
At5g38420	ribulose biphosphate carboxylase small chain 2B
At1g29930	chlorophyll A-B binding protein 2, chloroplast /
At3g57190	peptide chain release factor, putative similar to
At1g12900	glyceraldehyde 3-phosphate dehydrogenase,
At1g32060	phosphoribulokinase (PRK) / phosphopentokinase
At1g42970	glyceraldehyde-3-phosphate dehydrogenase B,
At2g25080	phospholipid hydroperoxide glutathione
At2g37220	29 kDa ribonucleoprotein, chloroplast, putative /
At3g04790	ribose 5-phosphate isomerase-related similar to
At3g16140	photosystem I reaction center subunit VI,
At3g26060	peroxiredoxin Q, putative similar to
At3g47520	malate dehydrogenase [NAD], chloroplast (MDH)
At3g63140	mRNA-binding protein, putative similar to mRNA
At4g04640	ATP synthase gamma chain 1, chloroplast (ATPC1)
At4g09650	ATP synthase delta chain, chloroplast, putative /
At4g20360	elongation factor Tu / EF-Tu (TUFA) identical to
At5g23060	expressed protein
At5g66570	oxygen-evolving enhancer protein 1-1, chloroplast
AtCg00540	petA cytochrome f
At1g03600	photosystem II family protein similar to
At1g20020	ferredoxin-NADP(+) reductase, putative /
At1g50250	cell division protein ftsH homolog 1, chloroplast
At1g55490	RuBisCO subunit binding-protein beta subunit,
At2g15620	ferredoxin-nitrite reductase, putative strong
At2g30950	FtsH protease (VAR2) identical to zinc dependent
At3g11630	2-cys peroxiredoxin, chloroplast (BAS1) identical
At3g52960	peroxiredoxin type 2, putative similar to type 2
AtCg00020	psbA PSII 32 KDa protein
AtCg00270	psbD PSII D2 protein
AtCg00340	psaB PSI P700 apoprotein A2
At4g32260	ATP synthase family contains Pfam profile:
At1g52230	photosystem I reaction center subunit VI,

Continued on next page

Table B.1: Continued

Locus	Annotation ^a
At2g05100	chlorophyll A-B binding protein / LHCII type II
At4g25050	acyl carrier family protein / ACP family protein
At5g20720	20 kDa chaperonin, chloroplast (CPN21) (CHCPN10)
At3g08940	chlorophyll A-B binding protein (LHCB4.2)
At2g47400	CP12 domain-containing protein contains Pfam
At1g05190	ribosomal protein L6 family protein Similar to
At2g33800	ribosomal protein S5 family protein contains Pfam
At3g46780	expressed protein
At4g12800	photosystem I reaction center subunit XI,
At5g40950	50S ribosomal protein L27, chloroplast, putative
At5g61410	ribulose-phosphate 3-epimerase, chloroplast,
At1g78630	ribosomal protein L13 family protein similar to
At2g44650	chloroplast chaperonin 10 (cpn10) identical to
At4g01050	hydroxyproline-rich glycoprotein family protein
At5g65220	ribosomal protein L29 family protein contains
At5g24490	30S ribosomal protein, putative similar to
At2g46820	expressed protein
At2g21170	triosephosphate isomerase, chloroplast, putative
At2g28190	superoxide dismutase [Cu-Zn], chloroplast (SODCP)
At2g43090	aconitase C-terminal domain-containing protein
At3g18890	expressed protein similar to UV-B and ozone
At1g60950	ferredoxin, chloroplast (PETF) identical to
At3g58990	aconitase C-terminal domain-containing protein
At5g58710	peptidyl-prolyl cis-trans isomerase, putative /
At3g27850	50S ribosomal protein L12-3, chloroplast (CL12-C)
At1g29920	chlorophyll A-B binding protein 165/180,
At3g23370	expressed protein ; expression supported by MPSS
At1g06430	FtsH protease, putative similar to zinc dependent
At1g67090	ribulose biphosphate carboxylase small chain 1A
At3g13470	chaperonin, putative similar SWISS-PROT:P21240-
At5g23120	photosystem II stability/assembly factor,
At5g35630	glutamine synthetase (GS2) identical to glutamine
At1g03130	photosystem I reaction center subunit II,
At1g56190	phosphoglycerate kinase, putative similar to
At3g54050	fructose-1,6-bisphosphatase, putative /
At4g02530	chloroplast thylakoid lumen protein SP:022773
At5g42270	FtsH protease, putative similar to FtsH protease
At3g55800	sedoheptulose-1,7-bisphosphatase, chloroplast /
At5g06290	2-cys peroxiredoxin, chloroplast, putative very
At5g09650	inorganic pyrophosphatase family protein similar
At1g07320	50S ribosomal protein L4, chloroplast (CL4)
At5g04140	glutamate synthase (GLU1) / ferredoxin-dependent
At1g09340	expressed protein
At1g31180	3-isopropylmalate dehydrogenase, chloroplast,
At3g27925	DegP protease, putative SP:022609; almost
At3g47070	expressed protein
At3g63190	ribosome recycling factor, chloroplast, putative
At1g08380	expressed protein

Continued on next page

Table B.1: Continued

Locus	Annotation ^a
At2g20890	expressed protein
At2g21330	fructose-bisphosphate aldolase, putative strong
At3g01480	peptidyl-prolyl cis-trans isomerase, putative /
At3g63490	ribosomal protein L1 family protein ribosomal
At4g03520	thioredoxin M-type 2, chloroplast (TRX-M2) nearly
At4g27440	protochlorophyllide reductase B, chloroplast /
At5g17170	rubredoxin family protein contains Pfam profile
At5g45390	ATP-dependent Clp protease proteolytic subunit
At1g35680	50S ribosomal protein L21, chloroplast / CL21
At1g54780	thylakoid lumen 18.3 kDa protein SP:Q9ZVL6
At2g05620	expressed protein
At4g24280	heat shock protein 70, putative / HSP70, putative
At1g63940	monodehydroascorbate reductase, putative similar
At4g22240	plastid-lipid associated protein PAP, putative
At5g53490	thylakoid lumenal 17.4 kDa protein, chloroplast
At3g21055	photosystem II 5 kD protein, putative identical
At5g64040	photosystem I reaction center subunit PSI-N,
At1g74470	geranylgeranyl reductase identical to
At1g02560	ATP-dependent Clp protease proteolytic subunit
At1g03680	thioredoxin M-type 1, chloroplast (TRX-M1) nearly
At1g34430	dihydrolipoamide S-acetyltransferase, putative
At1g55480	expressed protein
At2g21530	forkhead-associated domain-containing protein /
At2g35490	plastid-lipid associated protein PAP, putative
At3g44890	50S ribosomal protein L9, chloroplast (CL9)
At3g47470	chlorophyll A-B binding protein 4, chloroplast /
At3g58610	ketol-acid reductoisomerase identical to
AtCg00650	rps18 ribosomal protein S18
AtCg00750	rps11 ribosomal protein S11
AtCg00800	rps3.chloroplast ribosomal protein S3
At1g02780	60S ribosomal protein L19 (RPL19A) similar to
At2g28900	mitochondrial import inner membrane translocase
At2g40490	uroporphyrinogen decarboxylase, putative / UPD,
At2g43030	ribosomal protein L3 family protein contains Pfam
At2g43560	immunophilin / FKBP-type peptidyl-prolyl
At3g14930	uroporphyrinogen decarboxylase, putative / UPD,
At3g48420	haloacid dehalogenase-like hydrolase family
At4g16155	dihydrolipoamide dehydrogenase 2, plastidic /
At5g08280	hydroxymethylbilane synthase / porphobilinogen
At1g51400	photosystem II 5 kD protein 100% identical to
At1g23740	oxidoreductase, zinc-binding dehydrogenase family
At5g54600	50S ribosomal protein L24, chloroplast (CL24)
At1g13270	metallopeptidase M24 family protein similar to
At1g16720	expressed protein
At1g16880	uridylyltransferase-related similar to
At1g32990	ribosomal protein L11 family protein similar to
At1g49970	ATP-dependent Clp protease proteolytic subunit
At1g62780	expressed protein

