THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 277, No. 10, Issue of March 8, pp. 8354–8365, 2002

Printed in U.S.A.

Proteome Map of the Chloroplast Lumen of *Arabidopsis thaliana**

Received for publication, September 6, 2001, and in revised form, November 20, 2001
Published, JBC Papers in Press, November 21, 2001, DOI 10.1074/jbc.M108575200

Maria Schubert‡§, Ulrika A. Petersson§¶, Brian J. Haas‖, Christiane Funk¶, Wolfgang P. Schröder§‡‡, and Thomas Kieselbach‡§§**

From the Departments of ‡Medical Nutrition and §§Biosciences, Karolinska Institute, Novum, Huddinge SE-14186, Sweden, the §Department of Natural Science, Södertörns University College, Bipontus, Box 4101, Huddinge SE-14104, Sweden, the ¶Department of Biochemistry and Biophysics, Arrhenius Laboratories for Natural Sciences, Stockholm University, Stockholm SE-10691, Sweden, and ‖The Institute for Genomic Research, Rockville, Maryland 20850

The thylakoid membrane of the chloroplast is the center of oxygenic photosynthesis. To better understand the function of the luminal compartment within the thylakoid network, we have carried out a systematic characterization of the luminal thylakoid proteins from the model organism *Arabidopsis thaliana*. Our data show that the thylakoid lumen has its own specific proteome, of which 36 proteins were identified. Besides a large group of peptidyl-prolyl cis-trans isomerases and proteases, a family of novel PsbP domain proteins was found. An analysis of the luminal signal peptides showed that 19 of 36 luminal precursors were marked by a twin-arginine motif for import via the Tat pathway. To compare the model organism *Arabidopsis* with another typical higher plant, we investigated the proteome from the thylakoid lumen of spinach and found that the luminal proteins from both plants corresponded well. As a complement to our experimental investigation, we made a theoretical prediction of the luminal proteins from the whole *Arabidopsis* genome and estimated that the thylakoid lumen of the chloroplast contains ~80 proteins.

The ability to perform oxygenic photosynthesis belongs to the distinguishing characteristics of higher plants, algae, and cyanobacteria. In higher plants, the center of the photosynthetic process is the thylakoid membrane of the chloroplast. Here, in a synergistic series of reactions, four protein complexes, the photosystems I and II, the cytochrome b_6f complex, and the ATP-synthase, produce NADPH and ATP that fuel the further synthesis of carbohydrates (1, 2).

A key feature in the energy conversion of photosynthesis is the link between the electron transfer from photosystem II to I via the cytochrome b_6f complex and the generation of a proton gradient over the thylakoid membrane. To balance the flow of electrical charges during the formation of the proton gradient, there is a busy traffic of chloride and calcium ions from the stroma into the lumen and of magnesium ions from the lumen into the stroma (3–6). This ion traffic plays a fundamental role for the proper function of photosynthesis. For a long time it was believed that accumulating protons and balancing the ion currents over the thylakoid membrane was the main function of the luminal compartment. The ensemble of known luminal proteins was small and consisted of the three extrinsic photosystem II proteins (PsbO, PsbP, and PsbQ) and plastocyanin. This group was later joined by some new proteins such as violaxanthin de-epoxidase (7), polyphenol oxidase (8, 9), the extrinsic photosystem I protein PsaN (10), and the carboxyl-terminal processing protease for the D1 protein (11).

To achieve a more profound understanding of content and functions of the thylakoid lumen, we designed a method that enabled us to isolate a highly pure fraction of luminal proteins from spinach thylakoids. For the first time, we showed that the lumen of the thylakoid membrane contained at least 20 proteins and that the protein concentration of this compartment was similar to that of the stroma (12). Several new luminal proteins could be characterized in more detail. The 17.4-kDa protein (TL17) had a remarkable new pentapeptide motif and led to the discovery of a whole family of unknown pentapeptide proteins in *Synechocystis sp*. PCC 6803 (13). In addition, a novel 16-kDa protein (TL16) was found to be routed into the thylakoid lumen by the Tat translocation pathway (14). The luminal immunophilin TL40 was suggested to participate in signal transduction over the thylakoid membrane (15), and *Hcf136* was identified as a luminal assembly factor for photosystem II (16). Recently, the 29-kDa peroxidase homologue TL29 and a novel plastocyanin were added to the list of new luminal proteins (17) along with several proteins from the thylakoids of pea that were not further characterized (18).

The completion of the *Arabidopsis thaliana* genome sequencing project by the end of 2000 (19) started a new era in plant research. To apply the knowledge of the *Arabidopsis* genome to an investigation into luminal proteins, we designed a method to isolate the luminal proteins from *Arabidopsis* chloroplasts and studied them by proteomics. In this study, we performed the first systematic characterization of the chloroplast lumen of *Arabidopsis* and compared its proteins with those from the chloroplast lumen of spinach. We found 36 luminal proteins in *Arabidopsis*, of which 22 could be identified in spinach also. By comparing the experimentally identified lumen proteins of *Arabidopsis* with a theoretical prediction of a luminal proteome in this organism, we estimated that the chloroplast lumen of *Arabidopsis* comprises ~80 proteins.

EXPERIMENTAL PROCEDURES

Growth Conditions and Lumen Preparation—Plants were cultivated hydroponically according to Norén *et al.* (20). *A. thaliana* ecotype Columbia was grown for 13 weeks with 8 h light per day and a light intensity of 100 μmol of photons m^{-2} s^{-1}, and spinach (*Spinacia oleracea)* was raised for 5 weeks with 10 h of light per day. The lumen fraction from spinach chloroplasts was isolated according to Kieselbach *et al.* (12).

* This work was supported by grants of the Swedish Research Council and Södertörns Högskola and by the Protein Analysis Unit at the Center of Structural Biochemistry at the Karolinska Institute. 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.

** Supported by Cooperative Agreement DBI-9813586 from the National Science Foundation.

‡‡ To whom correspondence should be addressed. Tel.: 46-8-58588587; Fax: 46-8-58588510; E-mail: wolfgang.schroder@sh.se.

The luminal content from *Arabidopsis* chloroplasts was isolated in the same way with the following minor changes. To avoid protein degradation, 1 mM of EDTA was added to the last washing buffer, and 50 μg/ml of Pefablock was added to the thylakoids just prior to the Yedapress treatment. The lumen fraction was concentrated using Centriprep YM-3 concentrators (Millipore Corp.), and protein quantification was carried out according to Bradford (21) using bovine serum albumin as a standard.

Two-dimensional Electrophoresis—The luminal proteins were separated by isoelectric focusing in the first and by SDS-polyacrylamide gel electrophoresis (SDS-PAGE) in the second dimension. The samples for analytical two-dimensional gels contained 100 μg of protein, and those for Western blots for microsequencing contained 200 μg of protein. The luminal proteins were solubilized in 5 M urea, 2 M thiourea, 4% (w/v) CHAPS,[1] 50 mM dithiothreitol, and 0.8% (v/v) carrier ampholytes (IPG buffer, pH 3–10 nonlinear or 4–7 linear Amersham Biosciences, Inc.) and applied during rehydration to a nonlinear IPG strip, pH 3–10, or a to linear IPG strip, pH 4–7. The strips were allowed to rehydrate overnight at 20 °C and then transferred to IPGphor cup-loading strip holders and covered with paraffin oil. The proteins were focused for 10 min at 300 V followed by a 3-h gradient from 300 to 3500 V and a 30-min gradient from 3500 to 8000 V. The isoelectric focusing was then completed at a constant voltage of 8000 V until 60,000 V-h was reached. Subsequently, the strips were equilibrated first for 15 min in 50 mM Tris-HCl (pH 6.8), 6 M urea, 30% (v/v) glycerol, 2% (w/v) SDS, and 1% (w/v) dithiothreitol and then for 10 min in the same buffer without dithiothreitol but with 2.5% (w/v) iodoacetamide and a trace of bromphenol blue. In the second dimension, SDS-PAGE according to Laemmli (22) was carried out in a gradient slab gel (T = 9–16%) using a Protean II XL system from Bio-Rad. Before the polymerization of the gel, 5 mM of sodium thiosulfate was added to the monomer solution to decrease the background staining with silver. Proteins were detected by silver staining according to Bjellqvist *et al.* (23). The two-dimensional gels were scanned using an image scanner and evaluated with the Image Master two-dimensional software (both from Amersham Biosciences). The apparent masses of the proteins that were detected on the two-dimensional gels were determined manually or with the Image Master two-dimensional software using identified proteins of known masses as a reference.

MALDI-TOF Mass Spectrometry and Microsequencing—MALDI-TOF analysis of in-gel digested proteins was carried out with a Reflex III mass spectrometer from Bruker. The in-gel digests were performed using sequencing grade modified trypsin (Promega) and analyzed as described (24, 25). Data base searches were done with the MS BioTools software from Bruker using the Mascot search engine (available on the World Wide Web at www.matrixscience.com). If a protein could not unambiguously be identified by a fingerprint spectrum, its identity was confirmed by a postsource decay analysis of single peptides (26). Amino-terminal microsequencing was carried out with a Procise sequencer from Applied Biosystems. Proteins were sequenced from polyvinylidene difluoride membrane following resolution by two-dimensional electrophoresis essentially as described (27).

