Transit Peptide Cleavage Sites of Integral Thylakoid Membrane Proteins*

Stephen M. Gómez‡§, Karl Y. Bil'¶, Rodrigo Aguilera‡, John N. Nishio‖, Kym F. Faull‡, and Julian P. Whitelegge‡**

A set of 58 nuclearly encoded thylakoid-integral membrane proteins from four plant species was identified, and their amino termini were assigned unequivocally based upon mass spectrometry of intact proteins and peptide fragments. The dataset was used to challenge the Web tools ChloroP, TargetP, SignalP, PSORT, Predotar, and MitoProt II for predicting organelle targeting and transit peptide proteolysis sites. ChloroP and TargetP reliably predicted chloroplast targeting but only reliably predicted transit peptide cleavage sites for soluble proteins targeted to the stroma. SignalP (eukaryote settings) accurately predicted the transit peptide cleavage site for soluble proteins targeted to the lumen. SignalP (Gram-negative bacteria settings) reliably predicted peptide cleavage of integral thylakoid proteins inserted into the membrane via the "spontaneous" pathway. The processing sites of more common thylakoid-integral proteins inserted by the signal recognition peptide-dependent pathway were not well predicted by any of the programs. The results suggest the presence of a second thylakoid processing protease that recognizes the transit peptide of integral proteins inserted via the spontaneous mechanism and that this mechanism may be related to the secretory mechanism of Gram-negative bacteria. ***Molecular & Cellular Proteomics 2:1068–1085, 2003.***

Advances in genome sequencing (1–3) have provided a huge resource that is driving the field of bioinformatics. Besides providing a readout that approximates the primary structure of a particular gene product, there are many facets of expression, such as organellar targeting, that can be addressed via computational analysis of genomic data. Post-translational modification sites, including signal peptide cleavage sites, may be predicted, although the accuracy of such predictive algorithms is often rooted in the knowledge base used for their development.

Proteomics, the systematic study of large sets of proteins expressed by a particular cell type or tissue, is largely possible because of recent technological and computational advances. Contemporary proteomics is based on the following three components: analytical separation of proteins from complex starting material, identification of the proteins, and subsequent sorting and classification of the datasets using bioinformatics software tools. The availability of complete genome sequences, improvements in techniques for protein separation and displays (4–6), and new developments in mass spectrometry (MS)[1] for protein identification (7–10) have arguably outpaced developments in bioinformatics. Using the complete mitochondrial, plastid, and nuclear *Arabidopsis thaliana* genomes (2, 11, 12) and results from recent proteomics experiments (13–16), it is possible to test the performance of algorithms that are commonly used to predict organellar targeting of nuclear gene products and transit peptide cleavage sites. Thus, it is timely to reevaluate the various tools at our disposal.

The chloroplast is a fascinating organelle not only because of its vital photosynthetic function but also because its development requires the coordinated expression and interaction of all three genomes. Chloroplasts have three membrane systems, the outer and inner envelopes and the thylakoids, which form the boundaries of three separate soluble compartments, the interenvelope space, stroma, and thylakoid lumen. Chloroplast proteins, encoded by nuclear genes, have amino-terminal transit sequences that aid in targeting the polypeptides to the correct chloroplast membrane or compartment (17–22). After translation in the cytosol, the transit peptide of the proprotein is phosphorylated by a cytosolic Ser/Thr kinase before import into the chloroplast (23). The phosphorylated proproteins are bound by the Toc complex (translocon of the outer chloroplast envelope) (24), transported across the outer envelope membrane, dephosphorylated in the interenvelope space, and then transported across the inner envelope by the Tic complex (20). Homologs for the Toc and Tic complexes have been found in the genome of *Synechocystis* sp. PCC

From the ‡Pasarow Mass Spectrometry Laboratory, Department of Psychiatry and Biobehavioral Sciences, Department of Chemistry and Biochemistry, and Neuropsychiatric Institute, University of California, Los Angeles, California 90095, ¶Biosphere 2, Columbia University, Oracle, Arizona 85623, and ‖College of Natural Sciences, California State University, Chico, California 95929

Received, July 1, 2003

Published, MCP Papers in Press, August 5, 2003, DOI 10.1074/mcp.M300062-MCP200

[1] The abbreviations used are: MS, mass spectrometry; MSMS, tandem mass spectrometry; 2D, two-dimensional; Gm+, Gram-positive; Gm−, Gram-negative; IMP, integral membrane protein; IMT, intact mass tag; LCMS, liquid chromatography coupled to electrospray mass spectrometry; LCMS+, LCMS with fractions collected at a flow splitter between liquid chromatography and MS for MSMS; MALDI, matrix-assisted laser desorption ionization; TOF, time-of-flight; PS, photosystem; SPP, stromal processing protease.

6803 (20), which supports the model that chloroplasts are reduced cyanobacterial endosymbionts (25–28). Once in the stroma the transit peptide is cleaved by a stromal processing protease (SPP).

Proteins targeted to the thylakoid lumen have bipartite transit peptides. The stromal targeting information is in the amino-proximal portion of the transit peptide, and the carboxyl-proximal portion, similar to signal peptides of secreted proteins in bacteria, contains the information for targeting to the lumen. Two different pathways for targeting to the lumen have been characterized. The first is a Sec-dependent pathway related to the SecYEG export mechanism in bacteria (29, 30). The second is a ΔpH-dependent mechanism characterized by two conserved sequential arginines in the transit peptide, the Tat (twin arginine translocase) pathway (31–33). Proteins targeted to the lumen via the Sec pathway are generally translocated in an unfolded state, whereas proteins imported via the Tat pathway are translocated in a folded state (22). Proteins imported into the lumen by either pathway are processed by a thylakoid processing protease that removes the carboxyl-proximal portion of the transit peptide.

Integral membrane proteins (IMPs) destined for the thylakoid membrane are targeted by one of two additional translocation mechanisms. It is thought that most thylakoid-bound proteins are targeted by information contained within the mature protein sequence. This insertion mechanism, which orients the amino terminus of the protein into the stroma, requires a signal recognition particle and a putative unidentified thylakoid-bound translocation complex (34–37). A few integral thylakoid proteins have bipartite transit peptides that are cleaved first in the stroma and again after membrane insertion. These proteins are inserted via a novel signal recognition particle-independent mechanism that appears to require no protein or nucleotide co-factors (the spontaneous pathway) and has the unique characteristic of leaving the amino terminus of the mature protein oriented into the lumen (38–42). Chloroplast thylakoids have been used as a model system for membrane proteomics because of the size of the proteome, the limited number of post-translational modifications (16), the ease of collecting large amounts of membranes, and the ability to easily subfractionate detergent-resistant domains (43).

Proteomics studies using two-dimensional (2D) gel electrophoresis have provided useful insights for soluble proteins of the lumen and stroma as well as peripheral thylakoid proteins (13, 44–47) and integral thylakoid proteins (48), but these methods do not allow convenient determination of the amino terminus where blocked, as is common in the chloroplast. In this work, a dataset of 35 nuclear encoded integral thylakoid membrane proteins from *A. thaliana,* where the sites of secondary amino-terminal processing were explicitly determined by MS, was used to challenge several bioinformatics tools. After initial trials the dataset was expanded to include photosystem (PS) II thylakoid-associated proteins from pea and spinach published previously (16, 49) and a small number of tobacco IMPs, providing a total of 58 nuclear encoded integral thylakoid membrane proteins, each with an experimentally determined amino terminus. The results demonstrate that although some programs (ChloroP, TargetP) reliably predict trafficking to the chloroplast, they variably predict the processing sites of the transit peptides. The results highlight inadequacies of currently available bioinformatics tools for prediction of secondary amino-terminal processing of transit peptides of integral thylakoid proteins and demonstrate the need for improvements in the datasets used to train the algorithms.

MATERIALS AND METHODS

Plant Growth Conditions—A. thaliana var. *Columbia* seeds were soaked in water (4 °C) for 1–2 d before sowing. Seeds were sown in trays, and seedlings were transferred to 144 cm^2 × 12 cm high pots (~5 plants/pot) containing Scotts Pro-Gro professional potting mix (The Scotts Company, Marysville, OH) approximately 1 week after germination. Plants were grown in a growth chamber maintained at 23 ± 2 °C in the light and 16 ± 2 °C in the dark. The plants were illuminated with a 12-h light period and constant relative humidity of 70%. Fluorescent lamps (Sylvania F72T12/CW/VHO, 160 watts) supplemented with Sylvania 100-W incandescent bulbs were used as the light source. Quantum (400–700 nm) flux density was 700–800 μmol of photons m^{-2} s^{-1} at the level of the leaves. Plants were watered as necessary and supplemented once per week with 0.25× Hoagland's solution (50).

Leaves were harvested from 5-week-old *A. thaliana* plants, placed in square glass beakers, covered with grinding buffer (51), and ground with a chilled Polytron homogenizer (Brinkmann Instruments/Kinematica). The homogenate was filtered over 8 layers of diaper liners (Gerber), and the green filtrate was layered over (40%) Percoll gradients in 1.6-ml microfuge tubes (52). Thylakoids were collected from the Percoll pads, diluted in grinding buffer, and pelleted at full speed in a tabletop centrifuge. The green membrane pellets were resuspended to ~1.1 mg/ml chlorophyll in extraction buffer (51) and frozen in liquid nitrogen. This procedure typically required 10–15 min, and all manipulations were performed on ice or in a 4 °C refrigerator under low light. Tobacco PS II thylakoid proteins were isolated as described previously (16).

LCMS+—A. thaliana thylakoid protein samples (~66 μg of chlorophyll) were prepared by acetone precipitation prior to dissolution in 60% acetic acid. Reverse-phase chromatography was performed as described previously (5, 49) using a poly(styrene-divinylbenzene) copolymer (Polymer Labs PLRP/S, 5 μm × 300 Å, 2.1 × 300 mm) stationary-phase column equilibrated in aqueous 0.1% trifluoroacetic acid containing 5% acetonitrile and eluted (100 μl/min at 40 °C) with a stepped increasing concentration of acetonitrile (min/% acetonitrile: 0/5, 5/5, 10/25, 130/75, and 150/100) initiated at the moment of sample injection (100 μl/injection). The absorption of the effluent from the column was recorded at 280 nm. The effluent was split, and a portion was sent to a fraction collector (typically about 75%, 2-min fractions); the remainder was directed straight into the MS ion source (LCMS+). Tobacco protein samples were measured by LCMS as described previously (16).

Mass spectra were recorded on a PerkinElmer Life Sciences Sciex API *III* triple-quadrupole mass spectrometer with an Ionspray™ source as described by Whitelegge *et al.* (5). The instrument was scanned from *m/z* 600–2300 (0.3 step size, 1-ms dwell time, 6-s scan speed, orifice at 65 V). Manufacturer-supplied software was used for the computations of measured protein molecular weight (MacSpec

TABLE I
A. thaliana translated gene sequences (indicated by chromosome locus) were compared to the following experimentally verified amino-terminal sequences

A. thaliana locus	Gene product	Species	Reference
		Amino terminus determined by peptide sequencing	
At4g04640	AtpC	spinach	55, 56
		tobacco (*Nicotiana tabacum*)	57
At4g09650	AtpD	pea	58
		spinach	59
At4g03280	PetC	spinach	60
At4g02770	PsaD	pea	61
		tobacco	62
		Nicotiana tomentosiformis	62
		Nicotiana sylvestris	62
		cucumber	63
		barley	64
At4g28750/At2g20260	PsaE	tobacco	62
		N. sylvestris	65
		barley	66
		pea	61
		cucumber	67
At3g16140/At1g52230	PsaH	cucumber	63
		barley	68
		pea	61
		tobacco	62
At3g61470	Lhca2	pea	69
		spinach	69
		barley	70, 71
At3g47470	Lhca4	pea	69
		spinach	69
		barley	70, 72
At3g50820/At5g66570	PsbO	spinach	49, 73, 74
		A. thaliana	49
At1g06680	PsbP	pea	75
		spinach	49, 75
		Pinus pinaster	76
		tobacco	77
		A. thaliana	49
At4g21280/At4g05180	PsbQ	pea	75
		spinach	49, 75, 78–80
		A. thaliana	49
At1g79040	PsbR	spinach	81
		wheat	82
At2g30570	PsbW	spinach	79, 83, 84
		wheat	83
At3g21055	PsbX	spinach	83
		wheat	83
At2g06520	$PsbX_c$	spinach	83
At5g54270	Lhcb3	*A. thaliana*	85
		wheat	86
		spinach	87
At1g15820	Lhcb6	barley	88
		spinach	87
		Amino terminus determined by mass spectrometry	
At4g03280	PetC	spinach	18
		pea	18
At2g26500	PetM	spinach	51
At4g02770	PsaD	spinach	18
		pea	18
At4g28750/At2g20260	PsaE	spinach	18
At3g16140/At1g52230	PsaH	spinach	18

TABLE I—*continued*

A. thaliana locus	Gene product	Species	Reference
At5g64040	PsaN	pea	14
At3g50820/At5g66570	PsbO	pea	18
		spinach	18
At1g06680	PsbP	pea	18
		spinach	18
At4g21280/At4g05180	PsbQ	spinach	18
At1g79040	PsbR	spinach	18
At1g44575	PsbS	spinach	18
At2g30570	PsbW	spinach	18, 84
At3g21055	PsbX	spinach	18
At1g29910/At1g29920/At1g29930/ At2g34420/At2g34430	Lhcb1	spinach	18, 89
		petunia	90
		tomato	90
		pea	18
At2g05070/At2g05100/ At3g27690	Lhcb2	spinach	18, 89
		pea	18
At5g54270	Lhcb3	pea	18
		tomato	90
		spinach	91
At1g15820	Lhcb6	spinach	18
		tomato	90

3.3) and zero-charge molecular weight reconstructions (BioMultiView 1.3.1). Calculated average molecular weights were generated from translated gene sequences using PeptideMass (expasy.cbr.nrc.ca/tools/peptide-mass.html) after manually removing the transit peptide.