Continued on next page

Table B.1: Continued

Locus	Annotation ^a
At1g74970	ribosomal protein S9 (RPS9) identical to
At2g13360	serine-glyoxylate aminotransferase-related
At2g19940	semialdehyde dehydrogenase family protein similar
At2g31670	expressed protein similar to NADH-ubiquinone
At2g43750	cysteine synthase, chloroplast / O-acetylserine
At3g08740	elongation factor P (EF-P) family protein similar
At3g25920	50S ribosomal protein L15, chloroplast (CL15)
At4g25650	Rieske [2Fe-2S] domain-containing protein similar
At4g31990	aspartate aminotransferase, chloroplast /
At4g33680	aminotransferase class I and II family protein
At4g34120	CBS domain-containing protein contains Pfam
At5g07020	proline-rich family protein
At5g14320	30S ribosomal protein S13, chloroplast (CS13)
At5g35360	acetyl-CoA carboxylase, biotin carboxylase
At5g54770	thiazole biosynthetic enzyme, chloroplast (ARA6)
AtCg00380	rps4 chloroplast ribosomal protein S4
AtCg00790	rpl16.chloroplast ribosomal protein L16
At1g20630	catalase 1 identical to catalase 1 GI:2511725
At2g44920	thylakoid lumenal 15 kDa protein, chloroplast
At3g48110	aminoacyl-t-RNA synthetase, putative similar to
AtCg00140	atpH ATPase III subunit
At1g52670	biotin/lipoyl attachment domain-containing
At5g09660	malate dehydrogenase, glyoxysomal identical to
AtCg00730	petD cytochrome b/f
At1g10450	paired amphipathic helix repeat-containing
At1g16760	protein kinase family protein contains protein
At2g17680	expressed protein contains Pfam profile PF03087:
At2g44050	6,7-dimethyl-8-ribityllumazine synthase / DMRL
At3g17450	hAT dimerisation domain-containing protein
At4g25450	ABC transporter family protein similar to
At5g08680	ATP synthase beta chain, mitochondrial, putative
AtCg01300	rpl2.2_chloro ribosomal protein L2
At2g05070	chlorophyll A-B binding protein / LHCII type II
At3g27830	50S ribosomal protein L12-1, chloroplast (CL12-A)
At3g15360	thioredoxin M-type 4, chloroplast (TRX-M4) nearly
At3g54890	chlorophyll A-B binding protein / LHCI type I
At5g17710	co-chaperone grpE family protein similar to
At1g54630	acyl carrier protein 3, chloroplast (ACP-3)
At3g23400	plastid-lipid associated protein PAP / fibrillin
At2g42220	rhodanese-like domain-containing protein contains
At4g29060	elongation factor Ts family protein similar to
At2g37660	expressed protein
At3g25860	dihydrolipoamide S-acetyltransferase (LTA2)
At3g48730	glutamate-1-semialdehyde 2,1-aminomutase 2 (GSA
At4g09010	L-ascorbate peroxidase, chloroplast, putative
At5g13510	ribosomal protein L10 family protein ribosomal
At1g54580	acyl carrier protein, chloroplast, putative /
At1g62750	elongation factor Tu family protein similar to

Continued on next page

Table B.1: Continued

Locus	Annotation ^a
At1g75350	ribosomal protein L31 family protein similar to
At2g05990	enoyl-[acyl-carrier protein] reductase [NADH],
At3g61870	expressed protein hypothetical protein -
At4g34620	ribosomal protein S16 family protein ribosomal
At5g66190	ferredoxin-NADP(+) reductase, putative /
At1g06950	chloroplast inner envelope protein-related
At1g53240	malate dehydrogenase [NAD], mitochondrial
At1g68590	plastid-specific 30S ribosomal protein 3,
At2g38040	acetyl co-enzyme A carboxylase
At3g15190	chloroplast 30S ribosomal protein S20, putative
At3g61470	chlorophyll A-B binding protein (LHCA2) identical
At5g41670	6-phosphogluconate dehydrogenase family protein
At5g49910	heat shock protein 70 / HSP70 (HSC70-7) identical
At5g58330	malate dehydrogenase [NADP], chloroplast,
At1g30380	photosystem I reaction center subunit psaK,
At2g41220	glutamate synthase, chloroplast (GLU2) /
At3g13120	30S ribosomal protein S10, chloroplast, putative
At4g12060	Clp amino terminal domain-containing protein
At4g22890	expressed protein
At4g31870	glutathione peroxidase, putative glutathione
At3g19170	peptidase M16 family protein / insulinase family
At4g24620	glucose-6-phosphate isomerase, putative similar
At4g27700	rhodanese-like domain-containing protein contains
At5g16710	dehydroascorbate reductase, putative Strong
At1g44575	photosystem II 22kDa protein, chloroplast / CP22
At5g55220	trigger factor type chaperone family protein
AtCg00820	rps19 ribosomal protein S19
At1g06690	aldo/keto reductase family protein contains Pfam
At1g66430	pfkB-type carbohydrate kinase family protein
At1g69740	porphobilinogen synthase, putative /
At2g24820	Rieske [2Fe-2S] domain-containing protein similar
At2g25840	tRNA synthetase class I (W and Y) family protein
At2g30200	expressed protein
At2g38140	chloroplast 30S ribosomal protein S31 (PSRP4)
At3g16950	dihydrolipoamide dehydrogenase 1, plastidic /
At3g23940	dehydratase family contains Pfam profile: PF00920
At3g59980	tRNA-binding region domain-containing protein
At4g33030	UDP-sulfoquinovose synthase / sulfite:UDP-glucose
At4g34290	SWIB complex BAF60b domain-containing protein
At4g35090	catalase 2 identical to catalase 2 SP:P25819,
At5g01600	ferritin 1 (FER1) identical to ferritin
At5g12040	carbon-nitrogen hydrolase family protein similar
At5g30510	30S ribosomal protein S1, putative similar to
At5g35100	peptidyl-prolyl cis-trans isomerase
At5g38530	tryptophan synthase-related low similarity to
At5g46290	3-oxoacyl-[acyl-carrier-protein] synthase I
AtCg00330	rps14 ribosomal protein S14
At1g03475	coproporphyrinogen III oxidase, putative /

Continued on next page

Table B.1: Continued

Locus	Annotation ^a
At1g15140	oxidoreductase NAD-binding domain-containing
At1g20620	catalase 3 (SEN2) almost identical to catalase 3
At1g24360	3-oxoacyl-[acyl-carrier protein] reductase,
At1g29900	carbamoyl-phosphate synthase family protein
At1g74880	expressed protein
At1g80560	3-isopropylmalate dehydrogenase, chloroplast,
At2g35040	AICARFT/IMPCHase bienzyme family protein similar
At3g22890	sulfate adenylyltransferase 1 / ATP-sulfurylase 1
At3g52150	RNA recognition motif (RRM)-containing protein
At4g01310	ribosomal protein L5 family protein contains Pfam
At4g01690	protoporphyrinogen oxidase (PPOX) identical to
At4g34200	D-3-phosphoglycerate dehydrogenase, putative /
At5g14200	3-isopropylmalate dehydrogenase, chloroplast,
At5g47190	ribosomal protein L19 family protein similar to
At5g48300	glucose-1-phosphate adenylyltransferase small
At5g54190	protochlorophyllide reductase A, chloroplast /
At5g63890	histidinol dehydrogenase, putative / HDH,
AtCg00160	rps2 ribosomal protein S2
AtCg00780	rpl14 ribosomal protein L14
At1g56500	haloacid dehalogenase-like hydrolase family
At1g79850	30S ribosomal protein S17, chloroplast / CS17
At4g02930	elongation factor Tu, putative / EF-Tu, putative
At5g14780	formate dehydrogenase (FDH) identical to
At3g54660	glutathione reductase, chloroplast nearly
At4g09040	RNA recognition motif (RRM)-containing protein
At1g11860	aminomethyltransferase, putative similar to
At1g12250	thylakoid lumenal protein-related weak similarity
At1g21350	expressed protein
At1g34000	light stress-responsive one-helix protein (OHP2)
At1g63970	2C-methyl-D-erythritol 2,4-cyclodiphosphate
At1g76080	thioredoxin family protein low similarity to
At2g04400	indole-3-glycerol phosphate synthase (IGPS)
At2g26140	FtsH protease, putative contains similarity to
At2g37500	arginine biosynthesis protein ArgJ family
At3g01440	oxygen evolving enhancer 3 (PsbQ) family protein
At3g02730	thioredoxin, putative similar to SP—P29450
At3g26070	plastid-lipid associated protein PAP / fibrillin
At3g44880	Rieske [2Fe-2S] domain-containing protein similar
At3g52380	33 kDa ribonucleoprotein, chloroplast, putative /
At3g57610	adenylosuccinate synthetase (ADSS) identical to
At4g01900	P II nitrogen sensing protein (GLB I) identical
At4g04020	plastid-lipid associated protein PAP, putative /
At4g08390	L-ascorbate peroxidase, stromal (sAPX) identical
At4g13010	oxidoreductase, zinc-binding dehydrogenase family
At4g13430	aconitase family protein / aconitate hydratase
At4g14210	phytoene dehydrogenase, chloroplast / phytoene
At4g24930	thylakoid lumenal 17.9 kDa protein, chloroplast
At4g29840	threonine synthase, chloroplast identical to