Bioinformatics—The individual analysis of protein sequences by similarity searches (28, 29), pattern and profile searches (30, 31), alignments (32), and hydrophobicity plots (33) was carried out using the ExPaSy tools (available on the World Wide Web at www.expasy.ch). The prediction of chloroplast-targeted proteins encoded within the *A. thaliana* genome was performed by subjecting the latest version of the proteome, currently consisting of 25,657 proteins, to an analysis using the program TargetP via the World Wide Web interface (34). The NH_2-terminal portion of each protein sequence, not exceeding 140 residues in length, was analyzed by the plant version of TargetP, and all chloroplast-predicted proteins (rank 1–5) were used. The prediction of signal peptides and the peptide cleavage products was performed using the portable version of SignalP-2.0 (35, 36). The programs TargetP and SignalP are available on the World Wide Web at www.cbs.dtu.dk/services, and the *Arabidopsis* proteome is accessible on the Internet at ftp://ftp.tigr.org/pub/data/a_thaliana/ath1/SEQUENCES/ATH1.pep.

RESULTS

Isolation of the Thylakoid Lumen, Reproducibility of the Two-dimensional Electrophoresis, and Protein Identification—While *Arabidopsis* is an excellent model organism for molecular biological studies, biochemical work has been difficult due to the small leaf size and the small amount of material. The introduction of a hydroponical culture made it possible to overcome this drawback and to grow plants with considerably larger leaves that provided sufficient amounts of plant material in a high quality well suited for biochemical preparations. An essential element in this technique was the restriction of light to 8 h/day, which enabled us to grow the plants for 13 weeks without flowering. Using *Arabidopsis* plants cultivated in this way, we were able to purify the luminal fraction in the same high quality as was obtained in the original method for the isolation of thylakoid lumen from spinach chloroplasts (12). A typical lumen preparation from *Arabidopsis* started with 100–200 g of wet weight leaf material and yielded in ~2 mg protein/100 g of leaves.

An important prerequisite for a concise mapping of the luminal *Arabidopsis* proteins was a reproducible two-dimensional electrophoresis system that was capable of resolving the major part of the luminal proteins. To meet these requirements, we used a combination of nonlinear pH gradients (pH 3–10) and a polyacrylamide gradient gel that allowed us to detect proteins with isoelectric points between 4 and 9 and masses between 200 and 9 kDa. Samples from 14 different lumen preparations were analyzed in more than 20 experiments. The two-dimensional maps of the luminal *Arabidopsis* proteins showed an excellent reproducibility, and a representative experiment is shown in Fig. 1*A*. From the complete set of experiments, 13 two-dimensional gels were selected for a detailed image analysis. The total number of protein spots that were detected on the two-dimensional gels was between 400 and 700, and 277 of those were present on all 13 gels.

The protein pattern of the two-dimensional gel in Fig. 1*A* shows clearly that most proteins were detected in the acidic region of the pH gradient, while the group of basic proteins was relatively small. It should be noticed, too, that there is a distinct gap between the proteins in the acidic and the basic region of the pH gradient, where only few proteins could be detected. Since 80% of the luminal proteins were found in the acidic range of the pH gradient, we analyzed these proteins in more detail using a two-dimensional electrophoresis system with a linear pH gradient from 4 to 7. A typical two-dimensional map of the luminal proteins with isoelectric points in this pH range is shown in Fig. 1*B*. An evaluation of the images of 10 independent experiments showed that ~200 protein spots were present on each two-dimensional gel of the pH range from 4 to 7.

Having established a reproducible two-dimensional map of

[1] The abbreviations used are: CHAPS, 3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonic acid; MALDI-TOF, matrix-assisted laser desorption/ionization time-of-flight; PPIase, peptidyl-prolyl cis-trans isomerase.

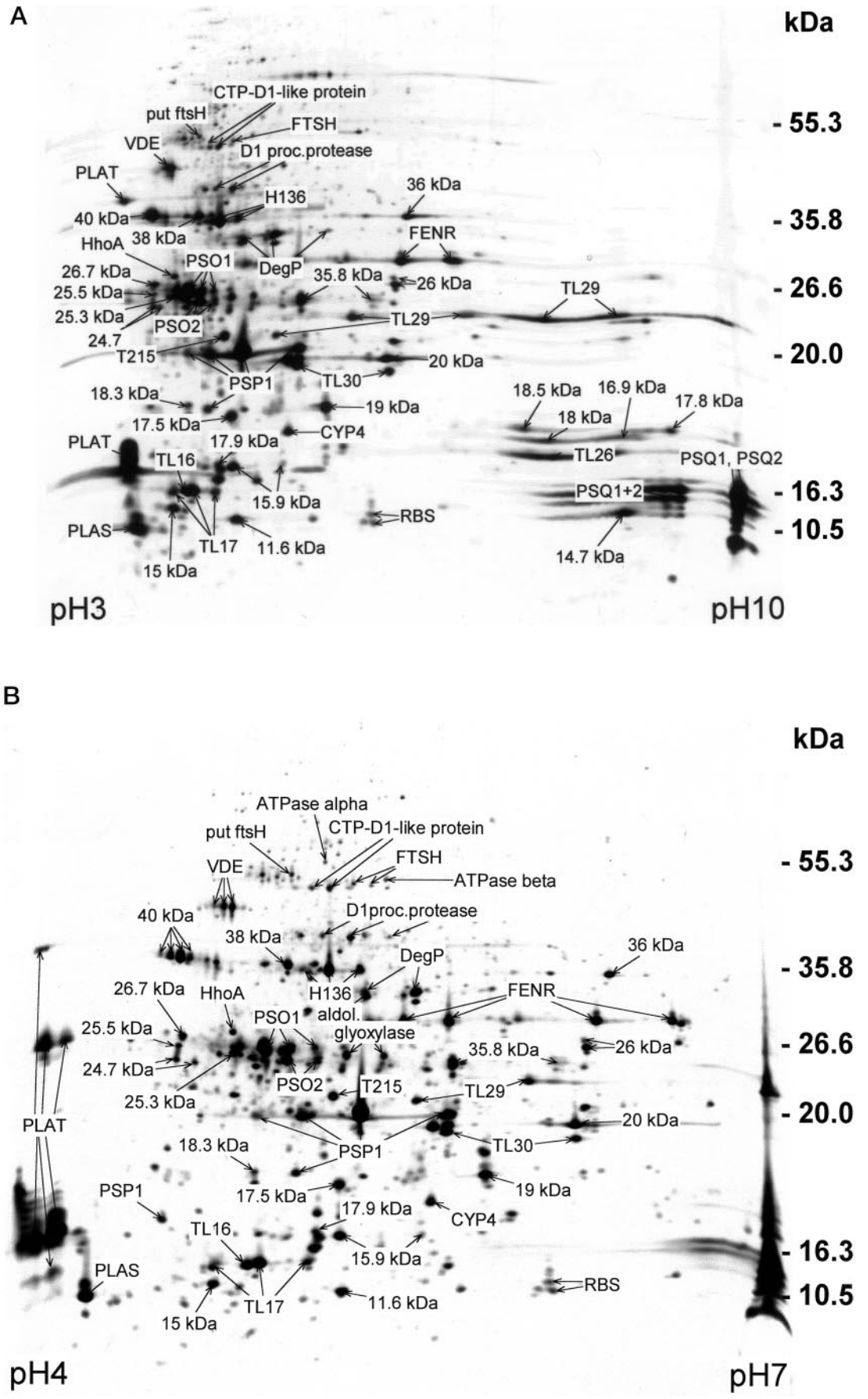

FIG. 1. **Silver-stained two-dimensional gels of 100 μg of soluble luminal proteins from the chloroplast of *Arabidopsis*.** The proteins were resolved by SDS electrophoresis in a 9–16% polyacrylamide gradient gel subsequent to isoelectric focusing in a nonlinear immobilized pH gradient from pH 3 to 10 (*A*) and linear immobilized pH gradient from pH 4 to 7 (*B*).

the chloroplast lumen from *Arabidopsis*, we systematically analyzed the luminal proteins by both MALDI-TOF mass spectrometry and amino-terminal microsequencing. Using this combination of methods, we were able to determine the amino termini of the mature proteins and to correct errors in the gene models of several proteins. This approach could be successfully applied to 90 protein spots, but others of the 277 spots that were detected in all two-dimensional gels were too weak for an analysis by both mass spectrometry and microsequencing. In total, 49 proteins were detected, and each protein was identified in at least two independent experiments except for the three putative fibrillins with apparent masses at 25.5, 25.3, and 24.7 kDa. Although these proteins were analyzed only once, the identification was specific in each single case, and the corresponding protein spots were detected in all two-dimensional gels analyzed. Hence, there is no doubt that these proteins were correctly identified.

The two-dimensional gel in Fig. 1*A* shows that 40 of the 49 identified proteins were detected in the acidic range of the pH gradient, while nine proteins were found in the basic region. The analysis of the proteins by mass spectrometry and microsequencing showed that the major part of the proteins from the luminal fraction was intact. There were only few degradation products found. As Fig. 1*B* shows, there are two fragments of PsbP1 that were detected in all two-dimensional gels analyzed. In addition, we also detected a degradation product for PsaN, but this fragment occurred less frequently and did not appear in a distinct pattern as those from PsbP1.

The Protein Content of the Chloroplast Lumen of Arabidopsis—One of the principal objectives of this study was to identify a representative group of proteins from the chloroplast lumen of *Arabidopsis* and to find indications for their possible functions. Hence, it was important to confirm that the proteins that were identified in the luminal fraction from *Arabidopsis* chloroplasts were truly lumen-located. The purification method alone was no proof for a luminal location. As Fig. 1*B* shows, we detected among the luminal proteins a putative glyoxylase and the Cyp4 cyclophilin that are putative stroma proteins but were present in similar amounts as known luminal proteins such as TL29 and Hcf136.