Protein Identification—Thirty-three intact mass tags (IMTs) from the *A. thaliana* LCMS experiments were correlated with the predicted masses calculated by modifying the entries at The Institute for Genomic Research *A. thaliana* database (www.tigr.org/tdb/e2k1/ath1/) with experimentally verified amino termini of orthologs (Table I).

Cyanogen Bromide Cleavage—Aliquots of fractions (10 μl) collected during LCMS+ were treated with saturated cyanogen bromide (CNBr) (1 μl, 1 g/ml) for 4 h at room temperature in the dark. The reaction mixture was either spotted directly (0.3 μl plus 0.5 μl of matrix solution) or dried by centrifugal evaporation (SpeedVac). Dried reaction mixtures were redissolved in 5 μl of 70% acetic acid and analyzed (0.2 μl plus 0.5 μl of matrix) by matrix-assisted laser desorption ionization (MALDI) coupled to delayed extraction time-of-flight MS in the reflector mode (Voyager DE STR, Applied Biosystems) using α-cyano-4-hydroxycinnamic acid as matrix (10 mg/ml of solution in water/acetonitrile/trifluoroacetic acid 30/70/0.1) and internal/external calibration with bovine insulin. Manufacturer-supplied default settings for a method optimized for peptides less than 6000 Da were used for all samples. Some proteins have no internal Met or no CNBr peptides less than 8000 Da (MALDI in the reflectron mode has an upper limit for high resolution data acquisition of approximately 8000 Da); consequently some abundant proteins will not be detected during MALDI analysis, and peptides from low abundance proteins may appear to be more highly expressed. The MALDI data was compared with the fragmentation pattern predicted by the MS-Tag tool in ProteinProspector, version 4.0.4 (prospector.ucsf.edu/) (90). MS-Tag calculates the predicted fragment masses with the transit peptide sequence still included. There are very few, if any, amino-terminal peptide masses for chloroplast-targeted proteins predicted by MS-Tag. The MS-Digest tool was used to predict the fragment masses after manually removing the transit peptide from each prospective match assigned by MS-Tag.

Trypsin Cleavage—Selected fractions collected during LCMS+ were reduced, alkylated, and treated with trypsin (Promega sequencing grade modified by reductive methylation). Dithiothreitol (15 μl, 10 mM in 50 mM ammonium bicarbonate; 30 min, 24 °C), iodoacetamide (15 μl, 55 mM in 50 mM ammonium bicarbonate; 20 min, 24 °C), and finally trypsin (12.5 μl, 6 ng/μl in 50 mM ammonium bicarbonate; 3 h, 37 °C) were added to aliquots of fractions (10 μl). After incubation, samples were dried by centrifugal evaporation and stored at −20 °C prior to analysis by microliquid chromatography-MSMS.

Analysis of Tryptic Peptide Sequence Tags by Tandem Mass Spectrometry—Samples were analyzed by microliquid chromatography-MSMS with data-dependent acquisition (LCQ-DECA, ThermoFinnigan, San Jose, CA) after dissolution in 5 μl of 70% acetic acid (v/v). A reverse-phase column (200 μm × 10 cm, PLRP/S 5 μm, 300 Å; Michrom Biosciences, San Jose, CA) was equilibrated for 10 min at 1.5 μl/min with 95% A, 5% B (A, 0.1% formic acid in water; B, 0.1% formic acid in acetonitrile) prior to sample injection. A linear gradient was initiated 10 min after sample injection ramping to 60% A, 40% B after 50 min and 20% A, 80% B after 65 min. Column eluent was directed to a coated glass electrospray emitter (TaperTip, TT150-50-50-CE-5, New Objective) at 3.3 kV for ionization without nebulizer gas. The mass spectrometer was operated in "triple-play" mode with a survey scan (400–1500 *m/z*), data-dependent zoom scan, and MSMS. Individual sequencing experiments were matched to a custom *A. thaliana* sequence database using Sequest software (ThermoFinnigan). "No enzyme" was set such that Sequest considered all possible peptide sequence permutations rather than just tryptic ones. To identify *N*-acetylated peptides, a static modification of +42 Da was set for "amino terminus peptides."

Data Analysis—ChloroP, version 1.1 (www.cbs.dtu.dk/services/ChloroP/) (91), is a neural network method for identifying probable chloroplast transit peptide sequences and predicting the proteolytic cleavage site of each transit peptide. ChloroP presents its prediction of chloroplast targeting as a "Y" or "N" output based upon the predicted presence of a chloroplast transit peptide. TargetP, version 1.01 (www.cbs.dtu.dk/services/TargetP/) (92), is a layered neural network method for predicting subcellular targeting based upon the type of targeting/transit peptide predicted to be at the amino terminus of each protein. TargetP predicts whether the protein in question is trafficked to the chloroplast, mitochondria, secretory pathway, or

TABLE II
Experimental mass measurements of nuclear encoded thylakoid-associated proteins from A. thaliana

Protein[a]	*A. thaliana* locus	Retention time	Intact mass tag[b]	Calculated intact mass[c]	Amino-terminal CNBr or MSMS tag[d]	Calculated amino-terminal CNBr tag[e]	Internal CNBr tag[f]
		min	*Da*	*Da*	*Da*	*Da*	
PsbX^{III}	At3g21055	20.6	3347.6	3345.94	N/D		
PsaE^{II}	At2g20260	26.4	10546.4	10546.93	X		
PsaE^{IV}	At4g28750	29.3	10469.8	10469.76	X		
PsaN^{V}	At5g64040	38.1	9702.7	9705.03	X		
PsbO^{III}	At3g50820	Fraction 44		26571.71	2082.01	2082.05	Y
PsaD^{IV}	At4g02770	45.1	17848.7	17847.39			Y
PsbO^{V}	At5g66570	45.6	26566.7	26565.70	2082.01	2082.05	Y
PsbP^{I}	At1g06680	47.1	20212.7	20212.41	6892.25	6890.38	Y
PsbQ^{IV1}	At4g21280	48.7	16310.2	16309.57	X		
PsbQ$^{IV\alpha}$	At4g05180	50.6	16349.2	16349.51	X		
PsaH^{III}	At3g16140	57.6	10355.0	10355.82	X		
PsaH^{I}	At1g52230	57.6	10355.0	10354.79	X		
PetC^{IV}	At4g03280	58.8	19000.9	19000.68	1210.65	1210.61	Y
AtpD^{IV}	At4g09650	72.4	20573.7	20570.59			Y
PsbR^{I}	At1g79040	72.9	10340.9	10341.64	1859.00	1859.04	Y
AtpC^{IV}	At4g04640	73.8	35709.3	35709.19	2626.58	2626.43	Y
Lhca3^{I}	At1g61520	74.9	25058.6	25056.64			Y
Lhcb1^{II2}	At2g34420	Fraction 76		24916.20	7830.7	7830.59	Y
Lhcb1^{II1}	At2g34430	Fraction 76		24815.97			Y
Lhca4^{III}	At3g47470	76.6	22256.0	22255.36	N/D		
Lhcb1$^{I\alpha}$	At1g29930	77.0	24906.9	24905.18			Y
Lhcb1$^{I\beta}$	At1g29920	77.0	24906.9	24905.18			Y
Lhcb1$^{I\gamma}$	At1g29910	77.0	24906.9	24905.18			Y
Lhcb2$^{II\alpha}$	At2g05100	79.1	24935.0	24930.18	7890.6	7890.72	Y
Lhcb2^{III}	At3g27690	79.1	24935.0	24932.26	7890.6	7890.72	Y
Lhca2^{III}	At3g61470	80.3	23201.8	23200.47	6250.4	6247.98	Y
Lhcb6^{I}	At1g15820	80.8	23107.8	23106.32			Y
Lhcb3^{II}	At5g54270	83.1	24280.6	24280.65	6845.8	6845.57	Y
Lhcb2$^{II\beta}$	At2g05070	83.6	24945.0	24944.25	7890.6	7890.72	Y
Lhcb5^{IV}	At4g10340	83.8	26449.0	26447.35	9630.9	9632.84	Y
Lhcb4^{V}	At5g01530	83.9	28212.7	28212.07	acVFGFGK		Y
PsbX$_c$^{II}	At2g06520	84.4	4183.7	4183.94	4182.36	4182.39	Y
PetM^{II}	At2g26500	98.7	4188.6	4188.89			Y
PsbS^{I}	At1g44575	100.6	22457.6	22457.23	4900.75	4900.75	Y
PsbW^{II}	At2g30570	124.7	6004.0	6007.76	N/D		

[a] *A. thaliana* protein annotation using the standard gene names. Superscript roman numerals indicate the chromosome where the gene is located. For paralogs on the same chromosome, Arabic numerals indicate the long arm of the chromosome with 1 closest to the centromere, and lowercase Greek letters indicate the short arm of the chromosome with α closest to the centromere. See footnote 2 in the text concerning confusion about the use of *psbX* as a gene name.

[b] Intact mass tag of zero charge average mass determined by LCMS.

[c] Average zero charge mass predicted from translations of the indicated gene sequences using PeptideMass (us.expasy.org/tools/peptide-mass.html).

[d] Amino-terminal CNBr mass tags. All masses are monoisotopic unless underlined to indicate average mass. X indicates proteins that have no internal Met or have no CNBr peptides smaller than 10,000 Da. N/D, not detected. Amino termini confirmed by MSMS sequence data are listed.

[e] Predicted CNBr peptide mass calculated using MS-Digest program in ProteinProspector (prospector.ucsf.edu/ucsfhtml4.0/msdigest.htm). All masses are monoisotopic unless underlined to indicate average mass.

[f] Intact mass tags confirmed by internal CNBr mass tags other than or in addition to the amino-terminal CNBr mass tag.

"other" subcellular location. PSORT (World Wide Web version, Oct 8, 1999, psort.ims.u-tokyo.ac.jp/) (93) is an expert system using a knowledge-base setup as an "if-then" cascade. PSORT predicts subcellular localization with much finer resolution than any of the other programs examined. The four subcellular/suborganellar localizations with highest scores are ranked in order. The output of PSORT is listed in the footnotes to Table IV. PSORT includes a hydrophobic moment analysis for chloroplast proteins as one of its expert analysis programs. The usefulness of this calculation is based on the assumption that all chloroplast proteins have a similar stromal targeting domain in the amino-terminal targeting peptide. The hydrophobic moment analysis in PSORT distinguishes chloroplast protein status as negative, positive, or undetermined. Predotar, version 0.5 (www.inra.fr/predotar/), is a program still under development designed to be a Web-based method for distinguishing chloroplast- from mitochondria-targeting sequences. Predotar predicts localization to the chloroplast,

TABLE III
Intact mass tags obtained from tobacco PS II membrane preparations by LCMS

Protein[a]	*N. tabacum* GenBank™/EBI accession numbers	Intact mass tag[b]	Calculated intact mass[c]
PsbR	X70088	10319.8	10319.63
Lhcb1–21	X52743	24749.6	24749.15
Lhcb1–7	X58229	24949.4	24950.38
Lhcb1–40	X52744	24913.2	24912.38
Lhcb1–50	X52742	24987.2	24988.43

[a] Tobacco protein annotation using the standard gene names.
[b] Intact mass tag of zero charge average mass determined by LCMS as described in Table II.
[c] Average zero charge mass predicted as described in Table II.

mitochondria, both organelles, or neither organelle. MitoProt II, version 1.0a4 (www.mips.biochem.mpg.de/cgi-bin/proj/medgen/mitofilter) (94), is a computational method for predicting mitochondrial targeting sequences and for predicting the proteolytic cleavage sites of the targeting peptides. MitoProt predicts targeting based upon a calculated probability score that the protein being examined is localized to the mitochondria. SignalP (www.cbs.dtu.dk/services/SignalP/) (95) allows discrimination between eukaryotic, Gram-positive bacterial and Gram-negative bacterial signal peptides. All three cases were tested with the dataset. Four parameters are calculated for a yes/no prediction of the presence of a signal peptide. Each organism group tested gave a predicted cleavage site.