Continued on next page

Table B.1: Continued

Locus	Annotation ^a
At4g37930	glycine hydroxymethyltransferase / serine
At5g08410	ferredoxin-thioredoxin reductase, putative
At5g13120	peptidyl-prolyl cis-trans isomerase
At5g16660	expressed protein
At5g19940	plastid-lipid associated protein PAP-related /
At5g35970	DNA-binding protein, putative similar to
At5g45680	FK506-binding protein 1 (FKBP13) identical to
At5g63310	nucleotide diphosphate kinase II, chloroplast
AtCg00720	petB cytochrome B6
AtCg01120	rps15 ribosomal protein S15
At1g01790	K ⁺ efflux antiporter, putative (KEA1) identical
At1g08520	magnesium-chelatase subunit chlD, chloroplast,
At1g21270	wall-associated kinase 2 (WAK2) identical to
At3g44620	low molecular weight phosphotyrosine protein
At4g00620	tetrahydrofolate dehydrogenase/cyclohydrolase,
At4g18440	adenylosuccinate lyase, putative /
At4g18810	expressed protein similar to UV-B and ozone
At4g25130	peptide methionine sulfoxide reductase, putative
AtCg01230	rps12.2 ribosomal protein S12 (trans-splice part
At5g60390	elongation factor 1-alpha / EF-1-alpha identical
At4g14680	sulfate adenylyltransferase 3 / ATP-sulfurylase 3
At3g02360	6-phosphogluconate dehydrogenase family protein
At1g29910	chlorophyll A-B binding protein 2, chloroplast /
At5g35170	adenylate kinase family protein contains Pfam
AtCg00830	rpl2.1_chloro ribosomal protein L2
At1g79550	phosphoglycerate kinase, putative similar to
At5g14740	carbonic anhydrase 2 / carbonate dehydratase 2
At5g56500	chaperonin, putative similar to
At5g63570	glutamate-1-semialdehyde 2,1-aminomutase 1 (GSA
At4g18480	magnesium-chelatase subunit chlI, chloroplast /
At2g45290	transketolase, putative strong similarity to
AtMg00180	ccb452 cytochrome c biogenesis orf452
AtCg00710	psbH PSII 10KDa phosphoprotein
At3g56130	biotin/lipoyl attachment domain-containing
At3g20390	endoribonuclease L-PSP family protein contains
At3g62410	CP12 domain-containing protein contains Pfam
At5g04590	sulfite reductase / ferredoxin (SIR) identical to
At5g45930	magnesium-chelatase subunit chlI, chloroplast,
At5g47840	adenylate kinase, chloroplast, putative / ATP-AMP
At1g07920	elongation factor 1-alpha / EF-1-alpha identical
At2g04030	heat shock protein, putative strong similarity to
At3g60370	immunophilin / FKBP-type peptidyl-prolyl
At5g08670	ATP synthase beta chain 1, mitochondrial
At5g17870	plastid-specific ribosomal protein-related
At1g19920	sulfate adenylyltransferase 2 / ATP-sulfurylase 2
At1g77490	L-ascorbate peroxidase, thylakoid-bound (tAPX)
At2g14880	SWIB complex BAF60b domain-containing protein
At2g28820	ribosomal protein L16 family protein contains

Continued on next page

Table B.1: Continued

Locus	Annotation ^a
At2g31810	acetolactate synthase small subunit, putative
At3g52930	fructose-bisphosphate aldolase, putative similar
At3g57560	aspartate/glutamate/uridylate kinase family
At4g23890	expressed protein hypothetical protein,
AtCg00840	rpl23.1 ribosomal protein L23
AtCg00900	rps7.1_chloroplast ribosomal protein S7
At1g23310	glutamate:glyoxylate aminotransferase 1 (GGT1)
At1g71500	Rieske [2Fe-2S] domain-containing protein
At2g22230	beta-hydroxyacyl-ACP dehydratase, putative
At2g47730	glutathione S-transferase 6 (GST6) identical to
At3g53460	29 kDa ribonucleoprotein, chloroplast /
At3g53900	uracil phosphoribosyltransferase, putative / UMP
At4g23100	glutamate-cysteine ligase /
At4g33760	tRNA synthetase class II (D, K and N) family
At5g17530	phosphoglucosamine mutase family protein low
At5g17560	BolA-like family protein contains Pfam profile:
At5g19220	glucose-1-phosphate adenylyltransferase large
At5g42650	allene oxide synthase (AOS) / hydroperoxide
At5g43780	sulfate adenylyltransferase 4 / ATP-sulfurylase 4
At5g60600	1-hydroxy-2-methyl-2-(E)-butenyl 4-diphosphate
At1g17650	6-phosphogluconate dehydrogenase NAD-binding
At1g51560	expressed protein
At1g56590	clathrin adaptor complexes medium subunit family
At1g64510	ribosomal protein S6 family protein similar to
At1g65260	PspA/IM30 family protein contains Pfam PF04012:
At1g80480	PRLI-interacting factor L, putative similar to
At3g04260	SAP domain-containing protein contains Pfam
At3g15020	malate dehydrogenase [NAD], mitochondrial,
At4g25370	Clp amino terminal domain-containing protein
At4g30910	cytosol aminopeptidase family protein contains
At4g33520	metal-transporting P-type ATPase, putative (PAA1)
At5g03880	expressed protein
At5g36700	phosphoglycolate phosphatase, putative similar to
At5g42070	expressed protein similar to unknown protein
At5g45170	CbbY protein-related low similarity to SP—P40119
At5g53460	glutamate synthase [NADH], chloroplast, putative
At5g53850	haloacid dehalogenase-like hydrolase family
At5g55280	cell division protein FtsZ, chloroplast, putative
At5g59750	riboflavin biosynthesis protein, putative similar
AtCg00810	rpl22 ribosomal protein L22
At1g29700	expressed protein
AtCg01110	ndhH NADH dehydrogenase 49KDa protein
At1g01090	pyruvate dehydrogenase E1 component alpha
At1g03630	protochlorophyllide reductase C, chloroplast /
At1g11750	ATP-dependent Clp protease proteolytic subunit
At1g12800	S1 RNA-binding domain-containing protein contains
At1g17220	translation initiation factor IF-2, chloroplast,
At1g18170	immunophilin / FKBP-type peptidyl-prolyl

Continued on next page

Table B.1: Continued

Locus	Annotation ^a
At1g48350	ribosomal protein L18 family protein similar to
At1g56050	GTP-binding protein-related similar to
At1g56070	elongation factor 2, putative / EF-2, putative
At1g58290	glutamyl-tRNA reductase 1 / GluTR (HEMA1)
At1g64770	expressed protein
At1g66520	formyltransferase, putative similar to
At1g67700	expressed protein
At1g68010	glycerate dehydrogenase / NADH-dependent
At1g74640	expressed protein contains Pfam profile: PF00561
At1g76100	plastocyanin identical to plastocyanin GI:1865683
At1g77090	thylakoid lumenal 29.8 kDa protein identical to
At1g77590	long-chain-fatty-acid-CoA ligase family protein
At2g24020	expressed protein contains Pfam domain PF02575:
At2g30170	expressed protein
At2g33150	acetyl-CoA C-acyltransferase, putative /
At2g33450	50S ribosomal protein L28, chloroplast (CL28)
At2g40300	ferritin, putative similar to ferritin subunit
At2g41680	thioredoxin reductase, putative / NADPH-dependent
At2g43945	expressed protein
At3g03910	glutamate dehydrogenase, putative similar to
At3g08580	ADP, ATP carrier protein 1, mitochondrial /
At3g14415	(S)-2-hydroxy-acid oxidase, peroxisomal, putative
At3g15520	peptidyl-prolyl cis-trans isomerase TLP38,
At3g17810	dihydroorotate dehydrogenase family protein /
At3g24430	expressed protein contains Pfam profile PF01883:
At3g27890	NADPH-dependent FMN reductase family protein
At3g32930	expressed protein
At3g45140	lipxygenase (LOX2) identical to SP—P38418
At3g46740	chloroplast outer envelope protein, putative
At3g47650	bundle-sheath defective protein 2 family / bsd2
At3g51140	expressed protein
At3g54090	pfkB-type carbohydrate kinase family protein
At3g54900	CAX-interacting protein 1 (CAXIP1) identical to
At3g59760	cysteine synthase, mitochondrial, putative /
At3g63410	chloroplast inner envelope membrane protein,
At4g02510	chloroplast outer membrane protein, putative
At4g12830	hydrolase, alpha/beta fold family protein low
At4g21210	expressed protein contains Pfam domain PF03618:
At4g25100	superoxide dismutase [Fe], chloroplast (SODB) /
At4g32520	glycine hydroxymethyltransferase, putative /
At4g33010	glycine dehydrogenase [decarboxylating], putative
At4g34090	expressed protein
At4g35630	phosphoserine aminotransferase, chloroplast
At5g01590	expressed protein
At5g01650	macrophage migration inhibitory factor family
At5g05000	translocate of chloroplast 34 (TOC34) /
At5g05740	peptidase M50 family protein / sterol-regulatory
At5g08540	expressed protein similar to unknown protein

Continued on next page

Table B.1: Continued

Locus	Annotation^a
At5g14260	SET domain-containing protein low similarity to
At5g14910	heavy-metal-associated domain-containing protein
At5g16390	biotin carboxyl carrier protein 1 (BCCP1)
At5g22800	aminoacyl-tRNA synthetase family protein contains
At5g46110	phosphate/triose-phosphate translocator, putative
At5g48220	indole-3-glycerol phosphate synthase, putative
At5g51070	ATP-dependent Clp protease ATP-binding subunit
At5g51110	expressed protein
At5g52520	tRNA synthetase class II (G, H, P and S) family
AtCg00580	psbE PSII cytochrome b559
AtCg00770	rps8 ribosomal protein S8
At3g05470	formin homology 2 domain-containing protein / FH2
At4g17560	ribosomal protein L19 family protein similar to
At5g11880	diaminopimelate decarboxylase, putative / DAP

^aArabidopsis Genome Initiative (AGI, www.tair.org), annotation.