To resolve whether the proteins that were identified in the luminal fraction were resident in the chloroplast lumen or not, we performed a concise analysis of their transit peptides. All proteins that are targeted to the chloroplast lumen are synthesized in the cytosol as precursors and cross the chloroplast envelope and the thylakoid membrane in a two-stage import via a stromal intermediate. Accordingly, the signal peptides of luminal precursors comprise two parts, of which one is designed for the transit through the envelope and the second one for the import into the thylakoid lumen. Once the precursor of a luminal protein has reached the chloroplast stroma, the envelope transit region is cleaved off, and the intermediate is routed into the thylakoid lumen by either the Sec pathway or the ΔpH-dependent twin arginine translocation (Tat) pathway. While the Sec machinery is used by proteins that tolerate partial unfolding during the import, proteins that need to be imported in a folded state cross the thylakoid membrane via the Tat complex (37).

The bipartite transit peptides for the import into the thylakoid lumen are a distinguishing feature of the luminal proteins. They are marked by a hydrophilic serine- and threonine-rich region for the transit through the envelope and a thylakoid targeting region with a typical hydrophobic core close to the processing site. In addition, the transit peptides of proteins that are routed by the Tat complex reveal a distinctive twin arginine motif in the beginning of the hydrophobic core region

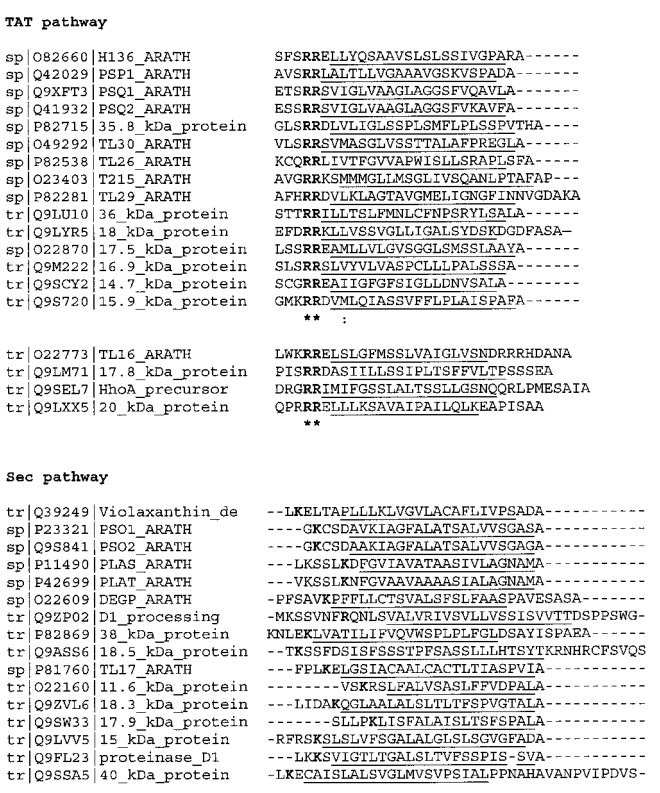

FIG. 2. **Transit peptides of luminal proteins from *A. thaliana*.** The alignment shows the hydrophobic core region (*underlined*) and the processing site of the bipartite transit peptides of the luminal proteins from Table I. Transit peptides of proteins that are routed by the Tat complex possess a distinctive twin arginine motif. By contrast, transit peptides of proteins that are targeted by the Sec pathway have a lysine residue close to the hydrophobic core region.

and a highly hydrophobic residue two or three positions after the twin arginine motif. By contrast, the transit peptides of proteins that are translocated by the Sec machinery do not have a twin arginine motif but a single lysine residue next to the amino-terminal end of the hydrophobic core region (37, 38).

The transit peptides of all proteins that were identified in the luminal fraction of *Arabidopsis* chloroplasts were comprehensively analyzed whether or not they revealed the features of bipartite signal peptides. Of the 49 proteins that were identified in the luminal fraction, 35 had, indeed, a bipartite transit peptide. Fig. 2 shows the thylakoid targeting regions of these transit peptides aligned with the program ClustalW. The alignment shows that all transit peptides possess a hydrophobic core; 19 signal peptides have a twin-arginine motif that marks them for translocation by the Tat pathway, and 16 have a lysine residue close to the hydrophobic core, which is a characteristic of signal peptides routed by the Sec machinery. Only the D1-processing protease appears to be an exception from this rule and has an arginine instead of lysine residue next to the amino-terminal end of the hydrophobic core.

The only protein for which it was hard to decide whether the transit peptide really contained a targeting signal for the thylakoid lumen was the 20-kDa protein (Q9LXX5). The prepeptide of this protein had a rather indistinct hydrophobic core region and an arginine triplet instead of a conventional twin-arginine motif. To resolve this apparent conflict, we searched the expressed sequence tag databases for homologues and examined their transit peptides. The soybean clone Gm-c1032-2020 and the tomato clone cLET42E20 encoded for full-length homologues of the 20-kDa protein, and both had typical bipartite transit peptides with a plain twin arginine motif (not

TABLE I

Proteins from the chloroplast lumen of A. thaliana

Tigr, Tigr *Arabidopsis* db; SP, Swiss-Prot/TrEMBL; MALDI, MALDI-TOF-MS; PSD, postsource decay analysis; Micro, microsequencing; ND, not determined.

Protein name	Gene locus (Tigr), accession (SP)	Identification	Mass exper./theor.[a]	Experimental NH_2-terminal sequence
			kDa	
Xanthophyll cycle				
Violaxanthin de-epoxidase	AT1g08550, Q39249	MALDI, Micro	44.2/39.8	VDALKTCACLLKGCRIELAK CIANPACA*X*N
Photosystem II assembly				
H136_ARATH	AT5g23120, O82660	MALDI, Micro	ND/35.8	DEQLSEWERVFLPID
Photosystem II subunits				
PSO1_ARATH	AT5g66570, P23321	MALDI, Micro	ND/26.6	EGAPKRLT
PSO2_ARATH	AT3g50820, Q9S841	MALDI, Micro	24.7/26.6	EGAPKRLT
PSP1_ARATH	AT1g06680, Q42029	MALDI, Micro	ND/20.2	AYGEAANVFGKPKTN
PSP2_ARATH	AT2G30790, O49344	Not detected		
PSQ1_ARATH	AT4g21280, Q9XFT3	MALDI, Micro	ND/16.3	DAISIKVGPPPAPS
PSQ2_ARATH	AT4g05180, Q41932	MALDI, Micro	15.3/16.3	EAIPIKVGGPPLPS
Proteins with a PsbP domain				
35.8 kDa protein	AT5g11450, P82715	MALDI, Micro	24.3/25.6	EEDVKMSGEELKMGTMVDDI
TL30_ARATH	AT1g77090, O49292	MALDI, Micro	19.4/22.2	VVKQGLLAGRVPGLSEPDE
T215_ARATH	AT4g15510, O23403	MALDI, Micro	21.5/21.3	STPVFREYIDTFDGYSFKYP
TL26_ARATH	AT3g55330, P82538	MALDI, Micro	16.3/17.8	AESKKGFLAVSDNKDAYAFLYPFGWQEV-VIEGQDKV
20-kDa protein	AT3g56650, Q9LXX5	MALDI, Micro	20.0/21.5	REVEVGSYLPLSPSDP*X*FVL
15.9-kDa protein	AT1g76450, Q9S720	MALDI, Micro	15.9/15.8	ETNASEAFRVYTDETNKFEISIPQ
Plastocyanins				
PLAS_ARATH	AT1g76100, P11490	Micro	ND/10.5	MEVLLGSD
PLAT_ARATH	AT1g20340, P42699	Micro	16.2/10.5	IEVLLGGGDGSLAFIPNDFSIAKGEKIVF
Putative peroxidase				
TL29_ARATH	AT4g09010, P82281	MALDI, Micro	23.2/29.3	ADLNQRRQRSEFQSKIKILLSTTIKAKPEL
Proteases				
Tail-specific proteases				
Carboxyl-terminal proteinase D1-like protein	AT5g46390, Q9FL23	MALDI, Micro	49.0/45.8	ATNDPYLS
D1-processing protease	AT4g17740, O23614	MALDI, Micro	39.6/41.9	LTEENLLF*X*EA
Serine proteases, trypsin family				
36-kDa protein	AT5g39830, Q9LU10	MALDI, PSD, Micro	31.3/37.5	LGDPSVATVEDVSPTVFPAGPLF
DegP protease (DEG1_ARATH)	AT3g27925,[b] O22609	MALDI, Micro	31.7/35.2	FVVSTPKKLQTDELA
HhoA protease	AT4g18370,[b] Q9SEL7	MALDI, Micro	27.4/23.5	LEQFKEXEE*X*L
Putative immunophilins				
Cyclophilin-type PPIases				
40-kDa protein	AT3g01480, Q9SSA5	MALDI, PSD, Micro	36.1/38.2	VLISGPPIKDPEALLRYALPID
38-kDa protein	AT3g15520,[b] P82869	MALDI, PSD, Micro	38.0/37.3	VLYSPDTKVPRTGELALRRAIPAN
18.5-kDa protein	AT5G13120,[b] Q9ASS6	Micro	18.5/20.0	NAEVVTEPQSKI
FKBP-type PPIases				
18-kDa protein	AT5g13410, Q9LYR5	MALDI, Micro	18.0/18.7	SQFADMPALKGKDYGKTKMKYPDY
17.8-kDa protein	AT1g20810, Q9LM71	MALDI, Micro	17.8/17.9	RERRSRKVIP
17.5-kDa protein	AT2g43560,[b] O22870[b]	MALDI, Micro	17.5/15.7	AGLPPEDKPRLCEAXCXKXL
16.9-kDa protein	AT3g60370, Q9M222	MALDI, Micro	16.9/16.4	KTKSKSPYDERRLLEQN
14.7-kDa protein	AT5g45680, Q9SCY2	MALDI, Micro	14.7/13.6	ETTSCEFSVSPSGLAFCDKV
Pentapeptide proteins				
TL17_ARATH	AT5g53490, P81760	MALDI, Micro	15.4/17.4	ANQRLPPLSTEPDR
11.6-kDa protein	AT2g44920, O22160	MALDI, Micro	14.7/11.5	FKGGGPYGQGVTRGQDLSGK
Proteins with unknown function				
19-kDa protein	Not available, P82658	Micro	19.0/ND	EGNQTYKIYYGTAASAANYGG
18.3-kDa protein	AT1g54780, Q9ZVL6	MALDI, Micro	18.3/22.1	SEFNILNDGP
17.9-kDa protein	AT4g24930, Q9SW33	MALDI, Micro	17.9/18.0	IPSLSSSQPLTTPFTQSKFVQTGLLNGKIR
TL16_ARATH	AT4g02530, O22773	MALDI, Micro	15.3/17.6	AILEADDDEELLEKV
15.0-kDa protein	AT5g52970, Q9LVV5	MALDI, Micro	15.0/11.5	KVGVNKPELLPKEFTSVIDV
Proteins with undecided location				
Putative fibrillins				
26.7-kDa protein	AT4g04020, O81439	MALDI, Micro	26.7/28.8	ATDIDDE
25.5-kDa protein	AT4g22240, O49629	MALDI	25.5/ND	ND
25.3-kDa protein	AT3g23400, Q9LW57	MALDI	25.3/ND	ND
24.7-kDa protein	AT3g58010, Q9M2P7	MALDI	24.7/ND	ND
Putative immunophilins				
26-kDa protein	AT5g35100, O65220	MALDI, Micro	25.5/27.9	TVTPPPPAKPPSPDITITDR
Stromal and other proteins				
ATPase α-chain	AtpA, P56757	MALDI	ND/55.3	ND
ATPase β-chain	AtpB, P19366	MALDI	ND/53.9	ND
Putative ftsh chloroplast protease	AT2g30950, O80860	MALDI, Micro	51.8/ND	FGQS*X*A*X*F (residues 209–216)
FTSH_ARATH	AT1g50250, Q9SX43	MALDI	49.9/ND	ND
Ferredoxin-NADP+ reductase	AT5g66190, Q9FKW6	MALDI	27.9/ND	ND
Putative glyoxylase	AT1g67280, Q9FYG5	MALDI, Micro	25.0/33.2	GVAESGKAAQ
Fructose bisphosphate aldolase	AT2g21330, Q93WF5	MALDI, Micro	31.3/38.1	ASAYADELV*X*TA
CYP4_ARATH	AT3g62030, P34791	MALDI, Micro	16.8/19.8	AAEEEEVIEPQA*X*VT
RBS small subunit	AT1g67090, Q9SAV4	MALDI	14.5/ND	ND

