

**Experimental and Computational Analysis of
Chloroplast Transit Peptide
Domain Architecture and Function**

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

Prakitchai Chotewutmontri
May 2013

Copyright © 2013 by Prakitchai Chotewutmontri
All rights reserved.

This dissertation is dedicated to my parents

Rotjana Chotewutmontri, D.D.S.

and

Pravit Chotewutmontri, Pharm.D.

Who provide endless support and encouragement

Acknowledgements

This work could not be completed without all the valuable help that I have received over the years. I would like to acknowledge all those who have assisted me in this project. I am particularly grateful to my advisor, Dr. Barry Bruce, for giving me the opportunity to work in his laboratory, for all the enthusiastic advices and his role as a model scientist. I would also like to thank my doctoral committee, Drs. Albrecht von Arnim, Jerome Baudry, Brad Binder and Andreas Nebenführ for their time and support. Drs. Albrecht von Arnim and Andreas Nebenführ also provided the biolistic device and microscope for my experiments. I thank Dr. John Dunlap for helping me with confocal microscopy. I also like to thank the organizations that funded my research and travels including the National Science Foundation, the Gibson Family Foundation, the Tennessee Solar Conversion and Storage using Outreach, Research and Education program, the Tennessee Plant Research Center, the Science Alliance and the Graduate School of Genome Science and Technology. Finally, I would like to thank the Bruce laboratory members and especially Sarah Wright, Natalie Myers, Michael Vaughn, Dr. Evan Reddick, Ian Campbell, Richard Simmerman and Khoa Nguyen for providing help with experimental design, instrument operation, data analysis, discussion and support.

Abstract

The Majority of chloroplast proteins are nuclear-encoded and utilize an N-terminal transit peptide (TP) to target into chloroplasts via the general import pathway. Bioinformatic and proteomic analyses provide thousands of predicted TPs, which show low sequence similarity. How the common chloroplast translocon components recognize these diverse TPs is not well understood. Previous results support either sequence- or physicochemical-specific recognitions. To further address this question, a reverse sequence approach was utilized such that the reverse TP contains the same amino acid composition as wild-type TP but lack similar sequence motifs. Using both native and reverse TPs of the two well-studied precursors, we explored these two modes of recognition. We found that reverse TPs behaved similar to wild-type TPs during binding but failed to support protein translocation. We further showed the importance of the N-terminal domain of TPs in governing protein translocation into plastids. When the TP N-termini were replaced with unrelated peptides with varying Hsp70 affinities, we showed that a subset of TP N-termini functions as Hsp70-interacting domains. We proposed that these domains interact with the stromal motor Hsp70. We further identified a conserved spacer distance between these N-terminal Hsp70 domains to the translocon receptor Toc34 binding sites called FGLK motifs. Using mutants with varying spacer lengths, we observed that the most efficient translocation occurred only at an optimal spacer length of around 28 to 31 aa. These results led us to propose the bimodal interaction model of TP architecture and function where a TP contains an N-terminal stromal interacting domain that is linked to a Toc interacting domain via an optimal spacer length. This configuration permits a temporal and/or spatial coupling between a "capturing step" by a TOC receptor and a "trapping/pulling" step by a stromal ATP-dependent molecular motor that is required for productive translocation.

Table of Contents

Chapter 1 Literature Review.....	1
1.1 Origin of Plastids.....	1
1.2 Plastid Protein Targeting Routes.....	4
1.3 The General Import Pathway.....	8
1.4 Roles of the Components of the General Import Pathway.....	12
1.4.1 Cytosolic factors.....	12
1.4.2 Outer envelope lipids.....	13
1.4.3 TOC apparatus.....	14
1.4.4 TIC apparatus.....	20
1.4.5 Stromal motors.....	21
1.4.6 Peptidases.....	24
1.4.7 Multiple Paralogs of Translocon Subunits.....	25
1.5 Regulation of Plastid Protein Import.....	26
1.5.1 Expression control.....	26
1.5.2 Precursor-specific import pathways.....	27
1.5.3 Redox regulation.....	29
1.5.4 Phosphorylation regulation.....	30
1.5.5 Regulation by ubiquitin-proteasome system.....	31
1.6 Transit Peptides.....	31
1.7 Interacting Domains in Transit Peptides.....	32
Chapter 2 Materials and Methods.....	36
2.1 Polymerase Chain Reaction.....	36
2.1.1 Amplification of DNA inserts.....	36
2.1.2 Screening of <i>E. coli</i> colonies.....	36
2.2 Construction of Vectors.....	37
2.2.1 <i>E. coli</i> expression vectors.....	37
2.2.2 Plant expression vectors.....	39
2.3 Site-directed Mutagenesis.....	42
2.4 Expression and Purification of Proteins.....	42
2.4.1 Forward and reverse peptides.....	42
2.4.2 Precursor and mature proteins.....	43
2.4.3 Radiolabeled proteins.....	44
2.4.4 Toc34G.....	45
2.5 Plant Growth.....	46
2.5.1 <i>Arabidopsis</i>	46
2.5.2 Pea.....	46
2.5.3 Tobacco.....	46
2.6 Transient Expression of Protein in Plants.....	47
2.6.1 Biolistic transformation.....	47

2.6.2	<i>Agrobacterium</i> -mediated transient transformation	47
2.7	Chloroplast Isolation.....	48
2.8	Chlorophyll Measurement.....	49
2.9	Autoradiography.....	49
2.10	Liquid Scintillation Counting.....	50
2.11	<i>In Vitro</i> Competitive Chloroplast Protein Binding Assay	50
2.12	<i>In Vitro</i> Chloroplast Protein Import Assay	52
2.13	<i>In Vitro</i> Competitive Chloroplast Protein Import Assay	54
2.14	<i>In Vitro</i> Stromal Processing Assay	55
2.15	<i>In Vivo</i> Chloroplast Protein Import Assays	56
2.15.1	Qualitative analysis using fluorescent imaging	56
2.15.2	Qualitative analysis using immunoblotting	57
2.15.3	Quantitative analysis using fluorescent imaging.....	57
2.16	MALDI-TOF Mass Spectrometry	58
2.17	Analytical Ultracentrifugation.....	59
2.18	Immunoblotting.....	60
2.19	Bioinformatic Analysis	61
2.19.1	Condon optimization	61
2.19.2	Similarity analysis of forward and reverse TPs	61
2.19.3	Generation of <i>Arabidopsis</i> TP dataset.....	61
2.19.4	Calculation of percentage of uncharged amino acids	62
2.19.5	Hsp70 binding site prediction	62
2.19.6	FGLK motif prediction.....	64
2.19.7	Clustering of TPs based on Hsp70 binding patterns	65
2.19.8	Comparison of TP datasets	65
2.19.9	Hydrophobicity	66
2.19.10	Sequence logo.....	66
2.19.11	Evaluation of the N-terminal sequence using the position-specific scoring matrices.....	66
2.19.12	Hsp70-FGLK spacer length distribution and amino acid composition	69
2.19.13	Random sequence generator	70
2.19.14	Hsp7-FGLK spacer design	70
2.19.15	Amino acid distribution.....	71
Chapter 3 Differential Recognition of Transit Peptide during Binding and Translocation into Plastids.....		72
3.1	Disclosure.....	72
3.2	Abstract.....	72
3.3	Introduction.....	73
3.4	Results	76
3.4.1	Production of forward and reverse peptides	76
3.4.2	Similarity analysis of forward and reverse peptides.....	79

3.4.3	Bioinformatic analysis of TP domains	81
3.4.4	Analytical ultracentrifugation analysis of Toc34 with the peptides	83
3.4.5	Development of a liquid scintillation-based <i>in vitro</i> chloroplast protein competitive binding assay	86
3.4.6	<i>In vitro</i> competitive binding assay of the peptides	90
3.4.7	<i>In vitro</i> competitive import assay of the peptides	94
3.4.8	<i>In vivo</i> imaging of forward and reverse-peptide fusion proteins	98
3.4.9	Development of a intensity ratio measurement for <i>in vivo</i> import assay	104
3.4.10	Immunoblotting analysis of import and processing of forward and reverse-peptide fusion proteins	106
3.4.11	Role of the transit peptide N-termini in protein import	108
3.5	Discussion	113
3.5.1	Flexible recognition of TPs by Toc34	113
3.5.2	Chaperone interactions with TPs	115
3.5.3	Bimodal model of TP design	118
3.5.4	Spatial requirements for concurrent TP recognition	123
3.6	Conclusions	125
Chapter 4 Role of the Transit Peptide N-terminus in Plastid Protein Import		126
4.1	Abstract	126
4.2	Introduction	126
4.3	Results	129
4.3.1	Bioinformatic analysis of TP datasets	129
4.3.2	Construction of TP N-terminal mutants	136
4.3.3	Bioinformatic analysis of the the Hsp70-interacting and non-interacting peptides	139
4.3.4	Prediction of protein localization using amino acid distributions of the N-terminal sequences	144
4.3.5	<i>In vivo</i> import assays	151
4.3.6	<i>In vitro</i> import assays	159
4.4	Discussion	163
4.5	Conclusions	175
Chapter 5 Role of the Hsp70-FGLK Spacer Length in Plastid Protein Import		176
5.1	Abstract	176
5.2	Introduction	177
5.3	Results	179
5.3.1	Bioinformatic analysis of the Hsp70-FGLK spacer	179
5.3.2	Design of novel Hsp70-FGLK spacers	181

5.3.3	Construction of TP mutants containing different Hsp70-FGLK spacer lengths	186
5.3.4	Analysis of spacer length requirement using the designed spacer sequences.....	190
5.3.5	Analysis of spacer length requirement using the wild-type spacer sequences	197
5.4	Discussion	204
5.5	Conclusions	207
Chapter 6 Conclusions and Future Directions.....		209
6.1	Chloroplast transit peptide domain architecture and function	209
6.1.1	Prior understanding.....	209
6.1.2	This project achievement.....	211
6.1.3	Future directions	215
6.2	Toward designing novel chloroplast transit peptides.....	217
Bibliography.....		219
Appendices		252
	DNA Sequences.....	253
	<i>Arabidopsis</i> TP Datasets	266
	The Position-specific Scoring Matrices	286
	Perl Script Codes	301
Vita.....		354

List of Tables

Table 3-1.	Codon Optimization Indices of the Synthetic Peptides	78
Table 3-2.	Curve Fitting Parameters of prSSU Homologous Binding.....	93
Table 3-3.	Curve Fitting Parameters of Peptide Competitive Binding	96
Table 3-4.	Curve Fitting Parameters of Competitive Import	99
Table 4-1.	Sensitivity and Specificity of the PSSM Classification of Protein Localization.....	152
Table A1-1.	General Primers.....	253
Table A1-2.	Primers for the Construction of pET-30a-based Vectors.....	254
Table A1-3.	Primers for the Construction of pAN187-based Vectors.....	255
Table A1-4.	Oligonucleotides for Cloning of the First 10 Amino Acids of TPs.....	257
Table A1-5.	Oligonucleotides for Cloning of the 14-aa Designed Spacer Mutants.....	258
Table A1-6.	Primers for the Construction of the Designed Spacer Mutant Vectors.....	259
Table A1-7.	Primers for the Construction of the Spacer Mutant Vectors based on the Native Spacers.....	262
Table A1-8.	Codon Optimized Synthetic DNAs.....	265
Table A2-1.	TargetP-predicted <i>Arabidopsis</i> TP Dataset.....	266
Table A2-2.	Results of Hsp70 Binding Site Clustering and TargetP Prediction of the 208-TP Dataset.....	282
Table A3-1.	The Components of the TP PSSM for Analysis of the N-terminal 10 Residues.....	286
Table A3-2.	The Components of the TP PSSM for Analysis of the N-terminal 30 Residues.....	287
Table A3-3.	The Components of the mTP PSSM for Analysis of the N-terminal 30 Residues.....	289
Table A3-4.	The Components of the SP PSSM for Analysis of the N-terminal 30 Residues.....	295

List of Figures

Figure 1-1.	Endosymbiont Gene Transfer and Protein Targeting to the Plastid	3
Figure 1-2.	Plastid Protein Targeting Routes	5
Figure 1-3.	The General Import Pathway	9
Figure 1-4.	Summary of All Experimentally Determined Interacting Domains in TPs.	34
Figure 3-1.	Relative Adaptiveness Plots of the Forward and Reverse Peptides	77
Figure 3-2.	Purification of Forward and Reverse Peptides	80
Figure 3-3.	Analysis of the N-terminal Uncharged Domain in TPs	82
Figure 3-4.	Effect of the Peptides on psToc34G Monomer:dimer Ratio	85
Figure 3-5.	Purification of ³⁵ S-prSSU, prSSU and mSSU, and Time Course Analysis of ³⁵ S-prSSU Binding to the Chloroplasts	87
Figure 3-6.	Binding of ³⁵ S-prSSU After Pre-treatments and Effect of Nucleotides on Binding	89
Figure 3-7.	Effect of DTT and BSA on Binding of ³⁵ S-prSSU	91
Figure 3-8.	Homologous Competitive Binding of prSSU to the Chloroplasts....	92
Figure 3-9.	Competitive Binding of ³⁵ S-prSSU with the Peptides	95
Figure 3-10.	Competitive Import of ³⁵ S-prSSU with the Peptides	97
Figure 3-11.	Targeting of Fluorescent Proteins Directed by Forward and Reverse Peptides into the Chloroplasts of Tobacco and <i>Arabidopsis</i>	100
Figure 3-12.	Targeting of Fluorescent Proteins Directed by Forward and Reverse Peptides into the Plastids of Onions	102
Figure 3-13.	Targeting of Fluorescent Proteins Directed by the Peptides at the C-terminus in Onion Cells	103
Figure 3-14.	Effect of Onion Cultivar on the Targeting of Fluorescent Proteins into the Plastids	105
Figure 3-15.	Immunoblotting Analysis of Import and Processing of Forward and Reverse-peptide Fusion Proteins	107
Figure 3-16.	Plastid Import Efficiency of N-terminal-altered Fusion Proteins	109
Figure 3-17.	Plastid Targeting of N-terminal-altered Fusion Proteins in Onion Cells	110
Figure 3-18.	Time-dependent Targeting of Fusion Proteins into the Plastid of Onion Cells	111
Figure 3-19.	Hsp70 Binding Site Prediction of the Forward and Reverse Peptides	112

Figure 3-20.	Bimodal Model of TP Design	119
Figure 4-1.	Hsp70 Binding Site and TargetP Predictions of the 208-TP Dataset	130
Figure 4-2.	Analysis of the N-terminal Uncharged Domain of the 208-TP Dataset	132
Figure 4-3.	Hsp70 Affinity versus Percentage of Uncharged Amino Acids of the N-terminal domains from the 208-TP Dataset	134
Figure 4-4.	Comparison of TP Datasets.....	135
Figure 4-5.	The N-terminal Peptide Sequences and the Mutant Constructs ..	137
Figure 4-6.	Hsp70 Binding Predictions of the N-terminal Peptide Mutants ..	140
Figure 4-7.	Amino Acid Distribution of the TP N-termini and the Classification of the Peptide Mutant N-termini	142
Figure 4-8.	Amino Acid Distributions of the N-terminal Sequences of Chloroplastic, Mitochondrial and Secretory Pathway Proteins	145
Figure 4-9.	Comparison of the Amino Acid Distributions of the 30 aa N-terminal Sequences of Chloroplastic, Mitochondrial and Secretory Pathway Proteins to the UniProt sequences.....	147
Figure 4-10.	PSSM Score Distribution of the Proteins from Different Localizations	149
Figure 4-11.	Targeting of the N-terminal Mutants in Onion Cells	153
Figure 4-12.	Plastid Targeting Efficiency of the N-terminal Mutants	155
Figure 4-13.	Relationships Between Plastid Targeting Efficiency and the Properties of N-terminal Domains of TPs	157
Figure 4-14.	Targeting of the N-terminal Mutants in <i>Arabidopsis</i> Cells.....	160
Figure 4-15.	Autoradiographs of <i>In Vitro</i> Translation Products of the N-terminal Mutants.....	162
Figure 4-16.	Import Time Course of Precursor Produced in the First Batch	164
Figure 4-17.	Import Time Course of Precursor Produced in the Second Batch.....	166
Figure 4-18.	Maximal <i>In Vitro</i> Import Rates of the N-terminal Mutant Precursors	167
Figure 4-19.	Comparison of Plastid Targeting Efficiencies and Import Rates of the N-terminal Mutants.....	168
Figure 4-20.	Steps in Plastid Protein Import.....	174
Figure 5-1.	Predicted Hsp70-interacting Sites and FGLK Motifs in the 208-TP Dataset	180
Figure 5-2.	The Hsp70-FGLK Spacer Length Distribution.....	182
Figure 5-3.	Amino Acid Distributions of the Hsp70-FGLK Spacers	184
Figure 5-4.	Sequences and the Hsp70 and FGLK Prediction Scores of the Wild-type and Designed Spacers	185
Figure 5-5.	Constructs Based on the Designed Spacers	187
Figure 5-6.	Constructs Based on the Wild-type Spacers.....	188

Figure 5-7.	Targeting of YFP Directed by the Spacer Mutant TPs Based on the Designed Spacers.....	191
Figure 5-8.	Plastid Import Efficiency of the Deletion Constructs of the Designed Spacers	192
Figure 5-9.	Plastid Import Efficiency of the Spacer Mutant TPs Based on the Designed Spacers.....	194
Figure 5-10.	Hsp70 and FGLK Prediction Scores of the Designed Spacer Mutants Containing Additional Predicted Sites.....	195
Figure 5-11.	Plastid Import Efficiency of the Spacer Mutant TPs Based on the Wild-type SSF Spacer Sequences	199
Figure 5-12.	Plastid Import Efficiency of the Spacer Mutant TPs Based on the Wild-type FDF Spacer Sequences.....	200
Figure 5-13.	FGLK Prediction Scores of the SSF Mutants Containing Additional FGLK Motifs	201
Figure 5-14.	RPPD Prediction Scores of the Wild-type Spacer Mutants	202
Figure 6-1.	Transit Peptide Domain Architecture Model and Function	214

List of Program Codes

Code A4-1.	Percentage of Uncharged Amino Acids Calculator	301
Code A4-2.	Hsp70 Binding Site Prediction based on Random Peptide-display Phage Library Derived Algorithm.....	305
Code A4-3.	Hsp70 Binding Site Prediction based on Cellulose-bound Peptide Library Derived Algorithm.....	307
Code A4-4.	FGLK Motif Prediction	310
Code A4-5.	TP PSSM Calculator Using the N-terminal 10 Residues.....	313
Code A4-6.	PSSM Calculator Using the N-terminal 30 Residues.....	316
Code A4-7.	FGLK Peak Finder	344
Code A4-8.	Hsp70-FGLK Distance Calculator	345
Code A4-9.	Random Sequence Generator.....	347
Code A4-10.	Mutant TP Generator Using Random Spacer Sequences	350
Code A4-11.	Amino Acid Distribution Calculator.....	352

List of Abbreviations

aa	Amino acid(s)
AUC	Analytical ultracentrifugation
BME	β -mercaptoethanol
BSA	Bovine serum albumin
CAI	Codon adaptive index
CFP	Cyan fluorescent protein
d35S	Double 35S promoter
DGDG	Digalactosyldiacylglycerol
ER	Endoplasmic reticulum
FDF	TP of ferredoxin in forward direction
FDR	TP of ferredoxin in reverse direction
FDtp	TP of ferredoxin
FNRtp	TP of ferredoxin-NADPH reductase
GAP	GTPase activating protein
GB	Grinding buffer
GEF	Guanine nucleotide exchange factor
GFP	Green fluorescent protein
IB	Import buffer
IC ₅₀	Inhibitor concentration at half maximum
IDP	Intrinsically disordered protein
IMS	The chloroplast intermembrane space
K_d	Equilibrium dissociation constant
K_i	Equilibrium dissociation constant of competitors
MALDI-TOF MS	Matrix-assisted laser-desorption ionization–time of flight mass spectrometry
MGD1	Monogalactosyldiacylglycerol synthase 1
MGDG	Monogalactosyldiacylglycerol
mSSU	Mature domain of prSSU
mTP	Mitochondrial targeting peptide
NMR	Nuclear magnetic resonance
OEP14	Outer envelope protein of 14 kDa
PC	Phosphatidylcholine
PCR	Polymerase chain reaction
PG	Phosphatidylglycerol
PI	Phosphatidylinositol
PIRAC	Protein import related anion channel
PMSF	Phenylmethanesulfonyl fluoride
PSSM	Position-specific scoring matrix
PreP1/PreP2	Presequence peptidase 1/2

prFD	Precursor of ferredoxin
prSSU	Precursor of the small subunit of RuBisCO
RCMLA	Reduced carboxy-methyl lactalbumin
RuBisCO	Ribulose-1,5-bis-phosphate carboxylase/oxygenase
SL	Sulfolipid
SP	Signal peptide of secretory pathway
SSB	SDS sample buffer
SSF	TP of the small subunit of RuBisCO in forward direction
SSR	TP of the small subunit of RuBisCO in reverse direction
TIC	Translocon at the inner envelope of the chloroplasts
TOC	Translocon at the outer envelope of the chloroplasts
TP	Transit peptide
YFP	Yellow fluorescent protein

Chapter 1

Literature Review

1.1 Origin of Plastids

Over a century ago, Mereschkowsky proposed that plastids derived from cyanobacteria (Martin and Kowallik, 1999). It is now widely accepted based on phylogenetic analysis that the primary plastids originated from endosymbiosis of a cyanobacterial ancestor (Baum, 2013; Criscuolo and Gribaldo, 2011; Douglas, 1998; McFadden and van Dooren, 2004). Fossil record and molecular clock analysis date the endosymbiosis event over 1.2 billion years ago (Butterfield, 2000; Yoon et al., 2004). The comparisons of genome organization and sequences of the primary plastids from Archaeplastida which includes Viridiplantae (green algae and plants), rhodophytes, and glaucophytes suggest a single endosymbiosis event (Keeling, 2004; Keeling, 2010; Martin et al., 1998; Price et al., 2012; Stoebe and Kowallik, 1999; Turner et al., 1999). Despite the advantage of photosynthetic capability, the only other endosymbiosis event resulting in primary plastids was reported in a protist *Paulinella* (Marin et al., 2005; Nowack et al., 2008; Reyes-Prieto et al., 2010; Theissen and Martin, 2006). To explain this incredible rare occurrence of primary plastid endosymbiosis, Ball et al. (2013) showed that the key enzymes required for the host to utilize photosynthetic carbon were derived from another bacterium, the pathogenic Chlamydiales, and proposed that the endosymbiosis occurs only in the infected host where the carbon utilization becomes beneficial. Many attempts pinpointing the plastid ancestor suggest unicellular nitrogen-fixing cyanobacteria based on 16S RNA, photosynthetic and metabolic gene sequences (Criscuolo and Gribaldo, 2011; Falcon et al., 2010; Gupta, 2009; Hackenberg et al., 2011; Kern et al., 2011; Pascual et al., 2011).

The most comprehensive study using a large-scale comparison of 241 complete genomes including 9 cyanobacteria and 4 photosynthetic eukaryotes indicates that the nitrogen-fixing heterocyst cyanobacteria share the largest number of genes with the photosynthetic eukaryotes (Deusch et al., 2008). Although modern heterocyst cyanobacteria, *Nostoc* sp. PCC7120 and *Anabaena variabilis* ATCC29143 harbor around 5,500 protein-encoding genes (Kaneko et al., 2001; Markowitz et al., 2012), plastid genomes of land plants only encode around 80 proteins (Timmis et al., 2004). This reduction of plastid genomes was found to occur mainly by transferring plastid DNA to the nuclear genome (Kleine et al., 2009). It was proposed that the ability of plastids to import nuclear-encoded proteins enabled the plastid ancestry to transfer its genes to the nucleus without compromising its metabolic capacity (Allen, 2003). During the endosymbiosis process, the import of proteins encoded by endosymbiont-to-nucleus genes permits the endosymbiont gene copies to undergo pseudogenization and later loss (Martin et al., 1993). In *Arabidopsis*, bioinformatics analysis estimated that around 4,500 protein-encoding genes in the nucleus are originated from the endosymbiont as shown in Figure 1-1 (Martin et al., 2002). Less than half of these endosymbiont-derived proteins (about 1,800) have targeted to plastids and the rest functions elsewhere (Martin et al., 2002). Interestingly, non-endosymbiont derived proteins (about 750) also localize to plastids (Martin et al., 2002) and function in photosynthesis, respiration and metabolic pathways (Kleine et al., 2009). While plastid protein import was crucial in establishing endosymbiosis in the past, it now is an indispensable function of the cells where over 2,500 proteins in *Arabidopsis* require this process to gain access to plastids.

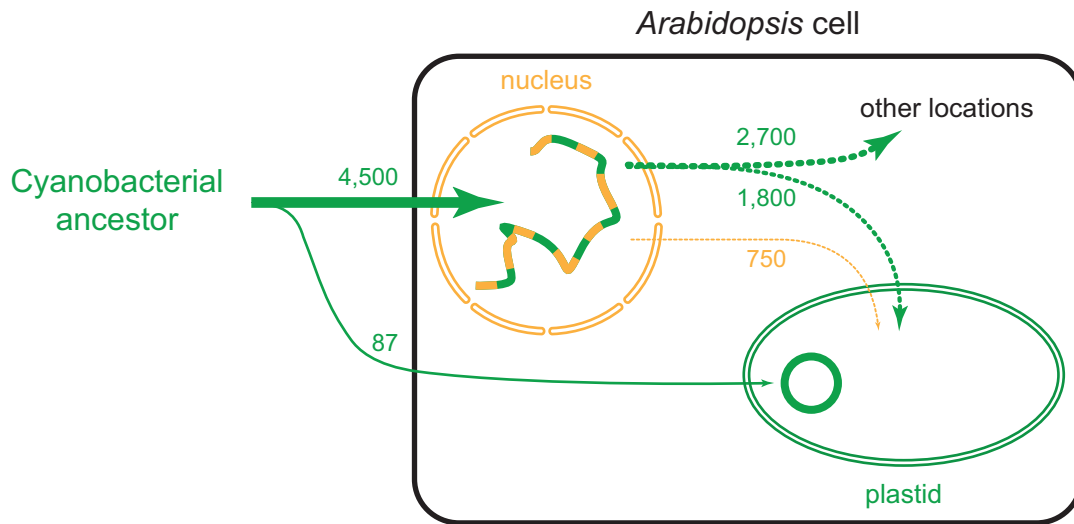


Figure 1-1. Endosymbiont Gene Transfer and Protein Targeting to the Plastid

Bioinformatic analysis of *Arabidopsis* genome sequence indicated that around 4,500 nuclear genes were transferred from the cyanobacterial ancestor and only 87 genes remained in the plastid genome. Although the majority of the proteins encoded by the endosymbiont-to-nucleus genes are targeted to other locations in the cell, about 1,800 of these proteins are targeted to the plastid. In addition, around 750 proteins encoded by non-endosymbiont nuclear genes are also targeted to the plastid.

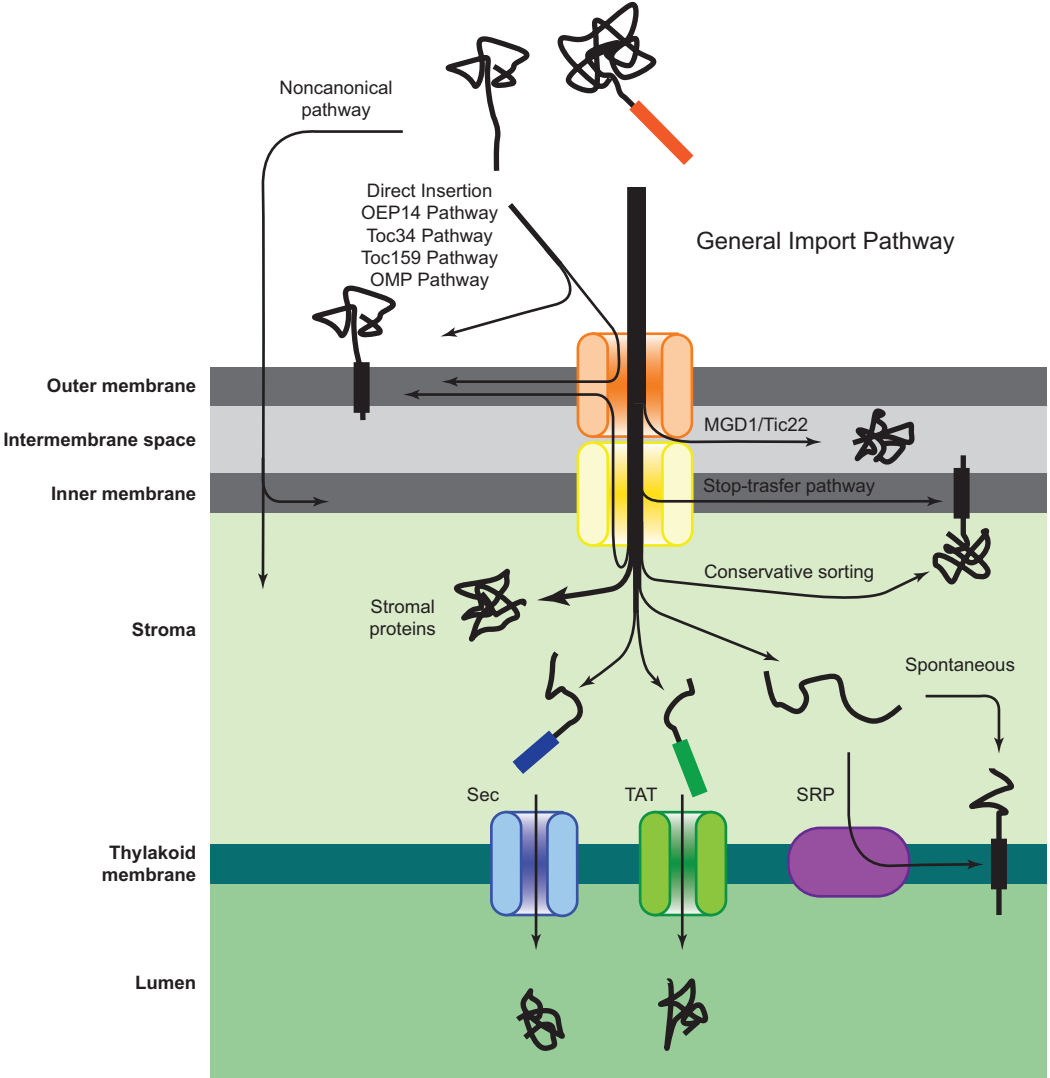
1.2 Plastid Protein Targeting Routes

Proteins targeted to plastids can be delivered post-translationally to six plastid locations including the outer envelope membrane, the intermembrane space (IMS), the inner envelope membrane, the stroma, the thylakoid membrane and the thylakoid lumen as shown in Figure 1-2 (Keegstra and Cline, 1999; Li and Chiu, 2010). Most of these processes involve protein translocation across the plastid membranes: the outer envelope, the inner envelope and the thylakoid membranes. In general, protein translocation across or insertion into membrane is mediated by oligomeric membrane complexes termed translocons (Walter and Lingappa, 1986). The majority of plastid-targeted proteins are synthesized as precursors containing the N-terminal targeting sequences called transit peptides (TPs) (Dobberstein et al., 1977; Kleffmann et al., 2004). Similar to the signal hypothesis (Blobel, 1980), TP acts as an intrinsic signal on the precursor protein that is recognized by the targeting receptors associated to the translocons. TPs direct translocation of precursor proteins across the double membranes of plastids via the translocon at the outer envelope of chloroplasts (TOC) and the translocon at the inner envelope of chloroplasts (TIC) in a process described as the general import pathway (Cline and Henry, 1996; Schnell et al., 1997; van 't Hof and de Kruijff, 1995a). After translocation into the stroma, TP is readily cleaved allowing the mature domain to fold into its native conformation or to be further targeted to the thylakoid (Richter and Lamppa, 1998).

At least six pathways mediate protein targeting to the outer envelope of plastids (Hofmann and Theg, 2005a; Jarvis, 2008). The most studied pathway requires an N-terminal transmembrane domain targeting signal, which is found in proteins such as the outer envelope protein of 14 kDa (OEP14) and the TOC subunit of 64 kDa (Toc64) (Hofmann and Theg, 2005b; Lee et al., 2001). Another pathway involves a C-terminal transmembrane domain targeting sequence such as those present in Toc34 and Toc159 (Smith et al., 2002; Tsai et al., 1999).

Figure 1-2. Plastid Protein Targeting Routes

Proteins targeted to plastids are delivered to six locations: outer envelope membrane, intermembrane space, inner envelope membrane, stroma, thylakoid membrane and thylakoid lumen. The outer envelope proteins use multiple pathways for targeting while the interior proteins containing TP pass through the envelope(s) via the General Import Pathway. Some of the proteins contain a second targeting signal for thylakoid lumen targeting through the secretory (Sec) or the twin-arginine translocase (TAT) pathways. The thylakoid membrane proteins utilize the signal recognition particle-dependent pathway (SRP) or the spontaneous pathway. In addition, experimental evidence and proteomic analysis reveal that many proteins utilize noncanonical pathway.



Toc75 utilizes a bipartite targeting sequence composed of TP and envelope targeting signal as an alternative approach (Inoue and Keegstra, 2003; Tranel and Keegstra, 1996). The insertions of OEP14, Toc159 and Toc75 require some components of the general import pathway. Additional outer envelope targeting pathways have been shown but are not well characterized (Hofmann and Theg, 2005a).

Targeting of proteins into the IMS is not well understood. Only Tic22 and monogalactosyldiacylglycerol synthase 1 (MGD1) have been studied (Kouranov et al., 1999; Vojta et al., 2007). Both of these proteins contain TP, but TP of Tic22 is not cleaved during targeting unlike all other TPs. Whereas MGD1 utilizes the general import pathway, Tic22 targeting is unclear.

The inner membrane proteins are targeted using two pathways comparable to those of the mitochondria (Jarvis, 2008). All of the studied inner membrane proteins contain TP and utilize the general import pathway. The first pathway utilized by the phosphate translocator requires the stop-transfer signal, which allows lateral exit of the protein through the TIC complex (Knight and Gray, 1995). Another pathway, the conservative sorting, employed by Tic40 and Tic110 involves the complete translocation of the proteins into the stroma before re-insertion into the inner membrane (Li and Schnell, 2006; Tripp et al., 2007).

The thylakoid luminal proteins utilize a bipartite targeting sequence containing a TP followed by a second targeting sequence for translocation through the secretory (Henry et al., 1994; Knott and Robinson, 1994; Nakai et al., 1994; Yuan et al., 1994) or the twin-arginine translocase (Bogsch et al., 1997; Chaddock et al., 1995; Henry et al., 1997) pathways. The thylakoid membrane proteins utilize TP to translocate into the stroma and insert into the membrane via the signal recognition particle-dependent pathway (Li et al., 1995; Payan and Cline, 1991; Schuenemann et al., 1998) or the spontaneous insertion pathway (Kim et al., 1998; Lorkovic et al., 1995; Michl et al., 1994; Thompson et al., 1999).

So far, TP has been utilized in targeting precursor proteins to all of the six locations of plastids and are considered to utilize the general import pathway. However, in recent years, many noncanonical targeting pathways have been reported. The targeting of TP-less quinone oxidoreductase homolog to the inner membrane was shown to require an internal targeting signal (Miras et al., 2007) whereas TP-less Tic32 utilizes a novel N-terminal signal (Nada and Soll, 2004). Two proteins containing signal peptide for ER targeting, carbonic anhydrase 1 (Villarejo et al., 2005) and nucleotide pyrophosphatase/ phosphodiesterase 1 (Nanjo et al., 2006), were shown to localize to plastids. The chloroplast proteomic data suggests that around 20% of chloroplast proteins lack any predictable targeting sequence (Kleffmann et al., 2004) and around 1-8% of chloroplast targeted precursors contain an ER signal peptide (Kleffmann et al., 2004; Zybailov et al., 2008). Thus, a small fraction of protein lacking TP is able to target to the plastids.

In summary, the general import pathway functions similar to a central transit hub where the majority of plastid-targeted proteins pass through before reaching their final locations. This pathway recognizes diverse TP sequences from various functional groups of proteins. Nevertheless, other pathways function in delivery TP-less to the plastids.

1.3 The General Import Pathway

The general import pathway is a working model describing a TP-directed translocation of proteins into the stroma of plastids (Bruce, 2000). Figure 1-3 illustrates the sequential steps in this pathway.

The outer envelope of plastids has been shown to recruit cytosolic precursors using multiple pathways. Work performed using isolated chloroplasts shows that the precursors are able to directly bind to TOC on the surface of chloroplasts (1a). The binding is reversible when lacking ATP/GTP and denoted

Figure 1-3. The General Import Pathway

Cytosolic precursors can be targeted to plastids directly by interacting with TOC (1a), or by interacting with the lipids (1e) before transferring to membrane Toc159 (2f) and reaching Toc34 (3a). The precursors can also interact with cytosolic Toc159 (1b) before transferring to Toc34 (2a). TPs can interact with Hsp90 (1d) before being targeted to Toc64 (2d) and further transferred to Toc34 (3b). Phosphorylated TPs can interact with the guidance complex (1c) composed of 14-3-3 and Hsp70c in cytosol. The precursors from guidance complexes can be transferred to membrane receptors Toc159 (2c) or Toc34 (2b). The binding state of precursor to the chloroplast can be subdivided based on GTP/ATP level and temperature of the system (4-5). When the ATP level is greater than 1 mM, the translocation process is initiated by Hsp93/cpHsp70 (6). When the precursors emerge into the stroma, stromal processing peptidase (SPP) will cleave TP from precursor proteins releasing the mature domain (7). TP will be further degraded by PreP1/PreP2 peptidases (8).

as energy-independent binding (Jarvis, 2008; Kouranov and Schnell, 1997; Perry and Keegstra, 1994). Addition of GTP also promotes binding (Inoue and Akita, 2008a; Young et al., 1999). TP interaction with chloroplast lipids suggests that the precursor can directly bind to the lipid surface (1e) before being transferred to the membrane receptors Toc159 and Toc34 (2f, 3a) (Pilon et al., 1995; Pinnaduwege and Bruce, 1996). Cytosolic factors were shown to interact and recruit precursors to the membrane receptors. Cytosolic Hsp90 captures precursor (1d) and delivers it to the membrane receptor Toc64 (2d) (Qbadou et al., 2006). Phosphorylated TPs interact with 14-3-3 and form the guidance complex with cytosolic Hsp70 (1c) (May and Soll, 2000). The guidance complex was proposed to dock to membrane receptors Toc159 (2c, 3a) or Toc34 (2b) (May and Soll, 2000; Qbadou et al., 2006). Another pathway utilizes the cytosolic Toc159 pool to deliver precursor to the membrane (1b, 2a) (Hiltbrunner et al., 2001; Smith et al., 2004).

When low levels of ATP (<0.1 mM) are present with or without GTP, the precursor engages in an irreversible binding state which are referred to as the early import intermediate (Inoue and Akita, 2008a; Kessler et al., 1994; Kouranov and Schnell, 1997; Olsen and Keegstra, 1992; Perry and Keegstra, 1994; Young et al., 1999). In the presence of 0.1 mM ATP at 4 °C, about 110 aa are buried in the translocons (4) (Akita and Inoue, 2009; Inoue and Akita, 2008a). When the temperature is increased to 25 °C, about 130 aa are buried (5) (Akita and Inoue, 2009; Inoue and Akita, 2008a). At this step, the precursor spans the double membrane via the Toc75 and Tic20/21 channels and TP interacts with Tic110 (Akita et al., 1997; Chen and Li, 2007; Inoue and Akita, 2008a; Nielsen et al., 1997) forming the contact site between the double membranes (Schnell and Blobel, 1993). The translocation process is initiated by the ATPase chaperones Hsp93/cpHsp70 when the ATP level is above 1 mM (6) (Shi and Theg, 2010; Su and Li, 2010). When the precursor emerges into the stroma, stromal processing peptidase (SPP) will cleave TP from the precursor,

releasing the mature domain (7) (Richter and Lamppa, 1998). TP will be further degraded by presequence peptidases, PreP1/PreP2 (8) (Glaser et al., 2006).

1.4 Roles of the Components of the General Import Pathway

1.4.1 Cytosolic factors

The precursors were proposed to maintain import-competency by interacting with cytosolic chaperones (Jackson-Constan et al., 2001). In fact, multiple Hsp70 binding sites have been shown in the majority of TPs (Ivey et al., 2000; Rial et al., 2000; Zhang and Glaser, 2002). While cytosolic Hsp70 is essential for the *in vitro* import of a membrane protein, the precursor of the light harvesting chlorophyll a/b-binding protein (Waegemann et al., 1990), it is not necessary for the *in vitro* import of precursors of soluble proteins such as ferredoxin (prFD) (Pilon et al., 1990) and the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase (prSSU) (Dabney-Smith et al., 1999).

Another biochemically identified cytosolic factor, the guidance complex composed of 14-3-3 and Hsp70 proteins, improves the efficiency of import (May and Soll, 2000). 14-3-3 recognizes phosphorylated TPs (May and Soll, 2000). This guidance complex was proposed to deliver the precursor to membrane receptors Toc64, Toc159 or Toc34 (May and Soll, 2000). Later experiments showed that the complex associates with Toc34 but not with Toc64 (Qbadou et al., 2006).

Some precursors associate with Hsp90. This chaperone delivers precursors to Toc64 through the interaction between tetratricopeptide repeats of Toc64 and Hsp90 (Qbadou et al., 2006). The precursor is subsequently transferred to Toc34 (Qbadou et al., 2006).

In addition, the primary TP receptor Toc159 (Ma et al., 1996; Perry and Keegstra, 1994) has been shown to distribute both in soluble cytosolic and

membrane-bound forms (Hiltbrunner et al., 2001) and it was proposed that cytosolic Toc159 interacts with precursor proteins and shuttle them to the translocon (see section 1.4.3.3)

Although cytosolic factors were shown to associate with some precursors and in certain cases increase import efficiencies, these factors were found to be non-essential both *in vitro* (Dabney-Smith et al., 1999; Pilon et al., 1990) and *in vivo* (Aronsson et al., 2007; Hofmann and Theg, 2005c; Nakrieko et al., 2004). The rate of *in vitro* protein import without cytosolic factors was also suggested to be sufficient to sustain chloroplast development (Pilon et al., 1992b).

1.4.2 Outer envelope lipids

The composition of lipids making up plastid envelope membranes resembles that of cyanobacteria (Joyard et al., 1991). These membranes contain high levels of glycerolipids, monogalactosyldiacylglycerol (MGDG), digalactosyldiacylglycerol (DGDG) and sulfolipid (SL), and low levels of phospholipids, phosphatidylcholine (PC), phosphatidylglycerol (PG) and phosphatidylinositol (PI). The composition of the lipids in plastids seems to be maintained over all plastid forms (Joyard et al., 1991). The presence of galactolipids, MGDG and DGDG make the outer membrane of the plastids unique among the cytosolically exposed membranes (Bruce, 1998).

The roles of lipids in the general import pathway have been probed mainly through the interactions with TPs (Pilon et al., 1995; Pinnaduwaage and Bruce, 1996; van 't Hof and de Kruijff, 1995b; van 't Hof et al., 1991; van 't Hof et al., 1993). The TPs interact specifically with MGDG containing membranes (Pilon et al., 1995; Pinnaduwaage and Bruce, 1996), which was proposed to be mediated by hydroxylated aa (Pilon et al., 1995). A study using an *Arabidopsis dgd1* mutant containing less than 10% wild-type DGDG levels showed that mutant chloroplasts import proteins to the stroma slower than wild-type chloroplasts

(Chen and Li, 1998). This import rate correlated with the reduction of the early import intermediate suggests that the increased proportion of MGDG in *dgd1* mutant envelope might trap the precursor at the energy-independent binding state (Chen and Li, 1998). Another study using an *Arabidopsis* MGDG synthase mutant (*mgd1-1*) producing 42% wild-type MGDG levels showed that protein import was not affected, which may be due to the high levels of remaining MGDG (Aronsson, 2008).

How TP-lipid interactions function in import remains unclear. But the ability of TP to reorient MGDG containing bilayers was proposed to be involved (Bruce, 1998; Chupin et al., 1994). Another possibility is based on the observations that TPs only adopt secondary structures upon binding to lipids. It was proposed that the TP-chloroplast membrane interaction triggers the folding of a specific recognition element for protein import (Bruce, 1998).

1.4.3 TOC apparatus

1.4.3.1 Core complex

The TOC core complex is composed of 3 subunits, Toc159, Toc75 and Toc34, and forms a hetero-oligomeric complex ranging from 500 to 1,000 kDa (Chen and Li, 2007; Kikuchi et al., 2006; Schleiff et al., 2003b). The subunits are named based on their predicted molecular weights in kDa (Schnell et al., 1997). The stoichiometry of these subunits is still debated. Ratios of 4:4:1 and 6:6:2 for Toc34:Toc75:Toc159 have been proposed (Kikuchi et al., 2006; Schleiff et al., 2003b). The complex size variability also indicates the possibility of a dynamic complex forming higher order structures (Jarvis, 2008). Both Toc159 and Toc34 contain a GTPase domain (Kessler et al., 1994) while Toc75 is a beta-barrel protein (Hinnah et al., 1997).

Toc86, the proteolytic form of Toc159, was identified as a major interacting partner of prSSU during binding (Ma et al., 1996; Perry and

Keegstra, 1994). Also, antibodies against Toc86 were found to inhibit prSSU binding to chloroplasts (Hirsch et al., 1994). These findings support a role of Toc159 as primary receptor for precursor proteins (Hirsch et al., 1994; Kessler et al., 1994). Yet, chloroplasts lacking Toc159 binding activity are able to import prSSU with lower efficiency (Chen et al., 2000). The intact form of Toc86, Toc159 was discovered only after the *Arabidopsis* genome became available (Bolter et al., 1998; Chen et al., 2000). Toc159 contains 3 domains, the N-terminal acidic or A domain, the central GTPase or G domain, and the C-terminal membrane anchor or M domain (Chen et al., 2000). The A domain is intrinsically disordered (Richardson et al., 2009) and provides selectivity for precursor protein binding (Inoue et al., 2010) which is discussed in section 1.5.2.1. The G domain is required for insertion of Toc159 into the outer envelope (Bauer et al., 2002; Smith et al., 2002; Wallas et al., 2003). The function of the G domain in the import pathway is discussed in section 1.4.3.2. Recently, the M domain was found to function as a reverse TP. Unlike TP, which locates at the N-terminus of the proteins, the M domain locates at the C-terminus of Toc159. It has been shown that the reverse sequence from C- to N-termini of the M domain fused to the N-terminal of GFP functioned as TP. It was able to directed GFP into the chloroplast stroma (Lung and Chuong, 2012).

Toc34 was identified from isolated translocons in complex with a precursor protein (Schnell et al., 1994) and was shown to contain 2 domains, the N-terminal GTPase domain (G domain) and the C-terminal membrane anchor domain (M domain) (Kessler et al., 1994; Seedorf et al., 1995; Sun et al., 2002). The crystal structures of the G domain of Toc34 (Toc34G) are related to those of the small GTPases of the Ras family (Koenig et al., 2008a; Sun et al., 2002). Toc34G also formed dimers in the crystal where the Arg133 was proposed to function similarly to the Arg finger of GTPase activating proteins (GAPs) of Ras GTPases (Kessler and Schnell, 2002). The role of GTPase regulation of protein import is discussed in section 1.4.3.2.

Toc75 was discovered at the same time as Toc86 (Perry and Keegstra, 1994). It was shown to be an integral membrane protein (Schnell et al., 1994) forming a 16 or 18 beta-strand barrel structure (Hinnah et al., 2002; Inoue and Potter, 2004; Schleiff et al., 2003a; Sveshnikova et al., 2000a). Electrophysiological measurements show that Toc75 has an intrinsic ability to specifically recognize TP and has a pore with a diameter at the restriction zone of about 14 Å (Hinnah et al., 2002). The mature Toc75 harbors three repeats of polypeptide transport associated (POTRA) domains followed by the C-terminal beta-barrel domain (Reddick et al., 2008; Sanchez-Pulido et al., 2003). A recent study showed that the N-terminal POTRA domains are localized in the cytosol (Sommer et al., 2011). However, the role of POTRA domains is still unclear.

1.4.3.2 GTPase cycle of TOC receptors

Both Toc34 and Toc159 belong to the paraseptin group of GTPases, which is most related to the septin GTPase group (Weirich et al., 2008). These 2 groups together form the septin family composed of oligomer-forming GTPases (Weirich et al., 2008). GTPases regulate a variety of processes in the cell by cycling between a GTP-bound active form and a GDP-bound inactive form (Gasper et al., 2009). The dimerization of Toc34 has been shown to influence chloroplast protein import (Aronsson et al., 2010; Lee et al., 2009b). A large number of researchers have focused on elucidating the GTPase cycles of TOC receptors to further understand the protein import process, however, some of the results were contradictory.

Homodimerization of TOC receptors are common; atToc33, psToc34 and psToc159 have been shown to dimerize (Koenig et al., 2008a; Koenig et al., 2008b; Oreb et al., 2011; Reddick et al., 2007; Sun et al., 2002; Yeh et al., 2007). Heterodimerizations between atToc33 and atToc159 and between psToc34 and psToc159 were also observed (Bauer et al., 2002; Becker et al., 2004b; Hiltbrunner et al., 2001; Rahim et al., 2009; Smith et al., 2002). Dimerization of

atToc33 was determined to be involved in the transition of precursor proteins from binding to translocation state of import (Lee et al., 2009b). In addition, three factors were found to affect dimerization. (i) Monomer-dimer equilibrium of atToc33 and psToc34 depends on their concentrations with the dissociation constants (K_d) around 400 μ M and 50 μ M, respectively (Koenig et al., 2008a; Oreb et al., 2011; Reddick et al., 2007). Based on the close proximity of Toc34 to each other in the core TOC complex (Schleiff et al., 2003b), Toc34 is likely to favor a dimer form in the complex. (ii) Nucleotide loading state is another factor that affects dimerization. In both atToc33 and psToc34, the GDP-bound forms produce more dimers than the GTP-bound forms (Koenig et al., 2008a). The same effect was also reported in the heterodimerization between atToc159 and atToc33 (Smith et al., 2002). Surprisingly, the addition of a GTP transition state analog, aluminum fluoride, shifted the equilibrium of both atToc33 and psToc34 to become exclusively dimers (Koenig et al., 2008b). These findings indicate that the dimerization is likely triggered by GTP hydrolysis. Furthermore, it can also be speculated that with the presumed active TOC receptor, the monomeric GTP-bound state hydrolyzes GTP to become a dimer, which subsequently traps it in the dimeric GDP-bound inactive state. The dimeric conformation was shown to prevent nucleotide exchange (Oreb et al., 2011) even though TOC receptors have higher affinities to GTP than GDP (Jelic et al., 2003; Reddick et al., 2007). Lastly, (iii) the effect of TP on TOC receptor functions has been the subject of many studies. GTP-bound forms of atToc33 (Gutensohn et al., 2000), psToc34 (Jelic et al., 2002; Schleiff et al., 2002) and psToc159 (Becker et al., 2004b; Kouranov and Schnell, 1997) have higher affinities to TP than the GDP-bound forms. TP is well known to stimulate TOC receptor GTP hydrolysis (Becker et al., 2004b; Jelic et al., 2003; Oreb et al., 2011; Reddick et al., 2007). Two separate roles of TP in GTP hydrolysis have been reported. First, it was shown that TP stimulated GTP hydrolysis of psToc34 while maintaining the same nucleotide exchange rate suggesting a role of TP as a GAP but not guanine

nucleotide exchange factors (GEF) (Reddick et al., 2008; Reddick et al., 2007) since GAP and GEF can both increase GTP hydrolysis but GAP lowers the transition state energy (Scheffzek et al., 1997) while GEF stimulates GDP exchange (Bos et al., 2007). Interestingly, another study found that TP stimulated GTP hydrolysis of atToc33 only when atToc33 was in the dimeric form but not in the monomeric form (Oreb et al., 2011). Further analysis found that TP stimulated atToc33 GTP hydrolysis through interaction with the dimer and increased the nucleotide exchange rate, which suggests a role of TP as a GDP-dissociation inhibitor-displacement factor where each atToc33 in the dimer acts as GDP-dissociation inhibitor and TP disrupts the dimer (Oreb et al., 2011).

Other discrepancies are found in the GTP hydrolysis measurements of TOC receptors. While one laboratory found that regardless of concentration, monomer or homodimer, psToc34, atToc33, atToc34 and atToc159 maintain similar GTP hydrolytic rates (Reddick et al., 2007), another laboratory found that the dimers of psToc159 and atToc33 hydrolyzes GTP faster than their monomers (Yeh et al., 2007). These differences may be due to two reasons. First, GTP hydrolysis and dimerization are diminished over time as can be seen in both atToc33 and psToc34 (Koenig et al., 2008a; Yeh et al., 2007). Second, the method based on the capture of TOC receptor on a surface may change the monomer-dimer equilibrium compared to what is seen in solution.

When the Arg130 of atToc33 and the Arg133 of psToc34, which were proposed to function similarly to the Arg finger of Ras GAP (Kessler and Schnell, 2002) were mutated to Ala, it was clear that these mutants were unable to dimerize (Koenig et al., 2008a; Lee et al., 2009b; Reddick et al., 2007; Yeh et al., 2007). While two reports found that the atToc33 mutant (R130A) was able to hydrolyze GTP at 0.6-1 fold of the wild-type level (Koenig et al., 2008a; Lee et al., 2009b), two reports showed that psToc34 mutant (R133A) hydrolyzed GTP at much lower rate, much less than 0.3 fold of the wild-type level (Koenig et al., 2008a; Reddick et al., 2007). These results may indicate a structural difference

between atToc33 and psToc34 functions. While psToc34 has high dimerization affinity (Koenig et al., 2008a; Reddick et al., 2007), atToc33 has lower affinity for dimerization (Koenig et al., 2008a; Oreb et al., 2011) and is able to support GTP hydrolysis in monomeric form (Koenig et al., 2008a; Lee et al., 2009b).

1.4.3.3 The models of TOC receptor function

How TOC receptors function together in protein import is at the early state of understanding. Three different models have been proposed based on available data (Bedard and Jarvis, 2005; Kessler and Schnell, 2004; Reddick, 2010; Smith, 2006).

The first model is called the ‘targeting model’ (Keegstra and Froehlich, 1999). It implicates cytosolic Toc159 in the capture of precursor proteins in the cytoplasm (step 1b of Figure 1-3) and their targeting to the plastid surface (Bauer et al., 2002; Hiltbrunner et al., 2001; Lee et al., 2003; Smith et al., 2002). Toc159-precursor complex is proposed to interact with Toc34 (step 2a) through the G domains (Bauer et al., 2002). This process is possibly controlled by the GTPase cycle, which results in transfer of precursor to Toc75 channel to initiate translocation (Smith et al., 2002).

The second model is referred to as the ‘motor model’ (Becker et al., 2004b). Toc34 on the outer envelope is proposed to act as the primary receptor (step 1a of Figure 1-3). Toc159 in this case is proposed to function as a GTPase motor pushing precursor proteins through Toc75 channel (Soll and Schleiff, 2004).

The third model is termed the ‘Toc Clock model’ (Reddick, 2010). By analysis of TP, psToc159 and psToc34 interactions together with GTP hydrolysis, it was proposed that in the inactive state, both Toc159 and Toc34 are in GDP-bound forms. Toc159 when loaded with GTP increases its affinity for TP. The TP-Toc159 has higher affinity to form heterodimer with GDP-bound Toc34. The hetero-dimerization induces GDP to GTP exchange of Toc34. The

GTP-loaded Toc34 has higher affinity to TP than GTP-loaded Toc159 resulting the transfer of TP to Toc34. TP acts as GAP stimulating Toc34 GTP hydrolysis and generates GDP-bound Toc34 dimer. Finally, TP is release from low affinity GDP-bound Toc34 into Toc75 channel.

1.4.4 TIC apparatus

The TIC complex was identified to be composed of Tic20, Tic21, Tic22, Tic32, Tic40, Tic55, Tic62 and Tic110 (Bedard and Jarvis, 2005; Kessler and Schnell, 2006; Smith, 2006; Soll and Schleiff, 2004). Their functions in protein import range from facilitating precursor translocation through IMS, TOC-TIC formation, TIC channel, stromal motor and regulation of import.

Tic22 is a peripheral inner membrane protein in the IMS (Kouranov et al., 1998). It is the only known soluble IMS protein in the translocons (Tripp et al., 2012). Based on its location, two functions have been proposed. First, Tic22 may act to facilitate the translocation of the precursor across the IMS (Becker et al., 2004a; Schnell et al., 1994). Tic22 was also suggested to mediate the formation of TOC-TIC supercomplexes at the contact site (Becker et al., 2004a; Schnell and Blobel, 1993).

Many TIC subunits were proposed to form the translocation channel of the inner envelope. However, the first TIC channel was proposed to be the protein import related anion channel (PIRAC) identified via electrophysiological measurement (van den Wijngaard and Vredenberg, 1999). PIRAC channel opening is regulated by TP (Dabney-Smith et al., 1999; van den Wijngaard et al., 1999) and it was found to associate with Tic110 (van den Wijngaard and Vredenberg, 1999). The other electrophysiological experiment identified TP regulated cation channel Tic110 as a TIC channel (Heins et al., 2002). This result excludes PIRAC as a TIC channel since all other protein translocation channels are cation selective (Heins et al., 2002). It was later found that Tic110 contains 6

transmembrane helices and the opening was regulated by Ca^{2+} (Balsera et al., 2009a). Tic110 is also essential for TIC complex formation (Inaba et al., 2005). The other potential TIC channels are Tic20 and Tic21, which are related to the pore-forming subunits of the mitochondrial translocon in the inner envelope (Reumann and Keegstra, 1999; Teng et al., 2006). *attic20* RNAi knockdown and *attic21* knockout plants showed defective translocation across inner membranes (Chen et al., 2002; Teng et al., 2006). While Tic20 is expressed mainly in early development, Tic21 is expressed at a later stage indicating a shared function of these proteins (Teng et al., 2006).

Tic40 is related to the co-chaperone Hip (Hsp70-interacting protein) and Hop (Hsp70/Hsp90-organizing protein) family (Bedard et al., 2007; Chou et al., 2003; Stahl et al., 1999) and plays a part in a stromal motor complex (section 1.4.5). The stromal domain of Tic110 was additionally identified to interact with TP (Akita et al., 1997; Inaba et al., 2003; Nielsen et al., 1997) and also functions in the same stromal motor complex (section 1.4.5).

Tic62, Tic55 and Tic32 were proposed to function in redox regulation of protein import, which is discussed in section 1.5.3.

Recently, a novel TIC complex has been identified (Kikuchi et al., 2013). Using tagged atTic20, four novel TIC proteins were isolated from the same complex including Tic214, Tic100, Tic56 and an unnamed protein under investigation (Kikuchi et al., 2013). The complex was shown to form a trimer based on electrophysiological measurement and could be purified together with the TOC components. Interestingly, neither Tic110, Hsp93 nor Tic40 were shown to be stably associate with this complex.

1.4.5 Stromal motors

It has long been shown that internal ATP hydrolysis is the driving force for plastid protein translocation (Theg et al., 1989). Since the identification of

Hsp70 family of proteins involvement in protein import into endoplasmic reticulum (ER) and mitochondria (Chirico et al., 1988; Deshaies et al., 1988; Murakami et al., 1988; Zimmermann et al., 1988), ATPase chaperones have been proposed to drive the translocation of precursor proteins into plastids (Keegstra and Cline, 1999; Marshall et al., 1990). Two classes of chaperones, Hsp70 and Hsp100 have now been shown to associate and facilitate the protein translocation.

Three Hsp70 proteins associated with chloroplasts were first identified from pea outer envelope and stroma (Marshall et al., 1990). Later, a Hsp70 was identified in the pea chloroplast outer membranes as an import intermediate associate proteins (IAP) (Schnell et al., 1994). This IAP can be detected by antibodies against mammalian cytosolic Hsp70, mAb SPA-820 (Schnell et al., 1994). Because the Hsp70 IAP was protected from an outer envelope impermeable protease thermolysin, it was proposed to localize in the IMS and interact with incoming precursor proteins translocated through Toc75 (Schnell et al., 1994). Since then the gene corresponding to the Hsp70 IAP has not yet been identified. A later study, however, found that the protein recognized by mAb SPA-820 is located in the stroma (Ratnayake et al., 2008). The other outer membrane Hsp70s from pea and spinach, the thermolysin-sensitive Com70s, were shown to associate with various precursors (Ko et al., 1992; Kourtz and Ko, 1997). Com70 also has higher sequence similarity with mammalian cognate Hsp70s than *E. coli* Hsp70, DnaK (Ko et al., 1992). It was later suggested that Com70 is the cytosolic Hsp70-1 (Guy and Li, 1998; Li et al., 1994). A stromal Hsp70 was later shown to associate with the translocation complexes (Nielsen et al., 1997).

In *Arabidopsis*, the only 2 Hsp70s predicted to harbor TP were shown to localize in the stroma (Ratnayake et al., 2008; Su and Li, 2008; Sung et al., 2001). These stromal Hsp70s are closely related to the bacterial Hsp70, DnaK (Ratnayake et al., 2008) and were named atcpHsc70-1 and atcpHsc70-2 (Su and Li, 2008). These cpHsc70s were shown to be directly involved in protein import.

The *atcpHsc70-1 atcpHsc70-2* double knockout is lethal and the single knockout plants showed reduced protein import efficiencies (Su and Li, 2010). Biochemical analysis also found atcpHsc70s stably associate with the translocons and the precursors (Su and Li, 2010). The stromal Hsp70, ppHsp70-2 involved in protein import was concurrently identified in moss *Physcomitrella patens* (Shi and Theg, 2010).

The stromal Hsp93 (ClpC), a member of the Hsp100 chaperone family, was first identified to be stably associated with the translocons (Nielsen et al., 1997). Two homologs were found in *Arabidopsis*, atHsp93-V and atHsp93-III (Kovacheva et al., 2005). Cross-linking data suggested that Hsp93, Tic110 and Tic40 function in concert during protein import (Chou et al., 2003), consistent with genetic analysis results (Kovacheva et al., 2005). The molecular interactions between these proteins have been studied. The binding of TP to Tic110 triggers the association of Tic110 with Tic40, which in turn induces the release of TP from Tic110 (Chou et al., 2006). Tic40 was also shown to stimulate Hsp93 ATP hydrolysis. Because it has lower affinity to ADP-bound Hsp93, Tic40 was suggested to act as an ATPase activating protein of Hsp93 (Chou et al., 2006).

Both Hsp70 and Hsp100 systems are essential for viability of plants (Shi and Theg, 2011). Knockout of each system entirely are lethal such as those observed in *atcpHsc70-1 atcpHsc70-2* double knockout (Su and Li, 2008), *athsp93-V athsp93-III-2* double knockout (Kovacheva et al., 2007), and *pphsp70-2* knockout (Shi and Theg, 2010). However, these results may reflect not only their roles in protein import but also other roles in plant development (Constan et al., 2004; Lee et al., 2007; Su and Li, 2008). In minimally invasive cases, the knockout lines of a single homolog of each system, such as *atcpHsp70-1*, *atcpHsp70-2*, and *athsp93-V* single mutants, import efficiencies dropped to about 40-60% (Su and Li, 2010) indicating important roles in protein import of each system. When the functions of both systems were reduced by knocking out a single homolog from each system such as the *atcpHsp70-1 athsp93-V* double

knockout plant, import was further reduced to 30% (Su and Li, 2010). This additional reduction compared to the effect of each single mutant, together with the coimmunoprecipitation of Hsp70 and Hsp93 suggested that both systems act in concert in the same complex (Shi and Theg, 2011).

1.4.6 Peptidases

The processing of prSSU was discovered over 30 years ago (Dobberstein et al., 1977). This finding led to the identification of two SPPs from pea (Oblong and Lamppa, 1992; VanderVere et al., 1995) and one in *Arabidopsis* (Richter and Lamppa, 1998; Zhong et al., 2003). SPP is classified as a member of the metalloprotease M16B subfamily with a zinc-binding motif (Aleshin et al., 2009; Richter and Lamppa, 1998; Richter et al., 2005; VanderVere et al., 1995). It was shown that SPP is the general processing enzyme of plastid protein import by its ability to proteolyze various precursors (Richter and Lamppa, 1998). SPP specifically binds to the C-terminal 12 residues of TP of prFD (Richter and Lamppa, 1999; Richter and Lamppa, 2002; Richter and Lamppa, 2003). *In vitro*, TP directs precursor binding to SPP (Richter and Lamppa, 1999). The mature domain is released immediately after cleavage of TP while TP remains attached to SPP (Richter and Lamppa, 1999). TP is further fragmented before release from SPP (Richter and Lamppa, 1999). The SPP recognition sites at the C-terminus of TP seem to share a weak motif (Emanuelsson et al., 1999; Gavel and von Heijne, 1990). Later, the physicochemical properties at specific residues were proposed to form a SPP binding motif (Richter and Lamppa, 2002). In fact, the regions at the C-termini of TPs show a positive net charge and are conserved at the position -1 for basic residue and positions -4, -3, -2 and +1 for uncharged residues relative to the SPP cleavage sites (Richter and Lamppa, 2002).

Free cleaved TPs in the stroma are potentially harmful (Glaser et al., 2006) and are degraded by presequence peptidases (PrePs) (Bhushan et al.,

2005). PrePs were originally identified from the mitochondrial matrix as mitochondrial presequence degradation enzymes (Stahl et al., 2002). They are classified as members of the metalloprotease M16C subfamily (Glaser et al., 2006). *Arabidopsis* has 2 PrePs, atPreP1 and atPreP2. Both proteins were found to have dual localizations in both chloroplasts and mitochondria (Bhushan et al., 2003; Bhushan et al., 2005; Moberg et al., 2003; Stahl et al., 2002). In addition, these proteins have different tissue-specific expression patterns; atPreP1 is expressed highly in flowers and can be detected in siliques while atPreP2 is expressed in leaves, shoots and roots (Bhushan et al., 2005).

1.4.7 Multiple Paralogs of Translocon Subunits

Although single copies of each TOC subunit were discovered in pea, the *Arabidopsis* genome sequences revealed multiple genes for most TOC/TIC subunits (Jackson-Constan and Keegstra, 2001). Much evidence now supports the proposal that these isoforms perform different functions in protein import (Jarvis et al., 1998), which is discussed later in section 1.5.2.

In *Arabidopsis*, two genes are coding for Toc34, *atTOC33* and *atTOC34* (Gutensohn et al., 2000; Jarvis et al., 1998), while Toc159 is encoded by 4 genes, *atTOC159*, *atTOC132*, *atTOC120* and *atTOC90* (Kubis et al., 2004).

Toc75 in *Arabidopsis* is encoded by 4 genes, *atTOC75-I*, *atTOC75-III*, *atTOC75-IV* and *atTOC75-V* (Baldwin et al., 2005). The atTOC75-V paralog was shown to be distinct from the other atToc75 proteins based on protein size, phylogenetic analysis and mechanism of insertion into the outer envelope (Inoue and Potter, 2004). It was renamed to *Arabidopsis* outer envelope protein of predicted 80 kDa (atOEP80) and suggested that this protein may function in the insertion of beta-barrel proteins into the outer envelope of plastids similar to the function of the Omp85 protein family (Inoue and Potter, 2004). Another paralog, *atTOC75-I* was shown to be a pseudogene disrupted by a transposon (Baldwin et

al., 2005). Out of the remaining two genes, *atTOC75-III* encodes for full length Toc75 containing TP while *atTOC75-IV* encodes a N-terminally truncated Toc75 having only the beta-barrel domain (Baldwin et al., 2005). The role of atToc75-IV was found to be limited to etioplast development while the *attoc75-III* mutant is embryo-lethal (Baldwin et al., 2005). This indicates the important role of atToc75-III as a common channel for all of the TOC receptors (Jarvis, 2008).

Two of the TIC subunit genes, *atTIC20* and *atTIC21* were found to encode for the proteins related to the pore-forming subunits of the mitochondria translocon at the inner envelope (Reumann and Keegstra, 1999; Teng et al., 2006). In addition, four isoforms of Tic20, *atTIC20-I*, *atTIC20-II*, *atTIC20-IV* and *atTIC20-V* are encoded in *Arabidopsis* (Kasmati et al., 2011).

1.5 Regulation of Plastid Protein Import

1.5.1 Expression control

Spatial and temporal expressions of the nuclear-encoded plastid precursor proteins under different internal and external conditions are well documented (Drea et al., 2001; Gesch et al., 2003; Harmer et al., 2000; Knight et al., 2002; Plumley and Schmidt, 1989; Vorst et al., 1990; Zhou et al., 2006). These regulations alter the cytosolic levels of precursors, which affect the rates of the precursor protein import. Because most nuclear-encoded plastid proteins utilize the general import pathway, changing the expression level of any of these proteins can also potentially affect the import rates of other proteins (Row and Gray, 2001a). Recent evidence indicates that some of the gene expression regulation of nuclear-encoded plastid proteins also involves retrograde signaling from the plastids (Estavillo et al., 2011; Gray et al., 2003; Kakizaki et al., 2009; Pesaresi et al., 2006).

In addition to the precursor proteins, the expressions of the translocon components are also regulated. While green tissues express higher levels of atToc33, atToc159, atToc64-III, atTic55, atTic62 and atTic40, non-green tissues express higher levels of atToc34, atToc132, atToc120, atTic20-I and atTic20-IV (Gutensohn et al., 2000; Vojta et al., 2004). A transcription factor CIA2 was also shown to upregulate atToc33 and atToc75-III expressions in leaves (Sun et al., 2001; Sun et al., 2009). Thus, cells control spatial and temporal protein import rates by altering precursor proteins level and generating different combinations of translocon components.

1.5.2 Precursor-specific import pathways

1.5.2.1 Photosynthetic and nonphotosynthetic precursors

It was proposed that the multiple paralogs of the translocon subunits perform different functions (Jarvis et al., 1998). Currently, much evidence is available to support this hypothesis.

Knockout mutant phenotype analysis and biochemical characterization found that atToc33 associates with atToc159 in the TOC complex functioning in the import of photosynthetic proteins while atToc34, atToc132, atToc120 are found in the TOC complex that functions in the import of nonphotosynthetic proteins (Ivanova et al., 2004; Kubis et al., 2004; Smith et al., 2004). In spinach, two Toc34 isoforms were also identified (Voigt et al., 2005) suggesting that other plants may also utilize specialized TOC receptors. In addition, the A domains of Toc159 have been shown to function in precursor selectivity of atToc159 and atToc132 (Inoue et al., 2010). This selectivity is further shown to depend on the TP sequence of the precursors (Wan et al., 1996; Yan et al., 2006).

However, the elements of TP corresponding to precursor-class selection are still largely unknown (Jarvis, 2008). One of the element discovered was a segment on the TP of *Arabidopsis* small subunit of RuBisCO (atSStp) from

residue 41 to 49, which governs the Toc159-dependent pathway (Lee et al., 2009a). Another element was identified from microarray analysis of nuclear-encoded plastid protein genes in *ppi1* mutant, the *attoc33* knockout plant (Vojta et al., 2004). Only the down-regulated genes were shown to contain positively charged aa at the C-terminal of TPs (-8 and -1 positions) suggesting this element is involved in atToc34 recognition (Vojta et al., 2004).

The precursor-specific pathway seems to merge at the TIC complex where TIC components were found to associate with both photosynthetic and nonphotosynthetic proteins (Chen et al., 2002; Jarvis, 2008; Kovacheva et al., 2005). Nevertheless, atTic20-IV was suggested to function in the alternative import pathway for housekeeping proteins (Kikuchi et al., 2013).

1.5.2.2 Age-specific precursors

Recently, the age-dependent regulation of protein import has been discovered (Teng et al., 2012). Precursors can be classified into 3 groups based on the optimal import rates into different ages of chloroplasts: young chloroplast specific, old chloroplast specific, and age-independent. The import efficiency into different ages of chloroplasts was shown to depend on the sequence of TP (Teng et al., 2012). Import competition assays also found that TP competed better within its own group suggesting each group utilizes a specific pathway (Teng et al., 2012). The attempt to determine the age-specific signal of TP identified two consecutive positive charged residues as signal for old chloroplast specific pathway (Teng et al., 2012). It is still unknown whether specific TOC receptor combinations participate in this recognition similar to that of section 1.5.2.1 or whether post-translational modification is involved in creating the age-dependent signal of TP (Teng et al., 2012). Although the physiological relevance of age-dependent import was shown by analyzing the precursor gene families, where each precursor contains TP from different age-selective group (Teng et al., 2012),

it is unknown whether the aging of chloroplast only depends on the age-selective import and/or differential expression of the precursors.

Nevertheless, the only reported components of translocons that differentially function at different ages are atTic20 and atTic21. Whereas atTic20 function is important in the early development stages, Tic21 function becomes dominant in the mature stage (Li and Chiu, 2010; Teng et al., 2006).

1.5.3 Redox regulation

The redox regulation of plastid protein import was shown to occur at both TOC and TIC translocons (Balsera et al., 2010). Earlier studies found that Cys-modifying agents (Friedman and Keegstra, 1989; Row and Gray, 2001b) and disulfide reducing agents (Pilon et al., 1992a; Stengel et al., 2009), inhibit and stimulate protein import, respectively. Protein import in *Physcomitrella* and *Chlamydomonas* were also enhanced in the presence of reducing agents (Stengel et al., 2009). In addition, the oxidant CuCl_2 was found to inhibit protein import by inducing disulfide bridge formation between Toc34, Toc75 and Toc159 (Seedorf and Soll, 1995). Disulfide bridge dimerization of Toc34 with a single conserved Cys has also been shown both *in vitro* and *in organello* (Lee et al., 2009b). These findings indicate the possibility of redox-dependent disulfide bridge regulation of protein import.

Another level of redox regulation involves TIC subunits. It was proposed that TIC components containing redox-related domains might be involved in regulation (Bedard and Jarvis, 2005). While both dehydrogenases Tic62 and Tic32 harbor NADPH-binding sites (Chigri et al., 2006; Stengel et al., 2008), Tic55 has a Rieske 2Fe-2S center (Caliebe et al., 1997). Additionally, Tic62 contains a binding site for ferredoxin-NADP⁺ reductase (FNR) (Stengel et al., 2008). The ratios of stromal NADP⁺/NADPH have been shown to regulate the movement of Tic62 between stroma and inner envelope, and the interaction of

Tic62 with FNR (Stengel et al., 2008). In reducing conditions, Tic62 accumulates in the stroma and has a higher affinity to FNR (Stengel et al., 2008). Another study showed that NADPH abolished Tic62 and Tic32 interaction with Tic110 (Chigri et al., 2006). Lastly, the stromal $\text{NADP}^+/\text{NADPH}$ ratio has been linked to regulate chloroplast protein import of a subgroup of precursors where higher ratios stimulate import (Stengel et al., 2009). This result confirms the role of redox regulation in protein import. Further studies would be required to determine the exact mechanism controlling the redox regulation.

1.5.4 Phosphorylation regulation

Specific phosphorylation of Toc34 and Toc159 by outer envelope kinases have been reported (Fulgosi and Soll, 2002). *In vitro*, Toc34's ability to bind GTP and homodimerize were inhibited by phosphorylation (Jelic et al., 2002; Oreb et al., 2008; Sveshnikova et al., 2000b). However, point mutations of atToc33 abolishing or mimicking phosphorylation showed similar import efficiencies as that of wild type (Aronsson et al., 2006; Oreb et al., 2007). Furthermore, atToc33 but not atToc34 can be phosphorylated, indicating different regulation of the two receptors (Jelic et al., 2003). Recently, a proteolytic fragment of atToc159 was shown to be solubilized in the cytosol in a hyperphosphorylated form (Agne et al., 2010).

Phosphorylation of TPs has also been shown to regulate protein import. Phosphorylations of multiple precursors have been observed (Waagemann and Soll, 1996) together with the identification of the kinases (Martin et al., 2006). The guidance complex recognizes phosphorylated precursors before delivering them to the translocons (section 1.4.1) (May and Soll, 2000). A phosphorylation-dephosphorylation cycle was proposed where the incoming precursors are in phosphorylated forms and the translocon-associated phosphatase dephosphorylates the precursors to initiate translocation (Waagemann and Soll,

1996). However, the phosphatase has not yet been identified. Nevertheless, the mutants of three TPs lacking the phosphorylation sites were able to direct the import of GFP into the plastids, which indicates that phosphorylation of TP is not required for plastid targeting (Nakrieko et al., 2004).

1.5.5 Regulation by ubiquitin-proteasome system

The level of precursor proteins in cytosol has been shown to be regulated by the ubiquitin-proteasome system, specifically through the cytosolic Hsc70 and the carboxy terminus of Hsc70-interacting protein (CHIP) E3 ubiquitin ligase pathway (Lee et al., 2009c; Shen et al., 2007). Another evidence also indicated that a putative C3HC4-type really interesting new gene (RING) E3 ubiquitin ligase SP1 interacts with all of the TOC components and initiates their degradation *via* the proteasome (Ling et al., 2012). It was proposed that SP1 regulates the turnover of TOC components and in combination with the differential expression of TOC components resulting in alteration of the composition of TOC components (Ling et al., 2012). This TOC composition change was also suggested to control the transition between plastid types (Ling et al., 2012).

1.6 Transit Peptides

TP is necessary and sufficient to facilitate protein import into plastids; the mature domain alone fails to be imported while addition of TP can direct the import of non-plastid proteins into plastids (Bruce, 2001). Thus, TP contains information governing the import process. The length of TPs varies from 20 – 150 aa based on the position of the processing site (Balsera et al., 2009b). However, it has been shown recently that short TPs cannot direct the import

(Bionda et al., 2010). Import only occurred when short TPs were extended into their mature domains to reach at least 60-aa in length (Bionda et al., 2010).

Many attempts have been made to identify the conserved motifs within TP using primary sequence alignment, but the conservation was greatly reduced upon increasing the numbers of TPs in the alignment (Karlín-Neumann and Tobin, 1986; von Heijne et al., 1989). Aa composition and organization are also highly divergent (Bruce, 2000). Nevertheless, three regions were loosely defined: (i) N-terminal domain of about 10 residues, lacking charged aa, ending with Pro/Gly and preferably having Ala as the second residue, (ii) central domain, lacking acidic aa but rich in hydroxylated aa, and (iii) C-terminal domain, rich in Arg and possibly forming an amphiphilic β -strand (Bruce, 2001; von Heijne et al., 1989).

A few studies have determined NMR structures of TP including the TPs of RuBisCO activase (Krimm et al., 1999) and ferredoxin (Lancelin et al., 1994) from *Chlamydomonas*, and TP of ferredoxin from a higher plant, *Silene latifolia* (Wienk et al., 1999). Even though TPs are found to be unstructured in aqueous solution, they were shown to adopt alpha-helical structures in membrane-mimetic environments suggesting the possible involvement of an alpha-helix in TP recognition (Bruce, 2001). Notably, the most stable alpha-helix of higher plant ferredoxin TP contains a semi-conserved FGLK motif that was suggested to interact with the translocation apparatus (Schleiff et al., 2002; Wienk et al., 2000).

1.7 Interacting Domains in Transit Peptides

Many groups have determined precursor interactions with translocon components in intact chloroplast by cross-linking during different stages of the import process (Akita et al., 1997; Chen and Li, 2007; Inoue and Akita, 2008a; Inoue and Akita, 2008b; Kouranov and Schnell, 1997; Ma et al., 1996; Perry and

Keegstra, 1994). These interactions can be confirmed *in vitro* using purified translocon components and various precursors or TPs. Ten different precursors (prSSU/prRBCA, prFNR, prOE23, prOE33, prFD, prHsp21, prLHCP, prE1 α , prL11, prPORA) were used in these experiments. The *in vitro* assays have uncovered many interacting partners including the guidance complex (May and Soll, 2000), cytosolic Hsp90 (Qbadou et al., 2006), Toc159 (Smith et al., 2004), Toc34 (Schleiff et al., 2002; Sveshnikova et al., 2000b), Toc75 (Hinnah et al., 2002), Tic110 (Inaba et al., 2003), stromal Hsp70 (Ivey et al., 2000; Rial et al., 2000), stromal processing peptidase (Richter and Lamppa, 1999), and presequence peptidases PreP1/2 (Glaser et al., 2006).

While this interaction information confirmed the cross-linking results in identifying the combination of translocon components at each state of the import pathway, it does not provide direct identification of the interacting domains on TPs. Only a limited number of studies reported the interacting domains of TPs and are summarized in Figure 1-4. TPs of the small subunit of RuBisCO (SStp) and ferredoxin (FDtp) are the only TPs with their interacting domains mapped. The mapped domains include lipid, Toc159, Toc34, stromal Hsp70 CSS1, and SPP interacting domains (Becker et al., 2004b; Ivey et al., 2000; Lee et al., 2009a; Pilon et al., 1995; Pinnaduwege and Bruce, 1996; Schleiff et al., 2002; Sveshnikova et al., 2000b).

Mutagenesis (Lee et al., 2009a; Lee et al., 2002; Pilon et al., 1995), deletion (Kindle, 1998; Kindle and Lawrence, 1998; Rensink et al., 1998; Rensink et al., 2000), Ala scanning (Lee et al., 2008; Lee et al., 2006; Lee et al., 2009a; Lee et al., 2002), domain swapping (de Castro Silva Filho et al., 1996; Lee et al., 2009a; Smeeckens et al., 1986), bioinformatics analysis (Lee et al., 2008; von Heijne et al., 1989) and the use of synthetic peptides (Perry et al., 1991; Pinnaduwege and Bruce, 1996; Schnell et al., 1991) have identified a large

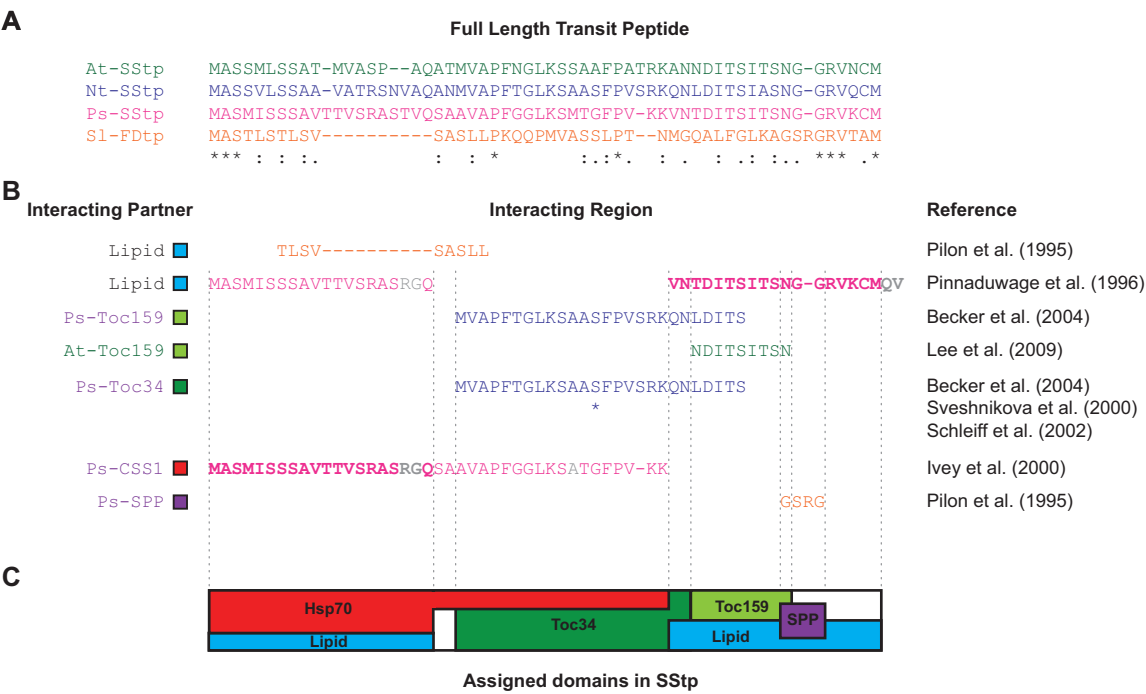


Figure 1-4. Summary of All Experimentally Determined Interacting Domains in TPs.

(A) Aa sequences of full-length TPs. (B) Interacting domains in TPs. For lipid and Ps-CSS1 interaction, the sequences in bold have stronger interactions than sequences with normal characters. Asterisk indicates phosphorylated Ser. (C) All the interacting domains found on SStp. At, *A. thaliana*; Nt, *N. tabacum*; Ps, *P. sativum*; Sl, *S. latifolia*.

number of critical regions in TPs containing information for the import process. However, the functions of most of these regions are still unknown. More importantly, these regions seem to be unique sequences; the same exact sequence could not be found in other TPs. This makes it difficult to elucidate their function. These identified regions also undermine the hypothesis that translocon component recognition of TPs is based on highly conserved sequence motifs. Instead of exact sequence motif, a few import-critical regions of TPs can be identified using algorithms that quantify aa composition. These regions are the N-terminal highly uncharged domain (von Heijne et al., 1989), the Hsp70 interacting domain (Ivey et al., 2000), and the FGLK domain (Chotewutmontri et al., 2012; Pilon et al., 1995). Thus, many steps in the general import pathway may recognize TP by physicochemical properties, which would also explain the lack of conserved motifs in TPs.

Chapter 2

Materials and Methods

2.1 Polymerase Chain Reaction

2.1.1 Amplification of DNA inserts

Some of the DNA inserts for cloning were produced by polymerase chain reaction (PCR) amplification of targeted sequence with specific primers. The template and primer combinations are listed in section 2.2. High efficiency DNA polymerase, Ex Taq (Takara), was used in PCR of DNA inserts. Fifty- μ l PCR reactions contained 0.4 ng/ μ l of DNA templates, 1x Ex Taq buffer, 0.03 U/ μ l Ex Taq, 0.2 mM of each dNTPs and 0.2 μ M of each primers. The PCR conditions were set with a step of denaturation for 2 min at 94 °C, followed by 35 cycles of denaturation-annealing-extension (15 sec at 94 °C, 15 sec at annealing temperature, 1 min for every 1 kb of amplicon at 72 °C), a step of extension for 7 min at 72 °C and a holding step at 4 °C. The annealing temperatures were generally chosen at the average melting temperatures of the primer pair. If the average melting temperatures exceeded 72 °C, the annealing temperature was set at 72 °C.

2.1.2 Screening of *E. coli* colonies

PCR was used to screen for *E. coli* colonies containing correct DNA inserts after cloning into a vector. The primers were chosen to target flanking regions of the insertion site on the vectors. To setup the PCR, an *E. coli* colony was suspended into 12 μ l sterile H₂O in a PCR tube. An aliquot of 2 μ l of suspension was removed and used as an inoculum of an over-night culture. Other

components were added to the PCR tube so that the final reaction contained 0.02 U/ μ l GoTaq DNA polymerase (Promega), 0.05 mM of each dNTPs, 0.2% Triton X-100, 0.2 μ M of each primers, 1 mM MgCl₂ and 1x Green GoTaq buffer. Similar PCR conditions to that of section 2.1.1 were used, except, the first denaturation step was set to 4 min at 94 °C. For convenience, the PCR reactions of GoTaq system can be directly loaded into the wells of agarose gel without addition of loading dye.

2.2 Construction of Vectors

2.2.1 *E. coli* expression vectors

2.2.1.1 Vectors for IMPACT expression system

For production of TPs, codon-optimized sequences from section 2.19.1 were synthesized (Epoch Biolabs). The synthetic DNAs were cloned into a pTYB2 vector (New England Biolabs) using NdeI and XmaI restriction sites. The transformant colonies were screened (section 2.1.2) using T7P and Intein-R primers (Table A1-1). The pTYB2 containing *S. latifolia* ferredoxin in forward and reverse direction (FDF and FDR) and *P. sativum* small subunit of RuBisCO in forward and reverse direction (SSF and SSR) were named pTYB2-FDF, pTYB2-FDR, pTYB2-SSF and pTYB2-SSR, respectively. These constructs were used in section 2.4.1.

2.2.1.2 Constructs based on pET-30a

To express TP-YFP fusion proteins in *E. coli*, the TP-YFP coding sequences in the plant expression vectors (section 2.2.2) were subcloned into *E. coli* expression vector pET-30a (Novagen). An NdeI site at the start codon was introduced during PCR amplification of the constructs from pBS-SSF-YFP, pBS-SSR-YFP, pBS-FDF-YFP, pBS-FDR-YFP and pAN187, using forward primers

SSF-NdeI-F, SSR-NdeI-F, FDF-NdeI-F, FDR-NdeI-F and ntSSF-NdeI-F (Table A1-2), respectively, in combination with M13F primer (Table A1-1). The PCR products were digested with NdeI and NotI and cloned into pET-30a digested with the same enzymes generating pET-SSF-YFP, pET-SSR-YFP, pET-FDF-YFP, pET-FDR-YFP and pET-ntSSF-YFP, respectively. The clones were verified by sequencing. Note that the N- and C-terminal His-tags on pET-30a were not part of the coding sequences.

For the negative control, vectors contain the first 20 aa of the mature domain of prSSU fused to YFP were made. An insert was amplified from pET-FDF-YFP using primers m20-NdeI-F (Table A1-2) and T7ter (Table A1-1), digested with NdeI and NotI and cloned into pET-30a producing pET-m20-YFP. In addition, C-terminal 6xHis tag was added to this construct. Two oligonucleotides, 6xHis-F and 6xHis-R (Table A1-2) were hybridized before ligation into pET-m20-YFP digested with BsrGI and XhoI to produce pET-m20-YFP6xHis.

To facilitate subcloning of expression cassettes from plant expression vectors into *E. coli* expression vectors, an NheI site at the ATG start was introduced into pET-30a-based vector. pET-SSF-YFP was digested with XbaI and NcoI to remove the ribosome binding site (RBS) sequence along with SSF and the mature protein regions. Two oligonucleotides, XbaI-RBS-NcoI-F and XbaI-RBS-NcoI-R (Table A1-2) were hybridized to generate a DNA adapter containing RBS and NheI. The adapter was ligated into the digested pET-SSF-YFP producing pET-NheI-YFP.

All remaining constructs utilized in Chapter 4 were generated by subcloning the expression cassettes into pET-NheI-YFP using NheI and NotI restriction sites.

The PCR protocol described in section 2.1.1 was used in the production of the DNA inserts. All generated constructs were verified by sequencing.

2.2.2 Plant expression vectors

To generate TP-YFP fusion constructs, a plastid marker pAN187 based on pBlueScript (Stratagene) was used as an expression plasmid backbone (Nelson et al., 2007). The expression cassette of pAN187 contains *N. tabacum* small subunit of RuBisCO TP with 20 aa of mature domain followed by YFP (ntSSF-20-YFP) under the control of a double 35S promoter (d35S) and a nos 3' terminator.

For FDF, FDR, SSF and SSR peptides, the original TP sequence in pAN187 was replaced by new TPs generated from the amplification of TP sequences in pTYB2 constructs (section 2.2.1.1) using primers listed in Table A1-3. The NheI site at the start codon of TP and the MscI site in the mature domain were used. The colonies were screened (section 2.1.2) using d35S-F and YFP-5ter-R primers (Table A1-1). The new pAN187 vectors containing FDF, FDR, SSF and SSR were named pBS-FDF-YFP, pBS-FDR-YFP, pBS-SSF-YFP and pBS-SSR-YFP, respectively.

For *Agrobacterium*-mediated transformation (section 2.6.2), the expression cassettes in pAN187-based vectors were subcloned into a binary vector pFGC19 (Nelson et al., 2007) using SacI and HindIII sites flanking the d35S promoter and nos 3' terminator. Note that this enzyme combination removed the plant selection marker from the plasmid. The new pFGC19 vectors containing FDF, FDR, SSF and SSR cassettes were named pFGC-FDF-YFP, pFGC-FDR-YFP, pFGC-SSF-YFP and pFGC-SSR-YFP, respectively.

The pBS-TP-YFP expression vectors containing an addition of the first 10 aa from the opposite TP pair at the N-terminus of pre-existing TP sequence were generated by ligation of the synthetic DNA fragments into the NheI site at the start codon. The DNA fragments containing the first 10 aa of TPs were made by hybridization of oligonucleotides listed in Table A1-4. The new constructs were

named pBS-SSF10-SSR-YFP, pBS-SSR10-SSF-YFP and pBS-FDF10-FDR-YFP. The pBS-FDR10-FDF-YFP was generated differently. The DNA fragment containing d35S promoter and the first 10 aa of FDR was amplified from pBS-FDR-YFP using T7P primer (Table A1-1) and FDR10-XbaI-R primer (Table A1-3). The fragment was digested with SacI and XbaI and cloned into pBS-FDF-YFP using SacI and NheI sites.

Based on the vectors containing the extra sequence of the opposite TP, the mutated constructs with the Met at N-terminus of pre-existing TP substituted by Ala/Ser were generated by using a overlap extension mutagenesis technique (Ho et al., 1989). For each construct, two rounds of PCR were performed to generate a mutant expression cassette using a set of 4 primers: 2 flanking primers and 2 mutagenic primers. M13R and nos-R primers (Table A1-1) were used as flanking primers in every mutagenesis. The pBS-SSF10-SSR-YFP vector was mutated by substituting the first Met residue of SSR to Ala with the mutagenic primers SSF10-MtoA-SSR-F and SSF10-MtoA-SSR-R. The pBS-SSR10-SSF-YFP vector was mutated by substituting the first and fourth Met residues of SSF to Ala and Ser, respectively, with mutagenic primers SSR10-MtoA-SSF-F and SSR10-MtoA-SSF-R. The pBS-FDF10-FDR-YFP vector was mutated by substituting the first Met of FDR to Ala with mutagenic primers FDF10-MtoA-FDR-F and FDF10-MtoA-FDR-R. The pBS-FDR10-FDF-YFP vector was mutated by substituting the first Met of FDF to Ala with mutagenic primers FDR10-MtoA-FDF-F and FDR10-MtoA-FDF-R. The mutagenic primers are listed in Table A1-3. The final PCR products containing the mutant cassettes were cloned into the same expression vectors using SacI and NotI sites to replace the former cassettes. The mutated constructs were named pBS-SSF10-MtoA-SSR-YFP, pBS-SSR10-MtoA-SSF-YFP, pBS-FDF10-MtoA-FDR-YFP and pBS-FDR10-MtoA-FDF-YFP.

The N-terminal mutants used to assess the N-terminal Hsp70 binding property of TPs in Chapter 4 were generated based on 2 previously generated

constructs, pBS-SSR10-MtoA-SSF-YFP and pBS-SSF10-MtoA-SSR-YFP. The DNA sequences of 8 peptides, pp38, pp9, PepG, V10, A6R, HbS, np09 and HA, were introduced on the forward primers to replace SSR10 and SSF10 in pBS-SSR10-MtoA-SSF-YFP and pBS-SSF10-MtoA-SSR-YFP, respectively. For the DRC8 peptide, a single primer could not cover the whole sequence. We designed a set of 2 forward primers that when used in 2 consecutive PCR reactions will introduce a complete sequence of DRC8. The primers were listed in Table A1-3. PCRs were performed together with nos-R (Table A1-1) to generate the DNA inserts. *NheI* and *NotI* were used for cloning the inserts into original vectors pBS-SSR10-MtoA-SSF-YFP and pBS-SSF10-MtoA-SSR-YFP.

The flipped and scrambled N-terminal mutants were produced by replacing the N-termini of SSF and FDF with the mutant sequences. The forward primers (Table A1-3) containing the DNA sequences of the flipped and scrambled sequences of the N-terminal 10 residues of SSF and FDF together with nos-R primer were used to produce mutant DNA inserts. The inserts were cloned into the original vectors pBS-SSF-YFP and pBS-FDF-YFP using *NheI* and *NotI*.

In Chapter 5, the mutant constructs containing the 14-aa designed Hsp70-FGLK spacers were produced from pBS-SSF-YFP. The vector was digested with *NheI* and *SphI* to remove the TP sequence. Oligonucleotides corresponding to the sequences of the mutant SSFs were hybridized to form the DNA inserts and cloned into the digested vector generating pBS-SSF-YFP-no92-14aa, pBS-SSF-YFP-no228-14aa and pBS-SSF-YFP-no296-14aa. The oligonucleotides were listed in Table A1-5. The DpnI-mediated mutagenesis was used repeatedly to insert the additional sequences generating the longer mutants spacers from 14 to 44 aa. Mutagenesis was also performed to delete the sequence at the N-terminus of the spacers of the 44-aa constructs to generate additional mutants with 34-aa spacers (the 44-10 mutants). To generate the mutants with equal TP length, the mutants with 34-aa spacer were used. First, the FGLK motifs in these constructs were deleted making the mFGLK mutants. Then the FGLK motifs were re-introduced

at different positions to generate the mutants with the spacer of lengths 14, 19, 24 and 29 aa. The FGLK insertion was performed in two steps because the insertion sequence was too long. Half of FGLK was introduced in each step. All of the mutagenic primers used to generate the designed spacer mutant were listed in Table A1-6.

For the construction of the spacer mutants based on the native spacers, the DpnI-mediated mutagenesis was used to mutate the pBS-SSF-YFP and pBS-FDF-YFP vectors. The sequence coding for 5 aa were added or deleted at each round of mutagenesis. The mutagenic primer pairs were listed in Table A1-7.