Our dataset was used to test each of the programs listed above. Default settings were used for ChloroP, MitoProt, and Predotar. User-defined settings for TargetP were selected as follows: origin of sequences, plant, perform cleavage site predictions, no cutoff, and winner-takes-all. PSORT was set to plant source of input sequence. SignalP was set to the default analysis by all three prediction routines.

RESULTS

Assembly of the Dataset

A dataset was assembled from 58 nuclear encoded thylakoid-associated proteins, each with experimentally verified amino termini. The dataset includes 18 proteins from pea and spinach that have been described already (16, 49), 35 proteins from *A. thaliana* (Table II), and five proteins from tobacco (Table III). The majority are integral membrane proteins isolated from thylakoids. For comparative purposes the set includes a small number of peripheral proteins localized to the stromal or luminal surface.

Assignment of Proteins and Post-translational Modifications

The 35 nuclear encoded proteins from *A. thaliana* thylakoids were defined by IMTs, and the identities were confirmed by CNBr peptide mass tags and/or MSMS sequencing of tryptic fragments (Table II). IMTs for chloroplast-encoded proteins targeted to the thylakoid will be presented independently. The results show that the amino termini of the Lhcb4 (LHCIIa/CP29) and Lhcb5 (LHCIIc/CP26) chlorophyll *a*/*b*-binding light-harvesting proteins are *N*-acetylvaline at position 33 (initiating Met is residue 1) and *N*-acetylleucine at position 38, respectively. The amino terminus of Lhcb4 was confirmed by MSMS sequencing of the amino-terminal tryptic peptide (acetyl-^{33}VFGFGK^{38}) from a fraction collected during LCMS+ concomitant with elution of the intact protein of 28,212.7 Da (LCMS+) (Fig. 1 and Table II) consistent with previous predictions (96). Note that this amino terminus gives a calculated mass of 28,212.07 Da for the gene product such that full agreement between measured and calculated mass (within measurement error, 0.01% or 2.8 Da) has been achieved.

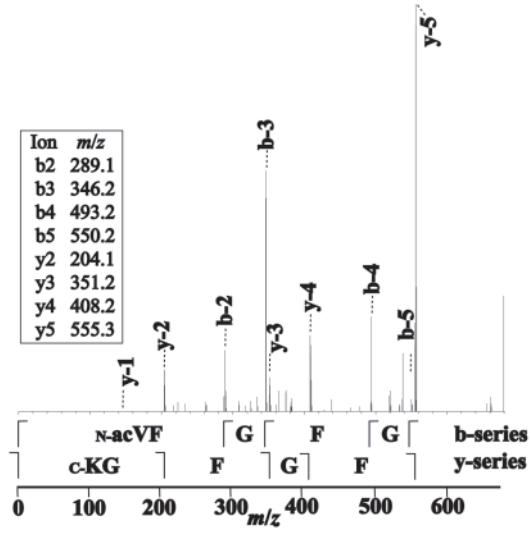

FIG. 1. **MSMS spectrum of the tryptic peptide sequence tag of the amino terminus of Lhcb4.** The complete b- and y-ion series are indicated on the spectrum, but only the observed ions are shown in the *inset*. The peptide sequence acVFGVGKc (*ac*, acetyl amino terminus) was assigned with an Xcorr value of (2.108) when the database was modified to include the correct Lhcb4 amino terminus and static modification of the amino terminus was set to +42 Da (acetylation). Identities of the b- and y-ion series are indicated below the spectra for ease of alignment. The presence of Lhbc4 in the sample was confirmed by the presence of several additional internal peptides (86–103, 3.779; 104–113, 3.317; 121–135, 3.815; 222–234, 3.482; and 235–244, 3.078 (peptides numbered from Met[i], Xcorr score)).

CNBr peptide mass tags (Fig. 2*A*) show that the amino-terminal amino acid of processed Lhcb5 is (*N*-acetyl) Leu^{38} rather than Lys^{41}, which was predicted previously (96). The assignment of Lhcb5 was confirmed by MSMS sequence of tryptic fragments in the appropriate liquid chromatography fraction (Fig. 2*B*).

The peptide at *m/z* 4900.7 (Fig. 3) is assigned as the acetylated amino terminus of *A. thaliana* PsbS providing full agreement of measured and calculated mass (Table II) in confirmation of the previous identification of a 22,457.6-Da IMT as PsbS in spinach (16). Thus in *A. thaliana* the amino-terminal residue of PsbS is *N*-acetylleucine at position 53 (Fig. 3) rather than Ala^{60} (Swiss-Prot accession number Q9XF91) or Ala^{61} (96).

Three proteins (Lhcb1^{II1}, Lhcb1^{II2}, and PsbO^{III}) were identified by CNBr peptide mass tags but not detected as IMTs during LCMS. Each of these proteins is encoded by one copy

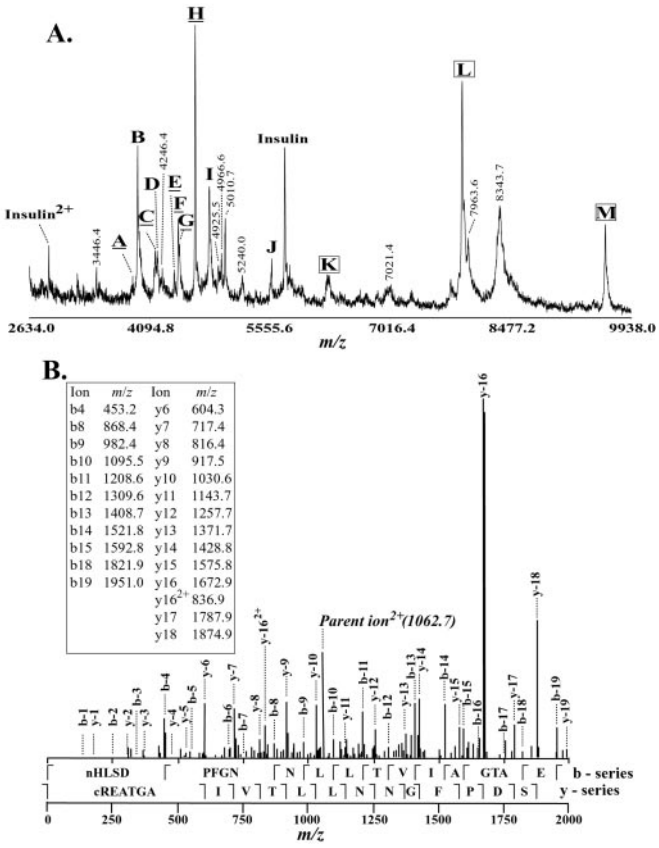

FIG. 2. **Assignment of proteolytic fragments from LCMS fraction 77 (77–78 min).** *A*, MALDI-TOF spectrum of CNBr fragments. To accommodate the mass range of interest the spectrum was collected in the linear mode; thus the ^{13}C isotopes of the peptide are not resolved, and average instead of monoisotopic masses were recorded. Identified peaks are labeled with *letters*, and unassigned peaks are labeled with their respective mass. The sample was spiked with an insulin standard for internal calibration. *Underlined* peak labels indicate carboxyl-terminal peptides, and *boxed* labels indicate amino-terminal peptides. The identified peaks are as follows: Peak A (measured mass (meas. mass) = 3885.8 Da, calculated mass (calc. mass) = 3885.4 Da, assignment (assgn.) = Lhca2^{III} peptide 222–257 (peptides numbered from Met^{1})), Peak B (meas. mass = 3946.5 Da, calc. mass = 3945.4 Da, assgn. = Peak L^{2+}), Peak C (meas. mass = 4161.7 Da, calc. mass = 4158.7 Da, assgn. = Lhcb4^{V} peptide 251–290), Peak D (meas. mass = 4186.1 Da, calc. mass = 4185.0 Da, assgn. = PsbX$_c$^{II} peptide 75–117) Peak E (meas. mass = 4390.5 Da, calc. mass = 4390.0 Da, assgn. = Lhcb1$^{I\alpha,I\beta,I\gamma,II2}$ peptide 227–267), Peak F (meas. mass = 4442.0 Da, calc. mass = 4441.0 Da, assgn. = Lhcb2$^{II\alpha}$ peptide 225–265), Peak G (meas. mass = 4455.2 Da, calc. mass = 4455.1 Da, assgn. = Lhcb2$^{II\beta}$ peptide 225–265), Peak H (meas. mass = 4646.1 Da, calc. mass = 4644.3 Da, assgn = Lhcb5^{IV} peptide 237–280), Peak I (meas. mass = 4818.9 Da, calc. mass = 4816.4 Da, assgn. = Peak M^{2+}), Peak J (meas. mass = 5575.2 Da, calc. mass = 5573.3 Da, assgn. = Lhcb2$^{II\alpha,II\beta,III}$ peptide 169–221), Peak K (meas. mass = 6251.0 Da, calc. mass = 6248.0 Da, assgn. = Lhca2^{III} peptide 46–102), Peak L (meas. mass = 7888.6 Da, calc. mass = 7890.7 Da, assgn. = Lhcb2$^{II\alpha,II\beta,III}$ + acetyl peptide 38–106), and Peak M (meas. mass = 9628.6 Da, calc. mass = 9632.9 Da, assgn. = Lhcb5^{IV} + acetyl peptide 38–122). *B*, MSMS spectrum of a tryptic peptide sequence tag of Lhcb5. The tryptic fragment corresponds to an internal peptide in Peak H above. The complete b- and y-ion series are indicated on the spectrum, but only the observed ions are shown in the *inset*. The peptide sequence HLSDPFGNNLLTVIAGDTER was assigned with an Xcorr value of (5.382). Identities of the b- and y-ion series are indicated below the spectra for ease of alignment.

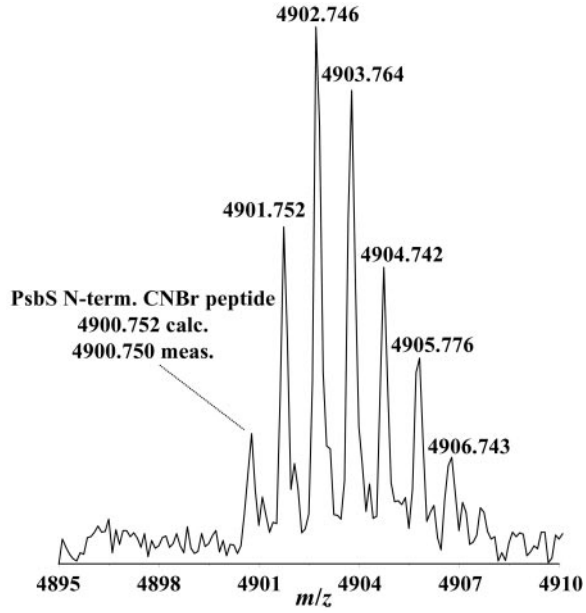

FIG. 3. **MALDI-TOF spectrum of the amino-terminal CNBr peptide of PsbS.** The spectrum was recorded in reflectron mode allowing resolution of ^{13}C isotopes, and all masses are monoisotopic. The measured mass of the peptide (4900.75 Da) is within 1 part/million of that calculated for amino-terminal peptide after cleavage of the signal sequence after position 52 and *N*-acetylation of residue 53. Thus this cleavage site is confirmed by intact protein mass and amino-terminal CNBr peptide mass as well as positive identification of PsbS by microliquid chromatography-MSMS (data not shown).

of a highly conserved multigene family and is present in low abundance compared with its respective paralog. The paralogs exhibit similar masses and similar retention times due to their high level of sequence conservation and therefore high conservation of biochemical/biophysical properties. Hence, the low abundance paralogs were "masked" by their higher abundance paralog(s) using LCMS (Fig. 4*A*). However, unique CNBr mass tags from each paralog could be detected in fractions collected during LCMS+. As an example, in *A. thaliana* the largest subunit of the oxygen-evolving enzyme (OEE1 or PsbO) is encoded on two paralogous genes on chromosomes III and V. The mature proteins encoded by these genes co-elute and differ by only 0.02% or 6 Da (26,571.7 Da (PsbO^{III}) *versus* 26,565.7 Da (PsbO^{V})). Molecular mass spectra of PsbO show a mass peak at 26,566.7 Da that corresponds to PsbO^{V} and a possible shoulder at 26,570 Da that may correspond to PsbO^{III} (Fig. 4*A*). When the LCMS+ data was analyzed unique internal CNBr fragments enabled discrimination between the two paralogs (Fig. 4*B*). Assuming similar ionization efficiency, it is estimated that PsbO^{V} is present in

