Current Biology, Vol. 14, 354–362, March 9, 2004, ©2004 Elsevier Science Ltd. All rights reserved. DOI 10.1016/j.cub.2004.02.039

The *Arabidopsis thaliana* Chloroplast Proteome Reveals Pathway Abundance and Novel Protein Functions

Torsten Kleffmann,[1] Doris Russenberger,[1] Anne von Zychlinski,[1] Wayne Christopher,[2] Kimmen Sjölander,[2] Wilhelm Gruissem,[1] and Sacha Baginsky[1,*]

[1]Institute of Plant Sciences and Functional Genomics Center Zurich
Swiss Federal Institute of Technology
ETH Zentrum, LFW E51.1
Universitätstrasse 2
CH-8092 Zurich
Switzerland

[2]Department of Bioengineering
University of California, Berkeley
Berkeley, California 94720

Summary

Background: Chloroplasts are plant cell organelles of cyanobacterial origin. They perform essential metabolic and biosynthetic functions of global significance, including photosynthesis and amino acid biosynthesis. Most of the proteins that constitute the functional chloroplast are encoded in the nuclear genome and imported into the chloroplast after translation in the cytosol. Since protein targeting is difficult to predict, many nuclear-encoded plastid proteins are still to be discovered.

Results: By tandem mass spectrometry, we identified 690 different proteins from purified *Arabidopsis* chloroplasts. Most proteins could be assigned to known protein complexes and metabolic pathways, but more than 30% of the proteins have unknown functions, and many are not predicted to localize to the chloroplast. Novel structure and function prediction methods provided more informative annotations for proteins of unknown functions. While near-complete protein coverage was accomplished for key chloroplast pathways such as carbon fixation and photosynthesis, fewer proteins were identified from pathways that are downregulated in the light. Parallel RNA profiling revealed a pathway-dependent correlation between transcript and relative protein abundance, suggesting gene regulation at different levels.

Conclusions: The chloroplast proteome contains many proteins that are of unknown function and not predicted to localize to the chloroplast. Expression of nuclear-encoded chloroplast genes is regulated at multiple levels in a pathway-dependent context. The combined shotgun proteomics and RNA profiling approach is of high potential value to predict metabolic pathway prevalence and to define regulatory levels of gene expression on a pathway scale.

Introduction

Chloroplasts are typical plant cell organelles that develop and differentiate from proplastids in a tissue-specific and signal-dependent manner. They are of central importance for cellular metabolism and have many unique roles in processes of global significance, including photosynthesis and amino acid biosynthesis. Chloroplasts are of cyanobacterial origin, but during evolution they lost their autonomy and transferred most of their genes to the nucleus [1]. To date, only limited information is available on the proteome that constitutes the chloroplast and its metabolic functions. First attempts to estimate the protein complement of *Arabidopsis thaliana* plastids by using prediction tools such as TargetP or ChloroP [2, 3] combined with a genome-wide search for genes of cyanobacterial origin [4–6] resulted in more than 3000 candidate proteins. Computer-assisted predictions based on transit peptides are unlikely to reveal the full chloroplast proteome, however, because import pathways of currently unknown mechanisms might exist that are not recognizable by available prediction algorithms. Several known mitochondrial and chloroplast proteins have already been identified that do not follow the canonical import pathways [7–9].

Proteomics is a powerful tool to reveal the protein complement of cell organelles and to obtain new insights into intracellular protein sorting and biochemical pathways. Progress has been made for the proteome analysis of plant mitochondria [10–13], peroxisomes [14], amyloplasts [15], and chloroplasts [16–20], but most of these studies focused on specific organelle compartments. Protein identification from the chloroplast thylakoid lumen and envelope has greatly improved the prediction of suborganelle protein localization [17, 19]. We used tandem mass spectrometry (MS/MS)-shotgun proteomics of chloroplast proteins to develop a map of all metabolic and regulatory pathways in *Arabidopsis thaliana* chloroplasts. Here, we report 690 proteins that we identified with high confidence. The prediction tools TargetP or ChloroP fail to identify the chloroplast location of many of these proteins. Several of the proteins have homologs in cyanobacteria, but not in yeast, supporting their cyanobacterial origin. Analysis of the chloroplast proteome therefore suggests that significantly more evolutionarily conserved proteins are transported into the chloroplast than previously expected from the prediction of transit peptides. For some chloroplast metabolic functions, the abundance of the identified proteins is also reflected by their RNA expression profile, which therefore provides a novel large-scale measure of chloroplast biochemical pathways during plant development when combined with proteomics data.