All of the DNA inserts were generated using the PCR protocol described in section 2.1.1 and the sequences of all of the generated constructs were verified by sequencing.

2.3 Site-directed Mutagenesis

Two different strategies were used to mutagenize the DNA constructs. One method was the overlap extension mutagenesis technique described by Ho et al. (1989) utilizing two rounds of PCR in generating mutant DNA insert. Another method was the DpnI-mediated mutagenesis described by Fisher and Pei (1997). The specific primers utilized in each mutagenesis were given in the description of vector constructions (section 2.2)

2.4 Expression and Purification of Proteins

2.4.1 Forward and reverse peptides

The expression and purification of FDF, FDR, SSF and SSR were performed similar to the previously reported protocol (Reddick et al., 2007) based on the IMPACT system (New England Biolabs). In this system, our proteins

were tagged at the C-terminus with an inducible self-cleavage Intein protease fused to a chitin-binding domain. In final purified peptides, only ProGly was remained tagged to the peptides. Briefly, *E. coli* ER2566 (New England Biolabs) harboring pTYB2 constructs (section 2.2.1.1) were grown in Luria-Bertani broth (LB: 1% tryptone, 0.5% yeast extract and 1% NaCl) at 37 °C and shaken at 225 rpm until OD₆₀₀ reached between 0.3 – 0.4. The cells were induced with 1 mM final concentration of IPTG and expressed overnight at 25 °C with shaking at 225 rpm. Cells were harvested and lysed in lysis buffer (20 mM Na-phosphate, pH 8.0, 1 mM EDTA, 500 mM NaCl, 0.1% Triton X-100, 1 mM PMSF, 1 µM leupeptin and 1 µM pepstatin) using a French press. The DNA was fragmented with sonication or digested with Benzonase nuclease (Sigma). Lysate was centrifuged at 20,000 x g for 30 min. Supernatant was collected and loaded onto a chitin matrix (New England Biolabs) column. The column was extensively washed with column buffer (20 mM Na-phosphate, pH 8.0, 500 mM NaCl and 1 mM EDTA). Cleavage was induced on the column by replacing the column buffer with an elution buffer (1 mM phosphate buffer, pH 7.5 and 50 mM BME) and incubated overnight at 4 °C. The peptides were eluted with elution buffer without BME, lyophilized to remove BME and stored at -80 °C.

2.4.2 Precursor and mature proteins

Precursor of the small subunit of RuBisCO (prSSU) and its mature domain (mSSU) from *Nicotiana tabacum* were expressed from pET-11d constructs (Klein and Salvucci, 1992) as inclusion bodies. The method described here has been published previously (Reddick et al., 2008). Briefly, *E. coli* BL21(DE3) harboring pET-11d constructs were grown in Terrific broth (TB: 1.2% tryptone, 2.4% yeast extract, 0.94% K₂HPO₄, 0.22% KH₂PO₄ and 0.4% glycerol) containing 150 µg/ml ampicillin at 37 °C with shaking at 250 rpm until OD₆₀₀ of 0.6 – 0.8 was reached. The cells were induced with 1 mM final

concentration of IPTG and expressed for 6 h at 37 °C with shaking at 250 rpm. Cells were harvested and lysed in lysis buffer (50 mM Tris-HCl, 2 mM EDTA, 1 mM DTT and 0.5% Triton X-100) using sonication. The lysate was centrifuged at 36,000 x g for 20 min. Pellets were washed with lysis buffer containing 300 mM NaCl, 3 times, using a Dounce homogenizer. The pellets were washed another 2 times with lysis buffer and finally washed with cold H₂O to remove salt and detergent. The pellets were resuspend in urea solubilization buffer (8 M urea, 50 mM DTT and 20 mM Tris-HCl, pH 8.0 by shaking 200 rpm overnight at 37 °C. The suspensions were centrifuged at 40,000 x g for 30 min to collect pellets. Most of the impurities remained in the supernatants and inclusion bodies remained as pellets as long as the volumn of urea solubilization buffer was kept minimal. The pellets were resuspended in another urea solubilization buffer. The suspensions were centrifuged again to remove insoluble components. Supernatants containing the proteins were stored at -80 °C.

2.4.3 Radiolabeled proteins

2.4.3.1 *In vivo* labeling of prSSU

To generate ³⁵S-prSSU, a variant method to section 2.4.2 was performed. *E. coli* BL21(DE3) carrying pET-11d containing prSSU (Klein and Salvucci, 1992) was grown in 5 ml of TB containing 150 µg/ml ampicillin at 37 °C with shaking at 250 rpm until OD₆₀₀ reached 0.6. A 3-ml inoculum was transferred to 30 ml of Dulbecco's Modification of Eagle's Medium deficient in Met, Cys, and Gln (MP Biomedicals) containing 150 µg/ml ampicillin and grown in the same conditions. When the OD₆₀₀ reached 0.6, the cells were induced by addition of IPTG to 1 mM final concentration. After 5 min, 7 mCi of Trans ³⁵S-Label metabolic labeling reagent (MP Biomedicals) was added. The expression was continued for 4–6 h. ³⁵S-prSSU was purified from the inclusion bodies similar to section 2.4.2.

2.4.3.2 *In vitro* labeling of proteins

For *in vitro* import assays, TP-YFP fusion proteins and the control mSSU-6xHis were produced from pET-30a constructs (section 2.2.1.2) using TNT T7 Coupled Wheat Germ Extract System (Promega). *In vitro* coupled transcription and translation was performed in 50 μ l reaction volumes containing 50% TNT Wheat Germ Extract, 4% TNT Reaction Buffer, 0.5% TNT T7 RNA polymerase, 20 μ M Amino Acid Mixture Minus Methionine, 0.8 U/ μ l RNasin Ribonuclease Inhibitor, 20 ng/ μ l DNA template, 0.8 μ Ci/ μ l 35 S-Met *In Vitro* Translation Grade (MP Biomedicals). The reactions were incubated at 30 °C for 2 h. The translation products were stored at -80 °C and further analyzed and utilized in section 2.12.

2.4.4 Toc34G

To study the interaction of Toc34 with TPs in section 2.17, Toc34 was expressed in *E. coli* and purified from previously described methods (Reddick et al., 2008; Reddick et al., 2007). Briefly, the GTPase domain of *P. sativum* Toc34 (Toc34G) fused to 6xHis tag was expressed from the pET-21d construct (Reddick et al., 2007). The cells of *E. coli* BL21(DE3)-RIL harboring the pET-21d construct were grown in LB at 37 °C until OD₆₀₀ of 0.6 was reached. The cells were induced with 1 mM final concentration of IPTG and allowed to express at 25 °C for 4 h. Cells were harvested and purified using PrepEase His-Tagged Protein Purification – High Specificity (USB Corporation). Eluted Toc34G protein was dialyzed into GBS buffer (20 mM Tricine-KOH, pH 7.65, 1 mM MgCl₂, 50 mM NaCl, and 1 mM BME) before further experiments were performed.

2.5 Plant Growth

2.5.1 *Arabidopsis*

Arabidopsis thaliana Col-0 seeds were surface sterilized by incubating in a solution containing 1.8% hypochlorite and 0.1% Triton X-100 for 10 min on a rotator and rinsed 3 times with sterile H₂O. The seeds were plated on an MS plate (0.7% agar, 0.225% Murashige and Skoog Basal Salt Mixture (Sigma), 1% sucrose, pH 6.0 with KOH). Seeds were stratified on the plate at 4 °C for at least 24 h before being moved to the growth chamber. *Arabidopsis* were grown in a growth chamber with illumination at 150 $\mu\text{E}/\text{m}^2/\text{sec}$ on a 16 h light and 8 h dark cycle at 22 °C.

2.5.2 Pea

Dwarf pea (*P. sativum*) seeds were ordered from J.W. Jung Seed Co. and stored at 4 °C. Seeds were imbibed overnight in running tap water with aeration before being planted in 35 cm x 50 cm x 39 cm metal flat with vermiculite. About 400 ml of dry seeds were used per flat. Peas were grown in a growth chamber with illumination at 160 $\mu\text{E}/\text{m}^2/\text{sec}$ on a 14 h light and 10 h dark cycle. The temperature was set at 17 °C and 19 °C for light and dark periods, respectively. Progress Number 9 and Green Arrow cultivars were used in the experiments as stated in the result sections.

2.5.3 Tobacco

Tobacco (*N. benthamiana*) seeds were planted on Super Fine Germination Mix (Fafard) and grown in a growth chamber with illumination at 160 $\mu\text{E}/\text{m}^2/\text{sec}$ on a 16 h light and 8 h dark cycle. The temperature was set at 26 °C and 22 °C for light and dark periods, respectively.

2.6 Transient Expression of Protein in Plants

2.6.1 Biolistic transformation

Biolistic transformation was used to transiently express proteins in *Arabidopsis* and onion (*Allium cepa*). The PDS-1000/He Biolistic Particle Delivery System (Bio-Rad) was used. The protocol was modified from instrument instruction to minimize the damage of plant tissue. First, 60 mg/ml of tungsten particles in 50% glycerol was prepared by washing with 70% ethanol for 10 min and rinsing 3 times with sterile H₂O. Tungstens M-10 and M-17 (Bio-Rad) were used for *Arabidopsis* and onion transformations, respectively. The plasmid vectors (section 2.2.2) were coated on tungsten particles by combining 10 µl of tungsten suspension with 1 µg of plasmid DNA (less than 10 µl), 25 µl of 2.5 M MgCl₂ and 5 µl of 200 mM spermidine. The coated particles were washed once in 70% ethanol and 3 times in absolute ethanol before being resuspended in 10 µl of absolute ethanol. The macrocarrier was spotted with 10 µl of coated tungsten particles. The bombardment was performed at 1,100 psi. For *Arabidopsis*, seedlings 10 – 12 day after germination on the plate (section 2.5.1) were used for transformation. The transformed *Arabidopsis* were transferred back to the growth chamber until further analysis. For onion, adaxial epidermal peels were used where the adaxial surface was placed against the surface of MS plate. The transformed onions were kept at room temperature in the dark until further analysis.

2.6.2 *Agrobacterium*-mediated transient transformation

Transient expression of proteins in tobacco leaf was done as previously described (Sparkes et al., 2006) using *Agrobacterium*-mediated transformation. Briefly, *Agrobacterium tumefaciens* strain GV3101 (pMP90) carrying the binary plasmid was grown overnight in YEB medium (0.1% yeast extract, 0.5% beef

extract, 0.5% peptone, 0.5% sucrose, 0.05% $\text{MgSO}_4 \cdot 7\text{H}_2\text{O}$, pH 7.0) containing appropriate antibiotic at 28 °C with shaking at 220 rpm. The cells were harvested by centrifugation at 3,500 x g for 5 min at 4 °C. The pellets were washed twice with infiltration medium (0.5% D-glucose, 50 mM MES, 2 mM Na_3PO_4 , 0.1 mM acetosyringone) to remove antibiotic and resuspended in infiltration medium to an OD_{600} of 0.1. The resuspended cells were drawn into a 1-ml syringe. To infiltrate, the leaf was held abaxial side up on a finger and the tip of the syringe was pressed against the leaf area above the finger. The plunger was pressed gently to deliver cell suspension into air spaces of the leaf. The infiltrated area was marked for identification. The plants were transferred back to the growth chamber until further analysis.

2.7 Chloroplast Isolation

The chloroplasts were isolated as previously described (Bruce et al., 1994). Briefly, plant tissues were diced with a food processor. All of the following steps were kept on ice or at 4 °C. Grinding buffer (GB: 330 mM sorbitol, 1 mM MgCl_2 , 1 mM MnCl_2 , 2 mM EDTA, 0.1% BSA, 50 mM HEPES-KOH, pH 7.3) was added to the diced tissues before further homogenization using Polytron®. The homogenate was filtered through 2 layers of Miracloth (Calbiochem) on top of 2 layers of cheesecloth. The filtrate was centrifuged at 2,000 x g for 7 min to collect the chloroplasts. Import buffer (IB: 300 mM sorbitol, 50 mM HEPES-KOH, pH 8.0) was added to the pellet and a paintbrush was used to suspend the chloroplasts before loading on top of a continuous Percoll (GE Healthcare) gradient prepared in IB. The loaded gradients were centrifuged at 5,800 x g for 15 min. Intact chloroplasts were collected from the lower band using a 14-gauge stainless steel needle. To remove Percoll, collected chloroplasts were diluted in a 2-fold volume of IB and pelleted by centrifugation at 3,500 x g for 10 min. The pelleted chloroplasts were resuspended in IB before loading onto a second Percoll

gradient. After removal of Percoll, isolated intact chloroplasts were used in further experiments.

2.8 Chlorophyll Measurement

Chlorophyll was extracted from chloroplasts by adding 10 μ l of chloroplast suspension into 990 μ l of 80% acetone, mixed by vortexing for 1 min and centrifuged at 21,000 $\times$ g for 1 min to removed insoluble materials. The absorbance of chlorophylls in 80% acetone was measured at 663 and 645 nm (A_{663} and A_{645}). The total chlorophyll in mg/ml was calculated based on the equation derived by Arnon (1949).

$$\text{Total chlorophyll mg/ml} = \frac{(8.02 \times A_{663}) + (20.2 \times A_{645})}{10 \mu\text{l} \times 1000 \mu\text{g/mg}} \times 1000 \mu\text{l}$$

2.9 Autoradiography

Digital autoradiography was used to quantify radioactivity of ^{35}S -labeled proteins separated by SDS-PAGE. First, Storage Phosphor Screen (Molecular Dynamics) was exposed to the dried SDS-PAGE gel for an appropriate time. The digital autoradiograph was produced by scanning the exposed screen using Storm 840 PhosphorImager (Molecular Dynamics) or Molecular Imager FX (Bio-Rad) with the highest resolution setting. Quantitation of band intensity was performed using ImageQuant software (Molecular Dynamics) or Quantity One software (Bio-Rad).

2.10 Liquid Scintillation Counting

To measure the radioactivity of ^{35}S -labelled proteins in chloroplast samples, the samples were bleached by adding 30 μl of sample into 90 μl of 30% H_2O_2 and incubated at 80 $^\circ\text{C}$ for 30 min. The bleached samples were transferred to a scintillation vial containing 4 ml of EcoLite(+) Liquid Scintillation Cocktail (MP Biomedicals). The samples were counted using an LS 6500 Scintillation Counter (Beckman Coulter).

2.11 *In Vitro* Competitive Chloroplast Protein Binding Assay

A competitive binding assay was used to determine the equilibrium dissociation constant of competitor protein (K_i) to chloroplasts. Varying amounts of non-labeled proteins were used as competitors to compete with radiolabeled precursor in chloroplast binding. To observe binding of proteins to the chloroplasts, the internal ATP level of the chloroplasts had to be minimized. We achieved ATP minimization by harvesting the plant tissues at the end of dark cycle and performing the assay in dim light to prevent ATP synthesis. We previously published the assay described below in detail (Reddick et al., 2008). Briefly, the chloroplasts were isolated from *P. sativum* cultivar Progress Number 9 as described in section 2.7 and diluted to 1 mg/ml chlorophyll as measured by the method of section 2.8. ^{35}S -prSSU was prepared using the method described in section 2.4.3.1. The binding assay was performed in a total volume of 300 μl containing 100 nM ^{35}S -prSSU, 10 mM DTT, 100 mM Na-ATP, 2 mM MgCl_2 , 1% BSA, 300 mM urea and 0.25 mg chlorophyll/ml chloroplasts in IB with various concentrations of competitor. The reactions were terminated after 30-min equilibration at room temperature by adding 700 μl of cold IB. Intact chloroplasts were re-isolated by centrifugation over 700 μl of cold 40% Percoll in

IB at 3,400 x g for 5 min. The supernatants containing broken chloroplasts and unbound ^{35}S -prSSU were discarded. Intact chloroplast pellets were gently resuspended in 1 ml of cold IB using a Pasteur pipette. Protein quantification was performed with 50 μl of resuspended chloroplasts using BCA protein assay (Pierce). The remaining 950 μl of suspensions were pelleted and resuspended in 60 μl of H_2O and mixed with 40 μl of 4x SDS sample buffer (4xSSB: 400 mM DTT, 10% glycerol, 4% SDS, 0.04% bromophenol blue, 40 mM Tris-HCl, pH 6.8). The samples were boiled for 4 min. The protein concentrations in each reaction sample were equalized using the concentrations determined from BCA assay by adding 2xSSB. Equal volume samples (30 – 50 μl) were used to determine the level of bound ^{35}S -prSSU. Samples were separated on 10 – 20% SDS-PAGE gel, fixed in a solution containing 40% methanol and 10% acetic acid, and dried before quantification via autoradiography (section 2.9). Liquid scintillation counting (section 2.10) was also used in quantification.

For the homologous competitive binding of prSSU competitor, two concentrations of ^{35}S -prSSU at 30 and 100 nM were used. Two independent assays of each concentration were performed, and the data were globally fitted to a one-site homologous competitive binding model shown below to determine the equilibrium dissociation constant (K_d) using Prism software (GraphPad).

$$B = \frac{B_{max} \times [Hot]}{[Hot] + [Cold] + K_d} + (NS \times [Hot])$$

The measured binding signals (B), the concentrations of ^{35}S -prSSU ($[Hot]$) and the concentrations of prSSU ($[Cold]$) were known based on the experimental setup. The maximal binding signal (B_{max}), the non-specific binding signal (NS) and the K_d were fitted by non-linear regression.

For other competitors, three independent assays were performed in the presence of 100 nM ^{35}S -prSSU. The data from heterologous competitive binding

were fitted to the one-site competitive binding model shown below to determine equilibrium dissociation constant (K_i) using Prism software.

$$B = \frac{B_{max} - NS}{1 + 10^{([logCold] - logIC50)}} + NS$$

Where

$$logIC50 = \log(10^{logK_i} \times (1 + [HotNM]/HotK_dNM))$$

The measured binding signals (B), the logarithms of concentrations of competitors ($[logCold]$) and the concentration of ^{35}S -prSSU in nM ($[HotNM]$) were known based on the experimental setup. The K_d of prSSU in nM ($[HotK_dNM]$) was previously determined from the homologous competitive binding of prSSU above. The maximal binding signal (B_{max}), the non-specific binding signal (NS) and the logarithm of K_i ($logK_i$) were fitted by non-linear regression. Determination of the logarithm of inhibitor concentration at half maximal binding ($logIC50$) was bypassed when the two equations were fitted together. When only the first equation was used in the fitting, the value of $logIC50$ was determined and the inhibitor concentration at half maximal binding (IC_{50}) of the competitor was derived.

2.12 *In Vitro* Chloroplast Protein Import Assay

Using the import assay, translocations of precursor proteins into the chloroplasts were measured directly and the precursor import rate was also determined. Generally, the precursor is radiolabeled or can be detected with antibody.

For the analysis of purified TP-YFP fusion proteins, spinach chloroplasts were used. Baby spinach was purchased from a local market. The spinach chloroplasts were isolated using the method in sections 2.7 and the chlorophyll

concentration was determined based on section 2.8. The import was performed in total volume of 300 μ l containing 0.25 mg chlorophyll/ml chloroplasts, 400 nM fusion protein, 10 mM DTT, 2 mM Mg-ATP, 0.5% BSA, 300 mM urea in IB. The reactions were stopped after a 20-min incubation at room temperature. The chloroplasts were re-isolated and prepared for SDS-PAGE as described in section 2.11. To determine the import level, 50 μ l of samples were separated on 10 – 15% SDS-PAGE gel and detected by immunoblotting (section 2.18). Three independent assays were performed.

For *in vitro* translated TP-YFP fusion proteins (section 2.4.3.2), the chloroplasts from *P. sativum* cultivar Green Arrow were used. Chloroplasts were isolated using the method in section 2.7 and the chlorophyll concentration was determined based on section 2.8. In each set of experiments, all of the proteins were labeled with the same stock of radioactive ^{35}S -Met. Thus, the ^{35}S -Met found in labeled proteins shared the same specific activity. SDS-PAGE gels were used to separate 1 μ l of the translation product of each protein. Relative radioactivity of each labeled precursor protein was quantified by autoradiography (section 2.9). Relative concentration of each protein was calculated from the following equation.

$$RA = \frac{RR}{N_{Met}}$$

Relative amount of precursor in 1 μ l of translation product (RA) was calculated from the division of its autoradiograph-derived relative radioactivity from 1 μ l of translation product (RR) by the total number of Met presented in its sequence (N_{Met}). The translation products of each precursor were diluted with 50% TNT Wheat Germ Extract (Promega) to equalize relative concentration based on the calculated relative amounts of precursors. Equal volume of the diluted labeled proteins was used in the assay. The import was performed in a

total volume of 500 μ l containing 0.25 mg chlorophyll/ml chloroplasts, labeled protein, 2 mM L-Met, 10 mM DTT, 2 mM Mg-ATP, 0.5% BSA, 300 mM urea in IB. L-Met was added to prevent novel synthesis of radiolabeled protein by the chloroplasts. The reactions were incubated at room temperature. A 150- μ l sample was taken at 5, 10 and 15 min after the reaction was started and mixed with 600 μ l of cold IB to terminate the import. The chloroplasts were re-isolated and the protein concentrations were determined as described in section 2.11. The re-isolated chloroplasts pellets were resuspended in 50 μ l of 2xSSB and boiled for 4 min before the protein concentrations were equalized by addition of 2xSSB. To determine the amount of import, 45 μ l of samples were separated on a 15% SDS-PAGE gel before being quantified by autoradiography (section 2.9). Two sets of independent assays were performed.

2.13 *In Vitro* Competitive Chloroplast Protein Import Assay

A competitive import assay was used to determine IC_{50} of TPs. Varying amounts of non-labeled TPs were used as competitor against ^{35}S -prSSU in chloroplast import. The assay described below was modified from a previously published protocol (Dabney-Smith et al., 1999). Briefly, the chloroplasts were isolated from *P. sativum* cultivar Progress Number 9 (section 2.7) and the chlorophyll concentration was determined (section 2.8). The assays were performed in total volumes of 300 μ l containing 0.125 mg chlorophyll/ml chloroplasts, 100 nM ^{35}S -prSSU, 1 mM BME, 2 mM Mg-ATP, 0.5% BSA, 250 mM urea in IB with various amounts of competitor. The reactions were stopped after a 15-min incubation at room temperature by adding 700 μ l of cold IB. The chloroplasts were re-isolated, solubilized in 2xSSB and separated on SDS-PAGE gel as described in section 2.11. The gels were quantified by autoradiography (section 2.9). At least three separate assays were performed. The values were

normalized to the values from the reaction with no competitor controls, and the data were fitted to one phase exponential decay model shown below using Prism software.

$$I = I_0 \cdot e^{-k \cdot [Cold]} + NS$$

Where

$$IC_{50} = \ln 2 / k$$

The measured import signals (I) and the concentrations of competitor ($[Cold]$) were known based on experimental setup. The maximal import signal at zero amount of competitor (I_0), the decay rate (k) and the non-specific signal (NS) were fitted with non-linear regression from the first equation. IC_{50} of the competitor was derived from the decay rate using the second equation.

2.14 *In Vitro* Stromal Processing Assay

To determine the stromal processed form of chloroplast-imported proteins, the precursors were directly processed with stromal extracts. Baby spinach was purchased from a local market. The spinach chloroplasts were isolated (sections 2.7) and the chlorophyll concentration was determined (section 2.8). The stromal processing assay was modified from a previously described method (Richter and Lamppa, 1998). Briefly, the intact chloroplasts were pelleted and resuspended at 0.8 mg/ml chlorophyll in 5 mM HEPES-KOH, pH 7.5, incubated at 4 °C for 30 min, and lysed using a Dounce homogenizer. The lysate was centrifuged at 137,000 x g for 30 min, and the supernatant was used as the stromal extract. The processing assay was performed in a total reaction volume of 100 µl containing 60 µl stromal extract, 250 nM TP-YFP fusion proteins, 2 mM PMSF, and 20 mM HEPES-KOH, pH 7.5. The reactions were incubated at room temperature. Samples of 50 µl were taken at 0, 10, and 60 min after the reaction was started.

The samples were immediately mixed with equal volume of 4xSSB and boiled for 3 min. The samples were used for immunoblotting (section 2.18). Two separated assays were performed.

2.15 *In Vivo* Chloroplast Protein Import Assays

2.15.1 Qualitative analysis using fluorescent imaging

In vivo protein import of TP-YFP fusion proteins was observed based on transient expression of the proteins in *Arabidopsis*, onion and tobacco. The sweet onion cultivar Vidalia was used if not stated otherwise. Plant tissues were transformed as described in section 2.6 using the constructs generated from section 2.2.2. The localizations of the proteins were determined from YFP fluorescent signals.

For epifluorescence imaging, an Axiovert 200 M microscope (Zeiss) equipped with YFP/cyan fluorescent protein filters (filter set 52017; Chroma) was used. The images were captured with a $\times 63$ (1.4 numerical aperture) plan-apo oil immersion objective unless stated otherwise. Image capture was done with a digital camera (Orca ER; Hamamatsu Photonics). The microscope was controlled by Openlab software (Improvision). Two images were captured with the same exposure time: one with excitation light on and another one with light off. To remove the camera noise, the signal from dark image was subtracted from the fluorescent signal pixel by pixel. For confocal imaging, an SP2 laser scanning confocal microscope (Leica) was used. YFP and chlorophyll were excited at 488 nm using an argon laser. Fluorescent signals from YFP and chlorophyll were recorded from 512 – 584 nm and 650 – 750 nm, respectively. The images were taken with an HC PL APO $\times 20$ (0.7 numerical aperture) objective. All images were captured 12 h after transformation unless otherwise stated. Resizing and cropping of the images for presentation were done using Photoshop (Adobe).

For control, fluorescent protein organelle markers (Nelson et al., 2007) were used for localization comparison. The pAN186 plasmid expressed *N. tabacum* TP of the small subunit of RuBisCO and the first 20 aa of the mature domain followed by CFP (ntSSF-20-CFP) was used as a plastid marker. The pAN83 plasmid expressed CFP containing C-terminal Ser-Lys-Leu tag (px-CFP) was used as a peroxisome marker.

2.15.2 Qualitative analysis using immunoblotting

When TP-YFP fusion proteins were transiently expressed in tobacco using the method of section 2.6.2, the amount of expressed proteins permits detection by immunoblotting. Total protein was extracted from the infiltrated areas of the leaf 2 days after transformation as previously described (Isaacson et al., 2006). Briefly, approximately 1 g of tissues was ground in liquid nitrogen in the presence of 100 mg polyvinylpolypyrrolidone into a fine powder. The frozen powder was transferred into a Dounce homogenizer containing 15 ml of ice-cold extraction buffer (10% trichloroacetic acid and 2% BME in acetone). The tissues were homogenized for 3 min, transferred into 50-ml centrifuge tube and incubated at -20 °C for at least 30 min. The extracted proteins were pelleted by centrifugation at 5,000 x g for 30 min at 4 °C. The pellets were washed 3 times with 10 ml of ice-cold acetone and dried in the fume hood. The pellets were then resuspended in 250 µl of 2xSSB containing 8 M urea, boiled for 3 min, centrifuged at 21,000 x g for 3 min to remove insoluble pellets and stored at -80 °C. The samples were separated on a 10 – 15% SDS-PAGE gel before detection by immunoblotting (section 2.18). Two separate transformations were analyzed.

2.15.3 Quantitative analysis using fluorescent imaging

When TP-YFP fusion proteins were transiently expressed in onion epidermal cells using the method in section 2.6.1, the low density and dispersed

distribution of plastids allowed measurement of each plastid separately. Analysis of the camera noise subtracted images taken from epifluorescence microscopy in section 2.15.1 was performed using ImageJ software (Abramoff et al., 2004). The intensity per pixel values from different areas in the images was calculated by dividing of the summation of intensity signals in the area (Integrated Density variable in ImageJ measurement function) by the total number of pixels in the area (Area variable in ImageJ measurement function). A circle area was drawn to fit around a selected plastid to measure the plastid intensity per pixel. The same circle was then enlarged to threefold diameter (using Specify function of ImageJ). While sharing the same center, the enlarged circle area extended to incorporate intensity signals from the cytosol. The summation of intensity signals and the total number of pixels of the combined plastid and cytosol were measured from the enlarged circle. The summation of intensity signals in the cytosol was calculated by subtracting the summation of the plastid area from the summation of the combined area. The cytosol intensity per pixel was calculated the same way and the cytosol intensity per pixel was calculated. A rectangular area outside the transformed cell in the same image was used to calculate the background intensity per pixel. The background intensity per pixel was subtracted from the plastid and cytosol intensity per pixel values. For each plastid, the ratio between the background removed plastid and cytosol intensity per pixel values was calculated. The ratio intensity of each cell is the average of all plastid ratio values. At least two plastids were used for measurement in each cell. The number of cells used in the measurement for each construct is stated in its figure legend.

2.16 MALDI-TOF Mass Spectrometry

Matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF MS) was performed using a Microflex mass spectrometer (Bruker Daltonics) similar to a previously described protocol

(Reddick et al., 2007). Briefly, the stainless steel target was covered by paraffin wax by streaking with 50 mg/ml paraffin wax in chloroform using a cotton swab and dried under vacuum. TP solutions were mixed 1:1 with the matrix solution (10 mg/ml α -cyano-4-hydroxycinnaminic acid, 50% acetonitrile and 0.1% trifluoroacetic acid) and 2 μ l were spotted on the target plate. After drying, the spots were washed twice with 2 μ l of 10 mM diammonium hydrogen citrate. Mass spectra were acquired in positive ion mode. The peaks were identified using flexAnalysis software (Bruker Daltonics) and analyzed with FindPept (Gattiker et al., 2002).

2.17 Analytical Ultracentrifugation

Analytical ultracentrifugation (AUC) was utilized to observed monomer-dimer equilibrium of Toc34 in the presence of TPs. The experiments were performed with some modification from a previously published method (Reddick et al., 2007). The sedimentation velocity of the proteins was analyzed on an Optima XL-I Analytical Ultracentrifuge (Beckman) using the interference mode. In this mode, the detector is very sensitive to mismatch of buffers in the sample and reference cells. First, we dialyzed Toc34 and TPs extensively into GBS buffer using 10,000 and 3500 molecular weight cutoffs, respectively. The analysis samples were prepared by mixing of Toc34, TP, GTP, and dialysis buffer to a final concentration of 13.5 μ M Toc34, 135 μ M TP, and 2 mM GTP. The samples were dialyzed again in GBS buffer containing 2 mM GTP from 1 h. The AUC cells, Epon charcoal-filled two sector 12-mm centerpieces with sapphire windows, were loaded with 400 μ l of sample using the final dialysis buffer as reference. The AUC cells, An-50 Ti rotor (Beckman) and interference detector were assembled into the centrifuge before the temperature was equilibrated at least for 1 h at 25 °C. The interference scans of sedimentation were obtained at 50,000 rpm (200,000 $\times$ g). The scan data was used to fit the distribution of the sedimentation

coefficients, $c(s)$, of the samples using Sedfit software (Schuck, 2000). The solvent viscosity and density were set to 0.00896 g/cm/sec and 1.0003 g/ml, respectively, at 25 °C based on previously determined values (Reddick et al., 2007). The partial specific volume of the protein mixture was estimated using Sednterp software (Lebowitz et al., 2002) to be 0.7441 ml/g using Toc34 sequence as an input. The best-fit $c(s)$ distribution was regularized as described previously (Dam and Schuck, 2004). The fractions of Toc34 monomer and dimer were calculated by integrating the area under the $c(s)$ distributions from 2.6 – 3.3 S and 3.3 – 4.4 S for monomer and dimer, respectively. Two separate experiments were performed.

2.18 Immunoblotting

The samples from *in vivo* import, *in vitro* import, and *in vitro* stromal processing assays using YFP tagged proteins were separated on 10 – 15% SDS-PAGE gels. The proteins were transferred to a polyvinylidene fluoride membrane, Immobilon-P (Millipore), by electroblotting at 4 °C in transfer buffer (0.3% Tris base, 1.45% glycine and 20% methanol) running at 24 V for at least 1 h. The membranes were blocked in TBS-T buffer (25 mM Tris-HCl, pH 8.0, 137 mM NaCl, 3 mM KCl and 0.1% Tween-20) containing 3% non-fat milk at room temperature for 1 h or at 4 °C overnight. The primary antibody, rabbit polyclonal anti-green fluorescent protein (GFP) (Jungwirth et al., 2010), was used at 1:5,000 in TBS-T buffer containing 3% non-fat milk. The membranes were incubated with the primary antibody for at least 1 h at room temperature. The membranes were then washed with TBS-T buffer containing 3% non-fat milk, 3 times with 15 times incubation each time. The secondary antibody, goat anti-rabbit IgG HRP conjugated antibody (Chemicon), was used at 1:50,000 in TBS-T buffer by incubating with membrane at room temperature for 1 h. The membranes were washed 3 times for 15 min each using TBS-T buffer. The

Immobilon Western Chemiluminescent HRP Substrate (Millipore) was used for detection. The chemiluminescent signal was detected using the ChemiDoc XRS system (Bio-Rad) using Quantity One software (Bio-Rad).

2.19 Bioinformatic Analysis

2.19.1 Condon optimization

Wild-type aa sequences of TPs of *S. latifolia* ferredoxin CAA26281) and *P. sativum* small subunit of RuBisCO (CAA25390) were codon optimized using Gene Designer software (Villalobos et al., 2006) based on the codon usage of highly expressed genes of *E. coli* (Hénaut and Danchin, 1996). The optimized sequences are shown in Table A1-8. The adaptiveness and codon adaptive index (CAI) values were calculated as previously described (Sharp and Li, 1987). For cloning, the NdeI and XmaI sites were added to the designed sequences. Cloning was performed as in section 2.2.1.

2.19.2 Similarity analysis of forward and reverse TPs

Global pairwise alignment of the TP aa sequences was performed with the Needleman-Wunsch algorithm using Needle program in EMBOSS package (Rice et al., 2000). A series of substitution scoring matrices, BLOSUM45 to 90 were tested. When gap opening and extension penalties were set as default setting, 10 and 5, respectively, BLOSUM55 generated the highest scores. The %identity and %similarity computed using BLOSUM55 were reported.

2.19.3 Generation of *Arabidopsis* TP dataset

For analysis of the N-terminal property of TPs, a dataset of *A. thaliana* TP sequences was generated. TargetP (Emanuelsson et al., 2007) was used to

predict TPs from the *Arabidopsis* genome. Only the sequences predicted to localize to chloroplasts and assigned reliability classes 1 or 2 were collected into the dataset. The TargetP reliability class is based on the difference between the highest and the second highest localization scores. If the difference is greater than 0.8, it is assigned as class 1. If the difference is greater than 0.6 but less than 0.8, it is assigned as class 2. The dataset was further reduced based on the distribution of predicted TP length. Only sequences with predicted lengths from 35 – 71 aa were kept, resulting in a final dataset of 912 sequences (Table A2-1).

2.19.4 Calculation of percentage of uncharged amino acids

To determine the uncharged property of the N-terminal region of TPs, the percentage of uncharged aa in the TP sequences was calculated. We assigned Lys, Arg, Asp, Glu and His as charged aa and other aa as uncharged aa. The percentage of uncharged aa was calculated within a sub-sequence. A window of length L was defined and moved along the whole length of the sequence, a single aa at a time. The calculated percentage in each window was assigned to the aa position at the center. The average percentage of uncharged aa at each position was the average from all of the sequences in the dataset. The calculation was repeated using window lengths L of 5 to 17 aa. A Perl script was written to perform this calculation (Code A4-1). The percentage of uncharged aa data within the first 30 residues was fitted to the inverted Boltzmann sigmoidal model with nonlinear regression using Prism.

2.19.5 Hsp70 binding site prediction

2.19.5.1 Random peptide phage display (RPPD) derived algorithm

An algorithm to predict Hsp70 binding sites was developed previously (Ivey et al., 2000) based on information derived from *E. coli* DnaK interaction

with a random 6-aa peptide-phage display library (Gragerov et al., 1994). For each aa, the ratio of occurrences between DnaK-interacting and non-interacting peptide sequences was measured (Gragerov et al., 1994). These ratios were used as the indices for each aa with the exception of Met, Cys and Glu. These aa were underrepresented in the phage library and the indices of 1 were assigned (Ivey et al., 2000). The Hsp70 binding score was calculated using a 6-aa window as described in an equation below (Ivey et al., 2000).

$$A_n = I_{n-2} \times I_{n-1} \times I_n \times I_{n+1} \times I_{n+2} \times I_{n+3}$$

Where the score at aa position n (A_n) was calculated from multiplication of all of the index scores (I) of each aa within the 6-aa window from position $n-2$ to $n+3$. The indices of aa Ala, Cys, Asp, Glu, Phe, Gly, His, Ile, Lys, Leu, Met, Asn, Pro, Gln, Arg, Ser, Thr, Val, Trp, and Tyr were 0.876, 1.000, 0.871, 1.000, 0.506, 0.567, 0.567, 1.772, 2.025, 2.015, 1.000, 0.754, 0.785, 0.547, 1.489, 1.362, 0.597, 0.800, 1.782 and 0.759, respectively. A Perl script was written based on this algorithm (Code A4-2) to analyze the whole length of TP sequences. The calculating window was moved along the length of TP.

2.19.5.2 Cellulose-bound peptide scanning (CBPS) derived algorithm

Previously, an algorithm to predict Hsp70 binding site was developed from *E. coli* DnaK interactions with a cellulose-bound peptide library (Rudiger et al., 1997b). For each aa, the statistical energy distribution was based on the equation below.

$$\Delta\Delta G_K = -RT \ln (P_b/P_n)$$

Where the statistical energy distribution (ΔG_K) of each aa was calculated from the relative occurrences of that aa in DnaK binding region (P_b) and non-binding region (P_n). The binding score was calculated based on 13-aa window comprised of leaf, core and right regions (Rudiger et al., 1997b). The ΔG_K of each aa in each region was derived separately. And a correction factor was assigned to each position in the window. The score can be calculated based on the equation

$$\begin{aligned}
 S_n = & (0.33 \times L_{n-6}) + (0.66 \times L_{n-5}) + (1.00 \times L_{n-4}) + (1.50 \times L_{n-3}) \\
 & + (1.00 \times C_{n-2}) + (1.00 \times C_{n-1}) + (1.00 \times C_n) + (1.00 \times C_{n+1}) + (1.00 \times C_{n+2}) \\
 & + (1.50 \times R_{n+3}) + (1.00 \times R_{n+4}) + (0.66 \times R_{n+5}) + (0.33 \times R_{n+6})
 \end{aligned}$$

Where the score at aa position n (S_n) was calculated from the summation of the multiplications of correction factors with the ΔG_K of each aa within the 13-aa window from position $n-6$ to $n+6$. The ΔG_K of the left, core and right regions were designated as L , C and R , respectively. The ΔG_K values are listed in a previously published article (Rudiger et al., 1997b). A variant algorithm where the score was only calculated from a 6-aa window ($n-2$ to $n+3$) was used in order to cover a longer sequence area as previously reported (Rudiger et al., 1997b). A Perl script was written based on this algorithm (Code A4-3) to analyze the whole length of TP sequences. The calculating window was moved along the length of TP.

2.19.6 FGLK motif prediction

Previously, a heuristic algorithm was developed by McWilliams to detect the FGLK motif within TPs (Chotewutmontri et al., 2012). Using an 8-aa window, each TP was classified as containing an FGLK motif when the aa sequence in the window satisfied all 5 criteria of Rule 22: (1) contains Phe, (2)

contains Pro or Gly, (3) contains Lys or Arg, (4) contains Ala, Leu or Val and (5) does not contain Asp or Glu. To be able to apply a sliding window prediction method similar to that used in section 2.19.5, a scoring scheme was developed. The criteria (1) to (4) were each given score of 2 and the criterion (5) was given score of 0 when satisfied. The FGLK score at aa position n was calculated from aa sequence from position $n-3$ to $n+4$ by multiplication of all the criteria score. Thus, the maximum score is 16 and the minimum score is 0. The Perl script used to predict FGLK motif based on sliding window calculation is shown as Code A4-4.

2.19.7 Clustering of TPs based on Hsp70 binding patterns

The N-terminal 80 aa sequences of the 208-TP dataset (Lee et al., 2008) were utilized. The RPPD algorithm (2.19.5.1) was used to predicted Hsp70 binding sites producing the Hsp70 binding scores from positions 3 to 77 for each TP. The MATLAB program (MathWorks) was used to cluster the prediction data. The *clustergram* function from the bioinformatics toolbox was applied to cluster data based on the hierarchical clustering method and generate a dendrogram along with a heat map of the clustering. To cluster the TPs according to their Hsp70 binding patterns, Pearson correlation was specified in *clustergram* function to be used in distance matrix calculation. Unweighted pair group method with arithmetic mean (UPGMA) was selected as a method to be used in dendrogram construction. The clustering results are listed in Table A2-2.

2.19.8 Comparison of TP datasets

To compare the TPs in different datasets, a standalone BLAST program (Altschul et al., 1997) version 2.2.25+ was utilized. The TP datasets in FASTA format were converted into different BLAST databases using *formatdb* command. The TP sequences from each TP dataset were searched against each database to

identify the same protein in the databases using *blastp* command. TPs were classified to be the same proteins when *blastp* aligned the proteins with 100% sequence identity over the length of at least 34 aa (the shortest TP length found in the datasets).

2.19.9 Hydrophobicity

The hydrophobicity of the peptides was estimated using an online program ProtScale on ExPASy server (Wilkins et al., 1999). The hydrophobic scale of aa by Kyte and Doolittle (1982) was applied.

2.19.10 Sequence logo

To generate sequence logo, the sequence datasets were submitted to an online program WebLogo (<http://weblogo.berkeley.edu/>) (Crooks et al., 2004).

2.19.11 Evaluation of the N-terminal sequence using the position-specific scoring matrices

In order to study the role of the TP N-terminal domains in Chapter 4, a series of peptides was used to replace the native N-terminal domains. To evaluate if these peptide sequences are similar to the TP N-terminal sequences, the position-specific scoring matrix (PSSM) method was used. This method is widely used in identifying the motifs or sequence patterns within the sequences (Henikoff, 1996).

Generally, the aa sequences containing the motif of interest were aligned. This multiple alignment of the motif of size M residues was then used to generate a PSSM with a dimension of 20 aa x M positions. Each row represents an aa (i) and each column represents an aa position (j) of the motif. The matrix element contains a score ($s_{i,j}$). The simplest type of scoring is to utilize the number of a particular aa found at a specific position of the motif (Hertz and Stormo, 1999).

To identify the motif within a sequence, a sliding window of size M is scanned through the sequence. At each location, the score of the subsequence is calculated from the summation of the score $s_{i,j}$ of each aa in the subsequence (Hertz and Stormo, 1999). Another simple type of scoring used the relative frequency ($f_{i,j}$) of the aa i at position j and the background frequency (p_i) of the aa i (Hertz and Stormo, 1999). The matrix element score is calculated as the log ratio of these frequencies.

$$s_{i,j} = \ln \frac{f_{i,j}}{p_i}$$

The subsequence score (SS) can then be calculated from the summation of the scores associated with the aa in the subsequences (Hertz and Stormo, 1999).

$$SS = \sum_{j=1}^M \sum_{i=1}^{20} s_{i,j}$$

Because the peptides in the series are 8-12 aa long, a TP PSSM corresponding to the TP N-terminal 10 residues was created. The N-terminal sequences of TPs from the 208-TP dataset (Lee et al., 2008) were used without aligning the sequences. Because the first residue is always Met (Figure 4-6A), this position is ignored. While residue 2 had a distinct aa distribution, residues 3-12 had approximately the same distributions (Figure 4-7B). For the first position of the matrix ($j = 1$), the calculated relative frequency of aa at residue 2 position ($f_{i,1}$) indicated that the 208-TP lacked Cys, His, Trp and Tyr at this position. We instead calculated the frequencies at residue 2 using the larger set, the 912-TP dataset (Chotewutmontri et al., 2012). However, we found no Trp at this position. We corrected this missing value by using the value of 0.01 as the number of Trp count at this position. To avoid the same problem, the aa from

residues 3-12 were combined to calculate an averaged frequency ($f_{i, a(3-12)}$), which was used as the frequencies of the matrix positions 2 to 9 (for example, $f_{i, 2} = f_{i, a(3-12)}$). The aa frequencies of UniProt Release 2012_10 were used as the background frequencies (p_i). We used log base 2 of the ratio of $f_{i, j}$ over p_i to calculate the matrix element scores $s_{i, j}$. The PSSM score of a sequence was then calculated from the summation of the $s_{i, j}$ scores from 9 positions corresponding to residues 2 to 10. The frequencies and the $s_{i, j}$ scores of this TP PSSM are reported in Table A3-1. A Perl script was written to perform this calculation (Code A4-5).

We extended this N-terminal sequence analysis using PSSM to the mitochondrial targeting peptides (mTPs) and secretory pathway signal peptides (SPs), however, this time the N-terminal 30 residues were used in creating the PSSMs. Based on the observed distributions (Figure 4-7B), the TP PSSM was expanded to include the frequencies from residues 13 to 19 ($f_{i, 12}$ to $f_{i, 18}$) and the averaged frequency of residues 20-30 ($f_{i, a(20-30)}$). The averaged frequency $f_{i, a(20-30)}$ was used as the values for $f_{i, 19}$ to $f_{i, 29}$. Note that the missing aa values were corrected by using the value of 0.01 as the number of count. The $s_{i, j}$ scores of the extended TP PSSM are reported in Table A3-2. For mTP and SP PSSMs, the sequences from TargetP training set (Emanuelsson et al., 2007) were used. In contrast to the TP PSSM that utilized two averaged frequencies, the mTP and SP PSSMs only contain the frequencies derived from individual position from residues 2 to 30. The $s_{i, j}$ scores of the mTP and SP PSSMs are reported in Table A3-3 and A3-4, respectively. To calculate the PSSM scores from the extended TP PSSM, the mTP, and SP PSSMs, a Perl script was written to perform this function (Code A4-6). The three scores were compared using Excel (Microsoft). The protein was predicted to localize to the location that gave the highest score. However, if the highest score is 0 or less, the protein was predicted to localize to other location than plastids, mitochondria and secretory pathway.

2.19.12 Hsp70-FGLK spacer length distribution and amino acid composition

The Hsp70-FGLK spacer length(s) of a TP was defined as the distance(s) in aa between the residue with the highest RPPD score within the N-terminal 15 aa (Hsp70 peak) to the residue(s) with the maximal FGLK score (FGLK peak). To identify the Hsp70 peaks, Excel (Microsoft) was used to extract the residue positions from the predicted RPPD scores of the TPs belonging to the cluster groups 1-3 (Figure 4-1, Table A2-2) of the 208-TP dataset (Lee et al., 2008). Each TP gave a single Hsp70 peak. To identify the FGLK peak(s), every residue containing the FGLK score higher than the cutoff value was extracted using a Perl script (Code A4-7). The FGLK motif prediction (section 2.19.6) produced the scores in value of 2, 4, 8 or 16. We set the cutoff value at 15 to extract the position(s) with the maximal FGLK score. Each TP may produce more than one position depending on the number of residues having the maximal FGLK score. Note that only the N-terminal 80-aa sequences were used in FGLK prediction.

Distance between the Hsp70 and FGLK peaks were measured using a Perl script (Code A4-8). The generated result files show the residue positions of the Hsp70 and FGLK peaks. Because some of the FGLK peaks were clustered together forming a plateau, Excel (Microsoft) was used to reduce the multiple distances belonging to the same plateau to a single distance value corresponding to the distance to the FGLK peak at the middle position on the plateau. Prism software (GraphPad) was used to construct the histogram of the Hsp70-FGLK distances and fitted to the Gaussian distribution. The averaged spacer distance was determined to be about 24 aa.

To determine aa composition of the Hsp70-FGLK spacers, the sequences of the Hsp70-FGLK spacers were extracted from the TP sequences. However, only the spacers with the closest length to the averaged spacer distance of 24 aa were used. In addition, 4 residues from both N- and C-termini of the sequences were

removed because they represented the shoulders of Hsp70 and FGLK peaks. The aa distributions within whole spacer sequences or part of sequences were calculated as described in 2.19.15.

2.19.13 Random sequence generator

In order to generate novel Hsp70-FGLK spacer sequences, a random sequence generator was written as Perl script (Code A4-9). The generator utilizes user-defined aa frequency distribution and generated random sequences with length and total numbers as specified by the user. To produce the random sequence, the script first generates a pool of 3,000 aa with the same aa frequencies as supplied by user. Then the generator randomly selects an aa from the pool to form a sequence. The selection is repeated until the defined length is reached. The process is repeated until the number of sequences reached the number specified by user.

2.19.14 Hsp7-FGLK spacer design

In attempts to generate novel Hsp70-FGLK spacers, a pool of 400 random sequences of 26 aa long was generated based on the aa distribution of Hsp70-FGLK spacer determined from section 2.19.12 using the random sequence generator (section 2.19.13).

In order to minimize the effect of additional Hsp70 and FGLK domains within the spacer sequence, we screened the sequences with Hsp70 and FGLK prediction programs (sections 2.19.5 and 2.19.6). First, the mutant TP sequences were generated from the SSF sequence by replacing the native spacer with the random sequences. This was done by using a Perl script program (Code A4-10). The mutant sequences were then submitted for Hsp70 and FGLK predictions. Three mutant sequences lacking positive Hsp70 and FGLK domains within the

spacer regions were selected which contain the random sequence numbers 92, 228 and 296.

2.19.15 Amino acid distribution

To count the number of each aa presented in the protein sequence dataset and calculate the frequency of each aa, a Perl script was written to perform this calculation as shown as Code A4-11.

Chapter 3

Differential Recognition of Transit Peptide during Binding and Translocation into Plastids

3.1 Disclosure

Most of the work reported in this chapter has been published in a research article by Chotewutmontri et al. (2012). Some of the methods developed here have also been published as part of a method chapter by Reddick et al. (2008). The results generated solely by other authors in the published articles are omitted from the result section but are included in the discussion to clarify the findings.

3.2 Abstract

Bioinformatic and proteomic analyses provide thousands of predicted TP sequences, which show low sequence similarity. How the common chloroplast translocon components recognize these diverse TPs is not well understood. Previous results support either sequence-specific or physicochemical-specific recognitions. To further address this question, a reverse sequence approach was utilized such that the reverse TPs contains the same aa composition as wild-type TP but lack similar sequence motifs. Using both native and reverse TPs of the two well-studied precursors, the small subunit of RuBisCO, and ferredoxin, we explored these two modes of recognition. We found that reverse TPs behaved similar to wild-type TPs during binding but failed to support protein translocation. We further showed the importance of the N-terminal 10-aa domain

of TPs in governing protein translocation into plastids. We linked these N-termini to the Hsp70 interacting domain and proposed a model of TP architecture based on this finding.

3.3 Introduction

The ability of plastids to import precursor proteins post-translationally from the cytosol has been known for over 30 years (Chua and Schmidt, 1979; Dobberstein et al., 1977). The key to this process is the role of an N-terminal extension, known as the transit peptide (TP), which directs the precursor to the plastid membrane and through the translocons at the outer and inner chloroplast envelope membranes (TOC/TIC) (Bruce, 2000; Bruce, 2001). Analysis of multiple plant and algal genomes using various TP identification tools, indicates that the number of nuclear-encoded precursors ranges from about 2,100 in *Arabidopsis thaliana* to as high as about 4,800 in rice (*Oryza sativa*) (Richly and Leister, 2004). Despite this large number of available sequences, fundamental understanding of how TPs function is still lacking.

Early analysis suggested that TPs might be composed of distinct homology blocks that share limited sequence similarity (Karlin-Neumann and Tobin, 1986). However, this hypothesis was challenged and replaced by a loose structural organization with three identifiable regions (von Heijne et al., 1989). Multiple efforts using mutagenesis (Lee et al., 2002; Pilon et al., 1995), deletion (Kindle, 1998; Kindle and Lawrence, 1998; Rensink et al., 1998; Rensink et al., 2000), Ala scanning (Lee et al., 2008; Lee et al., 2006; Lee et al., 2009a; Lee et al., 2002), domain swapping (de Castro Silva Filho et al., 1996; Lee et al., 2009a; Smeekens et al., 1986), and the use of synthetic peptides (Perry et al., 1991; Pinnaduwege and Bruce, 1996; Schnell et al., 1991) have investigated the structure and function of only a few TPs in detail. However, these results are not extendable to other TPs based on sequence analysis, and the elucidation of common TP

functional domains remains enigmatic. Although earlier attempts to identify homology blocks failed due to the high degree of sequence variation, it is still possible that TPs may contain a conserved motif or nonlinear peptide pattern that may provide some common mode of recognition (Lee et al., 2006). Moreover, a systematic approach involving *in vivo* targeting analysis indicates that individual aa do not contain specific targeting information, but the overall context of the aa sequence is critical for targeting to the chloroplast (Lee et al., 2002). Recent efforts to identify any universal signature motif in 208 experimentally confirmed TPs have not been fruitful, and it was concluded that these TPs are highly dissimilar (Lee et al., 2008). However, when these authors used a bioinformatics-based approach to pregroup TPs into seven subgroups, one or more conserved motifs were identified within a given subgroup but were not universal (Lee et al., 2008).

This suggests that TPs do not share any consensus motifs, yet each TP may contain different functional motifs that facilitate targeting and import. In light of the conserved nature of translocon components (Reumann and Keegstra, 1999) and the high fidelity of protein targeting *in vitro* and *in vivo*, it is difficult to reconcile how individual TPs can engage a common set of translocon components without some unifying information encoded within the TP. It is possible that sequence information does not define a TP, but rather the physicochemical properties of the TP determine its targeting activity. This may explain why TP prediction algorithms function in the absence of any detectable sequence similarity. These physicochemical properties may be environmentally sensitive and/or context specific, behaving differently as a function of pH, in a membrane-like environment, or upon receptor binding. One example of this is the tendency of TPs to convert from a random coil in an aqueous environment (von Heijne and Nishikawa, 1991) to an α -helix in the presence of membranes or membrane-mimetic environments (Bruce, 1998; Krimm et al., 1999; Wienk et al., 2000). Finally, it is possible that TP interaction with different components of the

TOC and TIC as well as the stromal-localized components, such as stromal processing peptidase (SPP) and molecular chaperones, uses multiple mechanisms of recognition ranging from general physicochemical properties to specific sequence recognition.

Here, we attempted to differentiate the role of TP sequence-specific contributions from the physicochemical properties using TP sequences that have been reversed with respect to their N- to C- sequence, termed reverse peptides. These reverse TPs share no more similarity to their parent sequences than any random sequence (Haack et al., 1997); however, they share with their parents many identical properties, including (1) ratio of hydrophobic/hydrophilic aa, (2) a global aa composition, (3) chirality, (4) spacing of their constituent aa, (5) placement of secondary structures, and (6) potential mirroring of three-dimensional structure (Battistutta et al., 1994; Guptasarma, 1992; Lacroix et al., 1998). Thus, they contain the identical aa composition and its associated physicochemical properties yet are sequence divergent. This inherent property has attracted considerable interest in using reverse peptides to study various structure function relationships of peptides/proteins, including antimicrobial peptides (Pellegrini and von Fellenberg, 1999) and Leu zippers (Holtzer et al., 2000). They have also been used to examine protein folding (Lacroix et al., 1998; Olszewski et al., 1996) and antibody recognition (Benkirane et al., 1995; Guichard et al., 1994).

We developed a series of assays to evaluate how the well-studied TPs of the small subunit of ribulose-1,5-bis-phosphate carboxylase/oxygenase (RuBisCO) and ferredoxin and their reverse peptides interact with different components of the plastid protein import machinery. These assays include *in vitro* analyses of the interaction of forward (native) and reverse peptides with the isolated cytosolic receptor domain of the GTPase Toc34, the stromal molecular motor ATPase Hsp70, and the SPP. We perform *in organello* analyses of isolated chloroplasts' ability to bind and import forward and reverse peptides, and carry

out *in vivo* analyses of onion (*Allium cepa*), *Arabidopsis*, and tobacco (*Nicotiana benthamiana*) cells' ability to sort and deliver forward and reverse peptides into plastids. Interestingly, we see that certain steps in the import process can recognize both the TP and its reverse peptide. Other steps are highly selective, such as *in vitro* and *in vivo* translocations. To reconcile these results, we further tested the requirements for the N-terminal sequence to be uncharged and largely nonpolar. This requirement seems to be a key determinant in the ability of a given sequence (forward or reverse) to mediate translocation. These results are discussed in light of a possible general mechanism of TP recognition given the lack of sequence similarity that is so pervasive in the plastid targeting sequences.

3.4 Results

3.4.1 Production of forward and reverse peptides

To characterize the biophysical, biochemical, and targeting activities of the TPs and their reverse peptides, an *E. coli* expression system that allows the production of these peptides without an attached epitope tag was utilized. This system used the self-cleavage activity of intein. Due to the heterologous nature of this system, the genes were first codon optimized. The DNA sequences of forward (native) and reverse peptides of TPs of the small subunit of RuBisCO (SSF and SSR) and ferredoxin (FDF and FDR) are shown in Table A1-8. The relative adaptiveness plots shown in Figure 3-1 indicates the use of *E. coli* preferred codons in the optimized DNA sequences. The optimized sequences have higher codon adaptive indices (CAI), the geometric mean of the relative adaptiveness, at about 0.8 whereas the native DNA sequences have average CAIs of about 0.36 (Table 3-1).

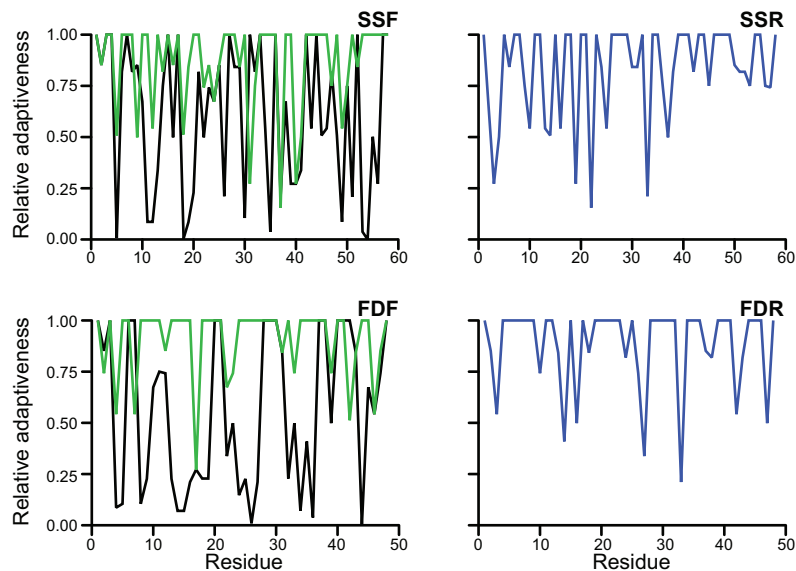


Figure 3-1. Relative Adaptiveness Plots of the Forward and Reverse Peptides

The relative adaptiveness was calculated based on a method by Sharp and Li (1987) using the codon usage table of the highly expressed genes in *E. coli* reported by Hénaut and Danchin (1996). Black and green lines represent the relative adaptiveness prior to and after codon optimization, respectively. Since the reverse sequences do not exist in nature, there is no black trace.

Table 3-1. Codon Optimization Indices of the Synthetic Peptides

DNA sequences	<i>E. coli</i> codon usage table			Mean	SD
	Hénaut and Danchin (1996).	Carbone et al. (2003)	Nakamura et al. (2000)		
Wild-type FDF	0.326	0.168	0.587	0.360	0.211
Optimized FDF	0.882	0.847	0.910	0.880	0.032
Optimized FDR	0.847	0.752	0.911	0.837	0.080
Wild-type SSF	0.337	0.214	0.542	0.365	0.166
Optimized SSF	0.823	0.746	0.853	0.808	0.055
Optimized SSR	0.781	0.678	0.832	0.764	0.078

The codon optimized synthetic DNAs were cloned into pTYB2 vector and verified by sequencing. The aa sequences of the peptides are presented in Figure 3-2A. The peptides were expressed in *E. coli* ER2566 cells, purified by chitin-affinity chromatography, intein-cleaved, eluted from the column and lyophilized. Figure 3-2B shows the purification profile of FDF peptide as a representative. The purity of the four peptides was confirmed by SDS-PAGE and matrix-assisted laser desorption/ ionization time-of-flight mass spectroscopy (MALDI-TOF MS) (Figure 3-2C and D). The MALDI spectrum of SSF shows a major species at 5901.49 m/z that corresponds to the mass of SSF from aa 2-60 (Figure 3-2D). The theoretical average mass of SSF₂₋₆₀ is 5903.81 D. The peak represents the ⁺¹ charge state that corresponds to a processed peptide missing the N-terminal Met residue, presumably due to the activity of the *E. coli* methionine aminopeptidase. In addition, there are multiple peaks that correspond to different levels of Met oxidation of SSF₂₋₆₀ following the major peak. As shown in Figure 3-2D for SSR, the major species at 5942.88 m/z corresponds to the ⁺¹ charge state of SSR from aa 1-58 with 4 Met oxidations. The theoretical average mass is 5944.83 D. The other peptides yielded similar spectra to confirm their sequences (data not shown).

3.4.2 Similarity analysis of forward and reverse peptides

Although this study investigates two of the best-characterized TPs from the small subunit of RuBisCO and ferredoxin, which are both localized in the stroma, highly abundant and associated with photosynthesis, they have very limited sequence similarity. In fact, SSF shares 21.2, 12.0, and 2.2% identity and 42.4, 20.0, and 10.9% similarity with SSR, FDF, and FDR, respectively. Likewise, FDF shares 14.3, 12.0, 2.2% identity and 42.9, 20.0, 10.9% similarity with FDR, SSF, and SSR, respectively.

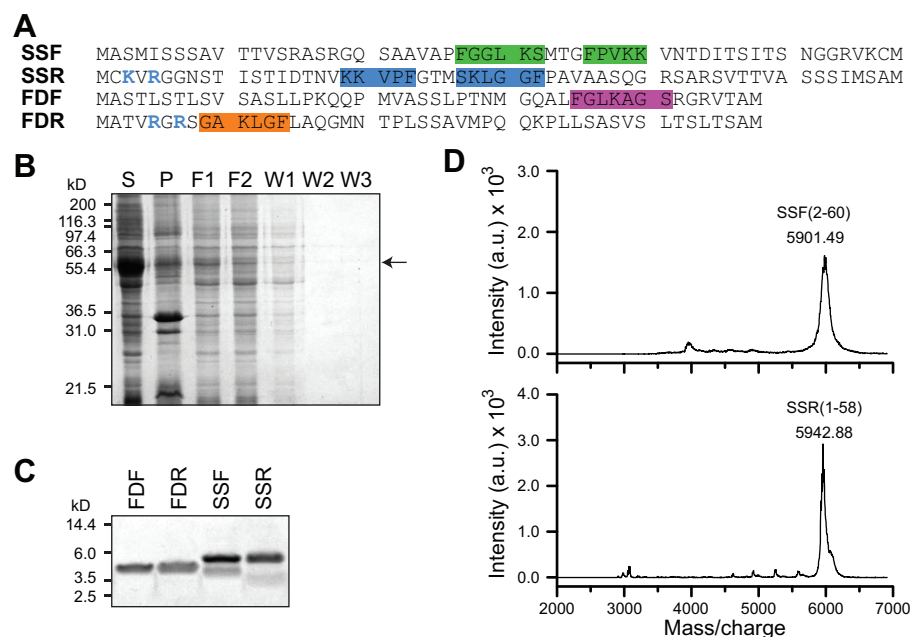


Figure 3-2. Purification of Forward and Reverse Peptides

(A) The aa sequences of SSF, SSR, FDF and FDR. The FGLK motifs are highlighted. The charged aa within the N-terminal 10 aa are colored blue. (B) SDS-PAGE shows the purification profile of FDF. The FDF-intein-chitin binding domain fusion protein (62.5 kDa) is indicated with arrow. S, soluble fraction; P, pelleted fraction; F1 and F2, soluble fraction after flow through the chitin column one and twice; W1, W2 and W3, the wash fractions 1, 2 and 3. (C) SDS-PAGE shows the purified peptides. (D) MALDI-TOF spectra of SSF and SSR peptides.

Although the forward and reverse peptides have identical physicochemical properties, they share very little sequence similarity and therefore any similarity in activity of the two TPs (SSF and FDF) must be based on properties beyond simple sequence similarity.

3.4.3 Bioinformatic analysis of TP domains

Despite the failure of bioinformatic algorithms to identify universally conserved motifs within TPs, there have been two short domains identified as highly characteristic of TPs. One is a short uncharged N-terminal segment that has been observed in most chloroplast TPs (von Heijne et al., 1989). This domain has also been suggested to be capable of functioning as a strong Hsp70-binding domain and is possibly a key to the formation of translocation intermediates via its recognition by IMS or stromal Hsp70s (Ivey and Bruce, 2000; Ivey et al., 2000; Pilon et al., 1995; Zhang and Glaser, 2002). To verify this N-terminal property of TPs, we have performed an analysis of this domain on a dataset of the 912 most confidently predicted TPs from the *Arabidopsis thaliana* genome (Figure 3-3 and Table A2-1). Only the TargetP predicted *Arabidopsis* TPs with reliability classes 1 and 2 were used (see method section 2.19.3 for detail). The percentage of uncharged aa within a specific residue length window between 5-17 was calculated along the length of TPs. The values shown were averaged across the dataset. Regardless of window size, the percentage of uncharged aa shows the transition from highly uncharged at almost 91% at the N-terminus to moderately uncharged at about 78% at the C-terminus confirming an uncharged bias of the N-terminus. When data within the first 30 aa was fitted to a sigmoidal curve, the transition point was determined to be 15.23 indicating the border of N-terminal uncharged region is within the first 15 aa; the transition actually starts at around residue 10 (Figure 3-3, inset).

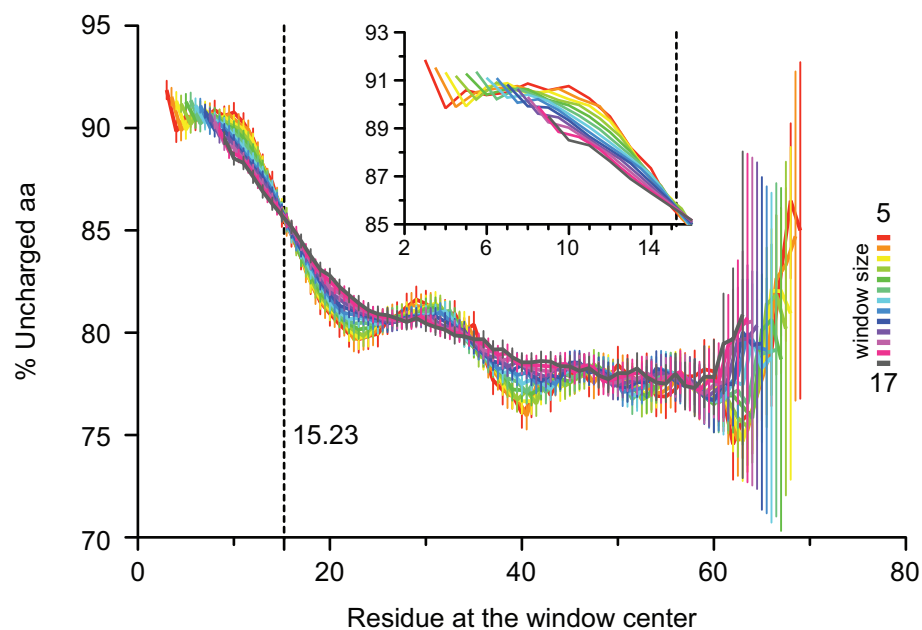


Figure 3-3. Analysis of the N-terminal Uncharged Domain in TPs

The percentage of uncharged aa was calculated from the dataset of confidently predicted TPs from *Arabidopsis thaliana*. $n = 912$. Means $\pm$ SE are shown. Inset shows zoom-in without error bars.

A second semi-conserved TP motif was first observed by Karlin-Neumann and Tobin (1986) and was suggested by Pilon and coworkers (1995) to be involved in the chloroplast recognition of TPs. This group identified a loose FGLK motif that was found at least once in each of 27 characterized TPs. Taking this analysis further, McWilliams (Chotewutmontri et al., 2012) developed a heuristic approach to identify the FGLK motif. SSF has two of these motifs while FDF has only one (Figure 3-2A). Interestingly, it is this region in SSF that shows the most sequence similarity with its reverse version (53.3% identity and 66.6% similarity) suggesting that evolution provided a motif with some targeting activity independent of the orientation of binding.

3.4.4 Analytical ultracentrifugation analysis of Toc34 with the peptides

One of the most specific and potentially mechanistic roles of TPs identified to date has been their ability to function as GTPase activating proteins (GAPs) and stimulate the GTP hydrolysis of the TOC receptor GTPases (Becker et al., 2004b; Jelic et al., 2003; Oreb et al., 2011; Reddick et al., 2007). As previously reported by our laboratory, SSF stimulates GTP hydrolysis of the cytosolic G domain of pea Toc34 (psToc34G) in a GAP-like manner, and does not function as a guanine nucleotide-exchange factor to modulate the rate of nucleotide exchange (Reddick et al., 2008; Reddick et al., 2007). Although this activity has also been observed for several peptides, its mapping has only been refined to the C-terminal 26 aa of SSF (Schleiff et al., 2002).

A second specific interaction of TPs has been their ability to disrupt the stability of the psToc34G homodimer. The concentration dependent monomer-dimer equilibrium of psToc34G has been well documented (Koenig et al., 2008a; Reddick et al., 2007; Sun et al., 2002; Weibel et al., 2003; Yeh et al., 2007). Although the specific mode of binding is not known, it is clear that TP

interaction with psToc34G dimer shifts the equilibrium towards a monomer. A dynamic equilibrium exists between the 3.0 S monomeric and 3.4 S dimeric species of purified psToc34G in solution when analyzed by analytical ultracentrifugation (AUC) as shown in Figure 3-4B and C, black line. Binding of peptides increases the sedimentation coefficient of psToc34G homodimer to greater than 3.4 S. Analysis of the areas under the curves of monomer and dimer peaks reveal that addition of SSF disrupts the dimer and increases the relative amount of psToc34G monomer (Figure 3-4B, green versus black lines and C). While quantitatively not to the same degree as SSF, SSR also stimulates the dimer to monomer transition of psToc34G (Figure 3-4A, blue line and B). SSF stimulates what is effectively about 50% monomer-dimer equilibrium to shift to essentially a 75/25% monomer-dimer distribution whereas the addition of SSR results in only a 60/40% monomer-dimer distribution (Figure 3-4C). Thus, reversing SSF to SSR only slightly impairs the wild type *in vitro* dimer disruption activity. This indicates that the isolated TOC component psToc34G is able to bind and be stimulated by both forward and reverse TPs. Interestingly, neither FDF nor FDR bias the monomer-dimer equilibrium of psToc34G (Figure 3-4C). The relationship between how TPs stimulate GTP hydrolysis and how this activity correlates with the disruption of psToc34G homodimer is unclear; however, the fact that the ferredoxin TPs contain only one FGLK motif may suggest that a very high local concentration of the two FDF/FDR peptides would be required to be functionally equivalent to the apparent concentration of the two tethered motifs found within a single SSF/SSR peptide. Nevertheless, no attempt was made to test the TP concentration dependence of these observations.

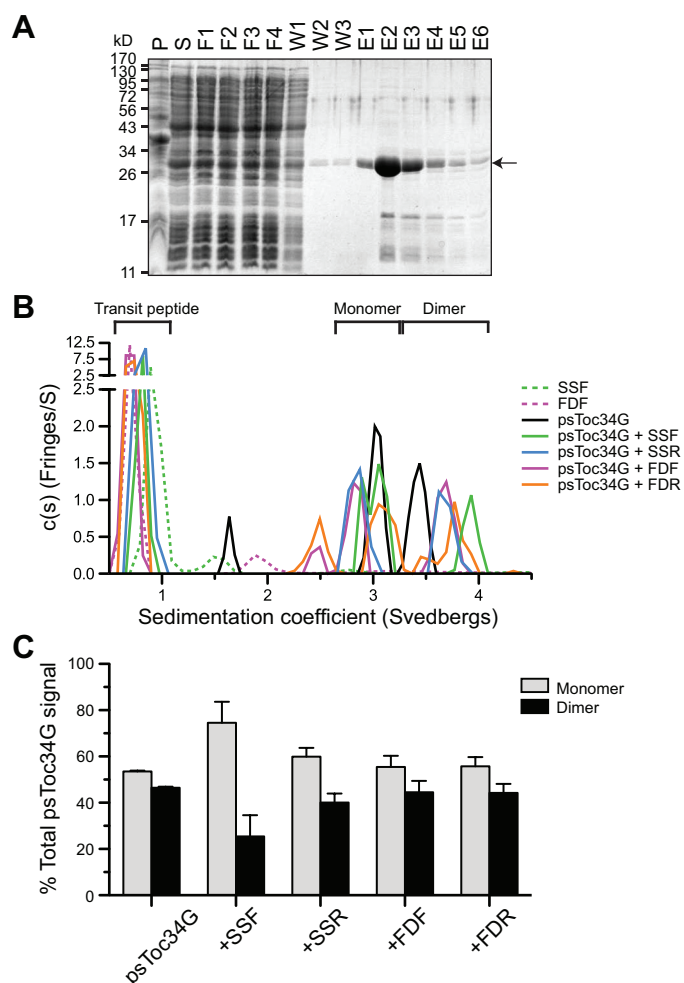


Figure 3-4. Effect of the Peptides on psToc34G Monomer:dimer Ratio

(A) Purification profile of psToc34G. E2 fraction was used in the assays. P, pellet fraction of the cell lysate; S, the soluble fraction; F1-4, soluble fractions after flow through 1-4 times over the column; W1-3, wash fractions 1-3; E1-6, elution fractions 1-6. (B) AUC analysis of psToc34G monomer-dimer equilibrium at 13.5 μ M in the absence and presence of the peptides at 135 μ M. (C) Quantitation of the monomeric and dimeric psToc34G species with and without various peptides. $n = 2$. Means $\pm$ SD are shown.

3.4.5 Development of a liquid scintillation-based *in vitro* chloroplast protein competitive binding assay

During chloroplast protein import, it is possible to trap an intermediate state specifically associated with the chloroplast, but not yet internalized (Olsen and Keegstra, 1992). This intermediate can be observed using radioactivity or by various fluorescence assays such as flow cytometry or confocal microscopy that permit the quantification and imaging of the bound precursor (Subramanian et al., 2001). In the past, our laboratory has employed quantitative import competition assays which used competitive inhibitors to determine specific inhibitory values of the import process such as the half maximal inhibitory concentration (IC_{50}) (Dabney-Smith et al., 1999). The method to determine the specific inhibitory values of the binding process is developed here. The radiolabeled precursor of small subunit of RuBisCO (^{35}S -prSSU) was employed in the assays. ^{35}S -prSSU was labeled *in vivo* and purified from inclusion bodies (Figure 3-5A and B).

First, the binding time course was determined in the presence of low levels of ATP which allows the formation of the early import intermediate but blocks the translocation (Inoue and Akita, 2008a; Kessler et al., 1994; Kouranov and Schnell, 1997; Olsen and Keegstra, 1992; Perry and Keegstra, 1994; Young et al., 1999). Binding of 100 nM ^{35}S -prSSU to the chloroplasts show the initial accumulation before decreasing and reaching equilibrium after 30 min (Figure 3-5C). Higher levels of ATP sustain the ^{35}S -prSSU accumulations longer, indicating the involvement of ATPase in the formation of intermediates. None of the binding conditions produce detectable levels of the import-processed mature form of ^{35}S -prSSU except for binding at 100 μM ATP for 15 and 24 min (data not shown). Because this binding assay was performed under light conditions permitting ATP synthesis by photosynthesis, it is possible that with a longer

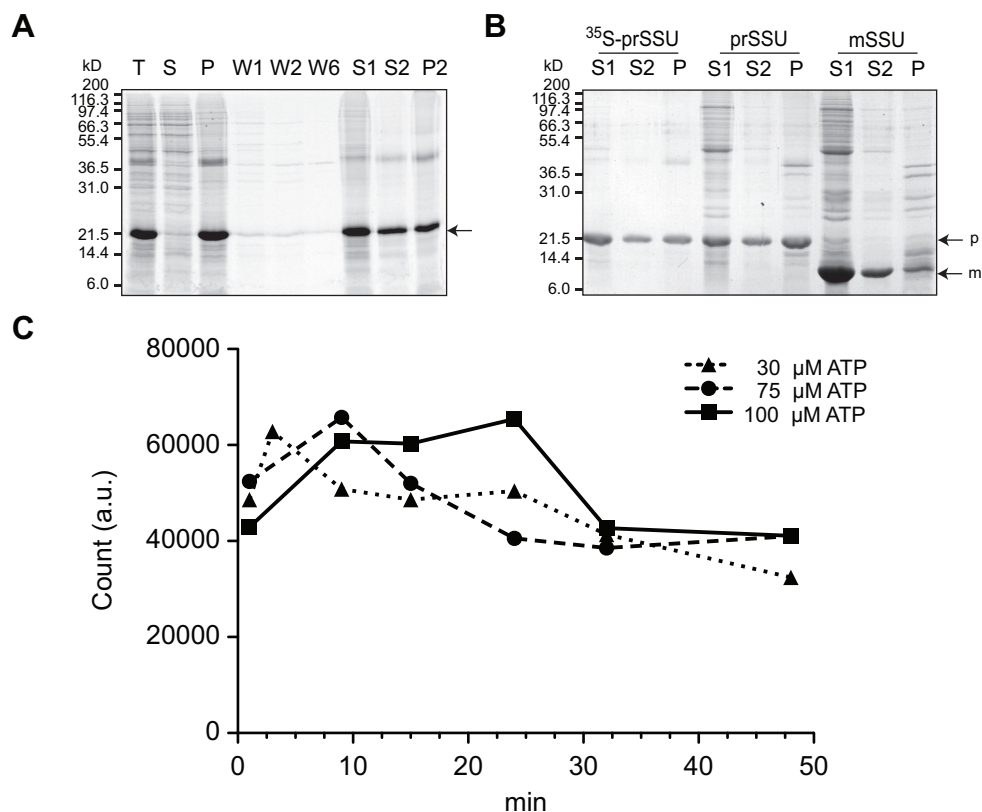


Figure 3-5. Purification of ^{35}S -prSSU, prSSU and mSSU, and Time Course Analysis of ^{35}S -prSSU Binding to the Chloroplasts

(A) Autoradiograph of ^{35}S -prSSU purification profile separated by SDS-PAGE. S2 fraction was used in the assays. Arrow indicates ^{35}S -prSSU. T, total protein; S, soluble fraction; P, pelleted fraction; W1, W2 and W6, the wash fractions 1, 2, and 6; S1 and S2, 8 M urea soluble fractions 1 and 2; P2, pelleted fraction of S2. (B) Coomassie stained SDS-PAGE gel of 8 M urea solubilized ^{35}S -prSSU, prSSU and mSSU. The S2 fractions were used in the assays. Arrows with p and m indicate prSSU and mSSU, respectively. S1 and S2, 8 M urea soluble fractions 1 and 2; P2, pelleted fraction of S2. (C) Time course of 100 nM ^{35}S -prSSU binding to the chloroplasts at different levels of ATP.

reaction time, the amount of synthesized ATP is high enough to facilitate translocation, which in turn reduced the accumulation of ^{35}S -prSSU.

Next, two pre-treatments that deplete the internal and external ATP of the chloroplasts were tested. The chloroplasts were incubated at room temperature for 10 min with or without 5 U/mg chlorophyll a pyrase. After the treatments, chloroplasts were re-isolated before used in the assays, which were performed under dim light. The binding of 100 nM ^{35}S -prSSU in different levels of ATP was measured after 10 min of incubation. Dark-treated chloroplasts show saturation of binding starting at 50 μM ATP (Figure 3-6A). The translocation of ^{35}S -prSSU was also observed at 150 and 450 μM ATP (data not shown). Surprisingly, pyrase-treated chloroplasts show decreasing levels of bound ^{35}S -prSSU with increasing amounts of ATP. Since the aim was to promote binding, pyrase was not used in further assays. In addition, further binding assays were performed under dim light to prevent ATP synthesis.

Because GTP has been reported to promote binding (Inoue and Akita, 2008a; Young et al., 1999), the effects of various nucleotides on the binding were examined. Dark-treated chloroplasts were incubated with 100 nM ^{35}S -prSSU for 15 min in presence of 100 μM ATP, or 500 μM GDP, GTP, GMP-PNP, XTP or GTP γ S (Figure 3-6B). Both GTP and GTP γ S at 500 μM seem to promote more binding than the other nucleotides. But because the low level of ATP at 100 μM also promotes binding almost the same level as 500 μM GTP, this condition is preferred.

A homologous competitive binding assay was performed using dark-treated chloroplasts in a reaction containing 100 nM ^{35}S -prSSU with various concentrations of prSSU (Figure 3-5B) from 0.25 to 3 mM in the presence of 100 μM ATP. After incubating for 10 min at room temperature under dim light, the level of bound ^{35}S -prSSU was analyzed. The results show that at the highest prSSU concentration, the binding of ^{35}S -prSSU decreased to 42% (data not shown) indicating a high level of non-specific binding. To reduce non-specific

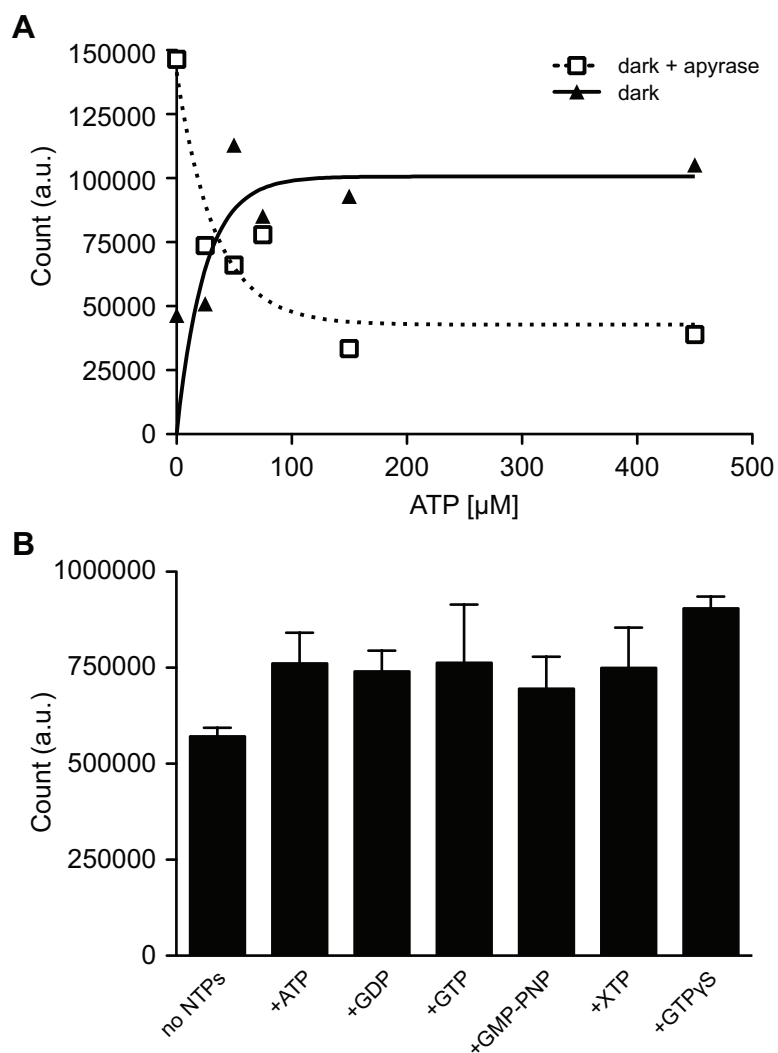


Figure 3-6. Binding of ^{35}S -prSSU After Pre-treatments and Effect of Nucleotides on Binding

(A) Binding of 100 nM ^{35}S -prSSU to pre-treated chloroplasts for 10 min. The chloroplasts were incubated at room temperature with or without 5 U/mg chlorophylls apyrase. (B) Binding of 100 nM ^{35}S -prSSU to dark-treated chloroplasts for 15 min under dim light in presence of 100 μM ATP, or 500 μM GDP, GTP, GMP-PNP, XTP or GTP γ S or absence of nucleotides. $n = 3$. Means $\pm$ SE are shown.

binding, binding assays were performed in the same settings with various amounts of DTT and BSA. The observed efficacies of prSSU in competing with ^{35}S -prSSU were increased with higher levels of DTT and BSA (Figure 3-7).

Based on the effects of pre-treatment, ATP, DTT and BSA determined above, binding assay conditions were adjusted. Binding assays were performed using chloroplasts isolated from plants at the end of their 14-h dark cycle instead of dark-pretreatment for simplicity. The chloroplasts were kept under dim light throughout the assay until mixing with 2xSSB. To minimize non-specific binding, the levels of 100 μM ATP, 10 mM DDT and 1% BSA were utilized. These conditions were tested by performing homologous competitive binding assays. Two concentrations of ^{35}S -prSSU of 30 and 100 nM were used. Figure 3-8A shows representative autoradiographs of SDS-PAGE gels of the binding reactions. The employed binding conditions are able to prevent the import of ^{35}S -prSSU. The non-specific bindings were reduced to 11% and 36% in 30 and 100 nM ^{35}S -prSSU, respectively (Figure 3-8B, Table 3-2). In addition, the ability to prevent translocation prompted us to develop a liquid scintillation-based assay that would be more rapid and robust than the traditional SDS-PAGE autoradiography assay. For comparison, scintillation counting and autoradiography were used to determine the equilibrium dissociation constant (K_d) of prSSU (Figure 3-8B). The scintillation-based assay yields nearly the same K_d as the SDS-PAGE autoradiography-based assay at 153.8 and 153.1 nM, respectively (Table 3-2). Thus, the separation by SDS-PAGE is not required and scintillation counting can be used to precisely and quickly determine the results of this assay in a quantitative manner.

3.4.6 *In vitro* competitive binding assay of the peptides

Using this more rapid binding assay, the ability of SSF, SSR, FDF and FDR to function as competitive inhibitors in binding was examined. As expected,

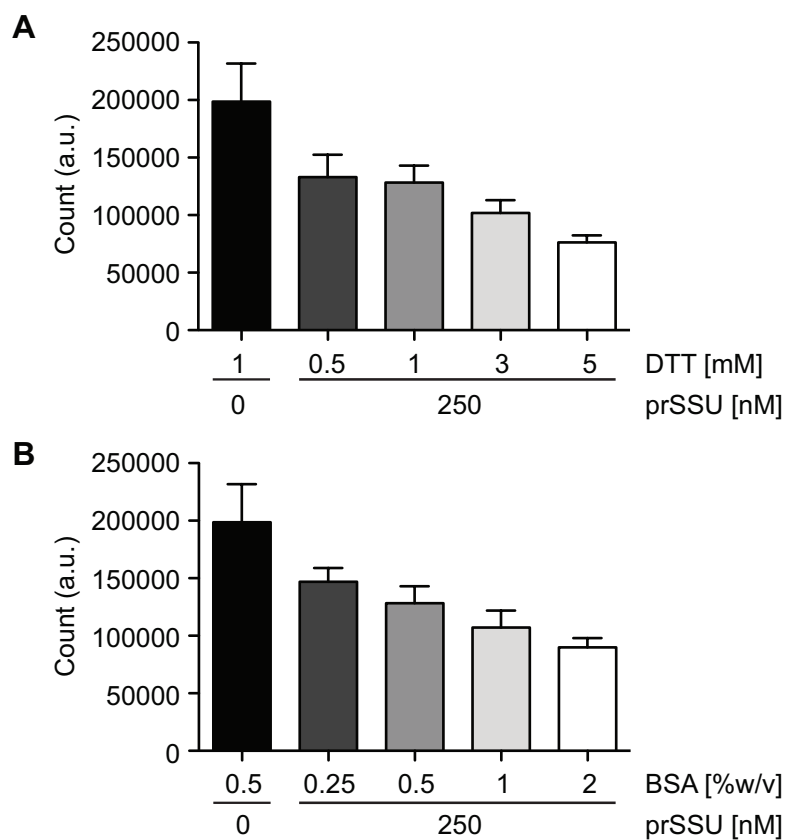


Figure 3-7. Effect of DTT and BSA on Binding of ^{35}S -prSSU

(A) Binding of 100 nM ^{35}S -prSSU in the reactions containing 0.5% BSA with various amounts of DTT and prSSU as indicated. $n = 3$. Means $\pm$ SE are shown.

(B) Binding of 100 nM ^{35}S -prSSU in the reactions containing 1 mM DTT with various amounts of BSA and prSSU as indicated. $n = 3$. Means $\pm$ SE are shown.

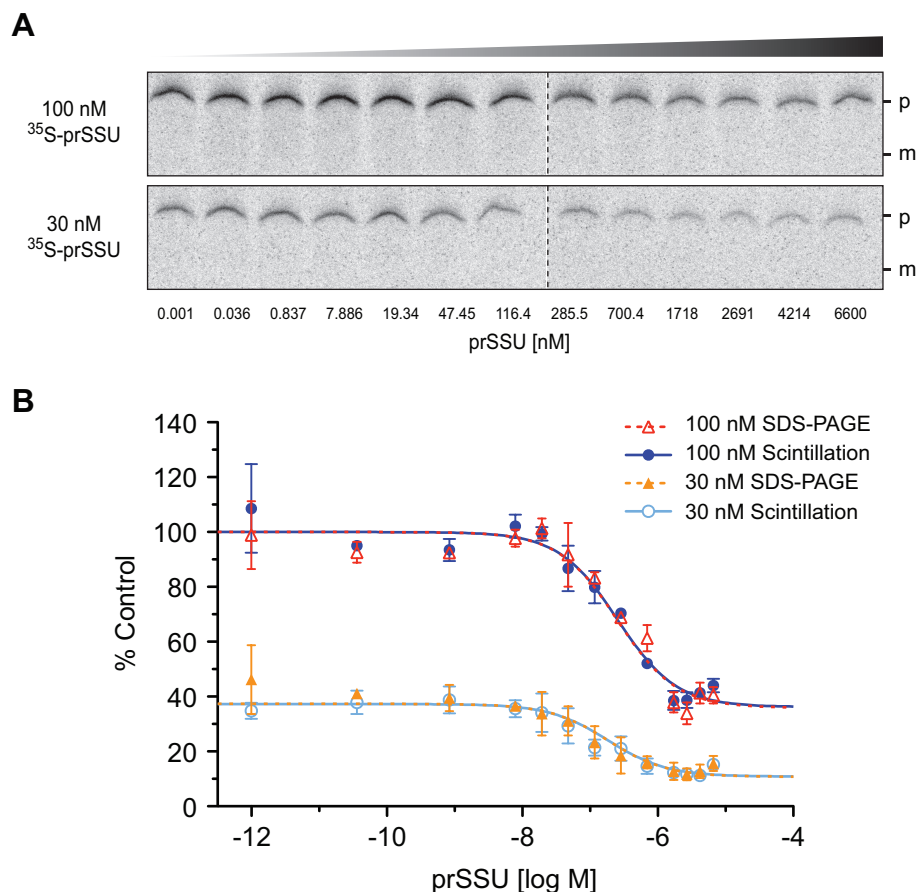


Figure 3-8. Homologous Competitive Binding of prSSU to the Chloroplasts

(A) Representative autoradiographs of SDS-PAGE gels from binding assays using 30 and 100 nM ^{35}S -prSSU in the presence of prSSU competitor as indicated. Only bound precursors (p) were detected. Import-processed mature proteins (m) are undetected. (B) Homologous binding assays of prSSU measuring by scintillation counting (solid lines) and SDS-PAGE autoradiography (dashed lines). To improve fitting confidence, two concentrations of ^{35}S -prSSU, 30 nM (lower traces) and 100 nM (upper traces), were globally fitted together. Fittings with data collected from autoradiography of SDS-PAGE gels were comparable to those collected using scintillation counting. Data is presented as means $\pm$ SE from two independent sets of assays.

Table 3-2. Curve Fitting Parameters of prSSU Homologous Binding

Fitted parameter	Dataset ^a		
	30 nM	100 nM	Global
K_d (nM)	153.8 (153.1)	153.8 (153.1)	153.8 (153.1)
95% CI of K_d (nM)	89.95 – 263.1 (92.12 – 254.4)	89.95 – 263.1 (92.12 – 254.4)	89.95 – 263.1 (92.12 – 254.4)
Fraction of Non-specific binding	0.3622 (0.3586)	0.3622 (0.3586)	0.3622 (0.3586)
Non-specific binding (%)	36.22 (35.86)	10.87 (10.76)	-
R^2	0.9209 (0.9141)	0.8397 (0.7655)	0.9626 (0.9490)
$R^{2\ b}$	0.9208 (0.9142)	0.8398 (0.7647)	-

^a The values are from scintillation counting and from autoradiograph (in parentheses).

^b The R^2 generated from swapping the data and fitted parameters between scintillation counting and autoradiograph.

the mature domain of the small subunit of RuBisCO (mSSU) does not compete with ^{35}S -prSSU for chloroplast binding (Figure 3-9, in black). Both forward TPs, SSF and FDF, compete for binding (Figure 3-9, in green and magenta), and surprisingly, both reverse peptides, SSR and FDR, also compete (Figure 3-9, in blue and orange). The equilibrium dissociation constants (K_i) and IC_{50} of the forward TPs are slightly lower than those of the reverse peptides (Table 3-3).

3.4.7 *In vitro* competitive import assay of the peptides

When incubated with precursor proteins in the presence of high levels of ATP (>1 mM), isolated chloroplasts import and process precursors to mature proteins, as evidenced by a shift in size from a larger precursor to a smaller TP-less mature protein (Figure 3-10) (Dabney-Smith et al., 1999; Friedman and Keegstra, 1989). First, import assays were performed to determine the time course of ^{35}S -prSSU imports (Figure 3-10A). Import of ^{35}S -prSSU using two concentrations of chloroplasts (0.125 and 0.25 mg chlorophyll/ml) were linear over a 30 min period. We used a 15 min import time for further experiments.

The import competition assay can be performed by titrating the competitors to a constant concentration of ^{35}S -prSSU. The amount of imported ^{35}S -prSSU determines the competitiveness of the competitors. The assays were performed with forward TPs and the reverse peptides, as well as mSSU and prSSU as negative and positive controls, respectively (Figure 3-10B). As expected, prSSU competes against ^{35}S -prSSU (Figure 3-10C, dark green) and mSSU does not (Figure 3-10C, black). Both forward TPs, SSF and FDF, compete for import (Figure 3-10C, green and magenta). Unlike the earlier *in vitro* monomerization and binding assays, these import assays reveal that SSR and FDR do not compete (Figure 3-10C, blue and orange), suggesting that the chloroplast translocons are able to effectively distinguish between forward and reverse sequences. The inhibition curves were fitted to one-phase exponential

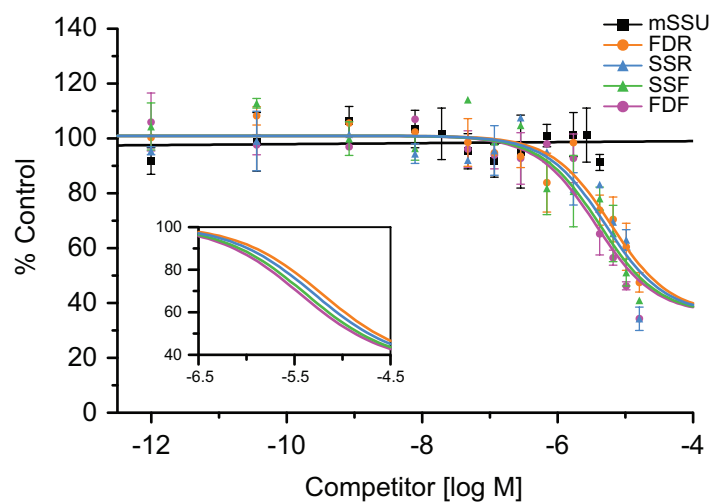


Figure 3-9. Competitive Binding of ^{35}S -prSSU with the Peptides

Competitive binding assays of competitors using scintillation counting. Inset shows close-up fitting curves of the peptides. Data is presented as means $\pm$ SD from three independent assays. The data of mSSU was fitted to a straight line.