Appendix C

Top-Hits to Proteobacteria

These data are the BLAST results for the 'Any2' protein data queried against the database of selected bacteria and yeast described in Chapter 2 for which the top hit was to one of the Proteobacteria.

Table C.1: 'Any2' Proteins with Top-Hit to Proteobacteria

Query	Hit	E	Annotation
At1g15140.1	NP_891173.1	4.0e-20	[Bordetella bronchiseptica RB50] CDP-6-deoxy-delta-3,4-glucoseen reductase
At1g17650.1	NP_952425.1	5.0e-77	[Geobacter sulfurreducens PCA] 3-hydroxyisobutyrate dehydrogenase family protein
At1g48860.1	NP_717994.1	5.3e-112	[Shewanella oneidensis MR-1] 3-phosphoshikimate 1-carboxyvinyltransferase
At1g53240.1	NP_417703.1	9.4e-83	[Escherichia coli K12] malate dehydrogenase
At1g62750.1	NP_772043.1	1.6e-245	[Bradyrhizobium japonicum USDA 110] translation elongation factor G
At1g63770.1	NP_951364.1	1.8e-200	[Geobacter sulfurreducens PCA] aminopeptidase N
At1g70410.1	NP_953356.1	1.6e-32	[Geobacter sulfurreducens PCA] carbonic anhydrase
At2g13360.1	NP_772679.1	5.1e-91	[Bradyrhizobium japonicum USDA 110] probable SgaA serine-glyoxylate aminotransferase
At2g14880.1	NP_890745.1	2.1e-14	[Bordetella bronchiseptica RB50] hypothetical protein BB4210
At2g17630.1	NP_415427.1	3.1e-91	[Escherichia coli K12] 3-phosphoserine aminotransferase
At2g19940.1	NP_953916.1	1.6e-80	[Geobacter sulfurreducens PCA] N-acetyl-gamma-glutamyl-phosphate reductase
At2g21330.1	NP_946293.1	7.3e-76	[Rhodopseudomonas palustris CGA009] fructose-bisphosphate aldolase
At2g25080.1	NP_948965.1	2.2e-44	[Rhodopseudomonas palustris CGA009] putative glutathione peroxidase
At2g44650.1	NP_771866.1	2.6e-09	[Bradyrhizobium japonicum USDA 110] heat shock protein
At3g01500.1	NP_953356.1	1.7e-37	[Geobacter sulfurreducens PCA] carbonic anhydrase
At3g04790.1	NP_947326.1	9.1e-46	[Rhodopseudomonas palustris CGA009] ribose 5-phosphate isomerase
At3g15020.1	NP_716401.1	2.5e-82	[Shewanella oneidensis MR-1] malate dehydrogenase
At3g17810.1	NP_416652.3	3.0e-63	[Escherichia coli K12] putative dihydropyrimidine dehydrogenase, FMN-linked
At3g20390.1	NP_223597.1	1.3e-28	[Helicobacter pylori J99] hypothetical protein jhp0879

Continued on next page

Table C.1: Continued

Query	Hit	E	Annotation
At3g27890.1	NP_418169.1	8.4e-27	[Escherichia coli K12] hypothetical protein b3713
At3g47520.1	NP_417703.1	2.0e-82	[Escherichia coli K12] malate dehydrogenase
At3g52930.1	NP_768160.1	1.7e-83	[Bradyrhizobium japonicum USDA 110] fructose biphosphate aldolase
At3g52960.1	NP_887880.1	2.1e-37	[Bordetella bronchiseptica RB50] AhpC/TSA-family protein
At3g54660.1	NP_888964.1	2.3e-118	[Bordetella bronchiseptica RB50] putative glutathione reductase
At3g55800.1	NP_952702.1	3.5e-44	[Geobacter sulfurreducens PCA] fructose-1,6-bisphosphatase
At3g58610.1	NP_772975.1	1.6e-57	[Bradyrhizobium japonicum USDA 110] similar to ketol-acid reductoisomerase
At4g01050.1	NP_220968.1	3.0e-09	[Rickettsia prowazekii str. Madrid E] hypothetical protein RP600
At4g18440.1	NP_415649.1	1.6e-149	[Escherichia coli K12] adenylosuccinate lyase
At4g21210.1	NP_951509.1	1.2e-50	[Geobacter sulfurreducens PCA] hypothetical protein GSU0450
At4g23100.1	NP_946175.1	2.5e-137	[Rhodopseudomonas palustris CGA009] gamma-glutamylcysteine synthetase
At4g31870.1	NP_948965.1	1.7e-44	[Rhodopseudomonas palustris CGA009] putative glutathione peroxidase
At4g31990.1	NP_888815.1	4.5e-92	[Bordetella bronchiseptica RB50] aromatic-amino-acid aminotransferase
At4g33510.1	NP_770418.1	1.7e-147	[Bradyrhizobium japonicum USDA 110] phospho-2-dehydro-3-deoxyheptonate aldolase
At4g34290.1	NP_890745.1	6.9e-16	[Bordetella bronchiseptica RB50] hypothetical protein BB4210
At4g34620.1	NP_221224.1	1.7e-19	[Rickettsia prowazekii str. Madrid E] 30S RIBOSOMAL PROTEIN S16 (rpsP)
At4g38970.1	NP_946293.1	1.5e-77	[Rhodopseudomonas palustris CGA009] fructose-bisphosphate aldolase
At5g08280.1	YP_026260.1	5.6e-62	[Escherichia coli K12] hydroxymethylbilane synthase (porphobilinogen deaminase)
At5g09660.1	NP_417703.1	8.2e-84	[Escherichia coli K12] malate dehydrogenase
At5g14740.1	NP_953356.1	4.4e-37	[Geobacter sulfurreducens PCA] carbonic anhydrase
At5g23060.1	NP_220968.1	4.6e-07	[Rickettsia prowazekii str. Madrid E] hypothetical protein RP600
At5g35170.1	NP_415007.1	9.7e-42	[Escherichia coli K12] adenylate kinase
At5g38530.1	NP_946940.1	4.7e-152	[Rhodopseudomonas palustris CGA009] putative tryptophan synthase beta chain
At5g47840.1	NP_415007.1	1.3e-39	[Escherichia coli K12] adenylate kinase

Appendix D

Top-Hits to Yeast

These data are the BLAST results for the 'Any2' protein data queried against the database of selected bacteria and yeast described in Chapter 2 for which the top hit was to one of the yeast *Sacchromyces cerevisiae*.

Table D.1: 'Any2' Proteins with Top-Hit to Yeast

Query	Hit	E	Annotation
At1g02780.1	YP_087095.1	7.0e-55	Protein component of the large (60S) ribosomal subunit, nearly identical to Rpl19Bp and has similarity to rat L19 ribosomal protein; rpl19a and rpl19b single null mutations result in slow growth, while the double null mutation is lethal; Rpl19bp
At1g03475.1	NP_010329.1	1.4e-79	Coproporphyrinogen III oxidase, an oxygen requiring enzyme that catalyzes the sixth step in the heme biosynthetic pathway; localizes to the mitochondrial inner membrane; transcription is repressed by oxygen and heme (via Rox1p and Hap1p); Hem13p
At1g07920.1	NP_009676.1	7.9e-182	functions in the binding reaction of aminoacyl-tRNA (AA-tRNA) to ribosomes; translational elongation factor EF-1 alpha; Tef2p
At1g11860.1	NP_010302.1	5.3e-73	T subunit of the mitochondrial glycine decarboxylase complex, required for the catabolism of glycine to 5,10-methylene-THF; expression is regulated by levels of levels of 5,10-methylene-THF in the cytoplasm; Gcv1p
At1g13270.1	NP_013345.1	2.0e-66	Methionine aminopeptidase, catalyzes the cotranslational removal of N-terminal methionine from nascent polypeptides; function is partially redundant with that of Map2p; Map1p
At1g20260.2	NP_009685.1	1.0e-181	Vacuolar H ⁺ ATPase regulatory subunit (subunit B) of the catalytic (V1) sector; Vma2p
At1g21270.1	NP_009411.1	4.7e-18	Protein kinase of the Mitotic Exit Network that is localized to the spindle pole bodies at late anaphase; promotes mitotic exit by directly switching on the kinase activity of Dbf2p; Cdc15p
At1g23310.1	NP_010396.1	2.9e-97	putative alanine transaminase (glutamyl pyruvic transaminase); Alt2p
At1g56070.1	NP_010673.1	2.6e-286	Elongation factor 2 (EF-2), also encoded by EFT1; catalyzes ribosomal translocation during protein synthesis; contains diphthamide, the unique posttranslationally modified histidine residue specifically ADP-ribosylated by diphtheria toxin; Eft2p