Results

Protein Identification, Predicted Localization, and RNA Expression Levels

We used purified chloroplasts (for details see Supplemental Experimental Procedures) for two different protein fractionation strategies based on multidimensional chromatography (MDC) of proteins and peptides (strat-

*Correspondence: sacha.baginsky@ipw.biol.ethz.ch

Table 1. Prediction of Protein Localization Based on TargetP

TargetP Prediction	Complete	Envelope	MDC
Plastid	376	58	318
Mitochondria	37	9	28
Secretory pathway	49	25	24
Any other location	142	39	103
Plastid genome	32	6	26
Total	636	137	499

All proteins identified from the MDC (MDC column) and the envelope (ENVELOPE column) fractionation strategies excluding contaminants (Table S2A) were analyzed for their predicted subcellular localization by using TargetP [3] and information retrieved from the MIPS database (http://mips.gsf.de/proj/thal/). The number of proteins identified by MS/MS analysis that are encoded in the chloroplast genome is shown in the row labeled plastid genome.

egy one) and enrichment of envelope membrane proteins (strategy two) because envelope and thylakoid membrane proteins cannot be separated by MDC (Figure S1). Altogether, 690 proteins were identified with high confidence. The MDC approach allowed the exclusive identification of 409 proteins, and 148 proteins were exclusively identified from the envelope fraction, while 133 proteins were identified by both fractionation procedures (Table 1 lists all identified proteins). We have compiled all relevant information on the 690 identified proteins into a database that is available at http://www.pb.ipw.biol.ethz.ch/proteomics/.

We determined to which extent TargetP [3] (http://www.cbs.dtu.dk/services/TargetP/) predicted the chloroplast localization of the identified proteins. Since most of the proteins from the outer envelope membrane do not contain a transit peptide, we separately analyzed the TargetP prediction of nuclear-encoded proteins for the proteins identified by the two fractionation strategies excluding all putative contaminants (a list of proteins that are known abundant proteins in other cell organelles and thus potentially not of chloroplast origin is provided in Table S2A; these proteins were excluded from all subsequent analyses). TargetP predicted a chloroplast localization of only 58 of the 137 envelope proteins, while 9 proteins were predicted to localize to mitochondria, 25 to the secretory pathway, and 39 to "any other location" (Table 1). In contrast, TargetP correctly predicted the chloroplast localization of 318 of the 499 proteins identified by the MDC approach. Only 28 proteins were predicted to localize to mitochondria, 24 to the secretory pathway, and 103 to any other location. Considering TargetP specificity and sensitivity [2], it is likely that the prediction for different subcellular localizations results from incorrect negative predictions of true chloroplast proteins. We cannot exclude, however, that these proteins are targeted to more than one organelle. Dual targeting has been described for several proteins (reviewed in [23]), and there is currently no evidence that the published list is exhaustive. In this context it is interesting to note that proteins with homology to components of the mitochondrial protein import machinery (TIM and TOM) have been identified from chloroplast envelope membranes [20]. Although we cannot completely rule out that some incorrect negative predictions of true chloroplast proteins are due to misannotated N termini, these findings suggest that organellar protein import and protein trafficking are more complex than previously anticipated.