[a] Experimental/theoretical mass (kDa) is shown.

[b] Corrected in this work.

shown). This ensured that the prepeptide of the 20-kDa protein contained only an unusual variant of the twin arginine motif and, indeed, was a true bipartite transit peptide.

In the case of the 19-kDa protein (P82658), no *Arabidopsis* gene was available that could be used for an examination of the signal peptide. It could only be mapped to the tentative consensus sequence TC115875 in the *Arabidopsis* Gene Index, which cannot be used to assess its subcellular location. Hence, we searched the expressed sequence tag databases and found several cDNAs from other plants that encoded the complete precursor of a homologue of the 19-kDa protein. Two representative examples were the tomato cDNA cTOF22B19 and the soybean cDNA Gm-c1013-3374. An analysis of the signal peptides of these homologues demonstrated clearly that they had all features of a bipartite transit peptide (not shown), which indicates that the 19-kDa protein from *Arabidopsis* is a lumen-located protein as well. The other 14 proteins that were identified in the luminal fraction from *Arabidopsis* came either from the chloroplast stroma such as the small subunit of ribulose-bisphosphate carboxylase or the thylakoid membrane such as the Ftsh protease or the α- and β-subunits of the ATP synthase. As for the four putative fibrillins and the 26-kDa protein, the subcellular location could not be predicted reliably, and, hence, it was left undecided.

As Table I shows, the entire range of the luminal proteins of *Arabidopsis* that were identified in this work covers both well established classical proteins such as violaxanthin de-epoxidase and the extrinsic subunits of photosystem II, as well as novel proteins such as the large group of cyclophilins and FKBP-type peptidyl-prolyl cis-trans isomerases. In addition, there are many new proteins for which yet no function is known, such as, for instance, the PsbP domain proteins.

The silver-stained two-dimensional gels in Figs. 1, *A* and *B*, show clearly that the major proteins in the luminal fraction from *Arabidopsis* were the isoforms of the extrinsic subunits of photosystem II and of plastocyanin. As shown in Table I, the *Arabidopsis* genome encodes for two similar forms of each of these proteins, and we identified all of them except for the protein PSP2, which, despite strong efforts, could not be found. Fig. 1*B* shows plainly that the protein pattern reveals a marked isoelectric heterogeneity for these proteins. The extrinsic protein PSP1, for instance, was found in as many as four spots excluding those two that contained degraded forms of PSP1. However, we have not yet investigated whether this heterogeneity represents post-translational modifications of the single protein or just an artificial effect of the two-dimensional electrophoresis system. Remarkably, we also found a group of six proteins with apparent masses between 15.9 and 35.8 kDa that have a PsbP domain that is related to the extrinsic PsbP protein of photosystem II.

Two further luminal proteins with functions for photosystem II were found. The xanthophyll cycle enzyme violaxanthin de-epoxidase participates in the protection of photosystem II from excess light (39), and Hcf136 is essential for photosystem II assembly (16). The positions of these proteins are marked in Fig. 1, *A* and *B*, at the apparent molecular masses of 44.2 and 34.8 kDa. As with the extrinsic subunits of photosystem II, both Hcf136 and violaxanthin de-epoxidase revealed isoelectric heterogeneity and were detected in three different spots (Fig. 1*B*). A distinct isoelectric heterogeneity was also observed for the putative ascorbate peroxidase TL29, the pentapeptide protein TL17, and for the 40-kDa protein Q9SSA5 that is a homologue of the TL40 immunophilin of spinach.

Proteases are important regulatory proteins and their presence in the thylakoid lumen was postulated for a long time. DegP and the D1-processing protease were the first proteases that were found to be lumen-located (11, 40). In this work, three novel luminal proteases were identified that include a carboxyl-terminal proteinase D1-like protein with an apparent mass of 49 kDa, a DegP-like 36-kDa protein, and the 27-kDa HhoA protease. As Fig. 1*B* shows, isoelectric heterogeneity was observed for this group of proteins, too, except for the HhoA protease.

Immunophilins play an important role in protein folding and comprise the largest group of the novel luminal proteins shown in Table I. They consist of two groups. There are three putative cyclophilin-type peptidyl-prolyl cis-trans isomerases (PPIases) that were identified at the apparent molecular masses of 40, 38, and 18.5 kDa and five putative FKBP-type PPIases with apparent molecular masses between 14.7 and 18.5 kDa. It is noteworthy that five of these putative isomerases were basic proteins. Fig. 1 reveals that only the 38-kDa and the 40-kDa proteins that belong to the putative cyclophilin-type PPIases and the 17.5-kDa FKBP-type PPIase isolog were detected in the acidic range of the two-dimensional gels.

As for the other seven luminal proteins that were identified, no functions are known yet. Two pentapeptide repeat proteins were identified, of which TL17 was the first higher plant protein that possessed this motif (13). The 11.6-kDa protein that was found in this study is a novel member of this group. In addition, we identified five further proteins with apparent molecular masses between 15 and 19 kDa that do not reveal any known motifs, patterns, or domains.

The Protein Content of the Chloroplast Lumen of Spinach—Since *Arabidopsis* is a prominent model organism, we wanted to know if its luminal protein ensemble was representative for other higher plants. The organism of choice for this study was spinach, because the preparation of the luminal content of the spinach chloroplast was thoroughly characterized and known to produce a highly pure fraction of luminal proteins (12). To obtain a typical two-dimensional map of the luminal spinach proteome, a set of nine independent two-dimensional experiments was carried out, and one representative two-dimensional gel is shown in Fig. 3. The protein pattern revealed more than 500 spots, of which ~250 were detected in all nine two-dimensional experiments. A comparison of Figs. 3 and 1*A* shows that the two-dimensional patterns of the luminal proteins from spinach and *Arabidopsis* chloroplasts are very similar. As with *Arabidopsis*, the major part of luminal proteins from spinach was detected in the acidic range of the pH gradient, and there is a similar gap between the acidic and the basic proteins as it was found for the luminal proteins from *Arabidopsis*. Although the two-dimensional pattern of the lumen proteins from spinach closely resembles the one from *Arabidopsis*, it is obvious that it reveals considerably less isoelectric heterogeneity.