Table 3-3. Curve Fitting Parameters of Peptide Competitive Binding

Fitted parameter	Competitor			
	FDF	FDR	SSF	SSR
K_i (μM)	2.220	3.735	2.537	3.091
95% CI of K_i (μM)	1.516 – 3.252	2.587 – 5.392	1.752 – 3.672	2.023 – 4.723
R^2	0.8763	0.8580	0.8565	0.8264
IC_{50} (μM)	3.663	6.164	4.186	5.101
95% CI of IC_{50} (μM)	2.501 – 5.366	4.270 – 8.898	2.891-6.060	3.338-7.794
R^2	0.8763	0.8580	0.8565	0.8264

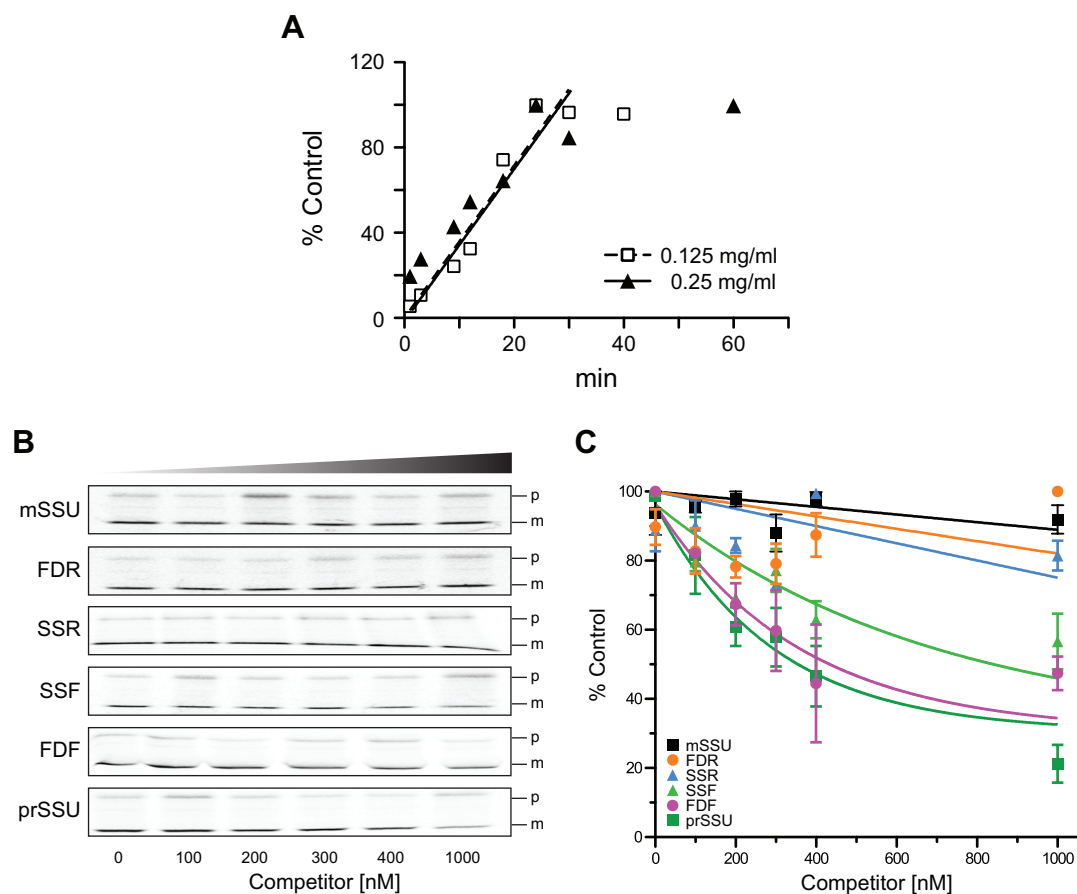


Figure 3-10. Competitive Import of ^{35}S -prSSU with the Peptides

(A) Import time course of ^{35}S -prSSU in 0.125 and 0.25 mg chlorophyll/ml chloroplasts. (B) Representative autoradiograms of SDS-PAGE gels from import assays in the presence of 100 nM ^{35}S -prSSU. Competitors are indicated on the left. Concentrations are shown at the bottom. p and m, precursor and mature protein sizes. (C) Quantification of imported ^{35}S -prSSU from the import assay partially represented in A. Data is presented as means $\pm$ SD from three independent assays. The data from mSSU, SSR and FDR were fitted to straight lines.

decays (Table 3-4). The IC_{50} of prSSU, FDF and SSF were determined to be 202.7, 247.8 and 480.3 nM, with 95% confident intervals of 130.4 – 454.6, 159.3 – 557.0 and 307.9 – 1092 nM, respectively.

3.4.8 *In vivo* imaging of forward and reverse-peptide fusion proteins

In addition to *in vitro* competition assays, the efficiencies of the forward and reverse TPs in directing the import of YFP into plastids were tested. The fusion proteins are chimeric proteins containing a TP fused to the first 20 aa of *N. tabacum* mSSU followed by YFP. The localization of transiently expressed TP-YFP fusion proteins was observed in *N. benthamiana* leaves, *Arabidopsis* seedlings and onion epidermis peels as shown in Figure 3-11A, B and Figure 3-12A, respectively. In agreement with *in vitro* import assays, the forward TPs, SSF and FDF, direct YFP into plastids as observed by co-localization with chlorophyll in tobacco (Figure 3-11A), co-localization with a plastid CFP marker in *Arabidopsis* (Figure 3-11B) and the punctate pattern in onion cells (Figure 3-12A). Note that in tobacco some YFP signals do not overlapped with chlorophyll signals since plastids in epidermal cells do not contain chlorophylls. Using the reverse-TP fusion proteins, both SSR and FDR show low efficiency in directing YFP into plastids. Most of the YFP signal was observed outside of the plastids (Figures 3-11A, B and 3-12A).

To determine if the reverse peptides could direct plastid protein import when fused to at the C-terminus, both SSR and SSF were fused to YFP at the C-terminus (YFP-SSR and YFP-SSF). Both C-terminal fusion proteins are not targeted into the plastids (Figure 3-13). While YFP-SSF localizes in the nucleus and the cytosol (Figure 3-13A), YFP-SSR localizes in peroxisomes (Figure 3-13B). Note that the plastid-localized signals detected in the YFP channel of YFP-SSF are the CFP signals from ntSSF-20-CFP (Figure 3-13A). The

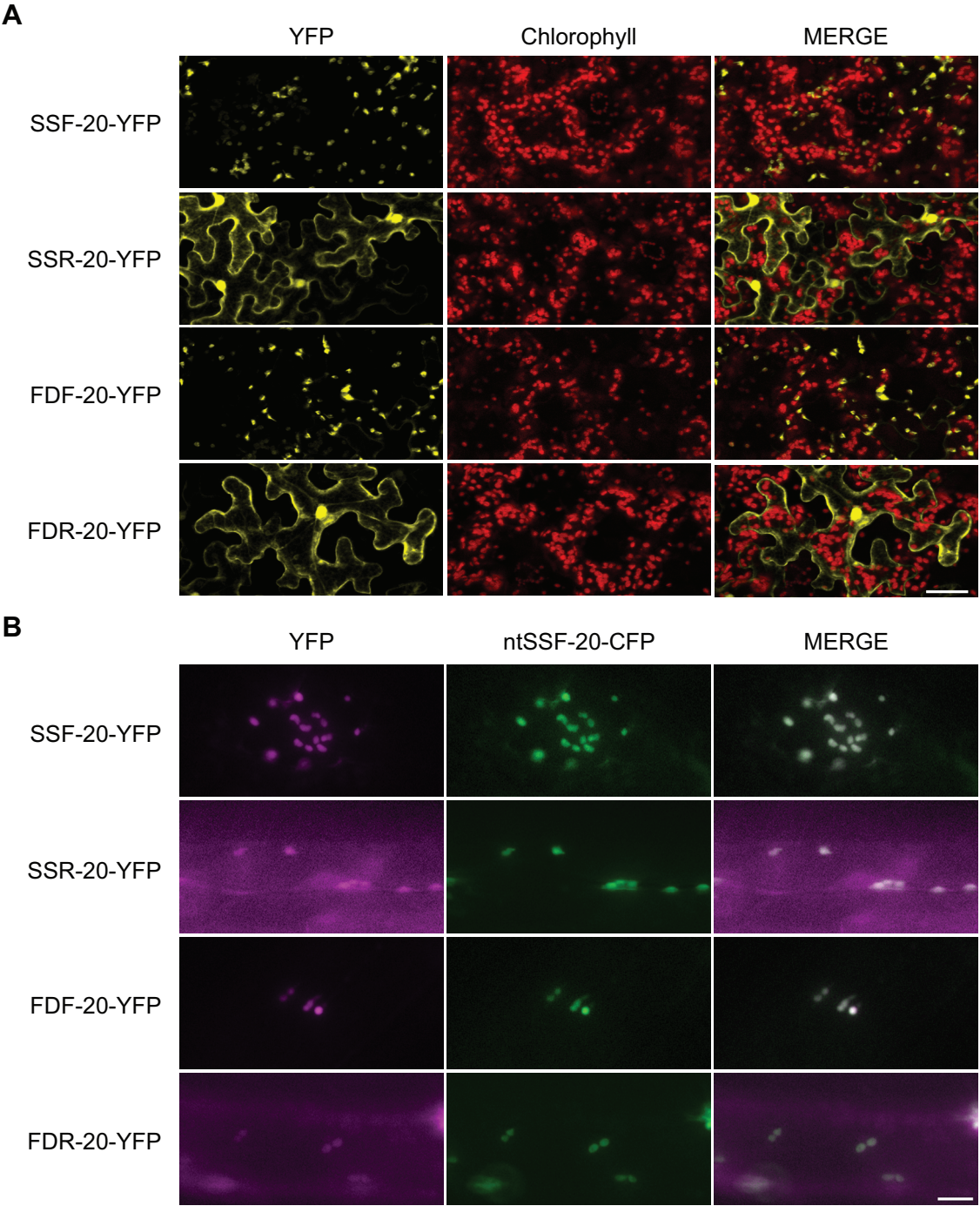
Table 3-4. Curve Fitting Parameters of Competitive Import

Fitted parameter	Competitor					
	prSSU	FDF	SSF	FDR	SSR	mSSU
Maximal import signal (I_0) ^a	65.91	65.91	65.91	-	-	-
Decay rate (k)	0.002454	0.004672	0.004600	-	-	-
Non-specific binding sigal ^a	30.43	30.43	30.43	-	-	-
R ²	0.8370	0.7451	0.7187	-	-	-
IC_{50} (nM)	355.06	434.14	841.50	No Inhibition	No Inhibition	No Inhibition

^a Shared parameters.

Figure 3-11. Targeting of Fluorescent Proteins Directed by Forward and Reverse Peptides into the Chloroplasts of Tobacco and *Arabidopsis*

In vivo plastid targeting functions of TPs were observed using N-terminus fusions of TPs linked to the first 20 aa sequence of mSSU from *N. tabacum* followed by YFP. (A) Localization patterns of transiently expressed YFP fusion proteins in *N. benthamiana* leaves. Auto-fluorescence of chlorophyll was used as a chloroplast marker. Only some YFP signals overlap with chlorophyll signal because plastids in epidermis cells do not contain chlorophylls. Left labels indicate TPs. Top labels indicate fluorescent signals. Bar, 50 μm . (B) Localization patterns of transiently expressed YFP fusion proteins in *A. thaliana* seedlings. The CFP plastid marker construct (ntSSF-20-CFP) was used as a plastid marker. Bar, 10 μm .



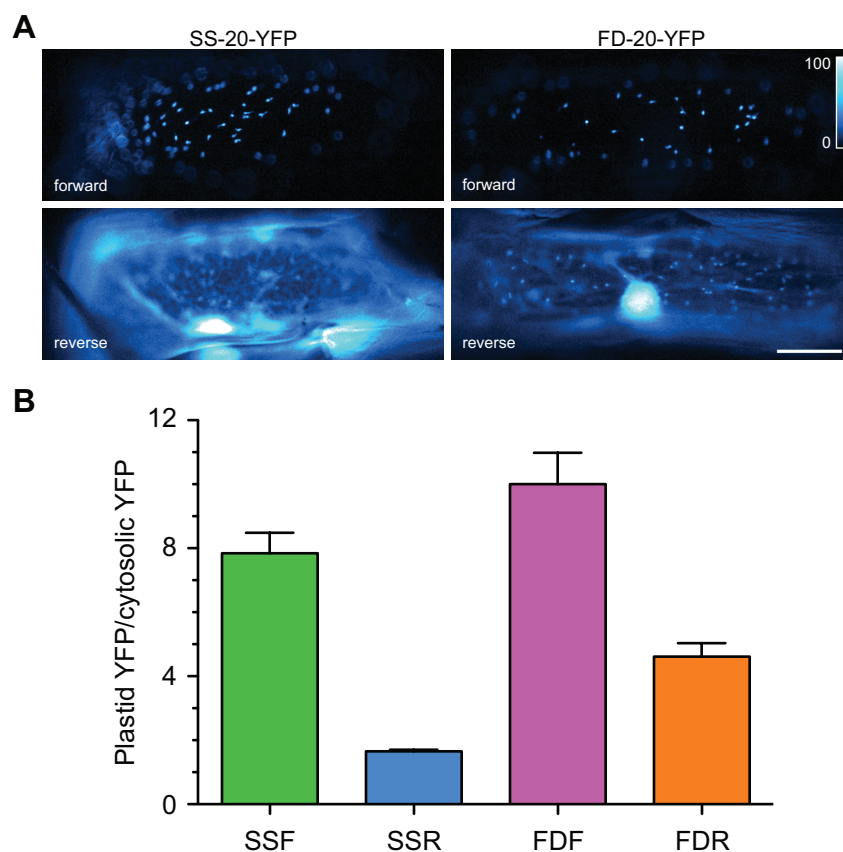


Figure 3-12. Targeting of Fluorescent Proteins Directed by Forward and Reverse Peptides into the Plastids of Onions

In vivo plastid targeting functions of TPs were tested using the N-terminus fusions of TPs linked to the first 20 aa sequence of mSSU from *N. tabacum* followed by YFP. (A) Localization patterns of transiently expressed YFP fusion proteins in onion epidermal cells observed under 20x objective. The forward peptide driven YFP proteins showed strong localization to the plastids, whereas reverse-peptide driven YFP proteins localized mostly outside of the plastids. Bar, 50 μ m; SS and FD, TPs of small subunit of RuBisCO and ferredoxin, respectively. (B) Quantitative analysis of plastid targeting represented in A. $n = 20$. Means $\pm$ SE are shown.

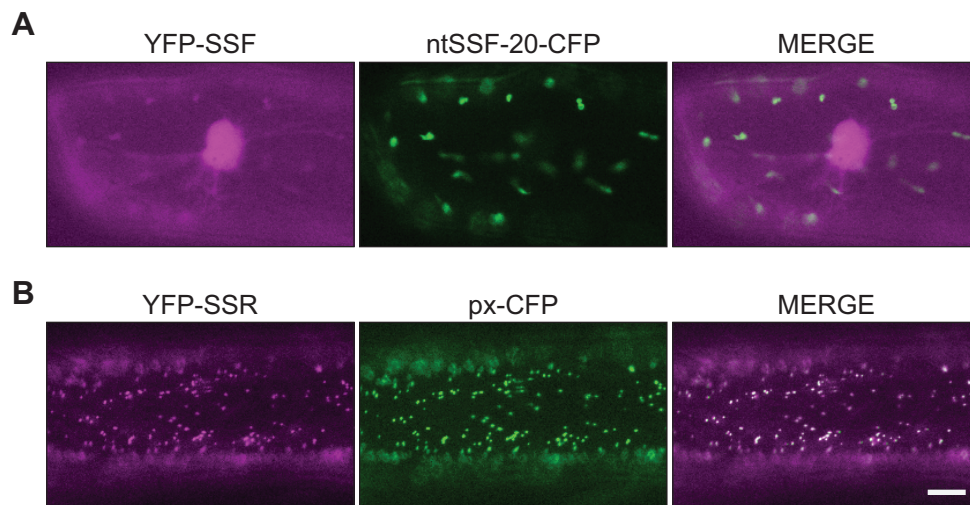


Figure 3-13. Targeting of Fluorescent Proteins Directed by the Peptides at the C-terminus in Onion Cells

(A) YFP-SSF fusion protein localization in comparison with a plastid CFP marker (ntSSF-20-CFP). (B) YFP-SSR fusion protein localization in comparison with a peroxisome CFP marker (px-CFP). Images were taken with 20x objective. Bar, 20 μ m.

peroxisome localization of YFP-SSR is possibly due to degradation at the C-terminus of 29 aa exposing the peroxisome-targeting signal Ser-Lys-Leu (Reumann, 2004) within SSR sequence.

3.4.9 Development of a intensity ratio measurement for *in vivo* import assay

One of the challenges in analyzing results from *in vivo* assays is the lack of a quantitative measure similar to the K_i and IC_{50} generated from *in vitro* assays. Here, a novel method was developed to quantify the efficiency of TP in directing the import of YFP into the plastids of onion cells. The efficiency is expressed as the relative intensity ratio between the plastid YFP and cytosolic YFP signals. The ratio measurements were performed on transiently transformed onion cells similar to those of Figure 3-12A and the results are shown in Figure 3-12B. As expected, the average ratios of SSF and FDF are high at 7.85 and 10.0, respectively, since most of the YFP was targeted to the plastids. SSR and FDR show lower average ratios at 1.65 and 4.61, respectively, indicating lower efficiency in directing the import.

To establish the reproducibility of the intensity ratio measurement, two sweet onion cultivars, Vidalia and Sunbrero, and a White onion cultivar were examined. Three independent assays were performed using the plastid marker construct pAN187 containing *N. tabacum* SStp and 20 aa of the mature domain followed by YFP (ntSSF-20-YFP). Figure 3-14 shows the ratios of ntSSF-20-YFP measured from these onions. Tukey's test showed that the average ratios between experiments within the same cultivar are not different ($p > 0.05$). However, when the data from the same cultivar were combined, the average ratios between cultivars were significantly different ($p < 0.001$). This result indicates that ratio measurements are reproducible within a cultivar.

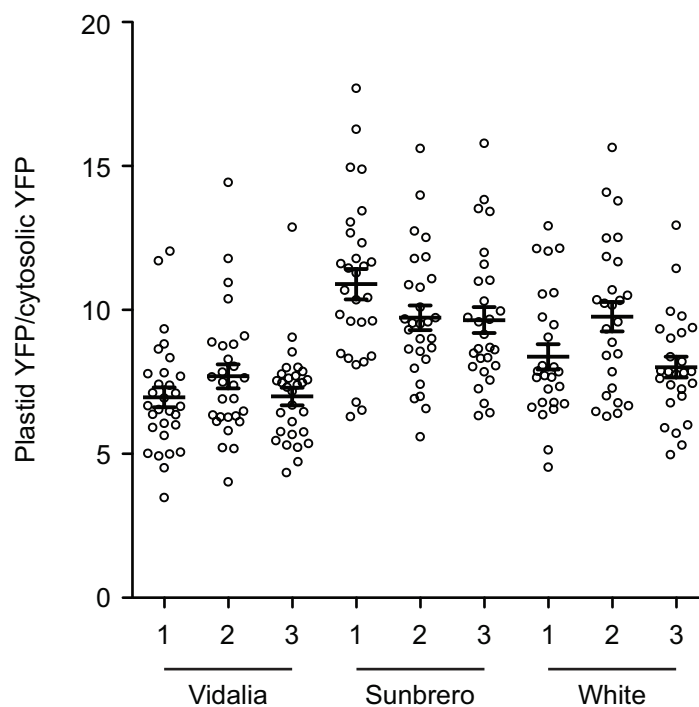


Figure 3-14. Effect of Onion Cultivar on the Targeting of Fluorescent Proteins into the Plastids

Quantitative analysis of *in vivo* targeting of ntSSF-20-YFP proteins. The ratios were determined using three independent assays (labeled 1, 2, 3) for each cultivar. $26 \leq n \leq 30$. Means $\pm$ SE are shown. Each spot represents the averaged ratio of each cell.

3.4.10 Immunoblotting analysis of import and processing of forward and reverse-peptide fusion proteins

To investigate the ability of chloroplast SPP to process chimeric fusion proteins and to confirm the *in vivo* import of the proteins, Western blotting was used to detect YFP at the C-terminus of fusion proteins. Western blotting of total protein extract from tobacco leaves transiently expressing fusion proteins show processed forms which are indicative of protein translocation into the chloroplasts (Figure 3-15A). The mature forms with sizes similar to YFP were observed in all fusion proteins expressed in tobacco indicating that the proteins were imported and that the cleavage sites were located near the beginning of the YFP domain. In addition, comparison with the *E. coli* expressed full-length precursors indicates that precursor and intermediate forms were present in the reverse-peptide fusion proteins. The reverse-peptide fusion proteins also showed lower amounts of mature forms, indicating a lower efficiency of import similar to the *in vivo* imaging results (Figures 3-11 and 3-12). *In vitro* import of fusion proteins was performed using spinach chloroplasts and shown in Figure 3-15B. Re-isolated chloroplasts showed processed forms similar to total protein extracts from tobacco for forward-TP fusion proteins, confirming that processing occurred in the chloroplasts. Because the same amount of re-isolated chloroplast proteins was loaded here, the antibody failed to detect the processed forms of reverse-peptide fusion proteins indicating that they were imported less efficiently. *In vitro* stromal processing assays were performed using the stromal extract from spinach chloroplasts (Figure 3-15C). Here, the SPP was able to cleave all fusion proteins into processed species similar irrespective of the orientation of the TP (Figure 3-15A and B).

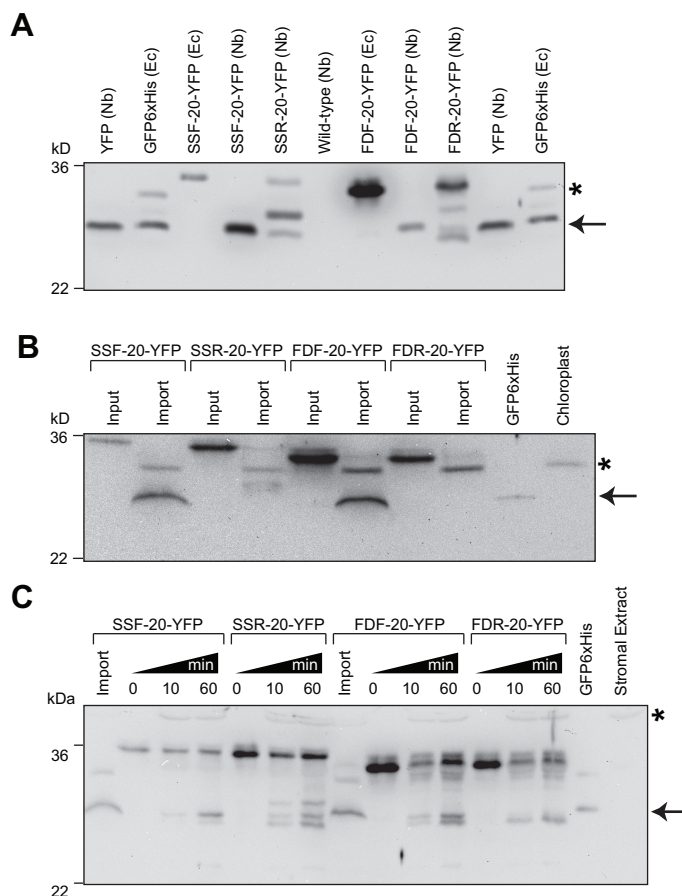


Figure 3-15. Immunoblotting Analysis of Import and Processing of Forward and Reverse-peptide Fusion Proteins

(A) Western blotting of YFP targeting in total protein extracts from tobacco leaves. Top labels indicate the expressed proteins. Nb and Ec, tobacco and *E. coli* extracts. Arrow and asterisk indicate the mature forms and non-specific bands from *E. coli* extract, respectively. Note that different amounts of protein extract were loaded in order to visualize the processed species. (B) Western blotting of *in vitro* imported fusion proteins. Equal protein amounts of re-isolated chloroplasts were loaded. Arrow and asterisk indicate the mature forms and non-specific bands from chloroplasts, respectively. (C) Western blotting of *in vitro* stromal processing assays. Arrow and asterisk indicate the processed forms and non-specific bands from the stromal extract, respectively.

3.4.11 Role of the transit peptide N-termini in protein import

Based on previous observations that TPs harbor an uncharged N-terminus (von Heijne et al., 1989) and our analysis in Figure 3-3, it is possible that the low import efficiency of reverse-peptides is due to charged aa present within their N-terminus (Figure 3-2A). To investigate the role of an uncharged N-terminus, the charged N-termini reverse-peptide YFP fusion proteins were altered by appending an extra 10 aa corresponding to the N-terminus of the forward TP as shown in Figure 3-16A. The alteration of the uncharged N-termini of forward-TP fusion proteins to the charged N-termini of the reverse TPs was performed in a similar manner. *In vivo* targeting using onion epidermal cells was performed (Figure 3-17) and the plastid/cytosolic YFP ratio was calculated (Figure 3-16B). The extra residues seem to invert the import efficiencies of the former TPs. Fusion proteins containing a Met following the extra sequence show moderate change compared to the fusion proteins with an Ala substitution of the Met. This result indicates the possibility of two translation start sites at the first and the internal Met residues that potentially lead to the production of a mixture of proteins with either low or high import efficiencies. In addition to the ratios obtained 12 h after transformation, in some cases, we calculated the ratio using the images captured 24 h after transformation (Figure 3-18). The results showed that ratios obtained after 24 h are always slightly higher than those from 12 h indicating the accumulation of YFP in the plastids over time.

Finally, two algorithms were used to predict the Hsp70 binding site in the peptides (Figure 3-19). All import-efficient peptides that we have studied seem to harbor a strong Hsp70 binding site at their N-termini (the first 10 residues) whereas the sites are not present in import-deficient peptides. The results generated from the algorithm developed by Ivey et al (2000) show strong agreement with every prediction (Figure 3-16 and 3-19A). However, the results generated from the algorithm developed by Rudiger et al (1997) show some

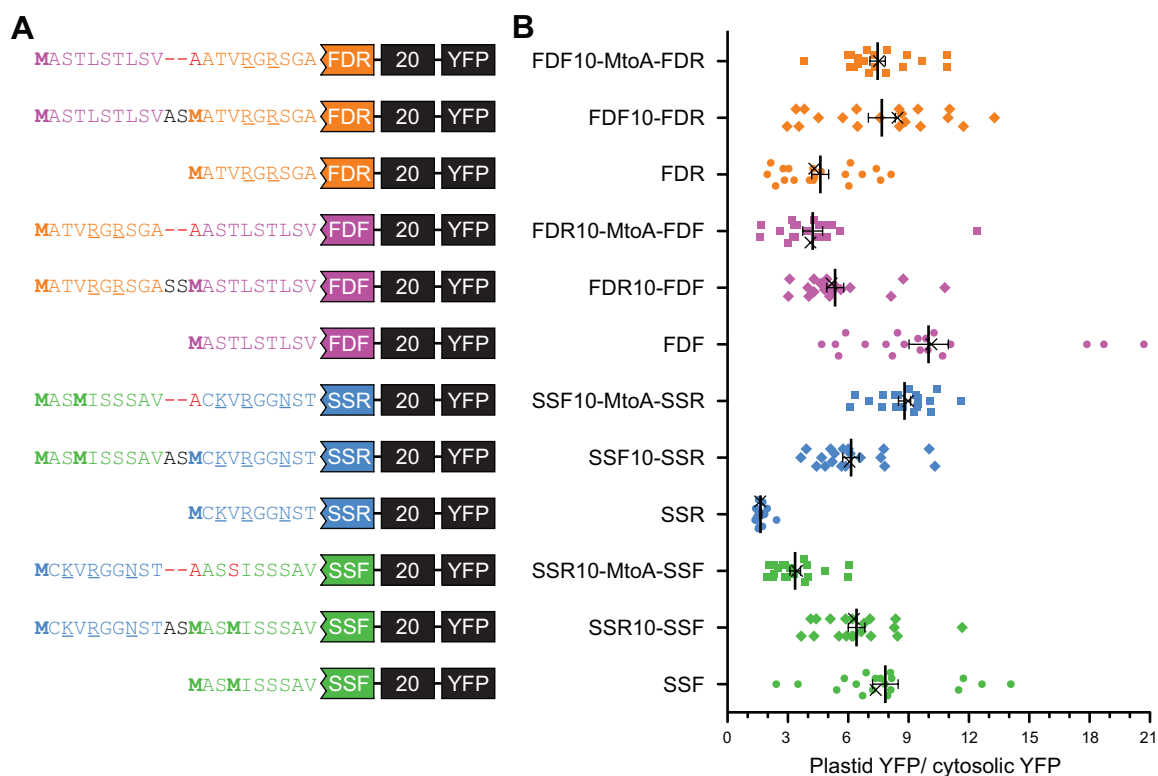


Figure 3-16. Plastid Import Efficiency of N-terminal-altered Fusion Proteins

Extra aa sequences representing the first 10 aa from the opposite TP were added to the N-terminus of each fusion construct.

(A) Representation of the constructs used in B. The partial N-terminal sequences are shown. Met are in bold. Substitutions are in red. Additional aa residues from restriction sites are in black. Charged aa are underlined. (B) Plastid targeting efficiencies of the fusion proteins in onion epidermal cells. Left labels indicate TPs in the constructs. The extra 10 residues were named based on the sourced TP and indicated with suffix 10. MtoA indicated the substitution of the internal Met with Ala. $n = 20$. The images of cells marked with X are shown in Figure 3-17.

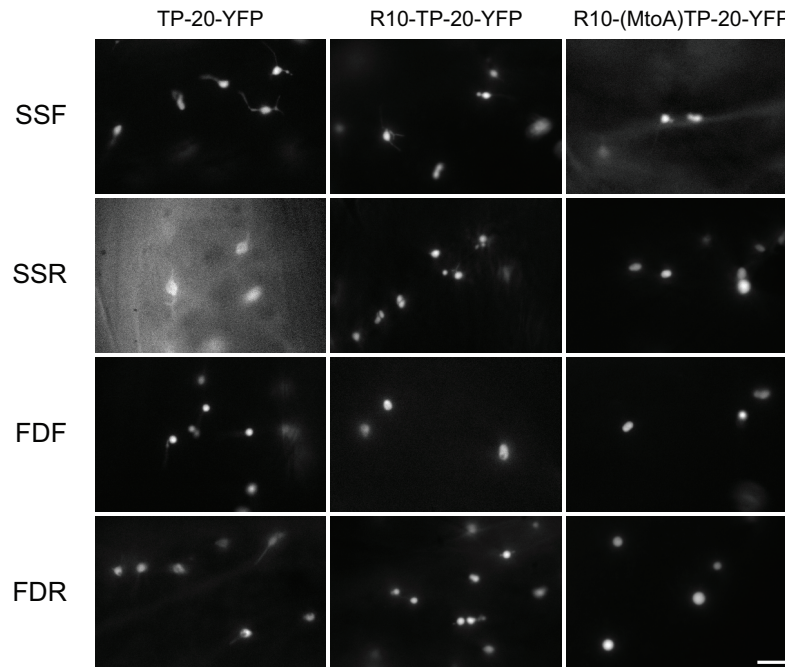


Figure 3-17. Plastid Targeting of N-terminal-altered Fusion Proteins in Onion Cells

An extension of 10 aa representing the beginning of the opposite TP were added at the N-terminus of each TP-YFP fusion construct. *In vivo* plastid targeting was observed in transiently transformed onion epidermal cells. Top labels indicate the fusion proteins. TP is shown on the left labels. R10 indicates the first 10 aa sequence from the opposite TP. MtoA indicates the substitution of internal Met with Ala. The calculated ratios of plastid YFP/ cytosolic YFP are shown in Figure 3-16B. Bar, 10 μ m.

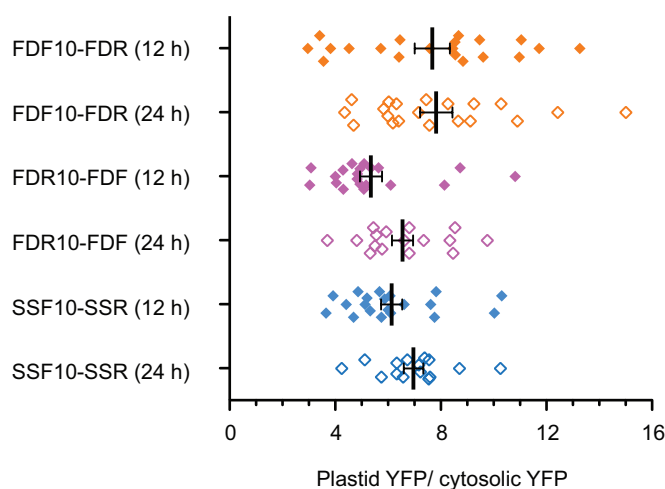


Figure 3-18. Time-dependent Targeting of Fusion Proteins into the Plastid of Onion Cells

The ratios of plastid YFP/cytosolic YFP were calculated from images taken at 12 and 24 h after transformations. Left labels show the TP in the constructs. Suffix 10 indicated only the first 10 aa sequence. Data was collected from 20 cells except SSF10-SSR (24 h) and FDR10-FDF (24 h) where 15 and 16 cells were used, respectively.

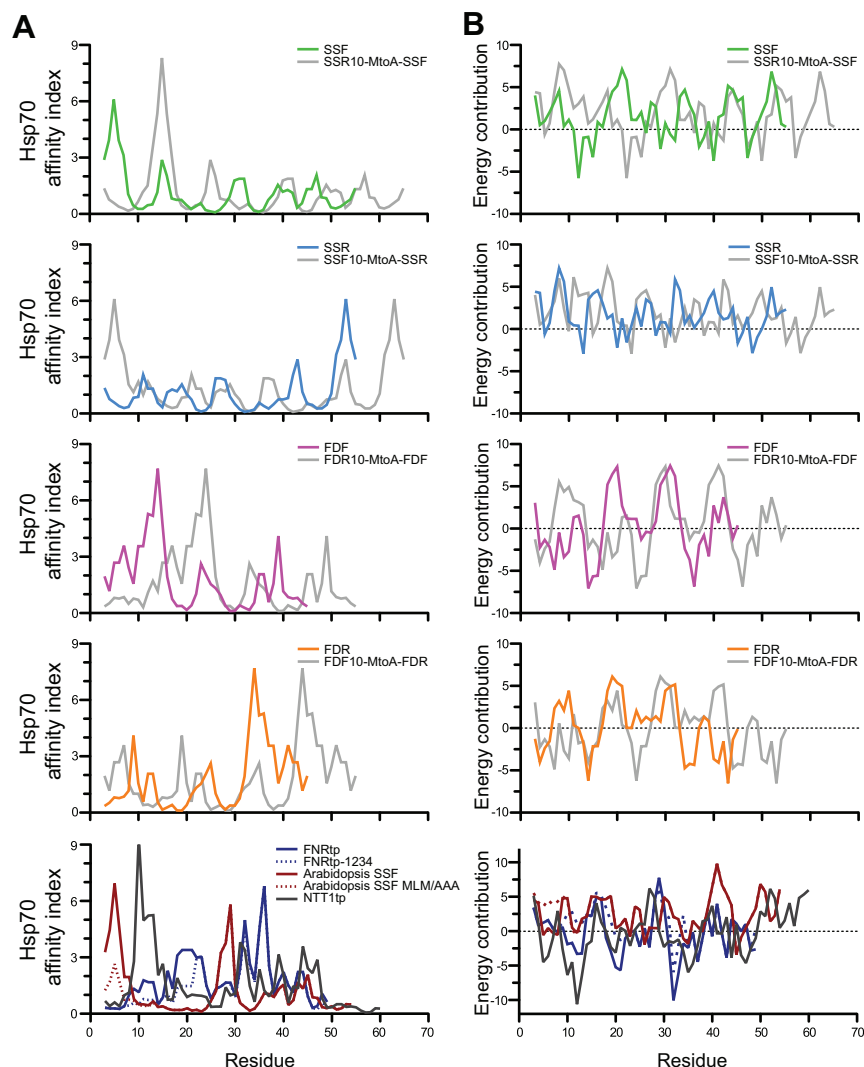


Figure 3-19. Hsp70 Binding Site Prediction of the Forward and Reverse Peptides

(A) and (B) show the prediction results generated using algorithms developed by Ivey et al. (2000) and Rudiger et al. (1997), respectively. The higher values in (A) or lower values in (B) are predicted to have higher affinity to Hsp70. FNRtp and FNRtp-1234 are wild-type and mutant TPs of ferredoxin-NADPH reductase reported by Rial et al. (2003). *Arabidopsis* SSF mutant MLM/AAA was generated by Lee et al. (2009). NTT1tp is the TP of *Arabidopsis* nucleotide transporter 1 reported utilized Bionda et al. (2010).

disagreement in the predictions of SSF and SSR where the calculated energy contribution within the first 10 residues shows no difference (Figure 3-16 and 3-19B).

3.5 Discussion

The lack of a consensus motif within TPs (Bruce, 2001; Lee et al., 2008) poses an interesting question as to how the TOC translocon can quickly and efficiently recognize and import highly variable TPs. To differentiate sequence specific contributions from physicochemical properties of TPs, we assayed two model TPs, the small subunit of RuBisCO from *P. sativum* (SSF) and ferredoxin from *S. latifolia* (FDF) as well as their respective N- to C- reversions, called reverse peptides, SSR and FDR, respectively. We showed that the forward (native) and reverse peptides share limited primary sequence similarity, while having identical aa composition. Using both forward and reverse peptides in several *in vitro*, *in organello*, and *in vivo* assays, we have revealed two modes of recognition for TPs. Toc34 recognition *in vitro*, along with preprotein binding *in organello* and *in vivo*, are specific for the physicochemical properties of the peptides. However, translocations *in organello* and *in vivo* demonstrate strong spatial and/or sequence specificity within the N-terminal uncharged region of TP that harbor a predicted Hsp70 binding site.

3.5.1 Flexible recognition of TPs by Toc34

Although there have been reports that cytosolic factors initially recognize TPs/precursors, kinetic arguments suggest that *in vitro* import can attain native import rates without cytosolic factors to support organelle biogenesis (May and Soll, 2000; Pilon et al., 1992b). This very rapid binding and translocation occurs despite a very low-density distribution of TOC complexes on the outer membrane

of plastids (≤ 1 TOC/13,600 nm²) (Friedman and Keegstra, 1989; Schleiff et al., 2003b). One potential strategy to increase the kinetics of productive binding of TP with the TOC complex is to relax the structural constraints of the recognition mechanism.

Although the structural basis of initial binding is poorly understood, it is probably mediated by one of the TOC GTPases. While the interaction between TP and Toc34 has been clearly shown (Jelic et al., 2002; Reddick et al., 2007; Schleiff et al., 2002; Sveshnikova et al., 2000b), it is not known how and where Toc34 binds to the TP. The shortest peptide shown to directly bind to Toc34 is the B1 peptide corresponding to aa 22-47 of *N. tabacum* SSF (Schleiff et al., 2002). In fact, these residues in pea and *Arabidopsis* SSFs were shown to contain two FGLK motifs (Pilon et al., 1995). Another evidence using a deletion mutant of pea SSF lacking the second FGLK motif found that the mutant peptide was unable to bind to the isolated chloroplasts (Subramanian, 2001). Hence, we proposed that the FGLK motif is required for Toc34 recognition. Since our heuristic analysis of targeting sequences previously showed the best discrimination of chloroplast TPs when the FGLK motif was required, this motif may be specifically recognized by Toc34 during chloroplast protein import (Chotewutmontri et al., 2012). We have shown that both forward and reverse peptides are able to stimulate Toc34 GTP hydrolysis *in vitro* (Chotewutmontri et al., 2012). In addition, this ability to productively interact with the TP independent of collision orientation is also supported by the observation that both SSF and SSR are able to shift the monomer-dimer equilibrium of psToc34G towards the monomer (Figure 3-4). Disruption of Toc34 dimerization by TPs may or may not be required for effective preprotein import but does appear to be a consequence of either TP binding and/or structural changes arising from GTP hydrolysis.

TGs are largely unstructured in solution forming a “perfect random coil” (von Heijne and Nishikawa, 1991) suggesting that TGs may be a new example of

intrinsically disordered proteins (IDPs) or at least contain significant segments that are predicted to be IDPs. Mechanistically, the ability of IDPs to assume different conformations when interacting with different binding partners (Kriwacki et al., 1996; Narayan et al., 2011; Tompa et al., 2005; Uversky et al., 2008) may explain how multiple interactions can be accommodated by the relatively short TPs during binding and translocation. It has also been proposed that the kinetics of favorable recognition between two components can be accelerated by increasing the encounter productivity using the so-called “Fly-casting model” (Shoemaker et al., 2000). This model explains how TPs may be rapidly and successfully recognized upon receptor recognition regardless of whether the peptide binds with one topological orientation or the opposite. Furthermore, this model of binding would accommodate both the forward and reverse peptides’ abilities to be recognized successfully by one or more of the TOC GTPases. Having the ability to successfully recognize TPs with either an N- or C-terminal orientation during binding would greatly accelerate preprotein binding and processing. This may explain how the kinetics of post-translational preprotein import *in vitro* may be able to match the maximum rates predicted *in vivo* during greening and chloroplast development (Pilon et al., 1992b) without the need for cytosolic factors or the topological organization present during the co-translational processes of protein transport.

3.5.2 Chaperone interactions with TPs

Our work has clearly implicated the N-terminal region of two well studied TPs. Prior work has already shown multiple interactions of TP N-termini with lipids (Pilon et al., 1995; Pinnaduwaage and Bruce, 1996) and the import receptor Toc159 (Lee et al., 2009b). However, our results which demonstrate that the reverse peptides were able to compete with prSSU for binding to the chloroplasts (Figure 3-9) yet are unable direct fusion protein import into the plastids (Figures

3-11, 3-12 and 3-15 to 3-17), suggest that the discrimination of TPs during translocation is mediated by components functioning *in trans* relative to the outer envelope components. Based on prior work in our lab (Ivey and Bruce, 2000; Ivey et al., 2000) and others (Rial et al., 2000; Zhang and Glaser, 2002), there is clear evidence that N-terminal sequences of TPs interact with the Hsp70 class of molecular chaperones that may be located in either the IMS or the stroma. Although TP interactions with chaperones localized in the IMS could not be ruled out, only two (cpHsc70-1 & cpHsp70-2) of the 14 *Arabidopsis* Hsp70s are predicted to be chloroplast localized and both of which are localized in the stroma (Ratnayake et al., 2008; Su and Li, 2008; Su and Li, 2010). Moreover, there is no evidence of any Hsp93 homologue to be localized to the IMS (Constan et al., 2004). Thus several labs have suggested that one role of the N-terminal residues is to mediate interaction with stromal chaperones.

Although Hsp70s were initially proposed to mediate multiple steps in chloroplast protein import (Marshall et al., 1990), similar to what has been observed in mitochondria (Tomkiewicz et al., 2007), it is a member of the Hsp100 family, stromal Hsp93 (also called ClpC) that has been proposed to drive chloroplast import (Cline and Dabney-Smith, 2008; Jarvis, 2008). The essentiality of this chaperone family has been shown in *Arabidopsis* where a double knockout of the two chloroplast homologues, atHSP93-V (ClpC1) and atHSP93-III (ClpC2), proved lethal (Kovacheva et al., 2007). Although Hsp93 has been shown to interact with translocon components as well as precursor proteins (Nielsen et al., 1997), there is still no direct evidence of its interaction with TPs. Furthermore, recent *in vivo* work has implicated a clear role of Hsp70s in protein import. In *Arabidopsis*, the mutants of the two stromal localized Hsp70s show reduced translocation efficiencies (Su and Li, 2010). Moreover, in moss (*P. patens*), stromal Hsp70-2 is an essential gene with temperature-sensitive mutants demonstrating reduced protein import (Shi and Theg, 2010). It was recently concluded that chloroplasts might have two separate chaperone systems

facilitating protein translocation into the stroma: the cpHsc70 system and the Hsp93/Tic40 system (Shi and Theg, 2011; Su and Li, 2010). Supporting this dual translocation model, it was shown that protein import into chloroplasts from the *cphsc70-1 hsp93-V* double mutant had a more severe import defect than either of the single mutants, suggesting that the two proteins function independently, possibly interacting with a discreet subset of substrates. Moreover, the *cphsc70-1 tic40* double knockout was lethal, confirming that cpHsc70-1 and Tic40 have overlapping yet essential functions.

Although Hsp93 and Hsp70 may both play vital roles in chloroplast protein import, considerably more is known about TP recognition by Hsp70 (Ivey et al., 2000; Rial et al., 2000; Zhang and Glaser, 2002). Bioinformatic and experimental approaches demonstrate that the N-terminal region of most TPs (>75%) has the highest affinity for Hsp70. Our current work confirms this observation since only the import-efficient TPs contain a strong Hsp70 binding site at their N-termini (Figures 3-16 to 3-19). Interestingly, the fact that all four of our TPs are able to productively interact with the stromal Hsp70 CSS1 in solution (Chotewutmontri et al., 2012), along with the observation that placement of a non-Hsp70 binding segment in front of an existing binding domain reduces the translocation of a precursor *in vivo*, evokes a specific placement requirement of where Hsp70 interacting sequences can function in driving translocation. Although we observed *in vitro* the ability of forward TPs and reverse peptides to interact with two individual translocon components, Toc34 (Figure 3-4) and CSS1 (Chotewutmontri et al., 2012), the sequence determinants for actual translocation of the precursors through TOC is more complex and strongly influenced by the N-terminal 10 aa. The importance of the highly uncharged N-terminus of TPs (von Heijne et al., 1989) in chloroplast import has been shown repeatedly both *in vitro* (Pilon et al., 1995; Pinnaduwaage and Bruce, 1996; Rensink et al., 1998) and *in vivo* (Lee et al., 2008; Lee et al., 2006; Lee et al., 2009a; Lee et al., 2002). These studies concluded that the hydrophobic N-

terminal region of FDtp and SStp was involved in directing the initial stages of the import process by binding to either envelope lipids or the Toc159/86 import receptor.

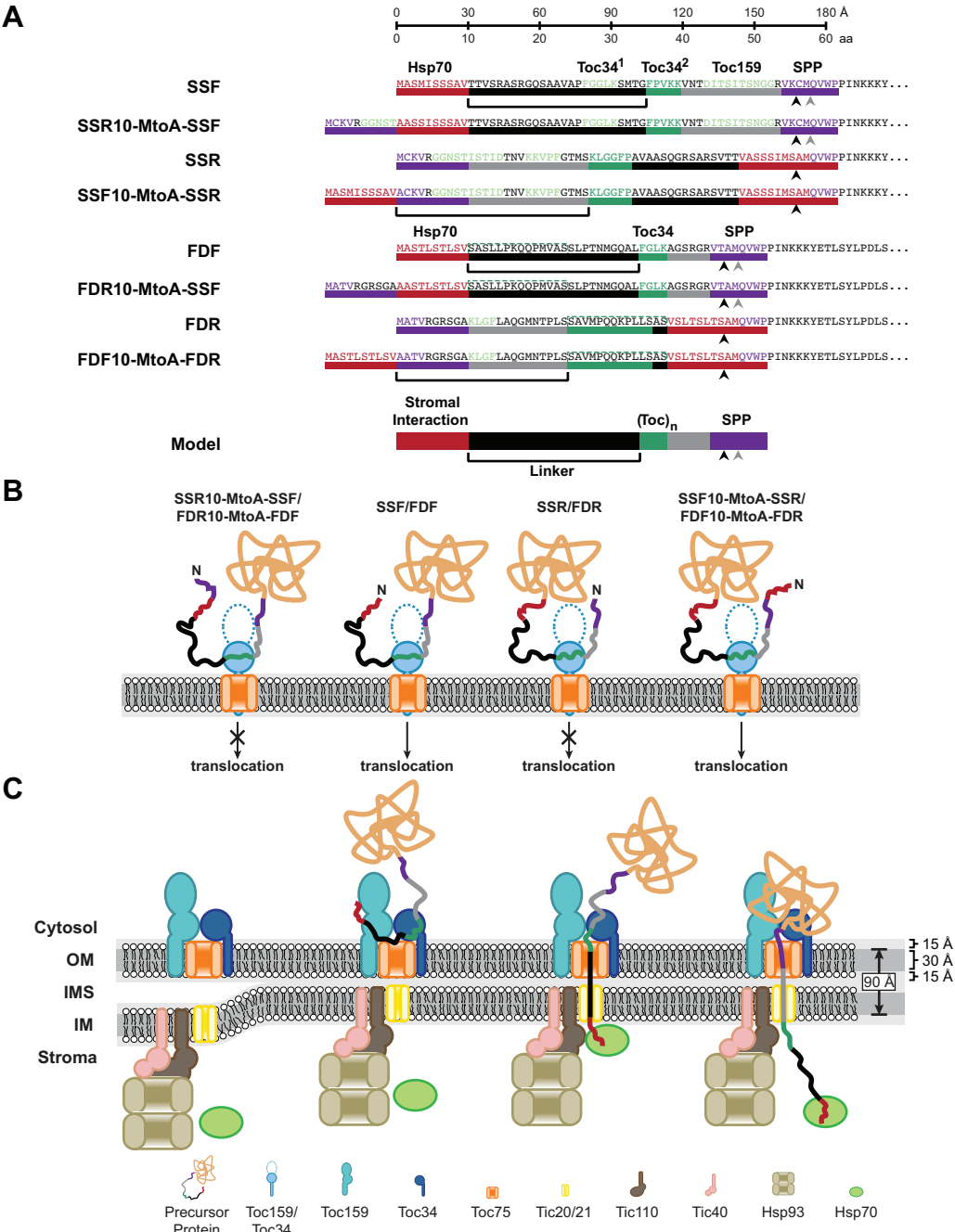
However, our findings provide strong evidence of separate recognition requirements during the binding and translocation. The reverse peptides behave differently from other N-terminal mutated TPs (Lee et al., 2006; Lee et al., 2009a; Lee et al., 2002; Pilon et al., 1995; Rensink et al., 1998) such that they are indistinguishable from forward TPs in preprotein binding (Figures 3-4 and 3-9) yet were unable to direct translocation *in vivo* (Figures 3-11 and 3-12) or *in vitro* (Figures 3-10 and 3-15B). Thus, although the reverse peptide can undergo successful binding, their sequence/organization does not permit translocation. Future work with these reverse peptides may permit us to further identify the critical elements of the N-terminal region of TP for stromal-protein interaction.

3.5.3 Bimodal model of TP design

Based on previous studies and our observation that the N-terminal Hsp70 binding site of TP is the major determinant for translocation, we propose a model describing a bimodal TP architecture (Figure 3-20A), which allows a TP to concurrently engage TOC receptors and stromal chaperones. The TP is proposed to contain an N-terminal stromal protein recognition site linked to a TOC receptor recognition site via a linker region along with a SPP recognition site at its C-terminus (Figure 3-20A). This work demonstrates that the N-terminal Hsp70 binding site of SSF and FDF determines the translocation of preproteins into chloroplasts (Figures 3-16 and 3-19). Many lines of evidence also suggest that the stromal Hsp70 is a chloroplast translocation motor (Marshall et al., 1990; Schnell et al., 1994; Shi and Theg, 2010; Su and Li, 2010; Theg et al., 1989). Based on the trapping and pulling model in ER and mitochondria protein import (Tomkiewicz et al., 2007), and the proposed unfolding and pulling model

Figure 3-20. Bimodal Model of TP Design

(A) The recognition elements in forward-TP, reverse-peptide, and MtoA constructs are shown. The sequence was highlighted and marked with colored bars to indicate different elements. Predicted Hsp70 binding sites are colored red. Proposed Toc34 (FGLK) and experimentally determined Toc159 binding sites are colored green. Predicted SPP recognition site are colored in purple with black and grey arrowheads indicating TargetP predicted and actual cleavage sites, respectively. Top ruler bar shows the aa length and the length of experimentally determined unfolded protein (~ 3.0 Å/aa). The black bar indicates the linker region. The general model of TP is shown at the bottom. (B) Depiction of translocation competent and incompetent fusion protein interactions with Toc34 receptor. As unstructured proteins, forward and reverse peptides can engage to Toc34 in opposite orientations. However, only proteins containing a N-terminal Hsp70 site are able to translocate across the membrane. (C) Depiction of concurrent TP recognition by Toc34 and stromal Hsp70. As detailed in the discussion, an import-efficient TP is proposed to harbor an N-terminal stromal interacting site and a TOC receptor binding site separated by a linker with a preferred length that allows the concurrent engagement of a TOC receptor and a stromal motor through the double membrane. The hydrophobic core of the double membrane at the contact site is estimated to be 90 Å.



in chloroplast import (Keegstra and Cline, 1999), it is tempting to predict that interactions of TP with the stromal chaperones is required to trap or pull the precursor into the chloroplasts. Our finding of the N-terminal Hsp70 binding site requirement in protein import supports this prediction. Apart from stromal Hsp70, stromal Hsp93 is another example of stromal protein involved in chloroplast import, although little is known regarding the Hsp93 recognition sequence on the TP (Chou et al., 2006; Kovacheva et al., 2007; Nielsen et al., 1997). Whether discrete subsets of TPs contain an N-terminal Hsp93-binding element still remains to be determined. With these pieces of information, we propose that TPs harbor a stromal protein recognition site at their N-termini, which allow the stromal proteins to trap and/or pull TP during the translocation process (RULE 1). In a study where all of the predicted Hsp70 binding sites of pea ferredoxin-NADPH reductase TP (FNRtp) were mutated, the mutant TP (FNRtp-1234) was able to direct protein import (Rial et al., 2003). Our Hsp70 predictions showed that FNRtp intrinsically lacks an N-terminal Hsp70 (Figure 3-19) suggesting that it may utilize Hsp93 in trapping. In fact, FNRtp has recently been shown to interact with Hsp93 (Bruch et al., 2012).

Other portions of TPs have been shown to interact with lipids and TOC components in order to be targeted to the chloroplast surface (Bruce, 2000). Manipulation of the Toc34 recognition site by deletion of the FGLK motif in FDF (Rensink et al., 1998) or Ala scanning of the second FGLK motif in *Arabidopsis* SSF (Lee et al., 2006) were shown to inhibit translocation but still permit binding. Thus, TP mediated precursor translocation is determined not only by the N-terminal Hsp70 binding site but also by the Toc34 binding site. Nevertheless, the N-terminal Hsp70 binding site seems to be the major determinant for translocation, as it seems that it can overrule Toc34 recognition. As evidence for this, both forward TPs and reverse peptides contain FGLK motifs but only those harboring an N-terminal Hsp70 binding site are efficiently translocated (Figures 3-2A, 3-10 to 3-12, 3-15 to 3-17, 3-19). Upon deletion of the

N-terminal Hsp70 binding site, Δ 1-14 (Pilon et al., 1995) and Δ 6-14 (Rensink et al., 1998) of FDF, and Δ T1 (aa 2-12) of E1 α -subunit of pyruvate dehydrogenase TP (Lee et al., 2009a), translocation was abrogated. In contrast, a variant of N-terminal Hsp70 binding site mutant of *Arabidopsis* SSF, the MLM/AAA mutant (Met5, Lue6 & Met11 were substituted with Ala) has decreased import efficiency (Lee et al., 2009a) suggesting that the MLM/AAA mutant has lower Hsp70 affinity than that of the wild type as also predicted in Figure 3-19. Interestingly, translocation was restored to the wild-type SSF level when the N-terminus of MLM/AAA mutant was fused to the C-terminus of E1 α TP and expressed in the wild-type *Arabidopsis* (Lee et al., 2009a) suggesting that this “new” C-terminus may provide a higher affinity binding site and compensate for the low affinity Hsp70 binding site in MLM/AAA. In fact, the translocation of MLM/AAA chimera TP was not restored to the wild-type SSF level when expressed in the Toc159 knockout mutant *ppi2* suggesting that Toc159 interaction with the C-terminal domain can substitute for reduced Hsp70 affinity. Thus, there is a direct connection between Hsp70 interaction and TOC receptor binding in regulating the translocation steps in chloroplast import. While the stromal protein interaction at the N-terminus could possibly provide the trapping and pulling mechanism, the initial interaction with TOC components may capture or trap the TP at the chloroplast surface. The TOC GTPases may provide a mechanical pushing force through domain movements in response to nucleotide status.

Surprisingly, it has recently been shown that the reverse peptide of the C-terminal membrane anchor (M domain) of Toc159 fused at the N-terminus of GFP can direct GFP import into the chloroplast stroma indicating that M domain is a reverse TP (Lung and Chuong, 2012). This is a mirror construct compared to our YFP-SSR construct (Figure 3-13B) where the reverse N-terminal SStp was fused at the C-terminus of YFP. Although our result showed that YFP-SSR is not targeted to the stroma, Lung and Chuong (2012) showed that the C-terminal fusions of the M domain, SSR and FDR to GFP direct the

precursors to the chloroplast membranes. None of the C-terminal reverse-TP fusion proteins so far is able to direct protein import into the stroma indicating that the placement of TPs at the N-termini of precursors is essential for the translocation process. This requirement can be explained in part by a study showing that the stromal Hsp93 can only recognize FNRtp when it located at the N-terminus but not the C-terminus (Bruch et al., 2012).

3.5.4 Spatial requirements for concurrent TP recognition

The interconnection between stromal and surface interactions can occur efficiently only if both are occurring concurrently or sequentially within the rapid timeframe of protein import. Thus, for a given unfolded preprotein to engage two binding elements concurrently during translocation, the relative spatial distance between interaction domains may be a critical feature of TP function. We observed that the FGLK motif location of FDF and SSF is relatively conserved, suggesting the existence of a preferred TOC interaction site. Assuming an N-terminal Hsp70 binding site within the first 10 aa of a TP, the linker length between the stromal interaction site and the TOC interaction site is 22, 24, and 25 aa in *Arabidopsis* SSF, FDF, and pea SSF, respectively. We propose that there is a linker with preferred length of ≥ 22 aa that connects the N-terminal stromal interaction site with the TOC receptor interaction site (RULE 2). Using the TP of *Arabidopsis* nucleotide transporter 1, the overall TP length requirement has been shown to be at least 60 aa to translocate titin protein (Bionda et al., 2010). However, based on the cleavage site predictions, the length of this TP is only 21 aa, which is much shorter than most TPs. To coincide with our model, we predicted that this TP also contains an N-terminal Hsp70 binding site (Figure 3-19). According to the cleavage site prediction, this TP could not accommodate a preferred linker size. However, when analyzed further, we observed an additional FGLK motif at aa 35-39, allowing a linker of 24 aa. Thus,

it is possible for the two rules to be accommodated, even if the TOC interacting domain falls within the mature domain C-terminal to the SPP cleavage site.

Prior work suggests that protein import takes place at contact sites between the chloroplast outer and inner envelope membranes (Schnell et al., 1990), which are rich in galactolipids (Bruce, 1998). The thickness of MGDG and DGDG bilayers has been measured to be 55 and 60 Å, respectively (Bottier et al., 2007; Marra, 1985). Within this distance, we assume an organization similar to a model bilayer containing a 30-Å hydrophobic core with two 15-Å polar regions on the two surfaces (White and von Heijne, 2008). For two tightly pressed membranes, the thickness of the two hydrophobic cores of outer and inner membranes would be around 90 Å (Figure 3-20C).

The conformation of a TP during translocation is unknown, but in other systems extended peptide lengths are known for fully unstructured peptides. Using atomic force microscopy, the extensibility of peptides under different forces has been measured and suggests an average length of 3.0 Å/aa (Chyan et al., 2004; Rief et al., 1997). Using the thickness of the chloroplast envelope and this value for the extended TP, the shortest linker of 22 aa would have a length of 66 Å. Although this observation suggests that the linker could not span 90-Å double membranes, the N-terminus of TP could reach the stromal side with the length of 96 Å through the addition of 10 aa of the Hsp70 binding site. Thus, the translocation in our model can occur only where the TP is bound deep inside the TOC complex or released from TOC receptor prior to interaction with a stromal protein (Figure 3-20C). In this model, the TOC receptor functions both to target TP from the cytosol to chloroplast and to prevent the TP from escaping the translocon prior to stromal capture. Evidence for this dual trapping model was observed previously using a modified TP with an N-terminal epitope tag (His-S), which introduced multiple charged residues at the N terminus (Subramanian et al., 2001). This TP is able to compete for binding but is unable to be imported

into the chloroplast. We now suspect that the His-S tag prevents the natural N-terminus from interacting with the stromal chaperones.

3.6 Conclusions

In this chapter we combine both quantitative *in vitro* and *in vivo* analyses of the efficacy of chloroplast TPs with a large-scale bioinformatics analysis of TP sequences. Our results converge on a new model of modular design of TP organization and function. Specifically this design requires the placement of a specific N-terminal domain of the TP that must be able to productively interact with one or more Hsp70 class of molecular chaperones. This domain is followed by a second element that interacts with one or more components of the TOC apparatus, which can promote binding, yet alone cannot support translocation. This supports the prior evidence that translocation is driven by a stromal ATP-dependent process which may include (but not be limited to) Hsp70-mediated recognition. Although these sequences are degenerate in nature, we do observe a key spacing requirement that may reflect the coordinated translocation of the preprotein across both membranes at contact sites where the TOC and TIC complexes are tightly oppressed. With this advance we are now in position to start designing TP variants that may allow these spatial and energetic requirements to be tested directly. Finally, this advance may provide new insight into the evolution of chloroplast preproteins and partially explain the high variability of TP length, composition and sequence.

Chapter 4

Role of the Transit Peptide N-terminus in Plastid Protein Import

4.1 Abstract

Previously, we have identified the N-terminal domain of transit peptides (TPs) as a major determinant for the translocation step in plastid protein import. This domain was reported to have two overlapping characteristics, highly uncharged and Hsp70-interacting. To distinguish between these two properties, we replaced the N-terminal domains of the TP of the small subunit of ribulose-1,5-bis-phosphate carboxylase/oxygenase and its reverse peptide with a series of unrelated peptides with varying Hsp70 affinities. Sequence analysis indicated that eight out of nine peptides in this series are not similar to TP N-termini. Using *in vivo* and *in vitro* protein import assays, we found that all of the precursors lacking the N-terminal Hsp70 binding property were not targeted to the plastids while most of the precursors containing N-terminal Hsp70 binding peptides were targeted to plastids. We also discuss why some N-terminal Hsp70 binding peptides failed to direct import. The ability of the unrelated Hsp70 interacting peptides in substituting the function of TP N-terminal domain indicates that at least a subset of TPs utilize an N-terminal Hsp70 binding domain in the translocation process.

4.2 Introduction

In *Arabidopsis*, around 2,100 nuclear-encoded proteins are predicted to be targeted to plastids (Richly and Leister, 2004). More than 70% of these plastid-

localized proteins harbor a TP in their precursor proteins (Kleffmann et al., 2004; Zybaïlov et al., 2008). The N-terminal targeting sequences, TPs, govern the post-translational targeting of precursor proteins into the plastid stroma through the translocons at the outer and inner envelope membranes of the chloroplasts (TOC/TIC) (Bruce, 2000; Bruce, 2001). However, little is known about how TPs accomplish their functions.

Bioinformatic analysis has not been fruitful in identifying the consensus motifs within TPs (Lee et al., 2008). Although when divided into small groups, a few short conserved peptide motifs were identified but their functions are still unclear (Lee et al., 2008). At the secondary structure level, TPs form random coils in aqueous solution (Bruce, 1998; von Heijne and Nishikawa, 1991) hindering direct structure-function analysis. Nevertheless, a conserved domain organization of TPs containing 3 loosely defined regions has been identified: (i) N-terminal domain of about 10 uncharged residues ending with Pro/Gly and preferably having Ala as the second residue, (ii) central domain, lacking acidic aa but rich in hydroxylated aa, and (iii) C-terminal domain, rich in Arg and possibly forming an amphiphilic beta-strand (Bruce, 2001; von Heijne et al., 1989).

The importance of the highly uncharged N-terminus of TPs (von Heijne et al., 1989) in plastid protein import has been shown numerous times both *in vitro* (Pilon et al., 1995; Pinnaduwa and Bruce, 1996; Rensink et al., 1998) and *in vivo* (Lee et al., 2008; Lee et al., 2006; Lee et al., 2009a; Lee et al., 2002). These studies have concluded that the N-terminal domain of TPs of the small subunit of ribulose-1,5-bis-phosphate carboxylase/oxygenase (SStp) and ferredoxin (FDtp) are involved in directing precursor binding state of the import process by interacting with either the envelope lipids or the Toc159 receptor. However, in Chapter 3 we have generated mutant TPs lacking the uncharged N-terminal domain but still having ability to bind to the plastids similar to that of the wild-type TPs. These mutants of both SStp and FDtp were constructed to contain the

reversed aa sequences from C- to N-termini. These mutants together with a series of their N-terminal mutants have identified a novel role of the TP N-terminal domain as a requirement for the protein translocation into plastid stroma (Chapter 3).

In addition to the uncharged properties of the N-terminal domain of TPs, this domain has been shown to harbor a strong Hsp70 binding site (Ivey et al., 2000; Rial et al., 2000; Zhang and Glaser, 2002). Based on the predictions, all of our import-competent mutants contain a strong Hsp70 binding site at their N-termini, which is lacking in the import-deficient mutants (Chapter 3). Because the stromal Hsp70 was shown to act as a plastid translocation motor (Marshall et al., 1990; Schnell et al., 1994; Shi and Theg, 2010; Su and Li, 2010; Theg et al., 1989), we suspected that the N-terminal Hsp70 binding site in our import-competent constructs is required for the stromal Hsp70 interaction in order to trap and/or pull the precursor into the plastids similar to the proposed unfolding and pulling model of chloroplast protein import (Keegstra and Cline, 1999).

To further address these two properties of the N-terminal domain of TPs, we utilized the forward (native) and reverse constructs of SStp, which are called SSF and SSR, respectively. The N-termini of these constructs were extended to include peptide sequences derived from an article where their affinities to Hsp70s have been determined (Fourie et al., 1994). Nine peptides were chosen ranging from strongly interacting to non-interacting with Hsp70s. The effect of the N-terminal domain alteration in precursor protein translocation was assessed by *in vivo* and *in vitro* import assays. In addition, the recognition of physicochemical properties in the N-terminal domain was assessed using the mutants of SSF and FDtp (FDF) generated by flipping and scrambling of the N-terminal 10 residues. These results together demonstrate that the Hsp70-interacting property of the N-terminal domain of TP is essential for precursor protein translocation into the plastid stroma.

4.3 Results

4.3.1 Bioinformatic analysis of TP datasets

Previously, the N-terminal uncharged domain of TPs was identified by von Heijne et al. (1989) and we have extended the analysis to a larger dataset containing 912 predicted TPs from *Arabidopsis thaliana* (Chapter 3). Our analysis showed that the N-terminal 15 aa domain of TPs is highly uncharged. The analysis by Ivey et al. (2000) indicated another property of the N-terminal domain where about 70% of TPs in the CHLPEP dataset (von Heijne et al., 1991) harbor a strong Hsp70 binding site at their N-termini. Because only 14 out of 260 TPs in CHLPEP dataset are from *Arabidopsis* and most of the TPs in the dataset are redundant (for example, 53 instances of SStp), we revisited the Hsp70 binding site analysis.

Using the 208-TP dataset collected previously from experimentally verified *Arabidopsis* proteins (Lee et al., 2008), Hsp70 binding sites were predicted by using the random peptide phage display-derived algorithm (RPPD) (Gragerov et al., 1994; Ivey et al., 2000). Figure 4-1 shows Hsp70 binding site analysis along with TargetP prediction. The orange-yellow heat map represents levels of Hsp70 affinity predicted via RPPD algorithm with higher score (brighter color) corresponding to higher affinity. The 208 TPs were clustered into 9 subgroups based on patterns of their Hsp70 binding sites using the hierarchical clustering method. These subgroups show pronounced differences in the highest Hsp70 affinity location (the brightest locations) where subgroups 1 to 9 contains the highest Hsp70 affinity sites at around aa positions 5, 10, 17, 23, 32, 46, 52, 58 and 70 of the TPs, respectively. This analysis also indicates that 32.21% and 46.63% of TPs in this dataset contains the strongest Hsp70 binding site within the first 20 and 30 aa, respectively. The clustering results are also shown in Table A2-2.

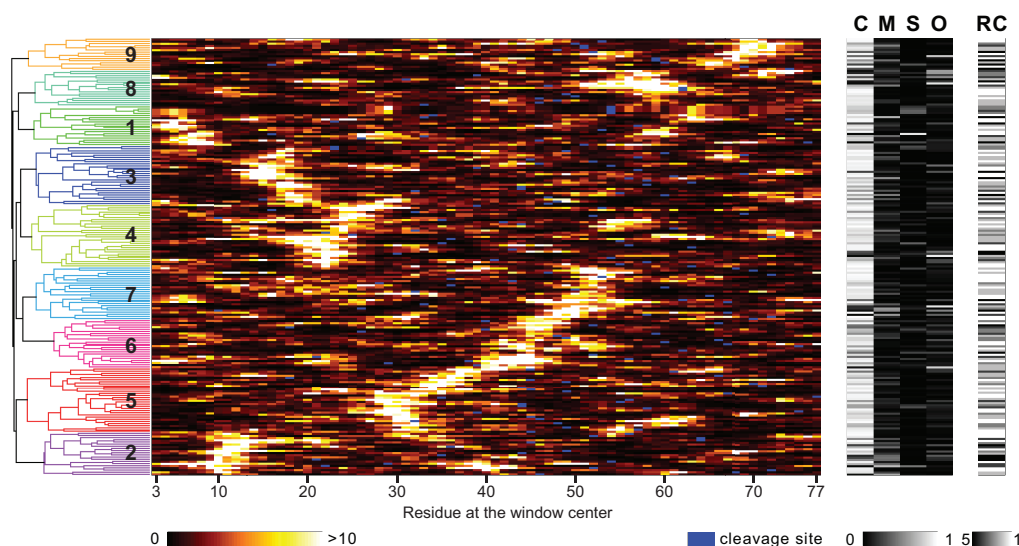


Figure 4-1. Hsp70 Binding Site and TargetP Predictions of the 208-TP Dataset

Each line represents a TP from the 208-TP dataset (Lee et al., 2008). Left panel shows a dendrogram of the hierarchical clustering of the TPs based on their predicted Hsp70 binding site patterns. Nine clusters were formed as indicated by numbering. The orange-yellow heat map panel represents the predicted Hsp70 binding score based on RPPD algorithm (Gragerov et al., 1994; Ivey et al., 2000) where the higher score (brighter color) has higher affinity. The black and white heat maps show TargetP (Emanuelsson et al., 2000) prediction results. The predicted cleavage sites were marked in blue in the RPPD heat map. C, M, S, O are the TargetP probability score for chloroplasts, mitochondria, secretory pathway and other localizations, respectively. RC is the TargetP reliability class where the lower value means higher confidence of prediction.

We further analyzed this dataset using the localization and cleavage site prediction program TargetP (Emanuelsson et al., 2000). About 90% of the TPs were predicted to localize to chloroplasts (Figure 4-1, black and white heat map) with only 68% of these predicted chloroplast TPs classified into the TargetP reliability classes 1 and 2. The reliability classes were determined by the differences between the highest and the second highest localization scores of TargetP and used to indicate the prediction confidence. Class 1 has the differences greater than 0.8 and class 2 has the differences between 0.6 and 0.8, respectively (Emanuelsson et al., 2000). In addition, TargetP prediction also provided predicted cleavage sites. The predicted chloroplast TPs from the 208-TP set have a mean TP length of 52.22 ± 17.15 (mean $\pm$ SD) aa.

We have previously determined the N-terminal uncharged domain in the 912-TP dataset (Chapter 3), the same analysis was also performed in the 208-TP dataset. Calculating the percentage of uncharged aa within a window size of 5-17 aa along the length of TP, we found that the % uncharged aa at the N-termini reached about 94% in average and decreased to 81% between residues 20-30 (Figure 4-2A). The C-termini of TPs after residue 40 contained about 75% uncharged aa. The transition mid-point between 94 to 81% of the N-terminal uncharged domain occurred at residue 14 when the data was fitted to the sigmoidal curves. This result is similar to the previous analysis of the 912-TP dataset. To show the % uncharged aa distribution of the N-terminal domain, we calculated % uncharged aa of the N-terminal domain with lengths of 5 to 17 aa. The N-terminal domains were separated into three % uncharged aa groups, 60-79, 80-99 and 100%. The fractions of TPs in each % uncharged aa group are shown in Figure 4-2B. The fraction of purely uncharged N-terminal region decreased from about 78% to 30% when the length of the N-terminal region increased from 5 to 17 aa. Inversely, the fraction of moderately uncharged N-termini increased from 18% to 65%. The fraction of 60-79% uncharged N-termini fluctuated

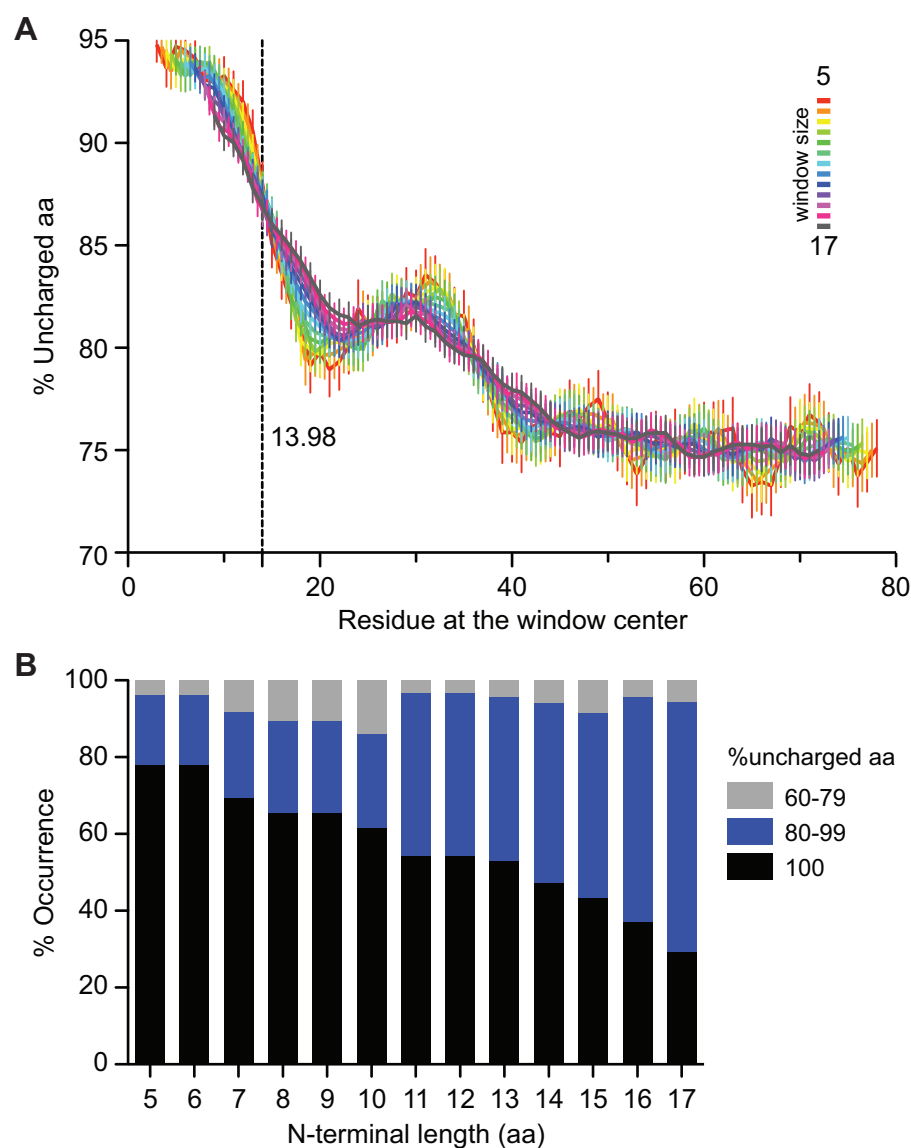


Figure 4-2. Analysis of the N-terminal Uncharged Domain of the 208-TP Dataset

(A) The percentage of uncharged aa within the window of size 5-17 aa along the length of TP was calculated from the dataset of 208 experimentally verified TPs from *A. thaliana*. $n = 208$. Means $\pm$ SE are shown. (B) Fractions of TPs in each of the three % uncharged groups observed in the TP N-terminal region of the 208-TP dataset based on the length of the N-terminal region from 5 to 17 aa.

between 3-13%. Thus, a large portion of TPs in the 208-TP dataset contains purely uncharged N-termini.

We then questioned if there is a correlation between the degree of uncharged-ness and the Hsp70 affinity. Using the 208-TP dataset, the accumulative RPPD scores from residues 1-15 of the TP N-termini were plotted against the % uncharged aa of the N-terminal 15 residues (Figure 4-3). The RPPD scores of the TPs with the % uncharged aa of the N-terminus greater than 85% distributed throughout the range of 2 to 140. The TPs with the % uncharged aa lower than 85% did not distributed in the low RPPD score area; this may be due to the small number of these TPs in the 208-TP dataset. When the TPs were grouped into the clusters determined from the RPPD patterns (Figure 4-1), TP clusters seemed to have no preference for any % uncharged aa range. Thus, we could not observed any direct correlation between the % uncharged aa and Hsp70 affinity of the TP N-termini.

Many published works have been performed using different TP datasets, yet the dataset comparison has not been done. We therefore compared three TP datasets, the 208-TP dataset (Lee et al., 2008), the 912-TP dataset (Chotewutmontri et al., 2012) and the 141-TP training set of TargetP (Emanuelsson et al., 2000). We found that the majority of TPs in each dataset did not overlap with the others (Figure 4-4). This was expected in case of the 141-TP dataset where TPs were derived from all plants in SWISS-PROT (Bairoch and Apweiler, 2000; Emanuelsson et al., 2000) but it was unexpected for the 208-TP and the 912-TP datasets that are both populated with *Arabidopsis* proteins. This discrepancy between the 208-TP and the 912-TP datasets may be due to the fact that the 912-TP set only contains TargetP reliability classes 1 and 2 proteins with predicted TP length between 35 -71 aa. Out of 125 TPs belonging to classes 1 and 2 in the 208-TP set, more than 66% (83 TPs) overlapped with the 912-TP dataset.

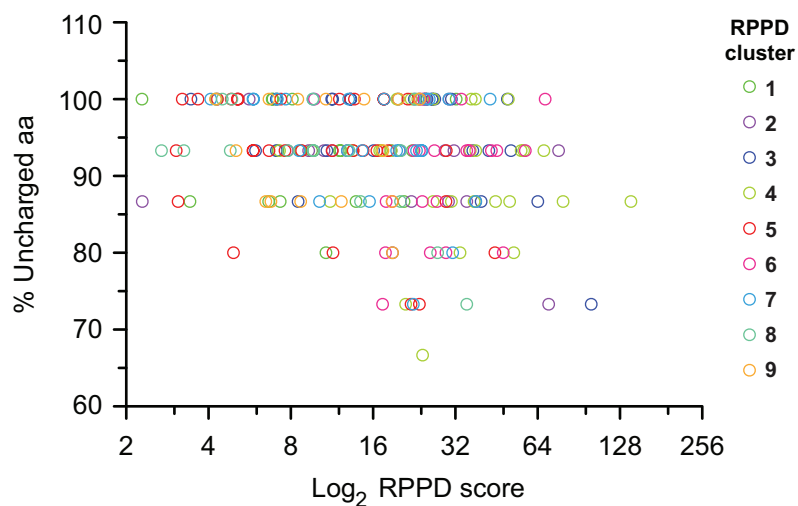


Figure 4-3. Hsp70 Affinity versus Percentage of Uncharged Amino Acids of the N-terminal domains from the 208-TP Dataset

The Hsp70 affinity is presented as the accumulative RPPD score calculated based on the first 15 aa. The % uncharged aa value was also calculated from the first 15 aa. The RPPD cluster groups were the same as described in Figure 4-1.

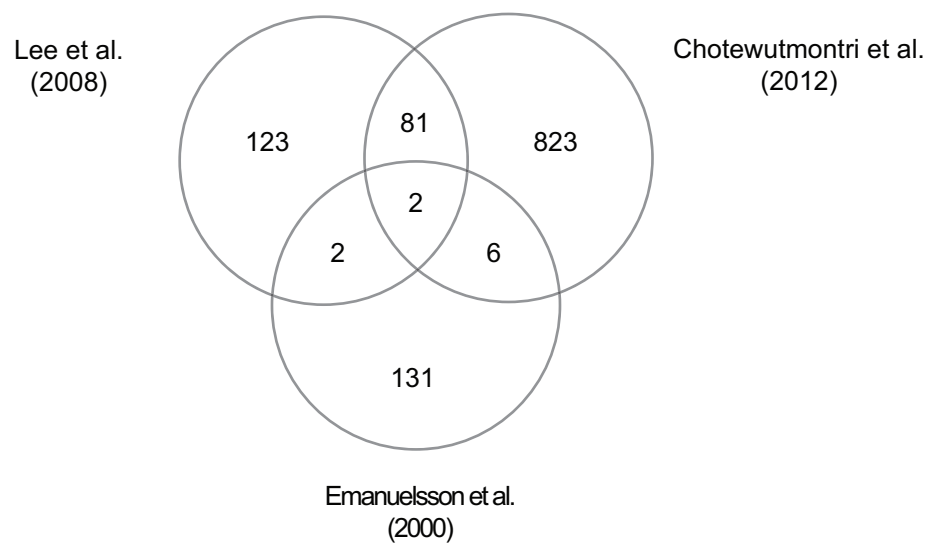


Figure 4-4. Comparison of TP Datasets

Three TP datasets were compared; the 208-TP dataset (Lee et al., 2008), the 912-TP dataset (Chotewutmontri et al., 2012) and 141-TP training set of TargetP (Emanuelsson et al., 2000). Number of TPs are shown.

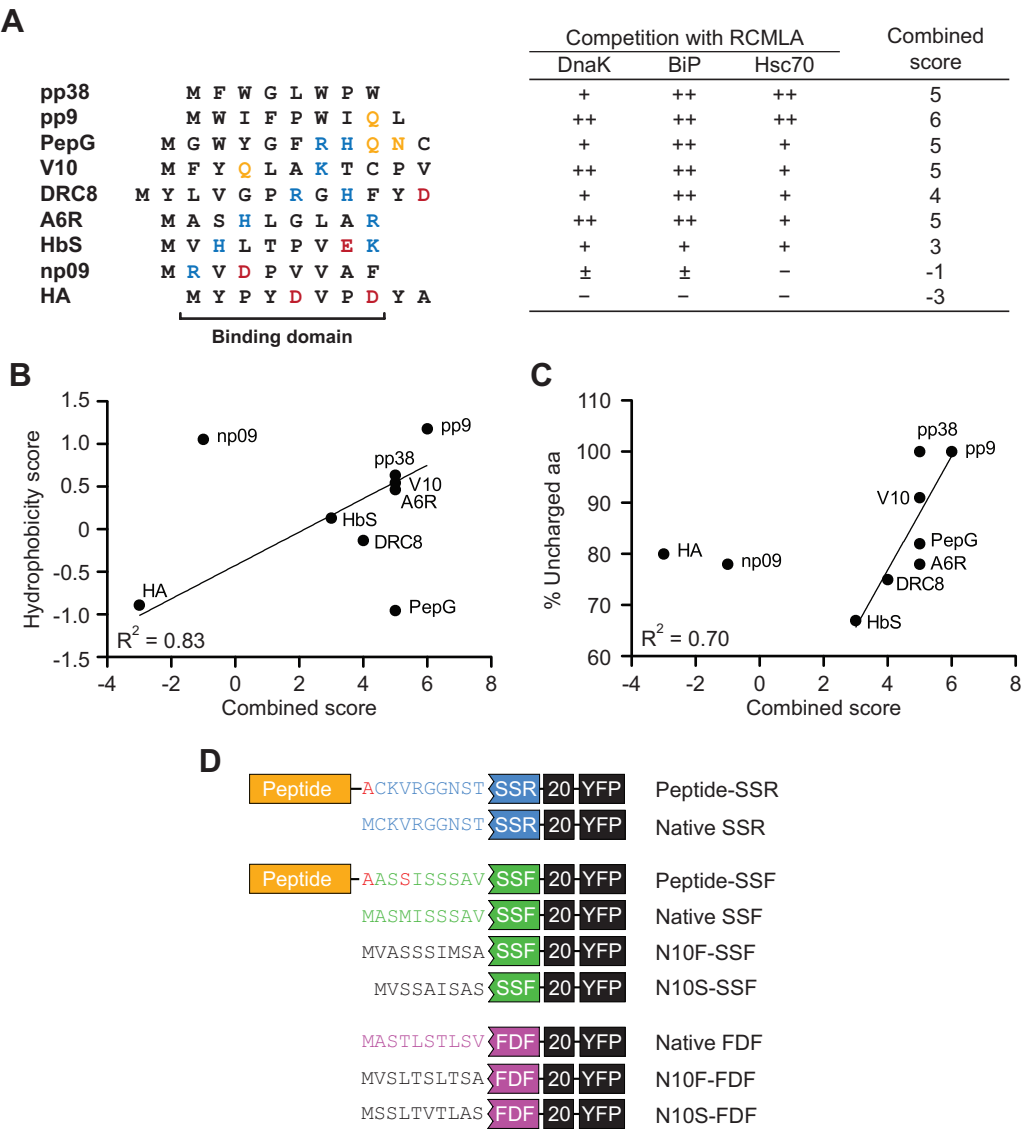
4.3.2 Construction of TP N-terminal mutants

We developed a quantitative *in vivo* plastid protein import assay as described in Chapter 3. This assay utilizes the transient expression of TP-YFP fusion proteins in onion epidermal cells to determine the plastid targeting efficiency of TPs as a ratio of plastid YFP signal to cytosolic YFP signal. In order to determine the role of the uncharged and strong Hsp70 binding properties of the N-terminal domain of TPs, two previously generated constructs were used to generate series of N-terminal mutants. These constructs were the SSF fused to 20 aa of the mature domain of prSSU followed by YFP (SSF-20-YFP) and the fusion protein of the reverse peptide of SSF (SSR-20-YFP). While SSF-20-YFP localized to the plastids, SSR-20-YFP did not target to the plastids (Chapter 3). Thus, the mutations were made in both import-competent and import-deficient TP constructs.

The N-termini of these two constructs were altered to contain additional peptide sequences derived from a study where their affinities to Hsp70s have been determined (Fourie et al., 1994). Nine peptides were chosen ranging from strongly interacting to non-interacting with Hsp70s (Figure 4-5A). We have truncated the original peptide sequences to only the Hsp70 binding domains creating peptides ranging from 8-12 aa. All truncated peptides contain polar and/or charged residues except for the pp38 peptide. The full-length peptide affinities to *E. coli* DnaK, bovine ER luminal BiP and bovine cytosolic Hsc70 had been determined as rating scores of the competitiveness against reduced carboxy-methyl lactalbumin (RCMLA) in binding to Hsp70s (Fourie et al., 1994). We calculated the summation of the rating scores where + and – were assigned values of 1 and –1, respectively, to generate a combined rating score for each peptide (Figure 4-5A). The N-terminal mutants of TPs were generated by extending their original N-termini to include the truncated peptide sequences together with the substitution of internal Met residues to Ala or Ser (Figure 4-5D).

Figure 4-5. The N-terminal Peptide Sequences and the Mutant Constructs

(A) Sequences of the Hsp70-interacting and non-interacting peptides are shown. Polar, negative and positive aa are colored yellow, red and blue, respectively. Table shows the published ratings of peptide competitiveness in binding competition with RCMLA to three Hsp70s: DnaK, BiP and Hsc70 (Fourie et al., 1994). The combined rating scores were calculated from the summation of the ratings where + and – were assigned values of 1 and -1, respectively. (B) The hydrophobicity of the peptides in (A) plotted against the combined rating scores. The hydrophobicity scores were calculated from the hydrophobicity scale of aa by Kyte and Doolittle (1982). The correlation line was determined without np09 and PepG values. (C) The % uncharged aa of the peptides in (A) plotted against the combined rating scores. The correlation line was determined without HA and np09 values. (D) Representation of the N-terminal mutant constructs. For the fusion peptide constructs, the internal Met were mutated to Ala or Ser (colored red). N10F and N10S denote the flipped and scrambled mutants. The second Met in N10S-SSF was deleted. The native sequences of SSR, SSF and FDF are colored blue, green and magenta, respectively.



To assess whether the TP N-terminal domain recognition is based on physicochemical properties, another series of mutants were generated. The N-terminal 10 residues of both SSF-20-YFP and FDF-20-YFP constructs were replaced with either the flipped sequences (reverse sequence from C- to N-terminus) or the scrambled sequences (Figure 4-5D). These mutants should retain the same physicochemical properties as the original constructs.

4.3.3 Bioinformatic analysis of the the Hsp70-interacting and non-interacting peptides

To further characterize the Hsp70-interacting and non-interacting peptides, we determined the hydrophobicity and % uncharged aa of the peptides. The hydrophobicity scale of aa by Kyte and Doolittle (1982) was used to calculate the hydrophobicity score. When the combined rating scores were plotted against the hydrophobicity scores (Figure 4-5B), seven peptides showed correlation between hydrophobicity and Hsp70 affinity ($R^2 = 0.83$). PepG and np09 are the exceptions. Despite having a strong affinity, PepG is hydrophilic. Inversely, np09 is one of the most hydrophobic peptides in this set but has a weak affinity. The combined rating scores versus the % uncharged aa plot (Figure 4-5C) also showed a correlation between Hsp70 affinity and the % uncharged aa value in seven peptides with high Hsp70 affinities ($R^2 = 0.70$). The non-interacting peptides HA and np09, however, do not follow this correlation.

We further analyzed the TP sequences of these N-terminal peptide mutants using two Hsp70 binding prediction algorithms, RPPD (Gragerov et al., 1994; Ivey et al., 2000) and the cellulose-bound peptide scanning-derived (CBPS) algorithm (Rudiger et al., 1997b). Figure 4-6 shows the prediction scores generated from the N-terminal 15 aa. While higher RPPD score indicates higher affinity, lower CBPS score indicates higher affinity. For each of the N-terminal peptides that is about 10-aa long (Figure 4-5A), a pair of mutant constructs was

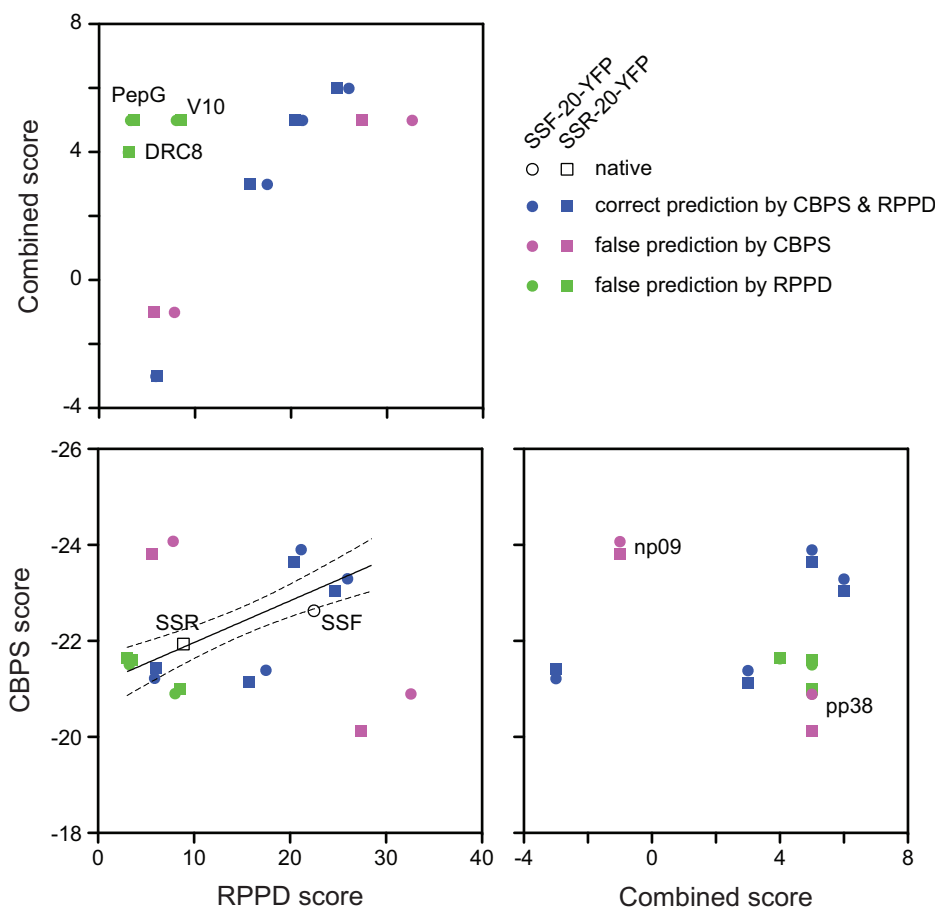


Figure 4-6. Hsp70 Binding Predictions of the N-terminal Peptide Mutants

The Hsp70 binding scores calculated from the N-terminal 15 aa of the mutant constructs are shown together with the experimentally derived combined rating score. Two Hsp70 binding prediction algorithms, RPPD (Gragerov et al., 1994; Ivey et al., 2000) and CBPS (Rudiger et al., 1997b), were used to calculate the accumulative scores within the N-terminal 15 aa. While higher RPPD score indicates higher affinity, lower CBPS score indicates higher affinity. Two constructs were made from each N-terminal peptide based on either SSF-20-YFP or SSR-20-YFP (shown as the closest pairs). In RPPD vs. CBPS plot, the scores of native N-termini of SSF and SSR are included and the correlation line along with its 95% CI lines is shown. Different classes of prediction results are colored differently.

made from SSF-20-YFP and SSR-20-YFP. The N-terminal 15 aa sequences of each pair are different because they incorporated either part of SSF or SSR, hence the pairs of each peptide showed a slightly different Hsp70 affinity. The RPPD-CBPS plot showed a moderate correlation between RPPD and CBPS scores with R^2 of 0.54. The CBPS scores versus the combined Hsp70 affinity rating scores plot indicated that the CBPS algorithm mispredicted the np09 pair as false positive and the pp38 pair as false negative. The RPPD scores versus the combined rating scores plot indicates that RPPD algorithm mispredicted PepG, V10 and DRC8 pairs as false negatives. Thus, although RPPD has high specificity, it lacks sensitivity to detect all of the interacting peptides.

We wanted to determine if the utilized N-terminal peptides are similar to TP N-termini. The position-specific scoring matrix (PSSM) method, which is heavily used in identification of motif or pattern in the sequences (Henikoff, 1996), was utilized, even though TP N-termini are weakly conserved (Figure 4-7A). The PSSM scores were calculated based on the aa distributions at specific positions in the sequence (Henikoff, 1996). Considering the TP N-terminal 10 aa as a pattern, a PSSM can be constructed based on the aa distributions of the N-terminal positions. Figure 4-7B shows that the first position of the TPs contains a conserved Met and while the residue 2 had a unique aa distribution, the residues 3-12 seemed to have approximately the same distribution. Thus, we constructed the TP PSSM containing 9 positions from residues 2 to 10. When the aa distribution of residue 2 was computed, we found that this position lacked Cys, His, Trp and Tyr. To obtain a better estimation of the aa distribution of residue 2, the larger 912-TP dataset was used instead. For residues 3 to 10, we combined all of the aa of residues 3-12 and calculated the averaged aa distribution of this region. Therefore, the TP PSSM only utilized 2 distributions, the residue 2 distribution from the 912-TP set for the matrix position 1 and the averaged distribution of residue 3-12 from the 208-TP for the matrix positions 2-9. The aa frequencies was converted into the matrix scores by dividing with the

Figure 4-7. Amino Acid Distribution of the TP N-termini and the Classification of the Peptide Mutant N-termini

(A) Logo plot of the N-terminal 15 aa of the 208-TP dataset. The total height of each residue position corresponds to the conservation in that position. (B) Logo plot showing the relative occurrence of each aa at each position in the N-terminal 15 aa of the 208-TP dataset. (C) The PSSM scores of the Hsp70-interacting and non-interacting peptides. The cumulative occurrence of the scores calculated from the N-terminal domain of the sequences in TargetP training set is shown together with the scores from the 208-TP dataset (the training set for this log-odds calculation). The lines were fitted based on a cumulative normal distribution.

background frequencies (the aa frequencies of UniProt database) followed by log base 2 transformation (see section 2.19.11 for detail). These scores indicate the degree of abundance of the aa in the TP N-termini compared to the average. A score greater than 0 indicates higher chance of being TP N-termini than random sequence. The total PSSM score was calculated from the summation of the corresponding scores from each matrix position.

About 90% of tested TP N-terminal sequences had the PSSM scores greater than 0 (Figure 4-7C). We found that DRC8, pp9, HA, pp38, PepG, HbS, np09 and V10 had the scores of -12.8, -12.7, -12.6, -8.8, -8.7, -5.2, -5.1 and -3.0, respectively, suggesting that they are not similar to TP N-termini (Figure 4-7C). But A6R peptide had a score of 1.4, which makes it possible to be a TP N-terminus. About 20% of TPs had the scores less than 1.4 (Figure 4-7C).