It is also possible that some of the proteins not predicted by TargetP to localize to the chloroplast were from other cell organelles and not removed during chloroplast isolation and purification. Most of these proteins are ribosomal proteins and were found in the preparation of outer envelope proteins. The envelope membrane is in close contact with other cell organelles (e.g., during photorespiration [24]) and the cytosol, and the interaction of the identified cytosolic proteins with the surface of the chloroplast envelope might be of biological relevance. Additional experiments and functional analyses are now required to substantiate the envelope association of the identified proteins and the biological relevance of potential interactions. For the subsequent analyses, we therefore focused our efforts on the proteins identified from the MDC approach. Recent proteome analyses with *Arabidopsis* mitochondria and peroxisomes identified the most abundant proteins of these organelles, which provided a reference to assess the level of potential cross contamination [10, 11, 14]. Based on these studies, we predict that two proteins in our MDC proteome analysis are more likely to be of peroxisomal origin, and five are more likely to be of mitochondrial origin (Table S2B). This indicates that a potential crosscontamination with proteins from these organelles is approximately 1% or less.

We performed RNA profiling analyses by using the *Arabidopsis* ATH1 GeneChip (Affymetrix, Santa Clara) to determine if a correlation exists between the chloroplast proteins identified by MS/MS and their expression levels. The analysis of RNA samples from leaves of the same developmental stage revealed that approximately 65% of the identified proteins were also highly expressed at the RNA level ($>$1000 arbitrary units) (Figure 1). Considering the specialized function of the chloroplast and the dynamic range of the proteins in the chloroplast proteome, this correlation between RNA levels and shotgun MS/MS protein identification provides a useful estimate of protein abundance (Table 2). Based on this correlation, we compared the distribution of TargetP-predicted chloroplast proteins and nonpredicted proteins among the five categories of RNA expression levels shown in Figure 1. As expected, we found a significantly higher proportion of proteins with a chloroplast signal peptide in the two categories of highest RNA expression levels. Proteins without a canonical transit peptide were also found in the categories of lower RNA expression levels. Since contaminations typically result from abundant proteins that are also expressed highly at the RNA level, it is therefore unlikely that our identification of chloroplast proteins without a canonical transit peptide resulted from a systematic contamination of the chloroplast protein fractions with abundant proteins from other cell compartments.

Homologies to Cyanobacterial Proteins

It is broadly accepted that during evolution genes were transferred from the cyanobacterial endosymbiont to

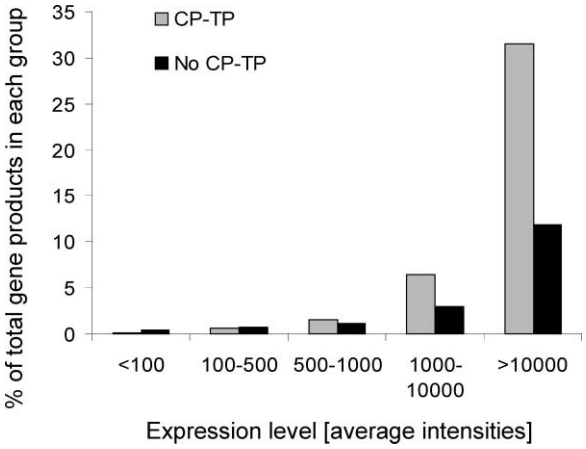


Figure 1. Relative RNA Expression Levels of Identified Proteins

The range of average intensities measured for the expression of genes on the GeneChip was divided into five categories. The number of all *Arabidopsis* genes that gave a presence call on the GeneChip was determined for each of the five expression level categories. The expression levels of proteins identified in our studies (except proteins listed in Table S2A) that are predicted by TargetP to contain a transit peptide (gray bars, CP-TP) were then compared with proteins without a canonical transit peptide (black bars, no CP-TP). The numbers of identified proteins are shown for each expression level category (x axis) as the percentage of the total number of genes that gave presence calls in the respective expression level category (y axis). Absolute numbers in each category were as follows: *Arabidopsis* genes with presence calls in <100, 1001; 100-500, 6094; 500-1000, 2552; 1000-10000, 3142; and >10000, 203; and proteins with a transit peptide versus proteins not predicted by TargetP in <100, 1 versus 4; 100-500, 40 versus 43; 500-1000, 39 versus 29; 1000-10000, 204 versus 94; and >10000, 64 versus 24.