Since no complete genome is available for spinach, protein identification was mainly carried out by microsequencing, which restricted the analysis of the luminal proteome from spinach to the more abundant proteins. Nevertheless, 25 luminal proteins from the spinach chloroplast were identified, and Table II shows that they correspond well to the luminal proteins from *Arabidopsis*. All groups of proteins that were found in the thylakoid lumen of *Arabidopsis* were also identified in the luminal compartment of the spinach chloroplast. Some differences were observed, however. Among the luminal proteins from spinach, we could not identify any isoforms of plastocyanin and the extrinsic photosystem II subunits PsbO and PsbQ. In addition, we did not find any *Arabidopsis* homologues to the 25.3-kDa protein and polyphenol oxidase. While we could not resolve experimentally if the 25.3-kDa protein was really absent in *Arabidopsis*, we could analyze the luminal content of

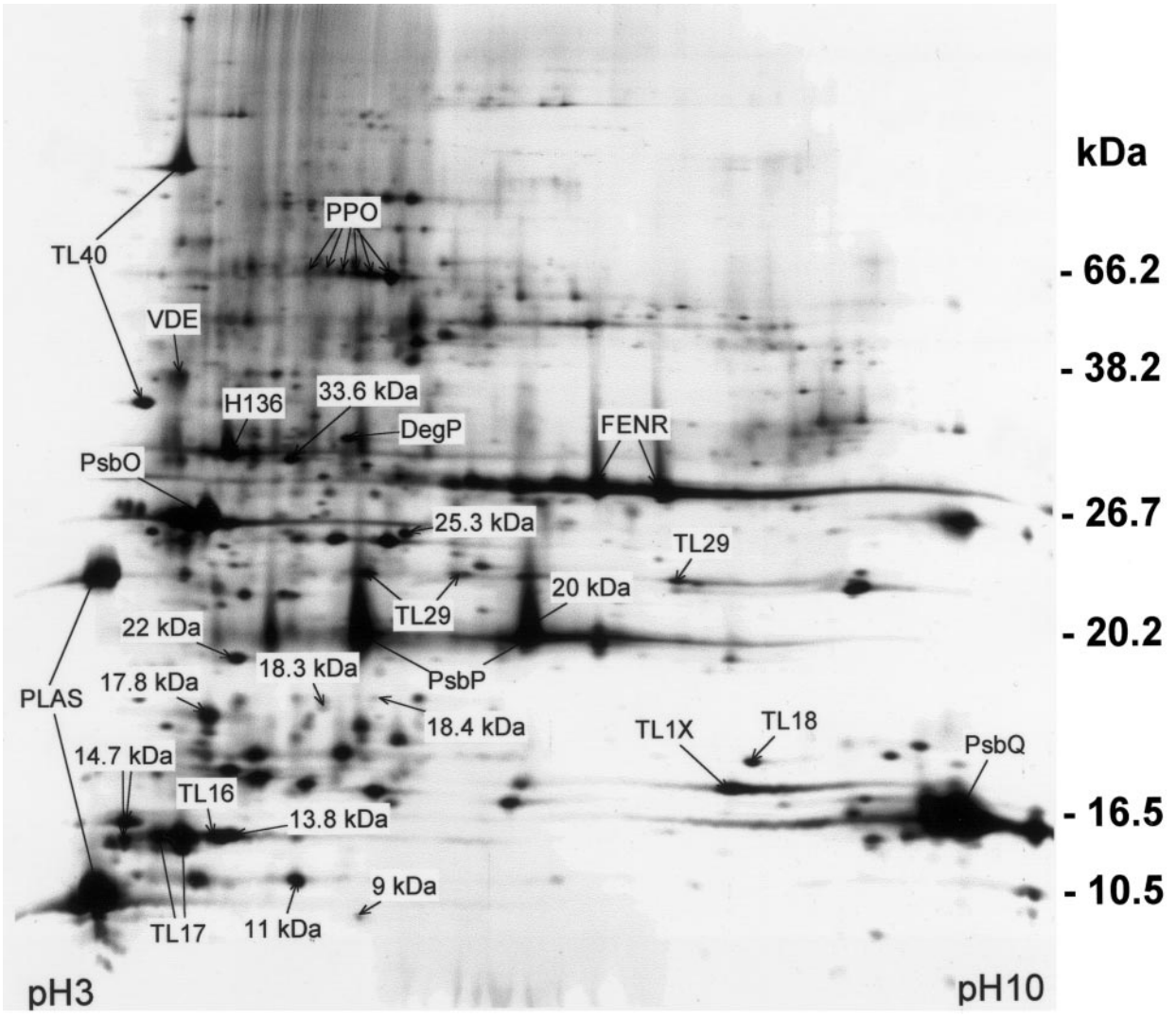

FIG. 3. **Silver-stained two-dimensional gel of 100 μg of soluble luminal proteins from the chloroplast of spinach.** The proteins were resolved by SDS electrophoresis in a 9–16% polyacrylamide gradient gel subsequent to isoelectric focusing in a nonlinear immobilized pH gradient from pH 3 to 10.

the *Arabidopsis* chloroplast for the presence of polyphenol oxidase activity. In two preparations of lumen fraction from Arabidopsis, no polyphenol oxidase activity was detectable (data not shown), and, as with garden pea, this enzyme seems not to be present in *Arabidopsis*. However, for 22 of the luminal proteins from the spinach chloroplast, a homologue in *Arabidopsis* was identified. In addition, there are 12 homologous proteins from garden pea (18) that can be related to both the luminal proteins from spinach and Arabidopsis. All of these data indicate that the luminal compartments of these three plants essentially comprise the same proteins.

Prediction of the Luminal Chloroplast Proteins of Arabidopsis—As a complement to our experimental characterization of the thylakoid lumen from *Arabidopsis*, we made a theoretical prediction of the luminal proteins from the whole *Arabidopsis* proteome to estimate the entire number of proteins in the thylakoid lumen. To detect the luminal proteins within the *Arabidopsis* proteome, we used the bipartite transit peptides as a marker because they contain the information that is necessary to target these proteins to the chloroplast lumen and, hence, are specific for their subcellular location. The prediction was carried out in three steps. First, we screened the latest version of the complete FASTA database of *Arabidopsis* proteins with the program TargetP for chloroplast-located proteins. From a total number of 25,657 *Arabidopsis* proteins, 3765 were predicted to have a subcellular location in the chloroplast. In the second stage, we screened the entire TargetP-predicted chloroplast proteome for potential signal peptides using the program SignalP 2.0 and obtained 514 signal peptides that were at least 30 residues in length. Among those, 358 had a length between 50 and 120 residues and thus were candidates for transit peptides of luminal proteins. Finally, the 358 preselected sequences were examined manually for typical features of bipartite transit peptides such as a hydrophilic serine- and threonine-rich amino-terminal region and a hydrophobic domain close to the processing site. In addition, the complete precursors were scanned for the presence of known pattern and profiles to exclude proteins that could not reside in the thylakoid lumen such as ion channels or chlorophyll-binding antennae proteins. To avoid false positives and to keep the number of missed luminal proteins to a minimum, the manual analysis of the SignalP-selected sequences was carried out two times, and conflicts between the two prediction cycles were examined again in each single case. From the 358 sequences that were individually analyzed, 303 were excluded, and 55 were predicted to belong to putative luminal proteins. The result from this evaluation is summarized in Table III, and a

TABLE II

Proteins from the chloroplast lumen of spinach

MALDI, MALDI-TOF-MS; PSD, postsource decay analysis, Micro, microsequencing; ND, not determined.