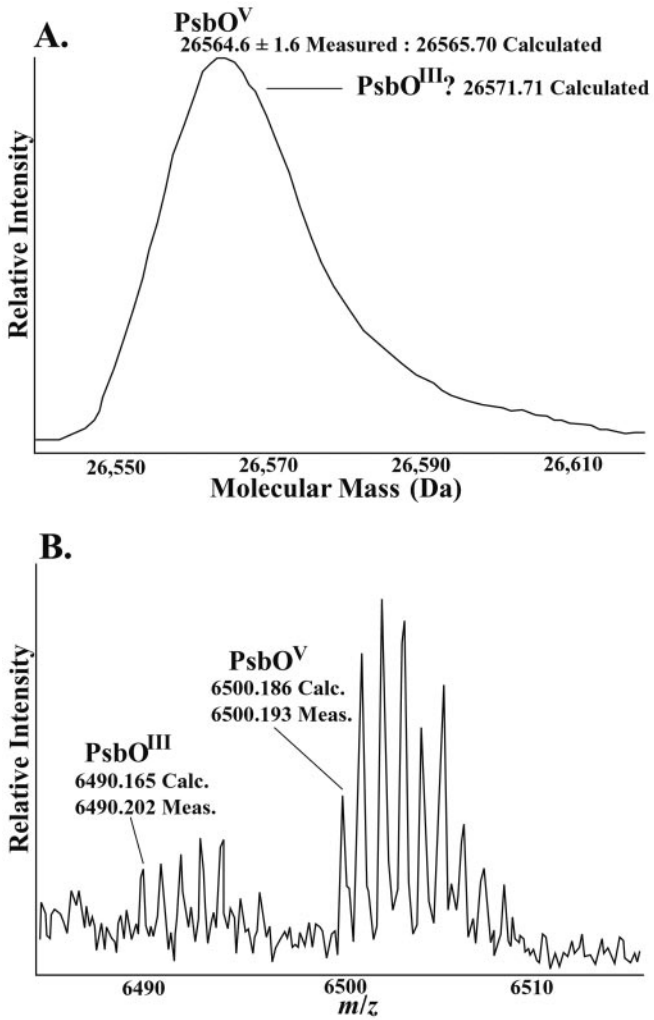

FIG. 4. **Resolution of PsbO isoforms by LCMS+.** *A*, reconstructed molecular mass spectrum of intact proteins recorded by electrospray ionization-MS. *B*, MALDI-TOF spectra recorded in reflectron mode of fraction 48 (48–50 min) CNBr peptides showing the presence of both PsbO isoforms: ^{103}EVKGTGTANQCPTIDGGSETFSFKAGKYTG-KKFCFEPTSFTVKADSVSKNAPPDFQNTKLh^{163}, PsbO^{III} *versus* ^{104}EVKGTGTANQCPTIDGGSETFSFKPGKYAGKKFCFEPTSFTVKA-DSVSKNAPPEFQNTKLh^{164}, PsbO^{V} (0.15% difference; *h*, homoserine lactone; sequence differences underlined). Note that analysis of LCMS+ fractions, in this case by CNBr treatment and MALDI-TOF, provided information that complemented the intact mass profiles generated in the original experiment.

steady-state amounts about 5× that of PsbO^{III} under the described growth conditions.

In addition, five membrane proteins from tobacco (*Nicotiana tabacum*) PS II membrane preparations were identified by their corresponding IMTs (Table III). The lack of sequence data from the nuclear genome complicates assignment of tobacco IMTs. Tentative identifications of these IMTs were based upon coincidence of measured and predicted mass after removal of the transit peptide (plus possible amino-terminal acetylation), and confidence in the assignments came from similarity between their liquid chromatography retention times and the retention times of orthologs from other plants (16).

Tests of Web-based Protein Analysis Programs

Chloroplast Targeting—The targeting predictions of five publicly available software packages listed on the ExPasy proteomic tools Web page (ca.expasy.org/tools/; ChloroP, TargetP, PSORT, Predotar, and MitoProt II) were tested with the 58-protein dataset (Table IV). ChloroP and TargetP correctly predicted chloroplast targeting in 57 and 56 instances, respectively. Previous proteomic surveys of soluble chloroplast proteins using 2D gel separation followed by tryptic digestion and assignment of the peptide fragments by matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) MS are in agreement with these results. ChloroP predicted that 31 of 34 (13) and TargetP predicted that 74 of 75 (46) chloroplast proteins identified by internal mass or sequence tags are targeted to the chloroplast. However, without an equivalent dataset of non-chloroplast proteins with explicitly determined amino termini, it is not possible to test for false positives using the same predictive criteria. A recent study of 80 proteins identified from the *A. thaliana* mitochondria proteome by 2D gel separation/MALDI-MS showed that TargetP predicted 54 to be targeted to the mitochondria. The remaining 16 (20%) proteins were predicted to be chloroplast-targeted, including mitochondrial proteins such as succinyl-CoA ligase, cytochrome *c* oxidase, and the mitochondrial elongation factor Tu (14).

Predotar predicted that 40 (69%) of the proteins in our dataset are targeted to the chloroplast, 2 are targeted to the mitochondria, 1 is targeted to both organelles, and 15 are targeted to neither. Predotar had similar success in predicting plastid targeting (58/75, 77%) when analyzing soluble chloroplast proteins identified by 2D gel/MALDI-MS (46). From our dataset, each protein predicted by Predotar not to be targeted to either organelle had a closely related paralog or ortholog (except for PsaH) predicted to be plastid-targeted. The proteins in this "neither" category appear to have no species bias. It is of interest that *A. thaliana* Lhcb1^{1a} and Lhcb1^{1b} were not predicted to be targeted to the same location despite having identical amino acid sequences except at position 20 in the transit peptide (Lys^{20} in Lhcb1^{1a} and Asn^{20} in Lhcb1^{1b}). In this case an additional positive charge (3 *versus* 2) in the transit peptide is apparently sufficient to predict an alternate location for the chloroplast protein.

PSORT attempts a more ambitious prediction of protein localization by assigning a suborganellar destination for nuclear encoded proteins. PSORT does not make a single prediction but ranks the four suborganellar compartments with the highest scores. One of the expert programs of PSORT is a hydrophobic moment analysis for predicting chloroplast proteins. The PSORT results can be interpreted in several

TABLE IV
Predicted organelle targeting by Web-based programs

Protein[a]	Species[b]	Localization[c]	ChloroP	TargetP	PSORT[d]	Predotar	MitoProt
PsaE^{IV}	At	S	Y	C	O:V:W:E/+	P	0.95
PsaE^{II}	At	S	Y	C	W:E:O:X/?	P	0.95
PsaE	So	S	Y	C	**S**:M:T:L/?	M	0.80
PsaD^{IV}	At	S	Y	C	N:**S**:M:X/?	P	0.96
PsaD	So	S	Y	C	N:**S**:M:**T**/?	P	0.94
Lhcb1	So	T+	N	C	X:S:M:**T**/?	N	0.45
PetC	Ps	T+	Y	C	P:**T**:S:L/+	P	0.36
Lhcb1*4	Ps	T+	Y	C	S:X:**T**:L/?	P	0.01
Lhcb1*2	Ps	T+	Y	C	S:**T**:L:X/+	N	0.81
Lhcb4^{V}	At	T+	Y	C	I:X:R:M/?	P	0.99
Lhcb5^{IV}	At	T+	Y	C	**T**:I:P:X/?	P	0.68
PsbS^{I}	At	T+	Y	C	**T**:P:G:I/−	N	0.92
Lhca3^{I}	At	T+	Y	C	S:M:**T**:X/+	P	0.96
Lhcb1–7	Nt	T+	Y	C	S:M:**T**:X/+	N	0.86
Lhcb1–40	Nt	T+	Y	C	S:**T**:L:X/+	P	0.70
Lhcb1^{II2}	At	T+	Y	C	S:X:**T**:L/+	N	0.85
Lhcb1–21	Nt	T+	Y	C	S:X:**T**:M/+	N	0.86
Lhcb2$^{II\alpha}$	At	T+	Y	C	**T**:X:S:L/+	P	0.86
Lhca4^{III}	At	T+	Y	C	X:S:M:**T**/?	P	0.97
AtpD^{IV}	At	T+	Y	C	S:**T**:L:M/?	P	0.88
PsbR^{I}	At	T+	Y	C	P:I:**T**:X/−	P	0.86
PetC^{IV}	At	T+	Y	C	**T**:S:R:L/+	P	0.63
PsbR	So	T+	Y	C	**T**:P:I:S/+	P	0.64
PsbR	Nt	T+	Y	C	P:**T**:I:X/?	N	0.78
Lhca2^{III}	At	T+	Y	C	X:O:W:E/+	P	0.98
Lhcb6I	At	T+	Y	M	X:O:W:E/?	P	0.99
Lhcb6	So	T+	Y	C	S:X:**T**:L/+	P	0.97
PetM^{II}	At	T+	Y	C	I:**T**:P:M/?	P	0.01
PsbS	So	T+	Y	C	P:G:W:E/−	P	0.60
Lhcb2	Ps	T+	Y	C	N:X:**T**:S/−	N	0.93
PsaH^{III}	At	T+	Y	C	S:X:**T**:L/+	M	0.84
PsaH^{I}	At	T+	Y	C	S:X:**T**:L/+	N	0.93
PsaH	So	T+	Y	C	C:M:X:**T**/−	P	0.74
Lhcb2$^{II\beta}$	At	T+	Y	C	**T**:X:S:L/+	P	0.93
PetC	So	T+	Y	C	L:S:**T**:M/+	P	0.16
AtpC^{IV}	At	T+	Y	C	S:**T**:L:M/+	P	0.80
Lhcb3	Ps	T+	Y	C	I:X:P:M/−	B	0.70
Lhcb2^{III}	At	T+	Y	C	W:X:**T**:P/?	P	0.96
Lhcb1$^{I\alpha}$	At	T+	Y	C	S:X:**T**:L/+	N	0.88
Lhcb1$^{I\beta}$	At	T+	Y	C	S:**T**:X:L/+	P	0.83
Lhcb1$^{I\gamma}$	At	T+	Y	C	S:**T**:X:L/+	P	0.83
Lhcb1^{II1}	At	T+	Y	C	X:S:M:T/?	P	0.84
Lhcb1–50	Nt	T+	Y	C	S:**T**:L:M/+	N	0.81
Lhcb3^{V}	At	T+	Y	C	**T**:I:X:S/+	P	0.32
PsbW	So	T−	Y	C	**T**:I:S:L/+	P	0.93
PsbX$_c$^{II}	At	T−	Y	C	I:**T**:P:R/?	P	0.84
PsbW^{II}	At	T−	Y	C	I:**T**:P:R/+	P	0.66
PsbO	Ps	Ls	Y	M	**L**:R:S:M/+	N	0.45
PsbO^{III}	At	Ls	Y	C	**L**:R:S:T/+	P	0.85
PsbO^{V}	At	Ls	Y	C	**L**:S:R:T/+	P	0.30
PsaN	At	Lt	Y	C	P:W:G:E/?	P	0.29
PsbQ	So	Lt	Y	C	**L**:R:S:T/+	N	0.54
PsbQ^{IV1}	At	Lt	Y	C	**L**:C:S:X/?	P	0.24
PsbQ$^{IV\alpha}$	At	Lt	Y	C	**L**:R:S:M/?	P	0.18
PsbX^{III}	At	Lt	Y	C	R:M:I:Q/−	P	0.45
PsbP	So	Lt	Y	C	**L**:S:T:N/+	N	0.39
PsbP^{I}	At	Lt	Y	C	**L**:S:R:T/+	P	0.63
PsbP	Ps	Lt	Y	C	R:C:M:T/−	N	0.44

ways depending upon the criteria used for determining a "hit." Under the strictest conditions, namely that the highest score must match the actual suborganellar destination and that the hydrophobic moment must predict that it is a chloroplast protein, PSORT does poorly, predicting 0/5 in the stroma, 6/42 in the thylakoid, and 6/11 in the lumen. If the requirement is that only one of the four highest scores matches the correct destination, then PSORT does much better (stroma, 3/5; thylakoid, 37/42; and lumen, 8/11). If the criteria for PSORT are similar to ChloroP and Predotar (chloroplast targeting or not), then the results are comparable with Predotar. PSORT correctly predicts 49/58 if at least one of the four highest scores is a chloroplast domain, 39/58 if at least two of the four high scoring results are to the chloroplast, and 25/58 if at least three of the four are in the chloroplast. PSORT was used previously to test a smaller set of soluble chloroplast proteins identified by mass and sequence tags. In this study PSORT predicted 13/34 proteins target to the chloroplast if only one of the four predicted destinations is in the chloroplast, 10/34 if two of the four are chloroplast domains, and 1/34 if three of the four predicted targets are chloroplast domains (13). Less strict interpretations are reliable if the suborganellar destination of the protein being examined is already known. It is therefore suggested that assignment of an unknown/hypothetical protein as a chloroplast protein should only be considered when three of the four PSORT suborganellar predictions are to chloroplast compartments. Furthermore, the tentative assignment should be confirmed by further experimentation.