Continued on next page

Table D.1: Continued

Query	Hit	E	Annotation
At1g66200.1	NP_015360.1	2.2e-106	Glutamine synthetase (GS), synthesizes glutamine from glutamate and ammonia; with Glt1p, forms the secondary pathway for glutamate biosynthesis from ammonia; expression regulated by nitrogen source and by amino acid limitation; Gln1p
At1g77490.1	NP_012992.1	4.3e-39	Cytochrome-c peroxidase; Ccp1p
At1g77590.1	NP_014962.1	1.4e-86	Long chain fatty acyl-CoA synthetase with a preference for C12:0-C16:0 fatty acids; involved in the activation of imported fatty acids; not required for growth on nonfermentable carbon sources; essential for stationary phase; Faa1p
At2g04030.1	NP_013911.1	1.9e-148	constitutively expressed heat shock protein; Hsc82p
At2g21170.1	NP_010335.1	4.6e-67	induced under stress conditions; triosephosphate isomerase; Tpi1p
At2g26140.1	NP_015349.1	1.9e-141	Mitochondrial inner membrane protease of the AAA family, responsible for degradation of unfolded or misfolded mitochondrial gene products; mutation causes an elevated rate of mitochondrial turnover; Yme1p
At2g28190.1	NP_012638.1	1.5e-43	Cu, Zn superoxide dismutase; Sod1p
At2g30170.1	NP_011943.1	3.7e-15	Ptc7p
At2g37220.1	NP_011675.1	3.2e-26	Nucleolar protein that binds nuclear localization sequences, required for pre-rRNA processing and ribosome biogenesis; Nsr1p
At3g02730.1	NP_011725.1	3.1e-13	thioredoxin; Trx2p
At3g08580.1	NP_009642.1	5.0e-116	Mitochondrial inner membrane ADP/ATP translocator, exchanges cytosolic ADP for mitochondrially synthesized ATP; expressed under anaerobic conditions; similar to Pet9p and Aac1p; has roles in maintenance of viability and in respiration; Aac3p
At3g09440.1	NP_013076.1	1.2e-242	member of 70 kDa heat shock protein family; Ssa2p
At3g52150.1	NP_011675.1	2.0e-18	Nucleolar protein that binds nuclear localization sequences, required for pre-rRNA processing and ribosome biogenesis; Nsr1p
At3g52380.1	NP_011675.1	2.2e-24	Nucleolar protein that binds nuclear localization sequences, required for pre-rRNA processing and ribosome biogenesis; Nsr1p
At3g57610.1	NP_014179.1	6.6e-114	Adenylosuccinate synthase, catalyzes the first committed step in the de novo biosynthesis of adenosine; Ade12p
At3g59980.1	NP_011410.1	5.8e-28	Protein that binds tRNA and methionyl- and glutamyl-tRNA synthetases (Mes1p and Ygl245wp), delivering tRNA to them, stimulating catalysis, and ensuring their localization to the cytoplasm; also binds quadruplex nucleic acids; Arc1p
At4g00620.1	NP_011720.1	4.0e-75	Cytoplasmic trifunctional enzyme C1-tetrahydrofolate synthase, involved in single carbon metabolism and required for biosynthesis of purines, thymidylate, methionine, and histidine; Ade3p
At4g02510.1	NP_116671.1	1.5e-11	Hypothetical ORF; Yfr016cp
At4g08390.1	NP_012992.1	6.3e-38	Cytochrome-c peroxidase; Ccp1p
At4g09040.1	NP_011675.1	4.0e-14	Nucleolar protein that binds nuclear localization sequences, required for pre-rRNA processing and ribosome biogenesis; Nsr1p
At4g24770.1	NP_011675.1	3.2e-34	Nucleolar protein that binds nuclear localization sequences, required for pre-rRNA processing and ribosome biogenesis; Nsr1p
At4g32520.1	NP_013159.1	9.8e-136	serine hydroxymethyltransferase; Shm2p
At4g37930.1	NP_013159.1	1.7e-145	serine hydroxymethyltransferase; Shm2p
At5g08670.1	NP_012655.1	8.0e-189	Beta subunit of the F1 sector of mitochondrial F1F0 ATP synthase, which is a large, evolutionarily conserved enzyme complex required for ATP synthesis; Atp2p

Continued on next page

Table D.1: Continued

Query	Hit	E	Annotation
At5g08680.1	NP_012655.1	2.1e-188	Beta subunit of the F1 sector of mitochondrial F1F0 ATP synthase, which is a large, evolutionarily conserved enzyme complex required for ATP synthesis; Atp2p
At5g09650.1	NP_009565.1	3.6e-58	Cytoplasmic inorganic pyrophosphatase (PPase), catalyzes the rapid exchange of oxygens from Pi with water, highly expressed and essential for viability, active-site residues show identity to those from E. coli PPase; Ipp1p
At5g12040.1	NP_013455.1	2.8e-67	Nit protein, one of two proteins in S. cerevisiae with similarity to the Nit domain of NitFhit from fly and worm and to the mouse and human Nit protein which interacts with the Fhit tumor suppressor; nitrilase superfamily member; Nit3p
At5g35100.1	NP_010439.1	9.5e-11	Cytoplasmic peptidyl-prolyl cis-trans isomerase (cyclophilin), catalyzes the cis-trans isomerization of peptide bonds N-terminal to proline residues; binds the drug cyclosporin A; Cpr1p
At5g35630.1	NP_015360.1	1.2e-105	Glutamine synthetase (GS), synthesizes glutamine from glutamate and ammonia; with Glt1p, forms the secondary pathway for glutamate biosynthesis from ammonia; expression regulated by nitrogen source and by amino acid limitation; Gln1p
At5g36700.1	NP_010045.1	5.8e-51	Alkaline phosphatase specific for p-nitrophenyl phosphate, involved in dephosphorylation of histone II-A and casein; Pho13p
At5g46110.1	NP_012342.1	6.6e-13	Hypothetical ORF; Yjl193wp
At5g53460.1	NP_010110.1	0.0e+00	NAD(+)-dependent glutamate synthase (GOGAT), synthesizes glutamate from glutamine and alpha-ketoglutarate; with Gln1p, forms the secondary pathway for glutamate biosynthesis from ammonia; expression regulated by nitrogen source; Glt1p
At5g54770.1	NP_011660.1	3.7e-75	Protein required for thiamine biosynthesis and for mitochondrial genome stability; Thi4p
At5g63890.1	NP_009900.2	1.9e-109	Multifunctional enzyme containing phosphoribosyl-ATP pyrophosphatase, phosphoribosyl-AMP cyclohydrolase, and histidinol dehydrogenase activities; catalyzes the second, third, ninth and tenth steps in histidine biosynthesis; His4p

Appendix E

Synechocystis Proteins

These are the *Synechocystis* PCC6803 proteins identified in the shotgun proteomics study.

Table E.1: Proteomics Results

Locus	Covg. ^a	N ^b	Annotation ^c
Distinct^d			
sll1577	84.3	21	phycocyanin beta subunit cpcB
slr1986	70.2	17	allophycocyanin beta subunit apcB
slr1329	67.9	32	ATP synthase beta subunit atpB
slr1847	67.5	7	hypothetical protein
slr1516	66.3	12	superoxide dismutase sodB
slr0012	65.5	7	ribulose biphosphate carboxylase small subunit rbcS
sll1099	64.4	33	elongation factor Tu tufA
sll1578	63.6	17	phycocyanin alpha subunit cpcA
sll0018	60.2	16	fructose-bisphosphate aldolase, class II fbaA, fda
ssl0707	57.1	7	nitrogen regulatory protein P-II glnB
ssl3093	56.6	3	phycobilisome small rod linker polypeptide cpcD
slr0923	56.2	6	hypothetical protein YCF65 ycf65
slr0623	56.1	8	thioredoxin trxA
ssl3364	54.1	2	CP12 polypeptide cp12
ssr3383	53.7	4	phycobilisome small core linker polypeptide apcC
slr1247	48.9	17	phosphate-binding periplasmic protein precursor (PBP)
slr0335	48.2	43	phycobilisome core-membrane linker polypeptide apcE
sll1342	48.1	16	NAD(P)-dependent glyceraldehyde-3-phosphate dehydrogenase gap2
sll1626	46.8	11	LexA repressor
sll1525	46.1	14	phosphoribulokinase prk
sll1801	45.5	6	50S ribosomal protein L23 rpl23
slr2075	44.7	8	10kD chaperonin groES
sll1810	44.1	6	50S ribosomal protein L6 rpl6
sll1712	44.0	4	DNA binding protein HU
ssr0330	44.0	2	ferredoxin-thioredoxin reductase, variable chain ftrV
slr2051	43.8	8	phycobilisome rod-core linker polypeptide cpcG1
sll0258	43.1	5	cytochrome c550 psbV
slr0585	41.7	14	argininosuccinate synthetase argG
sll1812	41.6	8	30S ribosomal protein S5 rps5