4.3.4 Prediction of protein localization using amino acid distributions of the N-terminal sequences

Based on the classification results of the peptides using PSSM method that only utilizes the short N-terminal 10 aa sequence, we expanded the classification to cover both mitochondrial and secretory pathway proteins. These proteins also contain the targeting signals in their N-terminal sequences (Schnell and Hebert, 2003). The analysis of the TP N-terminal uncharged domains (Chapter 3 and Figure 4-2) indicates that there is a transition of the sequence composition between residues 1-30. Therefore, the new TP PSSM was extended from residue 10 to residue 30. The aa distribution showed that from residue 15, positively charged aa, Arg and Lys, become more prevalent (Figure 4-8A). Unlike TPs, the N-terminal 15 aa of mitochondrial targeting peptides (mTPs) showed large numbers of Arg (Figure 4-8B). Interestingly, in the same region as the uncharged-to-charged transition found in TP, the signal peptides (SPs) of secretory pathway proteins became concentrated with aliphatic aa, Leu, Val and

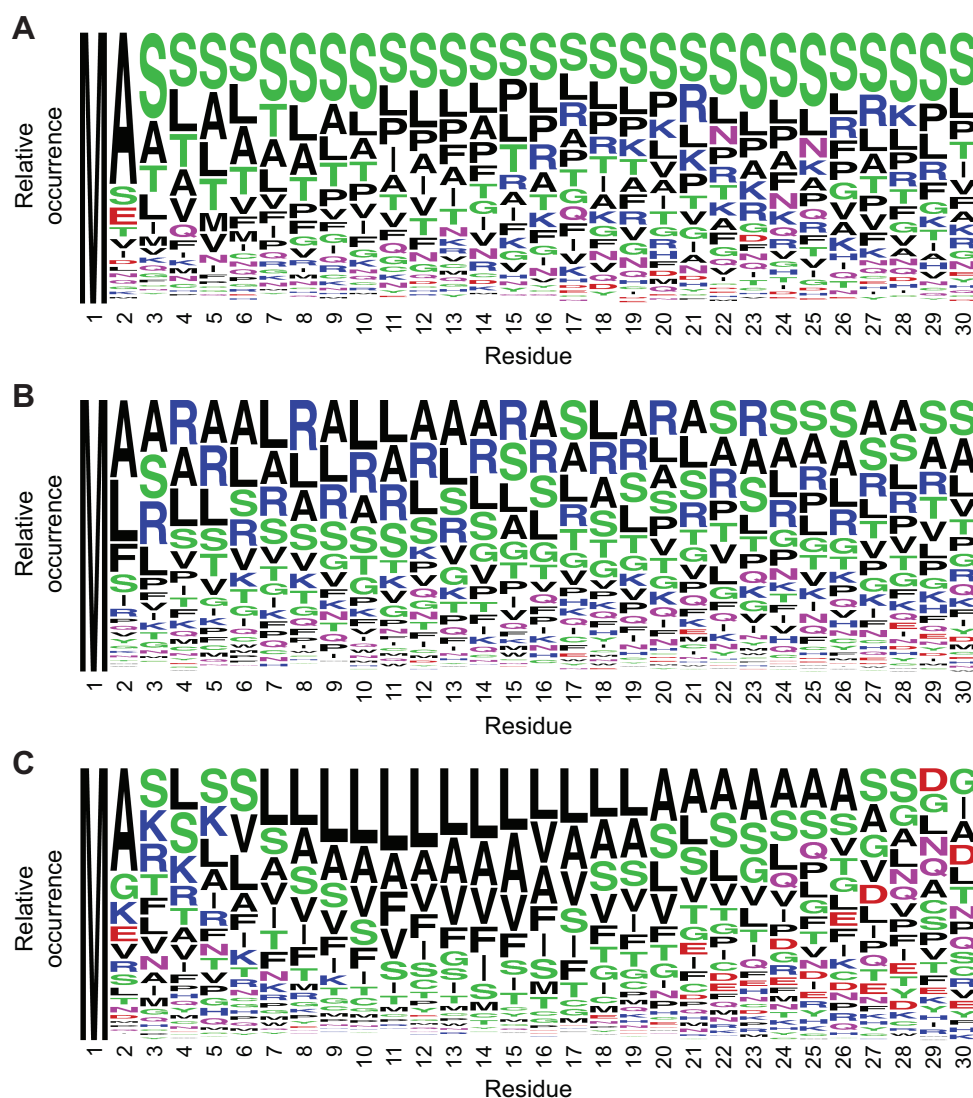


Figure 4-8. Amino Acid Distributions of the N-terminal Sequences of Chloroplastic, Mitochondrial and Secretory Pathway Proteins

(A)-(C) Logo plots of the N-terminal 30 aa of the chloroplastic, mitochondrial and secretory pathway proteins from the 208-TP dataset, the mitochondrial and secretory pathway training sets of TargetP, respectively.

Ala, and hydrophobic Phe (Figure 4-8C). Two separate PSSMs were generated from residues 2-30 of mTPs and SPs from the TargetP training set (see section 2.19.11 for detail).

We further compared the aa compositions of the N-terminal 15 aa domains from TP, mTP and SP sequences to the aa frequencies in the UniProt database (Figure 4-9). Two of the most pronounced differences were the large amount of Arg in mTPs and Ser in TPs. Considering the aa distribution in the UniProt database as an average aa abundance, we calculated the log ratio of the aa frequency in TP, mTP and SPs over the aa frequency in the UniProt database. The ratio indicates the aa frequency difference of the targeting signal from the average. TP, mTP and SP sequences showed higher abundances of Ala, Leu, Val, Met, Arg and Ser and lower levels of Trp, Ile, Gly, His, Asn, Asp, Glu and Cys (Figure 4-9B). TPs also showed higher levels of Ser and Thr and lower levels of Trp and Lys when compared to mTPs and SPs. Thus, the aa composition at the N-terminal domains of TPs, mTPs and SPs are different.

The sequences from the TargetP training set were used to evaluate the PSSM classification. Three PSSM scores were calculated based on TP, mTP and SP matrices for each protein sequence. The distribution of the scores from the chloroplast, cytosolic, mitochondrial, nuclear and secretory pathway proteins were plotted for each matrix (Figure 4-10). The cytosolic and nuclear proteins had the lowest scores in all matrices indicating that they have different N-terminal aa composition from those of TP, mTP and SP. Although the proteins belonging to the same category as the PSSM produced the highest scores, the score distributions from other protein categories still intersected with their scores. The score distributions from the chloroplast and mitochondrial proteins mostly intersected while the score distribution of the secretory pathway proteins partially intersected with both of the chloroplast and mitochondrial distributions. The SP PSSM yielded the best separation of the secretory pathway proteins from the others (Figure 4-10C).

Figure 4-9. Comparison of the Amino Acid Distributions of the 30 aa N-terminal Sequences of Chloroplastic, Mitochondrial and Secretory Pathway Proteins to the UniProt sequences

(A) Aa distribution of individual datasets. The 208-TP sequences were used to represent TPs. The mTP and SP sequences were from the TargetP training set.
(B) Log of the ratio between the aa frequency in TP, mTP or SP sequences over the aa frequency in UniProt database.

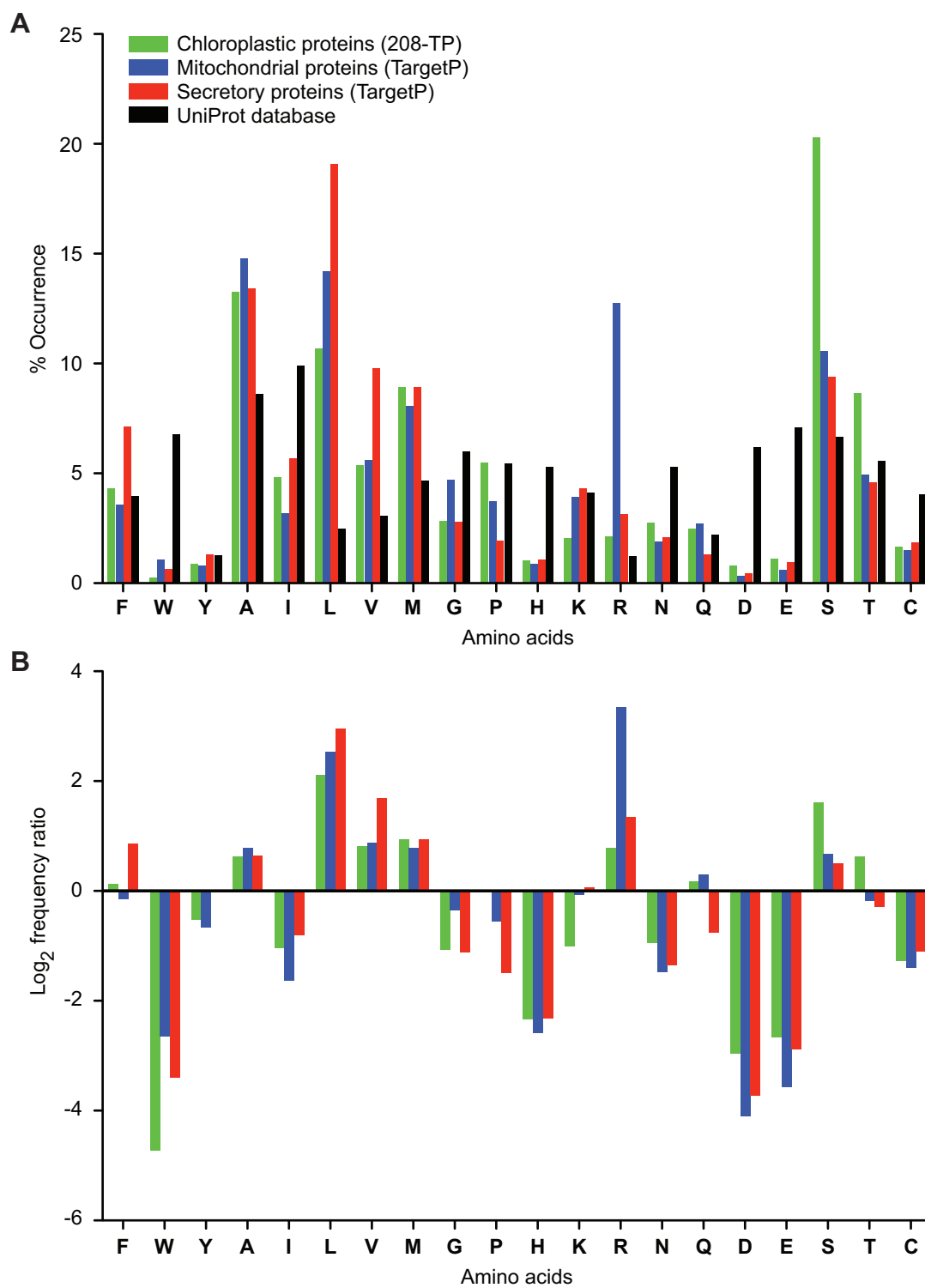
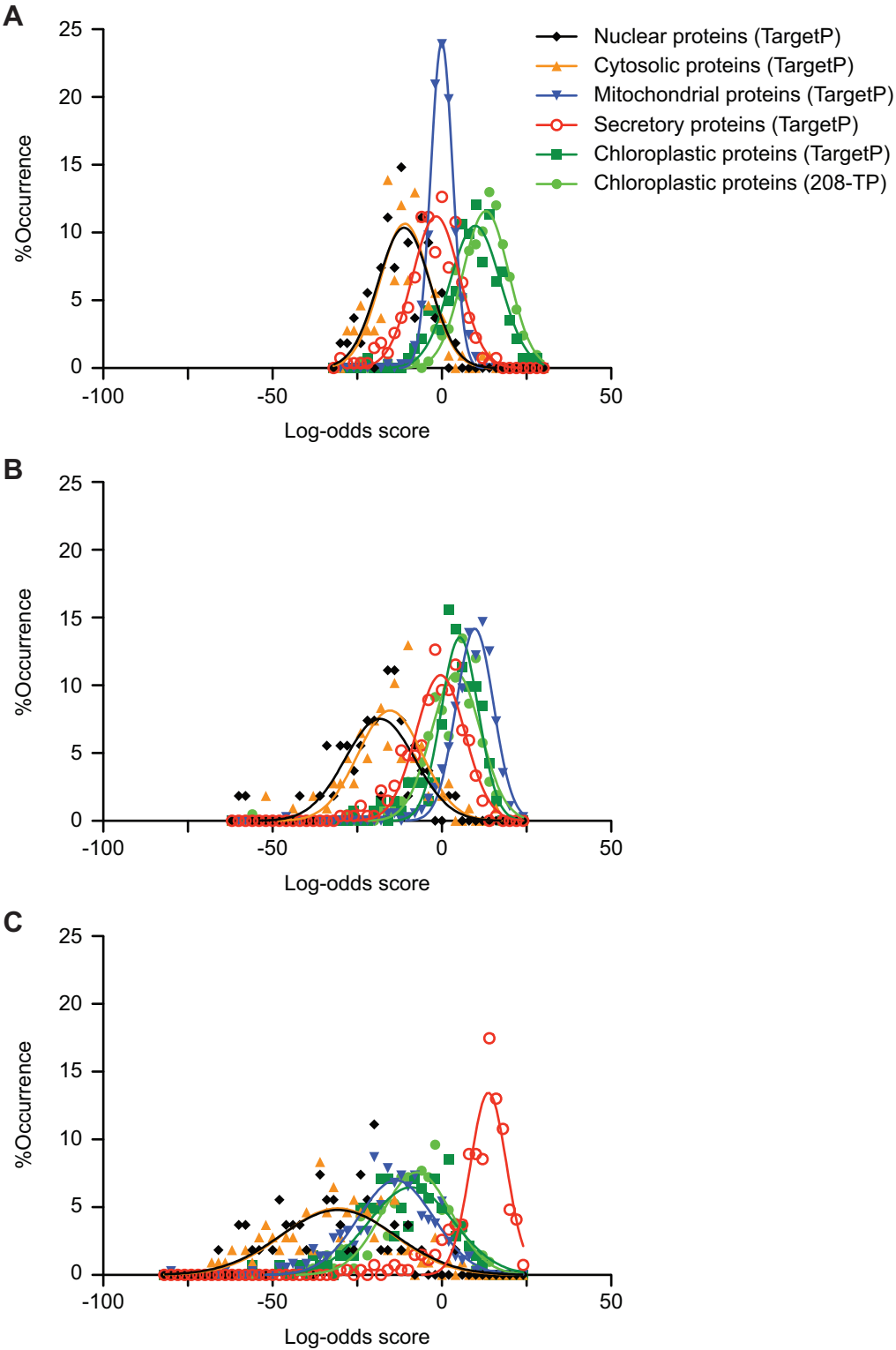


Figure 4-10. PSSM Score Distribution of the Proteins from Different Localizations

The distributions of PSSM scores of proteins from different localizations calculated based on TP (A), mTP (B) and SP (C) PSSMs. Proteins from the TargetP training set were used. The 208-TP dataset was used to generate the TP PSSM while the TargetP mTP and SP dataset were used in generating mTP and SP PSSMs. The lines represent curves fitted to normal distributions.



To improve the localization classification of proteins, we utilized all three PSSMs together. By comparing the scores calculated from all of the three matrices, a protein was predicted to localize to the location that produced the highest score. When the scores from all three matrices were 0 and less, the proteins were predicted to localize to the other locations. The scores of 0 and less indicate equal and less chances, respectively, for the protein to be in the same category as the matrix over random sequences. Table 4-1 shows the evaluation of the combined PSSM classification. Using the PSSM training sequences as a testing set, we found that our method has at least 86% sensitivity and at least 72% specificity. The sensitivity and specificity for chloroplast protein prediction dropped from 90% and 85% to 72% and 82%, respectively, when evaluated with different dataset (TargetP chloroplast proteins) than the PSSM training set (the 208-TP).

4.3.5 *In vivo* import assays

The ability of N-terminal peptide mutants to target to plastids was examined. Onion epidermal cells were transformed with the TP-YFP fusion protein constructs. Figure 4-11 shows representative images of the cells transiently expressing the proteins 12 h after transformation. While most of the constructs were able to target to plastids, pp9-SSR could not be detected in plastids and np09-SSF was not expressed. Four constructs, pp38-SSF, pp9-SSF, PepG-SSR and V10-SSF, had dual-localization to both plastids and mitochondria. The ratios of plastid YFP/cytosolic YFP were measured from at least 20 cells from each construct and are reported in Figure 4-12.

The ratios indicate that the mutants containing weak Hsp70 binding peptide np09 and non-binding peptide HA have the lowest targeting efficiencies at around 1.7 (Figure 4-12A). While the mutants containing strong Hsp70 binding peptides, PepG, V10, DRC8 and A6R have higher targeting efficiencies

Table 4-1. Sensitivity and Specificity of the PSSM Classification of Protein Localization

Dataset	Total proteins	Predicted localizations (%protein)			
		Chloroplast	Mitochondria	Secretory pathway	Others
208-TP	208	90.38	3.37	0.48	5.77
TargetP chloroplast proteins	141	72.34	18.44	0.71	8.51
TargetP mitochondrial proteins	368	3.80	87.77	0.82	7.61
TargetP secretory pathway proteins	269	2.97	2.97	86.62	7.43
TargetP cytosolic proteins	108	2.78	4.63	0.93	91.67
% Sensitivity ^b		90.38 (72.34) ^a	87.77	86.62	90.28
% Specificity ^c		85.68 (82.72) ^a	72.61	94.77	86.03

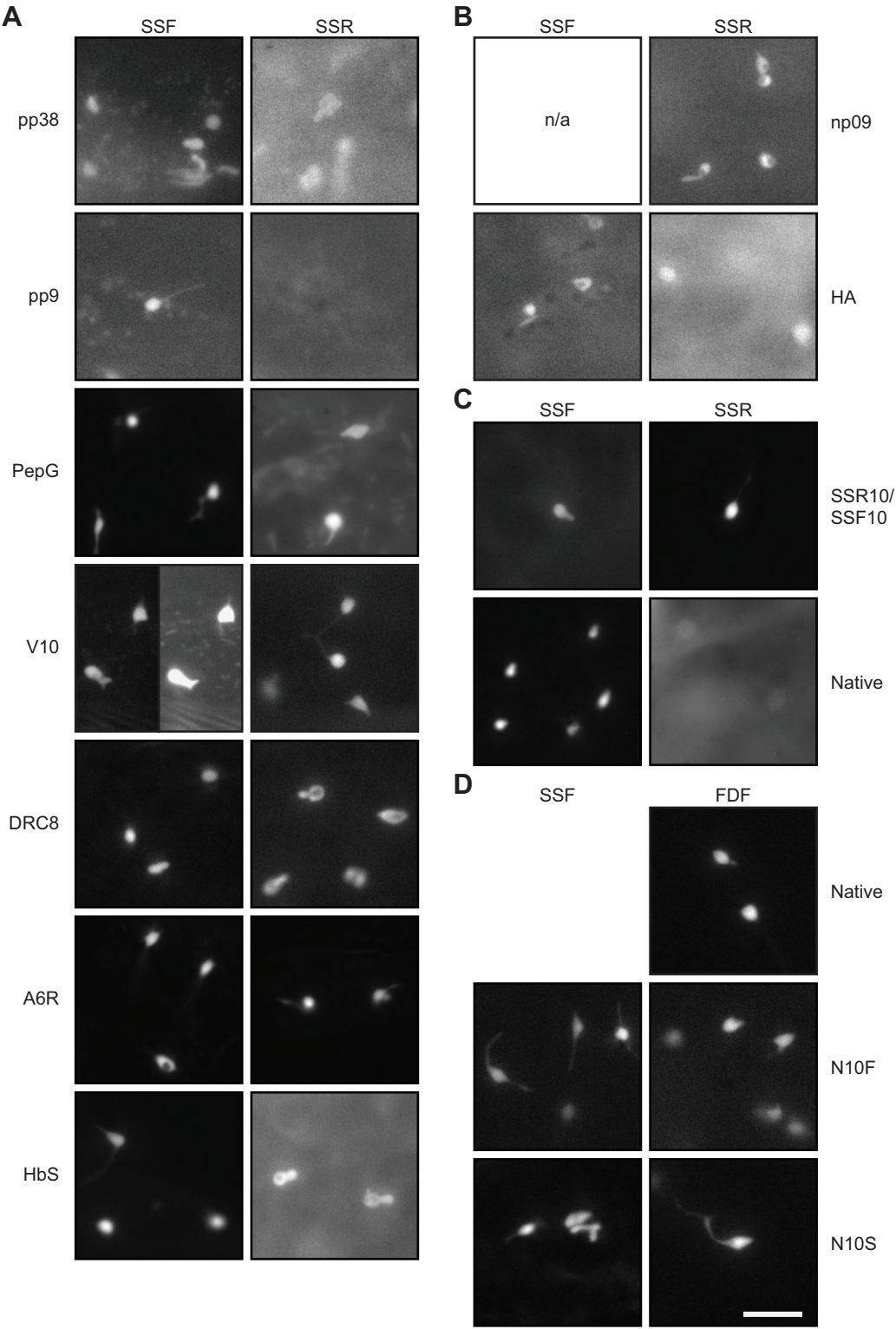
^a Values outside of parentheses were calculated from the 208-TP dataset and the values in parentheses were calculated from TargetP chloroplast proteins .

^b Sensitivity = true positive / (true positive + false negative).

^c Specificity = true positive / (true positive + false positive).

Figure 4-11. Targeting of the N-terminal Mutants in Onion Cells

Representative images of transiently expressed N-terminal mutant TP-YFP fusion proteins in onion epidermal cells. The mutants containing Hsp70-interacting peptides (A), the mutants containing non-interacting peptides (B), the control constructs (C), and the flipped and scrambled mutants (D) are shown. Left and right labels indicate the N-terminal peptides used to generate the mutants. Top labels indicate the original constructs. The pp38-SSF, pp9-SSF, PepG-SSR and V10-SSF fusion proteins show dual-localization to both plastids and mitochondria. N10F and N10S denote the flipped and scrambled mutants. Bar, 10 μm .



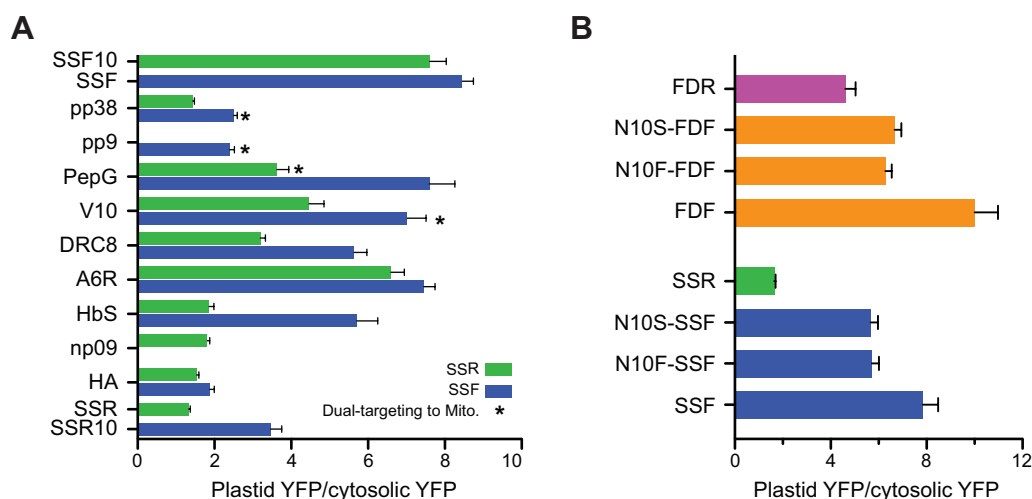


Figure 4-12. Plastid Targeting Efficiency of the N-terminal Mutants

The ratios of plastid YFP/cytosolic YFP were measured from onion epidermal cells transiently expressed the TP-YFP constructs similar to those shown in Figure 4-11. (A) The ratios from the N-terminal peptide constructs. SSF10 and SSR10 from a previous study (Chapter 3) are included. These constructs contain the N-terminal 10 aa of the opposite TP at their N-termini. (B) The ratios from the flipped (N10F) and scrambled (N10S) N-terminal constructs. Means $\pm$ SE are shown. $20 \leq n \leq 30$.

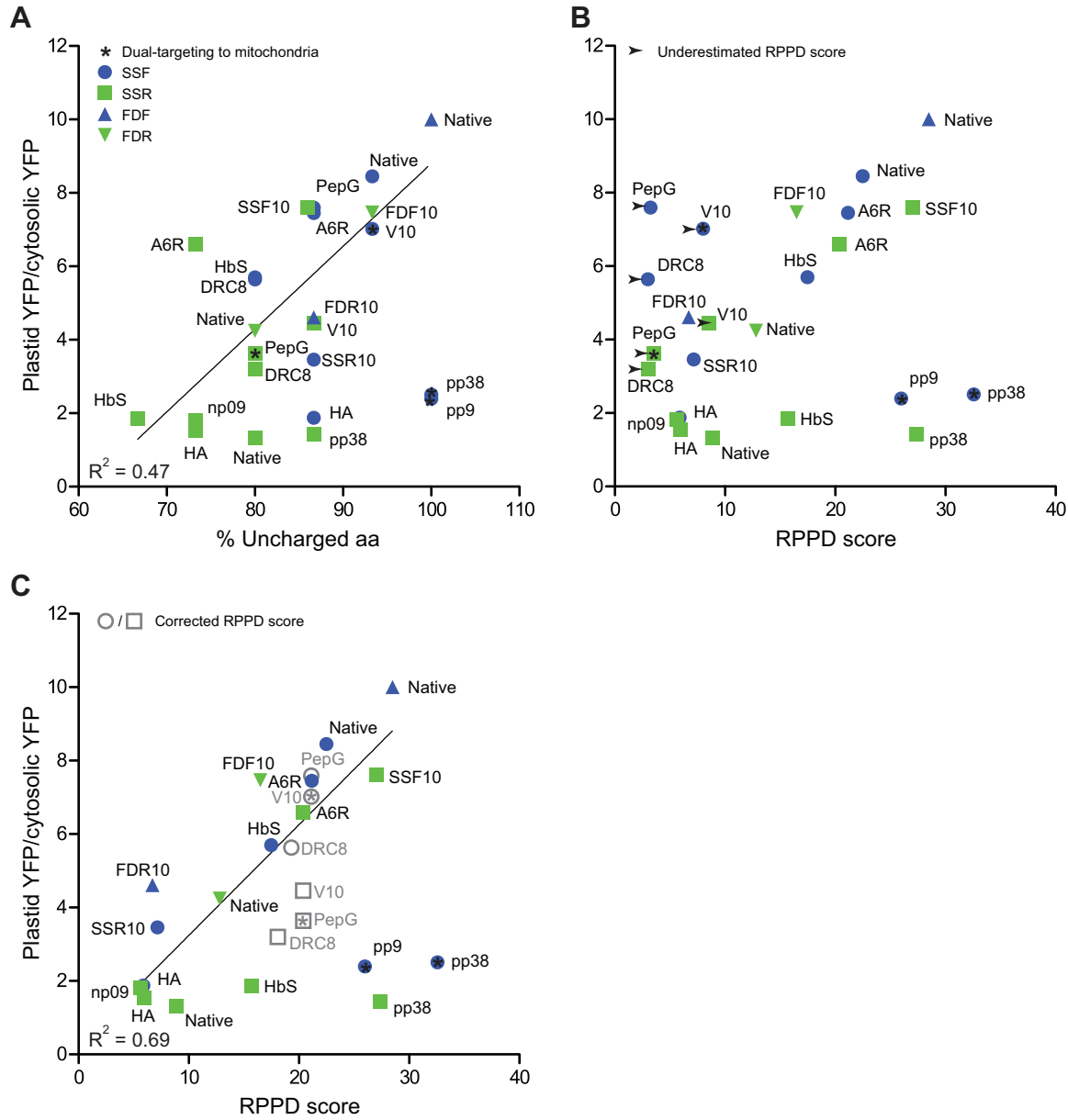
at around 5.5. The moderate Hsp70 binding peptide HbS directs protein imports at moderate efficiencies at around 4.0. Surprisingly, two strong Hsp70 binding peptides pp38 and pp9 have low targeting efficiencies at around 2.0. We also found that in all of the peptides, the constructs based on SSF-20-YFP have higher efficiencies than that based on SSR-20-YFP indicating the influence of the sequence following the N-terminal peptide.

The flipped and scrambled mutants show reduced targeting efficiencies compared to the original constructs (Figure 4-12B). However, these reduced efficiencies are still higher than those of the reverse TP constructs. These results suggest that recognition of the TP N-terminal domain is largely based on physicochemical properties.

When the ratios were plotted against the % uncharged aa values (Figure 4-13A), most of the proteins followed the trend where higher % uncharged aa of the N-terminal domain have higher targeting efficiencies. However, the highly uncharged peptide mutants (pp38-SSF, pp9-SSF and pp38-SSR), showed low import efficiencies. HA-SSF, another mutant with a moderate % uncharged N-terminus also showed low import efficiency. When the correlation was determined from the data excluding the values from the pp38 mutants, the pp9 mutant and the dual-localization mutants, the % uncharged aa correlates to the plastid targeting efficiencies with R^2 of 0.47. As shown in Figure 4-13B, at first, there seemed to be no correlation between the targeting efficiencies with the Hsp70 affinities. However, we knew from Figure 4-6 that RPPD under-predicted the affinities of PepG, V10 and DRC8 peptides. We then corrected the underestimated RPPD scores of PepG, V10 and DRC8 mutants based on experimentally derived affinities (the combined rating scores) shown in Figure 4-5C. Because PepG and V10 peptides have the same combined scores as A6R peptide, the RPPD scores of PepG/V10 SSF and SSR mutants were estimated to be equal to the RPPD scores of A6R-SSF and A6R-SSR mutants, respectively. DRC8 peptide has the combined scores of 4, which is mid-point between A6R

Figure 4-13. Relationships Between Plastid Targeting Efficiency and the Properties of N-terminal Domains of TPs

(A) The plastid targeting efficiency ratios were plotted against the % uncharged aa calculated from the N-terminal 15 aa of the precursors. In addition, FDF, FDR, SSF10, SSR10, FDF10 and FDR10 from previous study (Chapter 3) are included. These constructs with suffix 10 contain the N-terminal 10 aa of the opposite TP at their N-termini. The correlation line was determined without dual-localized, pp9 or pp38 mutants. (B) The plastid targeting efficiency ratios were plotted against the predicted Hsp70 affinity, the RPPD scores, which were calculated from the accumulation of RPPD scores of the N-terminal 15 aa. From Figure 4-6, RPPD algorithm underestimated the Hsp70 affinities of DRC8, PepG and V10 peptides. (C) The plastid targeting efficiency ratios were plotted against the corrected Hsp70 affinity. The corrected RPPD scores were estimated based on the combined rating affinities of DRC8, PepG and V10 to other peptides (Figure 4-5C). The correlation line was determined without dual-localized, pp9 or pp38 mutants.



and HbS peptides with the scores of 6 and 3, respectively. The RPPD scores of DRC8 mutants were estimated to be the averaged RPPD scores between A6R and HbS mutants. After the correction (Figure 4-13C), the targeting efficiencies and the Hsp70 affinities showed a correlation with R^2 of 0.69 with the exclusion of the values from the pp38 mutants, the pp9 mutant and the dual-localization mutants. If the RPPD scores of underestimated mutants were excluded together with the pp38 mutants, the pp9 mutant and the dual-localization mutants, the correlation analysis showed R^2 of 0.76 (if the same values were excluded from the % uncharged aa versus the targeting efficiency analysis, the correlation was determined to have R^2 of 0.49). The pp9 and pp38 mutants were the only group that diverted from the RPPD correlation. Unlike the % uncharged aa plot, the efficiencies of HA mutants agreed with the RPPD scores. Thus, in both the % uncharged aa and the RPPD score plots, the pp9 and pp38 mutants were different from other mutants.

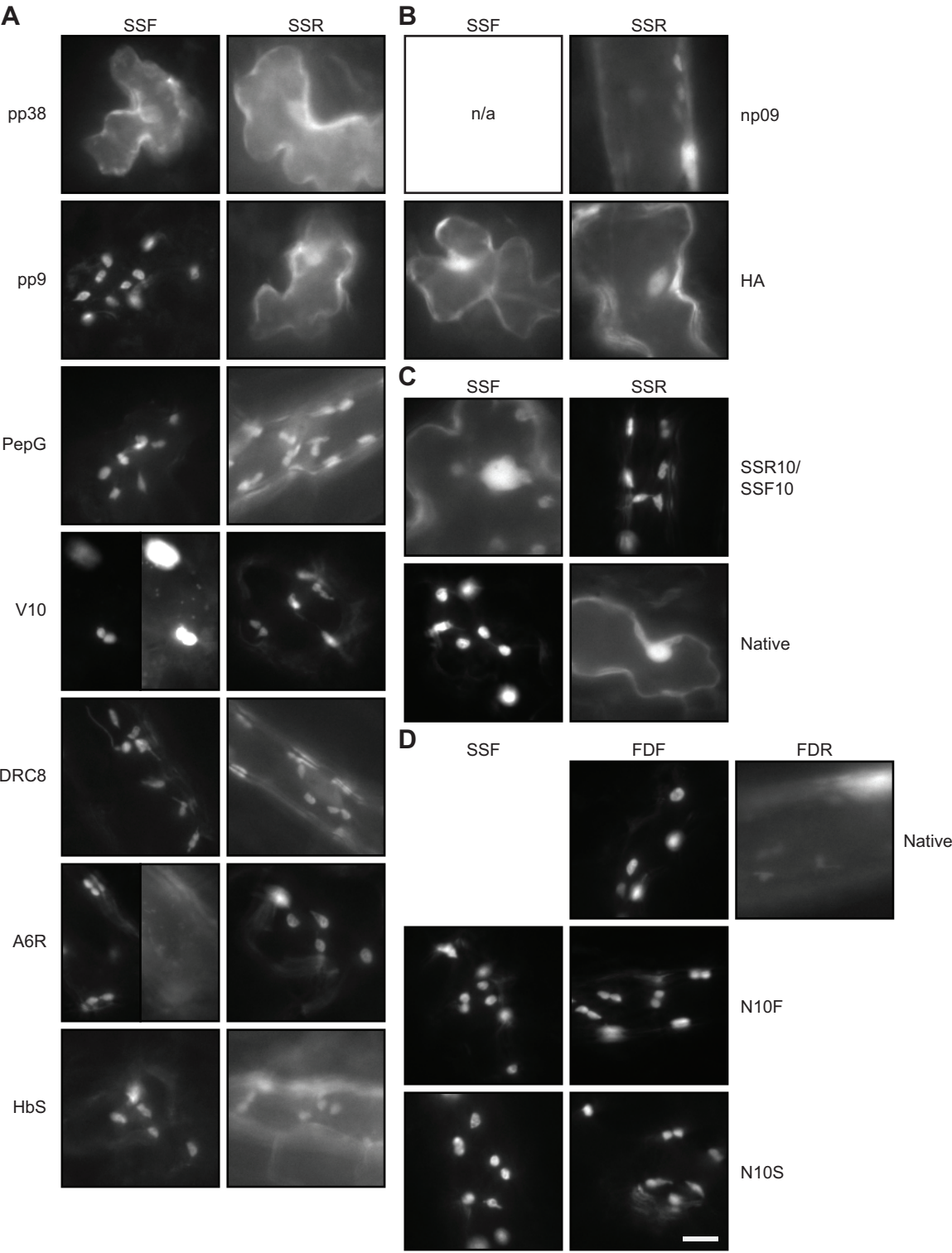
To confirm the targeting results from onion cells, the constructs were also transiently expressed in *Arabidopsis* seedlings. Figure 4-14 shows the localization patterns of these constructs, 2 days after transformation. The localizations are essentially similar to those of onion cells except that dual-localization was only observed from V10-SSF and A6R-SSF constructs.

4.3.6 *In vitro* import assays

To determine the import rate of the N-terminal mutant constructs, *in vitro* import assays using isolated pea chloroplasts were employed. The precursors were labeled with ^{35}S -Met via *in vitro* translation. We have performed the assays in 2 batches. The translation products from each batch are shown in Figure 4-15. Because *in vitro* translations produced not only precursor proteins but also degraded protein species similar to what has been observed in *E. coli* expressions (data not shown), the translation products were separated by SDS-PAGE

Figure 4-14. Targeting of the N-terminal Mutants in *Arabidopsis* Cells.

Representative images of cells transiently expressing the N-terminal mutant TP-YFP fusion proteins. The mutants containing Hsp70-interacting peptides (A), the mutants containing non-interacting peptides (B), the control constructs (C), and the flipped and scrambled mutants (D) are shown. Left and right labels indicate the N-terminal peptides used to generate the mutants. Top labels indicate the original constructs. The V10-SSF and A6R-SSF fusion proteins show dual-localization to both plastids and mitochondria. Bar, 10 μm .



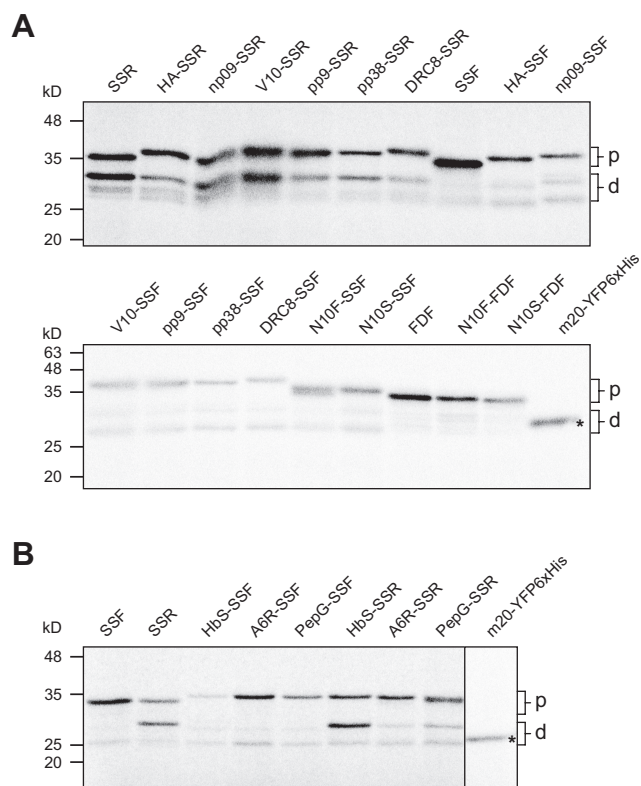


Figure 4-15. Autoradiographs of *In Vitro* Translation Products of the N-terminal Mutants

Two batches of translations were performed. SDS-PAGE was used to separate 1 μ l of the translation products. (A) and (B) show the autoradiographs of the translation products from batches 1 and 2, respectively. Top labels, constructs; p, precursor size; d, degraded products; asterisk, mature domain size.

followed by autoradiography before quantities of precursor proteins were determined from the autoradiographs. In each batch, equal quantities of precursors were used in the import assays. The quantities used between two batches were not the same. Figures 4-16 and 4-17 show the import time courses of the precursors where the amounts of import-processed mature domains were quantified over time. Some precursors showed an early plateau of import within 5 min while many precursors did not plateau at 15 min of import. Instead of using a fixed time point to determine the import rate, the highest import rates among the 3 time points were used. The import rates are shown in Figure 4-18. Surprisingly, V10 mutants import at higher rate than SSF.

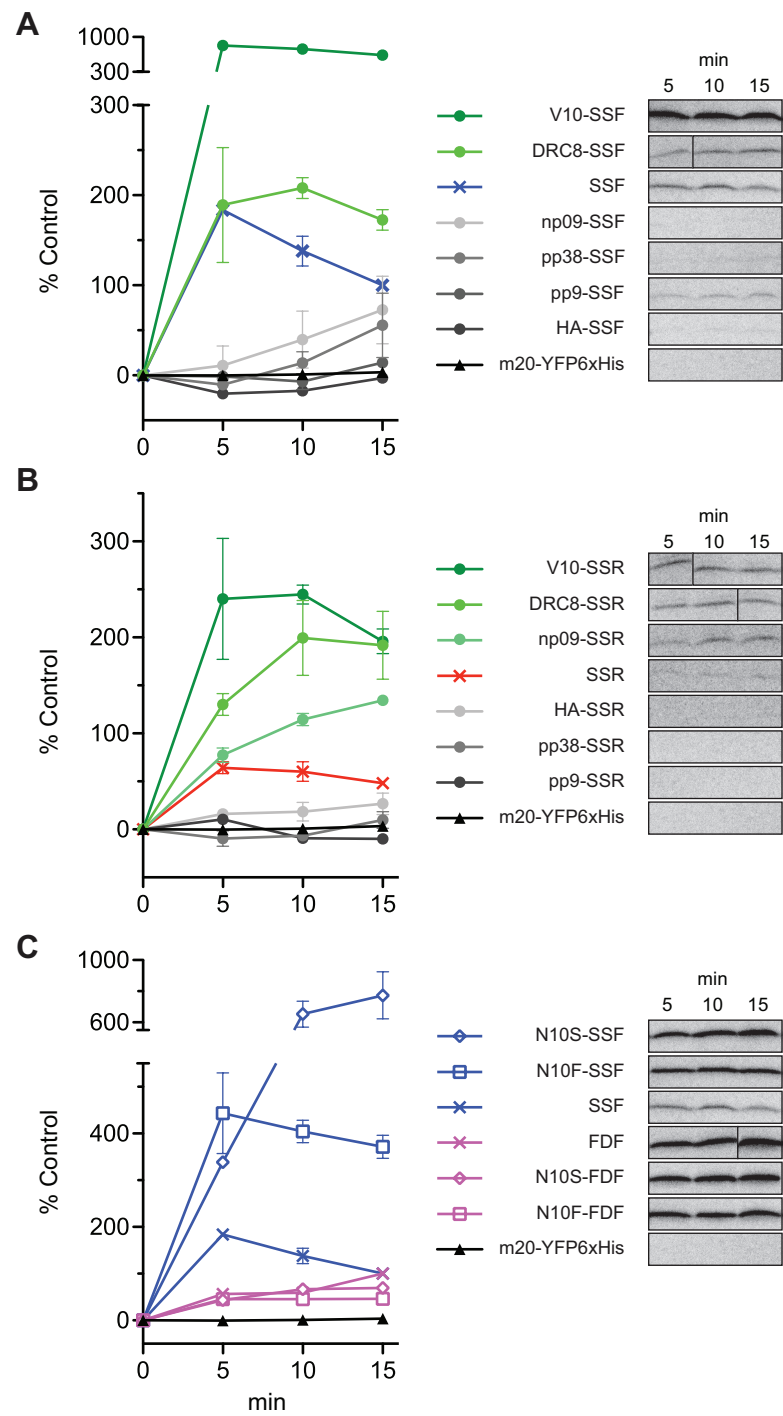
To compare the import rates with the targeting efficiencies, the targeting efficiency ratios of the precursors were plotted against the import rates (Figure 4-19A). This plot indicates that the high targeting efficiency precursors have a large distribution of import rates while the low efficiency precursors have a narrow range. We further plotted the targeting efficiencies against the log transformed import rates (Figure 4-19B and C). The precursors from translation batch 1 shows weak correlation with R^2 of 0.28 while the precursors from batch 2 shows a strong correlation of the targeting efficiency with the log import rate producing R^2 of 0.94 when PepG-SSF was excluded.

4.4 Discussion

The highly uncharged characteristic of the N-terminal domain of TPs (von Heijne et al., 1989) has been associated with the binding step in plastid protein import since the discovery of the specific interactions of this domain with chloroplast lipids and Toc159 receptor (Lee et al., 2008; Lee et al., 2006; Lee et al., 2009a; Lee et al., 2002; Pilon et al., 1995; Pinnaduwa and Bruce, 1996; Rensink et al., 1998). However, in the previous chapter we identified a novel role

Figure 4-16. Import Time Course of Precursor Produced in the First Batch

Equal molar concentration of precursor proteins were used in the import assay. Two sets of assays were performed. The constructs based on SSF-20-YFP (A), SSR-20-YFP (B) and the flipped and scrambled constructs (C) are shown. Representative autoradiographs of the import-processed mature domains are shown on the right. The amounts of mature domains were measured at 5, 10, and 15 min after import started. The values from the SSF and SSR series were normalized where the amount of mature domain of SSF construct at 15 min was assigned as 100%. The amount of mature domain of FDF at 15 min was assigned as 100% for FDF series.



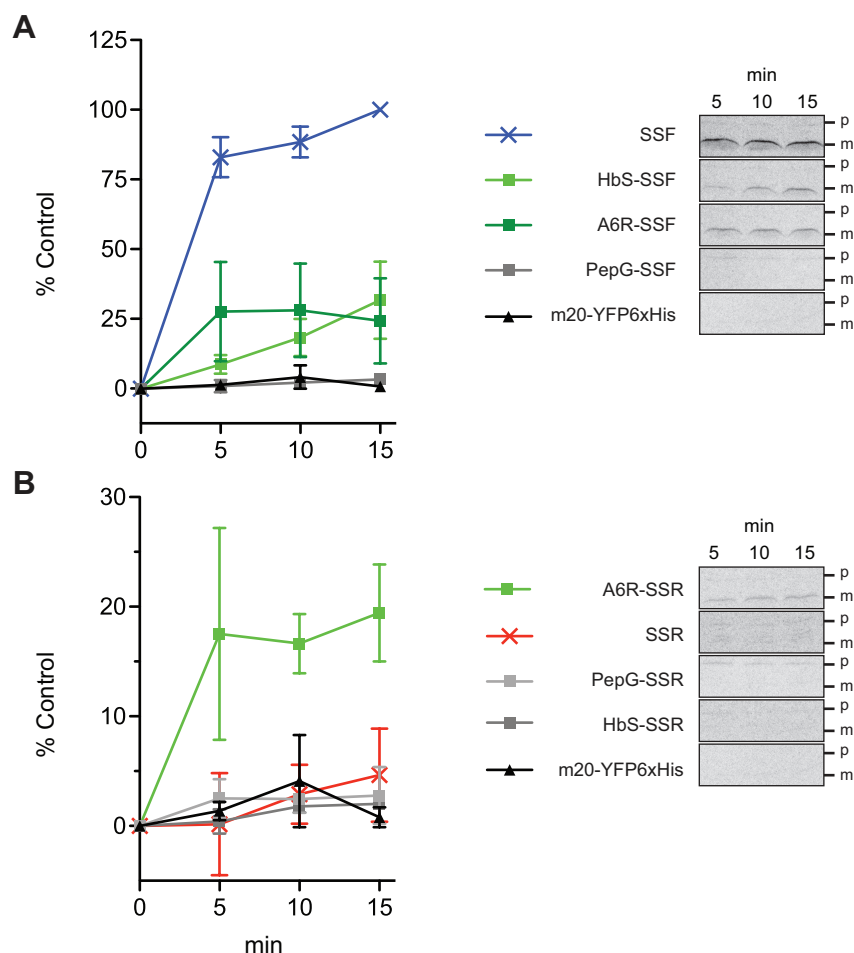


Figure 4-17. Import Time Course of Precursor Produced in the Second Batch

Equal molar concentration of precursor proteins were used in the import assay. Two sets of assays were performed. (A) and (B) are constructs based on SSF-20-YFP and SSR-20-YFP, respectively. Representative autoradiographs of the import-processed mature domains are shown on the right. The amounts of mature domains were measured at 5, 10, and 15 min after import started. The values were normalized where the amount of mature domain of SSF construct at 15 min was set as 100%.

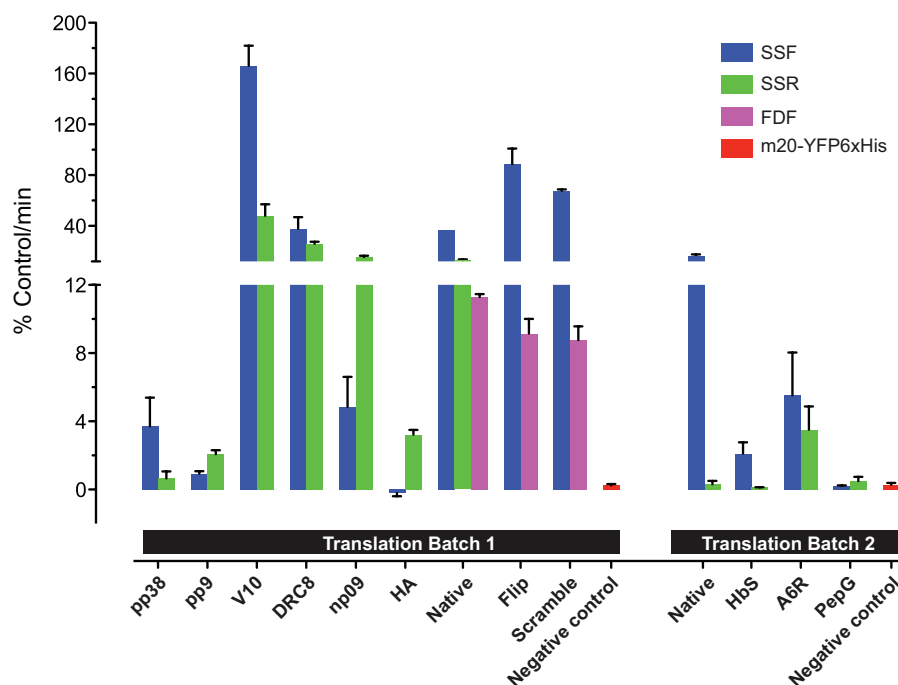
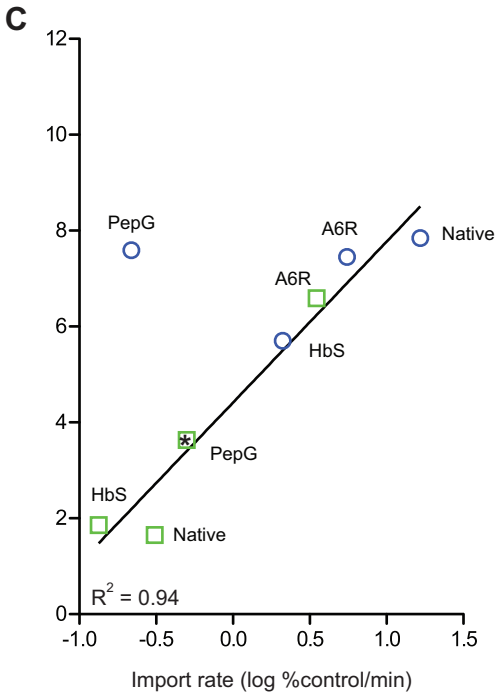
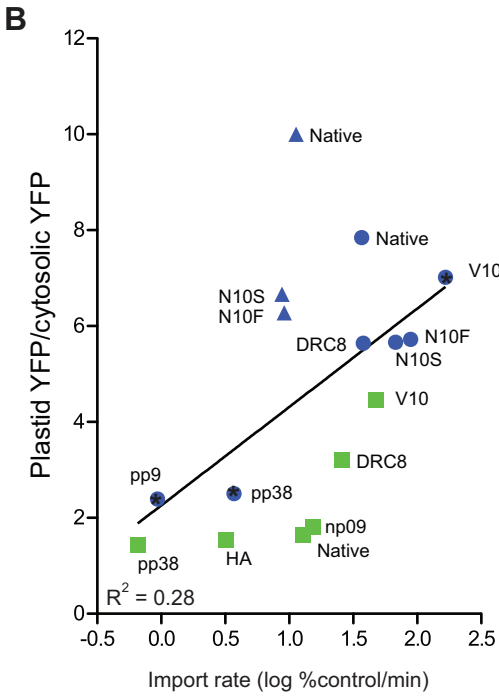
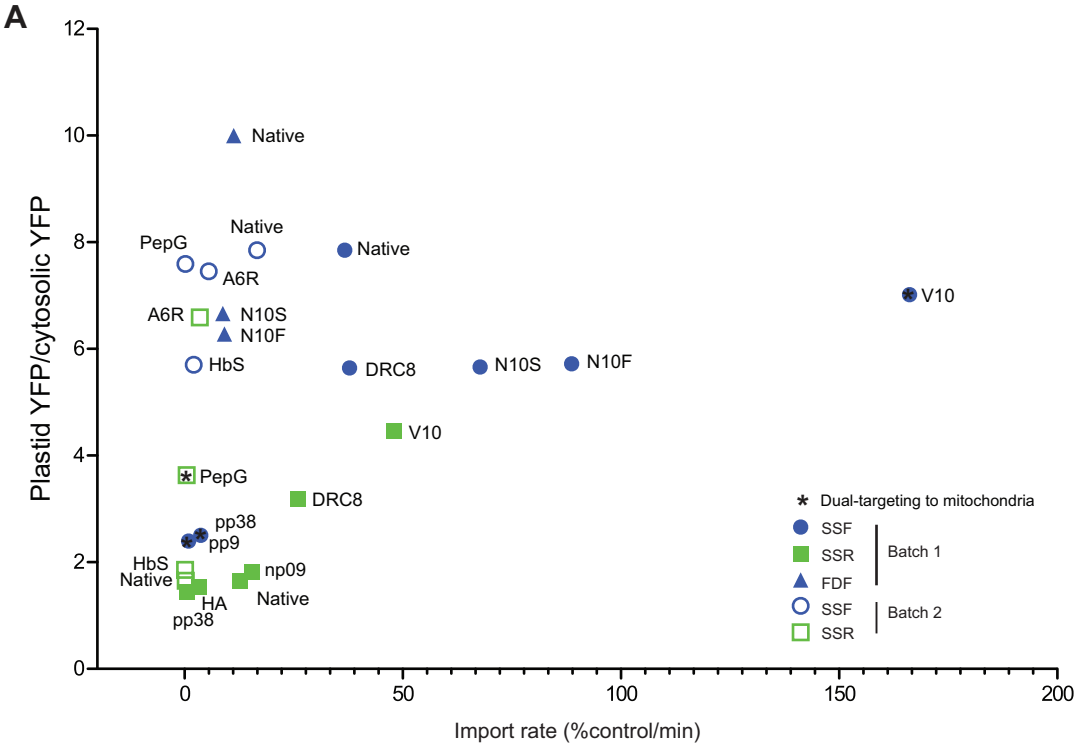


Figure 4-18. Maximal *In Vitro* Import Rates of the N-terminal Mutant Precursors

The *in vitro* import rates were derived from the highest import rates from 5 to 10 min of import. Two separate sets are shown based on the batch of translation. The values from m20-YFP6xHis, SSF and SSR series were normalized to the amount of mature domains from 15 min import of SSF-20-YFP (100%). The FDF series were normalized to the amount of mature domains from 15 min import of FDF-20-YFP. The precursor without TP, m20-YFP6xHis, was used as a negative control.

Figure 4-19. Comparison of Plastid Targeting Efficiencies and Import Rates of the N-terminal Mutants

(A) The plastid targeting efficiency ratios are plotted against the rates of import. (B) The plastid targeting efficiency ratios of protein translated in the first batch are plotted against the log transformed rates of import. (C) The plastid targeting efficiency ratios of protein translated in the second batch are plotted against the log transformed rates of import.



of the TP N-terminal domain as a major determinant for protein translocation into plastids. Based on the two well-known properties of the TP N-termini, highly uncharged (von Heijne et al., 1989) and Hsp70-interacting (Ivey and Bruce, 2000; Ivey et al., 2000; Rial et al., 2000), we proposed that the Hsp70-interacting property allows the TP N-termini to interact with the stromal Hsp70 chaperone in trans to initiate the translocation process (Chapter 3). To further characterize the function of TP N-terminus, we replaced this region with a series of peptides with varying affinities to Hsp70 and determined their effects on protein translocation into plastids both *in vivo* and *in vitro*.

In general, each TP contains multiple Hsp70 binding sites (Ivey et al., 2000; Rial et al., 2000; Zhang and Glaser, 2002). Using CBPS prediction (Rudiger et al., 1997b), more than 75% of TPs were found to contain at least one Hsp70 binding site (Rial et al., 2000; Zhang and Glaser, 2002). A more detailed analysis, using both RPPD (Gragerov et al., 1994; Ivey et al., 2000) and CBPS (Rudiger et al., 1997b) algorithms indicated that 75% of TPs in CHLPEP database (von Heijne et al., 1991) have the strongest Hsp70 binding site at their N-terminal regions (Ivey et al., 2000). However, we found that in the 208-TP dataset populated with *Arabidopsis* TPs, only 32.21% have the strongest RPPD Hsp70 binding site within the first 20 aa (Figure 4-1). This difference is possibly due to redundancy within the CHLPEP database that contains all of TPs known at that time (von Heijne et al., 1991). For example, we found at least 53 instances of SStp from different species in CHLPEP. Nonetheless, our 32.21% value may be an underestimated because RPPD predicts Hsp70 binding sites with high specificity but low sensitivity (Figure 4-6). Some of the Hsp70 binding sites may not be identified by RPPD. Although 75% of TPs contain at least one Hsp70 binding site, at least a third of TPs contain the strongest Hsp70 binding site at their N-termini.

TP N-termini are highly uncharged (von Heijne et al., 1989). We found that in average, the N-terminal 10 aa region of TPs in the 208-TP dataset has

94% uncharged aa (Figure 4-2A) which is similar to our analysis of the 912-TP dataset (Chapter 3). The transition point between the highly uncharged and the moderately uncharged regions was found to be at residue 14 (Figure 4-2A), which is almost the same as the transition point at residue 15 found in the 912-TP dataset (Chapter 3). Considering only this highly uncharged 15 aa region, we found that 43% of TPs contain purely uncharged aa in this region (Figure 4-2B). When the TPs from groups 1-3 (Figure 4-1) that contain the strongest Hsp70 binding site within the first 20 aa were analyzed, 36% of them are completely uncharged in their N-terminal 15 aa. These results indicate that more than a third of TPs have purely uncharged N-terminal regions regardless of their N-terminal Hsp70 affinities.

It has been shown multiple times that Hsp70 proteins recognize hydrophobic peptides (Fourie et al., 1994; Gragerov et al., 1994; Rudiger et al., 1997b). It was proposed that the hydrophobic region of proteins termed the hydrophobic core fits into the substrate cavity of Hsp70 proteins while the charged flanking regions interact with the surrounding area (Rudiger et al., 1997a). In the study by Fourie et al. (1994), all hydrophobic peptides interacted with Hsp70s while only some charged peptides interacted. Thus, it is not surprising that over a third of TPs in groups 1-3 containing strong Hsp70 binding N-termini have a purely uncharged N-terminus. This also indicates a challenge in separating the Hsp70-interacting function from the uncharged property.

The peptide set that we used in studying the function of TP N-termini is composed of 9 peptides that are 8-12 aa in length. We showed that all of these peptides, except A6R, were predicted to be different from the TP N-terminus (Figure 4-7C). Two of the peptides, pp38 and pp9, are purely uncharged and strongly interact with Hsp70s (Figure 4-5A and C). Other peptides contain charged aa and range from 67 to 91% uncharged. The % uncharged aa range of the peptides is comparable to the % uncharged range of the TP N-termini (Figure 4-3). Two of the peptides, np09 and HA, are weakly interacting and non-

interacting to Hsp70s, respectively (Figure 4-5A). Thus, our peptide set contains both Hsp70 interacting and non-interacting peptides but all of the purely uncharged peptides interact with Hsp70s.

For each of the chosen peptides, a pair of N-terminal mutants was generated from both SSF and SSR-20-YFP fusion constructs where the peptide sequence was fused at the N-termini of both constructs (Figure 4-5C). The effect of the peptides in directing YFP targeting was first observed using *in vivo* plastid protein import assays performed in both onion epidermal cells (Figure 4-11) and *Arabidopsis* seedlings (Figure 4-14). The results from both *in vivo* assays are essentially the same. We further determined the *in vivo* targeting efficiencies of the mutants from the images acquired from the *in vivo* assays in onion. In each pair, the mutant based on SSF-20-YFP had higher efficiency than the SSR-20-YFP mutant (Figure 4-12A). Thus, efficiency is not only affected by the N-terminal peptide sequence that was added to generate the mutant, but also the SSF or SSR sequences. This finding is in agreement with the sequence analysis (Figures 4-2A and 4-8A) where the TP N-terminal domain was shown to be about 15-aa long which is longer than the Hsp70 interacting peptides.

To compare the *in vivo* import efficiency with the % uncharged aa and the Hsp70 affinity values, we calculated the values based not only on the added peptide sequence in the mutants but rather the whole N-terminal 15 aa. We found a positive correlation in both the % uncharged aa and the Hsp70 affinity plots for most of peptide mutant precursors (Figure 4-13). In general, precursors with higher % uncharged aa or Hsp70 affinity N-terminal domains had higher import efficiencies. However, the pp9 and pp38 mutants containing N-terminal domains with purely uncharged aa and strong Hsp70 affinity, were import-deficient. Nonetheless, the HA mutants having moderate % uncharged N-termini similar to other import-competent peptides, PepG and A6R, failed to be imported into plastids. On the contrary, the RPPD scores of HA mutants predicting low Hsp70 affinity are in agreement with the low import efficiencies of

these mutants. Thus, the HA mutants indicate that the N-terminal domain of SS_{tp} utilizes Hsp70 interaction in the import process.

In contrast to other mutants, neither the % uncharged aa nor the Hsp70 affinity of the N-terminal domains of the pp9 and pp38 mutants predicted the low import efficiencies. This showed that their N-terminal domains do not meet all of the necessary functions for the plastid protein import. Figure 4-20 shows the process of plastid protein import, which involves reversible binding, irreversible import intermediate steps and translocation. The precursors containing the N-terminal Hsp70-interacting peptides were competent in targeting into plastids (Figure 4-13) indicating that the TPs had all of the requirements for every step. Because the TP N-terminal domain is required for the translocation step (Chapter 3), we hypothesize that the failure of pp9 and pp38 peptides occurs during binding and/or intermediate steps.

The sequences of pp9 and pp38 contain 2 and 3 Trp residues, respectively (Figure 4-5A). Trp is the rarest aa in the N-terminal domain of TP (Figure 4-9). We propose that Trp residues in pp9 and pp38 interfere with the binding or intermediate steps of the mutant precursor protein import possibly by interacting with the membrane lipids. Further mutagenesis analysis of these peptides may uncover the requirements for the binding or intermediate steps.

Our results indicate that at least a third of TPs utilize the stromal Hsp70 interactions to initiate the translocation process based on their N-terminal Hsp70-interacting domain, while the TPs lacking an N-terminal Hsp70-interacting domain may utilize other stromal chaperones such as Hsp93. Bruch et al. (2012) showed that a TP lacking the N-terminal Hsp70-interacting domain, the pea ferredoxin-NADPH reductase TP (FNR_{tp}), was found to interact with Hsp93. When all of the Hsp70 binding sites in FNR_{tp} were mutated, the mutant TP (FNR_{tp}-1234) was able to support protein import (Rial et al., 2003). These results suggest that FNR_{tp} may utilize the N-terminal Hsp93-interacting domain in initiating the translocation.

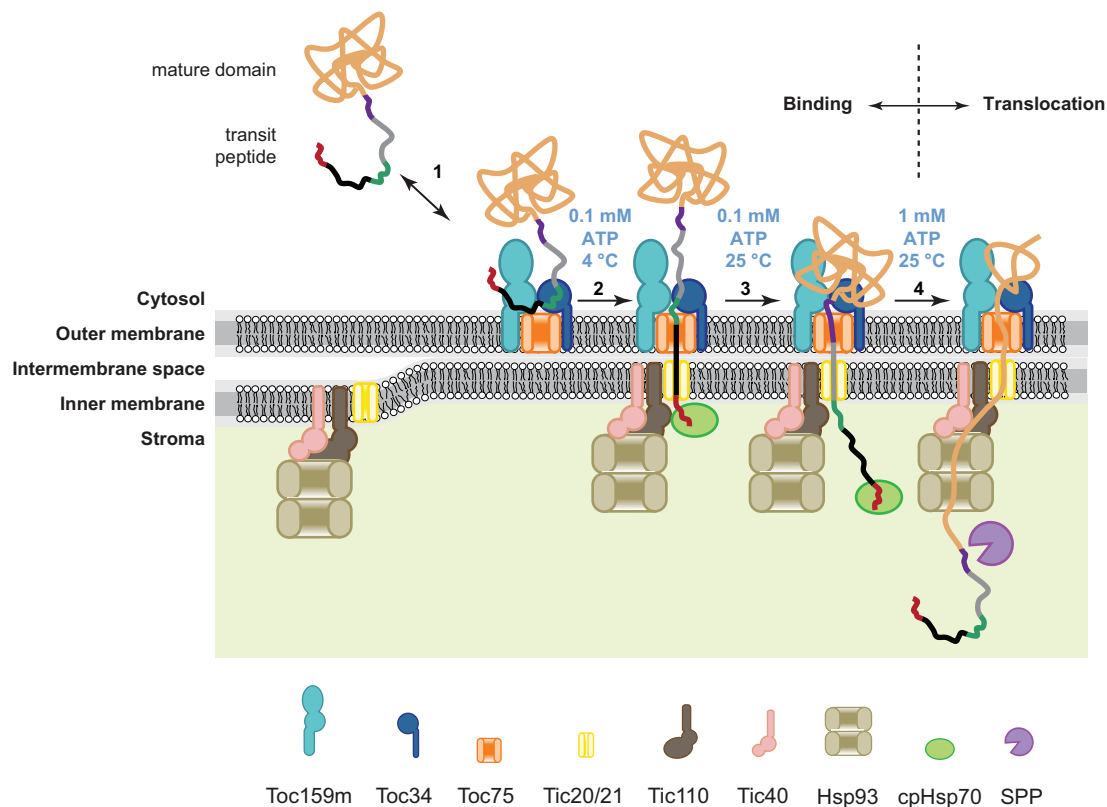


Figure 4-20. Steps in Plastid Protein Import

Cytosolic precursors bind to plastids by reversible energy-independent binding (step 1) when ATP/GTP is lacking (Kouranov and Schnell, 1997; Perry and Keegstra, 1994). In 0.1 mM ATP, the irreversible early import intermediates forms. At 4 °C, about 110 aa are buried in the translocons (step 2) (Akita and Inoue, 2009; Inoue and Akita, 2008a). At 25 °C, about 130 aa are buried (step 3) (Akita and Inoue, 2009; Inoue and Akita, 2008a). The translocation process is initiated by the ATPase chaperones Hsp93/cpHsp70 when the ATP level is above 1 mM (step 4) (Shi and Theg, 2010; Su and Li, 2010).

4.5 Conclusions

Our bioinformatic analyses, *in vitro*, and *in vivo* assays of SStp mutants revealed that SStp utilizes the N-terminal Hsp70-interacting domain during the translocation step in plastid protein import. This suggests that about 32% of TPs containing the strong N-terminal Hsp70-interacting domains may function in the same manner. Whether other TPs lacking an N-terminal Hsp70 binding site interact with other stromal chaperones such as cpHsp93 as we have proposed (Chapter 3) still has to be determined. Nevertheless, it is still unknown how this N-terminal Hsp70-interacting domain functions in relation to other TP domains such as the FGLK motif.

Chapter 5

Role of the Hsp70-FGLK Spacer Length in Plastid Protein Import

5.1 Abstract

The majority of chloroplast proteins are nuclear-encoded and utilize an N-terminal transit peptide (TP) to target and translocate into chloroplasts via the general import pathway. Although analysis of plant genomes was fruitful in providing over ten thousand predicted TP primary sequences, it is still poorly understood what constitutes a TP and how these components facilitate TP function. We have previously shown that the N-terminal sequence of TPs is a major determinant for the plastid protein translocation. A subset of TP N-terminal domains functions as Hsp70-interacting domains, which were proposed to interact with the stromal translocon motor Hsp70. Here, we identified the locations of the N-terminal Hsp70-interacting sites with respect to the proposed outer envelope translocon receptor Toc34 binding site (FGLK motif) and observed a conserved distance between these sites. When the Hsp70-FGLK spacer lengths were altered, we observed that the most efficient translocation occurred only at an optimal spacer length of around 28 to 31 aa. This result supports our proposed bimodal interaction model where the productive translocation requires a temporal and/or spatial coupling between a "capturing step" by a TOC receptor and a "trapping/pulling" step by a stromal ATP-dependent molecular motor.

5.2 Introduction

It was estimated that about 1,500 precursor proteins in *Arabidopsis* utilize N-terminal targeting sequences called transit peptides (TPs) in directing their post-translational targeting to plastids (Kleffmann et al., 2004; Richly and Leister, 2004; Zybailov et al., 2008). These proteins are crucial for plastid functions such as photosynthesis, respiration and metabolism (Kleffmann et al., 2004; Kleine et al., 2009). TPs are necessary and sufficient in governing the precursor protein translocation into the plastid stroma through the translocons at the outer and inner envelope membranes of the chloroplasts (TOC/TIC) (Bruce, 2000; Bruce, 2001). However, little is known about how TPs accomplish their function(s).

The primary sequences of TPs are highly divergent (Bruce, 2000). Sequence analysis only identified a few short conserved peptide motifs when TPs were grouped into small groups. No consensus motif has been identified from the entire set of TPs (Lee et al., 2008). Still, three weakly conserved domains have been identified: (i) N-terminal domain of about 10 uncharged residues ending with Pro/Gly and preferably having Ala as the second residue, (ii) central domain, lacking acidic aa but rich in hydroxylated aa, and (iii) C-terminal domain, rich in Arg and possibly forming an amphiphilic beta-strand (Bruce, 2001; von Heijne et al., 1989). TPs also lack any secondary structure. They form random coils in aqueous solution (Bruce, 1998; von Heijne and Nishikawa, 1991).

We showed that the N-terminal domain of TPs of the small subunit of ribulose-1,5-bis-phosphate carboxylase/oxygenase (SStp) and ferredoxin (FDtp) are major determinants for protein translocation into plastids (Chapter 3). Mutant TPs lacking the wild-type N-terminal domain failed to be translocated into the stroma suggesting that this domain interacts with an element in the stroma (Chapter 3). In Chapter 4, the role of this domain was further examined. Unrelated Hsp70-interacting peptides were able to substitute for the function of

the wild-type N-terminal domain of SStp indicating that at least a subset of TPs utilize the N-terminal Hsp70-interacting domain to interact with stromal Hsp70 to initiate the translocation process.

The interactions between TP and TOC receptors (Toc34 and Toc159 GTPases) have been shown repeatedly (Jelic et al., 2002; Reddick et al., 2007; Schleiff et al., 2002; Sveshnikova et al., 2000b). However, it is not known how and where Toc34 binds to TP. The shortest peptide shown to directly bind to Toc34 is the B1 peptide corresponding to aa 22-47 of SStp from tobacco (*Nicotiana tabacum*) (Schleiff et al., 2002). In fact, this region in SStp from pea (*Pisum sativum*) and *Arabidopsis* were shown to contain two FGLK motifs (Pilon et al., 1995). The deletion of the second FGLK motif from pea SStp diminished TP's ability to bind to the isolated chloroplasts (Subramanian, 2001). Sequence analysis of targeting sequences also found that FGLK motif can be used to discriminate chloroplast TPs from other targeting sequences (Chotewutmontri et al., 2012). Hence, we proposed that the FGLK motif is required for Toc34 recognition (Chotewutmontri et al., 2012).

We observed a conserved placement of the FGLK motif in relation to the N-terminal Hsp70 domain. The spacer distances between the N-terminal Hsp70 10-aa domains and the FGLK motifs in *Arabidopsis* and pea SStp, *Silene latifolia* FDtp, and *Arabidopsis* nucleotide transporter 1 range from 22 to 25 aa (Chapter 3). We proposed a “bimodal interaction model” describing this interconnection between stromal Hsp70 and surface Toc34 interacting domains as a key spacing requirement for coordinating translocation of the preprotein across both membranes at the contact sites where the TOC and TIC complexes are tightly compressed (Chapter 3).

Here, we expanded the Hsp70-FGLK spacer analysis to cover a set of 67 TPs containing a strong N-terminal Hsp70 interacting sequence (Chapter 4) derived from the dataset of 208 experimentally verified TPs (Lee et al., 2008). Series of SStp and FDtp mutants with varying spacer lengths were generated and

their import efficiencies were determined using *in vivo* protein import assays. The results indicate the importance of spatial coupling between the N-terminal Hsp70-interacting domain and the Toc34-interacting FGLK motif in plastid protein import.

5.3 Results

5.3.1 Bioinformatic analysis of the Hsp70-FGLK spacer

Previous analysis using only 4 TPs and their mutants indicated that the import-efficient TPs have similar spacer lengths. Their distances between the N-terminal Hsp70-interacting domain (10 aa) and the Toc34-interacting FGLK domain ranged from 22-25 aa (Chapter 3). To further investigate this property, a subset of the 208 experimentally verified TPs from *Arabidopsis* was used (Lee et al., 2008). This subset contains 67 TPs (the 67-TPs dataset) from the Hsp70 cluster groups 1-3 (Chapter 4) that harbor a strong Hsp70-interacting domain within their N-terminal 20 aa.

The random peptide phage display-derived algorithm (RPPD) (Gragerov et al., 1994; Ivey et al., 2000) was used to predict Hsp70 binding sites in TPs. RPPD calculated scores along the length of TP using a 6-aa sliding window and assigned RPPD scores to the residues at position 3 of the windows. The averaged RPPD scores along the lengths of TPs from each of the cluster groups are shown in Figure 5-1. Each group showed a distinct location of the strongest Hsp70 sites as reported in Chapter 4. The TPs from groups 1-3 contain a strong Hsp70-interacting domain in their N-termini.

The FGLK motif positions were identified using another algorithm. Based on the heuristic algorithm for detecting FGLK motif developed by McWilliams (Chotewutmontri et al., 2012), a modified version was developed to incorporate an 8-aa sliding window scoring function and to assign score values for each of the

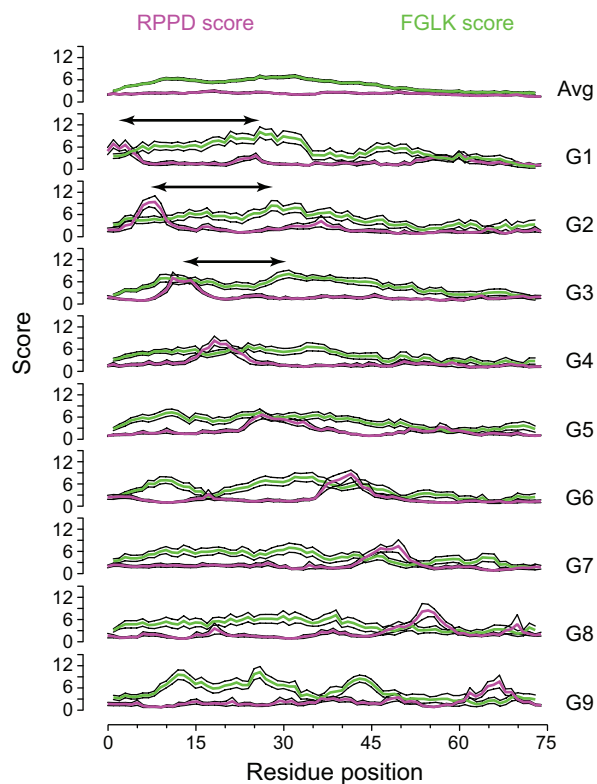


Figure 5-1. Predicted Hsp70-interacting Sites and FGLK Motifs in the 208-TP Dataset

TPs are shown as clusters G1 to G9 (right labels) based on the Hsp70 binding profile clustering (Chapter 4). The RPPD score (magenta lines) predicts Hsp70 affinity. Higher RPPD score indicates higher affinity. The FGLK score (green lines) indicates the level of the FGLK detection criteria that were satisfied. When all of the criteria are satisfied, the score of 16 is given indicating that the sequence contains a FGLK motif. Means $\pm$ SE are plotted (black lines). Black arrow bars indicate the distances between the N-terminal Hsp70 binding sites and FGLK motifs.

algorithm criteria. The FGLK score was assigned to the residues at position 4 of the windows. When a sequence satisfies all of the criteria, it will have the possible maximal score of 16 indicating that this sequence contains a FGLK motif. The FGLK scores are shown in Figure 5-1. Unlike the Hsp70 sites, the FGLK motif positions are more conserved. The FGLK motifs located between residues 10-15 and 25-30 seem to be conserved in all of TPs.

Instead of measuring the distance from the last residue of the Hsp70-interacting domain to the first residue of the FGLK motif as previously described (Chapter 3), we measured the distances between the Hsp70-interacting peaks and the FGLK peaks. For each TP, the Hsp70 peak is defined as the position within the N-terminal 15 aa with the highest RPPD score while the FGLK peak is defined as the position with the FGLK score of 16. This definition allows multiple FGLK peaks to be identified from a TP. To reduce the variation in the distance analysis, the multiple adjacent FGLK peaks that form a plateau were reduced to a single FGLK peak, represented by the center position of the plateau. The Hsp70-FGLK spacer lengths were calculated from the FGLK peaks (after the reduction) minus the Hsp70 peaks. Figure 5-2 shows the distribution of the Hsp70-FGLK spacer lengths. When the data was fitted to a Gaussian distribution, the mean of the Hsp70-FGLK spacer lengths was found to be about 24 aa.

5.3.2 Design of novel Hsp70-FGLK spacers

In order to study the spacer length effect, one of the aims was to design novel spacers for the generation of the spacer length mutants. Because TP sequences are highly divergent and lack any consensus motif (Bruce, 2000; Lee et al., 2008), the function of TPs is possibly depend on the composition of the sequences. As shown in the previous Chapters, the binding of TPs to the chloroplasts and the TP-Toc34 interactions (Chapter 3), and the Hsp70

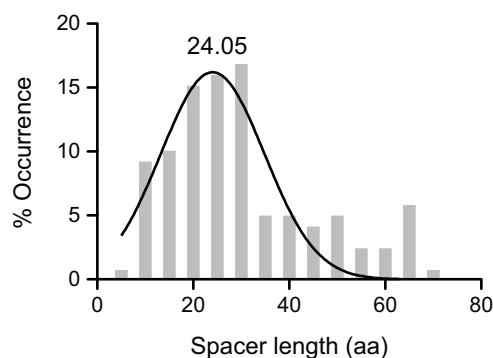


Figure 5-2. The Hsp70-FGLK Spacer Length Distribution

The distances between the Hsp70 peaks and the FGLK peaks were measured from the 67-TP dataset. The Hsp70 peak is defined as the position with the highest the RPPD score within the N-terminal 15 aa. The FGLK peak is defined as the position with the maximal FGLK score of 16. The distribution was fitted to a Guassian distribution represented by the black line. The mean of the Hsp70-FGLK spacer length distribution was determined to be 24.05 ± 10.67 (mean $\pm$ SD) aa. $n = 118$.

recognition of the TP N-termini (Chapter 4) were shown to be dependent on the sequence compositions. Thus, we proposed to design the novel spacers based on the observed aa composition of the wild-type TP spacers.

For each TP of the 67-TP dataset, only a sequence of the spacer with the closest length to 24 aa (the mean of spacer length) was extracted. These sequences contain the shoulder areas of both Hsp70 and FGLK peaks. Four residues were removed from both N- and C-termini of the sequences to eliminate these peak shoulders. The aa distribution of the whole spacer sequences are shown in Figure 5-3. The sequences were further separated into 3 equal regions and the aa distributions of the N-terminal, central and C-terminal regions were determined (Figure 5-3). The distributions from different regions were approximately the same indicating that there is no sequence bias in different regions of the spacers.

Based on the averaged spacer length of 24 aa, we planned to generate the spacer mutants with the lengths from 14 to 34 aa. Considering that part of the spacer length includes the shoulders of Hsp70 and FGLK peaks, only the non-shoulder sequence was altered to maintain the integrity of Hsp70 and FGLK peaks (Figure 5-4A). To generate a mutant of 34-aa spacer, the 26-aa spacer sequence is needed to replace the non-shoulder sequence. Thus, we first designed the 26-aa spacers, which later were shortened to generate the smaller spacers.

Using the aa frequencies of the entire spacer sequences, a pool of 400 random sequences of length 26 aa was generated (detail in section 2.19.12). To minimize the effect of additional Hsp70 and FGLK domains within the spacer sequences, we screened these sequences with the RPPD and the FGLK prediction programs. Three sequences, numbers 92, 228 and 296, were selected which lack additional Hsp70 or FGLK peaks. The sequences of these designed spacers and their predicted scores are shown in Figure 5-4B.

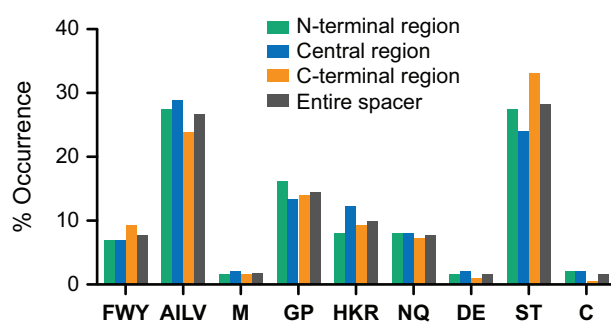


Figure 5-3. Amino Acid Distributions of the Hsp70-FGLK Spacers

The whole spacer sequences were divided into 3 equal regions: N-terminal, central and C-terminal. The aa distributions determined from the different regions and the entire spacer are shown.

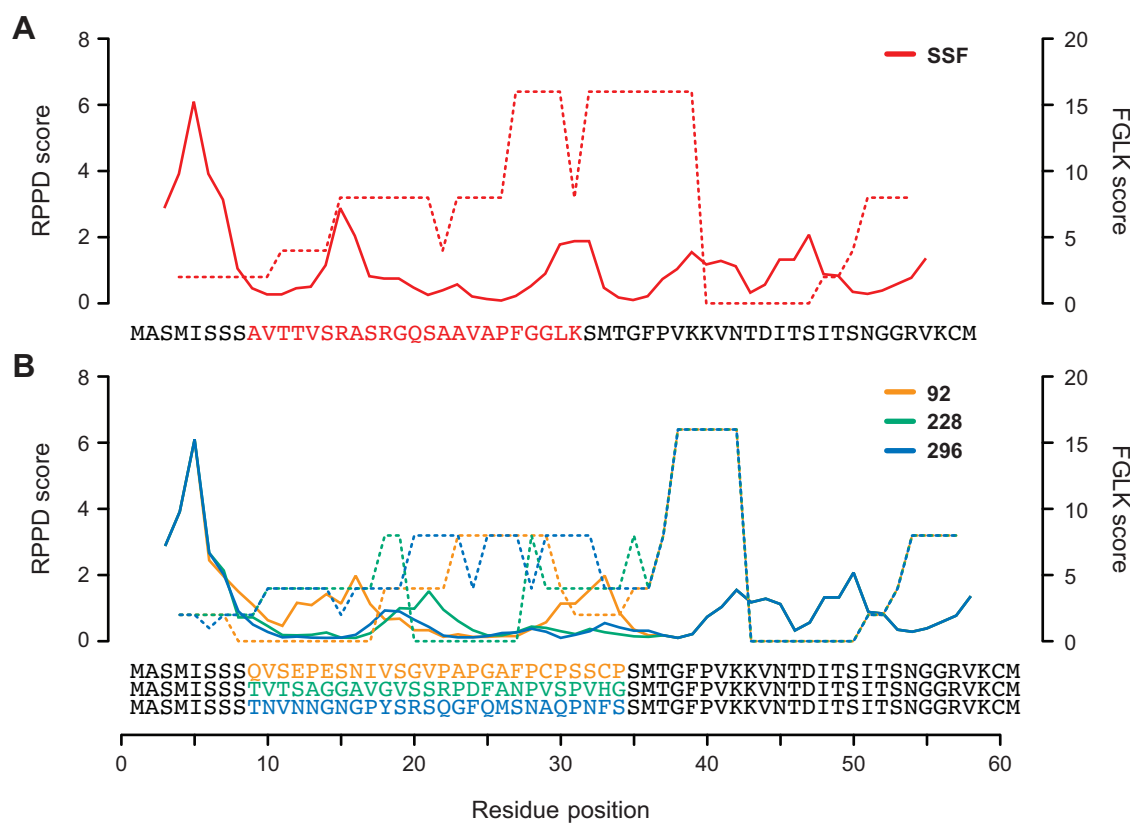


Figure 5-4. Sequences and the Hsp70 and FGLK Prediction Scores of the Wild-type and Designed Spacers

The wild-type spacer of SStp (A) was replaced with the designed spacers (B). The Hsp70-interacting site (RPPD score in solid lines) and the FGLK motif (dashed lines) were predicted. Out of 400 designed sequences, three sequences (numbers 92, 228 and 296) were selected because they lack the additional Hsp70 and FGLK peaks.

5.3.3 Construction of TP mutants containing different Hsp70-FGLK spacer lengths

To determine the effect of the Hsp70-FGLK spacer length in plastid protein import, we utilized TP-YFP fusion constructs previously generated in Chapter 3. The first set of mutants was constructed based on the designed spacers. The wild-type spacer of SSF (SSF) in the SSF-20-YFP construct (SSF fused to 20 aa of the mature domain of prSSU followed by YFP) was replaced with the sequences of the designed spacer numbers 92, 228 and 296 (Figures 5-4, 5-5A, and 5-5B). Note that the first FGLK of SSF was also removed to produce a single FGLK peak. The designed spacers are 26-aa long, which produced the 34-aa spacer constructs. The other spacer length mutants were generated in different ways. (i) The deletion constructs shown in Figure 5-5B were generated by C-terminal deletion of the spacer sequences. These constructs have different total TP lengths. (ii) The extended constructs (Figure 5-5C) were generated by addition of the extra sequence at the C-termini of the spacers. The N-terminal 10 aa of the spacers in the 44-aa constructs were removed to generate the 44-10 constructs that have the same spacer lengths as the 34-aa constructs but contain different sequences. (iii) The equal-length constructs were generated from the 34-aa constructs. The FGLK motifs were moved to different locations to produce different spacer lengths (Figure 5-5D). The equal-length constructs have the same total TP lengths as the 34-aa constructs.

Another set of the spacer length mutants was constructed using the wild-type spacer sequence of SSF. Figure 5-6A shows the wild-type SSF construct. The SSF spacer was divided into 5 regions. We generated the mutants with different spacer length by deletion and/or addition of the sequences of these regions (Figure 5-6B). While the first FGLK motifs were removed from the mutants containing longer spacer lengths than that of wild type, the first FGLK motifs are present in some of the shorter spacer length mutants. At least 3

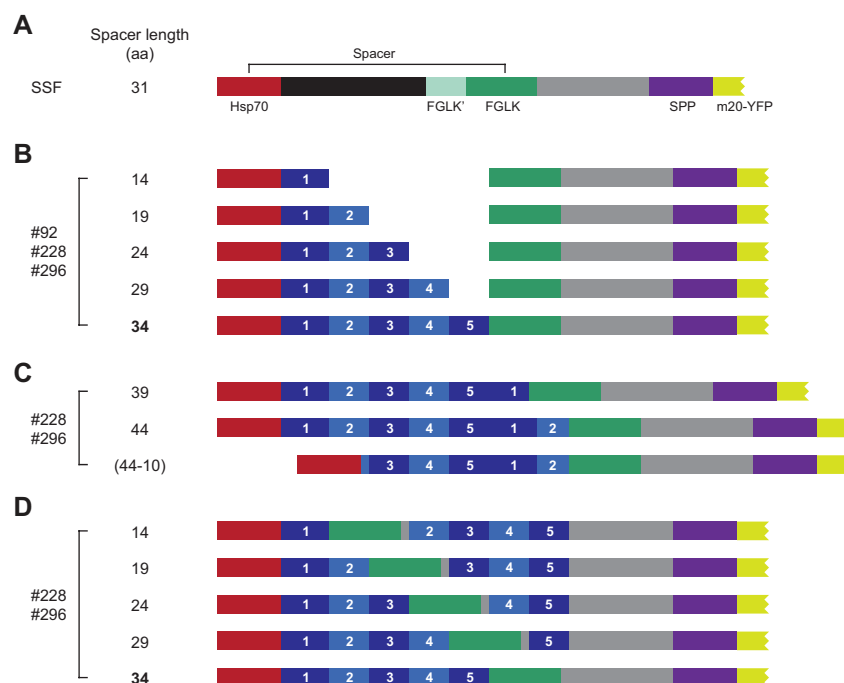
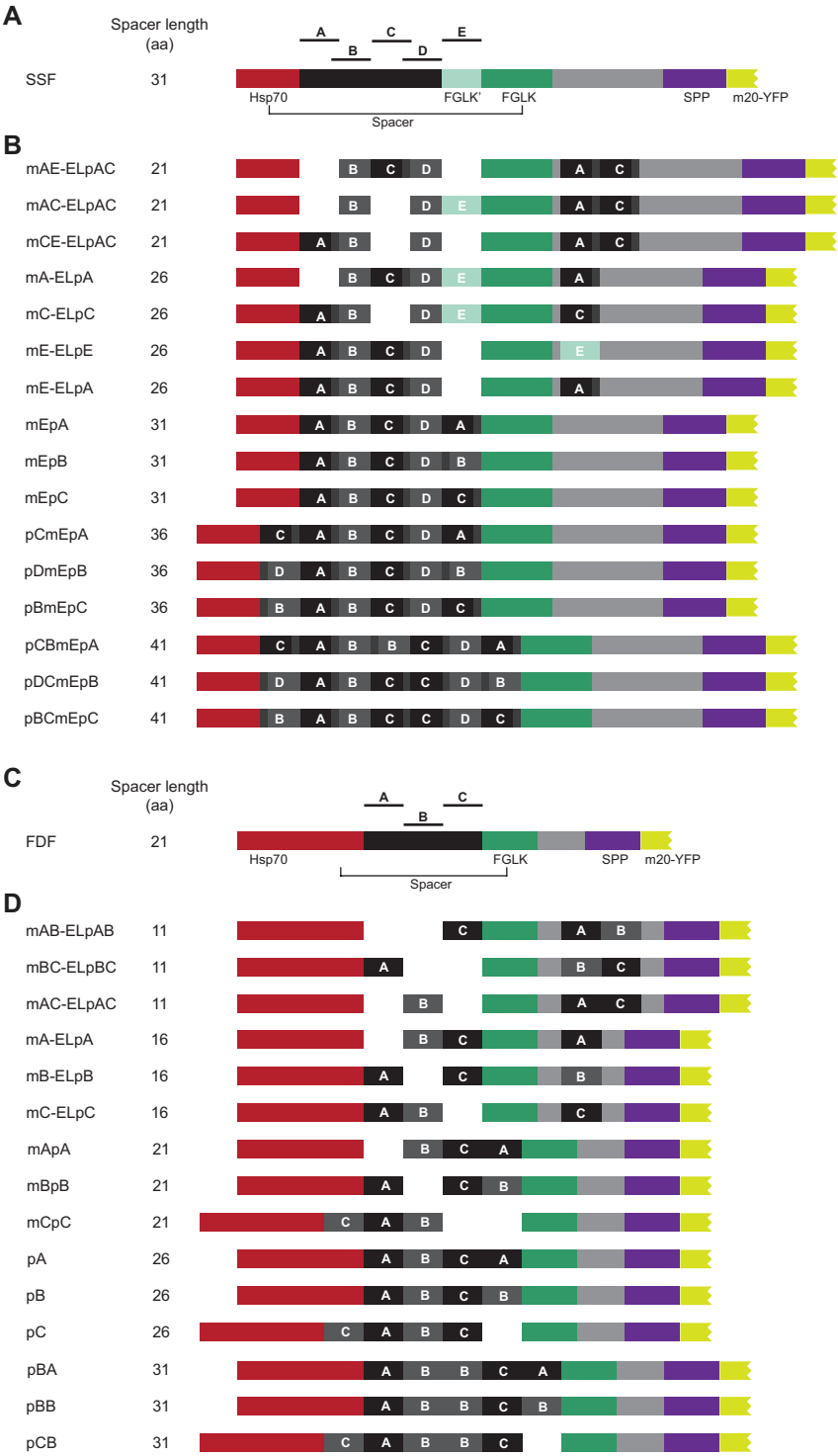


Figure 5-5. Constructs Based on the Designed Spacers

(A) The wild-type SSF-20-YFP construct containing SSF fused to 20 aa of the mature domain of prSSU followed by YFP. (B) The deletion constructs generated from three designed spacers. The 5-aa sequences were removed from the C-termini of the designed sequences at a time to generate the 14-aa to 29-aa spacer mutants. (C) The extended constructs generated from the spacer numbers 228 and 296 by addition of extra sequences in tandem. The N-terminal deletion of the spacer sequence also used to generate the variants of 34-aa mutants, the (44-10)-aa mutants. (D) The equal-length constructs. The FGLK motifs in the 34-aa constructs of both the 228 and 296 mutant series were moved to other locations to generate different spacer lengths. The mutants have the same total length of TP as the 34-aa mutants. Different regions in the designed spacer sequences are numbered. The N-terminal Hsp70-interacting domain, FGLK motifs, stromal processing peptidase recognition site (SPP) and the partial YFP are shown in different colors. The constructs are drawn to scale.

Figure 5-6. Constructs Based on the Wild-type Spacers

(A) The wild-type SSF-20-YFP construct. (B) The spacer length mutant constructs containing different combination of the SSF spacer regions. (C) The wild-type FDF-20-YFP construct. (D) The spacer length mutant constructs containing different combination of the FDF spacer regions. Different regions in the wild-type spacer sequences are labeled. The N-terminal Hsp70-interacting domain, FGLK motifs, stromal processing peptidase recognition site (SPP) and the partial YFP are shown in different colors. The constructs are drawn to scale. The mutant names were given based on the modifications that were applied to them. The small letter prefixes ‘m’ and ‘p’ denote minus (deletion) and plus (insertion) modification of the defined spacer region given as the followed capital letter.



mutants were also produced to share the same spacer lengths. The total TP lengths were kept constant in all of the shorter spacer mutants to the same length as that of the wild type. However, the longer spacer mutants have longer TP lengths than that of the wild type.

The last set of mutants was constructed based on the FDF-20-YFP construct (FDtp fused to 20 aa of the mature domain of prSSU followed by YFP). The FDF spacer was divided into 3 regions (Figure 5-6C). The constructs containing different spacer lengths were constructed in similar manner as those of SSF-20-YFP mentioned above (Figure 5-6D). Similar to the SSF spacer mutants, three FDF mutants shared the same spacer lengths. The total TP lengths in the shorter spacer mutants were kept constant as the wild-type length while the longer spacer mutants have longer TP lengths than that of the wild type.

5.3.4 Analysis of spacer length requirement using the designed spacer sequences

In vivo plastid protein import assays were used to determine the effect of the Hsp70-FGLK spacer length. The spacer mutant constructs were transiently expressed in onion (*Allium cepa*) epidermal cells for 12 h before the images of the transformed cells were taken. YFP targeted to the plastids was observed as punctate patterns while non-targeted YFP was observed in cytoplasm and nucleus. We quantified the plastid protein import efficiency in terms of the ratio between plastid YFP and cytosolic YFP signals. A higher ratio indicates higher accumulation of YFP in plastids. The ratios quantified from the images represented in Figure 5-7 are shown in Figure 5-8. These ratios show the import efficiencies of the deletion mutants generated from the three designed spacers. All of the mutants based on the number 92 spacer failed to target YFP to plastids, suggesting that this sequence does not function properly as a spacer. The mutants from the number 228 and number 296 spacers showed decreased import

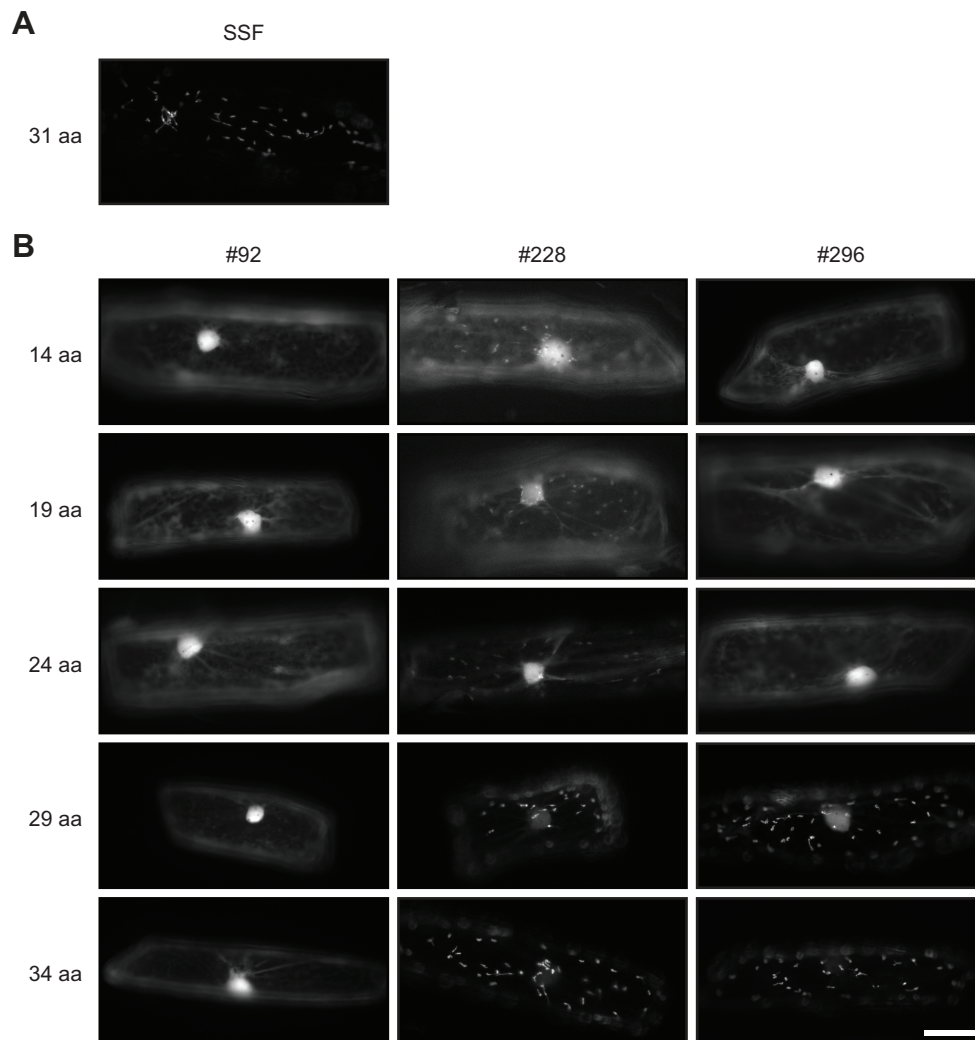


Figure 5-7. Targeting of YFP Directed by the Spacer Mutant TPs Based on the Designed Spacers

Localization patterns of transiently expressed YFP fusion proteins in onion epidermal cells observed under 20x objective. (A) The wild-type SSF-20-YFP construct. (B) The spacer length mutants based on the designed spacers. Top labels indicate the spacers used. Left label indicate lengths of the spacers. Bar, 60 μm .