MitoProt was initially used as a negative control, but surprisingly it predicted that a large number of stromal proteins and thylakoid IMPs (16/47, cutoff >0.90; 32/47, cutoff >0.80) would be targeted to the mitochondria (Table IV). MitoProt predicted that proteins targeted to the lumen would not go to the mitochondria. Although the size of the datasets for stromal and luminal proteins is small, the trend is striking. MitoProt identified 53 of 80 (14) and 39 of 48 (15) confirmed mitochondrial proteins as being trafficked to the mitochondria using a cutoff of >0.85. The presence of chloroplast proteins in tests of MitoProt reduced the reliability of the predictions (94). Here it is demonstrated that for certain chloroplast proteins the probability of MitoProt incorrectly predicting a mitochondrial destination is high, and thus one recommendation that emerged from these tests is that MitoProt results from plastid-containing eukaryotes be screened against ChloroP/TargetP to help reduce false positive predictions.

Transit Peptide Cleavage Prediction—The transit peptide is cleaved by SPP after translocation across the chloroplast envelope into the stroma. Proteins destined for the lumen have a bipartite transit peptide that is removed by TPP (22). ChloroP, TargetP, and MitoProt each contain transit peptide cleavage predictions. TargetP uses the ChloroP cleavage assignment if it predicts the protein is chloroplast-targeted. Only two of the test proteins were predicted by TargetP to go to the mitochondria, therefore only the ChloroP results are presented (Table V). SignalP was tested because the two targeting pathways to the lumen, Sec-dependent and ΔpH-dependent, have analogous systems in eukaryotic, Gram-positive (Gm+), and Gram-negative (Gm−) bacterial secretory pathways (22, 46).

The original test results suggested that 60% of the nuclear encoded chloroplast proteins tested by ChloroP have a predicted cleavage site within 2 amino acid residues of the actual cleavage site (91). We used this 2-amino acid threshold to determine whether a peptide cleavage prediction was "correct" or not (boxes in Table V). Using this cutoff and separating the proteins into stroma, thylakoid, and lumen domains, the successes of predicting the cleavage sites of stromal proteins were as follows: ChloroP, 5/5 (100%); MitoProt, 2/5 (40%); SignalP (Gm+), 1/5 (20%); and SignalP (eukaryotic/Gm−), 0/5 (0%). Predictions of the cleavage sites for luminal proteins were as follows: SignalP (eukaryotic/Gm+), 11/11 (100%); Signal P (Gm−), 9/11 (82%); and ChloroP and MitoProt, 0/11 (0%). The following are predictions for the thylakoid IMPs: MitoProt, 18/42 (43%); ChloroP, 9/42 (21%); SignalP (Gm−), 3/42 (7%); SignalP (Gm+), 1/42 (2%); and SignalP (eukaryotic), 0/42 (0%). The data for stromal and luminal proteins suggest that ChloroP is best for predicting the cleavage site of the transit peptide by SPP, whereas the eukaryotic settings of SignalP are best for predicting the processing site by thylakoid processing protease.

None of the programs adequately predict the transit peptide processing site of the membrane-spanning proteins inserted into the thylakoid via the signal recognition particle-dependent pathway. ChloroP correctly predicted the transit peptide cleavage site for one thylakoid IMP (*A. thaliana* PetC^{IV}, the cytochrome b_6/f complex Rieske Fe-S protein), but

[a] *A. thaliana* and tobacco protein annotations as described in Tables II and III. Pea and spinach annotations using standard genes.

[b] At, *A. thaliana*; Ps, *Pisum sativum*; So, *Spinacia oleracea*; Nt, *N. tabacum*.

[c] Chloroplast localization. S, stroma; T+, thylakoid membrane via the signal recognition particle-dependent pathway; T−, thylakoid membrane via the spontaneous pathway; Ls, lumen via the Sec pathway; Lt, lumen via the Tat pathway.

[d] Prediction of protein targeting using PSORT. E, endoplasmic reticulum lumen; W, endoplasmic reticulum membrane; X, peroxisome; P, plasma membrane; N, nucleus; Q, mitochondria outer membrane; I, mitochondria inner membrane; R, mitochondria intermembrane space; M, mitochondria lumen; S, chloroplast stroma; T, chloroplast thylakoid membrane; L, chloroplast lumen; O, secreted; V, vacuole; C, cytosol; G, Golgi body. The correct localizations are in bold. The PSORT hydrophobic moment analysis for chloroplast proteins output is indicated by positive (+), negative (−), or unknown (?).

TABLE V
Transit peptide cleavage prediction by Web-based programs

Protein[1]	Species[2]	Loc.[3]	ChloroP Δ[4]	SignalP Euk. Δ[4]	SignalP Gram+ Δ[4]	SignalP Gram- Δ[4]	MitoProt Δ[4]	N-term by MS[5]
PsaE^{IV}	At	S	0	23	22	-90	3	nA45
PsaE^{II}	At	S	0	25	-1	-31	0	nA47
PsaE	So	S	0	19	-7	-7	3	nA35
PsaD^{IV}	At	S	0	46	-9	46	12	nE46
PsaD	So	S	0	50	-4	31	-1	nA50
Lhcb1	So	T+	-11	-140	-198	21	36	acR36
PetC	Ps	T+	-10	-32	-119	-55	51	nA51
Lhcb1*4	Ps	T+	-9	-140	-198	21	37	acR37
Lhcb1*2	Ps	T+	-9	-140	-198	20	-2	acR38
Lhcb4^{V}	At	T+	-8	-230	-230	-228	2	acV33
Lhcb5^{IV}	At	T+	-7	-210	-10	4	15	acL38
PsbS^{I}	At	T+	-7	-98	-8	-98	7	acL53
Lhca3^{I}	At	T+	-7	-205	42	23	13	acA42
Lhcb1-7	Nt	T+	-6	-96	-198	21	-2	acR36
Lhcb1-40	Nt	T+	-6	-86	-198	21	-2	acR36
Lhcb1^{III2}	At	T+	-4	-139	-198	19	-2	acR34
Lhcb1-21	Nt	T+	-4	-138	-196	21	-2	acR36
Lhcb2$^{II\beta}$	At	T+	-4	-136	-194	-194	-2	acR38
Lhca4^{III}	At	T+	-3	-65	-173	9	11	nK53
AtpD^{IV}	At	T+	-1	48	48	-11	0	nS48
PsbR^{I}	At	T+	-1	-88	-88	-88	17	NS42
PetC^{IV}	At	T+	0	-47	-152	-152	51	nA51
PsbR	So	T+	1	-98	-98	-98	0	nS42
PsbR	Nt	T+	1	-98	-98	-98	2	nS38
Lhca2^{III}	At	T+	1	-108	-112	33	6	nV46
Lhcb6^{I}	At	T+	1	-76	-85	27	23	nA48
Lhcb6	So	T+	1	-201	-73	-3	-3	nA52
PetM^{II}	At	T+	1	-39	-37	86	-30	nN86
PsbS	So	T+	3	-97	-8	-8	63	acL63
Lhcb2	Ps	T+	3	-136	-194	-194	-2	acR38
PsaH^{III}	At	T+	6	-70	-70	-66	5	nK51
PsaH^{I}	At	T+	6	-70	-70	-66	5	nK51
PsaH	So	T+	6	-70	-72	-72	5	nK50
Lhcb2$^{II\alpha}$	At	T+	7	-136	-194	-194	-2	acR38
PetC	So	T+	7	-47	-118	-48	8	nA69
AtpC^{IV}	At	T+	8	51	7	32	7	nA51
Lhcb3	Ps	T+	9	-130	-189	-189	11	nG43
Lhcb2^{III}	At	T+	10	39	-194	-194	-2	acR39
Lhcb1$^{I\alpha}$	At	T+	12	-139	-198	21	-2	acR36
Lhcb1$^{I\beta}$	At	T+	12	-139	-198	21	-2	acR36
Lhcb1$^{I\gamma}$	At	T+	12	-139	-198	21	-2	acR36
Lhcb1^{III}	At	T+	12	-138	-197	-197	-2	acR36
Lhcb1-50	Nt	T+	12	-86	-198	21	-2	acR36
Lhcb3^{V}	At	T+	20	-130	-189	-189	11	nG43
PsbW	So	T-	14	-41	-41	0	42	nL84
PsbX$_c^{II}$	At	T-	16	-35	0	0	31	nA75
PsbW^{II}	At	T-	35	-43	52	0	80	nL80

for the spinach and pea orthologs it predicted a cleavage site 7 amino acids carboxyl-proximal and 10 amino acids amino-proximal, respectively (Table V). The majority of the correct predictions for MitoProt are for the Lhcb1 and Lhcb2 proteins. If MitoProt predicted the presence of a mitochondrial transit peptide, then it consistently predicted a cleavage site 2 amino acids carboxyl-proximal from the actual site in the Lhcb1 and Lhcb2 proteins. ChloroP predictions ranged from 11 amino

TABLE V—*continued*

Protein[1]	Species[2]	Loc.[3]	ChloroP Δ[4]	SignalP Euk. Δ[4]	SignalP Gram+ Δ[4]	SignalP Gram- Δ[4]	MitoProt Δ[4]	N-term by MS[5]
PsbO	Ps	Ls	37	0	0	0	62	nE82
PsbO^{III}	At	Ls	56	0	0	0	85	nE85
PsbO^{V}	At	Ls	56	0	0	0	61	nE86
PsaN^{V}	At	Lt	5	2	2	70	34	nG87
PsbQ	So	Lt	29	0	-2	1	44	nE84
PsbQ^{IV1}	At	Lt	31	0	-2	1	34	nD76
PsbQ$^{IV\alpha}$	At	Lt	34	0	-2	83	21	nE83
PsbX^{III}	At	Lt	41	2	2	2	41	nE75
PsbP	So	Lt	42	0	0	0	53	nA82
PsbP^{I}	At	Lt	47	0	0	0	48	nA78
PsbP	Ps	Lt	51	0	0	0	35	nA74

[1] *A. thaliana* and tobacco protein annotations as in Tables II and III. Pea and spinach annotations using standard gene names.

[2] Species abbreviations as in Table IV.

[3] Chloroplast localization as in Table IV.

[4] Δ, difference between observed and predicted amino-terminal residue. Positive numbers indicate the predicted cleavage site is amino-proximal to the observed amino terminus. Negative numbers indicate the predicted cleavage site is carboxyl-proximal to the observed amino-terminus. Boxes indicate that the predicted amino terminus is within ± 2 amino acids of the amino-terminal amino acid observed by MS. Gray boxes indicate program that best predicts the cleavage site for proteins in the stroma and the lumen. The special case of PsbW and PsbX$_c$ is discussed in the text. Euk, eukaryotic.

[5] Amino terminus determined by LCMS. n, unblocked alpha amine; ac, acetylated alpha amine.

acids amino-proximal to 12 amino acids carboxyl-proximal to the actual site for the same set of proteins (Table V). The appearance that MitoProt does better at predicting the transit peptide cleavage site of thylakoid IMPs is more likely due to its consistency at predicting the same cleavage site for orthologous proteins than due to recognition of some specific thylakoid-localizing domain. It is somewhat surprising that the MitoProt algorithm, which was not designed to work with chloroplast-targeted proteins, is much more consistent in its predictions of where a putative cleavage site is located than ChloroP, which was specifically created to analyze chloroplast proteins.

PsbW and PsbX$_c$[2] are the only thylakoid-bound proteins in the dataset that have amino termini oriented into the thylakoid lumen (39, 41). They are also the only proteins in the dataset that are targeted to the thylakoid by the spontaneous insertion pathway (42). The Gram-negative SignalP program correctly predicted the transit peptide processing sites for the two PsbW and single PsbX$_c$ examples in the study. We also tested the two other membrane proteins, PsbY and AtpG, which are inserted into the thylakoid via the spontaneous mechanism (42). Two PsbY proteins (spinach and *A. thaliana*, Swiss-Prot accession numbers P80470 and O49347, respectively) and two AtpG proteins (subunit II of the CF$_0$ ATPase; spinach, Swiss-Prot accession number P31853; and *A. thaliana*, Swiss-Prot accession number Q42139) are listed in the Swiss Protein/TrEMBL databases in addition to the single PsbX$_c$ and two PsbW proteins in the dataset. Bacterial signal peptides are cleaved according to the (−3, −1) rule where the −3 and −1 positions before the cleavage site are small neutral amino acids (usually Ala) (95, 97). Eukaryotic and Gm+ signal peptide consensus sequences generally have hydrophobic amino acids in the region between −16 and −25, whereas Gm− signal peptides are more likely to have hydrophilic or positively charged amino acids in this region, like the transit peptides of thylakoid proteins imported via the spontaneous path. The Gm− SignalP program correctly predicted the processing site of all seven proteins that are imported into the thylakoid membrane by this mechanism (data not shown). This unexpected observation suggests that the spontaneous thylakoid import mechanism may be related to the secretory mechanism of Gram-negative bacteria.