Continued next page

Table E.1: Continued

Locus	Covg.^a	N^b	Annotation^c
ssl2501	41.6	3	unknown protein
sll0427	41.2	10	photosystem II manganese-stabilizing polypeptide psbO
sll0534	40.7	4	ATP-dependent Clp protease proteolytic subunit 2 clpP2
slr0737	40.4	10	photosystem I subunit II psaD
sll1327	39.8	9	ATP synthase gamma chain atpC
ssr1951	39.7	2	hypothetical protein
slr0244	39.4	10	hypothetical protein
slr1198	39.3	9	antioxidant protein
sll1818	38.9	9	RNA polymerase alpha subunit rpoA
slr1756	38.9	15	glutamate-ammonia ligase glnA
ssl0563	38.3	3	photosystem I subunit VII psaC
sll0807	38.3	5	pentose-5-phosphate-3-epimerase rpe
sll1621	38.1	6	AhpC/TSA family protein
sll1580	38.1	16	phycobilisome rod linker polypeptide cpcC1
slr0542	37.9	6	ATP-dependent protease ClpP clpP
ssr2831	37.8	4	photosystem I subunit IV psaE
slr0164	36.4	6	ATP-dependent Clp protease proteolytic subunit clpP4
sll1908	36.3	14	D-3-phosphoglycerate dehydrogenase serA
sll1097	35.9	5	30S ribosomal protein S7 rps7
slr1722	35.9	9	inosine-5'-monophosphate dehydrogenase
slr0009	35.3	16	ribulose biphosphate carboxylase large subunit rbcL
slr0513	35.3	8	iron transport system substrate-binding protein, periplasmic protein
sll0395	34.9	4	phosphoglycerate mutase
sll1785	34.7	6	periplasmic protein, function unknown
sll1744	34.5	7	50S ribosomal protein L1 rpl1
sll1244	34.2	3	50S ribosomal protein L9 rpl9
slr1463	33.4	19	elongation factor EF-G fus
sll1326	33.0	20	ATP synthase alpha chain atpA
ssl3436	32.9	5	50S ribosomal protein L29 rpl29
sll0569	32.5	6	RecA gene product recA
sll0144	32.3	5	uridine monophosphate kinase pyrH
sll0982	32.0	4	unknown protein
sll0851	31.8	13	photosystem II CP43 protein psbC
slr1533	31.3	2	hypothetical protein
sll0145	31.3	4	ribosome releasing factor frr, rrf
sll1821	31.1	4	50S ribosomal protein L13 rpl13
sll0359	30.5	4	hypothetical protein
sll0017	30.0	7	glutamate-1-semialdehyde aminomutase hemL
slr2024	29.4	3	two-component response regulator CheY subfamily
ssr0482	29.3	3	30S ribosomal protein S16 rps16
slr1452	29.0	7	sulfate transport system substrate-binding protein sbpA
sll1106	28.7	4	hypothetical protein
slr1020	28.5	7	sulfolipid biosynthesis protein SqdB sqdB
sll1808	28.5	5	50S ribosomal protein L5 rpl5
sll1743	28.4	4	50S ribosomal protein L11 rpl11
sll0928	28.0	4	allophycocyanin-B apcD
slr1992	27.3	4	glutathione peroxidase-like NADPH peroxidase gpx2

Continued next page

Table E.1: Continued

Locus	Covg.^a	N^b	Annotation^c
sll1638	26.8	3	hypothetical protein
slr1051	26.0	3	enoyl-[acyl-carrier-protein] reductase
ssl2084	26.0	3	acyl carrier protein acpP
sll1363	25.9	6	ketol-acid reductoisomerase ilvC
ssr3451	25.9	3	cytochrome b559 alpha subunit psbE
sll1803	25.6	2	50S ribosomal protein L22 rpl22
slr0394	25.5	10	phosphoglycerate kinase pgk
slr0483	25.5	3	hypothetical protein
slr1437	25.0	1	unknown protein
slr1963	24.9	7	water-soluble carotenoid protein
sll0822	24.8	2	hypothetical protein
sll1815	24.6	4	adenylate kinase adk
sll1745	24.3	3	50S ribosomal protein L10 rpl10
sll0170	24.1	10	DnaK protein 2, heat shock protein 70, molecular chaperone dnaK2
slr1042	24.0	2	two-component response regulator CheY subfamily
sll1633	24.0	6	cell division protein FtsZ ftsZ
slr0172	23.9	2	hypothetical protein
sll0329	23.9	12	6-phosphogluconate dehydrogenase
slr2094	23.8	8	fructose-1,6-/sedoheptulose-1,7-bisphosphatase fbpI
slr0645	23.8	3	hypothetical protein
slr0233	23.8	1	thioredoxin M trxM1
slr1645	23.7	2	photosystem II 11 kD protein psb27, psbZ
slr0600	23.3	4	NADP-thioredoxin reductase
slr0821	23.1	2	hypothetical protein
sll0947	23.0	4	light repressed protein A homolog lrtA
sll1852	22.8	3	nucleoside diphosphate kinase
sml0008	22.5	2	photosystem I subunit IX psaJ
sll1680	22.2	2	hypothetical protein
sll1194	22.1	4	photosystem II 12 kDa extrinsic protein psbU
ssl2598	21.9	2	photosystem II PsbH protein psbH
sll1746	21.9	3	50S ribosomal protein L12 rpl12
sll1721	21.9	4	pyruvate dehydrogenase E1 component, beta subunit
slr1852	21.8	5	unknown protein
slr0947	21.8	4	response regulator for energy transfer from phycobilisomes photo-systems rpaB, ycf27
sll0230	21.8	1	hypothetical protein
sll0408	21.4	5	peptidyl-prolyl cis-trans isomerase
slr0426	21.4	4	GTP cyclohydrolase I folE
slr0506	21.4	7	light-dependent NADPH-protochlorophyllide oxidoreductase por
ssl3437	21.0	3	30S ribosomal protein S17 rps17
sll0872	20.8	4	unknown protein
slr0342	20.7	4	cytochrome b6 petB
slr0536	20.6	6	uroporphyrinogen decarboxylase hemE
slr1273	20.5	2	hypothetical protein
sll0576	20.2	5	putative sugar-nucleotide epimerase/dehydratase
slr0906	20.1	11	photosystem II core light harvesting protein psbB
sll1423	20.0	3	global nitrogen regulator ntcA, ycf28
sll1813	19.7	2	50S ribosomal protein L15 rpl15

Continued next page

Table E.1: Continued

Locus	Covg.^a	N^b	Annotation^c
sll1398	19.6	3	photosystem II reaction center 13 kDa protein psbW
slr1622	19.5	4	soluble inorganic pyrophosphatase ppa
slr0670	18.9	4	hypothetical protein
slr1839	18.8	2	carbon dioxide concentrating mechanism protein CcmK homolog
slr0469	18.8	3	30S ribosomal protein S4 rps4
slr1626	18.6	1	dihydroneopterin aldolase
sll1234	18.6	8	adenosylhomocysteinase
sll0039	18.5	1	positive phototaxis protein, pixH, pisH
sll0617	18.4	3	plasma membrane protein essential for thylakoid formation vipp1
slr0013	18.3	2	hypothetical protein
sll0819	18.2	4	photosystem I reaction center subunit III precursor (psaF)
sll1873	18.2	2	unknown protein
sll1816	18.1	2	30S ribosomal protein S13 rps13
slr1719	17.9	3	DrgA protein homolog
slr0952	17.9	5	fructose-1,6-bisphosphatase flpII
sll1282	17.7	1	riboflavin synthase beta subunit ribH
sll0320	17.5	3	probable ribonuclease D
slr1855	17.5	7	unknown protein
sll1323	17.5	2	ATP synthase subunit b' of CF(0) atpG
sll1317	17.4	5	apocytochrome f, component of cytochrome b6/f complex petA
slr1600	17.4	1	hypothetical protein
sll0368	17.4	2	uracil phosphoribosyltransferase
slr0447	17.3	5	periplasmic protein, ABC-type urea transport system, urtA
slr1324	17.3	3	ATP synthase B chain (subunit I) of CF(0) atpF
ssl2667	17.1	2	assembly factor for iron-sulfur clusters cnfU
slr1030	17.1	3	magnesium protoporphyrin IX chelatase subunit I chlI
ssr2998	16.9	1	hypothetical protein
slr1761	16.9	2	FKBP-type peptidyl-prolyl cis-trans isomerase, periplasmic
sll0752	16.7	2	hypothetical protein
sll1931	16.6	4	serine hydroxymethyltransferase glyA
slr1128	16.5	4	hypothetical protein
slr0376	16.4	1	hypothetical protein
slr1834	16.4	14	P700 apoprotein subunit Ia psaA
sll0902	16.2	6	ornithine carbamoyltransferase argF
slr0194	16.2	3	ribose 5-phosphate isomerase rpiA
slr1793	16.1	6	transaldolase
sll1261	16.1	5	elongation factor TS tsf
slr0924	16.1	2	periplasmic protein, function unknown
slr0333	16.0	2	unknown protein
sll0927	15.9	5	S-adenosylmethionine synthetase
sll1787	15.7	9	RNA polymerase beta subunit rpoB
sll0767	15.4	2	50S ribosomal protein L20 rpl20
slr1251	15.3	2	peptidyl-prolyl cis-trans isomerase
sll5080	15.3	3	78057 - 78881 complement non-heme chloroperoxidase
sll1325	15.1	3	ATP synthase delta chain of CF(1) atpD
slr1859	14.8	1	anti-sigma f factor antagonist
sll1987	14.5	7	catalase peroxidase cpx, katG
slr0879	14.4	1	glycine decarboxylase complex H-protein