the nucleus of the eukaryotic progenitor cells. Our proteomics data should indicate how many of these genes may encode proteins that are now transported into the chloroplast without a canonical transit peptide. We therefore searched the cyanobacterial (*Synechocystis sp*. strain PCC6803) and *Saccharomyces cerevisiae* genomes (NCBI GenBank) for the identified chloroplast proteins that lack a plastid-targeting signal (excluding plastid-encoded proteins and proteins listed in Table S2A) to identify those proteins that have a homolog in cyanobacteria, but not in yeast [5, 6]. The BLAST homology searches of cyanobacterial and yeast genomes by using both conservative and more permissive thresholds (E-value cutoff of e-10 and e-4, respectively) show a significantly larger fraction of these nonpredicted proteins to have a homolog in cyanobacteria, but not in yeast, than is typical of *Arabidopsis* proteins as a whole (7.4% versus 4.3% at cutoff e-4 and 7.4% versus 3.7% at cutoff e-10; see Figure 2). This result strongly suggests that the identified proteins are of cyanobacterial origin and imported into chloroplasts without a canonical transit peptide.

Functional Assignment of Chloroplast Proteins

We assigned the identified proteins to known chloroplast metabolic and regulatory pathways (http://mips.gsf.de/proj/thal) based on established functional categories (Tables S3–S5). Most of the proteins with known function were found in the categories "metabolism" (173) and "energy" (130), which include proteins with functions in amino acid metabolism, carbohydrate metabolism, and photosynthesis (Table S4). This reflects the main function of the chloroplast in cellular energy metabolism and is consistent with known chloroplast metabolic activities. We next classified the proteins that were identified from the MDC approach and were not predicted to localize to the chloroplast by their enzymatic function (Figure 3). As expected, the functional profiles of predicted and nonpredicted chloroplast proteins were similar. The majority of the proteins not predicted to localize to the chloroplast were of unknown function (Figure 3). These data also allowed us to estab-

Table 2. Correlation of Transcript Abundance and Protein Detection for Important Chloroplast Metabolic Functions

Pathway	RNA Expression Levels of Identified Proteins[a]	Average RNA Expression Levels of Pathway[b]	Proteins Identified/Proteins in Pathway[c]	SpC[d] Protein/RNA
Total	5275	n/a	n/a	0.54
Amino acid biosynthesis	3364	2028	23% (9/40)	0.72
Calvin cycle	13708	11241	93% (14/15)	0.66
Tetrapyrrole biosynthesis	4787	3439	50% (9/18)	−0.25
Nucleus-encoded PS membrane	16873	15205	65% (41/63)	0.42
Plastid-encoded gene expression	n/a	n/a	25% (8/32)	n/a

[a] Average RNA expression levels of all identified proteins from each pathway. Pathway assignment was done on the basis of the KEGG and the TAIR databases. Numbers indicate the arithmetic mean of RNA expression levels for proteins from the designated pathway on the GeneChip. Not applicable (n/a) to plastid-encoded genes.

[b] Average RNA expression levels of all proteins from each pathway including those that were not identified. Pathway assignment was the same as described in footnote a. All isoenzymes and orthologs with a chloroplast targeting signal were taken into account. Not applicable (n/a) to the complete proteome- and plastid-encoded genes.

[c] Pathway coverage achieved with the proteins identified in this study. Not applicable (n/a) to the complete chloroplast proteome.

[d] Spearman rank correlation coefficient (SpC) of transcript and protein levels. Protein levels were ranked by the number of identified tryptic peptides normalized to the molecular weight of the protein (number of tryptic peptides/molecular weight). Not applicable (n/a) to the plastid-encoded genes.