DISCUSSION

There is clearly a need for reliable bioinformatics software for predicting protein destinations and processing sites in large proteomics datasets. However, the requirement for ex-

[2] *psbT* is the name of the chloroplast gene encoding the PS II reaction center T protein. *psbT* was also used as the name for the nuclear gene encoding the soluble PS II M_r 5000 oxygen-encoding enzyme protein targeted to the lumen, later renamed *psbX*. Unfortunately, *psbX* also was used for a gene encoding a membrane-spanning PS II protein of unclear function in cyanobacteria and the plastids of rhodophyta and cryptophyta. Orthologs of this cyanobacterial gene have been found in *A. thaliana* (98) and rice (Sasaki, T., Matsumoto, T., and Yamamoto, K. (2001) GenBank™/EBI accession number AP004300), and this gene is UV-B-repressed in pea (Liu, L., White, M. J., and MacRae, T. H. (2001) GenBank™/EBI accession number AY065654). We label this gene as *psbX$_c$*. The use of PsbX for unrelated soluble and membrane-bound PS II proteins has created considerable confusion.

TABLE VI
Analysis of the teaching dataset used by ChloroP (www.cbs.dtu.dk/services/ChloroP/pages/datasets.html)

Examination of the ChloroP training dataset	
Amino terminus explicitly determined for the Swiss-Prot entry	49
Amino terminus determined in orthologous protein, "By Similarity" annotation omitted in Swiss-Prot	9
Amino terminus not experimentally determined, "Probable" or "Potential" omitted in Swiss-Prot	8
Amino terminus annotation is incorrect in Swiss-Prot when compared to the literature	6
Amino terminus from translation products imported into heterologous plastids or from transformed *E. coli*	3
Total	75
ChloroP training dataset is biased to proteins targeted to the stroma	
Stroma	62
Thylakoid	12
Envelope	1
Lumen	0
ChloroP training dataset is biased towards two species and has a significant number of algal sequences	
Streptophyta	62
Spinach	18
Pea	15
Chlorophyta	12
Chlamydomonas reinhardtii	8
Rhodophyta	1

perimentally verified amino and carboxyl termini of the mature proteins is rarely met because of the typically incomplete protein coverage obtained in standard peptide mass or sequence tag protein identification experiments. Thus, the LCMS+ approach is indispensable because it provides intact protein mass information as well as eliminating peptide losses associated with recovery of peptides from gels. The intact mass usually yields the amino-terminal cleavage site, although incorrect assignment of the carboxyl terminus could potentially confound assignments based upon the intact mass alone. Finding the amino-terminal peptide in digested LCMS+ fractions using standard software packages remains challenging because protein masses are generally calculated from translated mRNA and include the transit peptide in the mass calculation. By not specifying a particular digest it is possible to detect non-tryptic peptides potentially representing the amino terminus unless they are acetylated at the amino terminus, in which case it is necessary to modify all peptide amino termini for searching. Although a non-tryptic cleavage at the amino terminus of a peptide near the predicted cleavage site may locate the amino terminus, this is hypothetical unless it is also N_{α}-acetylated or in agreement with the measured intact mass. The current state of the art does not permit rapid prediction of the mass of putative chloroplast proteins after transit peptide cleavage, and there is no method for predicting the acetylation of the amino terminus that is common for thylakoid-bound proteins.

All the programs that were tested were deficient in one or more of their predictive algorithms, and none adequately predict the transit peptide processing site of the membrane-spanning proteins in the thylakoid. The training dataset used for the neural networks of ChloroP is the only one readily available for examination (www.cbs.dtu.dk/services/ChloroP/pages/datasets.html). This dataset was created by extracting from Swiss-Prot release 35 all entries with an FT line containing the key word "TRANSIT" and description word "CHLOROPLAST." Entries with a chloroplast transit peptide marked "BY SIMILARITY," "PROBABLE," or "POTENTIAL" were removed. The remaining entries were screened with SignalP, and those that showed the presence of the bipartite transit peptide for trafficking to the thylakoid or lumen were removed. A final set of 75 proteins was obtained after homology reduction and removal of a few mistargeted proteins (91). The removal of thylakoid- and lumen-targeted proteins essentially makes ChloroP a stroma-targeting and processing prediction algorithm (Table VI). Because the thylakoid- and lumen-localized proteins must pass through the chloroplast envelope to the stroma the chloroplast-targeting predictions are correct, but the transit peptide cleavage predictions of ChloroP are only for proteins cleaved by SPP. Thus the ChloroP Web page is somewhat misleading in presenting results of the transit peptide cleavage prediction as the amino terminus for all chloroplast-targeted proteins rather than just the stromal ones.

Our re-examination of the ChloroP dataset (Table VI) shows several problems that are probably common to all such training/testing datasets. Of the 75 proteins in the dataset only 43 (57%) had explicitly determined transit peptide cleavage sites. Nine of the remaining sequences had a cleavage site that should have been labeled in the Swiss-Prot entry "BY SIMILARITY" because the amino terminus indicated in the entry was predicted from an ortholog that had an experimentally verified amino terminus. There were 15 entries where no experimental evidence confirming the cleavage site had been obtained from any plant or algae and where the annotation "POTENTIAL" or "PROBABLE" was omitted. Five more entries that should have had the transit peptide labeled "PROBABLE" have been shown previously by experimental peptide sequence data to be cleaved at a site different from the annotation in the Swiss Protein Database (accession numbers

P11893, P17067, P09195, P12360, and P07370). Three of the five entries listed above had experimental confirmation of the amino terminus from orthologous proteins published several years prior to Swiss-Prot release 35 and are still incorrectly annotated (70, 87, 99). Of the three remaining entries, two were cleavage sites determined from proteins made in cell-free extracts and imported into chloroplasts of heterologous species (Swiss-Prot accession numbers P15102 and P00873), and one was based on the amino terminus found after recovery of protein expressed in transgenic *Escherichia coli* (Swiss-Prot accession number P12629).

Examples of misannotations include the entries for the tomato chloroplast biosynthetic threonine dehydratase (Swiss-Prot accession number P25306), the barley chloroplast photosystem I reaction center subunit XI (PsaL) (Swiss-Prot accession number P23993), and the safflower chloroplast acyl-[acyl-carrier protein] desaturase (Swiss-Prot accession number P22243). The tomato threonine dehydratase has Lys^{52} annotated as the experimentally confirmed amino terminus of the processed preprotein. The work referenced for this annotation did not include protein sequencing, and the 10-amino acid sequence annotated as the amino-terminal protein sequence in the Swiss Protein Database was from a figure where a 10-amino acid sequence was underlined to indicate the proposed region for the transit peptide cleavage site (100). The barley PsaL protein is blocked at the amino terminus of the processed protein, and the annotated amino terminus (Ala^{41}) found in the Swiss Protein Database was one of two proposed sites for the transit peptide cleavage site (101). The amino terminus of the processed PsaL protein has not been verified in vascular plants and should have been annotated "POTENTIAL."

The amino terminus of the safflower acyl-[acyl-carrier protein] desaturase is also blocked (102), and the annotation indicating an experimentally determined amino terminus (Ala^{34}) is from two overlapping peptide sequences at the amino-proximal end of a peptide map that begins with the same amino acid. This transit peptide should have been annotated as "PROBABLE." It is not surprising that the gargantuan task of annotating the Swiss Protein Database results in frequent unclear or misannotated entries. A suggested improvement in the Swiss Protein Database would be to link the "BY SIMILARITY" annotation to the entry to which it is referring. Although this would not solve the problems encountered when the annotation is missing, it would greatly speed up the ability to cross-reference predicted modifications and functional groups. The current situation requires time-consuming, careful, and diligent confirmation of the annotation of each entry before attempting to correlate intact protein mass tags with translated genomic sequence information, especially for proteins that have amino-terminal trafficking peptides. The same sort of diligence should also be utilized when creating software training/testing datasets, thereby minimizing ambiguities or errors that decrease confidence in the bioinformatics tools trained/tested with these datasets.

The range of species used in the ChloroP training set is narrow due to the lack of experimental verification of post-translational modifications for most proteins (see Table I), and the dataset is skewed toward proteins from pea and spinach. The training set also has a significant number of algal sequences (Table VI), but algal transit peptides are 32 amino acids shorter, on average, than those from vascular plants (103). Where direct comparisons between orthologous proteins are possible, the transit peptide cleavage sites are not conserved between the algae and vascular plants. It is reasonable that separate training datasets are needed for vascular plants and green algae, as is done in SignalP for the eukaryotes, Gram-positive bacteria, and Gram-negative bacteria. Additionally, it seems reasonable to use separate training sets for each suborganellar compartment or training sets for each characterized translocation mechanism. In this way the training sets would narrow the focus of the neural net programs and help to avoid the confusion created by attempting to find a generalized transit peptide cleavage motif that may not be universally applicable for all photosynthetic organisms.

The results in Table V suggest that there may be at least four proteases in the chloroplast involved in removing transit peptides: a stromal processing protease, for which the cleavage site can be predicted by ChloroP for proteins targeted to the stroma; a lumen processing protease, for which the cleavage site can be predicted by SignalP (eukaryotic) for proteins imported into the lumen via the Sec or Tat mechanisms; a thylakoid processing protease, for which the cleavage site is predicted by SignalP (Gm−) for proteins inserted into the membrane via the spontaneous mechanism; and a thylakoid protease, for which no reliable cleavage prediction program exists for proteins inserted into the membrane via a signal recognition particle-dependent mechanism. Attempts to identify common features of either transit peptides or mature proteins that might be useful for such predictions have thus far been unfruitful.

Previous proteomic studies of *A. thaliana* organelles that tested the trafficking predictions of various predictive programs (13, 14) are in general agreement with the results we observe for trafficking. However, because the earlier studies identified proteins based upon internal sequence or peptide mass tags, they were unable to assess the cleavage prediction routines. Subsequent studies of luminal proteins in *A. thaliana* analyzed the cleavage prediction routines of TargetP and SignalP (46, 47). Ten proteins overlap between these studies and ours: six lumen ($PsbQ^{VI1}$, $PsbQ^{VI\alpha}$, $PsbP^{I}$, $PsaN^{V}$, $PsbO^{V}$, and $PsbO^{III}$ with The Institute for Genomic Research (TIGR) chromosomes locus numbers t4g21280, At4g05180, At1g06680, At5g64040, At5g66570, and At3g50820, respectively), two stroma ($PsaD^{IV}$ and $PsaE^{IV}$ with TIGR chromosome locus numbers At4g02770 and At4g28570, respectively), and two thylakoid-integral ($AtpC^{IV}$ and $AtpD^{IV}$ with TIGR

chromosome locus numbers At4g04640 and At4g09650, respectively). SignalP was better at predicting the amino terminus of luminal proteins than TargetP (46), which is in agreement with our observations; however, only three ($PsbP^{I}$, $PsbO^{V}$, and $PsbO^{III}$) of the luminal proteins that overlapped with our study used the correct amino terminus when testing the programs. In contrast, our results agree completely with the experimentally determined amino termini for the oxygen-evolving enzyme proteins in the second study (47).

Integral membrane proteins are predicted to account for at least one third of the open reading frames in the *A. thaliana* genome (104). The 2D methods currently in use are biased against low abundance proteins (105, 106) and IMPs. However, IMPs with masses up to and exceeding 100,000 Da and containing up to 15 membrane-spanning α-helices have now been successfully characterized by LCMS using intact mass tags (5, 16, 49, 88, 107–111).[3] LCMS+ allows us to subject fractions to off-line digestion to generate peptide fragments for mass tag (112–115) and sequence tag (116, 117) experiments to confirm the IMT data (49, 118, 119). Characterizing a protein based upon its intact mass tag allows us to simultaneously determine secondary post-translational modifications for several proteins. LCMS+ is the only method currently available to rapidly determine such modifications for a large number of proteins. Consequently, LCMS+ should be the method of choice for obtaining datasets to train/test predictive bioinformatics tools for trafficking and post-translational processing.

* This work was supported by National Institutes of Health Grants AI-12601 and AI-29733 (to J. P. W.) and United States Department of Energy Grant DE-FG03-01ER15253 (to J. P. W. and K. F. F.). The National Institutes of Health, the Pasarow Foundation, and the W. M. Keck Foundation provided funds toward instrument purchases for this study. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

§ Present address: Pocagua Agricultural Systems, 718 10th St. SW, Albuquerque, NM 87102.