Continued next page

Table E.1: Continued

Locus	Covg.^a	N^b	Annotation^c
slr0941	14.3	1	hypothetical protein
sll1767	14.2	1	30S ribosomal protein S6 rps6
sll1663	14.1	3	phycocyanin alpha phycocyanobilin lyase related protein
ssl0242	14.1	1	hypothetical protein
ssr1399	14.1	1	30S ribosomal protein S18 rps18
sll0053	14.1	5	biotin carboxylase accC
slr1223	14.0	3	hypothetical protein
sll0180	13.8	5	hypothetical protein
slr1843	13.8	6	glucose 6-phosphate dehydrogenase zwf
sll0533	13.8	4	trigger factor
sll1031	13.7	8	carbon dioxide concentrating mechanism protein CcmM, putative
ssl2296	13.5	1	pterin-4a-carbinolamine
sll1182	13.5	2	cytochrome b6-f complex, petC3
sll1735	13.5	2	hypothetical protein
sll0470	13.4	2	hypothetical protein
slr0434	13.4	3	elongation factor P efp
sll1101	13.3	1	30S ribosomal protein S10 rps10
sll1322	13.3	2	ATP synthase A chain of CF(0) atpI
sll1316	13.3	1	cytochrome b6-f complex iron-sulfur subunit (Rieske iron
slr1234	13.2	1	protein kinase C inhibitor
sll1806	13.1	1	50S ribosomal protein L14 rpl14
slr0963	13.1	6	ferredoxin-sulfite reductase sir
slr1842	13.1	2	cysteine synthase cysK
slr1338	13.0	2	hypothetical protein
slr0752	13.0	4	enolase
sll1043	13.0	6	polyribonucleotide nucleotidyltransferase
slr0559	13.0	3	periplasmic binding protein of ABC transporter, natB
slr2136	12.9	4	GcpE protein homolog
slr0149	12.8	1	hypothetical protein
slr0609	12.8	3	hypothetical protein
slr1517	12.7	3	3-isopropylmalate dehydrogenase leuB
sll1415	12.7	2	hypothetical protein
sll0006	12.7	2	putative aminotransferase
sll1070	12.7	5	transketolase
sll1260	12.6	3	30S ribosomal protein S2 rps2
slr0165	12.4	2	ATP-dependent Clp protease proteolytic subunit clpP3
sll0248	12.4	2	flavodoxin isiB
slr1853	12.4	1	carboxymuconolactone decarboxylase
sll0057	12.4	2	heat shock protein GrpE grpE
slr0925	12.4	2	single-stranded DNA-binding protein ssb
slr1780	12.3	1	hypothetical protein YCF54 ycf54
sll1799	12.2	2	50S ribosomal protein L3 rpl3
slr0161	12.2	4	twitching motility protein PilT pilT1
ssr1600	12.1	1	similar to anti-sigma f factor antagonist
sll1579	12.1	4	phycobilisome rod linker polypeptide cpcC2
sll0080	12.0	3	N-acetyl-gamma-glutamyl-phosphate reductase argC
slr1356	11.9	3	30S ribosomal protein S1 rps1a
slr1046	11.9	1	putative TatA protein

Continued next page

Table E.1: Continued

Locus	Covg.^a	N^b	Annotation^c
sll0682	11.8	2	phosphate transport system permease protein PstA homolog
sll0172	11.8	1	periplasmic protein, function unknown
slr1133	11.7	3	L-argininosuccinate lyase argH
slr1655	11.5	2	photosystem I subunit XI psaL
sll1697	11.5	2	hypothetical protein
slr1470	11.5	1	hypothetical protein
slr1934	11.4	3	pyruvate dehydrogenase E1 component, alpha subunit
sll1800	11.4	2	50S ribosomal protein L4 rpl4
slr0073	11.4	2	two-component sensor histidine kinase hik36
slr0193	11.3	2	RNA-binding protein rbp3
ssl2874	11.2	1	hypothetical protein
sll0401	11.2	4	citrate synthase
slr1835	11.1	10	P700 apoprotein subunit Ib psaB
slr0435	11.0	1	biotin carboxyl carrier protein of acetyl-CoA carboxylase accB
sll2002	11.0	2	hypothetical protein
sll1341	10.9	2	bacterioferritin
sll0220	10.9	5	L-glutamine:D-fructose-6-P amidotransferase glmS
sll1789	10.9	9	RNA polymerase beta prime subunit rpoC2
slr1718	10.8	1	hypothetical protein
sll1545	10.7	2	glutathione S-transferase
sll1098	10.6	5	elongation factor EF-G fus
ssr1528	10.6	1	hypothetical protein
slr0370	10.4	3	succinate-semialdehyde dehydrogenase (NADP+)
sll1069	10.3	2	3-oxoacyl-[acyl-carrier-protein] synthase II
slr0708	10.2	2	periplasmic protein, function unknown
sll1559	10.2	3	soluble hydrogenase 42 kD subunit
slr0653	10.1	3	principal RNA polymerase sigma factor SigA sigA,rpoDI
slr0535	10.0	5	protease
slr1881	10.0	1	ATP-binding subunit of the ABC-type Nat permease, natE
sll7089	10.0	1	83905 - 85017 complement unknown protein
sll1077	10.0	3	agmatinase speB2
slr1643	9.9	3	ferredoxin-NADP oxidoreductase petH
slr1590	9.8	1	hypothetical protein
slr0943	9.7	2	fructose-bisphosphate aldolase, class I fda
sll0480	9.7	2	probable aminotransferase
sll1275	9.6	3	pyruvate kinase 2
slr1783	9.6	3	two-component response regulator NarL subfamily ycf29
sll1994	9.5	2	porphobilinogen synthase (5-aminolevulinate dehydratase) hemB
slr0049	9.5	2	hypothetical protein
slr0899	9.4	1	cyanate lyase cynS
slr1455	9.3	3	sulfate transport system ATP-binding protein cysA
sll1823	9.3	2	adenylosuccinate synthetase purA
sll0726	9.2	3	phosphoglucomutase
sll1292	9.1	1	two-component response regulator CheY subfamily
slr1322	9.0	3	putative modulator of DNA gyrase; TldD tldD
sll0019	8.9	3	1-deoxy-d-xylulose 5-phosphate reductoisomerase
slr0649	8.8	4	methionyl-tRNA synthetase metS
sll0998	8.7	2	LysR family transcriptional regulator ycf30

Continued next page

Table E.1: Continued

Locus	Covg.^a	N^b	Annotation^c
slr0629	8.6	3	alternative photosystem I reaction center subunit X psaK2
slr1890	8.5	1	bacterioferritin
slr1550	8.4	3	lysyl-tRNA synthetase lysS
slr0370	8.3	7	carbamoyl-phosphate synthase, pyrimidine-specific, pyrA
slr2034	8.2	2	putative homolog of plant HCF136
sll1023	8.2	1	succinyl-CoA synthetase beta chain
sll0422	8.2	2	asparaginase
sll1294	8.2	5	methyl-accepting chemotaxis protein
sll1654	8.1	1	hypothetical protein
slr1295	8.1	4	iron transport system substrate-binding protein futA1
sll0102	7.8	2	hypothetical protein
slr0721	7.8	2	malic enzyme
sll1450	7.8	4	nitrate/nitrite transport system substrate-binding protein nrtA
sll1109	7.7	1	hypothetical protein
slr1531	7.7	2	signal recognition particle protein ffh
slr0809	7.6	1	dTDP-glucose 4,6-dehydratase rfbB
slr0261	7.6	2	NADH dehydrogenase subunit 7 ndhH
sll0293	7.6	2	unknown protein
slr0543	7.5	2	tryptophan synthase beta subunit trpB
sll0173	7.4	2	virginiamycin B hydrolase, periplasmic protein vgb
slr0743	7.4	2	similar to N utilization substance protein
slr1734	7.4	3	glucose 6-phosphate dehydrogenase assembly protein opcA
slr1720	7.3	3	aspartyl-tRNA synthetase aspS
slr1349	7.3	4	glucose-6-phosphate isomerase
slr1926	7.2	1	hypothetical protein
sll1404	7.1	1	biopolymer transport ExbB protein homolog
slr1176	7.1	2	glucose-1-phosphate adenylyltransferase
slr0058	7.0	1	hypothetical protein
sll1568	6.9	1	fibrillin
sll2010	6.9	2	UDP-N-acetylmuramoylalanine-D-glutamate ligase murD
sll0051	6.9	2	hypothetical protein
slr2005	6.9	1	periplasmic protein, function unknown
sll1390	6.8	1	hypothetical protein
sll1981	6.7	2	acetolactate synthase ilvB
slr1277	6.6	4	pilus assembly protein homologous to general secretion
slr0477	6.5	1	phosphoribosylglycinamide formyltransferase purN
sll0065	6.4	1	acetolactate synthase small subunit ilvN
slr1856	6.4	1	phosphoprotein substrate of icfG gene cluster
slr1942	6.3	2	circadian clock protein KaiC homolog kaiC3
slr1342	6.3	2	hypothetical protein
slr0032	6.2	1	probable branched-chain amino acid aminotransferase
sll0502	6.2	2	arginyl-tRNA-synthetase argS
slr1044	6.2	3	methyl-accepting chemotaxis protein
slr1854	6.1	1	unknown protein
slr1226	6.1	1	phosphoribosyl aminoidazole succinocarboxamide synthetase purC
sll1080	6.0	1	ABC transport system substrate-binding protein
sll1641	6.0	2	glutamate decarboxylase
sll1272	5.9	1	unknown protein