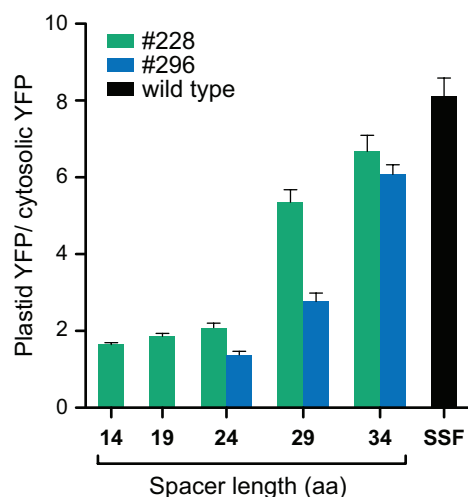


Figure 5-8. Plastid Import Efficiency of the Deletion Constructs of the Designed Spacers

The ratio of plastid YFP/ cytosolic YFP intensities were quantified from the images of onion epidermal cells 12 h after transformations. Representative images are shown in Figure 5-7. Mean $\pm$ SE are shown. $n = 30$. Note that only 4 cells of the 24-aa spacer mutant of spacer number 296 were used in the measurement because it was very difficult to identify YFP targeted plastids. None of the cells expressing YFP from the mutants of spacer number 92, the 14-aa and 19-aa mutants of spacer number 296 showed any observable YFP targeted plastids.

efficiencies when the spacer lengths were shortened. Surprisingly, the 24-aa spacer mutants that have the same spacer lengths as the averaged length found in the TPs performed badly in directing plastid import. These unexpected results may be explained by the change in total TP length. A recent article showed that the total TP length of about 60 aa is required to support plastid protein import (Bionda et al., 2010). Considering these findings together, the mutants with the longer spacer lengths (the extended constructs) and the mutants with the same total TP length (the equal-length constructs) generated from the number 228 and number 296 spacers were included.

When comparing the equal-length mutants with the deletion mutants of the same spacer lengths, the deletion mutants have lower import efficiencies (Figure 5-9) indicating that the decreased total TP lengths affected plastid import. Among the equal-length constructs, the constructs with lengths closer to the wild-type length have the higher ratios. The efficiencies were severely decreased when the spacer lengths of the mutants are at least ± 10 aa difference from the wild-type length. The extended mutants with a longer spacer and total TP lengths than those of the wild type also showed reduced import efficiencies.

When the RPPD prediction was applied to the mutant sequences, some of the mutants were shown to contain additional Hsp70 binding sites. The 44-10 mutants from both the number 228 and number 296 spacers had broader N-terminal Hsp70 binding sites (Figure 5-10A). The targeting efficiencies of these mutants were not different from those of the 34-aa equal-length mutants indicating that the broadenings of N-terminal Hsp70 regions do not change the targeting. Both of the 19-aa equal-length constructs of the number 228 and number 296 spacers contain a strong internal Hsp70 binding site (Figure 5-10B). We did not have enough mutants to determine the effect of the internal Hsp70 sites. However, TPs usually contain multiple Hsp70 sites along the sequences (Ivey et al., 2000) such as SSF shown in Figure 5-10B. Other mutants showed

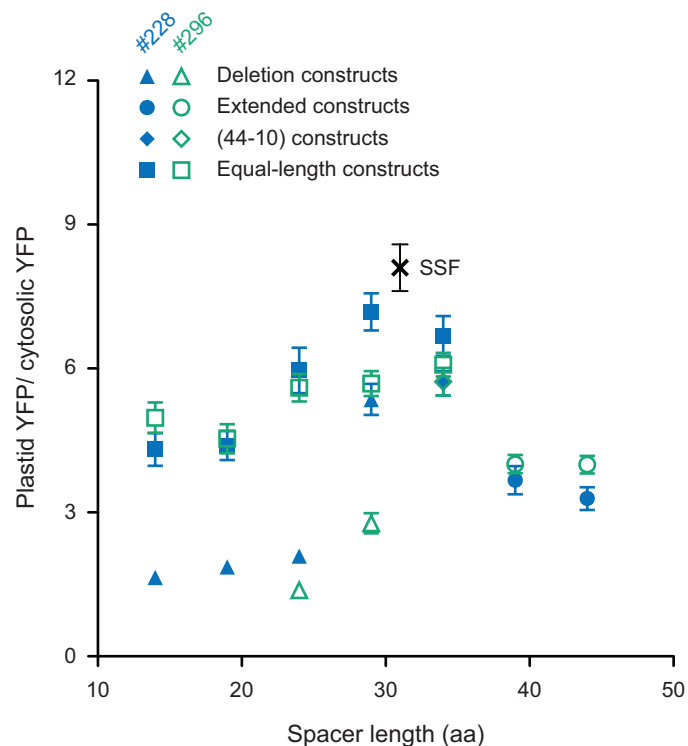
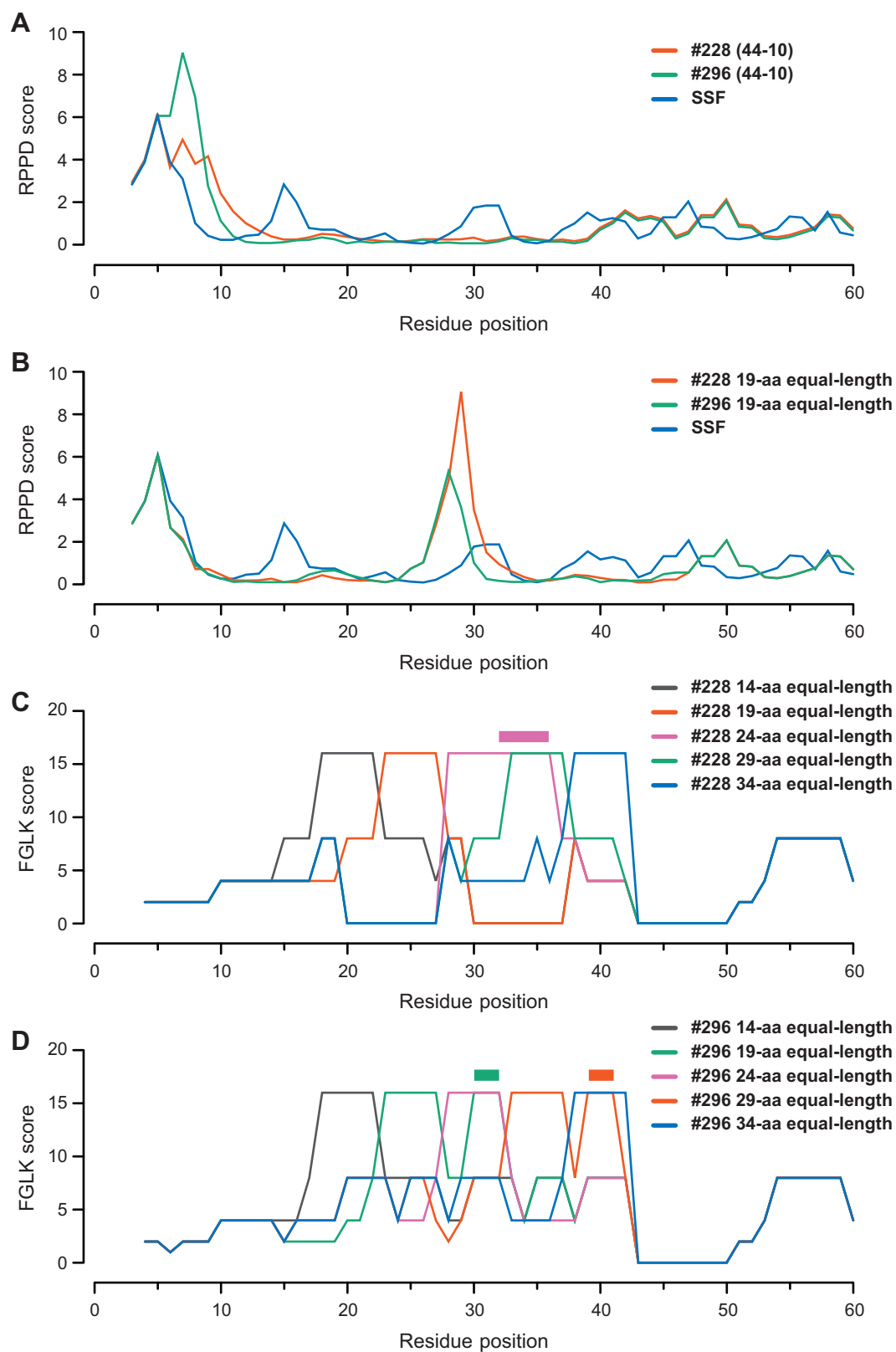


Figure 5-9. Plastid Import Efficiency of the Spacer Mutant TPs Based on the Designed Spacers

The ratio of plastid YFP/ cytosolic YFP intensities were quantified from the images of onion epidermal cells 12 h after transformations. Mean $\pm$ SE are shown. $n = 30$. Note that only 4, 23 and 25 cells of the 24-aa deletion, the 19-aa equal-length and the 14-aa equal-length constructs of spacer number 296 were used in the measurement, respectively.

Figure 5-10. Hsp70 and FGLK Prediction Scores of the Designed Spacer Mutants Containing Additional Predicted Sites

(A) and (B) show the RPPD scores of the mutants with broadened N-terminal Hsp70-interacting sites and the mutants with a strong internal Hsp70-interacting site, respectively. (C) and (D) show the FGLK scores of the mutants containing additional FGLK motifs from the number 228 and number 296 spacers, respectively. The color bars indicate the additional FGLK motifs.



that the N-terminal Hsp70 sites were preserved and no strong internal Hsp70-interacting sites were detected (data not shown).

The FGLK prediction showed that a few mutants contain additional FGLK motifs (Figure 5-10C and D). The 24-aa equal-length mutant of the number 228 spacer had additional FGLK motifs continued from the main motif plateau (Figure 5-10C). Compared to the 24-aa equal-length mutant of the number 296 spacer that do not contain an additional motif, both mutants showed the same level of efficiencies. The larger FGLK plateau in this case did not change the targeting efficiency. Note that the center of the plateau was only shifted by only 2 aa toward the C-terminus. The 19-aa and 29-aa equal-length mutants of the number 296 spacer had a second FGLK plateau following the main plateau (Figure 5-10D). When compared to the corresponding mutants of the number 228 spacer lacking the second plateau, the second plateau did not help to improve the targeting efficiencies. Note that the 19-aa equal-length mutant of the number 296 spacer also contains the internal Hsp70 site at the region between the two FGLK plateaus. Unlike the main FGLK plateau that is 5-aa long, the second plateau is only 3-aa long.

5.3.5 Analysis of spacer length requirement using the wild-type spacer sequences

Although we have shown the effects of spacer lengths in the mutants based on two designed spacers, the following experiments were performed to determine if the same effect could be observed using the wild-type spacers. Two wild-type constructs based on SSF and FDF TPs were used. The mutants were generated such that at least 3 mutants shared the same spacer lengths to ensure that the effects come from the differences in spacer lengths.

Images of onion epidermal cells transiently expressing the constructs were taken 12 h after transformations and used for the analysis (data not shown).

Figures 5-11 and 5-12 show the ratios of plastid YFP and cytosolic YFP signals from the SSF and FDF constructs, respectively. Similar to the designed spacer constructs, the import efficiencies were affected by the lengths of the Hsp70-FGLK spacers. The mutants with the spacer lengths closer to the wild-type lengths had higher import efficiencies. Surprisingly, two of the FDF mutants (mCpC and mBpB) showed much higher import efficiencies than that of the wild type (Figure 5-12).

The FGLK predictions revealed that some of the SSF mutants contain an additional FGLK motif at the -3 aa positions of the main FGLK motifs including mEpB, pDmEpB and pDCmEpB constructs (Figure 5-13A). These constructs seem to have the same efficiencies as those of the constructs with the same spacer lengths (Figure 5-11). Three SSF mutants contain a large patch of FGLK motifs prior to the main FGLK motifs including mAC-ELpAC, mA-ELpA and mC-ELpC constructs (Figure 5-13B). While mA-ELpA and mC-ELpC had the lowest efficiencies among the 26-aa spacer mutants, mAC-ELpAC had the highest efficiency among the 16-aa mutants (Figure 5-11). These inconsistent results indicate that the presence of internal FGLK motifs within the spacer regions did not cause drastic changes in the import efficiencies, compared to that of varying spacer length. No additional FGLK motifs were observed from the other SSF mutants or the FDF mutants (data not shown).

The Hsp70 affinity predictions using RPPD algorithms showed that some of the N-terminal Hsp70-interacting domains in the mutants were altered (Figure 5-14). Three types of alterations were found in the SSF mutants. (i) The stronger Hsp70 affinity mutants (pCmEpA and pCBmEpA) had slightly higher Hsp70 affinities and the Hsp70 peaks were shifted by 1 aa toward C-termini (Figure 5-14A). These mutants had the similar targeting efficiencies as those of mutants with preserved Hsp70 domains (Figure 5-11, pDmEpB and pDCmEpB). (ii) Two SSF mutants (pBmEpC and pBCmEpC) had broadened Hsp70-interacting domains (Figure 5-14B). These mutants had the comparable targeting efficiencies

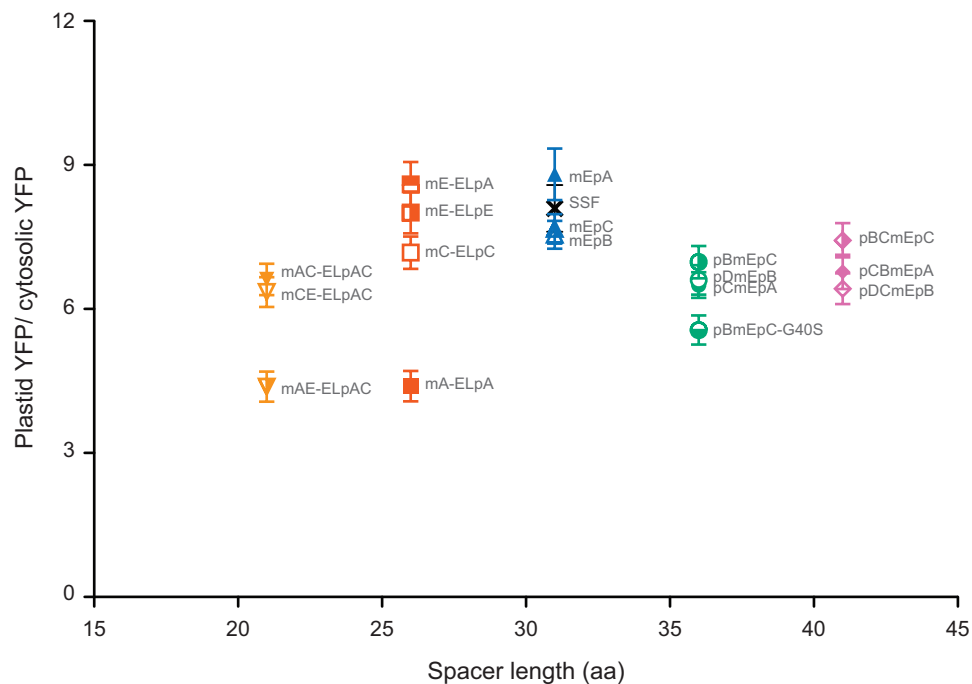


Figure 5-11. Plastid Import Efficiency of the Spacer Mutant TPs Based on the Wild-type SSF Spacer Sequences

The ratio of plastid YFP/ cytosolic YFP intensities were quantified from the images of onion epidermal cells 12 h after transformations. Mean $\pm$ SE are shown. $n = 30$. Note that the pBmEpC-G40S construct contains a single point mutation that changes the Gly40 to Ser.

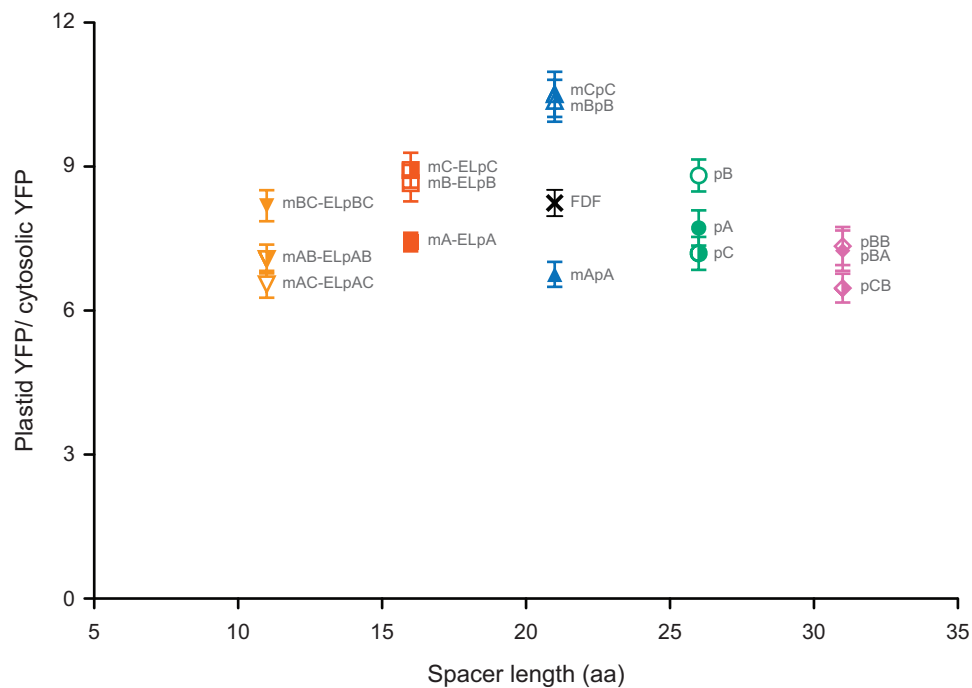


Figure 5-12. Plastid Import Efficiency of the Spacer Mutant TPs Based on the Wild-type FDF Spacer Sequences

The ratio of plastid YFP/ cytosolic YFP intensities were quantified from the images of onion epidermal cells 12 h after transformations. Mean $\pm$ SE are shown. $n = 30$.

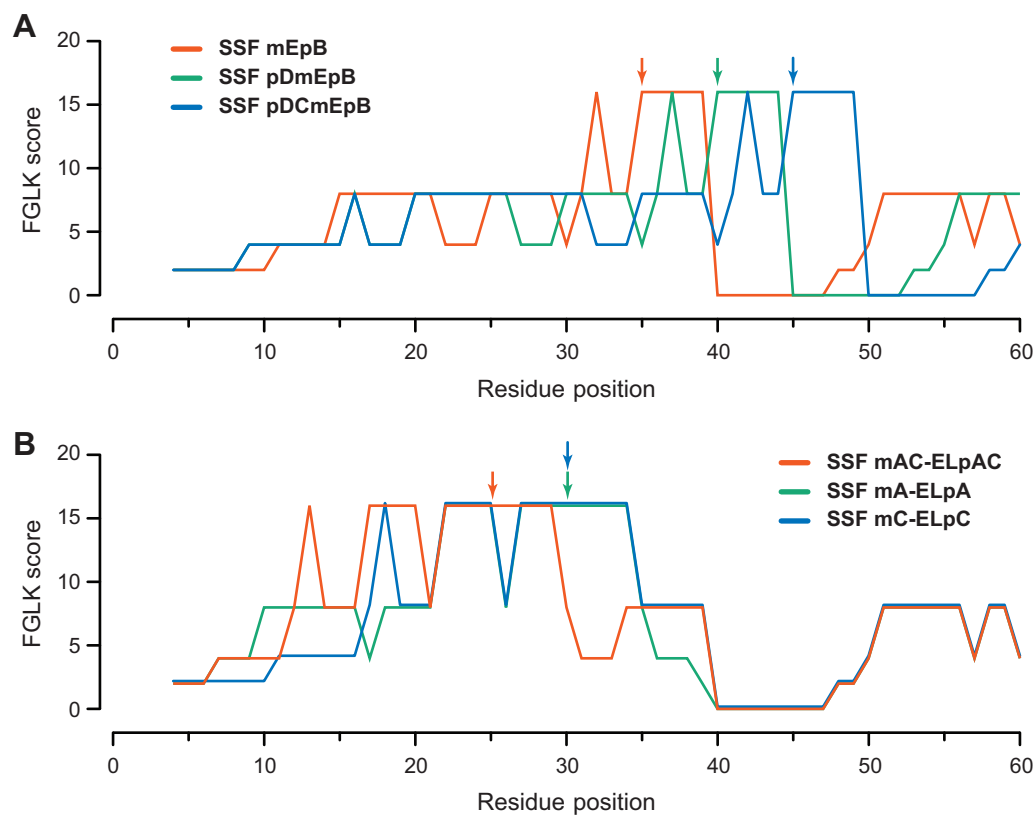


Figure 5-13. FGLK Prediction Scores of the SSF Mutants Containing Additional FGLK Motifs

The additional FGLK motifs located within the spacer sequences were identified in some of the mutants. (A) The mutants contain a single additional motif at -3 aa position in front of the main FGLK motifs. (B) The mutants contain patches of FGLK motifs in front of the main motifs. The start sites of the main FGLK motifs are indicated with arrows.

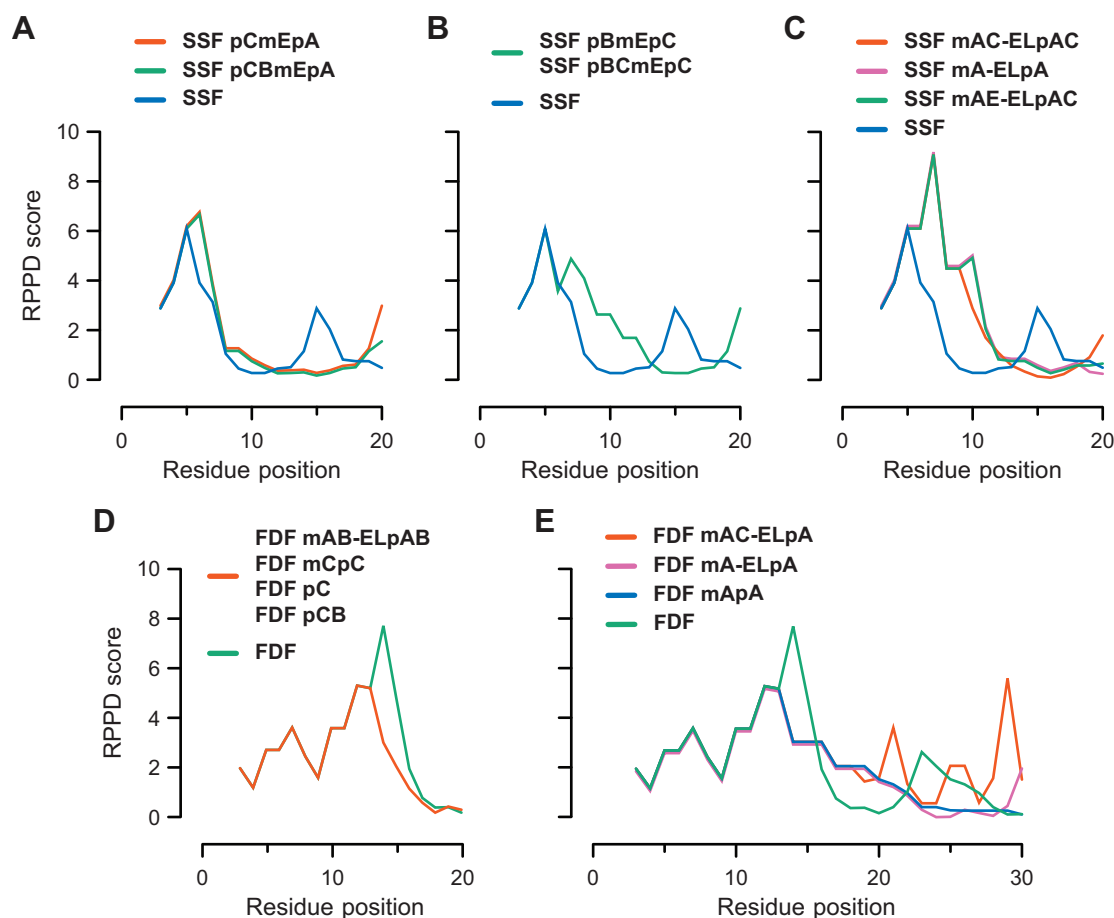


Figure 5-14. RPPD Prediction Scores of the Wild-type Spacer Mutants

The mutants containing the alterations of the N-terminal Hsp70-interacting domains are shown. The SSF mutants with the stronger N-terminal Hsp70 affinities (A), with the broadened N-terminal Hsp70 domains (B) and with the stronger and broadened Hsp70 domains (C) were plotted with the wild-type SSF as the control. The FDF mutants with reduced N-terminal Hsp70 affinities (D) and with the reduced and broadened Hsp70 domains (E) were plotted with the wild-type FDF as the control.

as those of the stronger affinity mutants and the mutants with preserved Hsp70 domains described above (Figure 5-11). (iii) Three of the SSF mutants (mAC-ELpAC, mA-ELpA, and mAE-ELpAC) had a stronger and broadened N-terminal Hsp70-interacting domain (Figure 5-14C). The highest RPPD scores were increased from 6 to 9 and the Hsp70 peaks were shifted by 2 aa toward C-termini. Two of these mutants (mAE-ELpAC and mA-ELpA) showed a severe reduction in targeting efficiencies compared to the other constructs with the same spacer lengths (Figure 5-11) indicating that the broadening of the strong Hsp70-interacting sites at the N-termini may affect targeting. The N-terminal Hsp70-interacting domains in other SSF mutants were not changed (data not shown).

Some of the N-terminal Hsp70-interacting domains of the FDF mutants were affected by the mutation performed (Figure 5-14D and E). (i) Four mutants (mAB-ELpAB, mCpC, pC, and pCB) had a reduced Hsp70 affinity (Figure 5-14D). The highest RPPD scores were dropped from 8 to 5 and the location of the Hsp70 peaks were shifted by 1 aa toward the N-termini. The mAB-ELpAB had the same targeting efficiencies to those of the mutants with preserved N-termini (Figure 5-12, mBC-ELpBC). While the mCpC had the highest efficiency among FDF constructs, pC and pCB had the lowest efficiencies among the mutants with the same spacer lengths (Figure 5-12). These results indicate that the alteration of the N-terminal domain in these mutants did not cause a strong effect in targeting. (ii) Three mutants (mAC-ELpAC, mA-ELpA, and mApA) had reduced and broadened N-terminal Hsp70-interacting domains (Figure 5-14E). These mutants had the lowest efficiencies among the mutants with the same spacer lengths (Figure 5-12) indicating that the broadening of the N-terminal Hsp70-interacting domain negatively affects the targeting activity. The N-terminal domains of other FDF mutants were similar to the wild-type FDF construct (data not shown).

5.4 Discussion

In Chapter 3, we proposed the “bimodal interaction model” describing the interconnection between stromal Hsp70 and surface Toc34 interacting domains as a key spacing requirement for coordinating translocation of the precursor proteins across both plastid membranes. We observed a conserved placement of the FGLK motif in relation to the N-terminal Hsp70-interacting domain. The Hsp70-FGLK spacer distances in *Arabidopsis* and pea SStp, *Silene latifolia* FDtp, and *Arabidopsis* nucleotide transporter 1 range from 22 to 25 aa (Chapter 3). In this chapter, we expanded the spacer analysis to cover a set of 67 TPs containing the strong N-terminal Hsp70 domains.

Sequence analysis found that the averaged length of Hsp70-FGLK spacers among the 67-TP sequences is 24 aa (Figure 5-2). The aa distributions of the spacers indicates that there is no sequence bias among the different regions in the spacer sequences (Figure 5-3). Based on the knowledge that TPs are highly divergent and lack any consensus motif (Bruce, 2000; Lee et al., 2008), and many of the TP recognitions are physicochemical-specific (Chapters 3 and 4), we attempted to design novel Hsp70-FGLK spacers using the aa frequencies determined from the spacer sequences (Figure 5-4). The designed spacer numbers 228 and 229, out of three designed spacers, were able to function as an Hsp70-FGLK spacer in place of the SStp spacer (Figure 5-9). The designed spacer number 92 failed to function as a spacer. This may be due to the incorporation of two Cys residues in its sequence because Cys is one of the rarest aa in the spacer sequences (Figure 5-3). Despite that these designed spacers were generated from a random sequence generator using the spacer aa distribution, the individual sequences of 26-aa long might not always represent the same aa distribution. How well the designed sequence match to the spacer sequences may have to be determined more quantitatively using the method such that developed in Chapter 4 for TP N-terminal sequence analysis.

The effects of Hsp70-FGLK spacer lengths were determined using the *in vivo* plastid protein import assays developed in Chapter 3. The mutants containing varying spacer lengths were generated in SSF and FDF background (Figures 5-5 and 5-6). In addition to the designed spacers, the wild-type spacer sequences of SSF and FDF were also utilized. The spacer length analysis included at least 8 sets of mutants spanning ± 10 aa of the wild-type lengths: 2 sets from the 228 and 296 spacer sequences, 3 sets from the SSF spacer and 3 sets from the FDF spacer. The results indicate that the mutants containing the spacer lengths similar to those of the wild types had the highest targeting efficiencies while the mutants with greater length deviation from the wild-type lengths had lower targeting efficiencies (Figures 5-9, 5-11 and 5-12). Thus, the wild-type spacer lengths are the most optimal lengths supporting the plastid protein targeting.

The shortest spacer mutants (-10 aa) and the longest spacer mutants ($+10$ aa) had the lowest targeting ratios at around 6 while the constructs with the spacer lengths equal to the wild-type lengths had the highest ratios at around 9 (Figures 5-9, 5-11 and 5-12). In term of efficiency, the ratio of 9 (9 plastid YFP signal/ 1 cytosolic YFP signal) corresponds to 90% (9 plastid YFP signal/ 10 total YFP signal) of YFP is in the plastids while ratio of 6 corresponds to 86% targeting efficiency. These reductions in protein import caused by the spacer length effect is much smaller than the reductions observed in the N-terminal mutants (Chapters 3 and 4). The worst N-terminal mutants had the ratios below 2 or 67% targeting efficiency. This lower reduction indicates that while the change in the N-terminal domain greatly affects import, change in the Hsp70-FGLK spacer is much more tolerable. In terms of the bimodal interaction model, this also suggests that the exchange of the TPs between the stromal and surface interactions can occur in a large acceptable timeframe.

In general, there are some Hsp70-interacting sites (Ivey et al., 2000) and FGLK motifs within and around the TP spacers (Figure 5-1). While 2 mutants were found to contain an internal Hsp70 site within their spacers (Figure 5-10B),

there is not enough data to draw any conclusion about the effect of the internal Hsp70 domains. Nevertheless, some of the mutants contain the alterations in their N-terminal Hsp70 domains. The SSF mutants (mAC-ELpAC, mA-ELpA and mAE-ELpAC, Figure 5-14C) and the FDF mutants (mAC-ELpAC, mA-ELpA and mApA, Figure 5-14E) containing a broadened N-terminal Hsp70-interacting domain showed reduced import efficiencies (Figures 5-11 and 5-12). These results suggest that the broadening of the N-terminal Hsp70 domains may negatively affect the function of TPs especially when causing large increases in the accumulative Hsp70 affinities.

Additional FGLK motifs were present in some of the mutants. The 24-aa equal-length mutant of the 228 spacer (Figure 5-10C), the 19-aa and 24-aa equal-length mutants of the 296 spacer (Figure 5-10D), and the mEpB, pDmEpB, pDCmEpB, mAC-ELpAC, mA-ELpAC and mC-ElpC mutants of the SSF spacer (Figure 5-13) contain extra FGLK motifs surrounding the main FGLK motifs. These mutants showed mild alterations, either positive or negative, of the targeting efficiencies from the preserved FGLK mutants (Figures 5-9 and 5-11). These results suggest that the additional FGLK motifs do not have any drastic effects on the targeting efficiencies.

Our results showed a large difference between the optimal spacer lengths of SSF and FDF. While the best spacers in SSF were 31-aa long, the best spacers in FDF were 21-aa long (Figures 5-11 and 5-12). It is without a doubt that the N-terminal Hsp70 domain of SSF is located within the N-terminal 10 aa (Figure 5-4). However, the N-terminal Hsp70 domain in FDF spans from residues 1 to 20 (Figure 5-14E). Our earlier experiments using only the first 10 aa of FDF showed that this 10-aa sequence had a similar function to the N-terminal 10 aa of SSF (Chapter 3). The FDF N-terminal 10 aa when fused to the N-terminus of a import-deficient TP (the FDR construct) can rescue the function of this TP (Chapter 3). Thus, although the N-terminal 10 aa of FDF has a much lower Hsp70 affinity (maximal RPPD score of about 4) than that of SSF (maximal

RPPD score of about 6), it can support plastid protein import. In fact, it was shown that Hsp70 recognize the peptides with RPPD scores above 2 (Ivey et al., 2000). In the context of the translocons, the incoming precursor proteins were translocated from the terminal region (Bruce, 2000; Chotewutmontri et al., 2012; Lung and Chuong, 2012). The first Hsp70 interaction would occur at the sites closest to the termini. If the local-maximal Hsp70 peak at the residue 7 of the N-terminal domain of FDF were considered, the spacer length in FDF would be 28 instead of 21 aa. We also revisited the averaged spacer length calculated from the sequence analysis. Although the average spacer length determined from the Gaussian distribution is 24 aa, the most frequent spacer length bin had the bin center at 30 aa (Figure 5-2). Therefore, we propose that the optimal Hsp70-FGLK spacer should be around 28 to 31-aa long.

As we had discussed in Chapter 3, the spacer was proposed to coordinate surface interaction (FGLK motif) with stromal interaction (N-terminal Hsp70 domain) while residing within the translocons. We also estimated the thickness between the two hydrophobic cores of outer and inner chloroplast membranes to be around 90 Å (Bottier et al., 2007; Bruce, 1998; Marra, 1985; White and von Heijne, 2008) while the extended peptide length was estimated from atomic force microscopy to be 3.0 Å/aa (Chyan et al., 2004; Rief et al., 1997). The proposed optimal Hsp70-FGLK spacer of 28 to 31 aa corresponds to 84 to 93 Å, that well agree with the estimated membrane thickness. Thus, this demonstration that the Hsp70-FGLK spacers showed preferable lengths similar to the membrane thickness supports the proposed role of the spacers in coordinating surface and stromal interactions during the import.

5.5 Conclusions

The sequence analysis showed that the TPs containing the strong N-terminal Hsp70-interacting domains had a conserved spacer length between the

Hsp70 domain and FGLK motifs. Using the aa distribution of the spacer sequences, we had designed three novel spacers. Two of these spacers can substitute for the SSF spacer. Mutants containing varying spacer lengths were generated based on the designed or wild-type spacers in both SSF and FDF background. *In vivo* import assays showed that the mutants containing the spacer lengths similar to those of the wild types had the highest targeting efficiencies while the mutants with greater length deviation from the wild-type lengths had lower targeting efficiencies. We propose that the optimal Hsp70-FGLK spacer should be around 28 to 31-aa long in coordinating the stromal Hsp70 and the surface Toc34 interactions during precursor protein translocation across both plastid membranes.

Chapter 6

Conclusions and Future Directions

6.1 Chloroplast transit peptide domain architecture and function

6.1.1 Prior understanding

Despite the ability to predict transit peptides (TPs) with high accuracy (Emanuelsson et al., 2000), it is still largely unknown what constitutes a TP and how these components facilitate TP function.

Because TPs direct the precursor protein translocation through the translocons at the outer and inner envelope membranes of the chloroplasts (TOC/TIC) into the stroma (Bruce, 2000), it is believed that TP interacts with most of the translocon components. Cross-linking experiments of precursor proteins with intact chloroplasts confirmed these interactions globally (Akita et al., 1997; Chen and Li, 2007; Inoue and Akita, 2008a; Inoue and Akita, 2008b; Kouranov and Schnell, 1997; Ma et al., 1996; Perry and Keegstra, 1994) while *in vitro* binding assays confirmed the direct interactions of TPs with many purified components including the guidance complex (May and Soll, 2000), cytosolic Hsp90 (Qbadou et al., 2006), Toc159 (Smith et al., 2004), Toc34 (Schleiff et al., 2002; Sveshnikova et al., 2000b), Toc75 (Hinnah et al., 2002), Tic110 (Chou et al., 2006; Inaba et al., 2003), stromal Hsp70 (Ivey et al., 2000; Rial et al., 2000), stromal processing peptidase (SPP) (Richter and Lamppa, 1999), and presequence peptidases PreP1/2 (Glaser et al., 2006). However, only some interacting domains were mapped on TP sequences (Figure 1-4) including lipid, Toc159, Toc34, stromal Hsp70 CSS1, and SPP (Becker et al., 2004b; Ivey et al., 2000; Lee et al., 2009a; Pilon et al., 1995; Pinnaduwaage and Bruce, 1996; Schleiff

et al., 2002; Sveshnikova et al., 2000b). These pieces of information were collectively utilized in building the general import pathway model (Bruce, 2000) of plastid protein import shown in Figure 1-3.

While most of the experimental findings support the existence of each step in the model, only a limited number of them shows links between these steps. Only the transfer of TP from cytosolic Hsp90 to Toc64 (Qbadou et al., 2006), the transfer of TP between Toc34 and Toc159 (Becker et al., 2004b), the transfer of TP from guidance complex to Toc34 (May and Soll, 2000; Qbadou et al., 2006), and the TP interactions within the Tic110-Tic40-Hsp93 complex (Chou et al., 2006; Chou et al., 2003; Kovacheva et al., 2005) were shown. The transfers of TP between the other steps have never been captured, especially between the binding and translocating steps.

In addition, mutagenesis of TP sequences in combination with chloroplast import assays have uncovered a large number of critical short sequences containing information that control the import process but most of their functions are still unknown (de Castro Silva Filho et al., 1996; Kindle, 1998; Kindle and Lawrence, 1998; Lee et al., 2008; Lee et al., 2006; Lee et al., 2009a; Lee et al., 2002; Perry et al., 1991; Pilon et al., 1995; Pinnaduwa and Bruce, 1996; Rensink et al., 1998; Rensink et al., 2000; Schnell et al., 1991; Smeekens et al., 1986; von Heijne et al., 1989). It is challenging to elucidate their functions because these regions seem to consist of unique sequences.

Attempts to identify the conserved motifs within TPs using primary sequence alignment were unsuccessful (Karlin-Neumann and Tobin, 1986; Lee et al., 2008; von Heijne et al., 1989), although three regions were loosely defined: (i) an N-terminal domain of about 10 uncharged residues ending with Pro/Gly and preferably having Ala as the second residue, (ii) a central domain, lacking acidic aa but rich in hydroxylated aa, and (iii) a C-terminal domain, rich in Arg and possibly forming an amphiphilic beta-strand (Bruce, 2001; von Heijne et al., 1989). At the secondary structure level, TPs are largely unstructured in solution

forming a “perfect random coil” (von Heijne and Nishikawa, 1991). TPs adopted alpha-helical structures in membrane-mimetic environments as seen in a few TP structures (Krimm et al., 1999; Lancelin et al., 1994; Wienk et al., 1999). Both the primary sequence and secondary structure of TPs were found to be highly divergent.

Bioinformatic approaches had been utilized to identify functional regions in many TP studies. A specialized multiple sequence alignment was developed to identify conserved motifs within a subset of TPs (Lee et al., 2008). Hsp70 prediction programs developed from the *E. coli* DnaK assays were used in the prediction of Hsp70 binding sites in TPs (Ivey et al., 2000; Rial et al., 2000; Zhang and Glaser, 2002). Lastly, McWilliams developed a heuristic algorithm to detect the semi-conserved FGLK motif (Chotewutmontri et al., 2012).

In summary, the experimental studies provide a complex picture of the plastid protein import, which involves multiple TP interactions with the translocon components. However, understanding on where the translocon components interact with TPs (and vice versa) and how TPs are transferred between each step during the import is still limited. While bioinformatics studies of TP are also hindered by the highly divergent nature of TP sequences, the low-throughput nature of the chloroplast import assays could not permit analysis of the large numbers of unique critical sequences that have been identified.

6.1.2 This project achievement

The initial intention of this project was to determine if TP recognition is based on physicochemical properties of TPs because there are conflict findings between sequence-specific and sequence-independent recognitions (Lee et al., 2006; Lee et al., 2002). To maintain the physicochemical properties and diminish the potential sequence motifs, the TP sequences were reversed from C- to N-

termini. Two reverse TPs were produced from the two well-studied TPs: the TPs of small subunit of RuBisCO (SStp) and ferredoxin (FDtp).

In Chapter 3, the reverse TPs were analyzed in comparison with the wild-type TPs. During the course of our study, two assays were developed. The liquid scintillation-based *in vitro* chloroplast binding assay allows direct quantification of the bindings without prior separation via SDS-PAGE. The quantitative *in vivo* chloroplast import assay allows the quantitative estimation of import efficiency from cells transiently expressing the mutant TP DNA constructs. These two assays gradually reduced the amount of work required for the analysis of the TP mutants. The *in vivo* assay was utilized repeatedly in Chapters 4 and 5.

We found that the reverse TPs were able to bind to chloroplasts at the same levels as the wild-type TPs but they failed to direct precursor protein import into chloroplasts. Using these TPs, Toc34 and stromal Hsp70 CSS1 were shown to recognize both wild-type and reverse TPs (Chotewutmontri et al., 2012). These findings indicate that while individual translocon components may utilize physicochemical-specific recognition to reach their interacting domains on TPs, the translocon complexes require a correct organization of these domains for TPs to function. These results also suggest that the discrimination between wild-type and reverse TP translocations occurs in the stroma after the binding step.

Sequence analysis revealed that the reverse TPs lack the N-terminal characteristics of native TPs. TP N-termini are highly uncharged (von Heijne et al., 1989) and contain strong Hsp70-interacting domains (Ivey et al., 2000). When the N-terminal 10 aa sequences of the wild-type TPs were fused to the N-termini of the reverse TP constructs, the modified TPs became import-competent. Vice versa, when the N-terminal 10 aa of the reverse TPs were fused to the wild-type TPs, the modified wild-type TPs became import-deficient. Despite the long-standing proposed role of the TP N-termini as a lipid interacting domain that is involved in the binding step of the chloroplast import (Lee et al., 2008; Lee et al., 2006; Lee et al., 2002; Pilon et al., 1995; Pinnaduwa and Bruce, 1996; Rensink

et al., 1998) or the recently identified role as Toc159 interacting domain (Lee et al., 2009a), we found that the TP N-terminal domain is a key determinant for the translocation step of plastid protein import.

Another property of the TP N-termini is a strong interaction with Hsp70 (Ivey et al., 2000), which suggest that the stromal Hsp70 may be the discriminating factor between the wild-type and reverse TPs. Sequence analysis also identified a conserved placement of FGLK motifs (Toc34 interacting domain) among the import-competent TPs. These TPs contain the spacer between their N-terminal domains and FGLK motifs that range from 22 to 25 aa. These observations together led us to propose the “biomodal interaction model” of TP architecture and function (Figure 6-1).

The biomodal interaction model predicts that an import-efficient TP harbors an N-terminal stromal interacting site and a TOC receptor binding site separated by a spacer with a preferred length that allows the concurrent engagement of a TOC receptor and a stromal motor through the double membrane of about 90 Å thickness (Figure 6-1B).

In Chapter 4, the analysis was focused on the TP N-termini. Using Hsp70 prediction, TPs were clustered into 9 groups. TPs in each group contain the strongest Hsp70-interacting domain at a specific location. We revealed that about a third of TPs have the strongest Hsp70-interacting domains located within their N-terminal 20 aa. The strong N-terminal Hsp70 domain of SStp was replaced with 9 unrelated peptides with varying Hsp70 affinities. We found a positive correlation between the N-terminal Hsp70 affinities and the import efficiencies in 7 peptide mutants. One of these 7 peptides (HA peptide), which does not bind to Hsp70s but has comparable uncharged aa level to those of TP N-termini, was found to be import-deficient. Based on these results, we concluded that a subset of TPs utilizes the N-terminal Hsp70-interacting domains in the translocation process during plastid protein import. Surprisingly, two peptides with the

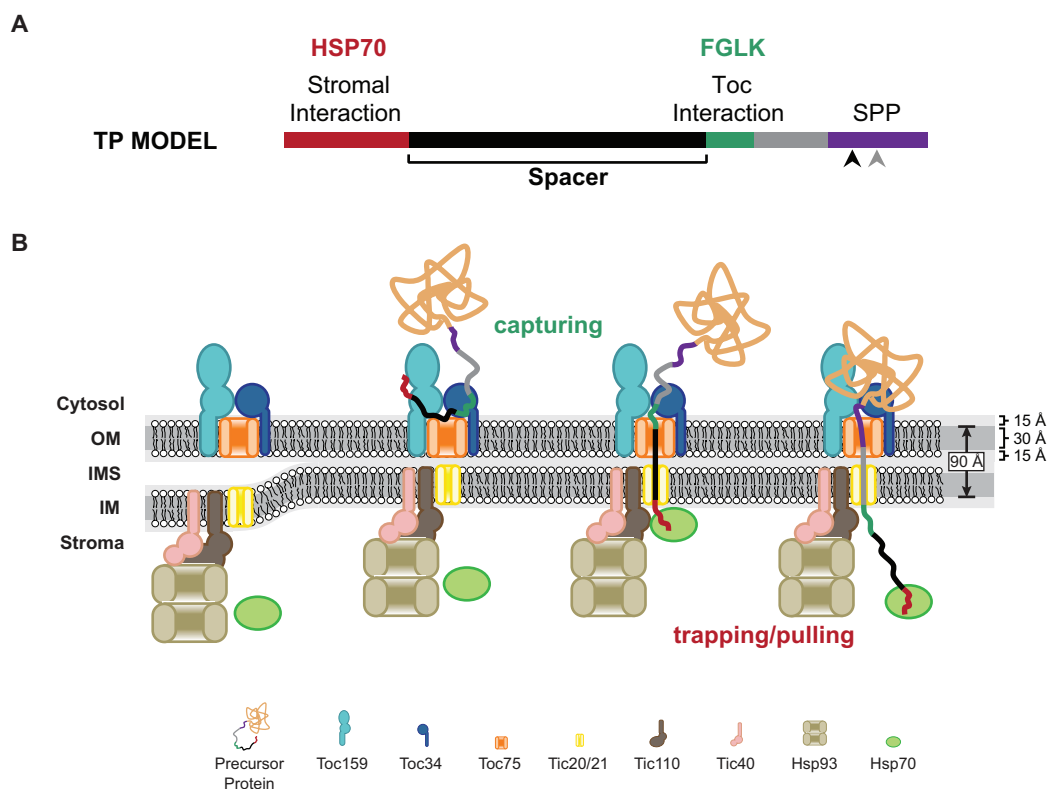


Figure 6-1. Transit Peptide Domain Architecture Model and Function

(A) Our findings showed that a subset of TPs contain an N-terminal Hsp70-interacting domain linked to a FGLK motif via a spacer with an optimal length.

(B) The mechanism of initiating protein translocation into the stroma. We proposed that the TOC interacting domain such as the FGLK motif captures TP and transfers it to the stromal interacting domain such as the N-terminal Hsp70 domain. The translocation proceeds only when the transfer occurs in a rapid timeframe.

strongest Hsp70 affinities failed to direct the import. Sequence analysis indicates that these peptides contain 2-3 residues of Trp which is the rarest aa in the TP N-termini. We proposed that although these peptides can interact with the stromal Hsp70 and initiate translocation, they may not function correctly in the preceding steps such as the binding or intermediate steps before translocation.

In Chapter 5, the spacer length between the strong N-terminal Hsp70 domain and FGLK motifs were studied. Mutants of SStp and FDtp were generated to contain spacers with lengths between ± 10 aa of wild-type lengths based on the designed spacer sequences or the wild-type SStp and FDtp spacer sequences. We found that the mutants containing the spacer lengths similar to those of the wild types had the highest targeting efficiencies while the mutants with greater length deviation from the wild-type lengths had lower targeting efficiencies. We proposed that the optimal Hsp70-FGLK spacer should be around 28 to 31-aa long.

The results from Chapters 4 and 5 provide supporting evidence for our biomodal interaction model. While we showed in Chapter 4 that some TPs utilized the N-terminal Hsp70 domain for targeting, Chapter 5 showed that the optimal import efficiencies only occur at specific Hsp70-FGLK spacer lengths. These results together indicate the link between binding and translocating steps in plastid protein import. While TPs are captured by Toc34 receptor via the FGLK motif, an optimal length spacer allows the stromal Hsp70 to trap/pull the N-terminal Hsp70 domain to initiate the translocation within a rapid timeframe.

6.1.3 Future directions

Although in Chapter 4, we showed that the unrelated Hsp70-interacting peptides could substitute for the TP N-terminal Hsp70 domain and mutants lacking N-terminal Hsp70 domain failed to direct import. We reasoned that the N-termini of these mutants were unable to interact with the stromal Hsp70

chaperone to initiate translocation. But there is no direct evidence to connect this domain with the stromal Hsp70 interaction. Analysis of these mutants in chloroplasts lacking stromal Hsp70 will provide supporting evidence. In addition, majority of TPs do not contain the strong N-terminal Hsp70 domain (Chapter 4). It is possible that these TP N-terminal domains may interact with other stromal proteins. Hsp93 is another translocon motor located in the stroma, however, its substrate recognition is still unknown (Chou et al., 2006; Chou et al., 2003; Kovacheva et al., 2005; Nielsen et al., 1997). Mutants of both stromal Hsp70 and Hsp93 have been identified in *Arabidopsis* (Kovacheva et al., 2007; Su and Li, 2008). Analysis of the precursor proteins containing either Hsp70-interacting or non-interacting (possibly Hsp93-interacting) N-terminal domain in both Hsp70 and Hsp93 mutants may provide evidence to support the role of stromal Hsp93 in initiating the translocation of precursor proteins lacking N-terminal Hsp70 interacting domain. However, it is also known that Hsp70 and Hsp93 function together during import (Shi and Theg, 2011) which potentially complicates this study. But, we previously concluded that the N-terminal domain is a key determinant for import (Chapter 3). The import of TPs utilizing the N-terminal Hsp70 domain is expected to be severely reduced in the Hsp70 mutant but much less reduced in the Hsp93 mutant. The import of TPs utilizing the N-terminal Hsp93 domain would behave oppositely. To reduce the effect from the sequences of different precursors, the experiments could be performed by replacing only the N-terminal domain of the model TP such as SStp with these potential Hsp70/Hsp93-interacting N-termini.

The TP mutants generated from two peptides, pp9 and pp38, were found to be import-deficient (Chapter 4). It was proposed that these mutants could not function in steps before the translocation step in the import. Experiments could be performed to investigate the formation of the binding and import intermediates of the precursors of these mutants using the previously reported methods (Akita and Inoue, 2009; Inoue and Akita, 2008a). Mutagenesis analysis

of these peptide sequences may also uncover the requirement of these steps in the import.

In Chapter 5, the experiments were performed exclusively using onion cells. It was proposed that the optimal spacer length corresponds to the thickness between the compressed chloroplast outer and inner membranes. Although the lipid composition of plastids are maintained over all plastid forms (Joyard et al., 1991) and our experiments using onion, *Arabidopsis*, pea, and tobacco in Chapter 3 were in agreement, the experiments using pea or *Arabidopsis* chloroplasts will provide additional supporting results. The *in vitro* import assays using either pea or *Arabidopsis* chloroplasts would provide an in-depth understanding about the kinetic implication caused by the spacer length.

6.2 Toward designing novel chloroplast transit peptides

Our findings suggest a possible domain architecture of TPs (Figure 6-1A). Many bioinformatics tools are currently available such that we can possibly design novel TPs. The Hsp70 prediction algorithm with high specificity is available (Gragerov et al., 1994; Ivey et al., 2000) while we have developed an Hsp70-FGLK spacer generator based on the observed spacer aa distribution in Chapter 4. The FGLK motif prediction based on the heuristic algorithm had been published (Chotewutmontri et al., 2012). Another important part of TPs is the SPP recognition site. Many articles had reported a SPP motif (Peltier et al., 2000; Richter and Lamppa, 2002; Zybaïlov et al., 2008). TargetP can also be used to detect TP and predict the SPP cleavage (Emanuelsson et al., 2000).

Similar to the designed Hsp70-FGLK spacers generated in Chapter 5, the specific-sequence/motif generators can be developed based on the aa frequency preferences employed in the scoring algorithm or the reported aa distribution of SPP motif. The Hsp70 binding peptides, the FGLK motifs, and the SPP motifs can be generated by the sequence generators. These generated sequences can be

screened using the prediction tools again to identify potential functional sequences. The last step is to combine these domain sequences together to form TPs. Note that our model has an unknown functional region between the FGLK motif and SPP motif (Figure 6-1A). We proposed that this region can be filled with the designed sequences generated from the aa frequencies from the entire TP sequences or more appropriately the frequencies found in the native sequence of this region. Lastly, the designed TPs should be at least 60 aa long to support the import (Bionda et al., 2010).

Bibliography

- Abramoff, M.D., P.J. Magalhaes, and S.J. Ram. 2004. Image processing with ImageJ. *Biophotonics Int.* 11:36-42.
- Agne, B., C. Andres, C. Montandon, B. Christ, A. Ertan, F. Jung, S. Infanger, S. Bischof, S. Baginsky, and F. Kessler. 2010. The acidic A-domain of *Arabidopsis* TOC159 occurs as a hyperphosphorylated protein. *Plant Physiol.* 153:1016-1030.
- Akita, M., and H. Inoue. 2009. Evaluating the energy-dependent "binding" in the early stage of protein import into chloroplasts. *Methods Enzymol.* 466:43-64.
- Akita, M., E. Nielsen, and K. Keegstra. 1997. Identification of protein transport complexes in the chloroplastic envelope membranes via chemical cross-linking. *J Cell Biol.* 136:983-994.
- Aleshin, A.E., S. Gramatikova, G.L. Hura, A. Bobkov, A.Y. Strongin, B. Stec, J.A. Tainer, R.C. Liddington, and J.W. Smith. 2009. Crystal and solution structures of a prokaryotic M16B peptidase: an open and shut case. *Structure.* 17:1465-1475.
- Allen, J.F. 2003. The function of genomes in bioenergetic organelles. *Philos Trans R Soc Lond B Biol Sci.* 358:19-37; discussion 37-18.
- Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389-3402.
- Arnon, D.I. 1949. Copper enzymes in isolated chloroplasts. Polyphenoloxidase in *Beta vulgaris*. *Plant physiology.* 24:1-15.
- Aronsson, H. 2008. The galactolipid monogalactosyldiacylglycerol (MGDG) contributes to photosynthesis-related processes in *Arabidopsis thaliana*. *Plant Signal Behav.* 3:1093-1095.
- Aronsson, H., P. Boij, R. Patel, A. Wardle, M. Topel, and P. Jarvis. 2007. Toc64/OEP64 is not essential for the efficient import of proteins into chloroplasts in *Arabidopsis thaliana*. *Plant J.* 52:53-68.
- Aronsson, H., J. Combe, R. Patel, B. Agne, M. Martin, F. Kessler, and P. Jarvis. 2010. Nucleotide binding and dimerization at the chloroplast pre-protein import receptor, atToc33, are not essential *in vivo* but do increase import efficiency. *Plant J.* 63:297-311.

- Aronsson, H., J. Combe, R. Patel, and P. Jarvis. 2006. *In vivo* assessment of the significance of phosphorylation of the *Arabidopsis* chloroplast protein import receptor, atToc33. *FEBS Lett.* 580:649-655.
- Bairoch, A., and R. Apweiler. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* 28:45-48.
- Baldwin, A., A. Wardle, R. Patel, P. Dudley, S.K. Park, D. Twell, K. Inoue, and P. Jarvis. 2005. A molecular-genetic study of the *Arabidopsis* Toc75 gene family. *Plant Physiol.* 138:715-733.
- Ball, S.G., A. Subtil, D. Bhattacharya, A. Moustafa, A.P. Weber, L. Gehre, C. Colleoni, M.C. Arias, U. Cenci, and D. Dauvillee. 2013. Metabolic effectors secreted by bacterial pathogens: essential facilitators of plastid endosymbiosis? *Plant Cell.* 25:7-21.
- Balsera, M., T.A. Goetze, E. Kovacs-Bogdan, P. Schurmann, R. Wagner, B.B. Buchanan, J. Soll, and B. Bolter. 2009a. Characterization of Tic110, a channel-forming protein at the inner envelope membrane of chloroplasts, unveils a response to Ca(2+) and a stromal regulatory disulfide bridge. *J Biol Chem.* 284:2603-2616.
- Balsera, M., J. Soll, and B. Bolter. 2009b. Protein import machineries in endosymbiotic organelles. *Cell Mol Life Sci.* 66:1903-1923.
- Balsera, M., J. Soll, and B.B. Buchanan. 2010. Redox extends its regulatory reach to chloroplast protein import. *Trends Plant Sci.* 15:515-521.
- Battistutta, R., A. Bisello, S. Mammi, and E. Peggion. 1994. Conformation of retro-bombolitin I in aqueous solution containing surfactant micelles. *Biopolymers.* 34:1535-1541.
- Bauer, J., A. Hiltbrunner, P. Weibel, P.A. Vidi, M. Alvarez-Huerta, M.D. Smith, D.J. Schnell, and F. Kessler. 2002. Essential role of the G-domain in targeting of the protein import receptor atToc159 to the chloroplast outer membrane. *J Cell Biol.* 159:845-854.
- Baum, D. 2013. The origin of primary plastids: a pas de deux or a menage a trois? *Plant Cell.*
- Becker, T., J. Hritz, M. Vogel, A. Caliebe, B. Bukau, J. Soll, and E. Schleiff. 2004a. Toc12, a novel subunit of the intermembrane space preprotein translocon of chloroplasts. *Mol Biol Cell.* 15:5130-5144.

- Becker, T., M. Jelic, A. Vojta, A. Radunz, J. Soll, and E. Schleiff. 2004b. Preprotein recognition by the Toc complex. *EMBO J.* 23:520-530.
- Bedard, J., and P. Jarvis. 2005. Recognition and envelope translocation of chloroplast preproteins. *J Exp Bot.* 56:2287-2320.
- Bedard, J., S. Kubis, S. Bimanadham, and P. Jarvis. 2007. Functional similarity between the chloroplast translocon component, Tic40, and the human co-chaperone, Hsp70-interacting protein (Hip). *J Biol Chem.* 282:21404-21414.
- Benkirane, N., G. Guichard, M.H. Van Regenmortel, J.P. Briand, and S. Muller. 1995. Cross-reactivity of antibodies to retro-inverso peptidomimetics with the parent protein histone H3 and chromatin core particle. Specificity and kinetic rate-constant measurements. *J Biol Chem.* 270:11921-11926.
- Bhushan, S., B. Lefebvre, A. Stahl, S.J. Wright, B.D. Bruce, M. Boutry, and E. Glaser. 2003. Dual targeting and function of a protease in mitochondria and chloroplasts. *EMBO reports.* 4:1073-1078.
- Bhushan, S., A. Stahl, S. Nilsson, B. Lefebvre, M. Seki, C. Roth, D. McWilliam, S.J. Wright, D.A. Liberles, K. Shinozaki, B.D. Bruce, M. Boutry, and E. Glaser. 2005. Catalysis, subcellular localization, expression and evolution of the targeting peptides degrading protease, AtPreP2. *Plant Cell Physiol.* 46:985-996.
- Bionda, T., B. Tillmann, S. Simm, K. Beilstein, M. Ruprecht, and E. Schleiff. 2010. Chloroplast import signals: the length requirement for translocation *in vitro* and *in vivo*. *J Mol Biol.* 402:510-523.
- Blobel, G. 1980. Intracellular protein topogenesis. *Proc Natl Acad Sci U S A.* 77:1496-1500.
- Bogsch, E., S. Brink, and C. Robinson. 1997. Pathway specificity for a delta pH-dependent precursor thylakoid lumen protein is governed by a 'Sec-avoidance' motif in the transfer peptide and a 'Sec-incompatible' mature protein. *EMBO J.* 16:3851-3859.
- Bolter, B., T. May, and J. Soll. 1998. A protein import receptor in pea chloroplasts, Toc86, is only a proteolytic fragment of a larger polypeptide. *FEBS Lett.* 441:59-62.
- Bos, J.L., H. Rehmann, and A. Wittinghofer. 2007. GEFs and GAPs: critical elements in the control of small G proteins. *Cell.* 129:865-877.

- Bottier, C., J. Gean, F. Artzner, B. Desbat, M. Pezolet, A. Renault, D. Marion, and V. Vie. 2007. Galactosyl headgroup interactions control the molecular packing of wheat lipids in Lahgmuir films and in hydrated liquid-crystalline mesophases. *Biochim Biophys Acta*. 1768:1526-1540.
- Bruce, B.D. 1998. The role of lipids in plastid protein transport. *Plant Mol Biol*. 38:223-246.
- Bruce, B.D. 2000. Chloroplast transit peptides: structure, function and evolution. *Trends Cell Biol*. 10:440-447.
- Bruce, B.D. 2001. The paradox of plastid transit peptides: conservation of function despite divergence in primary structure. *Biochim Biophys Acta*. 1541:2-21.
- Bruce, B.D., S. Perry, J. Froelich, and K. Keegstra. 1994. *In vitro* import of proteins into chloroplasts. In *Plant Molecular Biology Manual*. Vol. J1. S.B. Gelvin and R.A. Schilferoot, editors. Kluwer Academic Publishers, Dordrecht. 1-15.
- Bruch, E.M., G.L. Rosano, and E.A. Ceccarelli. 2012. Chloroplastic Hsp100 chaperones ClpC2 and ClpD interact *in vitro* with a transit peptide only when it is located at the N-terminus of a protein. *BMC Plant Biol*. 12:57.
- Butterfield, N.J. 2000. *Bangiomorpha pubescens* n. gen., n. sp.: implications for the evolution of sex, multicellularity, and the Mesoproterozoic/Neoproterozoic radiation of eukaryotes. *Paleobiology*. 26:386-404.
- Caliebe, A., R. Grimm, G. Kaiser, J. Lubeck, J. Soll, and L. Heins. 1997. The chloroplastic protein import machinery contains a Rieske-type iron-sulfur cluster and a mononuclear iron-binding protein. *EMBO J*. 16:7342-7350.
- Carbone, A., A. Zinovyev, and F. Kepes. 2003. Codon adaptation index as a measure of dominating codon bias. *Bioinformatics*. 19:2005-2015.
- Chaddock, A.M., A. Mant, I. Karnauchov, S. Brink, R.G. Herrmann, R.B. Klosgen, and C. Robinson. 1995. A new type of signal peptide: central role of a twin-arginine motif in transfer signals for the delta pH-dependent thylakoidal protein translocase. *EMBO J*. 14:2715-2722.
- Chen, K., X. Chen, and D.J. Schnell. 2000. Initial binding of preproteins involving the Toc159 receptor can be bypassed during protein import into chloroplasts. *Plant Physiol*. 122:813-822.

- Chen, K.Y., and H.M. Li. 2007. Precursor binding to an 880-kDa Toc complex as an early step during active import of protein into chloroplasts. *Plant J.* 49:149-158.
- Chen, L.J., and H.M. Li. 1998. A mutant deficient in the plastid lipid DGD is defective in protein import into chloroplasts. *Plant J.* 16:33-39.
- Chen, X., M.D. Smith, L. Fitzpatrick, and D.J. Schnell. 2002. *In vivo* analysis of the role of atTic20 in protein import into chloroplasts. *Plant Cell.* 14:641-654.
- Chigri, F., F. Hormann, A. Stamp, D.K. Stammers, B. Bolter, J. Soll, and U.C. Vothknecht. 2006. Calcium regulation of chloroplast protein translocation is mediated by calmodulin binding to Tic32. *Proc Natl Acad Sci U S A.* 103:16051-16056.
- Chirico, W.J., M.G. Waters, and G. Blobel. 1988. 70K heat shock related proteins stimulate protein translocation into microsomes. *Nature.* 332:805-810.
- Chotewutmontri, P., L.E. Reddick, D.R. McWilliams, I.M. Campbell, and B.D. Bruce. 2012. Differential transit peptide recognition during preprotein binding and translocation into flowering plant plastids. *Plant Cell.* 24:3040-3059.
- Chou, M.L., C.C. Chu, L.J. Chen, M. Akita, and H.M. Li. 2006. Stimulation of transit-peptide release and ATP hydrolysis by a cochaperone during protein import into chloroplasts. *J Cell Biol.* 175:893-900.
- Chou, M.L., L.M. Fitzpatrick, S.L. Tu, G. Budziszewski, S. Potter-Lewis, M. Akita, J.Z. Levin, K. Keegstra, and H.M. Li. 2003. Tic40, a membrane-anchored co-chaperone homolog in the chloroplast protein translocon. *EMBO J.* 22:2970-2980.
- Chua, N.H., and G.W. Schmidt. 1979. Transport of proteins into mitochondria and chloroplasts. *J Cell Biol.* 81:461-483.
- Chupin, V., R. van 't Hof, and B. de Kruijff. 1994. The transit sequence of a chloroplast precursor protein reorients the lipids in monogalactosyl diglyceride containing bilayers. *FEBS Lett.* 350:104-108.
- Chyan, C.L., F.C. Lin, H. Peng, J.M. Yuan, C.H. Chang, S.H. Lin, and G. Yang. 2004. Reversible mechanical unfolding of single ubiquitin molecules. *Biophys J.* 87:3995-4006.

- Cline, K., and C. Dabney-Smith. 2008. Plastid protein import and sorting: different paths to the same compartments. *Curr Opin Plant Biol.* 11:585-592.
- Cline, K., and R. Henry. 1996. Import and routing of nucleus-encoded chloroplast proteins. *Annu Rev Cell Dev Biol.* 12:1-26.
- Constan, D., J.E. Froehlich, S. Rangarajan, and K. Keegstra. 2004. A stromal Hsp100 protein is required for normal chloroplast development and function in *Arabidopsis*. *Plant Physiol.* 136:3605-3615.
- Criscuolo, A., and S. Gribaldo. 2011. Large-scale phylogenomic analyses indicate a deep origin of primary plastids within cyanobacteria. *Mol Biol Evol.* 28:3019-3032.
- Crooks, G.E., G. Hon, J.M. Chandonia, and S.E. Brenner. 2004. WebLogo: a sequence logo generator. *Genome Res.* 14:1188-1190.
- Dabney-Smith, C., P.W. van Den Wijngaard, Y. Treece, W.J. Vredenberg, and B.D. Bruce. 1999. The C terminus of a chloroplast precursor modulates its interaction with the translocation apparatus and PIRAC. *J Biol Chem.* 274:32351-32359.
- Dam, J., and P. Schuck. 2004. Calculating sedimentation coefficient distributions by direct modeling of sedimentation velocity concentration profiles. *Methods Enzymol.* 384:185-212.
- de Castro Silva Filho, M., F. Chaumont, S. Leterme, and M. Boutry. 1996. Mitochondrial and chloroplast targeting sequences in tandem modify protein import specificity in plant organelles. *Plant Mol Biol.* 30:769-780.
- Deshaies, R.J., B.D. Koch, M. Werner-Washburne, E.A. Craig, and R. Schekman. 1988. A subfamily of stress proteins facilitates translocation of secretory and mitochondrial precursor polypeptides. *Nature.* 332:800-805.
- Deusch, O., G. Landan, M. Roettger, N. Gruenheit, K.V. Kowallik, J.F. Allen, W. Martin, and T. Dagan. 2008. Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor. *Mol Biol Evol.* 25:748-761.
- Dobberstein, B., G. Blobel, and N.H. Chua. 1977. *In vitro* synthesis and processing of a putative precursor for the small subunit of ribulose-1,5-bisphosphate carboxylase of *Chlamydomonas reinhardtii*. *Proc Natl Acad Sci U S A.* 74:1082-1085.

- Douglas, S.E. 1998. Plastid evolution: origins, diversity, trends. *Curr Opin Genet Dev.* 8:655-661.
- Drea, S.C., R.M. Mould, J.M. Hibberd, J.C. Gray, and T.A. Kavanagh. 2001. Tissue-specific and developmental-specific expression of an *Arabidopsis thaliana* gene encoding the lipoamide dehydrogenase component of the plastid pyruvate dehydrogenase complex. *Plant Mol Biol.* 46:705-715.
- Emanuelsson, O., S. Brunak, G. von Heijne, and H. Nielsen. 2007. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc.* 2:953-971.
- Emanuelsson, O., H. Nielsen, S. Brunak, and G. von Heijne. 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol.* 300:1005-1016.
- Emanuelsson, O., H. Nielsen, and G. von Heijne. 1999. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Sci.* 8:978-984.
- Estavillo, G.M., P.A. Crisp, W. Pornsiriwong, M. Wirtz, D. Collinge, C. Carrie, E. Giraud, J. Whelan, P. David, H. Javot, C. Brearley, R. Hell, E. Marin, and B.J. Pogson. 2011. Evidence for a SAL1-PAP chloroplast retrograde pathway that functions in drought and high light signaling in *Arabidopsis*. *Plant Cell.* 23:3992-4012.
- Falcon, L.I., S. Magallon, and A. Castillo. 2010. Dating the cyanobacterial ancestor of the chloroplast. *ISME J.* 4:777-783.
- Fisher, C.L., and G.K. Pei. 1997. Modification of a PCR-based site-directed mutagenesis method. *BioTechniques.* 23:570-571, 574.
- Fourie, A.M., J.F. Sambrook, and M.J. Gething. 1994. Common and divergent peptide binding specificities of hsp70 molecular chaperones. *J Biol Chem.* 269:30470-30478.
- Friedman, A.L., and K. Keegstra. 1989. Chloroplast protein import: quantitative analysis of precursor binding. *Plant Physiol.* 89:993-999.
- Fulgosi, H., and J. Soll. 2002. The chloroplast protein import receptors Toc34 and Toc159 are phosphorylated by distinct protein kinases. *J Biol Chem.* 277:8934-8940.

- Gaspar, R., S. Meyer, K. Gotthardt, M. Sirajuddin, and A. Wittinghofer. 2009. It takes two to tango: regulation of G proteins by dimerization. *Nat Rev Mol Cell Biol.* 10:423-429.
- Gattiker, A., W.V. Bienvenut, A. Bairoch, and E. Gasteiger. 2002. FindPept, a tool to identify unmatched masses in peptide mass fingerprinting protein identification. *Proteomics.* 2:1435-1444.
- Gavel, Y., and G. von Heijne. 1990. A conserved cleavage-site motif in chloroplast transit peptides. *FEBS Lett.* 261:455-458.
- Gesch, R.W., I.H. Kang, M. Gallo-Meagher, J.C.V. Vu, K.J. Boote, L.H. Allen, and G. Bowes. 2003. Rubisco expression in rice leaves is related to genotypic variation of photosynthesis under elevated growth CO₂ and temperature. *Plant Cell Environ.* 26:1941-1950.
- Glaser, E., S. Nilsson, and S. Bhushan. 2006. Two novel mitochondrial and chloroplastic targeting-peptide-degrading peptidasomes in *A. thaliana*, AtPreP1 and AtPreP2. *Biol Chem.* 387:1441-1447.
- Gragerov, A., L. Zeng, X. Zhao, W. Burkholder, and M.E. Gottesman. 1994. Specificity of DnaK-peptide binding. *J Mol Biol.* 235:848-854.
- Gray, J.C., J.A. Sullivan, J.H. Wang, C.A. Jerome, and D. MacLean. 2003. Coordination of plastid and nuclear gene expression. *Philos Trans R Soc Lond B Biol Sci.* 358:135-144.
- Guichard, G., N. Benkirane, G. Zeder-Lutz, M.H. van Regenmortel, J.P. Briand, and S. Muller. 1994. Antigenic mimicry of natural L-peptides with retro-inverso-peptidomimetics. *Proc Natl Acad Sci U S A.* 91:9765-9769.
- Gupta, R.S. 2009. Protein signatures (molecular synapomorphies) that are distinctive characteristics of the major cyanobacterial clades. *Int J Syst Evol Microbiol.* 59:2510-2526.
- Guptasarma, P. 1992. Reversal of peptide backbone direction may result in the mirroring of protein structure. *FEBS Lett.* 310:205-210.
- Gutensohn, M., B. Schulz, P. Nicolay, and U.I. Flugge. 2000. Functional analysis of the two *Arabidopsis* homologues of Toc34, a component of the chloroplast protein import apparatus. *Plant J.* 23:771-783.
- Guy, C.L., and Q.B. Li. 1998. The organization and evolution of the spinach stress 70 molecular chaperone gene family. *Plant Cell.* 10:539-556.

- Haack, T., Y.M. Sanchez, M.J. Gonzalez, and E. Giralt. 1997. Structural comparison in solution of a native and retro peptide derived from the third helix of *Staphylococcus aureus* protein A, domain B: retro peptides, a useful tool for the discrimination of helix stabilization factors dependent on the peptide chain orientation. *J Pept Sci.* 3:299-313.
- Hackenberg, C., R. Kern, J. Huge, L.J. Stal, Y. Tsuji, J. Kopka, Y. Shiraiwa, H. Bauwe, and M. Hagemann. 2011. Cyanobacterial lactate oxidases serve as essential partners in N₂ fixation and evolved into photorespiratory glycolate oxidases in plants. *Plant Cell.* 23:2978-2990.
- Harmer, S.L., J.B. Hogenesch, M. Straume, H.S. Chang, B. Han, T. Zhu, X. Wang, J.A. Kreps, and S.A. Kay. 2000. Orchestrated transcription of key pathways in Arabidopsis by the circadian clock. *Science.* 290:2110-2113.
- Heins, L., A. Mehrle, R. Hemmler, R. Wagner, M. Kuchler, F. Hormann, D. Sveshnikov, and J. Soll. 2002. The preprotein conducting channel at the inner envelope membrane of plastids. *EMBO J.* 21:2616-2625.
- Hénaut, A., and A. Danchin. 1996. Analysis and predictions from *Escherichia coli* sequences, or *E. coli in silico*. In *Escherichia coli and Salmonella: cellular and molecular biology*. Vol. 2. F. Neidhardt and R. Curtiss, editors. ASM Press Washington, D.C. 2047-2066.
- Henikoff, S. 1996. Scores for sequence searches and alignments. *Curr Opin Struct Biol.* 6:353-360.
- Henry, R., M. Carrigan, M. McCaffrey, X. Ma, and K. Cline. 1997. Targeting determinants and proposed evolutionary basis for the Sec and the Delta pH protein transport systems in chloroplast thylakoid membranes. *J Cell Biol.* 136:823-832.
- Henry, R., A. Kapazoglou, M. McCaffery, and K. Cline. 1994. Differences between lumen targeting domains of chloroplast transit peptides determine pathway specificity for thylakoid transport. *J Biol Chem.* 269:10189-10192.
- Hertz, G.Z., and G.D. Stormo. 1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics.* 15:563-577.
- Hiltbrunner, A., J. Bauer, P.A. Vidi, S. Infanger, P. Weibel, M. Hohwy, and F. Kessler. 2001. Targeting of an abundant cytosolic form of the protein import receptor at Toc159 to the outer chloroplast membrane. *J Cell Biol.* 154:309-316.

- Hinnah, S.C., K. Hill, R. Wagner, T. Schlicher, and J. Soll. 1997. Reconstitution of a chloroplast protein import channel. *EMBO J.* 16:7351-7360.
- Hinnah, S.C., R. Wagner, N. Sveshnikova, R. Harrer, and J. Soll. 2002. The chloroplast protein import channel Toc75: pore properties and interaction with transit peptides. *Biophys J.* 83:899-911.
- Hirsch, S., E. Muckel, F. Heemeyer, G. von Heijne, and J. Soll. 1994. A receptor component of the chloroplast protein translocation machinery. *Science.* 266:1989-1992.
- Ho, S.N., H.D. Hunt, R.M. Horton, J.K. Pullen, and L.R. Pease. 1989. Site-directed mutagenesis by overlap extension using the polymerase chain-reaction. *Gene.* 77:51-59.
- Hofmann, N.R., and S.M. Theg. 2005a. Chloroplast outer membrane protein targeting and insertion. *Trends Plant Sci.* 10:450-457.
- Hofmann, N.R., and S.M. Theg. 2005b. Protein- and energy-mediated targeting of chloroplast outer envelope membrane proteins. *Plant J.* 44:917-927.
- Hofmann, N.R., and S.M. Theg. 2005c. Toc64 is not required for import of proteins into chloroplasts in the moss *Physcomitrella patens*. *Plant J.* 43:675-687.
- Holtzer, M.E., E. Braswell, R.H. Angeletti, L. Mints, D. Zhu, and A. Holtzer. 2000. Ultracentrifuge and circular dichroism studies of folding equilibria in a retro GCN4-like leucine zipper. *Biophys J.* 78:2037-2048.
- Inaba, T., M. Alvarez-Huerta, M. Li, J. Bauer, C. Ewers, F. Kessler, and D.J. Schnell. 2005. *Arabidopsis* tic110 is essential for the assembly and function of the protein import machinery of plastids. *Plant Cell.* 17:1482-1496.
- Inaba, T., M. Li, M. Alvarez-Huerta, F. Kessler, and D.J. Schnell. 2003. atTic110 functions as a scaffold for coordinating the stromal events of protein import into chloroplasts. *J Biol Chem.* 278:38617-38627.
- Inoue, H., and M. Akita. 2008a. Three sets of translocation intermediates are formed during the early stage of protein import into chloroplasts. *J Biol Chem.* 283:7491-7502.

- Inoue, H., and M. Akita. 2008b. The transition of early translocation intermediates in chloroplasts is accompanied by the movement of the targeting signal on the precursor protein. *Arch Biochem Biophys.* 477:232-238.
- Inoue, H., C. Rounds, and D.J. Schnell. 2010. The molecular basis for distinct pathways for protein import into *Arabidopsis* chloroplasts. *Plant Cell.* 22:1947-1960.
- Inoue, K., and K. Keegstra. 2003. A polyglycine stretch is necessary for proper targeting of the protein translocation channel precursor to the outer envelope membrane of chloroplasts. *Plant J.* 34:661-669.
- Inoue, K., and D. Potter. 2004. The chloroplastic protein translocation channel Toc75 and its paralog OEP80 represent two distinct protein families and are targeted to the chloroplastic outer envelope by different mechanisms. *Plant J.* 39:354-365.
- Isaacson, T., C.M. Damasceno, R.S. Saravanan, Y. He, C. Catala, M. Saladie, and J.K. Rose. 2006. Sample extraction techniques for enhanced proteomic analysis of plant tissues. *Nat Protoc.* 1:769-774.
- Ivanova, Y., M.D. Smith, K. Chen, and D.J. Schnell. 2004. Members of the Toc159 import receptor family represent distinct pathways for protein targeting to plastids. *Mol Biol Cell.* 15:3379-3392.
- Ivey, R.A., 3rd, and B.D. Bruce. 2000. *In vivo* and *in vitro* interaction of DnaK and a chloroplast transit peptide. *Cell Stress Chaperones.* 5:62-71.
- Ivey, R.A., 3rd, C. Subramanian, and B.D. Bruce. 2000. Identification of a Hsp70 recognition domain within the rubisco small subunit transit peptide. *Plant Physiol.* 122:1289-1299.
- Jackson-Constan, D., M. Akita, and K. Keegstra. 2001. Molecular chaperones involved in chloroplast protein import. *Biochim Biophys Acta.* 1541:102-113.
- Jackson-Constan, D., and K. Keegstra. 2001. *Arabidopsis* genes encoding components of the chloroplastic protein import apparatus. *Plant Physiol.* 125:1567-1576.
- Jarvis, P. 2008. Targeting of nucleus-encoded proteins to chloroplasts in plants. *New Phytol.* 179:257-285.

- Jarvis, P., L.J. Chen, H. Li, C.A. Peto, C. Fankhauser, and J. Chory. 1998. An *Arabidopsis* mutant defective in the plastid general protein import apparatus. *Science*. 282:100-103.
- Jelic, M., J. Soll, and E. Schleiff. 2003. Two Toc34 homologues with different properties. *Biochemistry*. 42:5906-5916.
- Jelic, M., N. Sveshnikova, M. Motzkus, P. Horth, J. Soll, and E. Schleiff. 2002. The chloroplast import receptor Toc34 functions as preprotein-regulated GTPase. *Biol Chem*. 383:1875-1883.
- Joyard, J., M.A. Block, and R. Douce. 1991. Molecular aspects of plastid envelope biochemistry. *Eur J Biochem*. 199:489-509.
- Jungwirth, M., M.L. Dear, P. Brown, K. Holbrook, and R. Goodchild. 2010. Relative tissue expression of homologous torsinB correlates with the neuronal specific importance of DYT1 dystonia-associated torsinA. *Hum Mol Genet*. 19:888-900.
- Kakizaki, T., H. Matsumura, K. Nakayama, F.S. Che, R. Terauchi, and T. Inaba. 2009. Coordination of plastid protein import and nuclear gene expression by plastid-to-nucleus retrograde signaling. *Plant Physiol*. 151:1339-1353.
- Kaneko, T., Y. Nakamura, C.P. Wolk, T. Kuritz, S. Sasamoto, A. Watanabe, M. Iriguchi, A. Ishikawa, K. Kawashima, T. Kimura, Y. Kishida, M. Kohara, M. Matsumoto, A. Matsuno, A. Muraki, N. Nakazaki, S. Shimpo, M. Sugimoto, M. Takazawa, M. Yamada, M. Yasuda, and S. Tabata. 2001. Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp strain PCC 7120. *DNA Res*. 8:205-213.
- Karlin-Neumann, G.A., and E.M. Tobin. 1986. Transit peptides of nuclear-encoded chloroplast proteins share a common amino acid framework. *EMBO J*. 5:9-13.
- Kasmati, A.R., M. Topel, R. Patel, G. Murtaza, and P. Jarvis. 2011. Molecular and genetic analyses of Tic20 homologues in *Arabidopsis thaliana* chloroplasts. *Plant J*. 66:877-889.
- Keegstra, K., and K. Cline. 1999. Protein import and routing systems of chloroplasts. *Plant Cell*. 11:557-570.
- Keegstra, K., and J.E. Froehlich. 1999. Protein import into chloroplasts. *Curr Opin Plant Biol*. 2:471-476.

- Keeling, P.J. 2004. Diversity and evolutionary history of plastids and their hosts. *Am J Bot.* 91:1481-1493.
- Keeling, P.J. 2010. The endosymbiotic origin, diversification and fate of plastids. *Philos Trans R Soc Lond B Biol Sci.* 365:729-748.
- Kern, R., H. Bauwe, and M. Hagemann. 2011. Evolution of enzymes involved in the photorespiratory 2-phosphoglycolate cycle from cyanobacteria via algae toward plants. *Photosynth Res.* 109:103-114.
- Kessler, F., G. Blobel, H.A. Patel, and D.J. Schnell. 1994. Identification of two GTP-binding proteins in the chloroplast protein import machinery. *Science.* 266:1035-1039.
- Kessler, F., and D.J. Schnell. 2002. A GTPase gate for protein import into chloroplasts. *Nat Struct Biol.* 9:81-83.
- Kessler, F., and D.J. Schnell. 2004. Chloroplast protein import: solve the GTPase riddle for entry. *Trends Cell Biol.* 14:334-338.
- Kessler, F., and D.J. Schnell. 2006. The function and diversity of plastid protein import pathways: a multilane GTPase highway into plastids. *Traffic.* 7:248-257.
- Kikuchi, S., J. Bedard, M. Hirano, Y. Hirabayashi, M. Oishi, M. Imai, M. Takase, T. Ide, and M. Nakai. 2013. Uncovering the protein translocon at the chloroplast inner envelope membrane. *Science.* 339:571-574.
- Kikuchi, S., T. Hirohashi, and M. Nakai. 2006. Characterization of the preprotein translocon at the outer envelope membrane of chloroplasts by blue native PAGE. *Plant Cell Physiol.* 47:363-371.
- Kim, S.J., C. Robinson, and A. Mant. 1998. Sec/SRP-independent insertion of two thylakoid membrane proteins bearing cleavable signal peptides. *FEBS Lett.* 424:105-108.
- Kindle, K.L. 1998. Amino-terminal and hydrophobic regions of the *Chlamydomonas reinhardtii* plastocyanin transit peptide are required for efficient protein accumulation *in vivo*. *Plant Mol Biol.* 38:365-377.
- Kindle, K.L., and S.D. Lawrence. 1998. Transit peptide mutations that impair *in vitro* and *in vivo* chloroplast protein import do not affect accumulation of the gamma-subunit of chloroplast ATPase. *Plant Physiol.* 116:1179-1190.

- Kleffmann, T., D. Russenberger, A. von Zychlinski, W. Christopher, K. Sjolander, W. Gruissem, and S. Baginsky. 2004. The *Arabidopsis thaliana* chloroplast proteome reveals pathway abundance and novel protein functions. *Curr Biol.* 14:354-362.
- Klein, R.R., and M.E. Salvucci. 1992. Photoaffinity labeling of mature and precursor forms of the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase after expression in *Escherichia coli*. *Plant Physiol.* 98:546-553.
- Kleine, T., U.G. Maier, and D. Leister. 2009. DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis. *Annu Rev Plant Biol.* 60:115-138.
- Knight, J.S., C.M. Duckett, J.A. Sullivan, A.R. Walker, and J.C. Gray. 2002. Tissue-specific, light-regulated and plastid-regulated expression of the single-copy nuclear gene encoding the chloroplast Rieske FeS protein of *Arabidopsis thaliana*. *Plant Cell Physiol.* 43:522-531.
- Knight, J.S., and J.C. Gray. 1995. The N-terminal hydrophobic region of the mature phosphate translocator is sufficient for targeting to the chloroplast inner envelope membrane. *Plant Cell.* 7:1421-1432.
- Knott, T.G., and C. Robinson. 1994. The secA inhibitor, azide, reversibly blocks the translocation of a subset of proteins across the chloroplast thylakoid membrane. *J Biol Chem.* 269:7843-7846.
- Ko, K., O. Bornemisza, L. Kourtz, Z.W. Ko, W.C. Plaxton, and A.R. Cashmore. 1992. Isolation and characterization of a cDNA clone encoding a cognate 70-kDa heat-shock protein of the chloroplast envelope. *J Biol Chem.* 267:2986-2993.
- Koenig, P., M. Oreb, A. Hofle, S. Kaltofen, K. Rippe, I. Sinning, E. Schleiff, and I. Tews. 2008a. The GTPase cycle of the chloroplast import receptors Toc33/Toc34: implications from monomeric and dimeric structures. *Structure.* 16:585-596.
- Koenig, P., M. Oreb, K. Rippe, C. Muhle-Goll, I. Sinning, E. Schleiff, and I. Tews. 2008b. On the significance of Toc-GTPase homodimers. *J Biol Chem.* 283:23104-23112.
- Kouranov, A., X. Chen, B. Fuks, and D.J. Schnell. 1998. Tic20 and Tic22 are new components of the protein import apparatus at the chloroplast inner envelope membrane. *J Cell Biol.* 143:991-1002.

- Kouranov, A., and D.J. Schnell. 1997. Analysis of the interactions of preproteins with the import machinery over the course of protein import into chloroplasts. *J Cell Biol.* 139:1677-1685.
- Kouranov, A., H. Wang, and D.J. Schnell. 1999. Tic22 is targeted to the intermembrane space of chloroplasts by a novel pathway. *J Biol Chem.* 274:25181-25186.
- Kourtz, L., and K. Ko. 1997. The early stage of chloroplast protein import involves Com70. *J Biol Chem.* 272:2808-2813.
- Kovacheva, S., J. Bedard, R. Patel, P. Dudley, D. Twell, G. Rios, C. Koncz, and P. Jarvis. 2005. *In vivo* studies on the roles of Tic110, Tic40 and Hsp93 during chloroplast protein import. *Plant J.* 41:412-428.
- Kovacheva, S., J. Bedard, A. Wardle, R. Patel, and P. Jarvis. 2007. Further *in vivo* studies on the role of the molecular chaperone, Hsp93, in plastid protein import. *Plant J.* 50:364-379.
- Krimm, I., P. Gans, J.F. Hernandez, G.J. Arlaud, and J.M. Lancelin. 1999. A coil-helix instead of a helix-coil motif can be induced in a chloroplast transit peptide from *Chlamydomonas reinhardtii*. *Eur J Biochem.* 265:171-180.
- Kriwacki, R.W., L. Hengst, L. Tennant, S.I. Reed, and P.E. Wright. 1996. Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity. *Proc Natl Acad Sci U S A.* 93:11504-11509.
- Kubis, S., R. Patel, J. Combe, J. Bedard, S. Kovacheva, K. Lilley, A. Biehl, D. Leister, G. Rios, C. Koncz, and P. Jarvis. 2004. Functional specialization amongst the *Arabidopsis* Toc159 family of chloroplast protein import receptors. *Plant Cell.* 16:2059-2077.
- Kyte, J., and R.F. Doolittle. 1982. A simple method for displaying the hydropathic character of a protein. *J Mol Biol.* 157:105-132.
- Lacroix, E., A.R. Viguera, and L. Serrano. 1998. Reading protein sequences backwards. *Fold Des.* 3:79-85.
- Lancelin, J.M., I. Bally, G.J. Arlaud, M. Blackledge, P. Gans, M. Stein, and J.P. Jacquot. 1994. NMR structures of ferredoxin chloroplastic transit peptide from *Chlamydomonas reinhardtii* promoted by trifluoroethanol in aqueous solution. *FEBS Lett.* 343:261-266.

- Lebowitz, J., M.S. Lewis, and P. Schuck. 2002. Modern analytical ultracentrifugation in protein science: a tutorial review. *Protein Sci.* 11:2067-2079.
- Lee, D.W., J.K. Kim, S. Lee, S. Choi, S. Kim, and I. Hwang. 2008. Arabidopsis nuclear-encoded plastid transit peptides contain multiple sequence subgroups with distinctive chloroplast-targeting sequence motifs. *Plant Cell.* 20:1603-1622.
- Lee, D.W., S. Lee, G.J. Lee, K.H. Lee, S. Kim, G.W. Cheong, and I. Hwang. 2006. Functional characterization of sequence motifs in the transit peptide of *Arabidopsis* small subunit of rubisco. *Plant Physiol.* 140:466-483.
- Lee, D.W., S. Lee, Y.J. Oh, and I. Hwang. 2009a. Multiple sequence motifs in the rubisco small subunit transit peptide independently contribute to Toc159-dependent import of proteins into chloroplasts. *Plant Physiol.* 151:129-141.
- Lee, J., F. Wang, and D.J. Schnell. 2009b. Toc receptor dimerization participates in the initiation of membrane translocation during protein import into chloroplasts. *J Biol Chem.* 284:31130-31141.
- Lee, K.H., D.H. Kim, S.W. Lee, Z.H. Kim, and I. Hwang. 2002. *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.* 14:388-397.
- Lee, K.H., S.J. Kim, Y.J. Lee, J.B. Jin, and I. Hwang. 2003. The M domain of atToc159 plays an essential role in the import of proteins into chloroplasts and chloroplast biogenesis. *J Biol Chem.* 278:36794-36805.
- Lee, S., D.W. Lee, Y. Lee, U. Mayer, Y.D. Stierhof, G. Jürgens, and I. Hwang. 2009c. Heat shock protein cognate 70-4 and an E3 ubiquitin ligase, CHIP, mediate plastid-destined precursor degradation through the ubiquitin-26S proteasome system in *Arabidopsis*. *Plant Cell.* 21:3984-4001.
- Lee, U., I. Rioflorido, S.W. Hong, J. Larkindale, E.R. Waters, and E. Vierling. 2007. The *Arabidopsis* ClpB/Hsp100 family of proteins: chaperones for stress and chloroplast development. *Plant J.* 49:115-127.
- Lee, Y.J., D.H. Kim, Y.W. Kim, and I. Hwang. 2001. Identification of a signal that distinguishes between the chloroplast outer envelope membrane and the endomembrane system *in vivo*. *Plant Cell.* 13:2175-2190.

- Li, H.M., and C.C. Chiu. 2010. Protein transport into chloroplasts. *Annu Rev Plant Biol.* 61:157-180.
- Li, M., and D.J. Schnell. 2006. Reconstitution of protein targeting to the inner envelope membrane of chloroplasts. *J Cell Biol.* 175:249-259.
- Li, Q.B., J.V. Anderson, and C.L. Guy. 1994. A cDNA clone encoding a spinach 70-kilodalton heat-shock cognate. *Plant Physiol.* 105:457-458.
- Li, X., R. Henry, J. Yuan, K. Cline, and N.E. Hoffman. 1995. A chloroplast homologue of the signal recognition particle subunit SRP54 is involved in the posttranslational integration of a protein into thylakoid membranes. *Proc Natl Acad Sci U S A.* 92:3789-3793.
- Ling, Q., W. Huang, A. Baldwin, and P. Jarvis. 2012. Chloroplast biogenesis is regulated by direct action of the ubiquitin-proteasome system. *Science.* 338:655-659.
- Lorkovic, Z.J., W.P. Schroder, H.B. Pakrasi, K.D. Irrgang, R.G. Herrmann, and R. Oelmuller. 1995. Molecular characterization of PsbW, a nuclear-encoded component of the photosystem II reaction center complex in spinach. *Proc Natl Acad Sci U S A.* 92:8930-8934.
- Lung, S.C., and S.D.X. Chuong. 2012. A Transit peptide-like sorting signal at the C terminus directs the *Bienertia sinuspersici* preprotein receptor Toc159 to the chloroplast outer membrane. *Plant Cell.* 24:1560-1578.
- Ma, Y., A. Kouranov, S.E. LaSala, and D.J. Schnell. 1996. Two components of the chloroplast protein import apparatus, IAP86 and IAP75, interact with the transit sequence during the recognition and translocation of precursor proteins at the outer envelope. *J Cell Biol.* 134:315-327.
- Marin, B., E.C. Nowack, and M. Melkonian. 2005. A plastid in the making: evidence for a second primary endosymbiosis. *Protist.* 156:425-432.
- Markowitz, V.M., I.M. Chen, K. Palaniappan, K. Chu, E. Szeto, Y. Grechkin, A. Ratner, B. Jacob, J. Huang, P. Williams, M. Huntemann, I. Anderson, K. Mavromatis, N.N. Ivanova, and N.C. Kyrpides. 2012. IMG: the Integrated Microbial Genomes database and comparative analysis system. *Nucleic Acids Res.* 40:D115-122.
- Marra, J. 1985. Controlled deposition of lipid monolayers and bilayers onto mica and direct force measurements between galactolipid bilayers in aqueous-solutions. *J Colloid Interf Sci.* 107:446-458.

- Marshall, J.S., A.E. DeRocher, K. Keegstra, and E. Vierling. 1990. Identification of heat shock protein hsp70 homologues in chloroplasts. *Proc Natl Acad Sci U S A*. 87:374-378.
- Martin, T., R. Sharma, C. Sippel, K. Waegemann, J. Soll, and U.C. Vothknecht. 2006. A protein kinase family in *Arabidopsis* phosphorylates chloroplast precursor proteins. *J Biol Chem*. 281:40216-40223.
- Martin, W., H. Brinkmann, C. Savonna, and R. Cerff. 1993. Evidence for a chimeric nature of nuclear genomes: eubacterial origin of eukaryotic glyceraldehyde-3-phosphate dehydrogenase genes. *Proc Natl Acad Sci U S A*. 90:8692-8696.
- Martin, W., and K.V. Kowallik. 1999. Annotated English translation of Mereschkowsky's 1905 paper 'Über Natur und Ursprung der Chromatophoren im Pflanzenreiche'. *Eur J Phycol*. 34:287-295.
- Martin, W., T. Rujan, E. Richly, A. Hansen, S. Cornelsen, T. Lins, D. Leister, B. Stoebe, M. Hasegawa, and D. Penny. 2002. 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*. 99:12246-12251.
- Martin, W., B. Stoebe, V. Goremykin, S. Hansmann, M. Hasegawa, and K.V. Kowallik. 1998. Gene transfer to the nucleus and the evolution of chloroplasts. *Nature*. 393:162-165.
- May, T., and J. Soll. 2000. 14-3-3 proteins form a guidance complex with chloroplast precursor proteins in plants. *Plant Cell*. 12:53-64.
- McFadden, G.I., and G.G. van Dooren. 2004. Evolution: red algal genome affirms a common origin of all plastids. *Curr Biol*. 14:R514-516.
- Michl, D., C. Robinson, J.B. Shackleton, R.G. Herrmann, and R.B. Klosgen. 1994. Targeting of proteins to the thylakoids by bipartite presequences: CFoII is imported by a novel, third pathway. *EMBO J*. 13:1310-1317.
- Miras, S., D. Salvi, L. Piette, D. Seigneurin-Berny, D. Grunwald, C. Reinbothe, J. Joyard, S. Reinbothe, and N. Rolland. 2007. Toc159- and Toc75-independent import of a transit sequence-less precursor into the inner envelope of chloroplasts. *J Biol Chem*. 282:29482-29492.