Protein name	Accession no.[a]	Identification	Exp. mass	Experimental NH_2-terminal sequence	Homologue in *Arabidopsis*	Homologue in garden pea
			kDa			
Xanthophyll cycle						
Violaxanthin de-epoxidase	Q9SM43	MALDI, Micro	43.1	VDALKTCTXL	Violaxanthin de-epoxidase (Q39249)	
Photosystem II assembly						
Putative HCF136	P82714	PSD, Micro	34.6	EDSLSDWERVYLPIDPGVVL	H136_ARATH (O82660)	37.3-kDa protein (P82342)
Photosystem II subunits						
PSBO_SPIOL	P12359	MALDI, Micro	ND	EGGKRLTY	PSO1_ARATH (P23321)	PSBO_PEA (P14226)
PSBP_SPIOL	P12302	MALDI, Micro	ND	AYGEA	PSP1_ARATH (Q42029)	PSBP_PEA (P16059)
PSBQ_SPIOL	P12301	MALDI, Micro	ND	EARPIVVGPPPPLSG	PSQ1_ARATH (Q9XFT3)	PSBQ_PEA (P19589)
Putative PsbP domain proteins						
22-kDa protein	P82796	Micro	22.0	EQSAGFREYIDFFDGYSLTY	T215_ARATH (O23403)	24-kDa protein (P82337)
TL1X_SPIOL	P82537	Micro	17.0	AESKKGFLPVIDKKDGYTFL YPFGGQEVSI	TL26_ARATH (P82538)	16.9-kDa protein (P82340)
20-kDa protein	P83090	Micro	20.0	RDVDVGSFLP KSPSDPSMVL	20-kDa protein (Q9LXX5)	24.1-kDa protein (P82329)
Plastocyanins						
PLAS_SPIOL	P00289	MALDI, PSD, Micro	ND	VEVLLGG	PLAT_ARATH (P42699)	PLAS_PEA (P16002)
Putative peroxidase						
TL29_SPIOL	P81833	Micro	23.1	ADLIQRRQXS	TL29_ARATH (P82281)	28.5-kDa protein (P82338)
Putative proteases						
Putative DEGP protease	P83091	Micro	36.3	FVVSTPXKLQ	DEGP_ARATH (O22609)	46-kDa protein
Immunophilins						
TL40_SPIOL	O49939	MALDI, Micro	ND	VLISGPP	40-kDa protein (Q9SSA5)	
Putative immunophilins						
33.6-kDa protein	P83092	Micro	33.6	VLYSPDTKVPR	38-kDa protein (P82869)	
TL18_SPIOL	P82536	Micro	17.7	SAEETPLQSKVTNKVVFDIS	18-kDa protein (Q9LYR5)	
17.8-kDa protein	P83061	Micro	17.8	AGLPPEEKPKLCDAACE	17.5-kDa protein (O22870)	
Putative pentapeptide proteins						
TL17_SPIOL	P81778	Micro	13.5	ANQRLPPLSNDPDRCERAFVGNT	TL17_ARATH (P81760)	18.3-kDa protein (P82326)
11-kDa protein	P82657	Micro	11.2	FKGGGPYGQGVTRGQDLSGKDF	11.6-kDa protein (O22160)	14.3-kDa protein (P82322)
18.3-kDa protein	P82806	Micro	18.3	DLNKFEAEMRGEFGIXSA	Putative 20.1-kDa protein (Q94C72)	
Proteins with unknown function						
Polyphenol oxidase	P43310	Micro	ND	APILPDVEKCTLSDALWDG	Not found	Not present in this plant
25.3-kDa protein	P83089	Micro	25.3	AIANAPLLDTTITDRVFFD	Not found	
18.4-kDa protein	P82799	Micro	18.4	DSPTPNTYNIYYGTAASAXN	19-kDa protein (P82658)	
TL16_SPIOL	P81834	Micro	13.3	APLEDEDDLELLEKVKRDRKK RLERQGAIN	TL16_ARATH (O22773)	16-kDa protein (P82323)
13.8-kDa protein	P82681	Micro	13.8	LDEFRVYSDDANKYKISIPQD	15.9-kDa protein (Q9S720)	
14.7-kDa protein	P82682	Micro	14.7	KTGVNKPELLPKEETTVIDV	15-kDa protein (Q9LVV5)	18.2-kDa protein (P82325)
9-kDa protein	P82671	Micro	9.1	GFLSGSTGIEXIPGPQL	Putative 11.9-kDa protein (AT2g03420)[b,c]	
Stromal and other proteins						
FENR_SPIOL	P00455	MALDI, Micro	31.2	QIASDVEA	FENR (Q9FKW6)	

[a] Swiss-Prot/TrEMBL.
[b] Tigr *Arabidopsis* db.
[c] Gene corrected in this work.

TABLE III
Prediction of proteins from the chloroplast lumen of A. thaliana

Gene locus	SignalP-NN prediction	Predicted processing site	Protein name and accession	Putative import pathway
Photosystem II assembly				
AT5g23120	78	78_79 ARA_DE	H136_ARATH (O82660)[a]	Tat
Photosystem II subunits				
AT3g50820	86	84_86 AGA_EG	PSO2_ARATH (Q9S841)[a]	Sec
AT5g66570	85	85_86 ASA_EG	PSO1_ARATH (P23321)[a]	Sec
AT4g05180	82	82_83 VFA_EA	PSQ2_ARATH (Q41932)[a]	Tat
AT1g06680	77	77_78 ADA_AY	PSP1_ARATH (Q42029)[a]	Tat
AT4g21280	75	75_76 VLA_DA	PSQ1_ARATH (Q9XFT3)[a]	Tat
AT2g30790	70	70_71 AIG_SK	PSP2_ARATH (O49344)[b]	Tat
Putative PsbP domain proteins				
AT4g15510	104	104_105 AFA_ST	T215_ARATH (O23403)[a]	Tat
AT5g11450	95	95_96 THA_EE	35.8-kDa protein (P82715)[a]	Tat
AT1g76450	80	80_81 AFA_ET	15.9-kDa protein (Q9S720)[a]	Tat
AT3g55330	74	74_75 SFA_AE	TL26_ARATH (P82538)[a]	Tat
AT2g39470	73	73_74 LLA_EE	P18.6 (O80634)[b]	Tat
Putative PsbQ domain proteins				
AT3g01440	67	74_75 SLA_QD	P24.8 (Q9SGH4)	Tat
AT1g14150	65	65_66 ALA_QE	P22.2 (Q9XI73)	Tat
Plastocyanins				
AT1g76100	72	72_73 AMA_ME	PLAS_ARATH (P11490)[a]	Sec
AT1g20340	68	68_69 AMA_IE	PLAT_ARATH (P42699)[a]	Sec
Photosystem I subunits				
AT5g64040	86	86_87 ANA_GV	PSAN_ARATH (P49107)	Tat
Proteases				
Tail-specific proteases				
AT4g17740	119	119_120 VTT_DS	D1 processing protease (O23614; Q9ZP02)[b]	Sec
AT3g57680	99	99_100 ALA_ES	P45.4 (Q9SVY2)	Tat
AT5g46390	68	68_69 VAA_TN	Carboxyl-terminal protease D1 like protein (Q9FL23)[a]	Sec
Serine proteases, trypsin family				
AT3g27925	102	105_106 ASA_FV	DEG1_ARATH (O22609)[a,b]	Sec
Immunophilins				
Cyclophilin-type peptidyl-prolyl cis-trans isomerases				
AT3g15520	103	114_115 ALA_VL	38-kDa protein (P82869)[a]	Sec
AT3g01480	82	91_92 DVS_VL	40-kDa protein (Q9SSA5)[a]	Sec
AT1g74070	74	74_75 AQA_DT	P22.2 (Q9C9C7)[b]	Tat
FKBP-type peptidyl-prolyl cis-trans isomerases				
AT1g18170	94	94_95 VIS_EQ	P16.5 (Q9LDY5)	Tat
AT3g10060	82	82_83 AEA_VS	P17.8 (Q9SR70)	Tat
AT4g19830	78	78_79 ALA_DF	P16.7 (O81864)	Tat
AT2g43560	73	76_77 AYA_AG	17.5-kDa protein (O22870)[a,b]	Tat
AT4g39710	73	73_74 ADA_TR	P12 (O65658)[b]	Sec
AT1g20810	71	71_72 SEA_RE	17.8-kDa protein (Q9LM71)[a,b]	Tat
AT3g60370	53	53_54 SSS_AK	16.9-kDa protein (Q9M222)[a]	Tat
Pentapeptide proteins				
AT2g44920	81	81_82 ALA_FG	11.6-kDa protein (O22160)[a,b]	Sec
AT1g12250	79	83_84 AMA_EL	P20.1 (Q94C72)[b]	Sec
AT5g53490	77	77_78 VIA_AN	TL17_ARATH (P81760)[a]	Sec
Unknown function				
AT3g44020	103	70_71 SSA_FD	P17.1 (Q9LXV9)	Sec
AT2g40400	93	93_94 ASA_ET	P71 (Q9SIY5)	Tat
AT2g37400	91	91_92 AIA_AP	P28.3 (Q9ZUS6)	Sec
AT3g56140	90	90_91 SLA_SE	P73.5 (Q9LYM7)	Tat
AT1g33780	88	88_89 GDA_SQ	P28.3 (Q9LQ30)	Sec
AT1g54780	84	84_85 ALA_SE	18.3-kDa protein (Q9ZVL6)[a]	Sec
AT3g09490	84	81_82 FSA_SF	P38.4 (Q9SF54)	Sec
AT5g02590	84	84_85 VTA_AT	P36.5 (Q9LZ43)[b]	Sec
AT2g26340	80	80_81 AMA_GG	P19.1 (O48702)[b]	Tat
AT1g51400	76	76_77 AMA_DD	P3.3 (Q9SYE2)	Tat
AT5g52970	75	75_76 ADA_KV	15.0-kDa protein (Q9LVV5)[a,b]	Sec
AT1g79450	74	71_72 LFA_SQ	P37.9 (Q9SAK3)[b]	Sec
AT1g51350	73	73_74 ADA_DD	P65.8 (Q9SYD7)[b]	Sec
AT2g23670	71	71_72 SMA_EN	P10 (O64835)	Tat
AT1g03600	68	68_69 VSA_AE	P11.7 (Q9LR64)	Tat
AT1g62140	66	64_65 EMA_AA	P40.1 (O04580)[b]	Tat
AT4g24930	63	63_64 ALA_IP	17.9-kDa protein (Q9SW33)[a]	Sec
AT1g14590	62	61_62 YRA_AD	P38 (Q9MA24)	Tat
AT5g46560	59	59_60 AVA_FT	P36.4 (Q9L527)	Sec
AT1g21500	59	59_60 ALA_AK	P13.2 (Q9LPK9)	Tat
AT3g08550	57	57_58 GLA_DP	P53.8 (Q9C9Z9)	Sec

[a] Identified in this work.
[b] Conflict between the sequence in the Tigr and in the Swiss-Prot/Trembl database.

TABLE IV
Summary of the prediction of luminal chloroplast proteins of A. thaliana

Protein	No. of sequences
Arabidopsis proteome	25,657
TargetP-predicted chloroplast proteome	3765
Predicted luminal proteins	55
Experimentally identified luminal proteins	36
Predicted and experimentally identified luminal proteins (overlap excluded)	66
Estimated number of luminal proteins	80[a]

[a] 0.3% of the *Arabidopsis* proteome, 2% of the putative chloroplast proteome.

summary of the single steps of the prediction is provided in Table IV.