** To whom correspondence should be addressed. Tel.: 310-794-5156; Fax: 310-206-2616; E-mail: jpw@chem.ucla.edu.

REFERENCES

1. The *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282,** 2012–2018
2. The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796–815
3. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921
4. Laemmli, U. K. (1970) Cleavage of structural proteins during the assembly of the head of the bacteriophage T4. *Nature* **227,** 680–685
5. Whitelegge, J. P., Gundersen, C. B., and Faull, K. F. (1998) Electrospray-ionization mass spectrometry of intact intrinsic membrane proteins. *Protein Sci.* **7,** 1423–1430
6. Washburn, M. P., Wolters, D., and Yates, J. R., III (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19,** 242–247
7. Whitehouse, C. M., Dreyer, R. N., Yamshita, M., and Fenn, J. B. (1985) Electrospray interface for liquid chromatographs and mass spectrometers. *Anal. Chem.* **57,** 675–679
8. Tanaka, K., Waki, H., Ido, Y., Akita, S., Yoshida, Y., and Yoshida, T. (1988) Protein and polymer analysis up to *m/z* 100,000 by laser ionization time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **2,** 151–153
9. Karas, M., and Hillenkamp, F. (1988) Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal. Chem.* **60,** 2299–2301
10. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., and Whitehouse, C. M. (1989) Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246,** 64–71
11. Marienfeld, J., Unseld, M., Brandt, P., and Brennicke, A. (1996) Genomic recombination of the mitochondrial *atp6* gene in *Arabidopsis thaliana* at the protein processing site creates two different presequences. *DNA Res.* **3,** 287–290
12. Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., and Tabata, S. (1999) Complete structure of the chloroplast genome of *Arabidopsis thaliana*. *DNA Res.* **6,** 283–290
13. Peltier, J.-B., Friso, G., Kalume, D. E., Roepstorff, P., Nilsson, F., Adamska, I., and van Wijk, K. J. (2000) Proteomics of the chloroplast: systematic identification and targeting analysis of lumenal and peripheral thylakoid proteins. *Plant Cell* **12,** 319–341
14. Millar, A. H., Sweetlove, L. J., Giegé, P., and Leaver, C. J. (2001) Analysis of the *Arabidopsis* mitochondrial proteome. *Plant Physiol.* **127,** 1711–1727
15. Kruft, V., Eubel, H., Jänsch, L., Werhahn, W., and Braun, H.-P. (2001) Proteomic approach to identify novel mitochondrial proteins in *Arabidopsis*. *Plant Physiol.* **127,** 1694–1710
16. Gómez, S. M., Nishio, J. N., Faull, K. F., and Whitelegge, J. P. (2002) The chloroplast grana proteome defined by intact mass measurements from liquid chromatography mass spectrometry. *Mol. Cell. Proteomics* **1,** 46–59
17. Cline, K., and Henry, R. (1996) Import and routing of nucleus-encoded chloroplast proteins. *Annu. Rev. Cell Dev. Biol.* **12,** 1–26
18. Fuks, B., and Schnell, D. J. (1997) Mechanism of protein transport across the chloroplast envelope. *Plant Physiol.* **114,** 405–410
19. Lübeck, J., Heins, L., and Soll, J. (1997) Protein import into chloroplasts. *Physiol. Plant.* **100,** 53–64
20. Heins, L., Collinson, I., and Soll, J. (1998) The protein translocation apparatus of chloroplast envelopes. *Trends Plant Sci.* **3,** 56–61
21. Dalbey, R. E., and Robinson, C. (1999) Protein translocation into and across the bacterial plasma membrane and the plant thylakoid membrane. *Trends Biochem. Sci.* **24,** 17–22
22. Robinson, C., Thompson, S. J., and Woolhead, C. (2001) Multiple pathways used for the targeting of thylakoid proteins in chloroplasts. *Traffic* **2,** 245–251
23. Waegemann, K., and Soll, J. (1996) Phosphorylation of the transit sequence of chloroplast precursor proteins. *J. Biol. Chem.* **271,** 6545–6554
24. Schnell, D. J., Blobel, G., Keegstra, K., Kessler, F., Ko, K., and Soll, J. (1997) A consensus nomenclature for the protein-import components of the chloroplast envelope. *Trends Cell Biol.* **7,** 303–304
25. Mereschkowsky, C. (1910) Theorie der zwei Plasmaarten als Grundiage der Symbiogenesis, einer neuen Lehre von der Entstehung der Organismen. *Biol. Zentralbl.* **30,** 278–303, 321–347, 353–367
26. Goksøyr, J. (1967) Evolution of eukaryotic cells. *Nature* **211,** 1161
27. Margulis, L. (1970) *Origin of Eukaryotic Cells*, pp. 276–293, Yale University Press, New Haven, CT
28. Flavell, R. (1972) Mitochondria and chloroplasts as descendants of prokaryotes. *Biochem. Genet.* **6,** 275–291
29. Yuan, J., Henry, R., McCaffery, M., and Cline, K. (1994) SecA homolog in protein transport within chloroplasts: evidence for endosymbiont-derived sorting. *Science* **266,** 796–798
30. Schuenemann, D., Amin, P., Hartmann, E., and Hoffman, N. E. (1999) Chloroplast SecY is complexed to SecE and involved in the translocation of the 33-kDa, but not the 23-kDa subunit of the oxygen-evolving

[3] J. P. Whitelegge and S. J. Karlish, unpublished data.

complex. *J. Biol. Chem.* **274,** 12177–12182
31. Mould, R. M., and Robinson, C. (1991) A proton gradient is required for the transport of two lumenal oxygen-evolving proteins across the thyalkoid membrane. *J. Biol. Chem.* **266,** 12189–12193
32. Cline, K., Ettinger, W. F., and Theg, S. (1992) Protein-specific energy requirements for protein transport across or into thylakoid membranes. Two lumenal proteins are transported in the absence of ATP. *J. Biol. Chem.* **267,** 2688–2696
33. Brink, S., Bogsch, E. G., Edwards, W. R., Hynds, P. J., and Robinson, C. (1998) Targeting of thylakoid proteins by the ΔpH-driven twin-arginine translocation pathway requires a specific signal in the hydrophobic domain in conjunction with the twin-arginine motif. *FEBS Lett.* **434,** 425–430
34. Lamppa, G. K. (1988) The chlorophyll *a*/*b*-binding protein inserts into the thylakoids independent of its cognate transit peptide. *J. Biol. Chem.* **263,** 14996–14999
35. Li, X., Henry, R., Yuan, J., Cline, K., and Hoffman, N. E. (1995) A chloroplast homologue of the signal recognition particle subunit SRP54 is involved in the post-translational integration of a protein into thylakoid membranes. *Proc. Natl. Acad. Sci. U. S. A.* **92,** 3789–3793
36. Tu, C. J., Schuenemann, D., and Hoffman, N. E. (1999) Chloroplast FtsY, chloroplast signal recognition particle, and GTP are required to reconstitute the soluble phase of light-harvesting chlorophyll protein transport into the thylakoid membrane. *J. Biol. Chem.* **274,** 27219–27224
37. Kogata, N., Nishio, K., Hirohashi, T., Kikuchi, S., and Nakai, M. (1999) Involvement of a chloroplast homologue of the signal recognition particle receptor protein, FtsY, in protein targeting to thylakoids. *FEBS Lett.* **329,** 329–333
38. Michl, D., Robinson, C., Shackleton, J. B., Herrmann, R. G., and Klösgen, R. B. (1994) Targeting of proteins to the thylakoids by bipartite presequences: CF_oII is imported by a novel, third pathway. *EMBO J.* **13,** 1310–1317
39. Lorkovic, Z. J., Schröder, W. P., Pakrasi, H. B., Irrgang, K.-D., Herrmann, R. G., and Oelmüller, R. (1995) Molecular characterisation of PSII-W, the only nuclear-encoded component of the photosystem II reaction centre. *Proc. Natl. Acad. Sci. U. S. A.* **92,** 8930–8934
40. Robinson, D., Karnauchov, I., Herrmann, R. G., Klösgen, R. B., and Robinson, C. (1996) Protease-sensitive thylakoidal import machinery for the Sec-, ΔpH-, and signal recognition particle-dependent protein targeting pathways, but not for CF_oII integration. *Plant J.* **10,** 149–155
41. Kim, S. J., Robinson, C., and Mant, A. (1998) Sec/SRP-independent insertion of two thylakoid membrane proteins bearing cleavable signal peptides. *FEBS Lett.* **424,** 105–108
42. Tissier, C., Woolhead, C. A., and Robinson, C. (2002) Unique structural determinants in the signal peptides of "spontaneously" inserting thylakoid membrane proteins. *Eur. J. Biochem.* **269,** 3131–3141
43. Berthold, D. A., Babcock, G. T., and Yocum, C. F. (1981) A highly resolved, oxygen-evolving photosystem II preparation from spinach thylakoid membranes. *FEBS Lett.* **134,** 231–234
44. Kieselbach, T., Hagman, Å., Andersson, B., and Schröder, W. P. (1998) The thylakoid lumen of chloroplasts. *J. Biol. Chem.* **273,** 6710–6716
45. Kieselbach, T., Bystedt, M., Hynds, P., Robinson, C., and Schröder, W. P. (2000) A peroxidase homologue and novel plastocyanin located by proteomics to the *Arabidopsis* thyalkoid lumen. *FEBS Lett.* **480,** 271–276
46. Peltier, J.-B., Emanuelsson, O., Kalume, D. E., Ytterberg, J., Friso, G., Rudella, A., Liberles, D. A., Söderberg, L., Roepstorff, P., von Heijne, G., and van Wijk, K. J. (2002) Central functions of the lumenal and peripheral thylakoid proteome of *Arabidopsis* determined by experimentation and genome-wide prediction. *Plant Cell* **14,** 211–236
47. Schubert, M., Petersson, U. A., Haas, B. J., Funk, C., Schröder, W. P., and Kieselbach, T. (2002) Proteome map of the chloroplast lumen of *Arabidopsis thaliana*. *J. Biol. Chem.* **277,** 8354–8365
48. Hippler, M., Klein, J., Fink, A., Allinger, T., and Hoerth, P. (2001) Towards functional proteomics of membrane protein complexes: analysis of thylakoid membranes from *Chlamydomonas reinhardtii*. *Plant J.* **28,** 595–606
49. Whitelegge, J. P., Zhang, H., Aguilera, R., Taylor, R. M., and Cramer, W. A. (2002) Full subunit coverage liquid chromatography electrospray ionization mass spectrometry (LCMS+) of an oligomeric membrane protein: cytochrome b_6f complex from spinach and the cyanobacterium *Mastigocladus laminosus*. *Mol. Cell. Proteomics* **1,** 816–827
50. Hoagland, D. R., and Arnon, D. I. (1950) The water culture method for growing plants without soil. *California Agriculture Experiment Station Circular 347*, University of California, Berkeley, CA
51. Peter, G. F., and Thornber, J. P. (1991) Biochemical composition and organization of higher plant photosystem II light-harvesting pigment-proteins. *J. Biol. Chem.* **266,** 16745–16754
52. Nishio, J. N., and Whitmarsh, J. (1991) Dissipation of the proton electrochemical potential in intact and lysed chloroplasts. I. The electric potential. *Plant Physiol.* **95,** 522–528
53. Hamasur, B., and Glaser, E. (1992) Plant mitochondrial F0F1 ATP synthase. Identification of the individual subunits and properties of the purified spinach leaf mitochondrial ATP synthase. *Eur. J. Biochem.* **205,** 409–416
54. Hightower, K. E., and McCarty, R. E. (1996) Proteolytic cleavage within a regulatory region of the gamma subunit of chloroplast coupling factor 1. *Biochemistry* **35,** 4846–4851
55. Larsson, K. H., Napier, J. A., and Gray, J. C. (1992) Import and processing of the precursor form of the gamma subunit of the chloroplast ATP synthase from tobacco. *Plant Mol. Biol.* **19,** 343–349
56. Hoesche, J. A., and Berzborn, R. J. (1992) Cloning and sequencing of a cDNA for the delta-subunit of photosynthetic ATP-synthase (EC 3.6.3.14) from pea (*Pisum sativum*). *Biochim. Biophys. Acta* **1171,** 201–204
57. Berzborn, R. J., Finke, W., Otto, J., and Meyer, H. E. (1987) Protein sequence and structure of N-terminal amino acids of subunit delta of spinach photosynthetic ATP-synthase CF1. *Z. Naturforsch.* **42,** 1231–1238
58. Pfefferkorn, B., and Meyer, H. E. (1986) N-terminal amino acid sequence of the Rieske iron-sulfur protein from the cytochrome b_6/*f*-complex of spinach thylakoids. *FEBS Lett.* **206,** 233–237
59. Dunn, P. P. J., Packman, L. C., Pappin, D., and Gray, J. C. (1988) N-terminal amino acid sequence analysis of the subunits of pea photosystem I. *FEBS Lett.* **228,** 157–161
60. Obokata, J., Mikami, K., Hayashida, N., Nakamura, M., and Sugiura, M. (1993) Molecular heterogeneity of photosystem I. *psaD*, *psaE*, *psaF*, *psaH* and *psaL* are all present in isoforms in *Nicotiana* spp. *Plant Physiol.* **102,** 1259–1267
61. Iwasaki, Y., Sasaki, T., and Takabe, T. (1990) Sequencing and expression of the gene that encodes a 20-kDa polypeptide of the PS I complex from cucumber cotyledon. *Plant Cell Physiol.* **31,** 871–879
62. Scheller, H. V., Hoej, P. B., Svendsen, I., and Moeller, B. L. (1988) Partial amino acid sequence of two nuclear-encoded photosystem I polypeptides from barley. *Biochim. Biophys. Acta* **933,** 501–505
63. Obokata, J., Mikami, K., Yamamoto, Y., and Hayashida, N. (1994) Microheterogeneity of PSI-E subunit of photosystem I in *Nicotaia sylvestris*. *Plant Cell Physiol.* **35,** 203–209
64. Anandan, S., Vainstein, A., and Thornber, J. P. (1989) Correlation of some published amino acid sequences for photosystem I polypeptides to a 17 kDa LHCI pigment-protein and to subunits III and IV of the core complex. *FEBS Lett.* **256,** 150–154
65. Iwasaki, Y., Ishikawa, H., Hibino, T., and Takabe, T. (1991) Characterization of genes that encode subunits of cucumber PS I complex by N-terminal sequencing. *Biochim. Biophys. Acta* **1059,** 141–148
66. Okkels, J. S., Scheller, H. V., Jepsen, L. B., and Moeller, B. L. (1989) A cDNA clone encoding the precursor for a 10.2 kDa photosystem I polypeptide of barley. *FEBS Lett.* **250,** 575–579
67. Ikeuchi, M., Hirano, A., and Inoue, Y. (1991) Correspondence or apoproteins of light-harvesting chlorophyll *a*/*b* complexes associated with photosystem 1 to *cab* genes: Evidence for a novel type 4 apoprotein. *Plant Cell Physiol.* **32,** 103–112
68. Knoetzel, J., Svendsen, I., and Simpson, D. J. (1992) Identification of the photosystem I antenna polypeptides in barley. Isolation of three pigment-binding antenna complexes. *Eur. J. Biochem.* **206,** 209–215
69. Welty, B. A., and Thornber, J. P. (1992) in *Research in Photosynthesis* (Murata, N., ed) Vol. 1, pp. 323–326, Kluwer Academic Publishers, Dordrecht, The Netherlands
70. Anandan, S., Morishige, D. T., and Thornber, J. P. (1993) Light-induced biogenesis of light-harvesting complex I (LHC I) during chloroplast