- Moberg, P., A. Stahl, S. Bhushan, S.J. Wright, A. Eriksson, B.D. Bruce, and E. Glaser. 2003. Characterization of a novel zinc metalloprotease involved in degrading targeting peptides in mitochondria and chloroplasts. *Plant J.* 36:616-628.
- Murakami, H., D. Pain, and G. Blobel. 1988. 70-kD heat shock-related protein is one of at least two distinct cytosolic factors stimulating protein import into mitochondria. *J Cell Biol.* 107:2051-2057.
- Nada, A., and J. Soll. 2004. Inner envelope protein 32 is imported into chloroplasts by a novel pathway. *J Cell Sci.* 117:3975-3982.
- Nakai, M., A. Goto, T. Nohara, D. Sugita, and T. Endo. 1994. Identification of the SecA protein homolog in pea chloroplasts and its possible involvement in thylakoidal protein transport. *J Biol Chem.* 269:31338-31341.
- Nakamura, Y., T. Gojobori, and T. Ikemura. 2000. Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res.* 28:292.
- Nakrieko, K.A., R.M. Mould, and A.G. Smith. 2004. Fidelity of targeting to chloroplasts is not affected by removal of the phosphorylation site from the transit peptide. *Eur J Biochem.* 271:509-516.
- Nanjo, Y., H. Oka, N. Ikarashi, K. Kaneko, A. Kitajima, T. Mitsui, F.J. Munoz, M. Rodriguez-Lopez, E. Baroja-Fernandez, and J. Pozueta-Romero. 2006. Rice plastidial N-glycosylated nucleotide pyrophosphatase/phosphodiesterase is transported from the ER-golgi to the chloroplast through the secretory pathway. *Plant Cell.* 18:2582-2592.
- Narayan, V., P. Halada, L. Hernychova, Y.P. Chong, J. Zakova, T.R. Hupp, B. Vojtesek, and K.L. Ball. 2011. A multi-protein binding interface in an intrinsically disordered region of the tumour suppressor protein interferon regulatory factor-1. *J Biol Chem.* 286:14291-303.
- Nelson, B.K., X. Cai, and A. Nebenführ. 2007. A multicolored set of *in vivo* organelle markers for co-localization studies in *Arabidopsis* and other plants. *Plant J.* 51:1126-1136.
- Nielsen, E., M. Akita, J. Davila-Aponte, and K. Keegstra. 1997. Stable association of chloroplastic precursors with protein translocation complexes that contain proteins from both envelope membranes and a stromal Hsp100 molecular chaperone. *EMBO J.* 16:935-946.

- Nowack, E.C., M. Melkonian, and G. Glockner. 2008. Chromatophore genome sequence of *Paulinella* sheds light on acquisition of photosynthesis by eukaryotes. *Curr Biol.* 18:410-418.
- Oblong, J.E., and G.K. Lamppa. 1992. Identification of two structurally related proteins involved in proteolytic processing of precursors targeted to the chloroplast. *EMBO J.* 11:4401-4409.
- Olsen, L.J., and K. Keegstra. 1992. The binding of precursor proteins to chloroplasts requires nucleoside triphosphates in the intermembrane space. *J Biol Chem.* 267:433-439.
- Olszewski, K.A., A. Kolinski, and J. Skolnick. 1996. Does a backwardly read protein sequence have a unique native state? *Protein Eng.* 9:5-14.
- Oreb, M., A. Hofle, P. Koenig, M.S. Sommer, I. Sinning, F. Wang, I. Tews, D.J. Schnell, and E. Schleiff. 2011. Substrate binding disrupts dimerization and induces nucleotide exchange of the chloroplast GTPase Toc33. *Biochem J.* 436:313-319.
- Oreb, M., A. Hofle, O. Mirus, and E. Schleiff. 2008. Phosphorylation regulates the assembly of chloroplast import machinery. *J Exp Bot.* 59:2309-2316.
- Oreb, M., M. Zoryan, A. Vojta, U.G. Maier, L.A. Eichacker, and E. Schleiff. 2007. Phospho-mimicry mutant of atToc33 affects early development of *Arabidopsis thaliana*. *FEBS Lett.* 581:5945-5951.
- Pascual, M.B., A. Mata-Cabana, F.J. Florencio, M. Lindahl, and F.J. Cejudo. 2011. A comparative analysis of the NADPH thioredoxin reductase C-2-Cys peroxiredoxin system from plants and cyanobacteria. *Plant Physiol.* 155:1806-1816.
- Payan, L.A., and K. Cline. 1991. A stromal protein factor maintains the solubility and insertion competence of an imported thylakoid membrane protein. *J Cell Biol.* 112:603-613.
- Pellegrini, A., and R. von Fellenberg. 1999. Design of synthetic bactericidal peptides derived from the bactericidal domain P(18-39) of aprotinin. *Biochim Biophys Acta.* 1433:122-131.
- Peltier, J.B., G. Friso, D.E. Kalume, P. Roepstorff, F. Nilsson, I. Adamska, and K.J. van Wijk. 2000. Proteomics of the chloroplast: systematic identification and targeting analysis of lumenal and peripheral thylakoid proteins. *Plant Cell.* 12:319-341.

- Perry, S.E., W.E. Buvinger, J. Bennett, and K. Keegstra. 1991. Synthetic analogues of a transit peptide inhibit binding or translocation of chloroplastic precursor proteins. *J Biol Chem.* 266:11882-11889.
- Perry, S.E., and K. Keegstra. 1994. Envelope membrane proteins that interact with chloroplastic precursor proteins. *Plant Cell.* 6:93-105.
- Pesaresi, P., S. Masiero, H. Eubel, H.P. Braun, S. Bhushan, E. Glaser, F. Salamini, and D. Leister. 2006. Nuclear photosynthetic gene expression is synergistically modulated by rates of protein synthesis in chloroplasts and mitochondria. *Plant Cell.* 18:970-991.
- Pilon, M., A.D. de Boer, S.L. Knols, M.H. Koppelman, R.M. van der Graaf, B. de Kruijff, and P.J. Weisbeek. 1990. Expression in *Escherichia coli* and purification of a translocation-competent precursor of the chloroplast protein ferredoxin. *J Biol Chem.* 265:3358-3361.
- Pilon, M., B. de Kruijff, and P.J. Weisbeek. 1992a. New insights into the import mechanism of the ferredoxin precursor into chloroplasts. *J Biol Chem.* 267:2548-2556.
- Pilon, M., P.J. Weisbeek, and B. Dekruijff. 1992b. Kinetic-analysis of translocation into isolated chloroplasts of the purified ferredoxin precursor. *FEBS Lett.* 302:65-68.
- Pilon, M., 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. 1995. Functional domains of the ferredoxin transit sequence involved in chloroplast import. *J Biol Chem.* 270:3882-3893.
- Pinnaduwa, P., and B.D. Bruce. 1996. *In vitro* interaction between a chloroplast transit peptide and chloroplast outer envelope lipids is sequence-specific and lipid class-dependent. *J Biol Chem.* 271:32907-32915.
- Plumley, F.G., and G.W. Schmidt. 1989. Nitrogen-dependent regulation of photosynthetic gene expression. *Proc Natl Acad Sci U S A.* 86:2678-2682.

- Price, D.C., C.X. Chan, H.S. Yoon, E.C. Yang, H. Qiu, A.P. Weber, R. Schwacke, J. Gross, N.A. Blouin, C. Lane, A. Reyes-Prieto, D.G. Durnford, J.A. Neilson, B.F. Lang, G. Burger, J.M. Steiner, W. Loffelhardt, J.E. Meuser, M.C. Posewitz, S. Ball, M.C. Arias, B. Henrissat, P.M. Coutinho, S.A. Rensing, A. Symeonidi, H. Doddapaneni, B.R. Green, V.D. Rajah, J. Boore, and D. Bhattacharya. 2012. *Cyanophora paradoxa* genome elucidates origin of photosynthesis in algae and plants. *Science*. 335:843-847.
- Qbadou, S., T. Becker, O. Mirus, I. Tews, J. Soll, and E. Schleiff. 2006. The molecular chaperone Hsp90 delivers precursor proteins to the chloroplast import receptor Toc64. *EMBO J*. 25:1836-1847.
- Rahim, G., S. Bischof, F. Kessler, and B. Agne. 2009. In vivo interaction between atToc33 and atToc159 GTP-binding domains demonstrated in a plant split-ubiquitin system. *J Exp Bot*. 60:257-267.
- Ratnayake, R.M.U., H. Inoue, H. Nonami, and M. Akita. 2008. Alternative processing of *Arabidopsis* Hsp70 precursors during protein import into chloroplasts. *Biosci Biotechnol Biochem*. 72:2926-2935.
- Reddick, L.E. 2010. Dynamics of the Toc GTPases: Modulation by Nucleotides and Transit Peptides Reveal a Mechanism for Chloroplast Protein Import. Doctoral Dissertation. University of Tennessee, Knoxville. 182 pp.
- Reddick, L.E., P. Chotewutmontri, W. Crenshaw, A. Dave, M. Vaughn, and B.D. Bruce. 2008. Nano-scale characterization of the dynamics of the chloroplast Toc translocon. *Methods Cell Biol*. 90:365-398.
- Reddick, L.E., M.D. Vaughn, S.J. Wright, I.M. Campbell, and B.D. Bruce. 2007. *In vitro* comparative kinetic analysis of the chloroplast Toc GTPases. *J Biol Chem*. 282:11410-11426.
- Rensink, W.A., M. Pilon, and P. Weisbeek. 1998. Domains of a transit sequence required for *in vivo* import in *Arabidopsis* chloroplasts. *Plant Physiol*. 118:691-699.
- Rensink, W.A., D.J. Schnell, and P.J. Weisbeek. 2000. The transit sequence of ferredoxin contains different domains for translocation across the outer and inner membrane of the chloroplast envelope. *J Biol Chem*. 275:10265-10271.

- Reumann, S. 2004. Specification of the peroxisome targeting signals type 1 and type 2 of plant peroxisomes by bioinformatics analyses. *Plant Physiol.* 135:783-800.
- Reumann, S., and K. Keegstra. 1999. The endosymbiotic origin of the protein import machinery of chloroplastic envelope membranes. *Trends Plant Sci.* 4:302-307.
- Reyes-Prieto, A., H.S. Yoon, A. Moustafa, E.C. Yang, R.A. Andersen, S.M. Boo, T. Nakayama, K. Ishida, and D. Bhattacharya. 2010. Differential gene retention in plastids of common recent origin. *Mol Biol Evol.* 27:1530-1537.
- Rial, D.V., A.K. Arakaki, and E.A. Ceccarelli. 2000. Interaction of the targeting sequence of chloroplast precursors with Hsp70 molecular chaperones. *Eur J Biochem.* 267:6239-6248.
- Rial, D.V., J. Ottado, and E.A. Ceccarelli. 2003. Precursors with altered affinity for Hsp70 in their transit peptides are efficiently imported into chloroplasts. *J Biol Chem.* 278:46473-46481.
- Rice, P., I. Longden, and A. Bleasby. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* 16:276-277.
- Richardson, L.G., M. Jelokhani-Niaraki, and M.D. Smith. 2009. The acidic domains of the Toc159 chloroplast preprotein receptor family are intrinsically disordered protein domains. *BMC Biochem.* 10:35.
- Richly, E., and D. Leister. 2004. An improved prediction of chloroplast proteins reveals diversities and commonalities in the chloroplast proteomes of *Arabidopsis* and rice. *Gene.* 329:11-16.
- Richter, S., and G.K. Lamppa. 1998. A chloroplast processing enzyme functions as the general stromal processing peptidase. *Proc Natl Acad Sci U S A.* 95:7463-7468.
- Richter, S., and G.K. Lamppa. 1999. Stromal processing peptidase binds transit peptides and initiates their ATP-dependent turnover in chloroplasts. *J Cell Biol.* 147:33-43.
- Richter, S., and G.K. Lamppa. 2002. Determinants for removal and degradation of transit peptides of chloroplast precursor proteins. *J Biol Chem.* 277:43888-43894.

- Richter, S., and G.K. Lamppa. 2003. Structural properties of the chloroplast stromal processing peptidase required for its function in transit peptide removal. *J Biol Chem.* 278:39497-39502.
- Richter, S., R. Zhong, and G. Lamppa. 2005. Function of the stromal processing peptidase in the chloroplast import pathway. *Physiol Plant.* 123:362-368.
- Rief, M., M. Gautel, F. Oesterhelt, J.M. Fernandez, and H.E. Gaub. 1997. Reversible unfolding of individual titin immunoglobulin domains by AFM. *Science.* 276:1109-1112.
- Row, P.E., and J.C. Gray. 2001a. Chloroplast precursor proteins compete to form early import intermediates in isolated pea chloroplasts. *J Exp Bot.* 52:47-56.
- Row, P.E., and J.C. Gray. 2001b. The effect of amino acid-modifying reagents on chloroplast protein import and the formation of early import intermediates. *J Exp Bot.* 52:57-66.
- Rudiger, S., A. Buchberger, and B. Bukau. 1997a. Interaction of Hsp70 chaperones with substrates. *Nat Struct Biol.* 4:342-349.
- Rudiger, S., L. Germeroth, J. Schneider-Mergener, and B. Bukau. 1997b. Substrate specificity of the DnaK chaperone determined by screening cellulose-bound peptide libraries. *EMBO J.* 16:1501-1507.
- Sanchez-Pulido, L., D. Devos, S. Genevrois, M. Vicente, and A. Valencia. 2003. POTRA: a conserved domain in the FtsQ family and a class of beta-barrel outer membrane proteins. *Trends Biochem Sci.* 28:523-526.
- Scheffzek, K., M.R. Ahmadian, W. Kabsch, L. Wiesmuller, A. Lautwein, F. Schmitz, and A. Wittinghofer. 1997. The Ras-RasGAP complex: structural basis for GTPase activation and its loss in oncogenic Ras mutants. *Science.* 277:333-338.
- Schleiff, E., L.A. Eichacker, K. Eckart, T. Becker, O. Mirus, T. Stahl, and J. Soll. 2003a. Prediction of the plant beta-barrel proteome: a case study of the chloroplast outer envelope. *Protein Sci.* 12:748-759.
- Schleiff, E., J. Soll, M. Kuchler, W. Kuhlbrandt, and R. Harrer. 2003b. Characterization of the translocon of the outer envelope of chloroplasts. *J Cell Biol.* 160:541-551.

- Schleiff, E., J. Soll, N. Sveshnikova, R. Tien, S. Wright, C. Dabney-Smith, C. Subramanian, and B.D. Bruce. 2002. Structural and guanosine triphosphate/diphosphate requirements for transit peptide recognition by the cytosolic domain of the chloroplast outer envelope receptor, Toc34. *Biochemistry*. 41:1934-1946.
- Schnell, D.J., and G. Blobel. 1993. Identification of intermediates in the pathway of protein import into chloroplasts and their localization to envelope contact sites. *J Cell Biol*. 120:103-115.
- Schnell, D.J., G. Blobel, K. Keegstra, F. Kessler, K. Ko, and J. Soll. 1997. A consensus nomenclature for the protein-import components of the chloroplast envelope. *Trends Cell Biol*. 7:303-304.
- Schnell, D.J., G. Blobel, and D. Pain. 1991. Signal peptide analogs derived from two chloroplast precursors interact with the signal recognition system of the chloroplast envelope. *J Biol Chem*. 266:3335-3342.
- Schnell, D.J., and D.N. Hebert. 2003. Protein translocons: multifunctional mediators of protein translocation across membranes. *Cell*. 112:491-505.
- Schnell, D.J., F. Kessler, and G. Blobel. 1994. Isolation of components of the chloroplast protein import machinery. *Science*. 266:1007-1012.
- Schuck, P. 2000. Size-distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and lamm equation modeling. *Biophys J*. 78:1606-1619.
- Schuenemann, D., S. Gupta, F. Persello-Cartieaux, V.I. Klimyuk, J.D. Jones, L. Nussaume, and N.E. Hoffman. 1998. A novel signal recognition particle targets light-harvesting proteins to the thylakoid membranes. *Proc Natl Acad Sci U S A*. 95:10312-10316.
- Seedorf, M., and J. Soll. 1995. Copper chloride, an inhibitor of protein import into chloroplasts. *FEBS Lett*. 367:19-22.
- Seedorf, M., K. Waagemann, and J. Soll. 1995. A constituent of the chloroplast import complex represents a new type of GTP-binding protein. *Plant J*. 7:401-411.
- Sharp, P.M., and W.H. Li. 1987. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res*. 15:1281-1295.

- Shen, G., Z. Adam, and H. Zhang. 2007. The E3 ligase AtCHIP ubiquitylates FtsH1, a component of the chloroplast FtsH protease, and affects protein degradation in chloroplasts. *Plant J.* 52:309-321.
- Shi, L.X., and S.M. Theg. 2010. A stromal heat shock protein 70 system functions in protein import into chloroplasts in the moss *Physcomitrella patens*. *Plant Cell.* 22:205-220.
- Shi, L.X., and S.M. Theg. 2011. The motors of protein import into chloroplasts. *Plant Signal Behav.* 6:1397-1401.
- Shoemaker, B.A., J.J. Portman, and P.G. Wolynes. 2000. Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. *Proc Natl Acad Sci U S A.* 97:8868-8873.
- Smeekens, S., C. Bauerle, J. Hageman, K. Keegstra, and P. Weisbeek. 1986. The role of the transit peptide in the routing of precursors toward different chloroplast compartments. *Cell.* 46:365-375.
- Smith, M.D. 2006. Protein import into chloroplasts: an ever-evolving story. *Can J Bot.* 84:531-542.
- Smith, M.D., A. Hiltbrunner, F. Kessler, and D.J. Schnell. 2002. The targeting of the atToc159 preprotein receptor to the chloroplast outer membrane is mediated by its GTPase domain and is regulated by GTP. *J Cell Biol.* 159:833-843.
- Smith, M.D., C.M. Rounds, F. Wang, K. Chen, M. Afithile, and D.J. Schnell. 2004. atToc159 is a selective transit peptide receptor for the import of nucleus-encoded chloroplast proteins. *J Cell Biol.* 165:323-334.
- Soll, J., and E. Schleiff. 2004. Protein import into chloroplasts. *Nat Rev Mol Cell Biol.* 5:198-208.
- Sommer, M.S., B. Daum, L.E. Gross, B.L. Weis, O. Mirus, L. Abram, U.G. Maier, W. Kuhlbrandt, and E. Schleiff. 2011. Chloroplast Omp85 proteins change orientation during evolution. *Proc Natl Acad Sci U S A.* 108:13841-13846.
- Sparkes, I.A., J. Runions, A. Kearns, and C. Hawes. 2006. Rapid, transient expression of fluorescent fusion proteins in tobacco plants and generation of stably transformed plants. *Nat Protoc.* 1:2019-2025.

- Stahl, A., P. Moberg, J. Ytterberg, O. Panfilov, H. Brockenhuus Von Lowenhielm, F. Nilsson, and E. Glaser. 2002. Isolation and identification of a novel mitochondrial metalloprotease (PreP) that degrades targeting presequences in plants. *J Biol Chem.* 277:41931-41939.
- Stahl, T., C. Glockmann, J. Soll, and L. Heins. 1999. Tic40, a new "old" subunit of the chloroplast protein import translocon. *J Biol Chem.* 274:37467-37472.
- Stengel, A., J.P. Benz, B.B. Buchanan, J. Soll, and B. Bolter. 2009. Preprotein import into chloroplasts via the Toc and Tic complexes is regulated by redox signals in *Pisum sativum*. *Mol Plant.* 2:1181-1197.
- Stengel, A., P. Benz, M. Balsera, J. Soll, and B. Bolter. 2008. TIC62 redox-regulated translocon composition and dynamics. *J Biol Chem.* 283:6656-6667.
- Stoebe, B., and K.V. Kowallik. 1999. Gene-cluster analysis in chloroplast genomics. *Trends in Genet.* 15:344-347.
- Su, P.H., and H.M. Li. 2008. *Arabidopsis* stromal 70-kD heat shock proteins are essential for plant development and important for thermotolerance of germinating seeds. *Plant Physiol.* 146:1231-1241.
- Su, P.H., and H.M. Li. 2010. Stromal Hsp70 is important for protein translocation into pea and *Arabidopsis* chloroplasts. *Plant Cell.* 22:1516-1531.
- Subramanian, C. 2001. Structural and Functional Analysis of a Chloroplast Transit Peptide: Interactions with the Chloroplast Translocation Apparatus. Doctoral Dissertation. University of Tennessee, Knoxville. 242 pp.
- Subramanian, C., R. Ivey, 3rd, and B.D. Bruce. 2001. Cytometric analysis of an epitope-tagged transit peptide bound to the chloroplast translocation apparatus. *Plant J.* 25:349-363.
- Sun, C.W., L.J. Chen, L.C. Lin, and H.M. Li. 2001. Leaf-specific upregulation of chloroplast translocon genes by a CCT motif-containing protein, CIA 2. *Plant Cell.* 13:2053-2061.
- Sun, C.W., Y.C. Huang, and H.Y. Chang. 2009. CIA2 coordinately up-regulates protein import and synthesis in leaf chloroplasts. *Plant Physiol.* 150:879-888.

- Sun, Y.J., F. Forouhar, H.M. Li, S.L. Tu, Y.H. Yeh, S. Kao, H.L. Shu, C.C. Chou, C. Chen, and C.D. Hsiao. 2002. Crystal structure of pea Toc34, a novel GTPase of the chloroplast protein translocon. *Nat Struct Biol.* 9:95-100.
- Sung, D.Y., E. Vierling, and C.L. Guy. 2001. Comprehensive expression profile analysis of the *Arabidopsis* Hsp70 gene family. *Plant Physiol.* 126:789-800.
- Sveshnikova, N., R. Grimm, J. Soll, and E. Schleiff. 2000a. Topology studies of the chloroplast protein import channel Toc75. *Biol Chem.* 381:687-693.
- Sveshnikova, N., J. Soll, and E. Schleiff. 2000b. Toc34 is a preprotein receptor regulated by GTP and phosphorylation. *Proc Natl Acad Sci U S A.* 97:4973-4978.
- Teng, Y.S., P.T. Chan, and H.M. Li. 2012. Differential age-dependent import regulation by signal peptides. *PLoS Biol.* 10:e1001416.
- Teng, Y.S., Y.S. Su, L.J. Chen, Y.J. Lee, I. Hwang, and H.M. Li. 2006. Tic21 is an essential translocon component for protein translocation across the chloroplast inner envelope membrane. *Plant Cell.* 18:2247-2257.
- Theg, S.M., C. Bauerle, L.J. Olsen, B.R. Selman, and K. Keegstra. 1989. Internal ATP is the only energy requirement for the translocation of precursor proteins across chloroplastic membranes. *J Biol Chem.* 264:6730-6736.
- Theissen, U., and W. Martin. 2006. The difference between organelles and endosymbionts. *Curr Biol.* 16:R1016-1017; author reply R1017-1018.
- Thompson, S.J., C. Robinson, and A. Mant. 1999. Dual signal peptides mediate the signal recognition particle/Sec-independent insertion of a thylakoid membrane polypeptide, PsbY. *J Biol Chem.* 274:4059-4066.
- Timmis, J.N., M.A. Ayliffe, C.Y. Huang, and W. Martin. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat Rev Genet.* 5:123-135.
- Tomkiewicz, D., N. Nouwen, and A.J. Driessen. 2007. Pushing, pulling and trapping--modes of motor protein supported protein translocation. *FEBS Lett.* 581:2820-2828.
- Tompa, P., C. Szasz, and L. Buday. 2005. Structural disorder throws new light on moonlighting. *Trends Biochem Sci.* 30:484-489.

- Tranel, P.J., and K. Keegstra. 1996. A novel, bipartite transit peptide targets OEP75 to the outer membrane of the chloroplastic envelope. *Plant Cell*. 8:2093-2104.
- Tripp, J., A. Hahn, P. Koenig, N. Flinner, D. Bublak, E.M. Brouwer, F. Ertel, O. Mirus, I. Sinning, I. Tews, and E. Schleiff. 2012. Structure and conservation of the periplasmic targeting factor Tic22 protein from plants and cyanobacteria. *J Biol Chem*. 287:24164-24173.
- Tripp, J., K. Inoue, K. Keegstra, and J.E. Froehlich. 2007. A novel serine/proline-rich domain in combination with a transmembrane domain is required for the insertion of AtTic40 into the inner envelope membrane of chloroplasts. *Plant J*. 52:824-838.
- Tsai, L.Y., S.L. Tu, and H.M. Li. 1999. Insertion of atToc34 into the chloroplastic outer membrane is assisted by at least two proteinaceous components in the import system. *J Biol Chem*. 274:18735-18740.
- Turner, S., K.M. Pryer, V.P. Miao, and J.D. Palmer. 1999. Investigating deep phylogenetic relationships among cyanobacteria and plastids by small subunit rRNA sequence analysis. *J Eukaryot Microbiol*. 46:327-338.
- Uversky, V.N., C.J. Oldfield, and A.K. Dunker. 2008. Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu Rev Biophys*. 37:215-246.
- van 't Hof, R., and B. de Kruijff. 1995a. Characterization of the import process of a transit peptide into chloroplasts. *J Biol Chem*. 270:22368-22373.
- van 't Hof, R., and B. de Kruijff. 1995b. Transit sequence-dependent binding of the chloroplast precursor protein ferredoxin to lipid vesicles and its implications for membrane stability. *FEBS Lett*. 361:35-40.
- van 't Hof, R., R.A. Demel, K. Keegstra, and B. de Kruijff. 1991. Lipid-peptide interactions between fragments of the transit peptide of ribulose-1,5-bisphosphate carboxylase/oxygenase and chloroplast membrane lipids. *FEBS Lett*. 291:350-354.
- van 't Hof, R., W. van Klompenburg, M. Pilon, A. Kozubek, G. de Korte-Kool, R.A. Demel, P.J. Weisbeek, and B. de Kruijff. 1993. The transit sequence mediates the specific interaction of the precursor of ferredoxin with chloroplast envelope membrane lipids. *J Biol Chem*. 268:4037-4042.

- van den Wijngaard, P.W., C. Dabney-Smith, B.D. Bruce, and W.J. Vredenberg. 1999. The mechanism of inactivation of a 50-pS envelope anion channel during chloroplast protein import. *Biophys J.* 77:3156-3162.
- van den Wijngaard, P.W.J., and W.J. Vredenberg. 1999. The envelope anion channel involved in chloroplast protein import is associated with Tic110. *J Biol Chem.* 274:25201-25204.
- VanderVere, P.S., T.M. Bennett, J.E. Oblong, and G.K. Lamppa. 1995. A chloroplast processing enzyme involved in precursor maturation shares a zinc-binding motif with a recently recognized family of metalloendopeptidases. *Proc Natl Acad Sci U S A.* 92:7177-7181.
- Villalobos, A., J.E. Ness, C. Gustafsson, J. Minshull, and S. Govindarajan. 2006. Gene Designer: a synthetic biology tool for constructing artificial DNA segments. *BMC Bioinformatics.* 7:285.
- Villarejo, A., S. Buren, S. Larsson, A. Dejardin, M. Monne, C. Rudhe, J. Karlsson, S. Jansson, P. Lerouge, N. Rolland, G. von Heijne, M. Grebe, L. Bako, and G. Samuelsson. 2005. Evidence for a protein transported through the secretory pathway en route to the higher plant chloroplast. *Nat Cell Biol.* 7:1224-1231.
- Voigt, A., M. Jakob, R.B. Klosgen, and M. Gutensohn. 2005. At least two Toc34 protein import receptors with different specificities are also present in spinach chloroplasts. *FEBS Lett.* 579:1343-1349.
- Vojta, A., M. Alavi, T. Becker, F. Hormann, M. Kuchler, J. Soll, R. Thomson, and E. Schleiff. 2004. The protein translocon of the plastid envelopes. *J Biol Chem.* 279:21401-21405.
- Vojta, L., J. Soll, and B. Bolter. 2007. Protein transport in chloroplasts - targeting to the intermembrane space. *FEBS J.* 274:5043-5054.
- von Heijne, G., T. Hirai, R. Klösigen, J. Steppuhn, B. Bruce, K. Keegstra, and R. Herrmann. 1991. CHLPEP—A database of chloroplast transit peptides. *Plant Mol Biol Rep.* 9:104-126.
- von Heijne, G., and K. Nishikawa. 1991. Chloroplast transit peptides. The perfect random coil? *FEBS Lett.* 278:1-3.
- von Heijne, G., J. Steppuhn, and R.G. Herrmann. 1989. Domain structure of mitochondrial and chloroplast targeting peptides. *Eur J Biochem.* 180:535-545.

- Vorst, O., F. van Dam, R. Oosterhoff-Teertstra, S. Smeekeens, and P. Weisbeek. 1990. Tissue-specific expression directed by an *Arabidopsis thaliana* preferredoxin promoter in transgenic tobacco plants. *Plant Mol Biol.* 14:491-499.
- Waagemann, K., H. Paulsen, and J. Soll. 1990. Translocation of proteins into isolated chloroplasts requires cytosolic factors to obtain import competence. *FEBS Lett.* 261:89-92.
- Waagemann, K., and J. Soll. 1996. Phosphorylation of the transit sequence of chloroplast precursor proteins. *J Biol Chem.* 271:6545-6554.
- Wallas, T.R., M.D. Smith, S. Sanchez-Nieto, and D.J. Schnell. 2003. The roles of *toc34* and *toc75* in targeting the *toc159* preprotein receptor to chloroplasts. *J Biol Chem.* 278:44289-44297.
- Walter, P., and V.R. Lingappa. 1986. Mechanism of protein translocation across the endoplasmic reticulum membrane. *Annu Rev Cell Biol.* 2:499-516.
- Wan, J., S.D. Blakeley, D.T. Dennis, and K. Ko. 1996. Transit peptides play a major role in the preferential import of proteins into leucoplasts and chloroplasts. *J Biol Chem.* 271:31227-31233.
- Weibel, P., A. Hiltbrunner, L. Brand, and F. Kessler. 2003. Dimerization of Toc-GTPases at the chloroplast protein import machinery. *J Biol Chem.* 278:37321-37329.
- Weirich, C.S., J.P. Erzberger, and Y. Barral. 2008. The septin family of GTPases: architecture and dynamics. *Nat Rev Mol Cell Biol.* 9:478-489.
- White, S.H., and G. von Heijne. 2008. How translocons select transmembrane helices. *Ann Rev Biophys.* 37:23-42.
- Wienk, H.L., M. Czisch, and B. de Kruijff. 1999. The structural flexibility of the preferredoxin transit peptide. *FEBS Lett.* 453:318-326.
- Wienk, H.L., R.W. Wechselberger, M. Czisch, and B. de Kruijff. 2000. Structure, dynamics, and insertion of a chloroplast targeting peptide in mixed micelles. *Biochemistry.* 39:8219-8227.
- Wilkins, M.R., E. Gasteiger, A. Bairoch, J.C. Sanchez, K.L. Williams, R.D. Appel, and D.F. Hochstrasser. 1999. Protein identification and analysis tools in the ExPASy server. *Methods Mol Biol.* 112:531-552.

- Yan, X., S. Khan, T. Hase, M.J. Emes, and C.G. Bowsher. 2006. Differential uptake of photosynthetic and non-photosynthetic proteins by pea root plastids. *FEBS Lett.* 580:6509-6512.
- Yeh, Y.H., M.M. Kesavulu, H.M. Li, S.Z. Wu, Y.J. Sun, E.H. Konozy, and C.D. Hsiao. 2007. Dimerization is important for the GTPase activity of chloroplast translocon components atToc33 and psToc159. *J Biol Chem.* 282:13845-13853.
- Yoon, H.S., J.D. Hackett, C. Ciniglia, G. Pinto, and D. Bhattacharya. 2004. A molecular timeline for the origin of photosynthetic eukaryotes. *Mol Biol Evol.* 21:809-818.
- Young, M.E., K. Keegstra, and J.E. Froehlich. 1999. GTP promotes the formation of early-import intermediates but is not required during the translocation step of protein import into chloroplasts. *Plant Physiol.* 121:237-244.
- Yuan, J., R. Henry, M. McCaffery, and K. Cline. 1994. SecA homolog in protein transport within chloroplasts: evidence for endosymbiont-derived sorting. *Science.* 266:796-798.
- Zhang, X.P., and E. Glaser. 2002. Interaction of plant mitochondrial and chloroplast signal peptides with the Hsp70 molecular chaperone. *Trends Plant Sci.* 7:14-21.
- Zhong, R., J. Wan, R. Jin, and G. Lamppa. 2003. A pea antisense gene for the chloroplast stromal processing peptidase yields seedling lethals in *Arabidopsis*: survivors show defective GFP import *in vivo*. *Plant J.* 34:802-812.
- Zhou, Y.H., J.Q. Yu, W.H. Mao, L.F. Huang, X.S. Song, and S. Nogues. 2006. Genotypic variation of Rubisco expression, photosynthetic electron flow and antioxidant metabolism in the chloroplasts of chill-exposed cucumber plants. *Plant Cell Physiol.* 47:192-199.
- Zimmermann, R., M. Sagstetter, M.J. Lewis, and H.R. Pelham. 1988. Seventy-kilodalton heat shock proteins and an additional component from reticulocyte lysate stimulate import of M13 procoat protein into microsomes. *EMBO J.* 7:2875-2880.
- Zybailov, B., H. Rutschow, G. Friso, A. Rudella, O. Emanuelsson, Q. Sun, and K.J. van Wijk. 2008. Sorting signals, N-terminal modifications and abundance of the chloroplast proteome. *PLoS One.* 3:e1994.

Appendices

Appendix 1

DNA Sequences

Table A1-1. General Primers

Primer name	Sequence (5'-3')
T7P	AATACGACTCACTATAGGG
T7ter	GCTAGTTATTGCTCAGCGG
Intein-R	ACCCATGACCTTATTACCAACCTC
d35S-F	CTATCCTTCGCAAGACCCTTCC
YFP-5ter-R	GAACAGCTCCTCGCCCTTGC
M13F	CGCCAGGGTTTCCCAGTCACGAC
M13R	TCACACAGGAAACAGCTATGAC
nos-R	CTTAACGTAATTCAACAGAA

Table A1-2. Primers for the Construction of pET-30a-based Vectors

Generated construct	Primer name	Primer sequence (5'-3')	Template
pET-SSF-YFP	SSF-NdeI-F	GGTGGTCATATGGCTTCTATGATTTCTTCTTCTGC	pBS-SSF-YFP
pET-SSR-YFP	SSR-NdeI-F	GGTGGTCATATGTGTAAGGTACGTGGCGGTAACAGC	pBS-SSR-YFP
pET-FDF-YFP	FDF-NdeI-F	GGTGGTCATATGGCATCTACTCTGTCTACTCTGTCTG	pBS-FDF-YFP
pET-FDR-YFP	FDR-NdeI-F	GGTGGTCATATGGCTACTGTTTCGTGGTCGTTCTGG	pBS-FDR-YFP
pET-ntSSF-YFP	ntSSF-NdeI-F	GGTAGATACATATGGCTTCCTCAGTTC	pAN187
pET-m20-YFP	m20-NdeI-F	GGTGGTCATATGCAGGTGTGGCCACC	pET-FDF-YFP
pET-m20-YFP6xHis	6xHis-F	GTACAAGGGCAGCCATCACCATCACCATCACTAAC	-
	6xHis-R	TCGAGTTAGTGATGGTGATGGTGATGGCTGCCCTT	
pET-NheI-YFP	XbaI-RBS-NcoI-F	CTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATG	-
	F	CTAGCC	
	XbaI-RBS-NcoI-R	CATGGGCTAGCATCTCCTTCTTAAAGTTAAACAAAA TTATTT	

Table A1-3. Primers for the Construction of pAN187-based Vectors

Generated construct	Primer name	Primer sequence (5'-3')	Template
pBS-FDF-YFP	FDF-F-NheI	GCTAGCATGGCATCTACTCTGTCTACTCTG TCTG	pTYB2-FDF
	FDF-R-MscI	TGGCCACACCTGCATAGCAGTAACACGGCC GCGAGAACC	
pBS-FDR-YFP	FDR-F-NheI	GCTAGCATGGCTACTGTTCGTGGTCGTTCT G	pTYB2-FDR
	FDR-R-MscI	TGGCCACACCTGCATGGCAGAGGTCAGGGA AGTC	
pBS-SSF-YFP	SSF-F-NheI	GCTAGCATGGCTTCTATGATTTCTTCTTCT G	pTYB2-SSF
	SSF-R-MscI	TGGCCACACCTGCATGCATTTAACACGACC GCCGTTGCTAG	
pBS-SSR-YFP	SSR-F-NheI	GCTAGCATGTGTAAGGTACGTGGCGGTAAC AGCACTATCTC	pTYB2-SSR
	SSR-R-MscI	TGGCCACACCTGCATTGCGCTCATGATGCT GGAGGAAGC	
pBS-FDR10-FDR-YFP	FDR10-XbaI-R	GGTGGTTCTAGATGCACCAGAACGACCACG	pBS-FDR-YFP
pBS-SSF10-MtoA-SSR-YFP	SSF10-MtoA-SSR-F	GCTTCTATGATTTCTTCTTCTGCCGTTGCG TGTAAGGTACGTGG	pBS-SSF10-SSR-YFP
	SSF10-MtoA-SSR-R	CCACGTACCTTACACGCAACGGCAGAAGAA GAAATCATAGAAGC	
pBS-SSR10-MtoA-SSF-YFP	SSR10-MtoA-SSF-F	CGGTAACAGCACTGCGGCTTCTAGCATTTCT TTCTTCTGCC	pBS-SSR10-SSF-YFP
	SSR10-MtoA-SSF-R	GGCAGAAGAAGAAATGCTAGAAGCCGCAGT GCTGTTACCG	
pBS-FDF10-MtoA-FDR-YFP	FDF10-MtoA-FDR-F	CTCTGTCTACTCTGTCTGTTGCGGCTACTG TTCGTGGTCG	pBS-FDF10-FDR-YFP
	FDF10-MtoA-FDR-R	CGACCAGAACAGTAGCCGCAACAGACAGA GTAGACAGAG	
pBS-FDR10-MtoA-FDF-YFP	FDR10-MtoA-FDF-F	GGTCGTTCTGGTGCAGCGCATCTACTCTG TCTACTCTG	pBS-FDR10-FDF-YFP
	FDR10-MtoA-FDF-R	CAGAGTAGACAGAGTAGATGCCGCTGCACC AGAACGACC	
pBS-pp38-MtoA-SSF-YFP	pp38-MtoA-SSF-F	GGTGGTGCTAGCATGTTCTGGGGTCTCTGG CCTTGGGCGGCTTCTAGC	pBS-SSR10-MtoA-SSF-YFP
pBS-pp9-MtoA-SSF-YFP	pp9-MtoA-SSF-F	GGTGGTGCTAGCATGTGGATCTTCCCTTGG ATTCAACTGCGGCTTCTAGC	pBS-SSR10-MtoA-SSF-YFP
pBS-PepG-MtoA-SSF-YFP	PepG-MtoA-SSF-F	GGTGGTGCTAGCATGGGTGGGTATGGTTTC CGTCATCAGAACTGCGCGCTTCTAGC	pBS-SSR10-MtoA-SSF-YFP
pBS-V10-MtoA-SSF-YFP	V10-MtoA-SSF-F	GGTGGTGCTAGCATGTTCTACCAACTTGCT AAGACCTGTCCAGTTGCGGCTTCTAGC	pBS-SSR10-MtoA-SSF-YFP
pBS-DRC8-MtoA-SSF-YFP	DRC8-MtoA-SSF-F1	GGACCAAGAGGTCATTTCTACGATGCAGCG GCTTCTAGC	pBS-SSR10-MtoA-SSF-YFP
	DRC8-MtoA-SSF-F2	GGTGGTGCTAGCATGTACCTTGTGGACCA AGAGGTCATTTCTAGC	
pBS-A6R-MtoA-SSF-YFP	A6R-MtoA-SSF-F	GGTGGTGCTAGCATGGCCAGCCATCTGGGT CTGGCCCGTGCGGCTTCTAGC	pBS-SSR10-MtoA-SSF-YFP
pBS-HbS-MtoA-SSF-YFP	HbS-MtoA-SSF-F	GGTGGTGCTAGCATGGTGCATCTGACCCCG GTGGAAGACGCGCTTCTAGC	pBS-SSR10-MtoA-SSF-YFP
pBS-np09-MtoA-SSF-YFP	np09-MtoA-SSF-F	GGTGGTGCTAGCATGCGTGTGATCCGGTT GTTGCTTTCGCGGCTTCTAGC	pBS-SSR10-MtoA-SSF-YFP
pBS-HA-MtoA-SSF-YFP	HA-MtoA-SSF-F	GGTGGTGCTAGCATGTACCCGTACGATGTT CCGGACTACGACGCGCTTCTAGC	pBS-SSR10-MtoA-SSF-YFP
pBS-pp38-MtoA-SSR-YFP	pp38-MtoA-SSR-F	GGTGGTGCTAGCATGTTCTGGGGTCTCTGG CCTTGGGCGTGTAAGGTACG	pBS-SSF10-MtoA-SSR-YFP
pBS-pp9-MtoA-SSR-YFP	pp9-MtoA-SSR-F	GGTGGTGCTAGCATGTGGATCTTCCCTTGG ATTCAACTGCGGTGTAAGGTACG	pBS-SSF10-MtoA-SSR-YFP
pBS-PepG-MtoA-SSR-YFP	PepG-MtoA-SSR-F	GGTGGTGCTAGCATGGGTGGTATGGTTTC CGTCATCAGAACTGCGCGTGTAAGGTACG	pBS-SSF10-MtoA-SSR-YFP

Table A1-3. (continued)

Generated construct	Primer name	Primer sequence (5'-3')	Template
pBS-V10-MtoA-SSR-YFP	V10-MtoA-SSR-F	GGTGGTGCTAGCATGTTCTACCAACTTGCT AAGACCTGTCCAGTTGCGTGTAAGGTACG	pBS-SSF10-MtoA-SSR-YFP
pBS-DRC8-MtoA-SSR-YFP	DRC8-MtoA-SSR-F1	GGACCAAGAGGTCATTTCTACGATGCAGC GTGTAAGGTACGTGG	pBS-SSF10-MtoA-SSR-YFP
	DRC8-MtoA-SSR-F2	GGTGGTGCTAGCATGTACCTTGTGGACCA AGAGGTCATTTCTACG	
pBS-A6R-MtoA-SSR-YFP	A6R-MtoA-SSR-F	GGTGGTGCTAGCATGGCCAGCCATCTGGGT CTGGCCCGTGCGTGTAAGGTACGTGG	pBS-SSF10-MtoA-SSR-YFP
pBS-HbS-MtoA-SSR-YFP	HbS-MtoA-SSR-F	GGTGGTGCTAGCATGGTGCATCTGACCCCG GTGAAAAAGCGTGTAAGGTACGTGG	pBS-SSF10-MtoA-SSR-YFP
pBS-np09-MtoA-SSR-YFP	np09-MtoA-SSR-F	GGTGGTGCTAGCATGAGAGTTGATCCAGTT GTGGCTTTTCGCGTGTAAGGTACGTGG	pBS-SSF10-MtoA-SSR-YFP
pBS-HA-MtoA-SSR-YFP	HA-MtoA-SSR-F	GGTGGTGCTAGCATGTACCCATACGATGTT CCTGACTACGCAGCGTGTAAGGTACGTGG	pBS-SSF10-MtoA-SSR-YFP
pBS-FlipN10SSF-YFP	Flip-SSF-F	GGTGGTGCTAGCATGGTTGCTTCTCCTCCAGC ATCATGAGCGCAACCACTGTTTCCCGTGC	pBS-SSF-YFP
pBS-FlipN10FDF-YFP	Flip-FDF-F	GGTGGTGCTAGCATGGTTTCTCTGACTTCC CTGACCTCTGCCTCTGCTTCTCTGCTGC	pBS-FDF-YFP
pBS-ScrambleN10SSF-YFP	Scramble-SSF-F	GGTGGTGCTAGCATGGTTTCTTCAGCAATC AGCGCATCCACCACTGTTTCCCG	pBS-SSF-YFP
pBS-ScrambleN10FDF-YFP	Scramble-FDF-F	GGTGGTGCTAGCATGTCTTCGCTGACCGTG ACCCTGGCATCTTCTGCTTCTCTGCTGC	pBS-FDF-YFP

Table A1-4. Oligonucleotides for Cloning of the First 10 Amino Acids of TPs

Generated construct	Oligonucleotide name	Primer sequence (5'-3')
pBS-SSF10-SSR-YFP	SSF10-F	CTAGCATGGCTTCTATGATTTCCTTCTGCCGTTG
	SSF10-R	CTAGCAACGGCAGAAGAAGAAATCATAGAAGCCATG
pBS-SSR10-SSF-YFP	SSR10-F	CTAGCATGTGTAAGGTACGTGGCGGTAACAGCACTG
	SSR10-R	CTAGCAGTGCTGTTACCGCCACGTACCTTACACATG
pBS-FDF10-FDR-YFP	FDF10-F	CTAGCATGGCATCTACTCTGTCTACTCTGTCTGTG
	FDF10-R	CTAGCAACAGACAGAGTAGACAGAGTAGATGCCATG

Table A1-5. Oligonucleotides for Cloning of the 14-aa Designed Spacer Mutants

Generated construct	Oligonucleotide name	Primer sequence (5'-3')
pBS-SSF-YFP-no92-14aa	Oligo1	CTAGCATGGCTTCTATGATTTCTTCTTCAGGT
	Oligo2	GGGAAACCGGTCATGCTTCCGGTTCGCTCACCT
	Oligo3	GAGAAGAAGAAATCATAGAAGCCATG GAGCGAACCGGAAAGCATGACCGGTTCCCTGTT AAAAAGGTAAACACCGACATCACCAG
	Oligo4	CATTTAACACGACCGCCGTGCTAGTAGTGCTGG TGATGTCGGTGTTTACCTTTTAAACA
	Oligo5	CATCACTAGCAACGGCGGTCGTGTTAAATGCATG
pBS-SSF-YFP-no228-14aa	Oligo1	CTAGCATGGCTTCTATGATTTCTTCTTACCGT
	Oligo2	GGGAAACCGGTCATGCTACCGGCGCTGGTCACGG
	Oligo3	TAGAAGAAGAAATCATAGAAGCCATG GACCAGCGCCGGTAGCATGACCGGTTCCCTGTT AAAAAGGTAAACACCGACATCACCAG
	Oligo4	CATTTAACACGACCGCCGTGCTAGTAGTGCTGG TGATGTCGGTGTTTACCTTTTAAACA
	Oligo5	CATCACTAGCAACGGCGGTCGTGTTAAATGCATG
pBS-SSF-YFP-no296-14aa	Oligo1	CTAGCATGGCTTCTATGATTTCTTCTTACCAAC
	Oligo2	GGGAAACCGGTCATGCTACCGTTGTTACGTTGGT
	Oligo3	AGAAGAAGAAATCATAGAAGCCATG GTGAACAACGGTAGCATGACCGGTTCCCTGTTAA AAAGGTAAACACCGACATCACCAGC
	Oligo4	CATTTAACACGACCGCCGTGCTAGTAGTGCTGGT GATGTCGGTGTTTACCTTTTAAACA
	Oligo5	ATCACTAGCAACGGCGGTCGTGTTAAATGCATG

Table A1-6. Primers for the Construction of the Designed Spacer Mutant Vectors

Generated construct	Primer name	Primer sequence (5'-3')	Template
pBS-SSF-YFP-no92-19aa	92-19-F	CTCAGGTGAGCGAACCAGGAAAGCAACATTG	pBS-SSF-YFP-no92-14aa
	92-19-R	TGAGCAGCATGACCGGTTTCCCTG CAGGGAAACCGGTCATGCTGCTCACAAATGT TGCTTTCCGGTTTCGCTCACCTGAG	
pBS-SSF-YFP-no92-24aa	92-24-F	GGAAGCAACATTGTGAGCGGTGTGCCGGC	pBS-SSF-YFP-no92-19aa
	92-24-R	CCCGAGCATGACCGGTTTCCCTG CAGGGAAACCGGTCATGCTCGGGGCCGGCA CACCGCTCACAAATGTTGCTTTCC	
pBS-SSF-YFP-no92-29aa	92-29-F	GAGCGGTGTGCCGGCCCGGGTGCCCTTTCC	pBS-SSF-YFP-no92-24aa
	92-29-R	GTGCAGCATGACCGGTTTCCCTG CAGGGAAACCGGTCATGCTGCACGGAAGG CACCCGGGGCCGGCACACCGCTC	
pBS-SSF-YFP-no92-34aa	92-34-F	GGGTGCCCTTTCCGTGCCGAGCAGCTGCC	pBS-SSF-YFP-no92-29aa
	92-34-R	GAGCATGACCGGTTTCCCTG CAGGGAAACCGGTCATGCTCGGGCAGCTGC TCGGGCACGGAAGGCACCC	
pBS-SSF-YFP-no228-19aa	228-19-F	CTACCGTGACGAGCGCCGGTGGTGCCGTGG	pBS-SSF-YFP-no228-14aa
	228-19-R	GTGTGAGCATGACCGGTTTCCCTG AAACCGGTCATGCTCACACCCACGGCACCA CCGGCGCTGGTCACGGTAGAAGAAGAAATC	
pBS-SSF-YFP-no228-24aa	228-24-F	GGTGGTGCCGTGGGTGTGAGCAGCCGTCCG	pBS-SSF-YFP-no228-19aa
	228-24-R	GATAGCATGACCGGTTTCCCTG CAGGGAAACCGGTCATGCTATCCGGACGGC TGCTCACACCCACGGCACCC	
pBS-SSF-YFP-no228-29aa	228-29-F	GTGTGAGCAGCGTCCGGATTTTGCCAACC	pBS-SSF-YFP-no228-24aa
	228-29-R	CGGTGAGCATGACCGGTTTCCCTG CAGGGAAACCGGTCATGCTCACCGGTTGG CAAAATCCGGACGGCTGCTCACAC	
pBS-SSF-YFP-no228-34aa	228-34-F	CGGATTTTGCCAACCCGGTGAGCCCGGTGC	pBS-SSF-YFP-no228-29aa
	228-34-R	ATGGTAGCATGACCGGTTTCCCTG CAGGGAAACCGGTCATGCTACCATGCACCG GGCTACCGGGTTGGCAAAATCCG	
pBS-SSF-YFP-no228-39aa	228-39-F	CGGTGAGCCCGGTGCATGGTACTGTTACTT	pBS-SSF-YFP-no228-34aa
	228-39-R	CAGCAAGCATGACCGGTTTCCCTG CAGGGAAACCGGTCATGCTTGCTGAAGTAA CAGTACCATGCACCGGGCTCACCG	
pBS-SSF-YFP-no228-44aa	228-44-F	GCATGGTACTGTTACTTCAGCAGGCGGAGC	pBS-SSF-YFP-no228-39aa
	228-44-R	CGTTGGCAGCATGACCGGTTTCCCTG CAGGGAAACCGGTCATGCTGCCAACGGCTC CGCTGCTGAAGTAACAGTACCATGC	
pBS-SSF-YFP-no228-(44-10)aa	228-(44-10)-F	GCATGGCTTCTATGATTTCTTCTTCTGTGA	pBS-SSF-YFP-no228-44aa
	228-(44-10)-R	GCAGCCGTCCGGATTTTGC GCAAAATCCGGACGGCTGCTCACAGAAGAA GAAATCATAGAAGCCATGC	
pBS-SSF-YFP-no296-19aa	296-19-F	CTACCAACGTGAACAACGGTAACGGTCCGT	pBS-SSF-YFP-no228-14aa
	296-19-R	ATAGCAGCATGACCGGTTTCCCTG CAGGGAAACCGGTCATGCTGCTATACGGAC CGTTACCGTTGTTACGTTGGTAG	
pBS-SSF-YFP-no296-24aa	296-24-F	CGTAACGGTCCGTATAGCCGTAGCCAGGG	pBS-SSF-YFP-no296-19aa
	296-24-R	TTTAGCATGACCGGTTTCCCTG CAGGGAAACCGGTCATGCTAAAACCCCTGGC TACGGCTATACGGACCGTTACCG	
pBS-SSF-YFP-no296-29aa	296-29-F	CGTATAGCCGTAGCCAGGGTTTTTCAGATGA	pBS-SSF-YFP-no296-24aa
	296-29-R	GCAACGCCAGCATGACCGGTTTCCCTG CAGGGAAACCGGTCATGCTGGCGTTGCTCA TCTGAAAACCCGTGGCTACGGCTATACG	

Table A1-6. (continued)

Generated construct	Primer name	Primer sequence (5'-3')	Template
pBS-SSF-YFP-no296-34aa	296-34-F	GGGTTTTTCAGATGAGCAACGCCAGCCGAA CTTTAGCAGCATGACCGGTTTCCTG	pBS-SSF-YFP-no296-29aa
	296-34-R	CAGGGAACCGGTCATGCTGCTAAAGTTCG GCTGGGCGTGTGCTCATCTGAAAACCC	
pBS-SSF-YFP-no296-39aa	296-39-F	CGCCCAGCCGAACCTTTAGCACTAATGTGAA CAATAGCATGACCGGTTTCCTG	pBS-SSF-YFP-no296-34aa
	296-39-R	CAGGGAACCGGTCATGCTATTGTTACAT TAGTGCTAAAGTTCGGCTGGGCG	
pBS-SSF-YFP-no296-44aa	296-44-F	GAACCTTTAGCACTAATGTGAACAATGGCAA TGCTCCTTACAGCATGACCGGTTTCCTG	pBS-SSF-YFP-no296-39aa
	296-44-R	CAGGGAACCGGTCATGCTGTAAGGACCAT TGCCATTGTTACATTAGTGCTAAAGTTC	
pBS-SSF-YFP-no296-(44-10)aa	296-(44-10)-F	GCATGGCTTCTATGATTCTTCTCTAGCC GTAGCCAGGGTTTTTCAGATGAGC	pBS-SSF-YFP-no296-44aa
	296-(44-10)-R	GCTCATCTGAAAACCCCTGGCTACGGCTAGA AGAAGAAATCATAGAAGCCATGC	
pBS-SSF-YFP-no228-34aa-mFGLK	228-34-mFGLK-F	GGTGAGCCCGGTGCATGGTAACACCGACAT CACCAGCATCAC	pBS-SSF-YFP-no228-34aa
	228-34-mFGLK-R	GTGATGCTGGTGATGTCGGTGTTACCATGC ACCGGGCTCACC	
pBS-SSF-YFP-no228-29aa-hFGLK	228-29-hFGLK-F	CGGATTTTGCCAAACCCGGTGAGCATGACCG GTTTAGCCCGGTGCATGGTAACAC	pBS-SSF-YFP-no228-34aa-mFGLK
	228-29-hFGLK-R	GTGTTACCATGCACCGGGCTAAACCGGTCA TGCTCACC GGTTGGCAAATCCG	
pBS-SSF-YFP-no228-29aa-EL	228-29-EL-F	GGTGAGCATGACCGGTTTCCTGTAAAAA GGTAAGCCCGGTGCATGGTAACAC	pBS-SSF-YFP-no228-29aa-hFGLK
	228-29-EL-R	GTGTTACCATGCACCGGGCTTACCTTTTAA ACAGGGAAACCGGTCATGCTCACC	
pBS-SSF-YFP-no228-24aa-hFGLK	228-24-hFGLK-F	GTGAGCAGCCGTCCGGATAGCATGACCGGT TTTTTGCCAAACCCGGTGAGC	pBS-SSF-YFP-no228-34aa-mFGLK
	228-24-hFGLK-R	GCTCACCGGGTTGGCAAAAACCGGTCATG CTATCCGGACGGCTGCTCAC	
pBS-SSF-YFP-no228-24aa-EL	228-24-EL-F	GGTAGCATGACCGGTTTCCTGTAAAAA GGTATTTGCCAAACCCGGTGAGCC	pBS-SSF-YFP-no228-24aa-hFGLK
	228-24-EL-R	GGCTCACCGGGTTGGCAAATACCTTTTTAA CAGGGAACCCGGTCATGCTATCC	
pBS-SSF-YFP-no228-19aa-hFGLK	228-19-hFGLK-F	CGGTGGTGCCGTGGGTGTGAGCATGACCGG TTTAGCAGCCGTCCGGATTTTGCCA	pBS-SSF-YFP-no228-34aa-mFGLK
	228-19-hFGLK-R	TGGCAAAATCCGGACGGCTGCTAAACCGGT CATGCTCACACCCACGGCACCACCG	
pBS-SSF-YFP-no228-19aa-EL	228-19-EL-F	GGTGTGAGCATGACCGGTTTCCTGTAAAAA AAGGTAAGCAGCCGTCCGGATTTTGCC	pBS-SSF-YFP-no228-19aa-hFGLK
	228-19-EL-R	GGCAAAATCCGGACGGCTGCTTACCTTTTT AACAGGGAAACCGGTCATGCTCACACC	
pBS-SSF-YFP-no228-14aa-hFGLK	228-14-hFGLK-F	CCGTGACCAGCGCCGGTAGCATGACCGGTT TGGTGCCGTGGGTGTGAGC	pBS-SSF-YFP-no228-34aa-mFGLK
	228-14-hFGLK-R	GCTCACACCCACGGCACCAACCGGTCATG CTACCGGCGCTGGTCACGG	
pBS-SSF-YFP-no228-14aa-EL	228-14-EL-F	GGTAGCATGACCGGTTTCCTGTAAAAAG GTAGGTGCCGTGGGTGTGAGC	pBS-SSF-YFP-no228-14aa-hFGLK
	228-14-EL-R	GCTCACACCCACGGCACCTACCTTTTTAAC AGGGAACCGGTCATGCTACC	
pBS-SSF-YFP-no296-34aa-mFGLK	296-34-mFGLK-F	GCCAGCCGAACCTTTAGCAACACCGACATC ACCAGC	pBS-SSF-YFP-no296-34aa
	296-34-mFGLK-R	GCTGGTGATGTCGGTGTTGCTAAAGTTCGG CTGGGC	
pBS-SSF-YFP-no296-29aa-hFGLK	296-29-hFGLK-F	GGGTTTTTCAGATGAGCAACGCCAGCATGAC CGGTTTCAGCCGAACCTTTAGC	pBS-SSF-YFP-no296-34aa-mFGLK
	296-29-hFGLK-R	GCTAAAGTTCGGCTGGAACCGGTCATGCT GGCGTTGCTCATCTGAAAACCC	

Table A1-6. (continued)

Generated construct	Primer name	Primer sequence (5'-3')	Template
pBS-SSF-YFP-no296-29aa-EL	296-29-EL-F	GCCAGCATGACCGGTTTCCCTGTTAAAAAG GTACAGCCGAACTTT	pBS-SSF-YFP-no296-29aa-hFGLK
	296-29-EL-R	AAAGTTCGGCTGTACCTTTTAAACAGGGAA ACCGGTCATGCTGGC	
pBS-SSF-YFP-no296-24aa-hFGLK	296-24-hFGLK-F	CCGTATAGCCGTAGCCAGGGTTTTCAGCATG ACCGGTTTCCAGATGAGCAACGCCAGC	pBS-SSF-YFP-no296-34aa-mFGLK
	296-24-hFGLK-R	GCTGGGCGTTGCTCATCTGGAAACCGGTCA TGCTAAAACCTGGCTACGGCTATACGG	
pBS-SSF-YFP-no296-24aa-EL	296-24-EL-F	GGGTTTTAGCATGACCGGTTTCCCTGTTAA AAAGGTACAGATGAGCAACGCC	pBS-SSF-YFP-no296-24aa-hFGLK
	296-24-EL-R	GGCGTTGCTCATCTGTACCTTTTAAACAGG GAAACCGGTCATGCTAAAACCC	
pBS-SSF-YFP-no296-19aa-hFGLK	296-19-hFGLK-F	CGGTAACGGTCCGTATAGCAGCATGACCGG TTTCCGTAGCCAGGGTTTTCAGATGAGC	pBS-SSF-YFP-no296-34aa-mFGLK
	296-19-hFGLK-R	GCTCATCTGAAAACCTGGCTACGGAAACC GGTCATGCTGCTATACGGACCGTTACCG	
pBS-SSF-YFP-no296-19aa-EL	296-19-EL-F	GCAGCATGACCGGTTTCCCTGTTAAAAAGG TACGTAGCCAGGGTTTT	pBS-SSF-YFP-no296-19aa-hFGLK
	296-19-EL-R	AAAACCTGGCTACGTACCTTTTAAACAGG GAAACCGGTCATGCTGC	
pBS-SSF-YFP-no296-14aa-hFGLK	296-14-hFGLK-F	CTACCAACGTGAACAACGGTAGCATGACCG GTTTCAACGGTCCGTATAGCCGTAGCC	pBS-SSF-YFP-no296-34aa-mFGLK
	296-14-hFGLK-R	GGCTACGGCTATACGGACCGTTGAAACCGG TCATGCTACCGTTGTTACGTTGGTAG	
pBS-SSF-YFP-no296-14aa-EL	296-14-EL-F	CAACGGTAGCATGACCGGTTTCCCTGTTAA AAAGGTAAACGGTCCGTATAGCCGTAGC	pBS-SSF-YFP-no296-14aa-hFGLK
	296-14-EL-R	GCTACGGCTATACGGACCGTTTACCTTTT AACAGGGAACCGGTCATGCTACCGTTG	

Table A1-7. Primers for the Construction of the Spacer Mutant Vectors based on the Native Spacers

Generated construct	Primer name	Primer sequence (5'-3')	Template
pBS-SSF-YFP-mA	SSF-mA-F	GCATGGCTTCTATGATTCTTCTTCTTCCC GTGCTTCTCGCGGCCAGTCTGC	pBS-SSF-YFP
	SSF-mA-R	GCAGACTGGCCGCGAGAAGCACGGGAAGAA GAAGAAATCATAGAAGCCATGC	
pBS-SSF-YFP-mC	SSF-mC-F	CCACTGTTTCCCGTGCTTCTGTGTGGCTC CGTTCGGTGG	pBS-SSF-YFP
	SSF-mC-R	CCACCGAACGGAGCCACAGCAGAAGCACGG GAAACAGTGG	
pBS-SSF-YFP-mE	SSF-mE-F	GCCAGTCTGCAGCTGTGGCTCCGAGCATGA CCGGTTTCCCTGTAAAAAGG	pBS-SSF-YFP
	SSF-mE-R	CCTTTTAAACAGGGAACCGGTCATGCTCG GAGCCACAGCTGCAGACTGGC	
pBS-SSF-YFP-mEpA	SSF-mEpA-F	GTCTGCAGCTGTGGCTCCGGCAGTGACTAC CGTGAGCATGACCGGTTTCCCTG	pBS-SSF-YFP-mE
	SSF-mEpA-R	CAGGGAACCGGTCATGCTCACGGTAGTCA CTGCCGGAGCCACAGCTGCAGAC	
pBS-SSF-YFP-mEpB	SSF-mEpB-F	GTCTGCAGCTGTGGCTCCGGTGAGCCGTGC ATCTAGCATGACCGGTTTCCCTG	pBS-SSF-YFP-mE
	SSF-mEpB-R	CAGGGAACCGGTCATGCTAGATGCACGGC TCACCGGAGCCACAGCTGCAGAC	
pBS-SSF-YFP-mEpC	SSF-mEpC-F	GTCTGCAGCTGTGGCTCCGGTGGTCAATC TGCGAGCATGACCGGTTTCCCTG	pBS-SSF-YFP-mE
	SSF-mEpC-R	CAGGGAACCGGTCATGCTCGCAGATTGAC CACGCGGAGCCACAGCTGCAGAC	
pBS-SSF-YFP-pCmEpA	SSF-pCmEpA-F	GGCTTCTATGATTCTTCTTCTCGTGGTCA ATCTCGCGCCGTACCAGTGTTCCTCG	pBS-SSF-YFP-mEpA
	SSF-pCmEpA-R	CGGGAACAGTGGTAACGGCCGAGATTGA CCACGAGAAGAAGAAATCATAGAAGCC	
pBS-SSF-YFP-pDmEpB	SSF-pDmEpB-F	GGCTTCTATGATTCTTCTTCTGTGAGCCGT TGCACCGGCCGTACCAGTGTTCCTCG	pBS-SSF-YFP-mEpB
	SSF-pDmEpB-R	CGGGAACAGTGGTAACGGCCGTGCAACT GCAGCAGAAGAAGAAATCATAGAAGCC	
pBS-SSF-YFP-pBmEpC	SSF-pBmEpC-F	GGCTTCTATGATTCTTCTTCTGTGAGCCGT TGCACTCGCCGTACCAGTGTTCCTCG	pBS-SSF-YFP-mEpC
	SSF-pBmEpC-R	CGGGAACAGTGGTAACGGCAGATGCACGG CTCACAGAAGAAGAAATCATAGAAGCC	
pBS-SSF-YFP-pCBmEpA	SSF-pCBmEpA-F	CCACTGTTTCCCGTGCTTCTGTGAGCCGTG CATCTCGCGCCAGTCTGCAGCTG	pBS-SSF-YFP-pCmEpA
	SSF-pCBmEpA-R	CAGCTGCAGACTGGCCGCGAGATGCACGGC TCACAGAAGCACGGGAACAGTGG	
pBS-SSF-YFP-pDCmEpB	SSF-pDCmEpB-F	CCACTGTTTCCCGTGCTTCTCGTGGTCAAT CTGCGCGCGCCAGTCTGCAGCTG	pBS-SSF-YFP-pDmEpB
	SSF-pDCmEpB-R	CAGCTGCAGACTGGCCGCGCGCAGATTGAC CACGAGAAGCACGGGAACAGTGG	
pBS-SSF-YFP-pBCmEpC	SSF-pBCmEpC-F	CCACTGTTTCCCGTGCTTCTCGTGGTCAAT CTGCGCGCGCCAGTCTGCAGCTG	pBS-SSF-YFP-pBmEpC
	SSF-pBCmEpC-R	CAGCTGCAGACTGGCCGCGCGCAGATTGAC CACGAGAAGCACGGGAACAGTGG	
pBS-SSF-YFP-mA-ELpA	SSF-mA-ELpA-F	CGGTTTCCCTGTAAAAAGGTAGCCGTAC CACTGTAAACACCGACATCACCAGC	pBS-SSF-YFP-mA
	SSF-mA-ELpA-R	GCTGGTGATGTGCGGTGTTAACAGTGGTAAAC GGCTACCTTTTAAACAGGGAACCG	
pBS-SSF-YFP-mAE-ELpA	SSF-mAE-ELpA-F	GCCAGTCTGCAGCTGTGGCTCCGAGCATGA CCGGTTTCCCTGTAAAAAGG	pBS-SSF-YFP-mA-ELpA
	SSF-mAE-ELpA-R	CCTTTTAAACAGGGAACCGGTCATGCTCG GAGCCACAGCTGCAGACTGGC	

Table A1-7. (continued)

Generated construct	Primer name	Primer sequence (5'-3')	Template
pBS-SSF-YFP-mAE-ELpAC	SSF-mAE-ELpAC-F	GGTAGCCGTTACCACTGTTTCGCGGCCAGTC	pBS-SSF-YFP-mAE-ELpA
	SSF-mAE-ELpAC-R	TGCAAACACCGACATCACCAGC GCTGGTGATGTCGGTGTTCGAGACTGGCC GCGAACAGTGGTAACGGCTACC	
pBS-SSF-YFP-mC-ELpC	SSF-mC-ELpC-F	CGGTTTCCTGTAAAAAGGTACGCGGCCA	pBS-SSF-YFP-mC
	SSF-mC-ELpC-R	GTCTGCAAACACCGACATCACCAGC GCTGGTGATGTCGGTGTTCGAGACTGGCC GCGTACCTTTTAAACAGGGAACCG	
pBS-SSF-YFP-mC-ELpAC	SSF-mC-ELpAC-F	CGGTTTCCTGTAAAAAGGTAGCCGTTAC	pBS-SSF-YFP-mC-ELpC
	SSF-mC-ELpAC-R	CACTGTTTCGCGGCCAGTCTGCAAACACC GGTGTTCGAGACTGGCCGCAACAGTGGT AACGGCTACCTTTTAAACAGGGAACCG	
pBS-SSF-YFP-mAC-ELpAC	SSF-mAC-ELpAC-F	GCTTCTATGATTTCTTCTTCTCCCGTGCT	pBS-SSF-YFP-mC-ELpAC
	SSF-mAC-ELpAC-R	TCTGCTGTGG CCACAGCAGAAGCACGGGAAGAAGAAGAAA TCATAGAAGC	
pBS-SSF-YFP-mE-ELpE	SSF-mE-ELpE-F	CGGTTTCCTGTAAAAAGGTATTCGGTGG	pBS-SSF-YFP-mE
	SSF-mE-ELpE-R	CCTGAAGAACACCGACATCACCAGC GCTGGTGATGTCGGTGTTCAGGCCACC GAATACCTTTTAAACAGGGAACCG	
pBS-SSF-YFP-mE-ELpA	SSF-mE-ELpA-F	CGGTTTCCTGTAAAAAGGTAGCCGTTAC	pBS-SSF-YFP-mE
	SSF-mE-ELpA-R	CACTGTAAACACCGACATCACCAGC GCTGGTGATGTCGGTGTAAACAGTGGTAAC GGCTACCTTTTAAACAGGGAACCG	
pBS-SSF-YFP-mCE-ELpAC	SSF-mCE-ELpAC-F	CGTGCTTCTGCTGTGGCTCCGAGCATGACC	pBS-SSF-YFP-mC-ELpAC
	SSF-mCE-ELpAC-R	GGTTTCCTGTAAAAAGG CCTTTTTAAACAGGGAACCGGTCATGCTCG GAGCCACAGCAGAAGCAGC	
pBS-FDF-YFP-mA	FDF-mA-F	GTCTGTTTCTGCTTCTCTGCTGCGGTGGC	pBS-FDF-YFP
	FDF-mA-R	ATCTTCTCTGCCGACC GGTCGGCAGAGAAGATGCCACCGGCAGCAG AGAAGCAGAAACAGAC	
pBS-FDF-YFP-mB	FDF-mB-F	GCCGAAGCAGCAGCCGATGCCACCAACAT	pBS-FDF-YFP
	FDF-mB-R	GGGCCAGGC GCCTGGCCCATGTTGGTCGGCATCGGCTGC TGCTTCGGC	
pBS-FDF-YFP-mC	FDF-mC-F	CCGATGGTGGCATCTTCTCTGCAGGCACTG	pBS-FDF-YFP
	FDF-mC-R	TTCGGTCTGAAAGC GCTTTCAGACCGAACAGTGCCTGCAGAGAA GATGCCACCATCGG	
pBS-FDF-YFP-mApA	FDF-mApA-F	CTCTGCCGACCAACATGGGCAAGCAGCAGC	pBS-FDF-YFP-mA
	FDF-mApA-R	CGATGCAGGCACTGTTTCGGTCTGAAAGC GCTTTCAGACCGAACAGTGCCTGCATCGGC TGCTGCTTGCCCATGTTGGTCGGCAGAG	
pBS-FDF-YFP-mBpB	FDF-mBpB-F	CGATGCCGACCAACATGGGCGTGGCATCTT	pBS-FDF-YFP-mB
	FDF-mBpB-R	CTCTGCAGGCACTGTTTCGGTCTGAAAGC GCTTTCAGACCGAACAGTGCCTGCAGAGAA GATGCCACGCCCATGTTGGTCGGCATCG	
pBS-FDF-YFP-mCpC	FDF-mCpC-F	CTGTTTCTGCTTCTCTGCTGCCCGGACCA	pBS-FDF-YFP-mC
	FDF-mCpC-R	ACATGGGCAAGCAGCAGCCGATGGTGGC GCCACCATCGGCTGCTGCTTGCCCATGTTG GTCGGCGGCAGCAGAGAAGCAGAAACAG	
pBS-FDF-YFP-pA	FDF-pA-F	CTCTGCCGACCAACATGGGCAACAACAAC	pBS-FDF-YFP
	FDF-pA-R	CGATGCAGGCACTGTTTCGGTCTGAAAGC GCTTTCAGACCGAACAGTGCCTGCATCGGT TGTGTTTGCCCATGTTGGTCGGCAGAG	
pBS-FDF-YFP-pB	FDF-pB-F	CTCTGCCGACCAACATGGGCGTGGCCTCTA	pBS-FDF-YFP
	FDF-pB-R	GCCTTCAGGCACTGTTTCGGTCTGAAAGC GCTTTCAGACCGAACAGTGCCTGAAGGCTA GAGGCAACGCCCATGTTGGTCGGCAGAG	

Table A1-7. (continued)

Generated construct	Primer name	Primer sequence (5'-3')	Template
pBS-FDF-YFP-pC	FDF-pC-F	CTGTTTCTGCTTCTCTGCTGCCGCTACTA	pBS-FDF-YFP
	FDF-pC-R	ATATGGGTAAGCAGCAGCCGATGGTGGC GCCACCATCGGCTGCTGCTTACCCATATTA GTAGGCGGCAGCAGAGAAGCAGAAACAG	
pBS-FDF-YFP-pBA	FDF-pBA-F	GCTGCCGAAGCAGCAGCCGATGGTTGCCTC	pBS-FDF-YFP-pA
	FDF-pBA-R	AAGCCTTGTGGCATCTTCTCTGCCGACC GGTCGGCAGAGAAGATGCCACAAGGCTTGA GGCAACCATCGGCTGCTGCTTCGGCAGC	
pBS-FDF-YFP-pBB	FDF-pBB-F	GCTGCCGAAGCAGCAGCCGATGGTTGCCTC	pBS-FDF-YFP-pB
	FDF-pBB-R	AAGCCTTGTGGCATCTTCTCTGCCGACC GGTCGGCAGAGAAGATGCCACAAGGCTTGA GGCAACCATCGGCTGCTGCTTCGGCAGC	
pBS-FDF-YFP-pCB	FDF-pC-F	GGGTAAGCAGCAGCCGATGGTTGCCTCAAG	pBS-FDF-YFP-pC
	FDF-pC-R	CCTTGTGGCATCTTCTCTGCCGACC GGTCGGCAGAGAAGATGCCACAAGGCTTGA GGCAACCATCGGCTGCTGCTTACCC	
pBS-FDF-YFP-mA-ELpA	FDF-mA-ELpA-F	CGGTCTGAAAGCAGGTTCTAAGCAGCAGCC	pBS-FDF-YFP-mA
	FDF-mA-ELpA-R	GATGCGCGGCCGTGTTACTGCTATGC GCATAGCAGTAACACGGCCGCGCATCGGCT GCTGCTTAGAACCTGCTTTCAGACCG	
pBS-FDF-YFP-mAB-ELpA	FDF-mAB-ELpA-F	CTGTTTCTGCTTCTCTGCTGCCGCCGACCA	pBS-FDF-YFP-mA-ELpA
	FDF-mAB-ELpA-R	ACATGGGCCAGG CCTGGCCCATGTTGGTCGGCGGCAGCAGAG AAGCAGAAACAG	
pBS-FDF-YFP-mAB-ELpAB	FDF-mAB-ELpAB-F	GGTTCTAAGCAGCAGCCGATGGTGGCATCT	pBS-FDF-YFP-mAB-ELpA
	FDF-mAB-ELpAB-R	TCTCTGCGCGGCCGTGTTACTGCTATGC GCATAGCAGTAACACGGCCGCGCAGAGAAG ATGCCACCATCGGCTGCTGCTTAGAACC	
pBS-FDF-YFP-mB-ELpB	FDF-mB-ELpB-F	CGGTCTGAAAGCAGGTTCTGTGGCATCTTC	pBS-FDF-YFP-mB
	FDF-mB-ELpB-R	TCTGCGCGGCCGTGTTACTGCTATGC GCATAGCAGTAACACGGCCGCGCAGAGAAG ATGCCACAGAACCTGCTTTCAGACCG	
pBS-FDF-YFP-mBC-ELpB	FDF-mBC-ELpB-F	GCCGAAGCAGCAGCCGATGCAGGCACTGTT	pBS-FDF-YFP-mB-ELpB
	FDF-mBC-ELpB-R	CGGTCTGAAAGC GCTTTCAGACCGAACAGTGCCTGCATCGGC TGCTGCTTCGGC	
pBS-FDF-YFP-mBC-ELpBC	FDF-mBC-ELpBC-F	GGTTCTGTGGCATCTTCTCTGCCGACCAAC	pBS-FDF-YFP-mBC-ELpB
	FDF-mBC-ELpBC-R	ATGGGCCCGCGCCGTGTTACTGCTATGC GCATAGCAGTAACACGGCCGCGGCCCATGT TGGTCGGCAGAGAAGATGCCACAGAACC	
pBS-FDF-YFP-mC-ELpC	FDF-mC-ELpC-F	CGGTCTGAAAGCAGGTTCTCCGACCAACAT	pBS-FDF-YFP-mC
	FDF-mC-ELpC-R	GGGCCCGCGCCGTGTTACTGCTATGC GCATAGCAGTAACACGGCCGCGGCCCATGT TGGTCGGAGAACCTGCTTTCAGACCG	
pBS-FDF-YFP-mAC-ELpC	FDF-mAC-ELpC-F	CTGTTTCTGCTTCTCTGCTGCCGCTGGCAT	pBS-FDF-YFP-mC-ELpC
	FDF-mAC-ELpC-R	CTTCTCTGCAGG CCTGCAGAGAAGATGCCACCGCAGCAGAG AAGCAGAAACAG	
pBS-FDF-YFP-mAC-ELpAC	FDF-mAC-ELpAC-F	CGGTCTGAAAGCAGGTTCTAAGCAGCAGCC	pBS-FDF-YFP-mAC-ELpC
	FDF-mAC-ELpAC-R	GATGCCGACCAACATGGGCCCGC GCGGCCCATGTTGGTCGGCATCGGCTGCTG CTTAGAACCTGCTTTCAGACCG	

Table A1-8. Codon Optimized Synthetic DNAs

Peptides	Sequence (5'-3')
FDF	GGTAGATACATATGGCATCTACTCTGTCTACTCTGTCTGTTTCTGCTTCTCTGCT GCCGAAGCAGCAGCCGATGGTGGCATCTTCTCTGCCGACCAACATGGGCCAGGCA CTGTTCCGGTCTGAAAGCAGGTTCTCGCGGCCGTGTTACTGCTATGCCCCGGTAAT GG
FDR	GGTAGATACATATGGCTACTGTTCGTGGTCGTTCTGGTGCAAACTGGGCTTTCT GGCCCAGGGCATGAACACCCCGCTGTCCTCTGCAGTCATGCCGCAGCAGAAACCA CTGCTGTCTGCTTCCGTTTCTCTGACTTCCCTGACCTCTGCCATGCCCCGGTAAT GG
SSF	GGTAGATACATATGGCTTCTATGATTTCTTCTCTGCCGTTACCACTGTTTCCCG TGCTTCTCGCGGCCAGTCTGCAGCTGTGGCTCCGTTCCGGTGGCCTGAAGAGCATG ACCGGTTTCCCTGTTAAAAAGGTAAACACCGACATCACCAGCATCACTAGCAACG GCGGTCGTGTTAAATGCATGCCCGGGTAATGG
SSR	GGTAGATACATATGTGTAAGGTACGTGGCGGTAACAGCACTATCTCTACTATTGA CACTAACGTTAAGAAAGTTCCTTTCGGCACTATGTCTAAACTGGGCGGCTTCCCA GCGGTTGCAGCCTCCCAGGGTCGTTCCGCGCGTAGCGTTACCACCGTTGCTTCCT CCAGCATCATGAGCGCAATGCCCGGGTAATGG

Appendix 2

Arabidopsis TP Datasets

Table A2-1. TargetP-predicted *Arabidopsis* TP Dataset

Locus identifier	TargetP cTP score	TargetP reliability class	TargetP TP length
Atlg01080	0.975	1	68
Atlg01090	0.962	1	61
Atlg01500	0.825	2	50
Atlg01690	0.843	2	46
Atlg01860	0.967	1	48
Atlg02560	0.901	1	62
Atlg02730	0.972	1	58
Atlg03130	0.951	1	43
Atlg03160	0.871	2	54
Atlg03480	0.931	2	48
Atlg03600	0.930	2	67
Atlg03680	0.971	1	48
Atlg03970	0.927	2	46
Atlg04420	0.983	1	45
Atlg05750	0.758	2	54
Atlg05860	0.813	2	51
Atlg06070	0.963	1	46
Atlg06510	0.819	2	50
Atlg06730	0.663	2	51
Atlg06950	0.941	1	50
Atlg07010	0.816	2	53
Atlg07160	0.939	1	67
Atlg07460	0.753	2	51
Atlg07900	0.948	2	43
Atlg08050	0.894	2	63
Atlg08130	0.873	2	53
Atlg08380	0.974	1	40
Atlg08490	0.925	2	35
Atlg08510	0.957	1	49
Atlg08520	0.822	2	49
Atlg08640	0.959	1	58
Atlg08850	0.902	2	57
Atlg09130	0.863	2	43
Atlg09420	0.895	2	49
Atlg10390	0.883	2	48
Atlg10500	0.960	1	55
Atlg10510	0.790	2	60
Atlg10700	0.924	1	39
Atlg10830	0.972	1	58
Atlg10890	0.950	2	49
Atlg10960	0.932	2	69
Atlg11430	0.808	2	58
Atlg11750	0.930	2	51
Atlg12230	0.969	2	49
Atlg12250	0.913	2	66
Atlg12520	0.852	2	67
Atlg12800	0.939	1	49
Atlg12900	0.874	2	48

Table A2-1. (continued)

Locus identifier	TargetP cTP score	TargetP reliable class	TargetP TP length
Atlg13200	0.882	2	35
Atlg13280	0.977	1	52
Atlg13990	0.983	1	61
Atlg14030	0.937	1	57
Atlg15140	0.806	2	47
Atlg15510	0.960	1	52
Atlg15700	0.938	1	60
Atlg15730	0.676	2	50
Atlg15980	0.867	2	44
Atlg16300	0.860	2	66
Atlg16630	0.955	1	64
Atlg17440	0.938	2	56
Atlg17870	0.937	1	50
Atlg18440	0.758	2	55
Atlg18500	0.930	1	57
Atlg19150	0.978	1	49
Atlg19480	0.853	2	37
Atlg19740	0.862	2	49
Atlg20310	0.923	2	58
Atlg20340	0.874	1	52
Atlg20830	0.833	2	52
Atlg20990	0.928	1	36
Atlg21060	0.937	2	59
Atlg21350	0.952	1	59
Atlg21600	0.802	2	59
Atlg21910	0.976	1	66
Atlg22110	0.915	2	63
Atlg22170	0.846	2	48
Atlg22230	0.855	2	64
Atlg22410	0.872	2	51
Atlg22660	0.785	2	45
Atlg23400	0.818	2	43
Atlg24040	0.963	1	55
Atlg24280	0.926	2	57
Atlg24310	0.930	2	59
Atlg25500	0.933	2	39
Atlg26160	0.892	2	50
Atlg26230	0.951	2	37
Atlg26760	0.946	2	47
Atlg27210	0.844	2	54
Atlg27510	0.948	1	69
Atlg27610	0.966	1	65
Atlg28530	0.929	2	37
Atlg29040	0.961	1	59
Atlg29410	0.914	1	38
Atlg29700	0.893	1	65
Atlg29900	0.963	1	62
Atlg30100	0.920	2	45
Atlg30120	0.970	1	44
Atlg31220	0.945	1	65
Atlg32190	0.911	1	67
Atlg32200	0.679	2	62
Atlg32220	0.827	2	57
Atlg32380	0.931	1	44
Atlg32440	0.930	2	55
Atlg32500	0.897	2	36
Atlg32990	0.953	1	60
Atlg33250	0.798	2	55
Atlg34000	0.957	1	42

Table A2-1. (continued)

Locus identifier	TargetP cTP score	TargetP reliable class	TargetP TP length
Atlg34380	0.930	1	61
Atlg35340	0.956	1	39
Atlg36280	0.965	1	50
Atlg36390	0.979	1	66
Atlg41610	0.981	1	49
Atlg41640	0.982	1	49
Atlg41670	0.987	1	49
Atlg42960	0.903	2	59
Atlg43220	0.913	2	54
Atlg43840	0.930	1	36
Atlg48350	0.863	2	47
Atlg48450	0.839	2	71
Atlg48490	0.892	2	59
Atlg48850	0.892	2	39
Atlg48860	0.949	1	39
Atlg49000	0.814	2	65
Atlg49380	0.803	2	69
Atlg49970	0.933	1	41
Atlg50020	0.989	1	48
Atlg50030	0.889	2	51
Atlg50040	0.919	1	65
Atlg50170	0.919	2	46
Atlg50250	0.800	2	48
Atlg50260	0.798	2	67
Atlg50770	0.896	2	68
Atlg51100	0.854	2	38
Atlg51110	0.873	1	55
Atlg51350	0.873	2	36
Atlg51440	0.941	1	52
Atlg52220	0.941	1	55
Atlg52510	0.833	2	39
Atlg52550	0.955	1	54
Atlg52870	0.939	2	41
Atlg53050	0.976	1	36
Atlg54130	0.697	2	64
Atlg54350	0.845	2	54
Atlg54580	0.918	2	51
Atlg55040	0.823	2	39
Atlg55140	0.970	1	47
Atlg55490	0.979	1	53
Atlg55580	0.932	2	53
Atlg55670	0.963	1	59
Atlg55800	0.945	1	50
Atlg56200	0.989	1	68
Atlg56460	0.901	2	44
Atlg58060	0.948	1	37
Atlg58080	0.922	1	68
Atlg58290	0.862	2	64
Atlg59650	0.892	2	53
Atlg59660	0.846	2	39
Atlg59990	0.804	2	63
Atlg60000	0.872	1	63
Atlg60950	0.962	1	52
Atlg61520	0.891	1	48
Atlg61590	0.866	2	41
Atlg61800	0.972	1	68
Atlg61820	0.904	1	57
Atlg62140	0.934	2	60
Atlg62180	0.973	1	66

Table A2-1. (continued)

Locus identifier	TargetP cTP score	TargetP reliable class	TargetP TP length
Atlg63610	0.877	2	58
Atlg63720	0.899	2	54
Atlg63970	0.937	1	52
Atlg64350	0.825	2	53
Atlg64510	0.926	1	40
Atlg64970	0.950	2	50
Atlg65010	0.969	1	35
Atlg65260	0.965	1	64
Atlg66430	0.975	1	46
Atlg66670	0.904	1	71
Atlg67080	0.895	1	37
Atlg67280	0.903	2	63
Atlg67560	0.822	2	40
Atlg67740	0.871	2	50
Atlg67810	0.916	1	41
Atlg67930	0.879	2	71
Atlg68070	0.969	1	48
Atlg68160	0.845	2	71
Atlg68450	0.951	2	58
Atlg68460	0.974	1	71
Atlg68880	0.756	2	61
Atlg69240	0.978	1	58
Atlg69650	0.804	2	42
Atlg69740	0.835	2	52
Atlg69830	0.911	1	55
Atlg70070	0.868	2	58
Atlg70200	0.896	1	49
Atlg70820	0.890	1	49
Atlg71480	0.858	2	70
Atlg71500	0.889	2	62
Atlg71720	0.867	2	63
Atlg71920	0.957	1	39
Atlg72010	0.813	2	55
Atlg72520	0.938	1	58
Atlg72540	0.921	1	38
Atlg72640	0.763	2	44
Atlg72810	0.957	1	57
Atlg73060	0.812	2	58
Atlg73110	0.866	2	39
Atlg73150	0.896	2	51
Atlg73430	0.840	2	43
Atlg73470	0.781	2	44
Atlg73740	0.778	2	56
Atlg73760	0.883	2	48
Atlg74040	0.960	1	46
Atlg74070	0.908	1	45
Atlg74470	0.906	2	43
Atlg74600	0.912	1	71
Atlg74730	0.803	2	44
Atlg74850	0.821	2	66
Atlg74960	0.675	2	52
Atlg74980	0.867	2	55
Atlg75260	0.853	2	67
Atlg75330	0.968	1	53
Atlg75390	0.912	2	49
Atlg75400	0.851	2	45
Atlg75460	0.926	1	50
Atlg75690	0.907	2	43
Atlg76050	0.741	2	43

Table A2-1. (continued)

Locus identifier	TargetP cTP score	TargetP reliable class	TargetP TP length
Atlg76080	0.946	1	56
Atlg76100	0.963	1	66
Atlg76570	0.758	2	64
Atlg76620	0.963	1	44
Atlg77390	0.852	2	57
Atlg77640	0.956	1	59
Atlg77930	0.868	2	67
Atlg78310	0.973	2	54
Atlg78560	0.986	1	70
Atlg78620	0.909	2	65
Atlg78630	0.962	1	56
Atlg79050	0.841	2	51
Atlg79560	0.810	2	49
Atlg79600	0.773	2	42
Atlg79850	0.957	1	48
Atlg80040	0.932	1	71
Atlg80370	0.788	2	51
Atlg80480	0.935	1	68
Atlg80670	0.890	2	36
Atlg80920	0.934	1	46
At2g01590	0.971	1	54
At2g01940	0.918	2	50
At2g02070	0.836	2	50
At2g02800	0.978	1	69
At2g02980	0.843	2	38
At2g03140	0.955	1	69
At2g03400	0.950	1	65
At2g04530	0.970	1	68
At2g05070	0.801	2	41
At2g06520	0.936	1	58
At2g07370	0.968	2	57
At2g07710	0.954	1	49
At2g10550	0.919	1	53
At2g12900	0.979	1	52
At2g12980	0.958	1	46
At2g13130	0.965	1	52
At2g13150	0.958	1	46
At2g14880	0.982	1	43
At2g15570	0.979	1	67
At2g16570	0.973	1	58
At2g17220	0.975	1	41
At2g17240	0.961	1	59
At2g17300	0.873	2	56
At2g17540	0.872	1	53
At2g17630	0.956	1	49
At2g17880	0.969	1	39
At2g18470	0.965	1	66
At2g18710	0.946	1	67
At2g20020	0.820	2	37
At2g20080	0.972	1	63
At2g20260	0.917	1	46
At2g20270	0.905	1	61
At2g20890	0.927	1	67
At2g20920	0.877	2	59
At2g21170	0.950	2	58
At2g21340	0.969	1	56
At2g21530	0.862	2	61
At2g22880	0.894	2	57
At2g23070	0.951	1	55

Table A2-1. (continued)

Locus identifier	TargetP cTP score	TargetP reliability class	TargetP TP length
At2g23160	0.938	2	50
At2g23670	0.928	2	57
At2g23720	0.880	2	56
At2g24060	0.789	2	55
At2g24090	0.911	1	56
At2g24820	0.897	1	48
At2g25250	0.937	2	60
At2g26000	0.888	2	62
At2g26100	0.845	2	36
At2g26280	0.933	2	67
At2g26610	0.963	1	64
At2g26670	0.926	1	54
At2g27510	0.859	2	49
At2g27660	0.890	2	58
At2g27680	0.895	2	35
At2g27950	0.888	2	37
At2g28000	0.884	2	45
At2g28190	0.960	1	61
At2g28800	0.919	2	55
At2g28880	0.959	1	43
At2g28930	0.858	2	59
At2g29180	0.959	1	65
At2g29280	0.896	1	61
At2g29630	0.918	1	37
At2g29650	0.986	1	59
At2g29760	0.896	2	42
At2g30390	0.933	1	49
At2g30790	0.951	2	42
At2g30950	0.911	2	47
At2g31250	0.866	2	52
At2g31350	0.884	2	64
At2g31400	0.940	1	41
At2g31840	0.840	2	48
At2g32140	0.869	2	42
At2g33800	0.929	1	49
At2g34420	0.849	2	37
At2g34590	0.979	1	70
At2g35260	0.954	1	54
At2g35490	0.944	1	53
At2g35600	0.893	2	55
At2g36000	0.945	1	55
At2g36390	0.898	2	37
At2g37000	0.941	1	63
At2g37080	0.921	1	55
At2g37220	0.976	1	47
At2g37240	0.924	2	50
At2g37420	0.879	2	54
At2g37660	0.876	1	69
At2g38040	0.930	2	54
At2g38060	0.905	2	44
At2g38140	0.923	2	54
At2g38270	0.956	1	62
At2g38360	0.844	2	44
At2g38450	0.732	2	56
At2g38780	0.914	2	63
At2g39000	0.928	2	61
At2g39080	0.948	1	64
At2g39730	0.888	1	58
At2g39830	0.839	2	58

Table A2-1. (continued)

Locus identifier	TargetP cTP score	TargetP reliability class	TargetP TP length
At2g39990	0.797	2	40
At2g40300	0.920	1	52
At2g40380	0.936	1	36
At2g40490	0.885	2	35
At2g41040	0.949	1	53
At2g41120	0.937	1	39
At2g41180	0.967	1	38
At2g41680	0.952	1	67
At2g42130	0.874	1	48
At2g42220	0.913	2	53
At2g42520	0.956	2	40
At2g42620	0.868	2	51
At2g42750	0.930	1	59
At2g42940	0.923	2	48
At2g43030	0.912	1	49
At2g43090	0.973	1	59
At2g43100	0.968	1	59
At2g43180	0.956	1	59
At2g43710	0.969	1	35
At2g43750	0.938	1	58
At2g43970	0.960	2	48
At2g44050	0.878	2	68
At2g44650	0.908	1	39
At2g44700	0.957	1	51
At2g44940	0.921	1	43
At2g45290	0.970	1	65
At2g45770	0.847	2	40
At2g46100	0.764	2	40
At2g46590	0.940	2	56
At2g46820	0.966	1	45
At2g47390	0.814	2	62
At2g47450	0.943	2	49
At2g47730	0.914	1	49
At2g48090	0.964	1	43
At3g01170	0.896	1	44
At3g01180	0.938	2	55
At3g01200	0.964	2	69
At3g01370	0.720	2	45
At3g01480	0.974	1	36
At3g01500	0.977	1	47
At3g02060	0.952	1	52
At3g02450	0.895	2	42
At3g02610	0.850	2	44
At3g02660	0.860	2	64
At3g02690	0.981	1	68
At3g02730	0.856	2	57
At3g02750	0.924	2	59
At3g03880	0.930	2	50
At3g04340	0.951	1	43
At3g04510	0.945	2	62
At3g04550	0.846	2	61
At3g05020	0.931	1	53
At3g06180	0.904	2	69
At3g06430	0.926	1	36
At3g06590	0.973	2	36
At3g06660	0.839	2	64
At3g07560	0.920	2	60
At3g08630	0.918	2	58
At3g08640	0.892	2	59

Table A2-1. (continued)

Locus identifier	TargetP cTP score	TargetP reliability class	TargetP TP length
At3g08940	0.875	2	39
At3g09070	0.887	2	47
At3g09150	0.908	2	45
At3g09650	0.828	2	65
At3g10060	0.791	2	56
At3g10130	0.919	1	70
At3g10670	0.899	2	66
At3g10690	0.975	1	71
At3g10940	0.852	2	61
At3g11050	0.915	2	44
At3g11330	0.725	2	64
At3g11490	0.790	2	43
At3g11670	0.895	2	58
At3g11690	0.816	2	64
At3g12080	0.860	2	65
At3g12590	0.962	2	63
At3g12930	0.915	1	66
At3g13180	0.889	2	64
At3g13470	0.924	1	49
At3g14050	0.920	1	63
At3g14390	0.961	1	48
At3g14490	0.809	2	48
At3g14900	0.828	2	35
At3g15000	0.914	2	56
At3g15050	0.910	1	61
At3g15095	0.982	1	68
At3g15490	0.843	2	52
At3g15520	0.820	2	52
At3g15690	0.706	2	54
At3g15840	0.923	2	49
At3g15900	0.960	1	60
At3g15940	0.816	2	40
At3g16000	0.843	2	41
At3g16250	0.885	2	48
At3g16890	0.814	2	51
At3g16950	0.977	1	70
At3g17040	0.904	1	68
At3g17100	0.978	2	35
At3g17170	0.877	2	47
At3g17600	0.979	1	56
At3g17830	0.793	2	57
At3g18040	0.964	1	42
At3g18110	0.733	2	44
At3g18270	0.916	1	70
At3g18390	0.921	2	56
At3g18420	0.891	2	39
At3g18630	0.928	1	49
At3g18650	0.861	2	39
At3g18680	0.963	1	53
At3g18870	0.974	1	54
At3g19110	0.803	2	49
At3g19120	0.822	2	67
At3g19160	0.876	2	35
At3g19480	0.847	2	38
At3g19490	0.947	1	39
At3g20150	0.913	1	54
At3g20230	0.772	2	68
At3g20320	0.882	1	45
At3g20330	0.914	2	68