To evaluate the prediction, we used the 35 experimentally identified luminal proteins for which full-length sequences were available. The 19-kDa protein, for which no gene is known yet, had to be excluded. The TargetP-predicted chloroplast proteome contained 34 of the identified luminal proteins. Violaxanthin de-epoxidase was missed by TargetP, which probably was due to its unusually long transit peptide. From the 34 identified luminal proteins that were present in the predicted chloroplast proteome, 25 were predicted to be lumen-located. The main reason why nine proteins escaped prediction was the preselection of putative luminal proteins with SignalP that failed to recognize these proteins. Remarkably, all of their precursors but one had a signal peptide with a twin-arginine motif, which indicates that we particularly underestimated proteins of the Tat pathway. In summary, the prediction found 25 of the 35 experimentally identified luminal proteins that were present in the *Arabidopsis* proteome, which was a rate of 71%. If we assume that these proteins are a representative part of the luminal proteome, we can extrapolate from the number of 55 predicted luminal proteins and estimate that the thylakoid lumen comprises ~80 proteins.

DISCUSSION

The investigation into the chloroplast lumen from *Arabidopsis* and from spinach that was carried out in this work showed clearly that the luminal proteins from both plants correspond well, and the luminal proteome from *Arabidopsis* can serve as a model for other higher plants. To better understand the proteins from the chloroplast lumen, we compared our data with the annotated *Arabidopsis* proteome at the European Bioinformatics Institute that is available on the World Wide Web at www.ebi.ac.uk/proteome. This analysis showed that the proteins that were found in the chloroplast lumen do not contain any members of the 10 biggest protein families in *Arabidopsis* such as eukaryotic protein kinases or proteins of the F-box domain family. By contrast, the luminal proteins that were identified belong to families that consist of a rather low number of proteins, of which, however, a relatively large part is located in the thylakoid lumen. A good example for the specific pattern of the luminal proteome are the luminal immunophilins. From a total number of 22 FKBP-type PPIases that are annotated in the *Arabidopsis* proteome at the European Bioinformatics Institute, five were detected in the chloroplast lumen. In addition, four further FKBP-type PPIases were found among the predicted luminal proteins (Table III). The group of luminal proteases has a specific composition too. It includes all known members of the family of tail specific proteases: the D1-processing protease, the processing protease D1-like protein, and the putative 45-kDa protein Q9SVY2. The last was not identified in this work but was found among the predicted proteins (Table III). In addition, there are three luminal proteases that belong to the trypsin family of the serine proteases, of which 15 members are annotated in the *Arabidopsis* proteome. We did not find any member of any of the big families of the subtilisin-type serine proteases with 53 annotated proteins and the sumo-specific proteases with 62 annotated proteins. As with the pentapeptide repeat proteins that are a big family in bacteria and cyanobacteria (41), only four proteins with this motif are annotated in the *Arabidopsis* proteome. Two of those, the 17-kDa protein TL17 and the 11.6-kDa protein T116, were identified in the thylakoid lumen, and a third one, the putative 24-kDa protein Q9FWX1, was predicted to be lumen-located (Table III). Although the functions of these proteins are unknown, they probably play a role that is specifically needed in the thylakoid lumen. Furthermore, the photosystem II assembly factor Hcf136 and the enzyme violaxanthin de-epoxidase contribute to the specific composition of the luminal proteome. The sequence of Hcf136 possesses a BNR repeat motif that, for instance, is found in many bacterial and eukaryotic glycosyl hydrolases but only in two *Arabidopsis* proteins. Violaxanthin de-epoxidase has a lipocalin motif, which is very rare among the proteins of *Arabidopsis* but, for example, is present in many eukaryotic fatty acid binding proteins. In addition, there is a novel family of PsbP domain proteins of which no members in other subcellular compartments are known. Besides the known PsbP1 protein of photosystem II, six further members of this family were identified, and two more were found among the predicted luminal proteins. The large number of these proteins indicates that they fulfill an important function in the thylakoid lumen. In summary, all of these data show that the chloroplast lumen of *Arabidopsis* not only contains a large number of proteins but also has its own specific proteome.

The identified and predicted luminal proteins of *Arabidopsis* in Tables I and III correspond well, and this remarkable agreement indicates that both the experimental and the *in silico* approach have covered a representative part of the luminal proteins. An important result of the *in silico* approach was the discovery of two putative PsbQ-related proteins that provide a complement to the family of PsbP domain proteins. The 24.8-kDa precursor Q9SGH4 and the 22.2-kDa precursor Q9XI73 possess the Prodom domain PD007524 that also is a typical feature of the extrinsic photosystem II subunits PsbQ1 and PsbQ2 from *Arabidopsis*. This domain is only shared by the currently known PsbQ proteins from higher plants, which indicates that it is a characteristic of those proteins. The extrinsic proteins PsbO, PsbP, and PsbQ play an important role in the established model of photosystem II from higher plants (2, 42, 43). It is generally accepted that these proteins participate in the regulation of oxygen evolution and are present in photosystem II in a stoichiometric ratio of 1:1:1. While it has not been entirely clarified whether the photosystem II complex has one or two copies of each extrinsic protein (2), recent crystal structure data indicated that there is one copy of the PsbO protein per photosystem II (43). The discovery of novel PsbP domain proteins in this study suggests that the luminal surface of photosystem II might have a more complex composition than previously believed. Classical reconstitution experiments with thylakoids from spinach showed that photosystem II does not require the PsbP protein to produce oxygen. However, it has a considerably higher oxygen-evolving activity if the PsbP protein is bound to the thylakoid membrane (44). Consistent results were obtained from a study of two types of photosystem II complexes from spinach, of which one contained only the extrinsic PsbO protein and the other one contained both the PsbO and PsbP protein (45). The novel PsbP domain proteins of the

thylakoid lumen might fulfill similar functions and provide for photosystem II a tool to modulate its oxygen-evolving activity.

While the extrinsic PsbO protein is found in the photosystem II of both higher plants and cyanobacteria, the PsbP and PsbQ proteins are only known in higher plants, and it is believed that the cyanobacterial proteins PsbU and PsbV fulfill the function of the higher plant proteins PsbP and PsbQ (46). However, the hypothetical 20.7-kDa protein P73952 from the cyanobacterium *Synechocystis* sp. PCC 6803 has a distinct PsbP domain, and it is a homologue to the PsbP-like T215 protein from *Arabidopsis*. Currently, the function of the 20.7-kDa protein is unknown, but deletion mutants in *Synechocystis* sp. PCC6803 have been performed, and studies are under progress to find out whether this protein plays a role in photosynthesis.[2]

Besides the PsbP domain proteins, immunophilins from the families of the cyclophilin- and FKBP-type PPIases were the biggest group of proteins that was found in the chloroplast lumen of *Arabidopsis*. As a general feature, immunophilins have the ability to catalyze the cis-trans isomerization of proline-imidic peptide bonds and to accelerate protein folding (47, 48). Remarkably, the chloroplast lumen contains an unusually big group of FKBP-type PPIases. An analysis of the signal peptides of all TargetP-predicted chloroplast immunophilins from the annotated *Arabidopsis* proteome at the European Bioinformatics Institute indicated that the FKBP-type PPIases of the chloroplast only occur in the thylakoid lumen (data not shown). That implies that these proteins fulfill specific rather than general functions that are particularly needed in the thylakoid lumen. In contrast to their mammalian relatives, plant immunophilins are a relatively recent discovery, and few of them have been functionally characterized. The only luminal immunophilin that was studied in detail is the cyclophilin TL40 from spinach that regulates the dephosphorylation of thylakoid membrane proteins by binding to a PP2A-like phosphatase (49, 50). The functions of the other luminal immunophilins are currently unknown, but these proteins could participate in processes such as activation or inhibition of other thylakoid proteins and protein folding and assembly. A putative target group for the activity of luminal immunophilins is, for instance, the extrinsic proteins PsbO, PsbP, and PsbQ. These proteins show unusual behavior in that they not only bind to the thylakoid membrane but also exist in the lumen in an unassembled state, in which they are long lived and assembly-competent (51, 52). Since unassembled proteins usually are rapidly degraded, it could be hypothesized that luminal immunophilins participate in the protection of these unassembled extrinsic proteins and in their proper assembly to photosystem II.

The novel luminal proteins from *Arabidopsis* not only show that the chloroplast lumen has a characteristic proteome but also illustrate the impact of the ΔpH-dependent Tat pathway for the protein traffic into this compartment. As Fig. 2 demonstrates, 19 of 35 luminal protein precursors have a signal peptide with a twin arginine motif and are marked for translocation by the Tat complex. That indicates that more than half of the luminal proteins from *Arabidopsis* might be routed across the thylakoid membrane via the Tat pathway. In combining the experimentally identified 35 thylakoid lumen proteins for which genes are known with the 55 predicted ones, there are a total of 66 *Arabidopsis* proteins that are presumed to occupy the thylakoid lumen. The genes for these proteins appear to be near evenly distributed among all five *Arabidopsis* chromosomes: 20 within chromosome 1, 9 within chromosome 2, 14 within chromosome 3, 10 within chromosome 4, and 13 within chromosome 5. In addition, they appear to be evenly distributed within each chromosome as well, and only four genes are within 100 kb of another gene encoding a presumed thylakoid-targeted protein. It should be noted, too, that none of the predicted or identified luminal proteins possessed any known ATP binding sites; thus, this work does not confirm previous reports on the presence of Hsp70 and related proteins in the thylakoid lumen (18, 53). In earlier studies, we analyzed the thylakoid lumen of spinach for the presence of ATPase activity (12, 54), but the detectable activity was so low that it was not specific for the luminal proteins. That does not exclude the possibility that ATP or other nucleotides are used in the chloroplast lumen, but so far evidence for this has not been provided, and the source of energy for the luminal proteins is still unknown.