Continued next page

Table E.1: Continued

Locus	Covg.^a	N^b	Annotation^c
slr0942	5.8	1	alcohol dehydrogenase [NADP+]
sll0064	5.8	1	periplasmic protein, putative polar amino acid transport
slr1540	5.8	1	mRNA-binding protein
sll0421	5.8	2	adenylosuccinate lyase purB
sll1056	5.7	2	phosphoribosylformyl glycine synthetase
sll0838	5.6	1	orotidine 5' monophosphate
slr1887	5.6	2	prophobilinogen deaminase (hydroxymethylbilane synthase) hemC
slr0186	5.4	2	2-isopropylmalate synthase leuA
slr1751	5.4	2	periplasmic carboxyl-terminal protease
slr1194	5.4	1	hypothetical protein
sll1435	5.4	2	glutamyl-tRNA(Gln) amidotransferase subunit B
slr0877	5.4	1	glutamyl-tRNA(Gln) amidotransferase subunit A
sll0519	5.4	1	NADH dehydrogenase subunit 1 ndhA
sll1298	5.3	1	putative carboxymethylenebutenolidase
slr0661	5.2	1	pyrroline-5-carboxylate reductase proC
slr1763	5.2	1	probable methyltransferase
sll0588	5.2	1	unknown protein
sll0041	5.1	3	phytochrome-like photoreceptor protein, taxD1
sll1021	5.1	2	hypothetical protein
slr1276	5.1	1	hypothetical protein
sll1830	5.0	2	unknown protein
slr1031	4.7	1	tyrosyl tRNA synthetase tyrS
slr1521	4.7	1	GTP-binding protein
sll0887	4.7	1	putative modulator of DNA gyrase; PmbA homolog
sll1804	4.6	1	30S ribosomal protein S3 rps3
sll0228	4.6	1	arginase speB1
sll0402	4.6	1	aspartate aminotransferase aspC
slr0452	4.5	1	dihydroxyacid dehydratase ilvD
sll1150	4.4	1	hypothetical protein
slr1238	4.4	1	glutathione synthetase gshB
slr0546	4.4	1	indole-3-glycerol phosphate synthase trpC
slr1945	4.3	1	2,3-bisphosphoglycerate-independent phosphoglycerate mutase
slr1022	4.2	1	N-acetylornithine aminotransferase argD
slr1506	4.2	1	hypothetical protein
sll7085	4.2	1	80203 - 81933 complement unknown protein
slr1265	4.2	2	RNA polymerase gamma-subunit rpoC1
sll1688	4.2	1	threonine synthase thrC
slr2088	4.0	1	acetohydroxy acid synthase ilvG
slr0301	4.0	1	phosphoenolpyruvate synthase
slr0638	4.0	1	glycyl-tRNA synthetase alpha chain glyQ
slr1453	3.8	1	sulfate transport system permease protein cyst
sll0247	3.8	1	iron-stress chlorophyll-binding protein, homologous to psbC (CP43) isiA
sll1185	3.8	1	coproporphyrinogen III oxidase, aerobic (oxygen-dependent) hemF
slr0659	3.8	2	oligopeptidase A
slr1254	3.8	1	phytoene dehydrogenase (phytoene desaturase) pds, crtD, crtP
sll1172	3.7	1	threonine synthase thrC
slr0399	3.7	1	chaperon-like protein for quinone binding in photosystem II ycf39

Continued next page

Table E.1: Continued

Locus	Covg.^a	N^b	Annotation^c
slr2144	3.7	1	periplasmic protein, function unknown
sll0430	3.6	1	HtpG, heat shock protein 90, molecular chaperone htpG
sll0404	3.5	1	glycolate oxidase subunit GlcD glcD
slr1334	3.5	1	phosphoglucomutase/phosphomannomutase
sll1499	3.4	4	ferredoxin-dependent glutamate synthase glsF
slr0374	3.4	1	hypothetical protein
slr0331	3.4	1	NADH dehydrogenase subunit 4 ndhD1
sll0593	3.4	1	glucokinase
sll1135	3.3	1	unknown protein
sll0078	3.2	2	threonyl-tRNA synthetase thrS
sll1883	3.1	1	arginine biosynthesis bifunctional protein ArgJ argJ
slr0220	3.0	1	glycyl-tRNA synthetase beta chain glyS
slr0782	3.0	1	putative flavin-containing monoamine oxidase
sll0616	2.9	3	preprotein translocase SecA subunit secA
slr0940	2.9	1	zeta-carotene desaturase crtQ-2
sll1074	2.8	1	leucyl-tRNA synthetase leuS
slr0744	2.8	1	translation initiation factor IF-2 infB
slr0417	2.8	1	DNA gyrase subunit A gyrA
sll0945	2.7	1	glycogen synthase glgA
slr1207	2.7	1	hypothetical protein
sll1393	2.6	1	glycogen (starch) synthase glgA
slr1165	2.6	1	sulfate adenylyltransferase
slr2131	2.5	2	RND multidrug efflux transporter
slr1560	2.5	1	histidyl tRNA synthetase hisS
sll1498	2.4	1	carbamoyl-phosphate synthase small chain
sll2001	2.4	1	leucine aminopeptidase
sll1452	2.4	1	nitrate/nitrite transport system ATP-binding protein nrtC
slr0213	2.2	1	GMP synthetase guaA
slr1312	2.1	1	arginine decarboxylase
slr0738	2.0	1	anthranilate synthetase alpha-subunit trpE
slr0657	2.0	1	aspartate kinase
slr1729	2.0	1	potassium-transporting P-type ATPase B chain kdpB
sll1665	1.9	1	unknown protein
slr2002	1.9	1	cyanophycin synthetase cphA
sll0058	1.9	1	DnaK protein 1, heat shock protein 70, molecular chaperone dnaK1
sll0529	1.8	1	hypothetical protein
sll0772	1.8	1	probable porin; major outer membrane protein
sll1502	1.5	1	NADH-dependent glutamate synthase large subunit gltB
sll0446	1.4	1	unknown protein
sll1561	1.4	1	proline oxidase putA
slr1325	1.4	1	GTP pyrophosphokinase
sll1951	0.7	1	unknown protein
Differentiable^d			
slr2076	64.5	43	60kD chaperonin groEL1
sll1029	63.1	8	carbon dioxide concentrating mechanism protein CcmK ccmK1
sll1028	60.2	7	carbon dioxide concentrating mechanism protein CcmK ccmK2
sll0416	55.1	27	60 kDa chaperonin 2, GroEL2, molecular chaperone groEL-2
slr2067	50.3	13	allophycocyanin alpha subunit apcA

Continued next page

Table E.1: Continued

Locus	Covg.^a	N^b	Annotation^c
sll0020	22.7	13	ATP-dependent Clp protease ATPase subunit
slr1908	22.2	11	probable porin; major outer membrane protein
slr1841	21.6	11	probable porin; major outer membrane protein
slr1459	21.3	4	phycobilisome core component apcF
slr1641	20.6	13	ClpB protein clpB1
sll0683	13.0	4	phosphate transport ATP-binding protein PstB homolog
sll1463	10.8	5	cell division protein FtsH ftsH
slr1272	10.6	2	probable porin; major outer membrane protein
slr0228	7.5	3	cell division protein FtsH ftsH
slr1604	5.5	3	cell division protein FtsH ftsH
slr0042	3.5	2	probable porin; major outer membrane protein
Equivalent^d			
sll0849	23.6	8	photosystem II reaction center D2 protein psbD
Equivalent to: slr0927			
sll1867	15.3	5	photosystem II D1 protein psbA3
Equivalent to: slr1311			
Subset^d			
slr1181	8.6	2	photosystem II D1 protein psbA1
Subset of: sll1867			
sll1550	6.1	4	probable porin; major outer membrane protein
Subset of: slr1841			
sll0684	4.9	2	phosphate transport ATP-binding protein PstB homolog
Subset of: sll0683			
slr0436	2.7	2	carbon dioxide concentrating mechanism protein CcmO
Subset of: sll1028			
slr0156	2.6	2	ClpB protein clpB2
Subset of: sll0020			
sll1271	1.7	1	probable porin; major outer membrane protein
Subset of: sll1550			

^a%Coverage; ^bnumber of unique peptides; ^cCyanobase annotation (<http://www.kazusa.or.jp/cyano/cyano.html>); ^dsee section 4.3 for definition of categories

Vita

David R. McWilliams was born in Middlesboro, KY on February 3rd, 1953 and raised in Knoxville, TN. He attended Davidson College as a National Merit Scholar, graduating in 1975 with dual majors in Pre-Medicine and Chemistry. He later attended graduate school at Duke University where he obtained the Masters Degree in Organic Chemistry with specialization in natural products. For 20 years he worked for the Celanese Acetate Company at their research center in Charlotte, NC in analytical technology development, new product research, and project management. He had attained the level of Staff Chemist when he left to join the Graduate School of Genome Science and Technology at the University of Tennessee-Oak Ridge National Laboratory in 2001. He completed the requirements for the Ph.D. degree in November 2006.