Table A2-1. (continued)

Locus identifier	TargetP cTP score	TargetP reliability class	TargetP TP length
At3g20350	0.921	2	53
At3g20420	0.904	2	54
At3g20490	0.880	2	49
At3g20680	0.946	1	60
At3g20930	0.968	1	54
At3g21140	0.934	2	46
At3g21200	0.929	1	41
At3g21290	0.926	1	64
At3g21810	0.814	2	63
At3g22150	0.917	2	50
At3g22890	0.945	1	47
At3g22960	0.862	2	47
At3g23290	0.966	2	65
At3g23740	0.912	2	38
At3g23790	0.894	1	47
At3g23940	0.809	2	35
At3g24860	0.986	1	47
At3g25410	0.973	1	70
At3g25780	0.984	1	56
At3g25860	0.981	1	47
At3g25920	0.866	2	65
At3g26060	0.904	2	57
At3g26650	0.850	2	45
At3g26710	0.912	2	44
At3g26740	0.872	1	41
At3g26840	0.945	1	65
At3g26900	0.908	1	56
At3g27110	0.767	2	47
At3g27160	0.966	1	47
At3g27210	0.956	1	64
At3g27580	0.915	2	65
At3g27830	0.941	1	58
At3g27840	0.960	1	59
At3g27850	0.955	1	54
At3g28460	0.749	2	43
At3g29185	0.848	2	41
At3g29200	0.969	1	52
At3g29770	0.908	2	46
At3g30780	0.846	2	57
At3g32930	0.846	2	53
At3g43610	0.825	2	51
At3g44880	0.969	1	49
At3g44890	0.798	2	41
At3g45140	0.765	2	56
At3g45890	0.954	2	68
At3g46440	0.842	2	49
At3g46880	0.827	2	46
At3g47470	0.898	2	49
At3g47650	0.981	1	56
At3g47970	0.853	2	67
At3g48070	0.914	2	48
At3g48420	0.957	1	65
At3g48560	0.976	1	55
At3g48560	0.976	1	55
At3g49170	0.868	2	50
At3g49350	0.910	2	55
At3g49680	0.924	2	60
At3g50180	0.919	2	57
At3g50240	0.902	2	49

Table A2-1. (continued)

Locus identifier	TargetP cTP score	TargetP reliability class	TargetP TP length
At3g50770	0.970	1	46
At3g50880	0.843	2	63
At3g51820	0.849	2	57
At3g51930	0.953	1	57
At3g52150	0.859	1	56
At3g52960	0.936	1	70
At3g53130	0.936	1	36
At3g53460	0.903	1	65
At3g53580	0.803	2	51
At3g53860	0.932	1	70
At3g53900	0.927	2	61
At3g54050	0.941	1	57
At3g54210	0.987	1	53
At3g54220	0.973	1	47
At3g54290	0.911	2	41
At3g54320	0.975	1	57
At3g54610	0.946	2	66
At3g54680	0.886	2	59
At3g54900	0.818	2	63
At3g55040	0.772	2	56
At3g55250	0.883	2	45
At3g55270	0.835	2	59
At3g55400	0.834	2	58
At3g55560	0.809	2	37
At3g55800	0.937	1	59
At3g56090	0.960	1	48
At3g56110	0.845	2	53
At3g56130	0.970	1	56
At3g56160	0.951	1	47
At3g56410	0.967	1	53
At3g56650	0.954	1	65
At3g56700	0.851	2	47
At3g56810	0.913	2	68
At3g56910	0.978	1	64
At3g56940	0.940	1	36
At3g57050	0.935	1	54
At3g57070	0.916	1	52
At3g57560	0.954	1	49
At3g57950	0.828	2	38
At3g58010	0.949	1	53
At3g58140	0.844	2	53
At3g58610	0.980	1	70
At3g58830	0.964	1	58
At3g58850	0.883	2	38
At3g58990	0.945	1	56
At3g59400	0.931	1	69
At3g59870	0.957	1	67
At3g59890	0.763	2	53
At3g60000	0.928	2	71
At3g60210	0.890	1	61
At3g60410	0.922	2	39
At3g60750	0.960	1	65
At3g61470	0.966	2	44
At3g61680	0.892	2	67
At3g61780	0.928	2	47
At3g62910	0.907	2	50
At3g63410	0.983	1	51
At3g63490	0.937	1	70
At4g00150	0.829	2	48

Table A2-1. (continued)

Locus identifier	TargetP cTP score	TargetP reliability class	TargetP TP length
At4g00270	0.887	2	36
At4g00620	0.880	2	60
At4g01150	0.916	2	62
At4g01650	0.934	1	51
At4g01940	0.953	1	68
At4g02040	0.942	1	65
At4g02770	0.895	2	44
At4g02780	0.912	1	60
At4g02800	0.965	2	54
At4g04020	0.895	2	55
At4g04350	0.915	2	56
At4g04480	0.852	2	56
At4g04610	0.964	1	53
At4g04640	0.963	1	42
At4g04770	0.985	1	63
At4g05070	0.806	2	69
At4g05390	0.944	1	47
At4g08330	0.829	2	60
At4g08510	0.912	2	64
At4g08600	0.948	1	66
At4g08650	0.811	2	37
At4g10000	0.901	2	47
At4g10030	0.959	1	62
At4g10300	0.799	2	37
At4g10340	0.892	2	48
At4g10620	0.859	2	58
At4g10750	0.954	1	65
At4g10840	0.874	2	65
At4g11680	0.927	2	48
At4g11910	0.817	2	54
At4g12060	0.936	1	58
At4g12800	0.879	2	50
At4g13050	0.873	2	48
At4g13200	0.974	1	56
At4g13220	0.840	2	63
At4g13670	0.900	2	40
At4g14070	0.896	2	45
At4g14680	0.954	1	49
At4g14700	0.957	1	49
At4g14770	0.892	2	65
At4g14870	0.846	2	38
At4g14890	0.888	2	56
At4g15560	0.956	1	58
At4g16060	0.899	1	51
At4g17040	0.958	1	68
At4g17070	0.959	1	49
At4g17600	0.954	1	39
At4g18240	0.779	2	42
At4g18320	0.829	2	38
At4g18440	0.890	2	57
At4g18480	0.971	1	60
At4g19100	0.891	2	35
At4g20120	0.945	1	62
At4g20210	0.925	1	42
At4g20360	0.975	1	67
At4g21280	0.889	2	44
At4g21460	0.869	2	58
At4g21660	0.954	2	60
At4g21990	0.984	1	69

Table A2-1. (continued)

Locus identifier	TargetP cTP score	TargetP reliability class	TargetP TP length
At4g22240	0.920	2	59
At4g22260	0.965	2	56
At4g22370	0.942	1	35
At4g22890	0.959	2	60
At4g22920	0.959	1	48
At4g23450	0.929	2	61
At4g23940	0.984	1	53
At4g24090	0.806	2	45
At4g24390	0.911	2	60
At4g24620	0.949	1	48
At4g24750	0.856	2	56
At4g25050	0.939	1	48
At4g25080	0.866	2	39
At4g25130	0.913	1	68
At4g25270	0.920	2	47
At4g25370	0.974	1	63
At4g25650	0.981	1	55
At4g25700	0.891	2	51
At4g25770	0.888	2	59
At4g25970	0.849	2	48
At4g25990	0.979	2	69
At4g26370	0.887	2	61
At4g26500	0.827	2	66
At4g26550	0.968	1	39
At4g26900	0.965	1	55
At4g27070	0.963	1	50
At4g27370	0.788	2	70
At4g27440	0.878	1	43
At4g27670	0.868	2	43
At4g28030	0.964	1	65
At4g28240	0.855	2	39
At4g28730	0.833	2	61
At4g28750	0.900	2	44
At4g29840	0.883	2	38
At4g29890	0.839	2	47
At4g30620	0.823	2	48
At4g30740	0.963	1	40
At4g31040	0.967	2	47
At4g31530	0.846	2	56
At4g31560	0.820	2	48
At4g31870	0.969	1	69
At4g32020	0.881	2	70
At4g33170	0.779	2	47
At4g33480	0.966	1	64
At4g33540	0.883	2	45
At4g34020	0.949	1	60
At4g34100	0.957	1	50
At4g34120	0.879	1	71
At4g34190	0.823	2	61
At4g34200	0.967	1	53
At4g34590	0.950	2	41
At4g34740	0.994	1	53
At4g35600	0.950	1	38
At4g35630	0.938	1	63
At4g35680	0.945	1	52
At4g35890	0.975	1	57
At4g35980	0.939	2	56
At4g36040	0.899	2	63
At4g36530	0.962	1	51

Table A2-1. (continued)

Locus identifier	TargetP cTP score	TargetP reliability class	TargetP TP length
At4g36810	0.903	2	56
At4g36910	0.916	1	71
At4g37510	0.942	1	51
At4g38610	0.968	1	66
At4g38880	0.961	1	59
At4g38970	0.810	2	46
At4g39040	0.884	2	65
At4g39610	0.946	2	59
At4g39690	0.952	1	39
At4g39740	0.808	2	37
At4g39970	0.978	1	46
At4g39980	0.864	2	47
At5g01310	0.982	1	66
At5g01600	0.876	2	47
At5g01920	0.972	1	49
At5g02020	0.917	1	56
At5g02120	0.758	2	40
At5g02160	0.903	1	45
At5g02250	0.908	2	35
At5g02600	0.943	1	70
At5g03110	0.887	2	57
At5g03415	0.972	2	39
At5g03800	0.934	1	58
At5g03880	0.918	2	46
At5g04140	0.828	2	62
At5g04260	0.858	2	55
At5g04360	0.903	1	62
At5g04710	0.918	1	63
At5g04770	0.949	1	50
At5g04980	0.919	1	55
At5g05380	0.931	2	41
At5g05400	0.889	2	36
At5g05460	0.925	2	68
At5g05580	0.875	2	42
At5g06340	0.864	1	44
At5g06790	0.822	2	58
At5g06930	0.935	2	55
At5g07950	0.983	1	63
At5g08050	0.958	1	62
At5g08650	0.980	1	45
At5g08740	0.933	2	52
At5g09760	0.856	2	38
At5g09790	0.935	2	44
At5g09820	0.944	2	61
At5g10160	0.872	2	48
At5g10330	0.957	1	39
At5g10620	0.779	2	42
At5g10920	0.965	1	45
At5g11250	0.890	1	55
At5g11270	0.977	1	65
At5g11480	0.797	2	43
At5g11840	0.936	1	42
At5g11880	0.863	2	49
At5g12860	0.937	1	69
At5g13110	0.816	2	50
At5g13310	0.950	1	35
At5g13340	0.912	2	58
At5g13420	0.978	1	61
At5g13510	0.887	2	40

Table A2-1. (continued)

Locus identifier	TargetP cTP score	TargetP reliability class	TargetP TP length
At5g13720	0.872	2	60
At5g13730	0.871	2	52
At5g13800	0.878	2	46
At5g13840	0.928	1	53
At5g14010	0.962	1	64
At5g14100	0.891	2	49
At5g14590	0.916	2	47
At5g14910	0.955	1	38
At5g15390	0.855	1	59
At5g15450	0.920	1	67
At5g15760	0.973	1	45
At5g15980	0.878	2	56
At5g16110	0.886	2	60
At5g16230	0.966	2	35
At5g16440	0.963	1	52
At5g16620	0.856	2	42
At5g16670	0.945	1	56
At5g17230	0.911	1	70
At5g17520	0.940	1	47
At5g17630	0.901	2	55
At5g17660	0.860	1	52
At5g17710	0.788	2	64
At5g17840	0.975	1	42
At5g17990	0.810	2	63
At5g18660	0.896	2	49
At5g18910	0.958	2	60
At5g19020	0.788	2	67
At5g19050	0.946	1	57
At5g19220	0.923	1	54
At5g19380	0.807	2	48
At5g19460	0.904	2	49
At5g19540	0.964	1	54
At5g19940	0.974	1	50
At5g20190	0.878	2	64
At5g20720	0.901	2	50
At5g22090	0.922	1	43
At5g22340	0.966	1	45
At5g22510	0.846	2	45
At5g22630	0.869	2	38
At5g22830	0.888	2	62
At5g23010	0.940	1	49
At5g23040	0.971	1	47
At5g23120	0.906	2	60
At5g23240	0.942	1	38
At5g23310	0.945	2	41
At5g24000	0.937	1	51
At5g24300	0.938	2	49
At5g24700	0.885	1	41
At5g25380	0.973	1	69
At5g25510	0.851	2	47
At5g26030	0.869	2	35
At5g26570	0.992	1	51
At5g26940	0.776	2	63
At5g27200	0.926	1	54
At5g27280	0.937	1	39
At5g27380	0.973	1	55
At5g27390	0.899	2	58
At5g27860	0.971	1	42
At5g28430	0.909	2	44

Table A2-1. (continued)

Locus identifier	TargetP cTP score	TargetP reliability class	TargetP TP length
At5g28500	0.846	2	51
At5g28750	0.838	2	59
At5g35360	0.919	2	70
At5g35630	0.926	1	45
At5g35790	0.983	1	50
At5g36120	0.957	1	39
At5g36790	0.879	2	62
At5g37360	0.943	1	53
At5g38060	0.956	1	45
At5g39980	0.800	2	64
At5g40030	0.967	1	55
At5g40140	0.865	2	41
At5g40160	0.931	1	39
At5g40470	0.944	2	49
At5g41960	0.966	1	57
At5g42270	0.829	2	58
At5g42390	0.920	2	54
At5g42750	0.913	2	50
At5g43050	0.793	2	44
At5g43160	0.980	2	37
At5g43540	0.892	2	46
At5g43780	0.878	2	40
At5g43930	0.970	1	44
At5g44000	0.940	2	56
At5g45390	0.963	1	60
At5g45930	0.943	1	60
At5g46420	0.958	1	52
At5g47110	0.959	1	38
At5g47750	0.901	1	47
At5g47840	0.941	2	57
At5g47870	0.865	2	39
At5g48110	0.927	1	52
At5g48130	0.931	2	59
At5g48300	0.976	1	70
At5g48370	0.849	2	38
At5g48440	0.931	2	37
At5g48790	0.957	1	48
At5g48830	0.928	1	62
At5g48910	0.850	2	53
At5g48990	0.882	2	43
At5g50210	0.935	1	69
At5g50250	0.922	1	71
At5g50280	0.852	2	44
At5g50350	0.910	2	68
At5g51100	0.771	2	46
At5g51110	0.827	2	50
At5g51670	0.919	2	40
At5g51920	0.972	1	42
At5g52250	0.934	2	66
At5g52570	0.905	1	52
At5g52810	0.861	2	42
At5g52920	0.885	2	63
At5g52960	0.914	1	42
At5g53080	0.953	1	70
At5g53170	0.919	2	63
At5g53450	0.976	1	52
At5g53570	0.971	2	52
At5g54090	0.899	2	40
At5g54180	0.889	2	64

Table A2-1. (continued)

Locus identifier	TargetP cTP score	TargetP reliability class	TargetP TP length
At5g54190	0.957	1	53
At5g54430	0.938	2	43
At5g54600	0.920	2	49
At5g54730	0.919	2	48
At5g54770	0.925	1	45
At5g54800	0.959	2	64
At5g54810	0.948	1	52
At5g55570	0.940	1	67
At5g55740	0.928	1	59
At5g56050	0.867	2	63
At5g56250	0.837	2	43
At5g57180	0.920	1	59
At5g57960	0.861	2	46
At5g58330	0.951	1	52
At5g59080	0.917	2	68
At5g60050	0.919	2	63
At5g60750	0.934	2	51
At5g61120	0.913	2	55
At5g61410	0.925	1	45
At5g62840	0.909	2	56
At5g63300	0.967	1	37
At5g63310	0.954	1	62
At5g63420	0.877	2	70
At5g63570	0.930	1	40
At5g63830	0.948	2	61
At5g64090	0.937	1	45
At5g64280	0.829	2	54
At5g64290	0.819	2	68
At5g64300	0.953	2	56
At5g65220	0.924	2	58
At5g65480	0.833	2	49
At5g65530	0.939	2	57
At5g65780	0.896	2	54
At5g65840	0.959	1	65
At5g66090	0.968	1	59
At5g66190	0.888	2	64
At5g66480	0.764	2	48
At5g66530	0.884	2	37

Table A2-2. Results of Hsp70 Binding Site Clustering and TargetP Prediction of the 208-TP Dataset

Hsp70 cluster group	SWISS- PROT ID	Lee et al. (2008) Group	TargetP prediction		
			Localization	Reliability class	TP length
1	Q9SUI7	None	Chloroplast	3	44
1	Q8W0Y8	None	Chloroplast	2	49
1	Q949Y5	None	Chloroplast	3	18
1	Q9XI84	DnaJ-J8	Chloroplast	1	57
1	P27521	PORA	Chloroplast	2	49
1	P52032	None	Chloroplast	1	72
1	Q9LYU9	DnaJ-J8	Chloroplast	3	80
1	Q9SAG8	DnaJ-J8	Chloroplast	1	46
1	Q41932	None	Chloroplast	3	48
1	Q39195	DnaJ-J8	Chloroplast	1	69
1	P93009	RbcS	Chloroplast	3	70
1	P25856	PORA	Chloroplast	2	45
1	P10797	RbcS	Chloroplast	3	54
1	Q96242	BCCP	Chloroplast	2	32
1	P25697	DnaJ-J8	Chloroplast	3	54
1	O22170	None	Other	2	-
1	Q9C5W3	None	Chloroplast	2	61
1	O49347	Cab	Chloroplast	2	50
1	O22886	None	Chloroplast	2	35
2	O24621	BCCP	Chloroplast	1	32
2	Q9STE8	DnaJ-J8	Chloroplast	1	79
2	Q9SCY0	DnaJ-J8	Chloroplast	5	52
2	Q9FKP0	None	Secretory Pathway	5	16
2	O22056	GLU2	Chloroplast	3	39
2	Q93WC9	None	Other	4	-
2	P82538	None	Chloroplast	1	74
2	Q949X0	TOCC	Chloroplast	2	53
2	Q9SI53	TOCC	Mitochondrion	5	76
2	Q0WZN5	None	Chloroplast	5	33
2	Q8L493	RbcS	Chloroplast	2	57
2	Q9ZR03	None	Chloroplast	3	50
2	Q9C642	None	Chloroplast	2	48
2	P25702	None	Chloroplast	4	51
2	P25701	DnaJ-J8	Chloroplast	2	51
2	Q8LPN3	BCCP	Chloroplast	1	45
2	Q9S7N7	None	Chloroplast	1	59
2	Q9SR43	DnaJ-J8	Chloroplast	2	45
2	Q9SYI0	TOCC	Chloroplast	4	61
2	Q9FYC2	RbcS	Chloroplast	1	49
3	Q96291	BCCP	Chloroplast	1	83
3	O24600	None	Chloroplast	1	95
3	P04777	None	Chloroplast	3	23
3	Q8LCA1	GLU2	Chloroplast	1	45
3	Q9LS03	None	Chloroplast	1	78
3	Q93ZB2	None	Secretory Pathway	1	16
3	Q8S9K3	Cab	Chloroplast	4	33
3	P10796	RbcS	Chloroplast	3	54
3	P10798	RbcS	Chloroplast	4	54
3	P10795	RbcS	Chloroplast	3	54
3	Q9SRL5	RbcS	Chloroplast	2	44
3	Q9FYL3	PORA	Chloroplast	3	45
3	P50318	RbcS	Chloroplast	1	95
3	Q9SMW0	None	Chloroplast	3	49
3	Q93ZC5	DnaJ-J8	Chloroplast	1	52
3	Q93W77	RbcS	Chloroplast	1	69
3	Q39101	RbcS	Chloroplast	2	47
3	O24629	None	Other	5	-

Table A2-2. (continued)

Hsp70 cluster group	SWISS- PROT ID	Lee et al. (2008) Group	TargetP prediction		
			Localization	Reliability class	TP length
3	Q9SUI5	DnaJ-J8	Chloroplast	4	32
3	Q9SKP6	DnaJ-J8	Chloroplast	2	58
3	Q9SUI4	RbcS	Chloroplast	2	50
3	Q39194	RbcS	Chloroplast	3	44
3	Q9LQK7	BCCP	Chloroplast	2	36
3	Q02166	None	Chloroplast	2	63
3	O78310	None	Chloroplast	1	61
3	P04778	None	Chloroplast	3	23
3	Q9FUZ2	RbcS	Chloroplast	5	56
3	Q3EAJ6	Cab	Chloroplast	4	35
4	Q9LDH1	None	Chloroplast	3	54
4	Q9LTF4	None	Other	5	-
4	Q9ZS97	None	Chloroplast	1	63
4	Q9M7I7	None	Chloroplast	4	47
4	P57720	RbcS	Chloroplast	2	39
4	P46248	RbcS	Chloroplast	4	41
4	Q9SQK3	PORA	Chloroplast	1	39
4	P82658	None	Chloroplast	4	32
4	Q8LBP4	None	Chloroplast	2	55
4	P25851	RbcS	Chloroplast	1	57
4	Q9FGS4	DnaJ-J8	Chloroplast	1	69
4	Q9LER7	Cab	Chloroplast	1	64
4	P46644	None	Chloroplast	4	64
4	P21238	None	Chloroplast	2	45
4	Q05753	RbcS	Chloroplast	4	36
4	Q9LS01	GLU2	Chloroplast	1	56
4	Q9SA14	DnaJ-J8	Mitochondrion	5	35
4	Q944G9	None	Chloroplast	2	46
4	Q9C5U8	RbcS	Chloroplast	4	29
4	O48782	None	Chloroplast	1	54
4	Q39199	DnaJ-J8	Chloroplast	2	51
4	Q39002	None	Chloroplast	5	21
4	O80575	None	Chloroplast	2	68
4	Q9M5K4	None	Chloroplast	1	23
4	O82499	BCCP	Mitochondrion	2	46
4	Q94FY7	TOCC	Chloroplast	1	98
4	Q9XFH8	BCCP	Chloroplast	2	57
4	Q8W033	BCCP	Other	5	-
4	Q9LYN2	RbcS	Chloroplast	1	48
4	Q9S720	None	Chloroplast	2	26
5	Q9XF89	None	Chloroplast	2	48
5	Q9S714	None	Chloroplast	1	46
5	Q3E9T1	None	Chloroplast	2	33
5	Q94AY1	None	Chloroplast	5	74
5	Q9SRN1	None	Other	4	-
5	Q3EA16	RbcS	Other	4	-
5	P81760	GLU2	Chloroplast	4	33
5	Q9S831	None	Chloroplast	2	44
5	Q42029	None	Chloroplast	2	31
5	Q9ZNZ7	GLU2	Chloroplast	2	62
5	Q42536	PORA	Chloroplast	1	53
5	Q9ZU32	Cab	Mitochondrion	3	26
5	Q42588	Cab	Other	2	-
5	Q948R9	None	Chloroplast	2	47
5	Q00218	RbcS	Chloroplast	4	47
5	Q8LEB5	None	Other	1	-
5	O82730	None	Other	3	-
5	Q9S841	None	Chloroplast	3	28

Table A2-2. (continued)

Hsp70 cluster group	SWISS- PROT ID	Lee et al. (2008) Group	TargetP prediction		
			Localization	Reliability class	TP length
5	Q01908	None	Chloroplast	1	42
5	P21218	PORA	Chloroplast	1	43
5	P10896	DnaJ-J8	Chloroplast	1	58
5	Q5XET4	PORA	Chloroplast	2	60
5	Q84JH7	None	Chloroplast	2	46
5	Q9XIK3	None	Chloroplast	1	55
5	Q9XF91	DnaJ-J8	Chloroplast	5	59
5	Q39161	None	Chloroplast	3	25
5	O04921	None	Chloroplast	1	49
5	Q9T0P4	GLU2	Chloroplast	2	72
5	P69834	DnaJ-J8	Mitochondrion	5	92
5	O64903	None	Chloroplast	1	62
5	O22160	None	Chloroplast	1	34
6	O04130	None	Chloroplast	3	49
6	O48741	PORA	Chloroplast	3	66
6	Q9LD95	None	Chloroplast	1	55
6	Q9SW33	None	Chloroplast	4	21
6	O22773	None	Chloroplast	1	38
6	Q9M401	PORA	Chloroplast	2	60
6	Q9CAK8	RbcS	Chloroplast	1	52
6	P92935	None	Mitochondrion	5	80
6	Q84W65	None	Chloroplast	2	66
6	P46310	DnaJ-J8	Chloroplast	2	81
6	Q9LW57	GLU2	Chloroplast	1	72
6	Q9MBA1	None	Chloroplast	2	36
6	Q9XF88	None	Chloroplast	2	39
6	Q9SAK2	BCCP	Chloroplast	3	29
6	Q9SX22	None	Chloroplast	2	56
6	O49292	None	Chloroplast	5	52
6	O49196	None	Chloroplast	4	59
6	Q9M439	PORA	Mitochondrion	4	22
6	Q9S756	None	Chloroplast	1	52
6	Q9LR75	RbcS	Chloroplast	2	48
6	Q43316	RbcS	Chloroplast	1	86
6	Q9SI93	TOCC	Mitochondrion	5	10
6	Q9SZ52	None	Chloroplast	5	62
7	P37271	None	Chloroplast	1	70
7	Q9SEU8	None	Chloroplast	1	72
7	Q07473	None	Chloroplast	3	40
7	Q94IC9	TOCC	Mitochondrion	4	58
7	Q38802	BCCP	Chloroplast	1	60
7	Q9S7W1	None	Chloroplast	3	26
7	P42699	BCCP	Chloroplast	1	52
7	Q9SKT0	RbcS	Chloroplast	1	67
7	Q43307	BCCP	Chloroplast	2	62
7	Q42533	BCCP	Chloroplast	1	82
7	Q38933	None	Other	2	-
7	P49077	PORA	Chloroplast	2	68
7	Q66GR6	None	Chloroplast	3	75
7	Q8RWW6	None	Chloroplast	1	70
7	P49107	RbcS	Chloroplast	5	81
7	Q8LD49	None	Chloroplast	1	67
7	P27202	RbcS	Chloroplast	3	40
7	Q93VA3	None	Chloroplast	4	18
7	P11490	DnaJ-J8	Chloroplast	1	66
7	Q9SEU7	None	Chloroplast	1	67
7	Q9M591	Cab	Chloroplast	1	36
7	P42732	RbcS	Chloroplast	2	46

Table A2-2. (continued)

Hsp70 cluster group	SWISS- PROT ID	Lee et al. (2008) Group	TargetP prediction		
			Localization	Reliability class	TP length
7	Q9FMT1	GLU2	Other	5	-
7	Q43349	None	Chloroplast	1	65
7	P16127	RbcS	Chloroplast	1	60
8	Q84RQ7	None	Chloroplast	2	99
8	Q93WX6	DnaJ-J8	Chloroplast	2	35
8	Q9LUD9	None	Chloroplast	3	39
8	Q9LS02	None	Chloroplast	1	77
8	Q8GXU8	None	Chloroplast	4	56
8	Q9SJE1	None	Chloroplast	2	49
8	Q9S7H1	None	Chloroplast	2	44
8	Q9LFV0	BCCP	Chloroplast	1	45
8	Q9SA56	None	Chloroplast	1	43
8	Q38854	None	Chloroplast	1	58
8	Q9S7D1	None	Chloroplast	2	58
8	Q9CA35	DnaJ-J8	Chloroplast	1	71
8	Q8W105	Cab	Chloroplast	2	90
8	Q38885	RbcS	Chloroplast	1	67
8	O23403	None	Chloroplast	1	75
8	Q9SUI6	None	Chloroplast	4	44
8	O81439	DnaJ-J8	Chloroplast	2	55
9	Q39089	None	Chloroplast	2	22
9	O50039	RbcS	Chloroplast	1	53
9	Q9XFT3	None	Chloroplast	2	44
9	O22265	RbcS	Chloroplast	1	56
9	P42762	None	Chloroplast	1	89
9	Q9SCX9	None	Chloroplast	3	32
9	O82796	GLU2	Chloroplast	4	15
9	P32068	RbcS	Chloroplast	5	60
9	Q96255	PORA	Chloroplast	1	63
9	Q9CAP8	None	Chloroplast	5	30
9	P21240	None	Chloroplast	1	53
9	Q93W20	None	Chloroplast	1	16
9	Q9SJU4	None	Chloroplast	3	10
9	P16972	None	Chloroplast	1	52
9	O04090	PORA	Chloroplast	2	69

Appendix 3

The Position-specific Scoring Matrices

Table A3-1. The Components of the TP PSSM for Analysis of the N-terminal 10 Residues

Amino Acids	$f_{i,1}$	$f_{i,2}$ to $f_{i,9}$	p_i	$s_{i,1}$	$s_{i,2}$ to $s_{i,9}$
Ala	0.4693	0.1197	0.0863	2.4426	0.4716
Cys	0.0055	0.0192	0.0125	-1.1912	0.6193
Asp	0.0241	0.0053	0.0532	-1.1405	-3.3300
Glu	0.0811	0.0082	0.0620	0.3883	-2.9232
Phe	0.0143	0.0442	0.0403	-1.4980	0.1357
Gly	0.0515	0.0288	0.0709	-0.4604	-1.2975
His	0.0033	0.0106	0.0221	-2.7503	-1.0653
Ile	0.0230	0.0500	0.0601	-1.3839	-0.2652
Lys	0.0154	0.0183	0.0531	-1.7899	-1.5387
Leu	0.0296	0.1240	0.0992	-1.7452	0.3217
Met	0.0175	0.0279	0.0246	-0.4899	0.1786
Asn	0.0362	0.0274	0.0412	-0.1860	-0.5869
Pro	0.0164	0.0510	0.0469	-1.5108	0.1207
Gln	0.0219	0.0308	0.0396	-0.8511	-0.3625
Arg	0.0121	0.0178	0.0544	-2.1727	-1.6122
Ser	0.1075	0.2409	0.0666	0.6901	1.8546
Thr	0.0296	0.1014	0.0558	-0.9140	0.8627
Val	0.0373	0.0635	0.0678	-0.8629	-0.0954
Trp	0.0000	0.0029	0.0129	-10.2024	-2.1631
Tyr	0.0044	0.0082	0.0305	-2.8000	-1.9020

Table A3-2. The Components of the TP PSSM for Analysis of the N-terminal 30 Residues

Amino Acids	$s_{i,1}$	$s_{i,2}$ to $s_{i,11}$	$s_{i,12}$	$s_{i,13}$	$s_{i,14}$
Ala	2.4426	0.4716	-0.1664	0.1554	-0.5816
Cys	-1.1912	0.6193	0.9413	0.9412	-0.3808
Asp	-1.1405	-3.3300	-2.4675	-1.1456	-2.4676
Glu	0.3883	-2.9232	-3.6887	-2.6888	-10.3327
Phe	-1.4980	0.1357	1.0215	0.7413	0.3933
Gly	-0.4604	-1.2975	-1.8825	-0.2976	-0.7127
His	-2.7503	-1.0653	-0.2028	-1.2029	0.1190
Ile	-1.3839	-0.2652	0.3563	-0.0588	-0.0589
Lys	-1.7899	-1.5387	-0.6574	-1.1429	-0.1430
Leu	-1.7452	0.3217	0.1559	0.2762	0.1558
Met	-0.4899	0.1786	-0.0355	-0.7725	-1.3575
Asn	-0.1860	-0.5869	0.0720	0.2240	-0.7761
Pro	-1.5108	0.1207	0.9627	0.8846	1.4695
Gln	-0.8511	-0.3625	-0.7187	-3.0407	-1.4558
Arg	-2.1727	-1.6122	-0.9147	-0.3298	0.2006
Ser	0.6901	1.8546	1.6342	1.4558	1.3370
Thr	-0.9140	0.8627	0.2709	0.1639	0.9870
Val	-0.8629	-0.0954	-1.2329	-0.2330	-0.6481
Trp	-10.2024	-2.1631	-0.4261	-8.0700	-8.0701
Tyr	-2.8000	-1.9020	-0.6676	-1.6676	-0.3458

Table A3-2. (Continued)

Amino Acids	$s_{i, 15}$	$s_{i, 16}$	$s_{i, 17}$	$s_{i, 18}$	$s_{i, 19}$ to $s_{i, 29}$
Ala	-0.0791	-0.1665	-0.4661	-0.4661	-0.5935
Cys	0.2041	0.2042	-1.3807	0.6192	0.1047
Asp	-2.4676	-1.4675	-0.8826	-1.8826	-1.6195
Glu	-2.6888	-1.6888	-3.6888	-2.1039	-2.4257
Phe	0.5189	0.5189	0.1039	0.6344	0.3670
Gly	-0.7127	-0.1821	-0.5606	-0.4232	-0.7450
His	-1.2029	0.1191	-1.2029	0.1190	0.1757
Ile	-0.6439	-0.1843	0.1636	-1.6439	-0.8782
Lys	0.2356	-0.2949	-0.0054	0.4420	0.3008
Leu	0.2761	0.0917	0.1559	0.1558	-0.0822
Met	-9.0014	-9.0013	-9.0013	-9.0014	-2.0944
Asn	-0.0980	-1.5130	0.0720	0.2239	0.0396
Pro	1.1740	0.6216	0.9626	0.8021	0.7937
Gln	-1.4558	0.6598	-0.2333	-0.2334	-0.3402
Arg	0.8925	0.8222	0.4071	0.0851	0.5182
Ser	1.4557	1.1620	1.4173	1.5653	1.8632
Thr	0.2708	0.3704	0.3704	0.2708	-0.1719
Val	-1.4961	-0.4960	-0.6480	-0.2331	-0.5105
Trp	-8.0701	-1.4262	-0.4262	-1.4263	-2.7480
Tyr	-0.6677	-2.6676	-0.3457	-9.3116	-1.6676

Table A3-3. The Components of the mTP PSSM for Analysis of the N-terminal 30 Residues

Amino Acids	$s_{i,1}$	$s_{i,2}$	$s_{i,3}$	$s_{i,4}$	$s_{i,5}$
Ala	1.7558	1.2041	0.8216	0.9411	0.9916
Cys	-8.8438	0.1220	0.9700	1.1181	1.1220
Asp	-10.9306	-10.9306	-10.9306	-3.2906	-3.2867
Glu	-2.9230	-2.9230	-2.1860	-11.1557	-11.1518
Phe	1.4723	-0.0779	0.0217	-0.4297	-1.3002
Gly	-1.8944	-1.2423	-2.3798	-0.7988	-0.6143
His	-2.0221	-1.4371	-9.6659	-2.0260	-2.0220
Ile	-0.2931	-0.7625	-0.4630	-0.7664	-0.6556
Lys	-1.6991	-0.5836	-0.4767	-0.5875	0.1755
Leu	1.2394	0.2054	0.5680	0.5371	0.6460
Met	-1.1767	-0.5917	0.1452	-0.0106	-0.3693
Asn	-1.3322	-0.9172	-1.1098	-1.5991	-1.9171
Pro	-0.9346	0.1434	0.1434	-0.5235	-1.2971
Gln	-0.6900	-2.2749	-0.4004	-1.0564	-0.4004
Arg	-0.5116	1.6118	1.6583	1.5351	0.9665
Ser	0.1960	1.3659	0.2955	0.4722	0.7106
Thr	-2.0338	-0.6553	-0.3557	0.7279	0.1038
Val	-1.4673	-0.8298	0.1177	0.0028	0.4073
Trp	-0.2454	-0.6604	-0.2454	-1.2493	0.7546
Tyr	-0.3169	-10.1307	-0.9019	-2.4907	-1.9018

Table A3-3. (Continued)

Amino Acids	$s_{i,6}$	$s_{i,7}$	$s_{i,8}$	$s_{i,9}$	$s_{i,10}$
Ala	0.6031	0.5023	0.9212	0.3718	0.7422
Cys	0.9739	0.1181	-0.6150	-0.2000	0.6074
Asp	-10.9267	-2.7056	-10.9306	-10.9306	-2.7018
Glu	-11.1479	-3.5118	-2.9230	-3.5080	-11.1518
Phe	-0.1809	-0.3041	0.0217	-0.1848	-0.8852
Gly	-0.4499	-0.7056	-0.0579	0.0531	-0.4538
His	-0.2108	-0.7040	-1.4371	-0.7001	-1.4371
Ile	-0.3716	-1.4668	-1.2930	-1.2931	-1.4630
Lys	-0.3732	0.2357	-0.3771	-0.4767	0.0379
Leu	0.9045	0.4531	0.5411	0.9424	0.7200
Met	-1.5878	-1.5956	-1.1766	-0.5917	-0.3693
Asn	-0.7433	-2.3360	-0.0102	-1.1098	-0.5952
Pro	-0.4002	-0.7865	-0.4041	-0.1976	-0.5196
Gln	-0.1555	-0.0564	-0.0525	-0.8599	-1.0525
Arg	1.1112	1.8269	1.3249	1.5390	1.3815
Ser	0.6404	0.6702	0.8149	0.7105	1.0325
Thr	0.1718	0.0328	-0.4488	0.5023	0.0366
Val	0.0672	0.1663	-0.1136	-1.1777	-0.3153
Trp	-8.8854	0.3357	-1.2454	0.7546	0.5620
Tyr	-2.4829	-1.4907	-2.4868	-10.1307	-10.1307

Table A3-3. (Continued)

Amino Acids	$s_{i, 11}$	$s_{i, 12}$	$s_{i, 13}$	$s_{i, 14}$	$s_{i, 15}$
Ala	0.8433	1.0143	0.7108	0.1804	0.7422
Cys	0.7962	-1.2000	0.6035	0.1181	0.8001
Asp	-1.7057	-10.9306	-10.9345	-3.2906	-3.2867
Glu	-2.9269	-2.9230	-2.9269	-2.5118	-2.9230
Phe	-0.3041	-0.3003	0.1109	-0.5671	-0.0778
Gly	-0.7988	0.1056	0.1523	0.1017	0.2052
His	-1.0260	-0.4371	-1.0260	-2.0259	0.1479
Ile	-1.0074	-1.2931	-0.6595	-0.2969	-0.8780
Lys	0.0340	-0.0361	-0.5875	-1.1179	-0.6990
Leu	0.4239	0.5410	0.3942	0.2016	0.2394
Met	-0.3732	-9.8205	-2.1806	-0.0106	-1.5917
Asn	-0.1137	-1.1098	-1.3361	-1.5991	-0.4577
Pro	0.0615	-0.1045	0.2839	0.2839	-0.2971
Gln	0.1362	-0.0525	-0.4043	-0.2788	0.1402
Arg	1.3496	0.8905	1.3210	1.5101	1.2660
Ser	0.7067	0.7462	0.7771	1.1661	0.9432
Thr	-0.5522	-0.1858	-0.2721	0.2843	0.1679
Val	-0.4711	0.2208	0.2169	-0.5536	-0.5497
Trp	-8.8931	-0.2454	-1.2493	0.9207	-1.2454
Tyr	-1.1688	-1.1649	-2.4907	-2.4907	-10.1307

Table A3-3. (Continued)

Amino Acids	$s_{i, 16}$	$s_{i, 17}$	$s_{i, 18}$	$s_{i, 19}$	$s_{i, 20}$
Ala	0.4406	0.5023	0.7109	0.1397	0.5954
Cys	1.2595	0.3812	0.3812	0.1181	0.9661
Asp	-10.9306	-2.7056	-2.7056	-1.9687	-2.7056
Glu	-1.5079	-2.1899	-3.5118	-2.1899	-1.1899
Phe	-0.4257	-0.3041	-0.3041	-0.3041	-0.0817
Gly	-0.1168	0.0493	-0.1207	-0.3133	-0.1821
His	0.6784	0.2960	-0.0259	-0.2186	-0.7040
Ile	-0.8780	-1.4668	-0.8819	-0.6595	-1.4668
Lys	-0.5835	-0.5874	0.0341	0.1045	-0.4805
Leu	0.1351	0.6166	0.1312	0.0187	0.3328
Met	-2.1766	-0.5956	-2.1805	-0.8586	0.1414
Asn	-1.5952	-1.3360	-0.5991	-0.3360	-1.3360
Pro	-0.1976	-0.3010	0.0615	0.7985	0.0615
Gln	-0.1594	-0.1633	-0.0564	0.2237	0.3062
Arg	0.7255	1.2318	1.1690	1.3210	0.8866
Ser	1.2216	0.6702	0.8442	0.4722	0.8110
Thr	0.6443	0.4478	0.4478	0.2254	0.2254
Val	-0.1136	-0.7341	-0.3931	0.0594	-0.1175
Trp	0.3396	0.3357	-0.2492	-0.6643	0.0727
Tyr	-0.9018	-0.3208	-1.1688	-1.1688	-1.4907

Table A3-3. (Continued)

Amino Acids	$s_{i, 21}$	$s_{i, 22}$	$s_{i, 23}$	$s_{i, 24}$	$s_{i, 25}$
Ala	0.4406	0.6829	0.4699	0.4699	0.6829
Cys	-0.1999	0.1181	0.7962	0.7962	1.1181
Asp	-10.9306	-2.2906	-2.2906	-1.9687	-2.2906
Glu	-2.5079	-2.9268	-2.5118	-2.1899	-2.9268
Phe	0.0217	-1.5671	-0.0817	-0.1886	-0.8891
Gly	-0.3798	-0.5357	0.0493	-0.1821	-0.0618
His	-0.7001	-0.7040	0.6745	-0.2186	0.6745
Ile	-0.7625	-0.7664	-0.6595	-0.4668	-2.4668
Lys	-0.5835	-0.0399	-0.3810	-0.1179	-0.1179
Leu	-0.3795	-0.3328	0.1312	-0.2839	0.2016
Met	-1.5917	-0.3732	-1.1805	-1.1805	-0.5956
Asn	-0.4577	-0.7511	0.3269	0.0790	-0.4616
Pro	0.8024	0.4152	0.5355	0.8458	0.0615
Gln	0.0471	0.4582	-0.2788	0.0432	0.0432
Arg	1.1073	1.3776	0.9991	0.8065	0.9251
Ser	1.0034	1.0572	1.0286	0.8766	1.1397
Thr	0.5023	0.2254	-0.4527	0.1640	0.0328
Val	0.2208	-0.6410	-0.8337	-0.1816	-0.8337
Trp	-0.2454	0.3357	-1.2492	-0.2492	-0.2492
Tyr	-0.3169	-0.4907	-1.4907	-1.4907	-0.3208

Table A3-3. (Continued)

Amino Acids	$s_{i, 26}$	$s_{i, 27}$	$s_{i, 28}$	$s_{i, 29}$
Ala	0.6582	0.5023	0.5650	0.6543
Cys	0.8001	0.3812	0.1181	0.6036
Asp	-1.9648	-1.2906	-1.4832	-0.9687
Glu	-1.7006	-0.9268	-1.1899	-1.9268
Phe	-0.1847	0.1109	-0.1886	-0.0817
Gly	-0.2423	-0.2462	-0.5357	-0.3133
His	-0.0220	-0.0259	0.8810	-0.4410
Ile	-0.4629	-1.1449	-1.6595	-0.8819
Lys	-0.0360	-0.2879	-0.1179	-0.3810
Leu	-0.3289	0.0187	-0.6672	-0.1908
Met	-9.8205	-1.1805	-0.5956	0.1414
Asn	-0.2167	-0.7511	-0.9210	-1.3360
Pro	0.4191	0.6990	0.4152	0.3511
Gln	-0.8598	-0.1633	-0.4043	0.3842
Arg	0.8103	0.6772	0.7646	-0.0009
Ser	0.9736	0.7423	0.9995	1.0572
Thr	0.3992	0.1640	0.6848	0.2254
Val	0.1177	0.0594	0.2659	0.1139
Trp	0.0766	-0.2492	-0.6643	0.0727
Tyr	-1.4868	-0.4907	-0.4907	-0.0312

Table A3-4. The Components of the SP PSSM for Analysis of the N-terminal 30 Residues

Amino Acids	$s_{i,1}$	$s_{i,2}$	$s_{i,3}$	$s_{i,4}$	$s_{i,5}$
Ala	2.1596	-0.8318	-0.4500	0.0021	-0.1398
Cys	-8.3850	0.2535	0.5702	0.2643	0.8386
Asp	-1.2429	-10.4771	-2.8385	-10.4663	-2.8332
Glu	0.0384	-10.6984	-2.4747	-10.6876	-3.0544
Phe	-2.4264	0.8162	0.0225	0.4859	0.7383
Gly	0.5644	-1.2483	-2.6686	-0.9156	-1.6632
His	-1.5632	-0.2467	0.2335	0.7641	0.0165
Ile	-2.4192	-0.4246	-0.4298	0.1712	0.0780
Lys	0.7598	1.2138	1.0222	1.1803	0.2570
Leu	-1.4061	-0.4855	0.6876	0.1843	0.4761
Met	-0.7178	0.7362	-0.1435	-0.3905	-0.4012
Asn	-0.4583	0.5363	-1.1470	0.3545	-0.4636
Pro	-1.6457	-1.6511	-0.6563	-0.0553	-1.0660
Gln	-1.4010	-0.5991	-0.6043	-0.2257	-0.8213
Arg	-0.1597	1.0414	0.5887	0.2328	-0.5435
Ser	-0.4521	1.1996	1.2631	1.0623	1.3340
Thr	-0.7269	0.6213	0.3405	-0.0841	-0.3172
Val	-0.0909	0.0642	-0.1015	-0.4725	1.1383
Trp	-8.4304	-8.4358	-0.2122	-8.4250	-8.4357
Tyr	-9.6718	-1.0334	-0.2312	-1.4376	-1.0333

Table A3-4. (Continued)

Amino Acids	$s_{i,6}$	$s_{i,7}$	$s_{i,8}$	$s_{i,9}$	$s_{i,10}$
Ala	0.1171	0.7211	0.6827	0.7747	0.7264
Cys	-0.7410	-0.7410	1.0663	1.5970	0.5862
Asp	-10.4717	-10.4717	-2.8278	-10.4555	-2.8225
Glu	-10.6929	-2.0491	-10.6929	-10.6768	-3.0437
Phe	0.8216	0.9660	0.9660	1.1134	1.7489
Gly	-1.6579	-1.4355	-1.2429	-1.6418	-3.2376
His	-1.5632	-1.5632	-0.9782	-9.1909	-9.2017
Ile	0.5195	0.4554	0.0834	-0.0810	-0.2983
Lys	-0.3657	-1.2401	-0.2401	-2.2240	-2.2348
Leu	1.1300	1.1300	1.5008	1.5361	1.6523
Met	-0.1328	0.4522	-0.1328	-0.1167	0.0949
Asn	0.1267	-1.8733	-1.4583	-1.4421	-2.4529
Pro	-0.8383	-1.0607	-1.3237	-1.3076	-1.3184
Gln	-1.4010	-1.8160	-10.0448	-10.0287	-10.0395
Arg	-0.8601	-0.6901	-2.2751	-2.8439	-1.8548
Ser	0.7055	0.7544	0.6024	0.6710	0.2452
Thr	0.4955	0.0101	-0.5749	-0.4212	-0.4321
Val	0.4067	0.5766	0.5222	0.8823	0.8270
Trp	0.7985	-8.4303	-0.7865	-8.4142	-8.4250
Tyr	-1.0279	-0.7060	-2.0279	-1.4268	-9.6665

Table A3-4. (Continued)

Amino Acids	$s_{i, 11}$	$s_{i, 12}$	$s_{i, 13}$	$s_{i, 14}$	$s_{i, 15}$
Ala	0.5122	1.0277	0.9965	1.4200	0.8003
Cys	1.8385	0.5808	1.0661	0.8600	1.0715
Asp	-10.4771	-10.4718	-10.4719	-10.4556	-2.2377
Glu	-2.0544	-10.6930	-10.6931	-10.6768	-2.4589
Phe	1.2687	1.0971	1.3808	1.2902	1.3338
Gly	-2.2483	-0.1555	-2.6581	-2.2268	-1.4303
His	-1.5686	-9.2071	-9.2072	-9.1909	-9.2018
Ile	0.3828	-0.3037	0.8031	-0.0811	0.6450
Lys	-2.8305	-10.4691	-10.4692	-10.4529	-10.4638
Leu	1.5885	1.3594	1.4215	1.2654	0.9776
Met	-0.7232	0.0895	0.7414	-0.3798	1.1943
Asn	-1.8787	-1.4584	-10.1023	-10.0860	-10.0969
Pro	-0.8437	-1.3238	-2.6458	-2.6295	-1.3185
Gln	-1.4064	-2.4011	-10.0450	-1.3849	-10.0396
Arg	-2.2805	-1.8602	-2.8603	-10.4878	-2.2699
Ser	-0.0704	-0.3452	0.0172	-0.0489	0.3122
Thr	-0.7323	-0.5750	-1.0897	-0.0734	-0.0842
Val	0.7232	1.1792	1.0694	1.0857	1.2532
Trp	-8.4357	0.5353	-0.7867	-0.1854	-8.4251
Tyr	-0.2260	-2.0280	-1.0281	-0.4269	-9.6666

Table A3-4. (Continued)

Amino Acids	$s_{i, 16}$	$s_{i, 17}$	$s_{i, 18}$	$s_{i, 19}$	$s_{i, 20}$
Ala	0.9649	0.8253	0.8254	1.2281	0.9914
Cys	1.4288	1.4234	1.5755	0.5809	1.7130
Asp	-10.4718	-2.8333	-2.2482	-2.2429	-1.0259
Glu	-2.4642	-1.4695	-2.4694	-1.7271	-0.2471
Phe	1.0330	0.9606	0.6558	0.7436	0.1533
Gly	-1.0730	-0.1608	-0.6632	-0.1554	-0.2483
His	-9.2071	-9.2125	-0.2466	0.0218	0.0164
Ile	0.5808	-0.2022	0.3125	-0.6821	-0.2021
Lys	-2.2403	-2.8306	-1.2455	-2.8251	-10.4744
Leu	1.0792	0.9104	0.8814	0.2720	0.2208
Met	0.0895	0.4467	0.0842	-1.1328	-0.1382
Asn	-10.1022	-1.1418	-1.1417	-0.1363	-0.8787
Pro	-2.0608	-1.6511	-0.6510	-0.3237	-0.8437
Gln	-10.0449	-10.0503	-0.5990	-1.8160	-0.5990
Arg	-2.8602	-10.5094	-2.8654	-10.5039	-1.8655
Ser	0.6023	0.8421	0.6495	1.0570	0.7490
Thr	0.1906	0.2677	0.1853	0.2731	0.4901
Val	0.9915	0.7705	0.4602	0.8217	0.5712
Trp	-0.7866	-0.2070	-8.4357	-8.4303	-8.4357
Tyr	-2.0280	-1.0334	-1.4483	-2.0279	-1.4484

Table A3-4. (Continued)

Amino Acids	$s_{i, 21}$	$s_{i, 22}$	$s_{i, 23}$	$s_{i, 24}$	$s_{i, 25}$
Ala	0.8888	1.1958	0.8254	0.8254	0.9219
Cys	1.7077	0.8386	0.2536	0.5755	1.7077
Asp	-0.5166	-1.2482	-0.1327	-0.5112	-0.5166
Glu	-0.7378	-1.0544	-0.7325	-0.8845	-0.0598
Phe	-0.4370	-0.2617	-0.4316	0.4752	0.5630
Gly	-0.5531	0.5592	-0.5478	-0.2482	-0.1661
His	-0.9889	0.4315	-0.2466	0.2389	-0.2519
Ile	-0.4298	-0.2021	-0.1025	-0.6875	-0.2074
Lys	-1.5138	-1.2455	-1.0231	-1.0231	-0.2509
Leu	0.3057	-0.3410	0.0740	-0.5634	-0.6512
Met	0.0789	-0.4012	0.2769	-1.1381	-1.7285
Asn	-0.6616	-0.6562	-0.4636	-0.0042	-0.1470
Pro	0.0442	-1.0660	0.1564	0.5970	-0.3344
Gln	-0.4116	0.1787	0.7636	0.9156	-0.2418
Arg	-1.2858	-1.2804	-0.2804	-0.6955	-0.5488
Ser	1.0463	0.9297	0.8866	0.7964	0.4218
Thr	-0.0006	0.0048	-0.9021	0.0048	0.3404
Val	0.3960	0.2758	-0.0137	-0.4832	0.1329
Trp	-0.7971	-8.4357	-8.4357	-8.4357	-8.4411
Tyr	-2.0386	-2.0333	-0.7113	0.1367	-1.4537

Table A3-4. (Continued)

Amino Acids	$s_{i, 26}$	$s_{i, 27}$	$s_{i, 28}$	$s_{i, 29}$
Ala	0.3747	-0.2155	-0.2155	-0.0086
Cys	0.5755	1.0557	2.2483	1.8385
Asp	0.5591	-0.3791	0.8054	0.6262
Glu	-0.3540	-0.3593	-0.8898	-0.7325
Phe	-0.4317	0.3704	0.1480	-0.6243
Gly	0.3956	0.2700	0.1388	0.5591
His	0.0164	-1.5738	0.2335	0.2388
Ile	-0.1026	-0.2074	-1.2074	0.5755
Lys	-1.0231	-0.6658	-0.6658	-1.0231
Leu	-0.4855	-0.4167	-0.3463	-0.4855
Met	-9.3670	-1.7284	-1.1435	-9.3670
Asn	-0.2937	0.6186	0.9234	0.4432
Pro	0.1563	0.2506	0.2506	0.2559
Gln	0.4010	0.5884	0.9807	0.5005
Arg	-2.2805	-1.5488	-1.0634	-0.4060
Ser	0.9714	1.0068	-0.1631	-0.3505
Thr	-0.0948	-0.3225	-0.2070	0.3457
Val	0.4013	-0.1889	-0.3816	-0.7242
Trp	-8.4357	0.5248	-0.2122	-8.4357
Tyr	-0.4484	0.5464	0.1313	0.1366

Appendix 4

Perl Script Codes

Code A4-1. Percentage of Uncharged Amino Acids Calculator

```
#!/usr/bin/perl
#####
# Created by Prakitchai Chotewutmontri, 6 Apr 2010
# updated on 15 Aug 2010
#
# GST student
#
# OBJECTIVE
#   We try to identify the N-ter uncharged region of Transit peptide
#   This script will calculate #Uncharged & %uncharge within
#   a specific window length (w=5,...,20) along the whole lenght of
#   transit peptide sequence.
#
# Input: multiple fasta amino acid sequences
#   >header 1
#   seq1.....
#   >header 2
#   seq2.....
#
# Output: separate by tab
#   <w_len><header><window-start-residue><#UC><#C><%UC><%C>\n
#
#####
#Subroutine prototypes
sub CHARGEcal ();

#global vars
my $i = 0;
my @temp = ();
my $num = 0;
my $total = 0;
my $line = "";
my $seq = "";
my $linelimit = 60;
my $filenum = 1;
my $wsize = 0;
my $minw = 5; #can change
my $maxw = 17; #can change
my $wi = 0;
my $s1 = "";
my $s2 = "";
my $s3 = "";
my @ans1 = ();
my @ans2 = ();
my @ans3 = ();
my @percent = ();
my $header = "";
my $resi = 0;
my $start = 0;
my %sum_percent = ();
my %sum_percentSQ = ();
my %num_count = ();
my @max_start = ();
my $cal_sum = 0;
my $cal_sd = 0;

#usage
my $USAGE = "usage: $0 <input fasta filename> <out data filename> <out summary filename>\n\n";

#check argument
unless (@ARGV == 3) {
    print $USAGE;
    exit -1;
}

#store argv
my $infile = $ARGV[0];
my $outfile = $ARGV[1];
my $outsumfname = $ARGV[2];

#open infile
unless ( open(INFILE,"<$infile") ) {
```

```

        print "Can't open $infile\n\n";
        exit -1;
    }

    #open outfile, first time, overwrite it
    unless ( open(OUTFILE,">$outfname") ) {
        print "Can't create $outfname\n\n";
    }

    #open outfile, first time, overwrite it
    unless ( open(OUTSUM,">$outsumfname") ) {
        print "Can't create $outsumfname\n\n";
    }

    #write the outfile header line
    # <w_len><header><window-start-residue><#UC><#C><%UC><%C><subseq>
    print OUTFILE "window size\tseq ID\twindow-start-residue\t#Uncharged\t#Charged\t";
    print OUTFILE "%Uncharged\t#Charged\tsubsequence\n\n";

    #main part
    $total = 0; #count # sequence
    while ($line = <INFILE> ) { #read input file line by line

        chomp ($line); #remove end-line char

        if ($line =~ /^>/ ) {#found the > at the first char, mean the header line
            $num++; #count header
            if ($num != 1) { #not the first header, calculate & write out previous seq result
                #sequence of the previous header was collected before this point

                #now analyse the seq and print out

                for ($wi=$minw; $wi<=$maxw; $wi++) { #run the function below for every window size
                    $resi = 0; #reset the start residue to zero
                    # run the calculation from of window length = wi
                    # along the whole lenght of the seq
                    for ($resi=0; $resi<=length($seq)-$wi; $resi++) {

                        #(1) generate substring
                        $s1 = substr($seq,$resi,$wi);
                        print "$s1\n";

                        #(2) calculate #uncharged, #charged
                        # ans[0] = #Uncharged amino acid
                        # ans[1] = #Charged amino acid

                        @ans1 = &CHARGEcal ($s1);
                        print "$ans1[0]\t$ans1[1]\n";

                        #(3) calcualte %uncharged, %charged
                        # percent[0] = %Uncharged amino acid
                        # percent[1] = %Charged amino acid

                        $percent[0] = ($ans1[0]*100)/($ans1[0]+$ans1[1]);
                        $percent[1] = ($ans1[1]*100)/($ans1[0]+$ans1[1]);
                        #print "percentUC=$percent[0]\n";

                        #(4) write to output file
                        #<w_len><header><window-start-residue><#UC><#C><%UC><%C><subseq>
                        $start = $resi + 1;
                        print OUTFILE "$wi\t$header\t$start\t";
                        printf OUTFILE '%.4f', $ans1[0];
                        print OUTFILE "\t";
                        printf OUTFILE '%.4f', $ans1[1];
                        print OUTFILE "\t";
                        printf OUTFILE '%.4f', $percent[0]; # %UC
                        print OUTFILE "\t";
                        printf OUTFILE '%.4f', $percent[1];
                        print OUTFILE "\t$s1";
                        print OUTFILE "\n"; #end of line

                        #(5) keep number in hash table for summary
                        # hash table keys = window size & start residue
                        #sum %UC
                        if ( exists $sum_percent{ $wi, $resi } ) {
                            #print "exists = $sum_percent{$wi,$resi}\n";
                            $sum_percent{ $wi, $resi } = $sum_percent{ $wi, $resi } +
                                $sum_percentSQ{ $wi, $resi } + ($percent[0]**2);
                            #print "new value = $sum_percent{$wi,$resi}\n";
                        } else {
                            $sum_percent{ $wi, $resi } = $percent[0];
                            $sum_percentSQ{ $wi, $resi } = ($percent[0]**2);
                            #print "initiate = $percent[0]\n";
                        }
                        #sum number of each %UC based on wi & resi
                        if ( exists $num_count{ $wi, $resi } ) {
                            #print "count exists = $num_count{ $wi, $resi }\n";
                            $num_count{ $wi, $resi } = $num_count{ $wi, $resi } + 1;
                        }
                    }
                }
            }
        }
    }

    $percent[0];

```

```

        #print "count new value = $num_count{ $wi, $resi }\n";
    } else {
        $num_count{ $wi, $resi } = 1;
    }
    #keep track of max value of start residue
    if ( exists $max_start[$wi] ) {
        if ( $resi > $max_start[$wi] ) { $max_start[$wi] = $resi; }
    } else {
        $max_start[$wi] = $resi;
    }
}

    } # for loop running window for the whole length of seq
} # for loop every window size

$seq = ""; #clear seq after used
}
#keep header
$header = $line;
}
else { #otherline remove spaces, concatenate sequence
    $line =~ s/\s//g ;
    $seq .= $line; #concatenate
} #end if header line

} #end while line = infile

#analyse the last seq after while loop
for ($wi=$minw; $wi<=$maxw; $wi++) { #run the function below for every window size
    $resi = 0; #reset the start residue to zero
    # run the calculation from of window length = wi
    # along the whole length of the seq
    for ($resi=0; $resi<=length($seq)-$wi; $resi++) {

        # (1) generate substring
        $s1 = substr($seq,$resi,$wi);
        print "$s1\n";

        # (2) calculate #uncharged, #charged
        # ans[0] = #Uncharged amino acid
        # ans[1] = #Charged amino acid

        @ans1 = &CHARGEcal ($s1);
        print "$ans1[0]\t\t$ans1[1]\n";

        # (3) calculate %uncharged, %charged
        # percent[0] = %Uncharged amino acid
        # percent[1] = %Charged amino acid

        $percent[0] = ($ans1[0]*100)/($ans1[0]+$ans1[1]);
        $percent[1] = ($ans1[1]*100)/($ans1[0]+$ans1[1]);

        # (4) write to output file
        #<w_len><header><window-start-residue><#UC><#C><%UC><%C><subseq>
        $start = $resi + 1;
        print OUTFILE "$wi\t$header\t\t$start\t";
        printf OUTFILE "%.4f", $ans1[0];
        print OUTFILE "\t";
        printf OUTFILE "%.4f", $ans1[1];
        print OUTFILE "\t";
        printf OUTFILE "%.4f", $percent[0]; #%UC
        print OUTFILE "\t";
        printf OUTFILE "%.4f", $percent[1];
        print OUTFILE "\t\t$s1";
        print OUTFILE "\n"; #end of line

        # (5) keep number in hash table for summary
        # hash table keys = window size & start residue
        #sum %UC
        if ( exists $sum_percent{ $wi, $resi } ) {
            $sum_percent{ $wi, $resi } = $sum_percent{ $wi, $resi } + $percent[0];
            $sum_percentSQ{ $wi, $resi } += ($percent[0]**2);
        } else {
            $sum_percent{ $wi, $resi } = $percent[0];
            $sum_percentSQ{ $wi, $resi } = ($percent[0]**2);
        }
        #sum number of each %UC based on wi & resi
        if ( exists $num_count{ $wi, $resi } ) {
            $num_count{ $wi, $resi } = $num_count{ $wi, $resi } + 1;
        } else {
            $num_count{ $wi, $resi } = 1;
        }
        #keep track of max value of start residue
        if ( exists $max_start[$wi] ) {
            if ( $resi > $max_start[$wi] ) { $max_start[$wi] = $resi; }
        } else {
            $max_start[$wi] = $resi;
        }
    }
}

```

```

    } # for loop running window for the whole length of seq
} # for loop every window size

#write the out summary file
# <w_len><%UC_res=1><%UC_res=2>...

#write the out summary file header line
# <w_len><%UC_res=1><%UC_res=2>...
print OUTSUM "%Uncharged\n";
print OUTSUM "window size\tstart residue=1..end\n";

for ($wi=$minw; $wi<=$maxw; $wi++) {
    print OUTSUM "$wi";

    for ($resi=0; $resi <= $max_start[$wi]; $resi++) {
        #print "cal sum = $sum_percent{$wi, $resi } devide by $num_count{ $wi, $resi }\n";
        $cal_sum = $sum_percent{ $wi, $resi } / $num_count{ $wi, $resi};
        print OUTSUM "\t";
        printf OUTSUM '%.4f', $cal_sum;
    }
    print OUTSUM "\n"; #end of line
}

#write the out summary file header line
# <w_len><%UC_res=1><%UC_res=2>...
print OUTSUM "SD of %Uncharged\n";
print OUTSUM "window size\tstart residue=1..end\n";

for ($wi=$minw; $wi<=$maxw; $wi++) {
    print OUTSUM "$wi";

    for ($resi=0; $resi <= $max_start[$wi]; $resi++) {
        #SD = sqrt( (1/(num-1))*(sum_val^2) - mean^2 )
        $cal_sum = $sum_percent{ $wi, $resi } / $num_count{ $wi, $resi};
        $cal_sd=sqrt(( $sum_percentSQ{$wi,$resi}/($num_count{$wi,$resi} -1))-($cal_sum**2));
        print OUTSUM "\t";
        printf OUTSUM '%.4f', $cal_sd;
    }
    print OUTSUM "\n"; #end of line
}

print OUTSUM "Num of %Uncharged\n";
print OUTSUM "window size\tstart residue=1..end\n";

for ($wi=$minw; $wi<=$maxw; $wi++) {
    print OUTSUM "$wi";

    for ($resi=0; $resi <= $max_start[$wi]; $resi++) {
        print OUTSUM "\t";
        printf OUTSUM '%.0f', $num_count{$wi,$resi};
    }
    print OUTSUM "\n"; #end of line
}

close OUTFILE;
close OUTSUM;

close INFILE;

exit 0;

#####
#Subroutine programs
#####
#
sub CHARGEcal () {
# Calculate the uncharge-to-charge ratio
my $inseq = $_[0]; #input sequence
my $num_UC = 0; #number of uncharged residues
my $num_C = 0; #number of charged residues = seq_lenght - #UC
my $ratio = 0; #num_UC/C ratio

$num_C = ($inseq =~ tr/K|R|D|E|H//);
$num_UC = length($inseq) - $num_C;
if ($num_C == 0) {
    $ratio = ($num_UC+1)/($num_C+1); #correct for zero denominator
} else {
    $ratio = $num_UC/$num_C;
}

return ($num_UC, $num_C, $ratio);
}

```

Code A4-2. Hsp70 Binding Site Prediction based on Random Peptide-display Phage Library Derived Algorithm

```
#!/usr/bin/perl
#####
# RPPDscoring.pl
# Created by Prakitchai Chotewutmontri, 1 Feb 2010
# GST student
#
# OBJECTIVE
#   Calculate HSP70 binding score based on
#   Ivey et al (2000). Plant Physiol. 122(4):1289-99.
#   Using index derived from RPPD (Random peptide phage display)
#   published in Gragerov et al. (1994) J Mol Biol 235:848.
#   Scoring using 6 windows
#   Score(i) = I(aa(i-2))*I(aa(i-1))*I(aa(i))*I(aa(i+1))*I(aa(i+2))*I(aa(i+3))
#   Score at residue i is the multiplication of indices (I) corresponded to
#   amino acid at position i-2 to i+3.
#
# Input: multiple fasta amino acid sequences
#   >header 1
#   seq1.....
#   >header 2
#   seq2.....
#
# Output: score from residue 3 to (length-2) separate by tab
#   >header 1
#   score(3)<tab>score(4)<tab>...score(length-2)<end-line>
#   >header 2
#   score(3)<tab>score(4)<tab>...score(length-2)<end-line>
#
#####

use strict;
use warnings;

#Subroutine prototypes
sub RPPDscore();

#global vars
# RPPD - Random peptide phage display Indices
# Index values are derived from graph (Fig. 1A, manually measured by ruler)
my %RPPD = ( "A", "0.876", #Alanine
             "C", "1.000", #Cysteine
             "D", "0.871", #Aspartate
             "E", "1.000", #Glutamate
             "F", "0.506", #Phenylalanine
             "G", "0.567", #Glycine
             "H", "0.567", #Histidine
             "I", "1.772", #Isoleucine
             "K", "2.025", #Lysine
             "L", "2.015", #Leucine
             "M", "1.000", #Methionine
             "N", "0.754", #Asparagine
             "P", "0.785", #Proline
             "Q", "0.547", #Glutamine
             "R", "1.489", #Arginine
             "S", "1.362", #Serine
             "T", "0.597", #Threonine
             "V", "0.800", #Valine
             "W", "1.782", #Tryptophan
             "Y", "0.759" ); #Tyrosine

my $i = 0;
my @temp = ();
my $num = 0;
my $line = "";
my $seq = "";
my $linelimit = 60;
my $filenum = 1;
my $windowsize = 6;

#usage
my $USAGE = "usage: $0 <input fasta filename> <output filename>\n\n";

#check argument
unless (@ARGV == 2) {
    print $USAGE;
    exit -1;
}

#store argv
my $inname = $ARGV[0];
my $outname = $ARGV[1];

#open infile
unless ( open(INFILE,"<$inname") ) {
```

```

        print "Can't open $infile\n\n";
        exit -1;
    }

    #open outfile
    unless ( open(OUTFILE,">$outfile") ) {
        print "Can't create $outfile\n\n";
    }

    #read infile and write outfile
    while ( $line = <INFILE> ) {
        chomp ($line);
        if ( $line =~ /^>/ ) { #this line is a header line
            $num++; #count header
            if ($num != 1) { #not the first header, cal & write previous seq result
                my $seqlength = length($seq);
                # position used in Ivey paper -2 -1 0 +1 +2 +3
                # RPPD score of this string is give to pos 0 residue
                for ($i=2; $i < $seqlength-3; $i++) { #calculate from index 2..ln-4
                    print "pos ".($i-2)." to ".($i-2+$window-size-1)." \n";
                    my $subseq = substr($seq,$i-2,$window-size);
                    print "seq = $subseq\n";
                    my $score = &RPPDscore($subseq);
                    print "subseq : $score\n";
                    printf OUTFILE '%.4f', $score;
                    if ($i != $seqlength-4) {
                        print OUTFILE "\t";
                    } else {
                        print OUTFILE "\n"; }
                }
                $seq = ""; #clear seq
            }
            #write header
            print "\n\n$line\n";
            print OUTFILE $line." \t"; #same line header
            #print OUTFILE $line." \n"; #separate line header
        }
        else { #otherline remove spaces, concate seq
            $line =~ s/\s//g ;
            $seq .= $line;
        }
    }

    #last seq
    my $seqlength = length($seq);
    for ($i=2; $i < $seqlength-3; $i++) { #calculate from index 2..ln-4
        print "pos ".($i-2)." to ".($i-2+$window-size-1)." \n";
        my $subseq = substr($seq,$i-2,$window-size);
        print "seq = $subseq\n";
        my $score = &RPPDscore($subseq);
        print "subseq : $score\n";
        printf OUTFILE '%.4f', $score;
        if ($i != $seqlength-4) {
            print OUTFILE "\t";
        } else {
            print OUTFILE "\n"; }
    }

    close OUTFILE;

    print "Total = $num sequences\n\n";

    close INFILE;

    exit 0;

#####
#Subroutine programs
#####
#
sub RPPDscore () {
    # Calculate RPPDscore for the input string
    my $sixaa = $_[0];
    my $score = 1; #non bias, will be used in multiplication
    my $pos = 0;
    if (length($sixaa) != 6) { print "input for RPPDscore() is not 6 aa.\n"; }
    for ($pos = 0; $pos < length($sixaa); $pos++) {
        $score = $score*&RPPD(substr($sixaa,$pos,1));
    }
    return $score;
}
#
#####

```

Code A4-3. Hsp70 Binding Site Prediction based on Cellulose-bound Peptide Library Derived Algorithm

```
#!/usr/bin/perl
#####
# CBPSScoring.pl
# Created by Prakitchai Chotewutmontri, 1 Feb 2010
# GST student
#
# OBJECTIVE
# Calculate HSP70 binding score based on
# delta delta G derived from CBPS (Cellulose-bound peptide library)
# published in Rudiger et al. (1997) EMBO J 16(7):1501-07.
# Scoring using 13-aa windows
# Score(n) = (0.33*Ln-6) + (0.66*Ln-5) + (1.00*Ln-4) + (1.50*Ln-3) +
#           Cn-2 + Cn-1 + Cn + Cn+1 + Cn+2 +
#           (1.50*Rn+3) + (1.00*Rn+4) + (0.66*Rn+5) + (0.33*Rn+6)
# Score at residue n is the summation of weight*ddGof aa at position
# n-6 to n+6. There are 3 tables for ddG (left, core, right)
#
# Input: multiple fasta amino acid sequences
# >header 1
# seq1.....
# >header 2
# seq2.....
#
# Output: score from residue 7 to (length-6) separate by tab
# >header 1
# score(7)<tab>score(8)<tab>...score(length-6)<end-line>
# >header 2
# score(7)<tab>score(8)<tab>...score(length-6)<end-line>
#
#####

use strict;
use warnings;

#Subroutine prototypes
sub CBPSScore();

#global vars
# CBPS - Cellulose-bound peptide scanning
# Equation come with correction factor (can be seen as weight factor)
my @cf = ("0.33", "0.66", "1.00", "1.50", "1.00", "1.00", "1.00", "1.00", "1.00", "1.00",
          "1.50", "1.00", "0.66", "0.33");
# delta delta G values are copies from Table I (Left, Core, Right)
my %CBPS_L = ( "A", "-0.07", #Alanine
               "C", "4.87", #Cysteine
               "D", "0.44", #Aspartate
               "E", "1.48", #Glutamate
               "F", "0.14", #Phenylalanine
               "G", "-0.33", #Glycine
               "H", "-0.24", #Histidine
               "I", "1.04", #Isoleucine
               "K", "-0.85", #Lysine
               "L", "1.70", #Leucine
               "M", "0.08", #Methionine
               "N", "0.74", #Asparagine
               "P", "0.15", #Proline
               "Q", "-1.13", #Glutamine
               "R", "-1.19", #Arginine
               "S", "-0.13", #Serine
               "T", "-0.91", #Threonine
               "V", "-0.26", #Valine
               "W", "-0.43", #Tryptophan
               "Y", "0.19" ); #Tyrosine
my %CBPS_C = ( "A", "0.79", #Alanine
               "C", "6.35", #Cysteine
               "D", "4.91", #Aspartate
               "E", "5.14", #Glutamate
               "F", "-1.17", #Phenylalanine
               "G", "1.95", #Glycine
               "H", "1.74", #Histidine
               "I", "-2.05", #Isoleucine
               "K", "0.40", #Lysine
               "L", "-3.62", #Leucine
               "M", "1.10", #Methionine
               "N", "2.36", #Asparagine
               "P", "1.63", #Proline
               "Q", "1.60", #Glutamine
               "R", "-0.79", #Arginine
               "S", "1.27", #Serine
               "T", "0.27", #Threonine
               "V", "-1.75", #Valine
               "W", "3.49", #Tryptophan
               "Y", "-1.88" ); #Tyrosine
my %CBPS_R = ( "A", "0.46", #Alanine
```

```

"C", "0.25", #Cysteine
"D", "0.35", #Aspartate
"E", "1.65", #Glutamate
"F", "0.53", #Phenylalanine
"G", "0.03", #Glycine
"H", "0.09", #Histidine
"I", "0.11", #Isoleucine
"K", "-1.08", #Lysine
"L", "-0.02", #Leucine
"M", "0.17", #Methionine
"N", "-0.29", #Asparagine
"P", "-0.28", #Proline
"Q", "-0.15", #Glutamine
"R", "-1.72", #Arginine
"S", "-0.23", #Serine
"T", "-0.48", #Threonine
"V", "0.70", #Valine
"W", "0.12", #Tryptophan
"Y", "1.15" ); #Tyrosine

my $i = 0;
my @temp = ();
my $num = 0;
my $line = "";
my $seq = "";
my $linelimit = 60;
my $filenum = 1;
my $windowsize = 13;

#usage
my $USAGE = "usage: $0 <input fasta filename> <output filename>\n\n";

#check argument
unless (@ARGV == 2) {
    print $USAGE;
    exit -1;
}

#store argv
my $infilename = $ARGV[0];
my $outfilename = $ARGV[1];

#open infile
unless ( open(INFILE,"<$infilename") ) {
    print "Can't open $infilename\n\n";
    exit -1;
}

#open outfile
unless ( open(OUTFILE,">$outfilename") ) {
    print "Can't create $outfilename\n\n";
}

#read infile and write outfile
while ( $line = <INFILE> ) {
    chomp ($line);
    if ( $line =~ /^>/ ) { #this line is a header line
        $num++; #count header
        if ($num != 1) { #not the 1st header, cal & write previous seq result
            my $seqlength = length($seq);
            # position used -6...-1 0 +1...+6
            # CBPS score of this string is give to pos 0 residue
            for ($i=6; $i < $seqlength-6; $i++) { #calculate from index 6..ln-7
                print "pos ".$(i-6)." to ".$(i-6+$windowsize-1)." \n";
                my $subseq = substr($seq,$i-6,$windowsize);
                print "seq = $subseq\n";
                my $calscore = &CBPSScore($subseq);
                print "subseq : $calscore\n";
                printf OUTFILE '%.4f', $calscore;
                if ($i != $seqlength-7) {
                    print OUTFILE "\t";
                } else {
                    print OUTFILE "\n";
                }
            }
            $seq = ""; #clear seq
        }
        #write header
        print "\n\n$line\n";
        print OUTFILE $line."\t"; # use \t for same line or \n for new line
    }
    else { #otherline remove spaces, concate seq
        $line =~ s/\s//g;
        $seq .= $line;
    }
}

#last seq
my $seqlength = length($seq);
for ($i=6; $i < $seqlength-6; $i++) { #calculate from index 6..ln-7
    print "pos ".$(i-6)." to ".$(i-6+$windowsize-1)." \n";
    my $subseq = substr($seq,$i-6,$windowsize);
    print "seq = $subseq\n";
}

```



```

        my $score = &CBPSScore($subseq);
        print "$subseq : $score\n";
        printf OUTFILE "%.4f", $score;
        if ($i != $seqlength-7) {
            print OUTFILE "\t";
        } else {
            print OUTFILE "\n"; }
    }

close OUTFILE;

print "Total = $num sequences\n\n";

close INFILE;

exit 0;

#####
#Subroutine programs
#####
#
sub CBPSScore () {
# Calculate CBPSScore for the input string
    my $sixaa = $_[0];
    my $score = 0; #non bias, will be used in summation
    my $pos = 0;
    if (length($sixaa) != 13) { print "input for CBPSScore() is not 13 aa.\n"; }
    for ($pos = 0; $pos < length($sixaa); $pos++) {
        if ($pos < 4 ) {
            $score = $score+($cf[$pos]*$CBPS_L(substr($sixaa,$pos,1)));
        } elsif ($pos < 9 ) {
            $score = $score+($cf[$pos]*$CBPS_C(substr($sixaa,$pos,1)));
        } else {
            $score = $score+($cf[$pos]*$CBPS_R(substr($sixaa,$pos,1)));
        }
    }
    return $score;
}
#
#####

```

Code A4-4. FGLK Motif Prediction

```
#!/usr/bin/perl
#####
# FGLKscoring2.pl
# Created by Prakitchai Chotewutmontri, 20 September 2011
# GST student
#
# OBJECTIVE
#   Calculate FGLK motif score based on heuristic rules developed by
#   David McWilliams (2007). Dissertation. UTK.
#   The score scheme is novel from this work
#
# Input: multiple fasta amino acid sequences
#   >header 1
#   seq1.....
#   >header 2
#   seq2.....
#
# Output: score from residue 4 to (length-4) separate by tab
#   >header 1<tab>score(4)<tab>score(5)<tab>...score(length-4)<end-line>
#   >header 2<tab>score(4)<tab>score(5)<tab>...score(length-4)<end-line>
#
# Output: FASTA of FOUND SEQ(4 group matched)
#   >header1_match1
#   seq
#   >header1_match2
#   seq
#
#####

use strict;
use warnings;

#Subroutine prototypes
sub FGLKscore2();

#global vars
my $good = 2;
my $bad = 0;
my $other = 1; #don't need

my $win = 8; #the length of the window to calculate score
my $aai = 0; #hold the amino acid index

my $i = 0;
my @temp = ();
my $num = 0;
my $line = "";
my $seq = "";
my $nummatch = 0;
my $header = "";

#usage
my $USAGE = "usage: $0 <input fasta file> <out score file> <out FASTA matched seq file>\n\n";

#check argument
unless (@ARGV == 3) {
    print $USAGE;
    exit -1;
}

#store argv
my $infname = $ARGV[0];
my $outfname = $ARGV[1];
my $fasta = $ARGV[2];

#open infile
unless ( open(INFILE,"<$infname") ) {
    print "Can't open $infname\n\n";
    exit -1;
}

#open outfile
unless ( open(OUTFILE,">$outfname") ) {
    print "Can't create $outfname\n\n";
}

#open fasta
unless ( open(FASTA,">$fasta") ) {
    print "Can't create $fasta\n\n";
}

#read infile and write outfile
while ( $line = <INFILE> ) {
    chomp ($line);
    if ( $line =~ /^>/ ) { #this line is a header line
        $num++; #count header
```

```

        if ($num != 1) { #not the 1st header, cal & write previous seq result
            my $seqlength = length($seq);
            #keep track of match seq
            $nummatch = 0;
            # calculate score with in window the whole length of seq
            for ($aai=0; $aai <= $seqlength-$win; $aai++) { #cal from index 0 to length-win
                print "pos ".$aai." to ".$($aai+$win-1)."\n";
                my $subseq = substr($seq,$aai,$win);
                print "seq = $subseq\n";
                my $calscore = &FGLKscore2($subseq);
                #print match seq to FASTA out file
                if ( $calscore >= $good**4 ) {
                    $nummatch++; #count the match in this seq
                    print FASTA "$header";
                    print FASTA " ";
                    print FASTA "$nummatch\n"; #FASTA formatted header on separate line
                    print FASTA "$subseq\n";

                }
                print "$subseq : $calscore\n";
                printf OUTFILE '%.4f', $calscore;
                if ($aai != $seqlength-$win) {
                    print OUTFILE "\t";
                } else {
                    print OUTFILE "\n"; }
            }
            $seq = ""; #clear seq
        }
        #write header
        print "\n\n$line\n";
        $header = $line; #keep header for FASTA out file
        print OUTFILE $line."\t"; #same line header
        #print OUTFILE $line."\n"; #separate line header
    }
    else { #otherline remove spaces, concate seq
        $line =~ s/\s//g ;
        $seq .= $line;
    }
}
#last seq
my $seqlength = length($seq);
#keep track of match seq
$nummatch = 0;
# calculate score with in window the whole length of seq
for ($aai=0; $aai <= $seqlength-$win; $aai++) { #calculate from index 0 to length-win
    print "pos ".$aai." to ".$($aai+$win-1)."\n";
    my $subseq = substr($seq,$aai,$win);
    print "seq = $subseq\n";
    my $calscore = &FGLKscore2($subseq);
    #print match seq to FASTA out file
    if ( $calscore >= $good**4 ) {
        $nummatch++; #count the match in this seq
        print FASTA "$header";
        print FASTA " ";
        print FASTA "$nummatch\n"; #FASTA formatted header on separate line
        print FASTA "$subseq\n";
    }
    print "$subseq : $calscore\n";
    printf OUTFILE '%.4f', $calscore;
    if ($aai != $seqlength-$win) {
        print OUTFILE "\t";
    } else {
        print OUTFILE "\n"; }
}

close OUTFILE;
close FASTA;

print "Total = $num sequences\n\n";

close INFILE;

exit 0;

#####
#Subroutine programs
#####
#
sub FGLKscore2 () {
    # Calculate FGLK score for the input string
    my $winaa = $_[0]; #input amino acid seq from a window
    my $score = 1; #multiplication unbias value
    #my $score = 0; #summation unbias value

    #test the length
    if (length($winaa) != $win) { print "input for FGLKscore() is not $win aa.\n"; }

    #scoring
    #RULE 22
    # F AND P|G AND K|R AND A|L|V NOT D|E

```

```

print "Found: ";

#multiplication
if ( $winaa =~ /[Ff]/ ) { #match F or f
    $score = $score*$good;
    print "F";
}
if ( $winaa =~ /[PpGg]/ ) { #match P or p or G or g
    $score = $score*$good;
    print "P";
}
if ( $winaa =~ /[KkRr]/ ) { #match K or k or R or r
    $score = $score*$good;
    print "K";
}
if ( $winaa =~ /[AaLlVv]/ ) { #match A|a|L|l|V|v
    $score = $score*$good;
    print "L";
}
if ( $winaa =~ /[DdEe]/ ) { #match D|d|E|e
    $score = $bad; #as the last if, this would result in "$bad=0" score
    print "\tDorE";
}
print "\n";

return $score;
}
#
#####

```

Code A4-5. TP PSSM Calculator Using the N-terminal 10 Residues

```
#!/usr/bin/perl
#####
# Created by Prakitchai Chotewutmontri, 19 November 2012
#
# GST student
#
# OBJECTIVE
#   Calculate the score measuring how close is the N-terminal region
#   resemble the N-terminal domain of TP
#   From WebLogo, the N-terminal of TP contain: (i) the N-ter Met, (ii) the
#   second residue which is generally Ala, (iii) highly uncharged until about
#   aa 12.
#   AA freq distributions of aa2 and from aa3-12 were calculated. These will
#   be used to represent TP N-terminus.
#   AA freq distribution of UniProt release 2012_10 was used at background
#   frequency
#
# APPROACH
#   The log odd score similar to position-specific scoring matrix (PSSM)
#   scoring scheme will be used to calculate the score for the N-terminal
#   domain of the sequences.
#   However, only 2 position matrix will be made. One for the 2nd aa and
#   another one for aa3-12. This is because, the 2nd has its own distribution
#   and aa3-12 seems to have the same distribution.
#   The log odd table = log (freq of aai in TP/freq of aai in UniProt)
#   Score aa seq 1-10 = sum2 to 10 of log odds
#
# Input: multiple fasta amino acid sequences
#   >header 1
#   seq1.....
#   >header 2
#   seq2.....
#
# Output: separate by tab
#   <header><N-ter sequence><log odd score>\n
#
#####
#Subroutine prototypes
sub LOGODDAA ();

#global vars
my $total = 0;
my $line = "";
my $num = 0;
my $rlen = 10; #length of N-terminal domain, CHANGE IF NEEDED
my $seq = "";

#usage
my $USAGE = "usage: $0 <input fasta filename> <output filename>\n\n";

#check argument
unless (@ARGV == 2) {
    print $USAGE;
    exit -1;
}

#store argv
my $infile = $ARGV[0];
my $outfile = $ARGV[1];

#open infile
unless ( open(INFILE,"<$infile") ) {
    print "Can't open $infile\n\n";
    exit -1;
}

#open outfile, first time, overwrite it
unless ( open(OUTFILE,">$outfile") ) {
    print "Can't create $outfile\n\n";
}

#write the outfile header line
#   <header><N-ter sequence><log odd score>
print OUTFILE "seq ID\tN-ter sequence\tlog odd score\n";

#main part
$total = 0; #count # sequence
while ($line = <INFILE> ) { #read input file line by line

    chomp ($line); #remove end-line char

    if ($line =~ /^>/ ) {#found the > at the first char, mean the header line
        $num++; #count header
        if ($num != 1) { #not the first header, calculate & write out previous seq result
            #sequence of the previous header was collected before this point
```

```

#now calculate the score
my $nseq = substr($seq,0,$nlen);
$nseq =~ tr/[a-z]/[A-Z]/; #convert to uppercase only
print "$nseq\n";

my $score = 0;
my $resi = 0;
for ($resi = 1; $resi<length($nseq); $resi++){
    my $pos_i = $resi+1;
    my $aa_i = substr($nseq,$resi,1);
    #print "$pos_i\t$aa_i\t$score\t";
    $score = $score + &LOGODDAA ($aa_i,$pos_i);
    #print &LOGODDAA($aa_i,$pos_i);
    #print "\t$score\n";
}

#write out file
print OUTFILE "$header\t$nseq\t";
printf OUTFILE '%.4f', $score;
print OUTFILE "\n"; #end of line

$seq = ""; #clear seq after used
}
#keep header
$header = $line;
}
else { #otherline remove spaces, concate sequence
    $line =~ s/\s//g ;
    $seq .= $line; #concate
} #end if header line
} #end while line = infile

#last seq
#calculate the score
my $nseq = substr($seq,0,$nlen);
$nseq =~ tr/[a-z]/[A-Z]/; #convert to uppercase only
print "$nseq\n";

my $score = 0;
my $resi = 0;
for ($resi = 1; $resi<length($nseq); $resi++){
    my $pos_i = $resi+1;
    my $aa_i = substr($nseq,$resi,1);
    #print "$pos_i\t$aa_i\t$score\t";
    $score = $score + &LOGODDAA ($aa_i,$pos_i);
    #print &LOGODDAA($aa_i,$pos_i);
    #print "\t$score\n";
}

#write out file
print OUTFILE "$header\t$nseq\t";
printf OUTFILE '%.4f', $score;
print OUTFILE "\n"; #end of line

close OUTFILE;
close INFILE;

exit 0;

#####
#Subroutine programs
#####
#
sub LOGODDAA () {
# Return log odd value of the input aa based on position
    my $inaa = $_[0]; #input aa
    my $inpos = $_[1]; #input aa position

    my %logodd2 = ( "A", 2.442563807,
                    "C", -1.191215311,
                    "D", -1.140495553,
                    "E", 0.388298646,
                    "F", -1.497971094,
                    "G", -0.460375819,
                    "H", -2.750299179,
                    "I", -1.383860388,
                    "K", -1.789852748,
                    "L", -1.745212866,
                    "M", -0.489873626,
                    "N", -0.185959585,
                    "P", -1.510833245,
                    "Q", -0.851139869,
                    "R", -2.17273589,
                    "S", 0.690140193,
                    "T", -0.914005136,
                    "V", -0.862908303,
                    "W", -10.20244257,
                    "Y", -2.800024553);

```

```

my %logodd3up = ( "A", 0.471636772,
                  "C", 0.61932271,
                  "D", -3.329957533,
                  "E", -2.923153859,
                  "F", 0.135689164,
                  "G", -1.297536055,
                  "H", -1.065292041,
                  "I", -0.265200073,
                  "K", -1.538742137,
                  "L", 0.321664908,
                  "M", 0.178645389,
                  "N", -0.58692567,
                  "P", 0.120734634,
                  "Q", -0.362529944,
                  "R", -1.612176123,
                  "S", 1.854635162,
                  "T", 0.862744571,
                  "V", -0.095439005,
                  "W", -2.163085855,
                  "Y", -1.902023692);

my $outlogodd = 0;

if ($inpos == 2) {
    $outlogodd = $logodd2{$inaa};
} else {
    $outlogodd = $logodd3up{$inaa};
}

return ($outlogodd);
}

```

Code A4-6. PSSM Calculator Using the N-terminal 30 Residues

```
#!/usr/bin/perl
#####
# Created by Prakitchai Chotewutmontri, 19 November 2012
#
# GST student
#
# OBJECTIVE
#   Calculate the score measuring how close is the N-terminal region
#   resemble the N-terminal domain of TP
#   From WebLogo, the N-terminal of TP contain: (i) the N-ter Met, (ii) the
#   second residue which is generally Ala, (iii) highly uncharged until about
#   aa 12.
#   AA freq distributions of aa2 and from aa3-12 were calculated. These will
#   be used to represent TP N-terminus.
#   AA freq distribution of UniProt release 2012_10 was used at background
#   frequency
#
# APPROACH
#   The log odd score similar to position-specific scoring matrix (PSSM)
#   scoring scheme will be used to calculate the score for the N-terminal
#   domain of the sequences.
#   However, only 2 position matrix will be made. One for the 2nd aa and
#   another one for aa3-12. This is because, the 2nd has its own distribution
#   and aa3-12 seems to have the same distribution.
#   The log odd table = log (freq of aai in TP/freq of aai in UniProt)
#   Score aa seq 1-10 = sum2 to 10 of log odds
#
# Input: multiple fasta amino acid sequences
#   >header 1
#   seq1.....
#   >header 2
#   seq2.....
#
# Output: separate by tab
#   <header><N-ter sequence><cp score aa2-12><cp score aa13-19><cp score aa20-30>
#   <total cp score><total mt score><total sp score><total others score>
#   <cp prob><mt prob><sp prob>\n
#
#####
use strict;
use warnings;

#Subroutine prototypes
sub LOGODDAA ();
sub ERF ();
sub ZDIST ();
sub N_PDF ();

#global vars
my $total = 0;
my $line = "";
my $num = 0;
my $nlen = 30; #length of N-terminal domain, CHANGE IF NEEDED
my $llen = 12; #end residue at the left side
my $mmin = 13; #start of the middle
my $mmax = 19; #end of the middle
my $rmin = 20; #start of the right
my $rmax = 30; #end of the right
my $seq = "";
my $header = "";

#MEAN and SD from the prediction results, give probability
#Based on the calculated score of each category
#   score->%relative accumulative dist->get %amplitude, mean and sd
#   %amplitude, mean and sd were derived from true training set
#   eg. the %ampli, mean and sd of cp are from the scores of Lee208TP set (training set)

my $cpampl = 0.1158; #from 11.58%
my $cpmean = 12.98;
my $cpsd = 6.789;

my $mtampl = 0.1418;
my $mtmean = 9.772;
my $mtsd = 5.503;

my $spampl = 0.1345;
my $spmean = 13.79;
my $spsd = 5.332;

#usage
my $USAGE = "usage: $0 <input fasta filename> <output filename>\n\n";
#check argument
unless (@ARGV == 2) {
    print $USAGE;
    exit -1;
}
```



```

#store argv
my $infile = $ARGV[0];
my $outfile = $ARGV[1];

#open infile
unless ( open(INFILE,"<$infile") ) {
    print "Can't open $infile\n\n";
    exit -1;
}

#open outfile, first time, overwrite it
unless ( open(OUTFILE,">$outfile") ) {
    print "Can't create $outfile\n\n";
}

#write the outfile header line
# <header><N-ter sequence><log odd score>
print OUTFILE "seq ID\tN-ter sequence\tlog odd cp score aa2-12";
print OUTFILE "\tlog odd cp score aa13-19\tlog odd cp score aa20-30\ttotal cp score";
print OUTFILE "\tttotal mt score\ttotal sp score\ttotal others score";
print OUTFILE "\tcp prob\tmt prob\tsp prob\n\n";

#main part
$total = 0; #count # sequence
while ($line = <INFILE> ) { #read input file line by line

    chomp ($line); #remove end-line char

    if ($line =~ /^>/ ) {#found the > at the first char, mean the header line
        $num++; #count header
        if ($num != 1) { #not the first header, calculate & write out previous seq result
            #sequence of the previous header was collected before this point

            #now calculate the score
            my $nseq = substr($seq,0,$nlen);
            $nseq =~ tr/[a-z]/[A-Z]/; #convert to uppercase only
            print "$nseq\n\n";

            my $cplscore = 0;
            my $cpmscore = 0;
            my $cprscore = 0;
            my $cpttotal = 0;

            my $mttotal = 0;
            my $spttotal = 0;
            my $otttotal = 0;

            my $resi = 0;
            for ($resi = 1; $resi<$llen; $resi++){
                my $pos_i = $resi+1;
                my $aa_i = substr($nseq,$resi,1);
                my @temp1 = &LOGODDAA ($aa_i,$pos_i);
                $cplscore = $cplscore + $temp1[0];
                $mttotal = $mttotal + $temp1[1];
                $spttotal = $spttotal + $temp1[2];
                $otttotal = $otttotal + $temp1[3];
            }
            for ($resi = $mmin-1; $resi<$rmax; $resi++){
                my $pos_i = $resi+1;
                my $aa_i = substr($nseq,$resi,1);
                my @temp2 = &LOGODDAA ($aa_i,$pos_i);
                $cpmscore = $cpmscore + $temp2[0];
                $mttotal = $mttotal + $temp2[1];
                $spttotal = $spttotal + $temp2[2];
                $otttotal = $otttotal + $temp2[3];
            }
            for ($resi = $rmin-1; $resi<$rmax; $resi++){
                my $pos_i = $resi+1;
                my $aa_i = substr($nseq,$resi,1);
                my @temp3 = &LOGODDAA ($aa_i,$pos_i);
                $cprscore = $cprscore + $temp3[0];
                $mttotal = $mttotal + $temp3[1];
                $spttotal = $spttotal + $temp3[2];
                $otttotal = $otttotal + $temp3[3];
            }
            $cpttotal = $cplscore+$cpmscore+$cprscore;

            #my $cprob = &ZDIST((($cpttotal-$cpmean)/$cpsd));
            #my $mtprob = &ZDIST((($mttotal-$mtmean)/$mtsd));
            #my $spprob = &ZDIST((($spttotal-$spmean)/$spsd));

            my $cprob = &N_PDF($cpttotal, $cpampl, $cpmean, $cpsd);
            my $mtprob = &N_PDF($mttotal, $mtampl, $mtmean, $mtsd);
            my $spprob = &N_PDF($spttotal, $spampl, $spmean, $spsd);

            #write out file
            print OUTFILE "$header\t$nseq\t";
            printf OUTFILE "%.4f", $cplscore;
            print OUTFILE "\t";
        }
    }
}