The annotation of protein coding genes within the *Arabidopsis* genome is heterogeneous, and the accuracy of each annotated gene is based on the available evidence that supports it. That includes homology to protein and nucleotide sequences, and in the extreme case, it is limited to computationally predicted genes. Most gene annotations are based on a combination of methods. The importance of experimentally determined sequence data in providing more accurate annotations is greatly exemplified by this work. When we searched the *Arabidopsis* genome for the experimentally identified luminal proteins, we found six genes that required gene structure annotation refinements that altered the annotated protein coding sequence of the gene. TIGR's *Arabidopsis* genome reannotation efforts are greatly benefiting from such genomic and proteomic studies. The identification of these genes as proteins of the chloroplast lumen also provides experimentally determined cellular localization data, which can be represented within the context of Gene Ontology assignments (55). The luminal 19-kDa protein, which could not be mapped to the *Arabidopsis* genome, demonstrates the need for continued sequencing efforts to further complete the *Arabidopsis* genome. Such efforts are currently in progress at TIGR, and a high priority has been established to identify the sequence containing the gene for this 19-kDa protein.

In summary, this study has shown that the chloroplast lumen not only fulfills a function for the generation of the pH gradient that fuels ATP synthesis but also has its own specific proteome. In the chloroplast lumen from *Arabidopsis*, 36 proteins were identified, and the entire luminal proteome of *Arabidopsis* was estimated to comprise ~80 proteins. This suggests that the narrow luminal space of the thylakoid membrane is densely packed with proteins. The discovery of the novel PsbP domain proteins and immunophilins in the chloroplast lumen of *Arabidopsis* indicates that luminal proteins play an important role for the regulation of photosynthesis.

Acknowledgments—We thank Professor Jan-Åke Gustafsson (Karolinska Institute, Huddinge, Sweden) for support and Dr. Fredrik Nilsson (Astra Zeneca AB, Göteborg, Sweden) for advice and discussions concerning mass spectrometry analysis.

REFERENCES

1. Albertsson, P.-Å. (1995) *Photosynth. Res.* **46,** 141–149
2. Wollman, F.-A., Minai, L., and Nechushtai, R. (1999) *Biochim. Biophys. Acta* **1411,** 21–85
3. Hind, G., Nakatani, H. Y., and Izawa, S. (1974) *Proc. Natl. Acad. Sci.* **71,** 1484–1488
4. Pottosin, I. I., and Schönknecht, G. (1995) *J. Membr. Biol.* **148,** 143–156
5. Pottosin, I. I., and Schönknecht, G. (1996) *J. Membr. Biol.* **152,** 223–233
6. Ettinger, W. F., Clear, A. M., Fanning, K. J., and Peck, M. L. (1999) *Plant Physiol.* **119,** 1379–1385
7. Hager, H., and Holocher, K. (1994) *Planta* **192,** 581–589
8. Sommer, A., Ne'eman, E., Steffens, J. C., Mayer, A. M., and Harel, E. (1994) *Plant Physiol.* **105,** 1301–1311

[2] C. Funk and W. P. Schröder, unpublished data.

9. Sokolenko, A., Fulgosi, H., Gal, A., Altschmied, L., Ohad, I., and Herrmann, R. G. (1995) *FEBS Lett.* **371,** 176–180
10. He, W.-Z., and Malkin, R. (1992) *FEBS Lett.* **308,** 298–300
11. Oelmüller, R., Herrmann, R. G., and Pakrasi, H. B. (1996) *J. Biol. Chem.* **271,** 21848–21852
12. Kieselbach, T., Hagman, Å., Andersson, B., and Schröder, W. P. (1998) *J. Biol. Chem.* **273,** 6710–6716
13. Kieselbach, T., Mant, A., Robinson, C., and Schröder, W. P. (1998) *FEBS Lett.* **428,** 241–244
14. Mant, A., Kieselbach, T., Schröder, W. P., and Robinson, C. (1999) *Planta* **207,** 624–627
15. Fulgosi, H., Vener, A. V., Altschmied, L., Herrmann, R. G., and Andersson, B. (1998) *EMBO J.* **17,** 1577–1587
16. Meurer, J., Plücken, H., Kowallik, K. V., and Westhoff, P. (1998) *EMBO J.* **17,** 5286–5297
17. Kieselbach, T., Bystedt, M., Hynds, P., Robinson, C., and Schröder, W. P. (2000) *FEBS Lett.* **480,** 271–276
18. Peltier, J.-B., Friso, G., Kalume, D. E., Roepstorff, P., Nilsson, F., Adamska, I., and van Wijk, K. J. (2000) *Plant Cell* **12,** 319–341
19. The Arabidopsis Genome Initiative (2000) *Nature* **408,** 796–815
20. Norén, H., Svensson, P., and Andersson, B. (1999) *Biosci. Rep.* **19,** 499–509
21. Bradford, M. M. (1976) *Anal. Biochem.* **72,** 248–254
22. Laemmli, U. K. (1970) *Nature* **227,** 680–685
23. Bjellqvist, B., Pasquali, C., Ravier, F., Sanchez, J.-C., and Hochstrasser, D. (1993) *Electrophoresis* **14,** 1357–1365
24. Shevchenko, A., Wilm, M., Vorm, O., and Mann, M. (1996) *Anal. Chem.* **68,** 850–858
25. Pandey, A., Andersen, J. S., and Mann, M. (2000) *Science's stke*, 37/pl1
26. Gevaert, K., Demol, H., Martens, L., Hoorelbeke, B., Puype, M., Goethals, M., Van Damme, J., De Boeck, S., and Vandekerckhove, J. (2001) *Electrophoresis* **22,** 1645–1651
27. Matsudaira, P. (1987) *J. Biol. Chem.* **262,** 10035–10038
28. Pearson, W. R. (1990) *Methods Enzymol.* **183,** 63–98
29. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402
30. Hofmann, K., Bucher, P., Falquet, L., and Bairoch, A. (1999) *Nucleic Acids Res.* **27,** 215–219
31. Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L., and Sonnhammer, E. L. (2000) *Nucleic Acids Res.* **28,** 263–266
32. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) *Nucleic Acids Res.* **22,** 4673–4680
33. Kyte, J., and Doolittle, R. F. (1982) *J. Mol. Biol.* **157,** 105–132
34. Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G. (2000) *J. Mol. Biol.* **300,** 1005–1016
35. Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. (1997) *Protein Eng.* **10,** 1–6
36. Nielsen, H., Brunak, S., and von Heijne, G. (1999) *Protein Eng.* **12,** 3–9
37. Robinson, C., Thompson, S. J., and Woolhead, C. (2001) *Traffic* **2,** 1–7
38. Robinson, C., and Bolhuis, A. (2001) *Nat. Rev. Mol. Cell. Biol.* **2,** 350–356
39. Hieber, A. D., Bugos, R. C., and Yamamoto, H. Y. (2000) *Biochim. Biophys. Acta,*. **1482,** 84–91
40. Itzhaki, H., Naveh, L., Lindahl, M., Cook, M., and Adam, Z. (1998) *J. Biol. Chem.* **273,** 7094–7098
41. Bateman, A., Murzin, A. G., and Teichman, S. A. (1998) *Protein Sci.* **7,** 1477–1480
42. Debus, R. J. (1992) *Biochim. Biophys. Acta* **1102,** 269–352
43. Nield, J., Orlova, E. W., Morris, E. P., Gowen, B., van Heel, M., and Barber, J. (2000) *Nat. Struct. Biol.* **7,** 44–47
44. Ljungberg, U., Jansson, C., Andersson, B., and Åkerlund, H. E. (1983) *Biochem. Biophys. Res. Commun.* **113,** 738–744
45. Boekema, E. J., Nield, J., Hankamer, B., and Barber, J. (1998) *Eur. J. Biochem.* **252,** 268–276
46. Shen, J. R., Ikeuchi, M., and Inoue, Y. (1992) *FEBS Lett.* **301,** 145–149
47. Schiene-Fischer, C., and Yu, C. (2001) *FEBS Lett.* **495,** 1–6
48. Göthel, S. F., and Marahiel, M. A. (1999) *Cell. Mol. Life Sci.* **55,** 423–436
49. Vener, A. V., Rokka, A., Fulgosi, H., Andersson, B., and Herrmann, R. G. (1999) *Biochemistry* **38,** 14955–14965
50. Rokka, A., Aro, E.-M., Herrmann, R. G., Andersson, B., and Vener, A. V. (2000) *Plant Physiol.* **123,** 1525–1535
51. Ettinger, W. F., and Theg, S. M. (1991) *J. Cell Biol.* **115,** 321–328
52. Hashimoto, A., Yamamoto, Y., and Theg, S. M. (1996) *FEBS Lett.* **391,** 29–34
53. Schlicher, T., and Soll, J. (1996) *FEBS Lett.* **379,** 302–304
54. Schröder, W. P., Höflich, J., Hagman, Å., Andersson, B., and Kieselbach, T. (1998) *Photosynthesis: Mechanisms and Effects* (Garab, G. ed) Vol. 3, pp. 2075–2078, Kluwer Academic Publishers Group, Dordrecht, The Netherlands
55. The Gene Ontology Consortium (2000) *Nat. Genet.* **25,** 25–29