development in barley (*Hordeum vulgare*). Studies using cDNA clones of the 21- and 20-kilodalton LHC I apoproteins. *Plant Physiol.* **101,** 227–236
71. Yamamoto, Y., Hermodson, M. A., and Krogmann, D. W. (1986) Improved purification and N-terminal sequence of the 33-kDa protein in spinach PS II. *FEBS Lett.* **195,** 155–158
72. Oh-Oka, H., Tanaka, S., Wada, K., Kuwabara, T., and Murata, N. (1986) Complete amino acid sequence of 33 kDa protein isolated from spinach photosystem II particles. *FEBS Lett.* **197,** 63–66
73. Murata, N., Kajiura, H., Fujimura, Y., Miyao, M., Murata, T., Watanabe, A., and Shinozaki, K. (1987) in *Progress in Photosynthesis Research* (Biggins, J., ed) Vol. 1, pp. 701–704, Martinus Nijhoff Publishers, Dordrecht, The Netherlands
74. Costa, P., Pionneau, C., Bauw, G., Dubos, C., Bahrman, N., Kremer, A., Frigerio, J.-M., and Plomion, C. (1999) Separation and characterization of needle and xylem maritime pine proteins. *Electrophoresis* **20,** 1098–1108
75. Takahashi, H., Ehara, Y., and Hirano, H. (1991) A protein in the oxygen-evolving complex in the chloroplast is associated with symptom expression on tobacco leaves infected with cucumber mosaic virus strain Y. *Plant Mol. Biol.* **16,** 689–698
76. Kuwabara, T., Murata, T., Miyao, M., and Murata, N. (1986) Partial degradation of the 18-kDa protein of the photosynthetic oxygen-evolving complex: a study of a binding site. *Biochim. Biophys. Acta* **850,** 146–155
77. Schröder, W. P., Henrysson, T., and Aakerlund, H.-E. (1988) Characterization of low molecular mass proteins of photosystem II by N-terminal sequencing. *FEBS Lett.* **235,** 289–292
78. Soncini, F. C., and Vallejos, R. H. (1989) The chloroplast reductase-binding protein is identical to the 16.5-kDa polypeptide described as a component of the oxygen-evolving complex. *J. Biol. Chem.* **264,** 21112–21115
79. Lautner, A., Klein, R., Ljungberg, U., Reilaender, H., Bartling, D., Andersson, B., Reinke, H., Beyreuther, K., and Herrmann, R. G. (1988) Nucleotide sequence of cDNA clones encoding the complete precursor for the "10-kDa" polypeptide of photosystem II from spinach. *J. Biol. Chem.* **263,** 10077–10081
80. Webber, A. N., Packman, L. C., and Gray, J. C. (1989) A 10 kDa polypeptide associated with the oxygen-evolving complex of photosystem II has a putative C-terminal non-cleavable thylakoid transfer domain. *FEBS Lett.* **242,** 435–438
81. Ikeuchi, M., Takio, K., and Inoue, Y. (1989) N-terminal sequencing of photosystem II low-molecular-mass proteins. 5 and 4.1 kDa components of the O2-evolving core complex from higher plants. *FEBS Lett.* **242,** 263–269
82. Zheleva, D., Sharma, J., Panico, M., Morris, H. R., and Barber, J. (1998) Isolation and characterization of monomeric and dimeric CP47-reaction center photosystem II complexes. *J. Biol. Chem.* **273,** 16122–16127
83. Morishige, D. T., and Thornber, J. P. (1991) Correlation of apoproteins with the genes of the major chlorophyll *a/b* binding protein of photosystem II in *Arabidopsis thaliana.* Confirmation for the presence of a third member of the LHC IIb gene family. *FEBS Lett.* **293,** 183–187
84. Webber, A. N., and Gray, J. C. (1989) Detection of calcium binding by photosystem II polypeptides immobilised onto nitrocellulose membrane. *FEBS Lett.* **249,** 79–82
85. Zolla, L., Timperio, A.-M., Testi, M. G., Bianchetti, M., Bassi, R., Manera, F., and Corradini, D. (1999) Isolation and characterization of chloroplast photosystem II antenna of spinach by reversed-phase liquid chromatography. *Photosynth. Res.* **61,** 281–290
86. Morishige, D. T., Anandan, S., Jaing, J. T., and Thornber, J. P. (1990) Amino-terminal sequence of the 21 kDa apoprotein of a minor light-harvesting pigment-protein complex of the Photosystem II antenna (LHC IId/CP24). *FEBS Lett.* **264,** 239–242
87. Michel, H., Griffin, P. R., Shabanowitz, J., Hunt, D. F., and Bennett, J. (1991) Tandem mass spectrometry identifies sites of three post-translational modificaions of spinach light-harvesting chlorophyll-protein II. *J. Biol. Chem.* **266,** 17584–17591
88. Huber, C. G., Timperio, A.-M., and Zolla, L. (2001) Isoforms of photosystem II antenna proteins in different plant species revealed by liquid chromatography-electrospray ionization mass spectrometry. *J. Biol. Chem.* **276,** 45755–45761
89. Ouellette, A. J. A., and Barry, B. A. (2002) Tandem mass spectrometric identification of spinach photosystem II light-harvesting components. *Photosynth. Res.* **72,** 159–173
90. Clauser, K. R., Baker, P. R., and Burlingame, A. L. (1999) Role of accurate mass measurement (+/− 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem.* **71,** 2871–2882
91. Emanuelsson, O., Nielsen, H., and von Heijne, G. (1999) ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Sci.* **8,** 978–984
92. Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* **300,** 1005–1016
93. Nakai, K., and Kanehisa, M. (1992) A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics* **14,** 897–911
94. Claros, M. G., and Vincens, P. (1996) Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur. J. Biochem.* **241,** 779–786
95. Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10,** 1–6
96. Jansson, S. (1999) A guide to the Lhc genes and their relatives in *Arabidopsis*. *Trends Plant Sci.* **4,** 236–240
97. von Heijne, G. (1985) Signal sequences. The limits of variation. *J. Mol. Biol.* **184,** 99–105
98. Kim, S. J., Robinson, D., and Robinson, C. (1999) An *Arabidopsis thaliana* cDNA encoding PSII-X, a 4.1 kDa component of photosystem II: a bipartite presequence mediates SecA/ΔpH-independent targeting into thylakoids. *FEBS Lett.* **390,** 175–178
99. Carol, P., Li, Y. F., and Mache, R. (1991) Conservation and evolution of the nucleus-encoded and chloroplast-specific ribosomal proteins in pea and spinach. *Gene* (*Amst.*) **103,** 139–145
100. Samach, A., Hareven, D., Gutfinger, T., Ken-Dror, S., and Lifschitz, E. (1991) Biosynthetic threonine deaminase gene of tomato: isolation, structure, and upregulation in floral organs. *Proc. Natl. Acad. Sci. U. S. A.* **88,** 2678–2682
101. Okkels, J. S., Scheller, H. V., Svendsen, I., and Møller, B. L. (1991) Isolation and characterization of a cDNA encoding an 18-kDa hydrophobic photosystem I subunit (PSI-L) from barley (*Hordeum vulgare* L.). *J. Biol. Chem.* **266,** 6767–6773
102. Thompson, G. A., Scherer, D. E., Foxall-Van Aken, S., Kenny, J. W., Young, H. L., Shintani, D. K., Kridl, J. C., and Knauf, V. C. (1991) Primary structures of the precursor and mature forms of stearoyl-acyl carrier protein desaturase from safflower embryos and requirement of ferredoxin for enzyme activity. *Proc. Natl. Acad. Sci. U. S. A.* **88,** 2578–2582
103. Bruce, B. D. (2001) The paradox of the plastid transit peptides: conservation of function despite divergence in primary structure. *Biochim. Biophys. Acta* **1541,** 2–21
104. Liu, Y., Engelman, M., and Gerstein, M. (2002) Genomic analysis of membrane protein families: abundance and conserved motifs. *Genome Biol.* **3,** 1–12
105. Aebersold, R., Rist, B., and Gygi, S. P. (2000) Quantitative proteome analysis: methods and applications. *Ann. N. Y. Acad. Sci.* **919,** 33–47
106. Gygi, S. P., Corthals, G. L., Zhang, Y., Rochon, Y., and Aebersold, R. (2000) Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 9390–9395
107. Whitelegge, J. P., le Coutre, J., Lee, J. C., Engel, C. K., Prive, G. G., Faull, K. F., and Kaback, H. R. (1999) Toward the bilayer proteome, electrospray ionization-mass spectrometry of large, intact transmembrane proteins. *Proc. Natl. Acad. Sci. U. S. A.* **96,** 10695–10698
108. le Coutre, J., Whitelegge, J. P., Gross, A., Turk, E., Wright, E. M., Kaback, H. R., and Faull, K. F. (2000) Proteomics on full-length membrane proteins using mass spectrometry. *Biochemistry* **39,** 4237–4242
109. Turk, E., Kim, O., le Coutre, J., Whitelegge, J. P., Eskandari, S., Lam, J. T., Kreman, M., Zampighi, G., Faull, K. F., and Wright, E. M. (2000) Molecular characterization of *Vibrio parahaemolyticus* vSGLT: a model for sodium-coupled sugar cotransporters. *J. Biol. Chem.* **275,** 25711–25716
110. Corradini, D., Huber, C. G., Timperio, A. M., and Zolla, L. (2000) Resolution and identification of the protein components of the photosystem II antenna system of higher plants by reversed-phase liquid chromatography with electrospray-mass spectrometric detection. *J. Chromatogr. A* **886,** 111–121

111. Zolla, L., and Timperio, A. M. (2000) High performance liquid chromatography-electrospray mass spectrometry for the simultaneous resolution and identification of intrinsic thylakoid membrane proteins. *Proteins* **41,** 398–406
112. Henzel, W. J., Billeci, T. M., Stults, J. T., Wong, S. C., Grimley, C., and Watanabe, C. (1993) Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc. Natl. Acad. Sci. U. S. A.* **90,** 5011–5015
113. James, P., Quadroni, M., Carafoli, E., and Gonnet, G. (1993) Protein identification by mass profile fingerprinting. *Biochem. Biophys. Res. Commun.* **195,** 58–64
114. Mann, M., Hojrup, P., and Roepstorff, P. (1993) Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol. Mass Spectrom.* **22,** 338–345
115. Yates, J. R., III, Speicher, S., Griffin, P. R., and Hunkapiller, T. (1993) Peptide mass maps: a highly informative approach to protein identification. *Anal. Biochem.* **214,** 397–408
116. Yates, J. R., III, Eng, J. K., and McCormack, A. L. (1995) Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Anal. Chem.* **67,** 3202–3210
117. Dongre, A. R., Eng, J. K., and Yates, J. R., III (1997) Emerging tandem-mass-spectrometry techniques for the rapid identification of proteins. *Trends Biotechnol.* **15,** 418–425
118. Zhang, H., Whitelegge, J. P., and Cramer, W. A. (2001) Ferredoxin:$NADP^+$ oxidoreductase is a subunit of the chloroplast cytochrome b_6f complex. *J. Biol. Chem.* **276,** 38159–38165
119. Whitelegge, J. P., Gómez, S. M., Aguilera, R., Roberson, R. W., Vermaas, V. F., Crother, T. R., Champion, C. I., Nally, J. E., Blanco, D. R., Lovett, M. A., Miller, J. N., and Faull, K. F. (2002) Identification of proteins and intact mass measurements in proteomics. *Appl. Genomics Proteomics* **1,** 13–22