```

```

        printf OUTFILE '%.4f', $cpmscore;
        print OUTFILE "\t";
        printf OUTFILE '%.4f', $cprscore;
        print OUTFILE "\t";
        printf OUTFILE '%.4f', $cpttotal;
        print OUTFILE "\t";
        printf OUTFILE '%.4f', $mttotal;
        print OUTFILE "\t";
        printf OUTFILE '%.4f', $spttotal;
        print OUTFILE "\t";
        printf OUTFILE '%.4f', $otttotal;
        print OUTFILE "\t";
        printf OUTFILE '%.4f', $cpprob;
        print OUTFILE "\t";
        printf OUTFILE '%.4f', $mtprob;
        print OUTFILE "\t";
        printf OUTFILE '%.4f', $spprob;
        print OUTFILE "\n"; #end of line

        $seq = ""; #clear seq after used
    }
    #keep header
    $header = $line;
}
else { #otherline remove spaces, concatenate sequence
    $line =~ s/\s//g ;
    $seq .= $line; #concatenate
} #end if header line

} #end while line = infile

#last seq
#calculate the score
my $nseq = substr($seq,0,$nlen);
$nseq =~ tr/[a-z]/[A-Z]/; #convert to uppercase only
print "$nseq\n";

my $cplscore = 0;
my $cpmscore = 0;
my $cprscore = 0;
my $cpttotal = 0;

my $mttotal = 0;
my $spttotal = 0;
my $otttotal = 0;

my $resi = 0;
for ($resi = 1; $resi<$llen; $resi++){
    my $pos_i = $resi+1;
    my $aa_i = substr($nseq,$resi,1);
    my @temp1 = &LOGODDAA ($aa_i,$pos_i);
    $cplscore = $cplscore + $temp1[0];
    $mttotal = $mttotal + $temp1[1];
    $spttotal = $spttotal + $temp1[2];
    $otttotal = $otttotal + $temp1[3];
}
for ($resi = $mmin-1; $resi<$mmax; $resi++){
    my $pos_i = $resi+1;
    my $aa_i = substr($nseq,$resi,1);
    my @temp2 = &LOGODDAA ($aa_i,$pos_i);
    $cpmscore = $cpmscore + $temp2[0];
    $mttotal = $mttotal + $temp2[1];
    $spttotal = $spttotal + $temp2[2];
    $otttotal = $otttotal + $temp2[3];
}
for ($resi = $rmin-1; $resi<$rmax; $resi++){
    my $pos_i = $resi+1;
    my $aa_i = substr($nseq,$resi,1);
    my @temp3 = &LOGODDAA ($aa_i,$pos_i);
    $cprscore = $cprscore + $temp3[0];
    $mttotal = $mttotal + $temp3[1];
    $spttotal = $spttotal + $temp3[2];
    $otttotal = $otttotal + $temp3[3];
}
$cpttotal = $cplscore+$cpmscore+$cprscore;

#my $cpprob = &ZDIST((( $cpttotal-$cpmean)/$cpsd));
#my $mtprob = &ZDIST((( $mttotal-$mtmean)/$mtsd));
#my $spprob = &ZDIST((( $spttotal-$spmean)/$spsd));

my $cpprob = &N_PDF($cpttotal, $cpampl, $cpmean, $cpsd);
my $mtprob = &N_PDF($mttotal, $mtampl, $mtmean, $mtsd);
my $spprob = &N_PDF($spttotal, $spampl, $spmean, $spsd);

#write out file
print OUTFILE "$header\t$nseq\t";
printf OUTFILE '%.4f', $cplscore;
print OUTFILE "\t";
printf OUTFILE '%.4f', $cpmscore;
print OUTFILE "\t";

```

```

printf OUTFILE '%.4f', $cprrscore;
print OUTFILE "\t";
printf OUTFILE '%.4f', $cpttotal;
print OUTFILE "\t";
printf OUTFILE '%.4f', $mttotal;
print OUTFILE "\t";
printf OUTFILE '%.4f', $spttotal;
print OUTFILE "\t";
printf OUTFILE '%.4f', $otttotal;
print OUTFILE "\t";
printf OUTFILE '%.4f', $cpprob;
print OUTFILE "\t";
printf OUTFILE '%.4f', $mtprob;
print OUTFILE "\t";
printf OUTFILE '%.4f', $spprob;
print OUTFILE "\n"; #end of line

close OUTFILE;
close INFILE;

exit 0;

#####
#Subroutine programs
#####
#
sub LOGODDAA () {
# Return log odd value of the input aa based on position
my $inaa = $_[0]; #input aa
my $inpos = $_[1]; #input aa position

my %cpodd = (
    aa2 => {
        "A" => 2.442563807,
        "C" => -1.191215311,
        "D" => -1.140495553,
        "E" => 0.388298646,
        "F" => -1.497971094,
        "G" => -0.460375819,
        "H" => -2.750299179,
        "I" => -1.383860388,
        "K" => -1.789852748,
        "L" => -1.745212866,
        "M" => -0.489873626,
        "N" => -0.185959585,
        "P" => -1.510833245,
        "Q" => -0.851139869,
        "R" => -2.17273589,
        "S" => 0.690140193,
        "T" => -0.914005136,
        "V" => -0.862908303,
        "W" => -10.20244257,
        "Y" => -2.800024553,
    },
    aa3to12 => {
        "A" => 0.471636772,
        "C" => 0.61932271,
        "D" => -3.329957533,
        "E" => -2.923153859,
        "F" => 0.135689164,
        "G" => -1.297536055,
        "H" => -1.065292041,
        "I" => -0.265200073,
        "K" => -1.538742137,
        "L" => 0.321664908,
        "M" => 0.178645389,
        "N" => -0.58692567,
        "P" => 0.120734634,
        "Q" => -0.362529944,
        "R" => -1.612176123,
        "S" => 1.854635162,
        "T" => 0.862744571,
        "V" => -0.095439005,
        "W" => -2.163085855,
        "Y" => -1.902023692,
    },
    aa20to30 => {
        "A" => -0.593547396,
        "C" => 0.104749537,
        "D" => -1.61946415,
        "E" => -2.425654199,
        "F" => 0.36701471,
        "G" => -0.744995032,
        "H" => 0.175716058,
        "I" => -0.87817695,
        "K" => 0.300793191,
        "L" => -0.082232034,
        "M" => -2.094373105,
        "N" => 0.039615934,
        "P" => 0.793693616,
    },

```

```

"Q" => -0.340162131,
"R" => 0.518220514,
"S" => 1.863248292,
"T" => -0.171854091,
"V" => -0.510476504,
"W" => -2.748048356,
"Y" => -1.667558438,
),
aa13 => {
  "A" => -0.166437065,
  "C" => 0.941250804,
  "D" => -2.467461057,
  "E" => -3.688688605,
  "F" => 1.021518144,
  "G" => -1.882498556,
  "H" => -0.202795565,
  "I" => 0.356288304,
  "K" => -0.657386633,
  "L" => 0.155927703,
  "M" => -0.035479416,
  "N" => 0.072037412,
  "P" => 0.962669788,
  "Q" => -0.718673754,
  "R" => -0.914738893,
  "S" => 1.634161219,
  "T" => 0.270928399,
  "V" => -1.232942529,
  "W" => -0.426120261,
  "Y" => -0.667558438,
),
aa14 => {
  "A" => 0.155421671,
  "C" => 0.941181446,
  "D" => -1.14560232,
  "E" => -2.688757964,
  "F" => 0.741340866,
  "G" => -0.297605414,
  "H" => -1.202864924,
  "I" => -0.058818554,
  "K" => -1.142882819,
  "L" => 0.276152578,
  "M" => -0.772514369,
  "N" => 0.223971147,
  "P" => 0.884597917,
  "Q" => -3.040671208,
  "R" => -0.329845751,
  "S" => 1.455754619,
  "T" => 0.163943836,
  "V" => -0.233011887,
  "W" => -8.07004581,
  "Y" => -1.667627797,
),
aa15 => {
  "A" => -0.581613278,
  "C" => -0.380816004,
  "D" => -2.467599771,
  "E" => -10.33268351,
  "F" => 0.393348208,
  "G" => -0.712712268,
  "H" => 0.118993816,
  "I" => -0.05888791,
  "K" => -0.142952174,
  "L" => 0.155788989,
  "M" => -1.357546225,
  "N" => -0.776098209,
  "P" => 1.469491063,
  "Q" => -1.455778062,
  "R" => 0.200599611,
  "S" => 1.337040767,
  "T" => 0.986996719,
  "V" => -0.648118742,
  "W" => -8.070115165,
  "Y" => -0.345769057,
),
aa16 => {
  "A" => -0.079112938,
  "C" => 0.204146496,
  "D" => -2.467599771,
  "E" => -2.688827319,
  "F" => 0.51887909,
  "G" => -0.712712268,
  "H" => -1.202934279,
  "I" => -0.64385041,
  "K" => 0.23559449,
  "L" => 0.276083223,
  "M" => -9.001402415,
  "N" => -0.098026303,
  "P" => 1.174035179,
  "Q" => -1.455778062,
  "R" => 0.892477315,

```

```

        "S" => 1.455685263,
        "T" => 0.270789685,
        "V" => -1.496115649,
        "W" => -8.070115165,
        "Y" => -0.667697152,
    },
    aa17 => {
        "A" => -0.166506424,
        "C" => 0.204215852,
        "D" => -1.467530415,
        "E" => -1.688757964,
        "F" => 0.518948445,
        "G" => -0.182128196,
        "H" => 0.119063171,
        "I" => -0.184349436,
        "K" => -0.294885912,
        "L" => 0.091728007,
        "M" => -9.00133306,
        "N" => -1.512994447,
        "P" => 0.621563512,
        "Q" => 0.659768511,
        "R" => 0.822157343,
        "S" => 1.162023416,
        "T" => 0.370394714,
        "V" => -0.496046293,
        "W" => -1.42618962,
        "Y" => -2.667627797,
    },
    aa18 => {
        "A" => -0.466066705,
        "C" => -1.380746649,
        "D" => -0.882567915,
        "E" => -3.688757964,
        "F" => 0.103910946,
        "G" => -0.56063982,
        "H" => -1.202864924,
        "I" => 0.163573867,
        "K" => -0.005379295,
        "L" => 0.155858344,
        "M" => -9.00133306,
        "N" => 0.071968053,
        "P" => 0.962600429,
        "Q" => -0.233316285,
        "R" => 0.407119843,
        "S" => 1.417280471,
        "T" => 0.370394714,
        "V" => -0.648049387,
        "W" => -0.42618962,
        "Y" => -0.345699702,
    },
    aa19 => {
        "A" => -0.466136061,
        "C" => 0.619183996,
        "D" => -1.88263727,
        "E" => -2.103864818,
        "F" => 0.634356307,
        "G" => -0.423205651,
        "H" => 0.118993816,
        "I" => -1.64385041,
        "K" => 0.442010326,
        "L" => 0.155788989,
        "M" => -9.001402415,
        "N" => 0.223901791,
        "P" => 0.802066402,
        "Q" => -0.233385641,
        "R" => 0.085122393,
        "S" => 1.565309754,
        "T" => 0.270789685,
        "V" => -0.233081243,
        "W" => -1.426258975,
        "Y" => -9.311553342,
    },
},

);

my %mtodd = (
    aa2 => {
        "A" => 1.755754768,
        "C" => -8.843808634,
        "D" => -10.9305924,
        "E" => -2.923001258,
        "F" => 1.472332154,
        "G" => -1.894418787,
        "H" => -2.022070719,
        "I" => -0.293061848,
        "K" => -1.699054208,
        "L" => 1.239355348,
        "M" => -1.176682665,
        "N" => -1.332200242,
        "P" => -0.934607877,
        "Q" => -0.689952001,
    }
);

```

```

        "R" => -0.511621625,
        "S" => 0.195976232,
        "T" => -2.033773582,
        "V" => -1.467255181,
        "W" => -0.245395415,
        "Y" => -0.31690859,
    },
    aa3 => {
        "A" => 1.20411234,
        "C" => 0.121975651,
        "D" => -10.9305924,
        "E" => -2.923001258,
        "F" => -0.077864928,
        "G" => -1.242342091,
        "H" => -1.437108218,
        "I" => -0.762547132,
        "K" => -0.583576991,
        "L" => 0.205408016,
        "M" => -0.591720164,
        "N" => -0.917162743,
        "P" => 0.143394635,
        "Q" => -2.274914502,
        "R" => 1.61176079,
        "S" => 1.365901234,
        "T" => -0.655261959,
        "V" => -0.829825261,
        "W" => -0.660432914,
        "Y" => -10.13068978,
    },
    aa4 => {
        "A" => 0.821642704,
        "C" => 0.969972557,
        "D" => -10.9305924,
        "E" => -2.186035664,
        "F" => 0.021670745,
        "G" => -2.379845615,
        "H" => -9.665926909,
        "I" => -0.46298685,
        "K" => -0.476661787,
        "L" => 0.567978096,
        "M" => 0.14524543,
        "N" => -1.109807821,
        "P" => 0.143394635,
        "Q" => -0.400445384,
        "R" => 1.658303376,
        "S" => 0.295511906,
        "T" => -0.355701677,
        "V" => 0.117707319,
        "W" => -0.245395415,
        "Y" => -0.901871091,
    },
    aa5 => {
        "A" => 0.941138832,
        "C" => 1.118089363,
        "D" => -3.290622498,
        "E" => -11.15570624,
        "F" => -0.429674519,
        "G" => -0.798769401,
        "H" => -2.025957006,
        "I" => -0.766433419,
        "K" => -0.587463278,
        "L" => 0.537124761,
        "M" => -0.010643951,
        "N" => -1.599120936,
        "P" => -0.523456666,
        "Q" => -1.056408368,
        "R" => 1.53511816,
        "S" => 0.472197864,
        "T" => 0.727874877,
        "V" => 0.002789719,
        "W" => -1.249281702,
        "Y" => -2.490719879,
    },
    aa6 => {
        "A" => 0.991607014,
        "C" => 1.12201496,
        "D" => -3.286696901,
        "E" => -11.15178064,
        "F" => -1.300218041,
        "G" => -0.614271559,
        "H" => -2.02203141,
        "I" => -0.655592619,
        "K" => 0.175454219,
        "L" => 0.646019917,
        "M" => -0.369288434,
        "N" => -1.917123434,
        "P" => -1.297138648,
        "Q" => -0.400406075,
        "R" => 0.966464981,
        "S" => 0.710588714,
    },

```

```

        "T" => 0.103769251 ,
        "V" => 0.407253245 ,
        "W" => 0.754643894 ,
        "Y" => -1.901831782 ,
    },
    aa7 => {
        "A" => 0.603147067 ,
        "C" => 0.973869342 ,
        "D" => -10.92669562 ,
        "E" => -11.14792316 ,
        "F" => -0.180883348 ,
        "G" => -0.449949411 ,
        "H" => -0.210819012 ,
        "I" => -0.371627224 ,
        "K" => -0.373229329 ,
        "L" => 0.904450219 ,
        "M" => -1.587823379 ,
        "N" => -0.743340957 ,
        "P" => -0.400196376 ,
        "Q" => -0.1555405 ,
        "R" => 1.111184992 ,
        "S" => 0.640445608 ,
        "T" => 0.171757064 ,
        "V" => 0.06715632 ,
        "W" => -8.88535482 ,
        "Y" => -2.482936807 ,
    },
    aa8 => {
        "A" => 0.502293794 ,
        "C" => 0.118128566 ,
        "D" => -2.705620794 ,
        "E" => -3.511810843 ,
        "F" => -0.304104434 ,
        "G" => -0.705620794 ,
        "H" => -0.703989708 ,
        "I" => -1.466833934 ,
        "K" => 0.235698163 ,
        "L" => 0.453099699 ,
        "M" => -1.595567248 ,
        "N" => -2.336047327 ,
        "P" => -0.786451868 ,
        "Q" => -0.056369165 ,
        "R" => 1.826923488 ,
        "S" => 0.670176445 ,
        "T" => 0.032768662 ,
        "V" => 0.166327655 ,
        "W" => 0.335720002 ,
        "Y" => -1.490680676 ,
    },
    aa9 => {
        "A" => 0.921217686 ,
        "C" => -0.614950634 ,
        "D" => -10.93055309 ,
        "E" => -2.922961949 ,
        "F" => 0.021710054 ,
        "G" => -0.057878211 ,
        "H" => -1.437068909 ,
        "I" => -1.29302254 ,
        "K" => -0.377086804 ,
        "L" => 0.541050357 ,
        "M" => -1.176643356 ,
        "N" => -0.010232839 ,
        "P" => -0.404053852 ,
        "Q" => -0.052482771 ,
        "R" => 1.324918952 ,
        "S" => 0.814925374 ,
        "T" => -0.448771772 ,
        "V" => -0.113578918 ,
        "W" => -1.245356106 ,
        "Y" => -2.486794283 ,
    },
    aa10 => {
        "A" => 0.371839786 ,
        "C" => -0.199952444 ,
        "D" => -10.9305924 ,
        "E" => -3.507963759 ,
        "F" => -0.184780132 ,
        "G" => 0.053113793 ,
        "H" => -0.700142624 ,
        "I" => -1.293061848 ,
        "K" => -0.476661787 ,
        "L" => 0.94237361 ,
        "M" => -0.591720164 ,
        "N" => -1.109807821 ,
        "P" => -0.197642283 ,
        "Q" => -0.859877002 ,
        "R" => 1.539004448 ,
        "S" => 0.710549405 ,
        "T" => 0.502279318 ,
        "V" => -1.177748564 ,
    },

```

```

        "W" => 0.754604585 ,
        "Y" => -10.13068978 ,
    },
    aa11 => {
        "A" => 0.742208236 ,
        "C" => 0.607402478 ,
        "D" => -2.701773709 ,
        "E" => -11.15181995 ,
        "F" => -0.885219851 ,
        "G" => -0.453846196 ,
        "H" => -1.437108218 ,
        "I" => -1.46298685 ,
        "K" => 0.037911386 ,
        "L" => 0.719981189 ,
        "M" => -0.369327743 ,
        "N" => -0.595234648 ,
        "P" => -0.519570378 ,
        "Q" => -1.05252208 ,
        "R" => 1.381463171 ,
        "S" => 1.0324775 ,
        "T" => 0.036615746 ,
        "V" => -0.315252088 ,
        "W" => 0.561959507 ,
        "Y" => -10.13068978 ,
    },
    aa12 => {
        "A" => 0.843291508 ,
        "C" => 0.796161268 ,
        "D" => -1.705659997 ,
        "E" => -2.926887545 ,
        "F" => -0.304143637 ,
        "G" => -0.798769401 ,
        "H" => -1.025957006 ,
        "I" => -1.007441519 ,
        "K" => 0.034025099 ,
        "L" => 0.42391415 ,
        "M" => -0.37321403 ,
        "N" => -0.113694108 ,
        "P" => 0.061505835 ,
        "Q" => 0.13623671 ,
        "R" => 1.349562507 ,
        "S" => 0.706663118 ,
        "T" => -0.552233042 ,
        "V" => -0.471141469 ,
        "W" => -8.893137892 ,
        "Y" => -1.168791784 ,
    },
    aa13 => {
        "A" => 1.014287782 ,
        "C" => -1.199952444 ,
        "D" => -10.9305924 ,
        "E" => -2.923001258 ,
        "F" => -0.30025735 ,
        "G" => 0.105581213 ,
        "H" => -0.437108218 ,
        "I" => -1.293061848 ,
        "K" => -0.036089195 ,
        "L" => 0.541011048 ,
        "M" => -9.820538854 ,
        "N" => -1.109807821 ,
        "P" => -0.104532879 ,
        "Q" => -0.05252208 ,
        "R" => 0.890476818 ,
        "S" => 0.746173315 ,
        "T" => -0.185776675 ,
        "V" => 0.220800812 ,
        "W" => -0.245395415 ,
        "Y" => -1.164905497 ,
    },
    aa14 => {
        "A" => 0.710841212 ,
        "C" => 0.603516191 ,
        "D" => -10.93447869 ,
        "E" => -2.926887545 ,
        "F" => 0.110893862 ,
        "G" => 0.152320998 ,
        "H" => -1.025957006 ,
        "I" => -0.659518215 ,
        "K" => -0.587463278 ,
        "L" => 0.394166807 ,
        "M" => -2.180568952 ,
        "N" => -1.33608653 ,
        "P" => 0.283898256 ,
        "Q" => -0.404331671 ,
        "R" => 1.320993355 ,
        "S" => 0.777052446 ,
        "T" => -0.272125123 ,
        "V" => 0.216914525 ,
        "W" => -1.249281702 ,
        "Y" => -2.490719879 ,
    },

```



```

),
aa15 => {
    "A" => 0.180365699 ,
    "C" => 0.118128566 ,
    "D" => -3.290583295 ,
    "E" => -2.511810843 ,
    "F" => -0.56713884 ,
    "G" => 0.101734128 ,
    "H" => -2.025917803 ,
    "I" => -0.296908933 ,
    "K" => -1.117938792 ,
    "L" => 0.201560932 ,
    "M" => -0.010604748 ,
    "N" => -1.599081733 ,
    "P" => 0.283937459 ,
    "Q" => -0.278761586 ,
    "R" => 1.510066383 ,
    "S" => 1.166133939 ,
    "T" => 0.284307429 ,
    "V" => -0.553564426 ,
    "W" => 0.920682502 ,
    "Y" => -2.490680676 ,
},
aa16 => {
    "A" => 0.742247545 ,
    "C" => 0.800086865 ,
    "D" => -3.286696901 ,
    "E" => -2.922961949 ,
    "F" => -0.07782562 ,
    "G" => 0.205156195 ,
    "H" => 0.147893592 ,
    "I" => -0.87798504 ,
    "K" => -0.699014899 ,
    "L" => 0.239394657 ,
    "M" => -1.591680855 ,
    "N" => -0.457691816 ,
    "P" => -0.297138648 ,
    "Q" => 0.140162306 ,
    "R" => 1.266025262 ,
    "S" => 0.943249471 ,
    "T" => 0.167899588 ,
    "V" => -0.549678033 ,
    "W" => -1.245356106 ,
    "Y" => -10.13065047 ,
},
aa17 => {
    "A" => 0.440591845 ,
    "C" => 1.259518484 ,
    "D" => -10.93055309 ,
    "E" => -1.50792445 ,
    "F" => -0.425748923 ,
    "G" => -0.1167719 ,
    "H" => 0.678408308 ,
    "I" => -0.87798504 ,
    "K" => -0.583537682 ,
    "L" => 0.135057997 ,
    "M" => -2.176643356 ,
    "N" => -1.595195339 ,
    "P" => -0.197602974 ,
    "Q" => -0.159397975 ,
    "R" => 0.725456881 ,
    "S" => 1.221550633 ,
    "T" => 0.644337632 ,
    "V" => -0.113578918 ,
    "W" => 0.339606395 ,
    "Y" => -0.901831782 ,
},
aa18 => {
    "A" => 0.502293794 ,
    "C" => 0.381162972 ,
    "D" => -2.705620794 ,
    "E" => -2.189882748 ,
    "F" => -0.304104434 ,
    "G" => 0.049266708 ,
    "H" => 0.296010292 ,
    "I" => -1.466833934 ,
    "K" => -0.587424075 ,
    "L" => 0.616598431 ,
    "M" => -0.595567248 ,
    "N" => -1.336047327 ,
    "P" => -0.301025041 ,
    "Q" => -0.163284369 ,
    "R" => 1.23176522 ,
    "S" => 0.670176445 ,
    "T" => 0.447806161 ,
    "V" => -0.734136672 ,
    "W" => 0.335720002 ,
    "Y" => -0.320755675 ,
},
aa19 => {

```

```

"A" => 0.710880415 ,
"C" => 0.381162972 ,
"D" => -2.705620794 ,
"E" => -3.511810843 ,
"F" => -0.304104434 ,
"G" => -0.120658293 ,
"H" => -0.025917803 ,
"I" => -0.881871434 ,
"K" => 0.034064302 ,
"L" => 0.131171604 ,
"M" => -2.180529749 ,
"N" => -0.599081733 ,
"P" => 0.061545038 ,
"Q" => -0.056369165 ,
"R" => 1.169029465 ,
"S" => 0.844205845 ,
"T" => 0.447806161 ,
"V" => -0.393099754 ,
"W" => -0.249242499 ,
"Y" => -1.168752581 ,
},
aa20 => {
"A" => 0.139723714 ,
"C" => 0.118128566 ,
"D" => -1.9686552 ,
"E" => -2.189882748 ,
"F" => -0.304104434 ,
"G" => -0.313303371 ,
"H" => -0.218562881 ,
"I" => -0.659479012 ,
"K" => 0.10445363 ,
"L" => 0.018696875 ,
"M" => -0.858601654 ,
"N" => -0.336047327 ,
"P" => 0.798510632 ,
"Q" => 0.223738754 ,
"R" => 1.321032558 ,
"S" => 0.472237067 ,
"T" => 0.22541374 ,
"V" => 0.059412451 ,
"W" => -0.664279998 ,
"Y" => -1.168752581 ,
},
aa21 => {
"A" => 0.595403198 ,
"C" => 0.966125473 ,
"D" => -2.705620794 ,
"E" => -1.189882748 ,
"F" => -0.081712013 ,
"G" => -0.182058838 ,
"H" => -0.703989708 ,
"I" => -1.466833934 ,
"K" => -0.480508871 ,
"L" => 0.332805465 ,
"M" => 0.141398346 ,
"N" => -1.336047327 ,
"P" => 0.061545038 ,
"Q" => 0.306200915 ,
"R" => 0.886629734 ,
"S" => 0.811038981 ,
"T" => 0.22541374 ,
"V" => -0.117465311 ,
"W" => 0.072685596 ,
"Y" => -1.490680676 ,
},
aa22 => {
"A" => 0.440591845 ,
"C" => -0.199913135 ,
"D" => -10.93055309 ,
"E" => -2.50792445 ,
"F" => 0.021710054 ,
"G" => -0.379806306 ,
"H" => -0.700103315 ,
"I" => -0.762507823 ,
"K" => -0.583537682 ,
"L" => -0.379515176 ,
"M" => -1.591680855 ,
"N" => -0.457691816 ,
"P" => 0.802397026 ,
"Q" => 0.047052902 ,
"R" => 1.107327516 ,
"S" => 1.003370463 ,
"T" => 0.502318627 ,
"V" => 0.220840121 ,
"W" => -0.245356106 ,
"Y" => -0.316869281 ,
},
aa23 => {
"A" => 0.682866039 ,
"C" => 0.118128566 ,

```

```

"D" => -2.290583295 ,
"E" => -2.926848342 ,
"F" => -1.56713884 ,
"G" => -0.535695792 ,
"H" => -0.703989708 ,
"I" => -0.766394216 ,
"K" => -0.03993628 ,
"L" => -0.332775496 ,
"M" => -0.373174827 ,
"N" => -0.751084826 ,
"P" => 0.415181993 ,
"Q" => 0.458204008 ,
"R" => 1.377616087 ,
"S" => 1.057199568 ,
"T" => 0.22541374 ,
"V" => -0.641027267 ,
"W" => 0.335720002 ,
"Y" => -0.490680676 ,
},
aa24 => {
"A" => 0.469872316 ,
"C" => 0.796200472 ,
"D" => -2.290583295 ,
"E" => -2.511810843 ,
"F" => -0.081712013 ,
"G" => 0.049266708 ,
"H" => 0.674521915 ,
"I" => -0.659479012 ,
"K" => -0.380973198 ,
"L" => 0.131171604 ,
"M" => -1.180529749 ,
"N" => 0.326917686 ,
"P" => 0.535476226 ,
"Q" => -0.278761586 ,
"R" => 0.999104463 ,
"S" => 1.028630416 ,
"T" => -0.452658165 ,
"V" => -0.833672345 ,
"W" => -1.249242499 ,
"Y" => -1.490680676 ,
},
aa25 => {
"A" => 0.469872316 ,
"C" => 0.796200472 ,
"D" => -1.9686552 ,
"E" => -2.189882748 ,
"F" => -0.188627217 ,
"G" => -0.182058838 ,
"H" => -0.218562881 ,
"I" => -0.466833934 ,
"K" => -0.117938792 ,
"L" => -0.283865895 ,
"M" => -1.180529749 ,
"N" => 0.078990173 ,
"P" => 0.845816347 ,
"Q" => 0.043166509 ,
"R" => 0.806459385 ,
"S" => 0.876627322 ,
"T" => 0.164013195 ,
"V" => -0.181595649 ,
"W" => -0.249242499 ,
"Y" => -1.490680676 ,
},
aa26 => {
"A" => 0.682866039 ,
"C" => 1.118128566 ,
"D" => -2.290583295 ,
"E" => -2.926848342 ,
"F" => -0.889066935 ,
"G" => -0.061764604 ,
"H" => 0.674521915 ,
"I" => -2.466833934 ,
"K" => -0.117938792 ,
"L" => 0.201560932 ,
"M" => -0.595567248 ,
"N" => -0.461578209 ,
"P" => 0.061545038 ,
"Q" => 0.043166509 ,
"R" => 0.925103882 ,
"S" => 1.139661728 ,
"T" => 0.032768662 ,
"V" => -0.833672345 ,
"W" => -0.249242499 ,
"Y" => -0.320755675 ,
},
aa27 => {
"A" => 0.65818328 ,
"C" => 0.800086865 ,
"D" => -1.964768806 ,
"E" => -1.700569528 ,

```

```

        "F" => -0.184740824 ,
        "G" => -0.242302782 ,
        "H" => -0.02203141 ,
        "I" => -0.462947541 ,
        "K" => -0.036049886 ,
        "L" => -0.328889103 ,
        "M" => -9.820499546 ,
        "N" => -0.216683716 ,
        "P" => 0.419068386 ,
        "Q" => -0.859837694 ,
        "R" => 0.810345779 ,
        "S" => 0.97362312 ,
        "T" => 0.399225134 ,
        "V" => 0.117746628 ,
        "W" => 0.076571989 ,
        "Y" => -1.486794283 ,
    },
    aa28 => {
        "A" => 0.502293794 ,
        "C" => 0.381162972 ,
        "D" => -1.290583295 ,
        "E" => -0.926848342 ,
        "F" => 0.110933065 ,
        "G" => -0.246189175 ,
        "H" => -0.025917803 ,
        "I" => -1.144905839 ,
        "K" => -0.287863793 ,
        "L" => 0.018696875 ,
        "M" => -1.180529749 ,
        "N" => -0.751084826 ,
        "P" => 0.698974959 ,
        "Q" => -0.163284369 ,
        "R" => 0.677176368 ,
        "S" => 0.742326231 ,
        "T" => 0.164013195 ,
        "V" => 0.059412451 ,
        "W" => -0.249242499 ,
        "Y" => -0.490680676 ,
    },
    aa29 => {
        "A" => 0.565029549 ,
        "C" => 0.118128566 ,
        "D" => -1.483228372 ,
        "E" => -1.189882748 ,
        "F" => -0.188627217 ,
        "G" => -0.535695792 ,
        "H" => 0.880972792 ,
        "I" => -1.659479012 ,
        "K" => -0.117938792 ,
        "L" => -0.667194535 ,
        "M" => -0.595567248 ,
        "N" => -0.921009827 ,
        "P" => 0.415181993 ,
        "Q" => -0.404292468 ,
        "R" => 0.76463921 ,
        "S" => 0.99948407 ,
        "T" => 0.684845358 ,
        "V" => 0.265863328 ,
        "W" => -0.664279998 ,
        "Y" => -0.490680676 ,
    },
    aa30 => {
        "A" => 0.654296887 ,
        "C" => 0.603555394 ,
        "D" => -0.9686552 ,
        "E" => -1.926848342 ,
        "F" => -0.081712013 ,
        "G" => -0.313303371 ,
        "H" => -0.440955302 ,
        "I" => -0.881871434 ,
        "K" => -0.380973198 ,
        "L" => -0.190756491 ,
        "M" => 0.141398346 ,
        "N" => -1.336047327 ,
        "P" => 0.351051655 ,
        "Q" => 0.384203427 ,
        "R" => -0.000895537 ,
        "S" => 1.057199568 ,
        "T" => 0.22541374 ,
        "V" => 0.113860235 ,
        "W" => 0.072685596 ,
        "Y" => -0.031249058 ,
    },
    },
    my %spodd = (
        aa2 => {
            "A" => 2.159645157 ,
            "C" => -8.384951785 ,
            "D" => -1.242916861 ,
            "E" => 0.038355931 ,

```

```

        "F" => -2.426363002 ,
        "G" => 0.564438061 ,
        "H" => -1.56321387 ,
        "I" => -2.419167501 ,
        "K" => 0.75980264 ,
        "L" => -1.406124463 ,
        "M" => -0.717825816 ,
        "N" => -0.458305894 ,
        "P" => -1.64567603 ,
        "Q" => -1.401020154 ,
        "R" => -0.15967998 ,
        "S" => -0.452082123 ,
        "T" => -0.726919827 ,
        "V" => -0.090860493 ,
        "W" => -8.430394756 ,
        "Y" => -9.671832933 ,
    },
    aa3 => {
        "A" => -0.831808306 ,
        "C" => 0.253511751 ,
        "D" => -10.47712821 ,
        "E" => -10.69835575 ,
        "F" => 0.816171857 ,
        "G" => -1.248309515 ,
        "H" => -0.246678429 ,
        "I" => -0.424560155 ,
        "K" => 1.213841605 ,
        "L" => -0.485517699 ,
        "M" => 0.736213148 ,
        "N" => 0.536301452 ,
        "P" => -1.651068684 ,
        "Q" => -0.599057886 ,
        "R" => 1.041378243 ,
        "S" => 1.199637509 ,
        "T" => 0.621324474 ,
        "V" => 0.064211525 ,
        "W" => -8.43578741 ,
        "Y" => -1.033369397 ,
    },
    aa4 => {
        "A" => -0.449996868 ,
        "C" => 0.57022816 ,
        "D" => -2.838483701 ,
        "E" => -2.474748749 ,
        "F" => 0.022464277 ,
        "G" => -2.6685587 ,
        "H" => 0.233536712 ,
        "I" => -0.42977184 ,
        "K" => 1.022216795 ,
        "L" => 0.687607857 ,
        "M" => -0.143467655 ,
        "N" => -1.146982139 ,
        "P" => -0.65628037 ,
        "Q" => -0.604269571 ,
        "R" => 0.588707581 ,
        "S" => 1.263138574 ,
        "T" => 0.340478346 ,
        "V" => -0.101464833 ,
        "W" => -0.212180405 ,
        "Y" => -0.231226161 ,
    },
    aa5 => {
        "A" => 0.002119473 ,
        "C" => 0.264317291 ,
        "D" => -10.46632266 ,
        "E" => -10.68755021 ,
        "F" => 0.48594048 ,
        "G" => -0.915575879 ,
        "H" => 0.764127112 ,
        "I" => 0.171207887 ,
        "K" => 1.180253027 ,
        "L" => 0.184250924 ,
        "M" => -0.390484834 ,
        "N" => 0.354461914 ,
        "P" => -0.055300643 ,
        "Q" => -0.225682266 ,
        "R" => 0.232756029 ,
        "S" => 1.062344411 ,
        "T" => -0.084077019 ,
        "V" => -0.472470729 ,
        "W" => -8.424981869 ,
        "Y" => -1.437601356 ,
    },
    aa6 => {
        "A" => -0.139822945 ,
        "C" => 0.838581907 ,
        "D" => -2.83316436 ,
        "E" => -3.054391908 ,
        "F" => 0.738277001 ,
        "G" => -1.663239358 ,
    },

```

```

"H" => 0.016463632 ,
"I" => 0.078047842 ,
"K" => 0.257017983 ,
"L" => 0.476115809 ,
"M" => -0.401182719 ,
"N" => -0.463590893 ,
"P" => -1.065998528 ,
"Q" => -0.821342651 ,
"R" => -0.543476602 ,
"S" => 1.334046257 ,
"T" => -0.317167326 ,
"V" => 1.138319762 ,
"W" => -8.435679754 ,
"Y" => -1.033261741 ,
},
aa7 => {
  "A" => 0.117054848 ,
  "C" => -0.741041567 ,
  "D" => -10.47168152 ,
  "E" => -10.69290907 ,
  "F" => 0.82161854 ,
  "G" => -1.657900332 ,
  "H" => -1.563159842 ,
  "I" => 0.519485983 ,
  "K" => -0.365674213 ,
  "L" => 1.129982466 ,
  "M" => -0.132809287 ,
  "N" => 0.126710635 ,
  "P" => -0.83826708 ,
  "Q" => -1.400966125 ,
  "R" => -0.86006567 ,
  "S" => 0.705513182 ,
  "T" => 0.495526623 ,
  "V" => 0.406693195 ,
  "W" => 0.798477963 ,
  "Y" => -1.027922715 ,
},
aa8 => {
  "A" => 0.721126172 ,
  "C" => -0.741041567 ,
  "D" => -10.47168152 ,
  "E" => -2.049052882 ,
  "F" => 0.966008449 ,
  "G" => -1.43550791 ,
  "H" => -1.563159842 ,
  "I" => 0.455355646 ,
  "K" => -1.240143331 ,
  "L" => 1.129982466 ,
  "M" => 0.452153214 ,
  "N" => -1.873289365 ,
  "P" => -1.060659501 ,
  "Q" => -1.816003625 ,
  "R" => -0.690140669 ,
  "S" => 0.754422783 ,
  "T" => 0.010099796 ,
  "V" => 0.576618196 ,
  "W" => -8.430340728 ,
  "Y" => -0.70599462 ,
},
aa9 => {
  "A" => 0.682652024 ,
  "C" => 1.066313355 ,
  "D" => -2.827825333 ,
  "E" => -10.69290907 ,
  "F" => 0.966008449 ,
  "G" => -1.242862832 ,
  "H" => -0.978197341 ,
  "I" => 0.083386868 ,
  "K" => -0.240143331 ,
  "L" => 1.500820161 ,
  "M" => -0.132809287 ,
  "N" => -1.458251866 ,
  "P" => -1.323693907 ,
  "Q" => -10.04482232 ,
  "R" => -2.275103169 ,
  "S" => 0.602419689 ,
  "T" => -0.574862705 ,
  "V" => 0.522170412 ,
  "W" => -0.786484538 ,
  "Y" => -2.027922715 ,
},
aa10 => {
  "A" => 0.774737541 ,
  "C" => 1.597023191 ,
  "D" => -10.45554486 ,
  "E" => -10.67677241 ,
  "F" => 1.113389646 ,
  "G" => -1.641763668 ,
  "H" => -9.190879368 ,
  "I" => -0.081048714 ,

```

```

        "K" => -2.224006668 ,
        "L" => 1.536065647 ,
        "M" => -0.116672623 ,
        "N" => -1.442115203 ,
        "P" => -1.307557244 ,
        "Q" => -10.02868565 ,
        "R" => -2.843929007 ,
        "S" => 0.671023773 ,
        "T" => -0.421222518 ,
        "V" => 0.882261477 ,
        "W" => -8.414204064 ,
        "Y" => -1.426823551 ,
    },
    aa11 => {
        "A" => 0.726430801 ,
        "C" => 0.586191157 ,
        "D" => -2.822520704 ,
        "E" => -3.043748253 ,
        "F" => 1.748920657 ,
        "G" => -3.237558204 ,
        "H" => -9.201711403 ,
        "I" => -0.298331626 ,
        "K" => -2.234838702 ,
        "L" => 1.652345531 ,
        "M" => 0.094887763 ,
        "N" => -2.452947237 ,
        "P" => -1.318389278 ,
        "Q" => -10.03951769 ,
        "R" => -1.854761041 ,
        "S" => 0.245154239 ,
        "T" => -0.432054552 ,
        "V" => 0.827035323 ,
        "W" => -8.425036099 ,
        "Y" => -9.666474276 ,
    },
    aa12 => {
        "A" => 0.512199922 ,
        "C" => 1.838528078 ,
        "D" => -10.47707438 ,
        "E" => -2.054445737 ,
        "F" => 1.268737889 ,
        "G" => -2.248255688 ,
        "H" => -1.568552697 ,
        "I" => 0.382848594 ,
        "K" => -2.830498687 ,
        "L" => 1.58853671 ,
        "M" => -0.723164643 ,
        "N" => -1.878682221 ,
        "P" => -0.843659935 ,
        "Q" => -1.406358981 ,
        "R" => -2.280496025 ,
        "S" => -0.070397827 ,
        "T" => -0.732258654 ,
        "V" => 0.723228434 ,
        "W" => -8.435733583 ,
        "Y" => -0.225960648 ,
    },
    aa13 => {
        "A" => 1.027679455 ,
        "C" => 0.580778473 ,
        "D" => -10.47178958 ,
        "E" => -10.69301713 ,
        "F" => 1.097144928 ,
        "G" => -0.155508046 ,
        "H" => -9.207124086 ,
        "I" => -0.30374431 ,
        "K" => -10.46907008 ,
        "L" => 1.359356257 ,
        "M" => 0.089475079 ,
        "N" => -1.458359921 ,
        "P" => -1.323801962 ,
        "Q" => -2.40107418 ,
        "R" => -1.860173725 ,
        "S" => -0.345220946 ,
        "T" => -0.57497076 ,
        "V" => 1.179174644 ,
        "W" => 0.535335502 ,
        "Y" => -2.02803077 ,
    },
    aa14 => {
        "A" => 0.996544513 ,
        "C" => 1.066097253 ,
        "D" => -10.47189762 ,
        "E" => -10.69312517 ,
        "F" => 1.380829847 ,
        "G" => -2.658116433 ,
        "H" => -9.207232133 ,
        "I" => 0.803062847 ,
        "K" => -10.46917812 ,
        "L" => 1.421532488 ,
    },

```

```

        "M" => 0.741443729 ,
        "N" => -10.10232416 ,
        "P" => -2.645838104 ,
        "Q" => -10.04503842 ,
        "R" => -2.860281772 ,
        "S" => 0.017241087 ,
        "T" => -1.089651979 ,
        "V" => 1.069442106 ,
        "W" => -0.786700639 ,
        "Y" => -1.028138817 ,
    },
    aa15 => {
        "A" => 1.420018023 ,
        "C" => 0.860002961 ,
        "D" => -10.4555995 ,
        "E" => -10.67682704 ,
        "F" => 1.290212772 ,
        "G" => -2.226780805 ,
        "H" => -9.190934004 ,
        "I" => -0.08110335 ,
        "K" => -10.45287999 ,
        "L" => 1.265363421 ,
        "M" => -0.379761665 ,
        "N" => -10.08602603 ,
        "P" => -2.629539975 ,
        "Q" => -1.384884098 ,
        "R" => -10.48783983 ,
        "S" => -0.048922944 ,
        "T" => -0.07335385 ,
        "V" => 1.085740235 ,
        "W" => -0.18544001 ,
        "Y" => -0.426878187 ,
    },
    aa16 => {
        "A" => 0.800322929 ,
        "C" => 1.071509531 ,
        "D" => -2.237666656 ,
        "E" => -2.458894205 ,
        "F" => 1.333774705 ,
        "G" => -1.430311734 ,
        "H" => -9.201819856 ,
        "I" => 0.644976393 ,
        "K" => -10.46376585 ,
        "L" => 0.977637365 ,
        "M" => 1.194314984 ,
        "N" => -10.09691188 ,
        "P" => -1.318497731 ,
        "Q" => -10.03962614 ,
        "R" => -2.269906993 ,
        "S" => 0.312159982 ,
        "T" => -0.084239702 ,
        "V" => 1.253191625 ,
        "W" => -8.425144552 ,
        "Y" => -9.666582729 ,
    },
    aa17 => {
        "A" => 0.9649437 ,
        "C" => 1.42877538 ,
        "D" => -10.47178958 ,
        "E" => -2.464198436 ,
        "F" => 1.03301459 ,
        "G" => -1.073045886 ,
        "H" => -9.207124086 ,
        "I" => 0.580778473 ,
        "K" => -2.240251386 ,
        "L" => 1.079248338 ,
        "M" => 0.089475079 ,
        "N" => -10.10221611 ,
        "P" => -2.060767556 ,
        "Q" => -10.04493037 ,
        "R" => -2.860173725 ,
        "S" => 0.602311635 ,
        "T" => 0.190563987 ,
        "V" => 0.991547641 ,
        "W" => -0.786592593 ,
        "Y" => -2.02803077 ,
    },
    aa18 => {
        "A" => 0.825303981 ,
        "C" => 1.423436752 ,
        "D" => -2.833272016 ,
        "E" => -1.469537063 ,
        "F" => 0.960561767 ,
        "G" => -0.160846674 ,
        "H" => -9.212462714 ,
        "I" => -0.202167733 ,
        "K" => -2.830552514 ,
        "L" => 0.910410978 ,
        "M" => 0.446706531 ,
        "N" => -1.141770454 ,
    },

```



```

        "P" => -1.651068684 ,
        "Q" => -10.050269 ,
        "R" => -10.50936854 ,
        "S" => 0.842085505 ,
        "T" => 0.267687519 ,
        "V" => 0.770480322 ,
        "W" => -0.20696872 ,
        "Y" => -1.033369397 ,
    },
    aa19 => {
        "A" => 0.825411637 ,
        "C" => 1.575547501 ,
        "D" => -2.248201859 ,
        "E" => -2.469429408 ,
        "F" => 0.655814841 ,
        "G" => -0.663239358 ,
        "H" => -0.246570774 ,
        "I" => 0.312513095 ,
        "K" => -1.245482358 ,
        "L" => 0.881372288 ,
        "M" => 0.084244108 ,
        "N" => -1.141662798 ,
        "P" => -0.650961029 ,
        "Q" => -0.59895023 ,
        "R" => -2.865404697 ,
        "S" => 0.649548083 ,
        "T" => 0.185333015 ,
        "V" => 0.460247857 ,
        "W" => -8.435679754 ,
        "Y" => -1.448299241 ,
    },
    aa20 => {
        "A" => 1.228086161 ,
        "C" => 0.580886528 ,
        "D" => -2.242862832 ,
        "E" => -1.727124787 ,
        "F" => 0.743616028 ,
        "G" => -0.155399991 ,
        "H" => 0.021802659 ,
        "I" => -0.682147878 ,
        "K" => -2.825105832 ,
        "L" => 0.27200147 ,
        "M" => -1.132809287 ,
        "N" => -0.136323771 ,
        "P" => -0.323693907 ,
        "Q" => -1.816003625 ,
        "R" => -10.50392186 ,
        "S" => 1.056985553 ,
        "T" => 0.273134202 ,
        "V" => 0.821730694 ,
        "W" => -8.430340728 ,
        "Y" => -2.027922715 ,
    },
    aa21 => {
        "A" => 0.991367759 ,
        "C" => 1.712997196 ,
        "D" => -1.025863267 ,
        "E" => -0.247090815 ,
        "F" => 0.153260672 ,
        "G" => -0.248255688 ,
        "H" => 0.016409803 ,
        "I" => -0.202113906 ,
        "K" => -10.47435488 ,
        "L" => 0.220804925 ,
        "M" => -0.138202143 ,
        "N" => -0.878682221 ,
        "P" => -0.843659935 ,
        "Q" => -0.599004059 ,
        "R" => -1.865458526 ,
        "S" => 0.749029927 ,
        "T" => 0.490133768 ,
        "V" => 0.571225341 ,
        "W" => -8.435733583 ,
        "Y" => -1.44835307 ,
    },
    aa22 => {
        "A" => 0.888805045 ,
        "C" => 1.707731684 ,
        "D" => -0.516555606 ,
        "E" => -0.737783155 ,
        "F" => -0.436967341 ,
        "G" => -0.553081482 ,
        "H" => -0.988855709 ,
        "I" => -0.42977184 ,
        "K" => -1.513836105 ,
        "L" => 0.305737222 ,
        "M" => 0.078924766 ,
        "N" => -0.661555312 ,
        "P" => 0.044159348 ,
        "Q" => -0.411624493 ,
    },

```

```

        "R" => -1.285761537 ,
        "S" => 1.046327185 ,
        "T" => -0.000558572 ,
        "V" => 0.396034827 ,
        "W" => -0.797142906 ,
        "Y" => -2.038581083 ,
    },
    aa23 => {
        "A" => 1.195780086 ,
        "C" => 0.838581907 ,
        "D" => -1.248201859 ,
        "E" => -1.054391908 ,
        "F" => -0.261722999 ,
        "G" => 0.559153063 ,
        "H" => 0.431501132 ,
        "I" => -0.202060077 ,
        "K" => -1.245482358 ,
        "L" => -0.341020133 ,
        "M" => -0.401182719 ,
        "N" => -0.656235971 ,
        "P" => -1.065998528 ,
        "Q" => 0.178657349 ,
        "R" => -1.280442196 ,
        "S" => 0.929656002 ,
        "T" => 0.004760769 ,
        "V" => 0.275823286 ,
        "W" => -8.435679754 ,
        "Y" => -2.033261741 ,
    },
    aa24 => {
        "A" => 0.825411637 ,
        "C" => 0.253619406 ,
        "D" => -0.132724642 ,
        "E" => -0.732463813 ,
        "F" => -0.431648 ,
        "G" => -0.547762141 ,
        "H" => -0.246570774 ,
        "I" => -0.102524404 ,
        "K" => -1.023089936 ,
        "L" => 0.074017366 ,
        "M" => 0.276889186 ,
        "N" => -0.463590893 ,
        "P" => 0.156393893 ,
        "Q" => 0.763619849 ,
        "R" => -0.280442196 ,
        "S" => 0.88658728 ,
        "T" => -0.902129826 ,
        "V" => -0.013683331 ,
        "W" => -8.435679754 ,
        "Y" => -0.711333647 ,
    },
    aa25 => {
        "A" => 0.825411637 ,
        "C" => 0.575547501 ,
        "D" => -0.511236265 ,
        "E" => -0.884466907 ,
        "F" => 0.475242595 ,
        "G" => -0.248201859 ,
        "H" => 0.238856054 ,
        "I" => -0.687486905 ,
        "K" => -1.023089936 ,
        "L" => -0.563412555 ,
        "M" => -1.138148314 ,
        "N" => -0.004159274 ,
        "P" => 0.596966485 ,
        "Q" => 0.915622943 ,
        "R" => -0.695479695 ,
        "S" => 0.796389471 ,
        "T" => 0.004760769 ,
        "V" => -0.483168614 ,
        "W" => -8.435679754 ,
        "Y" => 0.13666326 ,
    },
    aa26 => {
        "A" => 0.921918279 ,
        "C" => 1.707678053 ,
        "D" => -0.516609237 ,
        "E" => -0.05976488 ,
        "F" => 0.562979028 ,
        "G" => -0.16611199 ,
        "H" => -0.251943746 ,
        "I" => -0.207433049 ,
        "K" => -0.25085533 ,
        "L" => -0.651247687 ,
        "M" => -1.728483786 ,
        "N" => -0.14703577 ,
        "P" => -0.334405906 ,
        "Q" => -0.241753123 ,
        "R" => -0.548849574 ,
        "S" => 0.421782689 ,
    },

```

```

        "T" => 0.340424715 ,
        "V" => 0.13294679 ,
        "W" => -8.441052726 ,
        "Y" => -1.453672213 ,
    },
    aa27 => {
        "A" => 0.374696399 ,
        "C" => 0.575493672 ,
        "D" => 0.559099234 ,
        "E" => -0.354006019 ,
        "F" => -0.431701829 ,
        "G" => 0.395600502 ,
        "H" => 0.016409803 ,
        "I" => -0.102578233 ,
        "K" => -1.023143765 ,
        "L" => -0.485463872 ,
        "M" => -9.367020833 ,
        "N" => -0.29371972 ,
        "P" => 0.156340065 ,
        "Q" => 0.400995941 ,
        "R" => -2.280496025 ,
        "S" => 0.971422349 ,
        "T" => -0.094828733 ,
        "V" => 0.401300339 ,
        "W" => -8.435733583 ,
        "Y" => -0.44835307 ,
    },
    aa28 => {
        "A" => -0.215531614 ,
        "C" => 1.055654987 ,
        "D" => -0.379052082 ,
        "E" => -0.359271531 ,
        "F" => 0.370387581 ,
        "G" => 0.270040756 ,
        "H" => -1.57381821 ,
        "I" => -0.207379419 ,
        "K" => -0.665839198 ,
        "L" => -0.416728803 ,
        "M" => -1.728430156 ,
        "N" => 0.618552607 ,
        "P" => 0.250610226 ,
        "Q" => 0.588375507 ,
        "R" => -1.548795943 ,
        "S" => 1.006798821 ,
        "T" => -0.322486667 ,
        "V" => -0.188927674 ,
        "W" => 0.524785189 ,
        "Y" => 0.546381418 ,
    },
    aa29 => {
        "A" => -0.215531614 ,
        "C" => 2.248300065 ,
        "D" => 0.805372489 ,
        "E" => -0.889786248 ,
        "F" => 0.147995159 ,
        "G" => 0.138796222 ,
        "H" => 0.233536712 ,
        "I" => -1.207379419 ,
        "K" => -0.665839198 ,
        "L" => -0.346339475 ,
        "M" => -1.143467655 ,
        "N" => 0.923407189 ,
        "P" => 0.250610226 ,
        "Q" => 0.980692929 ,
        "R" => -1.063369116 ,
        "S" => -0.163126181 ,
        "T" => -0.207009449 ,
        "V" => -0.381572752 ,
        "W" => -0.212180405 ,
        "Y" => 0.131343919 ,
    },
    aa30 => {
        "A" => -0.008632241 ,
        "C" => 1.838528078 ,
        "D" => 0.62621343 ,
        "E" => -0.732517642 ,
        "F" => -0.624346907 ,
        "G" => 0.559099234 ,
        "H" => 0.238802225 ,
        "I" => 0.575493672 ,
        "K" => -1.023143765 ,
        "L" => -0.485463872 ,
        "M" => -9.367020833 ,
        "N" => 0.443245874 ,
        "P" => 0.255875738 ,
        "Q" => 0.500531615 ,
        "R" => -0.406026907 ,
        "S" => -0.350505746 ,
        "T" => 0.345743858 ,
        "V" => -0.724230543 ,
    },

```

```

        "W" => -8.435733583 ,
        "Y" => 0.136609431 ,
    },
);
my %otodd = (
    aa2 => {
        "A" => 2.035438239 ,
        "C" => -0.011692001 ,
        "D" => -9.742331957 ,
        "E" => 0.380736403 ,
        "F" => -9.340815597 ,
        "G" => 1.130342924 ,
        "H" => -0.833810275 ,
        "I" => 1.117591016 ,
        "K" => -9.739612455 ,
        "L" => -10.64250515 ,
        "M" => 0.011577779 ,
        "N" => -0.143939799 ,
        "P" => -1.331309935 ,
        "Q" => -0.671616559 ,
        "R" => -2.130716104 ,
        "S" => 0.483772349 ,
        "T" => -2.167441233 ,
        "V" => -10.09277593 ,
        "W" => -7.700991161 ,
        "Y" => 0.023354946 ,
    },
    aa3 => {
        "A" => -0.194280698 ,
        "C" => 2.328444671 ,
        "D" => 0.241660905 ,
        "E" => -0.131569737 ,
        "F" => -9.32260702 ,
        "G" => 0.592158152 ,
        "H" => 0.184398301 ,
        "I" => -0.671555329 ,
        "K" => -0.492585188 ,
        "L" => -0.810515385 ,
        "M" => 1.35171445 ,
        "N" => 0.874268778 ,
        "P" => -9.541920049 ,
        "Q" => -0.331479888 ,
        "R" => -0.527545027 ,
        "S" => 0.843017844 ,
        "T" => 0.850767343 ,
        "V" => -10.07456735 ,
        "W" => -7.682782585 ,
        "Y" => 0.719635428 ,
    },
    aa4 => {
        "A" => 0.363988599 ,
        "C" => 2.149748374 ,
        "D" => -0.299605472 ,
        "E" => -1.006259847 ,
        "F" => -1.705444034 ,
        "G" => 0.478002107 ,
        "H" => 1.157705098 ,
        "I" => 0.524143889 ,
        "K" => -1.519278392 ,
        "L" => 0.162791411 ,
        "M" => 0.810448074 ,
        "N" => 0.262613073 ,
        "P" => -0.924757063 ,
        "Q" => -0.095138685 ,
        "R" => -0.817272636 ,
        "S" => 0.268836845 ,
        "T" => -0.17592586 ,
        "V" => -10.10126056 ,
        "W" => 0.519342902 ,
        "Y" => 0.692942224 ,
    },
    aa5 => {
        "A" => -1.203143267 ,
        "C" => 1.997654007 ,
        "D" => 0.232798336 ,
        "E" => 0.274605193 ,
        "F" => -9.331469589 ,
        "G" => 0.19627246 ,
        "H" => 0.497463827 ,
        "I" => 0.19405122 ,
        "K" => -1.086410258 ,
        "L" => -0.819377954 ,
        "M" => 1.020923786 ,
        "N" => 1.980883426 ,
        "P" => -0.584998333 ,
        "Q" => -0.340342456 ,
        "R" => -0.799442001 ,
        "S" => 0.493118357 ,
        "T" => -0.158095226 ,
        "V" => -1.854611231 ,
    },

```

```

        "W" => -7.691645153 ,
        "Y" => 0.518127781 ,
    },
    aa6 => {
        "A" => -0.475165336 ,
        "C" => 0.988666344 ,
        "D" => 0.071807579 ,
        "E" => 0.002583124 ,
        "F" => -1.696601063 ,
        "G" => -0.053723303 ,
        "H" => 0.166548069 ,
        "I" => -0.46701314 ,
        "K" => -0.288042999 ,
        "L" => -0.538859 ,
        "M" => 2.011936123 ,
        "N" => 0.271456045 ,
        "P" => -9.559770281 ,
        "Q" => 0.328741786 ,
        "R" => 0.039567242 ,
        "S" => 0.577240098 ,
        "T" => 0.154845206 ,
        "V" => -10.09241758 ,
        "W" => 0.528185874 ,
        "Y" => 1.701785196 ,
    },
    aa7 => {
        "A" => -0.337661812 ,
        "C" => 0.573628845 ,
        "D" => 0.071807579 ,
        "E" => -0.319344971 ,
        "F" => -1.696601063 ,
        "G" => 0.486845078 ,
        "H" => 0.488476164 ,
        "I" => 0.185063557 ,
        "K" => -0.095397921 ,
        "L" => -0.998290619 ,
        "M" => 1.181861125 ,
        "N" => 0.856418546 ,
        "P" => -9.559770281 ,
        "Q" => 0.328741786 ,
        "R" => -0.323002837 ,
        "S" => 0.38459502 ,
        "T" => 0.417879612 ,
        "V" => -10.09241758 ,
        "W" => -0.056776627 ,
        "Y" => 1.701785196 ,
    },
    aa8 => {
        "A" => -0.466177672 ,
        "C" => 2.319582102 ,
        "D" => -0.089129759 ,
        "E" => -0.140432306 ,
        "F" => -0.102650899 ,
        "G" => -0.182239163 ,
        "H" => 0.175535732 ,
        "I" => 0.904544603 ,
        "K" => -1.501447757 ,
        "L" => -0.66737486 ,
        "M" => 1.828278709 ,
        "N" => 1.08779863 ,
        "P" => -9.550782618 ,
        "Q" => 0.50765445 ,
        "R" => -1.121370096 ,
        "S" => -0.413772239 ,
        "T" => 0.426867275 ,
        "V" => -10.08342992 ,
        "W" => -0.047788964 ,
        "Y" => 0.880697861 ,
    },
    aa9 => {
        "A" => -0.627168429 ,
        "C" => 2.448097963 ,
        "D" => 0.602322296 ,
        "E" => -0.511990049 ,
        "F" => -9.340457252 ,
        "G" => 0.393735674 ,
        "H" => 1.336473071 ,
        "I" => 0.632522534 ,
        "K" => -1.095397921 ,
        "L" => -1.190935697 ,
        "M" => 1.333864218 ,
        "N" => 0.271456045 ,
        "P" => -1.33095159 ,
        "Q" => -0.671258214 ,
        "R" => 0.570081959 ,
        "S" => 0.664702939 ,
        "T" => -0.167082889 ,
        "V" => -10.09241758 ,
        "W" => -0.056776627 ,
        "Y" => 0.286747697 ,
    },

```

```

),
aa10 => {
    "A" => 0.193974551 ,
    "C" => 2.439165944 ,
    "D" => -0.299694519 ,
    "E" => -0.743314489 ,
    "F" => -0.705533081 ,
    "G" => -0.714732018 ,
    "H" => -0.257421449 ,
    "I" => 0.886624921 ,
    "K" => -2.10432994 ,
    "L" => -0.422260137 ,
    "M" => 1.9098947 ,
    "N" => 0.432449028 ,
    "P" => -1.339883609 ,
    "Q" => -0.680190233 ,
    "R" => 0.030635224 ,
    "S" => 0.56830808 ,
    "T" => -0.006089906 ,
    "V" => -10.1013496 ,
    "W" => -7.709564835 ,
    "Y" => 1.152284796 ,
},
aa11 => {
    "A" => -0.627078829 ,
    "C" => 1.988755944 ,
    "D" => -0.2906729 ,
    "E" => -0.319255371 ,
    "F" => -1.696511463 ,
    "G" => -0.053633703 ,
    "H" => 1.488565764 ,
    "I" => -0.274278462 ,
    "K" => -1.51034582 ,
    "L" => -1.413238518 ,
    "M" => 1.596988224 ,
    "N" => 1.441470646 ,
    "P" => -0.33086199 ,
    "Q" => -0.086206114 ,
    "R" => 0.039656842 ,
    "S" => 0.664792539 ,
    "T" => 0.154934806 ,
    "V" => -10.09232798 ,
    "W" => -0.056687027 ,
    "Y" => 0.871799798 ,
},
aa12 => {
    "A" => -0.337661812 ,
    "C" => 2.158591345 ,
    "D" => -0.776189327 ,
    "E" => 0.002583124 ,
    "F" => -1.696601063 ,
    "G" => 0.071807579 ,
    "H" => 1.75151057 ,
    "I" => 0.89555694 ,
    "K" => -9.739254111 ,
    "L" => -0.828365617 ,
    "M" => 1.819291045 ,
    "N" => 0.730887664 ,
    "P" => -0.33095159 ,
    "Q" => -0.671258214 ,
    "R" => 0.191570336 ,
    "S" => -0.100831807 ,
    "T" => -0.845154794 ,
    "V" => -10.09241758 ,
    "W" => 0.528185874 ,
    "Y" => 1.023713291 ,
},
aa13 => {
    "A" => 0.109886765 ,
    "C" => 2.689195662 ,
    "D" => -0.2906729 ,
    "E" => 0.140176248 ,
    "F" => -1.696511463 ,
    "G" => -0.053633703 ,
    "H" => 0.75160017 ,
    "I" => 0.632612134 ,
    "K" => -9.739164511 ,
    "L" => -0.998201019 ,
    "M" => 1.012025723 ,
    "N" => 0.730977264 ,
    "P" => -1.33086199 ,
    "Q" => -0.34924052 ,
    "R" => -0.545305658 ,
    "S" => 0.577329698 ,
    "T" => -0.359638367 ,
    "V" => -1.448471795 ,
    "W" => -0.056687027 ,
    "Y" => 1.161306415 ,
},
aa14 => {

```

```

"A" => 0.118694674 ,
"C" => 1.319491948 ,
"D" => 0.370211705 ,
"E" => 0.148984157 ,
"F" => -9.331559743 ,
"G" => -0.334332411 ,
"H" => 0.497373673 ,
"I" => 0.641420043 ,
"K" => -1.086500412 ,
"L" => -0.529961491 ,
"M" => 1.480265251 ,
"N" => 1.602281649 ,
"P" => -0.907016582 ,
"Q" => 0.337639294 ,
"R" => -1.12146025 ,
"S" => -0.828899892 ,
"T" => 0.163742715 ,
"V" => -10.08352008 ,
"W" => -7.691735308 ,
"Y" => 1.170114324 ,
},
aa15 => {
"A" => 0.019249155 ,
"C" => 1.582616508 ,
"D" => -0.281774837 ,
"E" => -0.310357307 ,
"F" => -1.687613399 ,
"G" => -0.334242257 ,
"H" => 0.982890655 ,
"I" => 1.258181557 ,
"K" => -2.086410258 ,
"L" => -1.181948033 ,
"M" => 1.605886287 ,
"N" => -0.134593791 ,
"P" => -9.550782618 ,
"Q" => -1.077308051 ,
"R" => 0.7855205 ,
"S" => 0.04565938 ,
"T" => 0.301336393 ,
"V" => -10.08342992 ,
"W" => -0.047788964 ,
"Y" => 1.617663455 ,
},
aa16 => {
"A" => -0.475165336 ,
"C" => 2.89555694 ,
"D" => -0.513154922 ,
"E" => -0.73438247 ,
"F" => -9.340457252 ,
"G" => -0.191226827 ,
"H" => -0.24848943 ,
"I" => 0.53298686 ,
"K" => -9.739254111 ,
"L" => -0.828365617 ,
"M" => 2.099398964 ,
"N" => 1.271456045 ,
"P" => -1.915914091 ,
"Q" => 0.913704286 ,
"R" => 0.329073859 ,
"S" => -0.2528349 ,
"T" => -0.167082889 ,
"V" => -2.448561395 ,
"W" => 1.528185874 ,
"Y" => 0.023713291 ,
},
aa17 => {
"A" => -0.627168429 ,
"C" => 2.448097963 ,
"D" => -0.513154922 ,
"E" => -0.511990049 ,
"F" => -9.340457252 ,
"G" => 0.071807579 ,
"H" => 0.488476164 ,
"I" => 0.973559452 ,
"K" => -1.51043542 ,
"L" => -1.413328118 ,
"M" => 2.181861125 ,
"N" => 0.730887664 ,
"P" => -9.559770281 ,
"Q" => 0.136096708 ,
"R" => 0.676997163 ,
"S" => 0.036671717 ,
"T" => -0.845154794 ,
"V" => -1.863598894 ,
"W" => -0.056776627 ,
"Y" => 1.286747697 ,
},
aa18 => {
"A" => -0.980750845 ,
"C" => 2.167579009 ,

```

```

"D" => 0.080795242 ,
"E" => -0.310357307 ,
"F" => -1.102650899 ,
"G" => 0.665757743 ,
"H" => 0.497463827 ,
"I" => 0.734619601 ,
"K" => -2.086410258 ,
"L" => -1.404340455 ,
"M" => 1.605886287 ,
"N" => 0.980883426 ,
"P" => -1.321963927 ,
"Q" => -9.306126741 ,
"R" => 0.200557999 ,
"S" => 0.28666748 ,
"T" => 0.301336393 ,
"V" => -10.08342992 ,
"W" => 1.759565958 ,
"Y" => 0.710772859 ,
},
aa19 => {
"A" => 0.010261492 ,
"C" => 2.158591345 ,
"D" => 0.486845078 ,
"E" => -0.511990049 ,
"F" => -9.340457252 ,
"G" => -0.513154922 ,
"H" => 0.488476164 ,
"I" => 0.310594439 ,
"K" => -0.773469826 ,
"L" => -1.413328118 ,
"M" => 2.181861125 ,
"N" => 0.441381046 ,
"P" => -1.915914091 ,
"Q" => -1.086295714 ,
"R" => 0.570081959 ,
"S" => 0.664702939 ,
"T" => -0.359727967 ,
"V" => -10.09241758 ,
"W" => 0.528185874 ,
"Y" => 1.161216815 ,
},
aa20 => {
"A" => -0.797093431 ,
"C" => 1.796021266 ,
"D" => -1.098117422 ,
"E" => 0.140086648 ,
"F" => -9.340457252 ,
"G" => -0.34322992 ,
"H" => 0.488476164 ,
"I" => 0.047560033 ,
"K" => -2.095397921 ,
"L" => -1.190935697 ,
"M" => 1.918826719 ,
"N" => 1.078810967 ,
"P" => -0.593985996 ,
"Q" => 1.136096708 ,
"R" => 0.039567242 ,
"S" => 0.38459502 ,
"T" => 0.417879612 ,
"V" => -10.09241758 ,
"W" => 0.528185874 ,
"Y" => 1.286747697 ,
},
aa21 => {
"A" => 0.363988599 ,
"C" => 0.564785873 ,
"D" => -0.521997893 ,
"E" => 0.256774558 ,
"F" => -9.349300224 ,
"G" => -0.937035392 ,
"H" => 0.479633192 ,
"I" => -0.283211034 ,
"K" => -0.104240893 ,
"L" => -1.199778668 ,
"M" => 0.810448074 ,
"N" => 1.069967995 ,
"P" => -0.602828968 ,
"Q" => -0.680101186 ,
"R" => 1.108726783 ,
"S" => 0.268836845 ,
"T" => 0.283505758 ,
"V" => -2.457404367 ,
"W" => 0.519342902 ,
"Y" => 1.152373843 ,
},
aa22 => {
"A" => 0.211804078 ,
"C" => 1.804918775 ,
"D" => -0.767291818 ,
"E" => 0.011480633 ,

```



```

        "F" => -9.331559743 ,
        "G" => 0.196182306 ,
        "H" => 0.760408079 ,
        "I" => 0.434969165 ,
        "K" => -1.086500412 ,
        "L" => -0.98939311 ,
        "M" => 1.828188554 ,
        "N" => 1.187244149 ,
        "P" => -1.907016582 ,
        "Q" => -1.077398205 ,
        "R" => 0.048464751 ,
        "S" => -0.091934298 ,
        "T" => 0.426777121 ,
        "V" => -10.08352008 ,
        "W" => -7.691735308 ,
        "Y" => 1.411122423 ,
    },
    aa23 => {
        "A" => -0.220973901 ,
        "C" => 1.979823372 ,
        "D" => -0.785032299 ,
        "E" => -1.328187942 ,
        "F" => -1.705444034 ,
        "G" => 0.178441825 ,
        "H" => 0.96506002 ,
        "I" => 0.623679562 ,
        "K" => -1.519278392 ,
        "L" => -0.547701972 ,
        "M" => 1.810448074 ,
        "N" => 0.847575574 ,
        "P" => -1.339794562 ,
        "Q" => -0.680101186 ,
        "R" => -0.139200731 ,
        "S" => -0.109674779 ,
        "T" => 0.146002234 ,
        "V" => -10.10126056 ,
        "W" => 1.519342902 ,
        "Y" => 1.59983282 ,
    },
    aa24 => {
        "A" => -1.220884849 ,
        "C" => 2.149837426 ,
        "D" => -0.299516419 ,
        "E" => 0.372340828 ,
        "F" => -0.705354982 ,
        "G" => -0.199980746 ,
        "H" => 0.479722245 ,
        "I" => 0.417317737 ,
        "K" => -1.519189339 ,
        "L" => -0.685116443 ,
        "M" => 2.090645045 ,
        "N" => 0.847664626 ,
        "P" => -1.92466801 ,
        "Q" => -1.680012134 ,
        "R" => -1.554149178 ,
        "S" => 0.027917798 ,
        "T" => 0.283594811 ,
        "V" => -2.457315314 ,
        "W" => 1.519431955 ,
        "Y" => 1.693031277 ,
    },
    aa25 => {
        "A" => 0.001507572 ,
        "C" => 1.564874926 ,
        "D" => 0.352560277 ,
        "E" => -0.006170795 ,
        "F" => -1.705354982 ,
        "G" => -0.199980746 ,
        "H" => -0.257243349 ,
        "I" => 1.038806114 ,
        "K" => -0.782223745 ,
        "L" => -0.547612919 ,
        "M" => 1.462613823 ,
        "N" => 0.070057048 ,
        "P" => -1.92466801 ,
        "Q" => -1.095049633 ,
        "R" => 0.030813323 ,
        "S" => 0.15344868 ,
        "T" => -0.175836808 ,
        "V" => -1.872352813 ,
        "W" => 0.934469454 ,
        "Y" => 1.152462896 ,
    },
    aa26 => {
        "A" => 0.281615492 ,
        "C" => 1.979912425 ,
        "D" => -0.106871341 ,
        "E" => 0.256863611 ,
        "F" => -1.705354982 ,
        "G" => 0.06305366 ,

```

```

"H" => 0.479722245 ,
"I" => -0.11319698 ,
"K" => -0.782223745 ,
"L" => -2.007044538 ,
"M" => 1.462613823 ,
"N" => 0.722133744 ,
"P" => -1.92466801 ,
"Q" => -0.680012134 ,
"R" => -0.331756756 ,
"S" => 0.568486179 ,
"T" => 0.146091287 ,
"V" => -2.457315314 ,
"W" => 1.256397549 ,
"Y" => 1.277993778 ,
},
aa27 => {
"A" => -0.105496684 ,
"C" => 2.301751467 ,
"D" => 0.214967701 ,
"E" => 0.256774558 ,
"F" => -1.705444034 ,
"G" => -0.521997893 ,
"H" => 0.479633192 ,
"I" => 0.176220585 ,
"K" => -0.519278392 ,
"L" => -1.199778668 ,
"M" => 1.173018153 ,
"N" => 1.262613073 ,
"P" => -0.924757063 ,
"Q" => 0.319898814 ,
"R" => -0.139200731 ,
"S" => 0.027828745 ,
"T" => -0.006000859 ,
"V" => -10.10126056 ,
"W" => -0.065619598 ,
"Y" => 1.152373843 ,
},
aa28 => {
"A" => 0.372741976 ,
"C" => 1.57353925 ,
"D" => -0.290852095 ,
"E" => -0.149509564 ,
"F" => -9.340546847 ,
"G" => -0.928282015 ,
"H" => 0.751420975 ,
"I" => 0.184973962 ,
"K" => -2.095487515 ,
"L" => -0.297940495 ,
"M" => 1.918737124 ,
"N" => 0.078721372 ,
"P" => -0.916003685 ,
"Q" => 0.328652191 ,
"R" => -0.323092432 ,
"S" => 0.484041099 ,
"T" => 0.533267235 ,
"V" => -10.09250718 ,
"W" => -7.700722411 ,
"Y" => 1.509050524 ,
},
aa29 => {
"A" => -0.998492428 ,
"C" => 2.88680302 ,
"D" => -0.299516419 ,
"E" => -0.158173888 ,
"F" => -1.705354982 ,
"G" => 0.178530878 ,
"H" => 0.742756651 ,
"I" => 0.30184052 ,
"K" => -2.10415184 ,
"L" => -1.199689616 ,
"M" => 1.462613823 ,
"N" => 0.847664626 ,
"P" => -1.92466801 ,
"Q" => -0.095049633 ,
"R" => -0.554149178 ,
"S" => 0.268925897 ,
"T" => 0.146091287 ,
"V" => -1.457315314 ,
"W" => 1.256397549 ,
"Y" => 1.393470995 ,
},
aa30 => {
"A" => -0.635922348 ,
"C" => 1.564874926 ,
"D" => -0.784943246 ,
"E" => -0.32809889 ,
"F" => -1.120392481 ,
"G" => 0.565554001 ,
"H" => 0.965149072 ,
"I" => 0.623768615 ,

```

```

        "K" => -0.519189339 ,
        "L" => -1.007044538 ,
        "M" => 2.003182204 ,
        "N" => 0.722133744 ,
        "P" => -0.339705509 ,
        "Q" => -0.680012134 ,
        "R" => -0.331756756 ,
        "S" => -0.109585726 ,
        "T" => -0.368481886 ,
        "V" => -2.457315314 ,
        "W" => -0.065530546 ,
        "Y" => 1.393470995 ,
    },
);

my $outcpodd = 0;
my $outmtodd = 0;
my $outspodd = 0;
my $outotodd = 0;

#chloroplasts
if ($inpos == 2) {
    $outcpodd = $cpodd{"aa2"}{$inaa};
} elsif ($inpos < 13) {
    $outcpodd = $cpodd{"aa3to12"}{$inaa};
} elsif ($inpos >= 20) {
    $outcpodd = $cpodd{"aa20to30"}{$inaa};
} else {
    my $k = "aa".$inpos;
    $outcpodd = $cpodd{$k}{$inaa};
}
#others
my $keyaa = "aa".$inpos;
$outmtodd = $mtodd{$keyaa}{$inaa};
$outspodd = $spodd{$keyaa}{$inaa};
$outotodd = $otodd{$keyaa}{$inaa};

return ($outcpodd,$outmtodd,$outspodd,$outotodd);
}

sub ERF () {
    #input value
    my $xx = $_[0]; #input value

    #A&S Formula 7.1.26
    my $a1 = 0.254829592;
    my $a2 = -0.284496736;
    my $a3 = 1.421413741;
    my $a4 = -1.453152027;
    my $a5 = 1.061405429;
    my $p = 0.3275911;
    $xx = abs($xx);
    my $t = 1 / (1 + $p * $xx);

    #Direct calculation using formula 7.1.26 is absolutely correct
    #But calculation of nth order polynomial takes O(n^2) operations
    #return 1 - (a1 * t + a2 * t * t + a3 * t * t * t + a4 * t * t * t * t + a5 * t * t * t * t * t) * Math.Exp(-1 * x
    * x);

    #Horner's method, takes O(n) operations for nth order polynomial
    return 1 - (((((( $a5*$t+$a4)*$t)+$a3)*$t+$a2)*$t)+$a1)*$t*exp(-1*$xx*$xx);
}

sub ZDIST () {
    #input value
    my $zz = $_[0]; #input value
    my $sign = 1;

    if ($zz < 0) { $sign = -1; }
    return 0.5 * (1.0+$sign*&ERF(abs($zz)/sqrt(2)));
}

sub N_PDF () {
    #input value
    my $xx = $_[0]; #x value
    my $amp = $_[1];
    my $mean = $_[2];
    my $sd = $_[3];

    #Y=Amplitude*exp(-0.5*((X-Mean)/SD)^2)
    #calculation
    my $e = exp( -0.5 * ((( $xx-$mean)/$sd)**2) );
    return $amp*$e;
}

```

Code A4-7. FGLK Peak Finder

```
#!/usr/bin/perl
#####
# Created by Prakitchai Chotewutmontri, 17 January 2012
# GST student
#
# OBJECTIVE
#   Assign the peaks based on a cut-off value. Return the column # that has the value greater
# than the cut-off
#
# Input file 1: out_FGLK2_w8_2_0_Lee-CTP208-80aa_arrangeByDataLine_g1-3.txt
#   FGLK prediction score
#       8      8      8      ...
#       2      2      2      ...
#       2      2      4      ...
#       ...
#
# Output:
#       10      25      26      ...
#       6       7       8       ...
#       ...
#####

use strict;
use warnings;

#global vars
my $line = "";
my @temp = ();
my @column = ();
my $i = 0;
my $j = 0;

#usage
my $USAGE = "usage: $0 <in FGLK outfile filename> <cut-off value> <out filename>\n\n";

#check argument
unless (@ARGV == 3) {
    print $USAGE;
    exit -1;
}

#store argv
my $fglkfname = $ARGV[0];
my $cutoff = $ARGV[1];
my $outfname = $ARGV[2];

#open FGLK infile
unless ( open(INFGLK,"<$fglkfname") ) {
    print "Can't create $fglkfname\n\n";
}

#open outfile
unless ( open(OUTFILE,">$outfname") ) {
    print "Can't create $outfname\n\n";
}

#will read the value in fglk outfile & process using cut-off value
#read infiles
while ( $line = <INFGLK> ) {
    chomp ($line); #remove end line character
    @temp = ();
    @temp = split ( /\t/, $line); # split the FGLK values

    #identify columns with value greater than the cut-off
    @column = ();

    for ($i=0; $i< scalar(@temp); $i++ ) {
        if ( $temp[$i] > $cutoff ) {
            push (@column, ($i+1) );
        }
    }

    #print the identified column# to outfile
    for ($j=0; $j< scalar(@column); $j++ ) {
        if ($j != 0) { print OUTFILE "\t"; }
        print OUTFILE "$column[$j]";
    }
    print OUTFILE "\n";
}

close OUTFILE;
close INFGLK;
print "DONE\n\n";

exit 0;
```

Code A4-8. Hsp70-FGLK Distance Calculator

```
#!/usr/bin/perl
#####
# Created by Prakitchai Chotewutmontri, 13 January 2012
# GST student
#
# OBJECTIVE
#   Calculate the residue distances between RPPD (Hsp70) site & FGLK site
#
# Input file 1: RPPD_locs.txt
#   location of the RPPD peaks, each line from a transit peptide
#   44      0      0      0      0      0
#   19      54      0      0      0      0
#   26      71      0      0      0      0
#
# Input file 2: FGLK_locs.txt
#   location of the FGLK peaks, each line from a transit peptide
#   20      60      0      0      0
#   0       0      0      0      0
#   4       18      0      0      0
#
# Caution!!!
#           Each row in both input files should represent the same protein
#
# Output:
#   Distance1
#   Distance2
#   ...
#####

use strict;
use warnings;

#global vars
my $line1 = "";
my $line2 = "";
my @temp1 = ();
my @temp2 = ();
my @rppdpks = ();
my @fglkpks = ();
my $i = 0;
my $j = 0;
my $distance = 0;
my $count = 0;

#usage
my $USAGE = "usage: $0 <in RPPD filename> <in FGLK filename> <out filename>\n\n";

#check argument
unless (@ARGV == 3) {
    print $USAGE;
    exit -1;
}

#store argv
my $rppdfname = $ARGV[0];
my $fglkfname = $ARGV[1];
my $outfname = $ARGV[2];

#open RPPD infile
unless ( open(INRPPD,"<$rppdfname") ) {
    print "Can't open $rppdfname\n\n";
    exit -1;
}

#open FGLK infile
unless ( open(INFGLK,"<$fglkfname") ) {
    print "Can't create $fglkfname\n\n";
}

#open outfile
unless ( open(OUTFILE,">$outfname") ) {
    print "Can't create $outfname\n\n";
}

#will read both input files a line at a time, measure the distance and then load next line.

#read infiles
while ( $line1 = <INRPPD> ) {
    chomp ($line1); #remove end line character
    @temp1 = ();
    @temp1 = split ( /\t/, $line1); # split the RPPD location
    #keep only non zero values
    @rppdpks = ();
    for ($i=0; $i< scalar(@temp1); $i++ ) {
```

```

        if ( $temp1[$i] != 0 ) {
            push (@rppdpks, $temp1[$i]);
        }
    }

    $line2 = <INFGK>; #read FGLK line
    chomp ($line2); #remove end line character
    @temp2 = ();
    @temp2 = split ( /\t/, $line2); #split the FGLK location
    #keep only non zero values
    @fglkpks = ();
    for ($i=0; $i< scalar(@temp2); $i++ ) {
        if ( $temp2[$i] != 0 ) {
            push (@fglkpks, $temp2[$i]);
        }
    }

    #print header for OUTFILE
    print OUTFILE "RPPD_AA\tFGLK_AA\tDISTANCE\n";

    #check there are indentified peaks in both RPPD & FGLK, continue ...
    if (scalar(@rppdpks) != 0 && scalar(@fglkpks) != 0) {
        for ($i=0; $i< scalar(@rppdpks); $i++) {
            for ($j=0; $j< scalar(@fglkpks); $j++) {
                #calculate the distance
                #RPPD calculated using 6-aa window, real position is pos+2
                #FGLK calculated using 8-aa window, real position is pos+3
                my $rppd_aa = $rppdpks[$i] + 2;
                my $fglk_aa = $fglkpks[$j] + 3;
                $distance = $fglk_aa - $rppd_aa;
                #print the output
                print "AA $rppd_aa to AA $fglk_aa => $distance\n";
                print OUTFILE "$rppd_aa\t$fglk_aa\t$distance\n";
                $count++;
            }
        }
    }

    close OUTFILE;
    close INRPPD;
    close INFGK;

    print "Total = $count distances\n\n";
    exit 0;

```

Code A4-9. Random Sequence Generator

```
#!/usr/bin/perl
#Created by Prakitchai Chotewutmontri, 5 June 2012
#GST student
#
#Modified from older version "random_seq_AAfreq_LNdist.pl"
#
#####
#OBJECTIVE
# Create random amino acid sequence(s) based on input amino acid frequency file.
# Will generate the sequence with the input length & number
#
#INPUT FILE
# [File 1] = amino acid frequency file
# Ask for file name
# Contain: <amino acid>\t<freq>\n
# Example
# A      0.0813
# C      0.0142
# D      0.0540
# E      0.0673
# F      0.0388
# G      0.0704
# H      0.0228
# I      0.0592
# K      0.0588
# L      0.0967
# M      0.0241
# N      0.0405
# P      0.0477
# Q      0.0395
# R      0.0550
# S      0.0667
# T      0.0535
# V      0.0682
# W      0.0109
# Y      0.0293
#
#OUTPUT FILES
# Contain: fasta-formatted seq, header contain tag+number
# Ask for file name
# Example
# >tag1
# random sequence
# >tag2
# random sequence
# .....
#
#####

use strict;
use warnings;
use POSIX; #for rounding number

#global var
my $line = "";
my %AAfreq = ();
my $temp = ();
my $MeanVal = 0;
my $SDVal = 0;
my @AAlist = ();
my $i = 0;
my $j = 0;
my $k = 0;
my $l = 0;
my @firstpool = ();
my $numneed = 0;
my $realnum = 0;
my $firststrand = "";
my $pi = 3.14159265;
my @secondpool = ();

my @firstlnpool = ();
my $firstlnrand = "";
my @secondlnpool = ();

#usage
my $USAGE = "usage: $0 <aaFreq file> <length> <num seq> <out filename> <tag>\n\n";

#####
#
#MAIN PROGRAM
#
#####
#check argument
unless (@ARGV == 5) {
    print $USAGE;
}
```

```

        exit -1;
    }

    #store argv
    my $aafname = $ARGV[0];
    my $seqlength = $ARGV[1];
    my $totalseq = $ARGV[2];
    my $outfname = $ARGV[3];
    my $headertag = $ARGV[4];

    #open input files
    unless ( open(AAFILE,"<$aafname") ) {
        print "Can't open $aafname\n\n";
        exit -1;
    }

    print "\n\n..START.....\n\n";

    #read aa freq file
    $i=0;
    while ( $line = <AAFILE> ) {
        chomp ($line); #remove new-line char
        @temp = split(/\t/, $line); #split data
        $AAfreq{$temp[0]} = $temp[1]; #AA = freq
        $i++; #count
        #print "read $temp[0] => $temp[1]\n";
        #print "hash $temp[0] => $AAfreq{$temp[0]}\n";
    }

    #test number of aa in the file
    if ( $i != 20 ) { print " Warning! not complete list of amino acid\n\n"; }
    print "..Read amino acid frequency values\n\n";

    #open outfile
    unless ( open(OUTFILE,">$outfname") ) {
        print "Can't create $outfname\n\n";
    }

    #####
    #Part 1
    #Generate first pool of AA = 3000 residues
    # Thus, frequency 1 = 3000; freq x = x*3000
    @AAlist = sort ( $AAfreq{$a} cmp $AAfreq{$b} ) keys %AAfreq; #alphabetical order single-coded AA
    for ($i=0; $i < scalar(@AAlist); $i++) {
        #print $AAlist[$i];
        #print " = $AAfreq{$AAlist[$i]}\n";

        $numneed = $AAfreq{$AAlist[$i]} * 3000; #calculate portion in 3000 from freq
        $realnum = $numneed; #keep real value
        $numneed = floor($numneed); #get integer lower
        if ( ($numneed % 2) == 0 ) { #rounded number is even
            if ( ($realnum-$numneed) > 0.5 ) { $numneed++; } #>0.5 shift to next int
        } else { #rounded number is odd
            if ( ($realnum-$numneed) >= 0.5 ) { $numneed++; } #from 0.5, shift to next
        }
        #print "Amino = $AAlist[$i]\n";
        #print "Need = $numneed\n";

        #add AA to the pool
        for ($j=0; $j < $numneed; $j++) { #number is based on freq
            push(@firstpool, $AAlist[$i]); #add new AA to array
        }
        #final array size maybe largeer that 3000; eg 3001
        #because of rounding function!

        #print "Legth = ";
        #print scalar(@firstpool);
        #print "\n";
    }

    #####
    #Part 2
    #Generate First Random Sequence; Make Random AA seq of frist pool of AA!!
    $firststrand = join("", @firstpool[ map { rand @firstpool } ( 1 .. 3000 ) ]);
    #print ">First random seq\n";
    #print "$firststrand\n\n";

    #move seq from firststrand seq to array for ease of use
    for ($i=0; $i<length($firststrand); $i++) {
        push(@secondpool, substr($firststrand,$i,1));
    }

    #####
    #Part 3
    #Write the sequence to the output file

    for ($i=0; $i<$totalseq; $i++) {
        #Write seq to output file
        print OUTFILE "\>$headertag"; #headertag
        print OUTFILE $i+1; #numbered
        print OUTFILE "\n";
        my $len = $seqlength; #length of seq
    }

```



```

#generate random seq with this lenght from the secondpool of AA
my $secondrand = join("", @secondpool[ map { rand @secondpool } ( 1 .. $len ) ]);
#write to out file with 60 residue limit per line
for ($k=0; $k < length($secondrand); $k = $k + 60) { #60 char per line
    print OUTFILE substr($secondrand,$k,60)."\n";
}
#print ">$headertag";
#print $i+1;
#print "\n";
#for ($k=0; $k < length($secondrand); $k = $k + 60) { #60 char per line
#    print substr($secondrand,$k,60)."\n";
#}
}

print "..DONE\n\n\n";

close (AAFILE);
close (OUTFILE);

exit 0;

```

Code A4-10. Mutant TP Generator Using Random Spacer Sequences

```
#!/usr/bin/perl
#Created by Prakitchai Chotewutmontri, 5 June 2012
#GST student
#
#####
#OBJECTIVE
# To test the generated spacer sequences between RPPD-FGLK peaks if it is suitable
# to be used in the constructs. This program will generated in silico constructs of
# SSF mutants containing the synthetic linkers by adding the fix sequence of SSF
# onto the N-ter and C-ter of the linkers
#
#INPUT FILE
# [File 1] = A fasta-formatted sequence file
# Example
# >header1
# sequence1
# >header2
# sequence2
# .....
#
#OUTPUT FILES
# Contain: fasta-formatted seq
# Ask for file name
# Example
# >header1
# [NterSeq]sequence1[CterSeq]
# >header2
# [NterSeq]sequence2[CterSeq]
# .....
#
#####

use strict;
use warnings;

#global var
my $line = "";
my $num = 0;
my $seq = "";
my $header = "";

#usage
my $USAGE = "usage: $0 <fasta file> <N-ter Seq> <C-ter Seq> <out filename> \n\n";

#####
#
#MAIN PROGRAM
#
#####

#check argument
unless (@ARGV == 4) {
    print $USAGE;
    exit -1;
}

#store argv
my $inname = $ARGV[0];
my $nter = $ARGV[1];
my $cter = $ARGV[2];
my $outname = $ARGV[3];

#open infile
unless ( open(INFILE,"<$inname") ) {
    print "Can't open $inname\n\n";
    exit -1;
}

#open outfile
unless ( open(OUTFILE,">$outname") ) {
    print "Can't create $outname\n\n";
}

#convert N-ter & C-ter to capital letter
$nter =~ tr/[a-z]/[A-Z]/;
$cter =~ tr/[a-z]/[A-Z]/;

#read infile and write outfile
while ( $line = <INFILE> ) {
    chomp ($line);
    if ( $line =~ /^>/ ) { #this line is a header line
        $num++; #count header
        if ($num != 1) { #not the first header, write out previous seq result
            $seq =~ tr/[a-z]/[A-Z]/;
            my $fullseq = $nter.$seq.$cter;
            print OUTFILE $fullseq."\n";
        }
    }
}
```

```

        $seq = ""; #clear seq
    }
    #write header
    print "\n\n$line\n";
    $header = $line; #keep header for FASTA out file
    #print OUTFILE $line."\t"; #same line header
    print OUTFILE $line."\n"; #separate line header
}
else { #otherline remove spaces, concate seq
    $line =~ s/\s//g ;
    $seq .= $line;
}
}
#last seq
$seq =~ tr/[a-z]/[A-Z]/;
print OUTFILE $nter.$seq.$cter."\n";

close OUTFILE;

print "Total = $num sequences\n\n";

close INFILE;

exit 0;

```

Code A4-11. Amino Acid Distribution Calculator

```
#!/usr/bin/perl
#Created by Prakitchai Chotewutmontri, 16 Feb 2010
#GST student
#####
#OBJECTIVE
# Count amino acid frequency in fasta input file
#INPUT FILE
# A fasta-formatted sequence file
# Ask for file name
# Example
# >header1
# sequence1
# >header2
# sequence2
#OUTPUT FILES
# Contain 3 columns, tab-delimited
# Ask for file name
# <Amino_acid>\t<frequency>\t<count>\n
# Example
# A      0.0000      0
# C      0.2872     2872
# ...
# Y      0.0191     191
#
#####

use strict;
use warnings;

#global var
my @Aalist = ("A","C","D","E","F","G","H","I","K","L","M","N","P","Q","R","S","T","V","W","Y");
my %AatoFreq = ( "A", "0", #Alanine
                  "C", "0", #Cysteine
                  "D", "0", #Aspartate
                  "E", "0", #Glutamate
                  "F", "0", #Phenylalanine
                  "G", "0", #Glycine
                  "H", "0", #Histidine
                  "I", "0", #Isoleucine
                  "K", "0", #Lysine
                  "L", "0", #Leucine
                  "M", "0", #Methionine
                  "N", "0", #Asparagine
                  "P", "0", #Proline
                  "Q", "0", #Glutamine
                  "R", "0", #Arginine
                  "S", "0", #Serine
                  "T", "0", #Threonine
                  "V", "0", #Valine
                  "W", "0", #Tryptophan
                  "Y", "0" ); #Tyrosine

my $line = "";
my $numseq = 0;
my $seq = "";
my $res_i = 0;
my $numkey = 0;
my $totalAcount = 0;
my $totalcount = 0;
my $i = 0;
my $temp = 0;

#usage
my $USAGE = "usage: $0 <input fasta filename> <output filename>\n\n";

#error infile check
my $foundgap = 0;

#####
#
#MAIN PROGRAM
#
#####
#check argument
unless (@ARGV == 2) {
    print $USAGE;
    exit -1;
}

#store argv
my $infile = $ARGV[0];
my $outfile = $ARGV[1];

#open infile
unless ( open(INFILE,"<$infile") ) {
    print "Can't open $infile\n\n";
    exit -1;
}
```

```

}

#open outfile
unless ( open(OUTFILE,">$outfname") ) {
    print "Can't create $outfname\n\n";
}

#read infile
$numseq = 0;
$seq = "";
while ( $line = <INFILE> ) {
    chomp ($line); #remove new-line char
    if ( $line =~ /^>/ ) { #this line is a header line
        $numseq++; #count header
        if ($numseq > 1) { #found next header, count freq
            $seq =~ tr/[a-z]/[A-Z]//; #convert seq to Capital Letter
            for ($res_i=0; $res_i<length($seq); $res_i++) { #read each aa residue
                #use AA as a key, count +1
                $AtoFreq[substr($seq,$res_i,1)] = $AtoFreq[substr($seq,$res_i,1)]+1;
            }
            $seq = ""; #reset seq
        }
    }
    else { #other lines, not header,
        $line =~ s/\s//g; #remove spaces
        $seq .= $line; #add line to previous seq
    }
}
#last seq
for ($res_i=0; $res_i < length($seq); $res_i++) {
    $seq =~ tr/[a-z]/[A-Z]//; #convert seq to Capital Letter
    for ($res_i=0; $res_i<length($seq); $res_i++) { #read each aa residue
        #use AA as a key, count +1
        $AtoFreq[substr($seq,$res_i,1)] = $AtoFreq[substr($seq,$res_i,1)]+1;
    }
}
#finished read infile
print " Number of Sequence in Input :";
print $numseq;
print "\n";

#To calculate frequency
#count number of key (char) found in seq
#sum total number of count
$numkey = 0;
$totalcount = 0;
while ( my ($key, $value) = each(%AtoFreq) ) { #run through every keys
    $numkey++; #count number of key
    $totalcount = $totalcount+$value; #sum values
}
#sum total number of count from aa
$totalAcount = 0;
for ($i=0; $i < scalar(@Aalist); $i++) {
    $totalAcount = $totalAcount+$AtoFreq{$Aalist[$i]}; #sum values from AA keys
}
print " Number of Amino Acids in Input :";
print $totalAcount;
print "\n";

#test if found non protein character in the count
$foundgap = 0;
if ($numkey > 20) {
    if (exists($AtoFreq{"-"}) ) {
        $foundgap = 1;
        print " Number of Gaps Found in Input: ";
        print $AtoFreq{"-"};
        print "\n";
        if ($numkey > 21) {
            print " Number of Non-Amino Acid Characters Found in Input: ";
            print $totalcount-$totalAcount-$AtoFreq{"-"};
            print "\n";
        }
    }
    else {
        print " Number of Non-Amino Acid Characters Found in Input: ";
        print $totalcount-$totalAcount;
        print "\n";
    }
}

#Calcualte freq for each AA & WRITE OUTFILE
for ($i=0; $i < scalar(@Aalist); $i++) {
    $temp = $AtoFreq{$Aalist[$i]}; #keep num of count
    $AtoFreq{$Aalist[$i]} = ($AtoFreq{$Aalist[$i]}/$totalAcount); #count values/total AA count
    print OUTFILE $Aalist[$i];
    print OUTFILE "\t";
    printf OUTFILE "%.4f", $AtoFreq{$Aalist[$i]};
    print OUTFILE "\t$temp\n";
}

print ".....DONE\n";
close (INFILE);
close (OUTFILE);
exit 0;

```

Vita

Prakitchai Chotewutmontri also known as “Non” was born in Nonthaburi, Thailand on June 7, 1980. He was raised in Mukdahan where he attended Mukdalai primary school. Non moved to Bangkok to attend secondary school at Nawaminthrachinuthit Bodindecha School before received the high school diploma in junior year from the Office of Non-formal and Informal Education in 1997. He obtained his Bachelor’s Degree in Biology (Genetics) and Master’s Degree in Genetics (Biochemistry) from Kasetsart University in Bangkok in 2001 and 2004, respectively. Non later attended New Jersey Institute of Technology – Rutgers in Newark, NJ and received a Master’s Degree in Computational Biology in 2006. He then joined the University of Tennessee – Oak Ridge National Laboratory Graduate School of Genome Science and Technology. He rotated in the Bruce Laboratory before joining in Fall 2007 to work on chloroplast transit peptides.