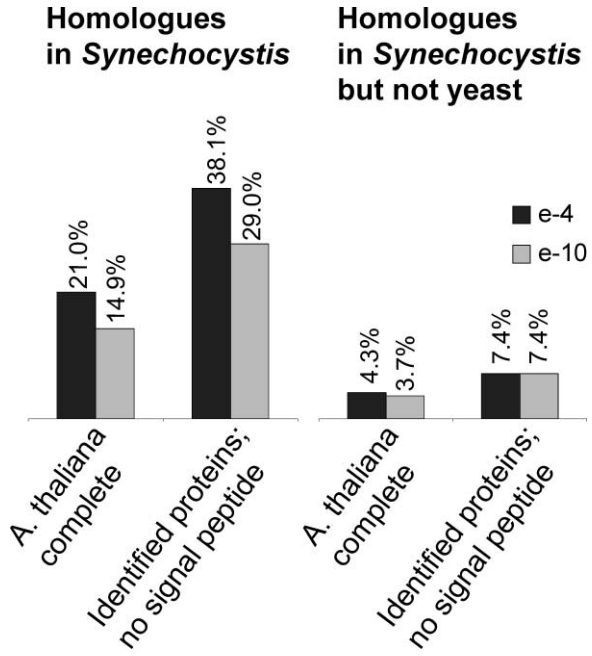


Figure 2. Proteins Identified from Chloroplasts, but Not Predicted, by TargetP to Localize to the Chloroplast Are Enriched for Proteins of Cyanobacterial Origin

Identified chloroplast proteins without a chloroplast transit peptide (identified proteins; no signal peptide) and all proteins from the complete *Arabidopsis* genome (*A. thaliana* complete) were BLAST-searched against the *Synechocystis* protein database (GenBank) and the yeast database. Provided is the percentage of proteins from both groups that have a homolog in *Synechocystis* (left diagram) and those that have a homolog in *Synechocystis*, but not in yeast (right diagram), as identified by BLAST searches by using conservative (e-10 [gray bars]) and more permissive (e-4 [black bars]) thresholds.

lish the chloroplast localization of an essential step in purine biosynthesis. We identified phosphoribosylformylglycinamidine synthase (At1g74260), which is predicted to localize to mitochondria (reliability class 3) in the chloroplast proteome. Phosphoribosylformylglycinamidine synthase is one of three enzymes that catalyze the conversion of glycinamide ribonucleotide to aminoimidazole ribotide. The two other enzymes are predicted to localize to the chloroplast, which strongly argues that these steps in purine biosynthesis occur in plastids. Moreover, phosphoribosylformylglycinamidine synthase was also consistently found in BY-2 plastids (S.B., A. Siddique, and W.G., unpublished data).

The significant number of proteins with unknown function makes it difficult to assign novel metabolic activities to the chloroplast. We therefore applied structure- and function-prediction methods to develop more informative annotations for proteins of unclassified function [25]. Of the 141 proteins in this set annotated as expressed, hypothetical, or unknown, we were able to predict a molecular function and/or 3D structure for at least one domain for 59 (or 42%) by using the bioinformatics approaches described in the Supplemental Experimental Procedures. We illustrate these predictions for four representative sequences: At3g14590, At5g62620, At3g23700, and At5g02180.

Analysis of At3g14590 suggested a potential N-terminal transmembrane domain, followed by two copies of a calcium binding domain (C2; the second domain appears to be truncated and possibly degenerate). At5g62620 has two different domains with related functions, a galectin-3 carbohydrate recognition domain (CRD) near the N terminus, and a C-terminal domain from the galactosyltransferase family. The third example, At3g23700, has four copies of the S1 RNA binding domain. Analysis of At5g02180, which was identified in the membrane fraction, revealed seven putative transmembrane helices. The automatically derived MIPS classification predicted this protein to be an amino acid transporter, which we could confirm by our analysis.

We identified 118 proteins with predicted functions in transport processes (Figure 4A, Tables S3 and S5), including the putative amino acid transporter At5g02180. Most of these proteins were identified in the envelope fraction, which would be consistent with their functions in the outer and inner envelope membranes. Of all identified proteins, 46 were predicted in silico to localize to the inner envelope membrane [26], and nine of these 46 proteins have putative transport functions. Our proteome analysis confirmed the chloroplast envelope localization of these 46 proteins and provides further support for the nine predicted transporter proteins in the chloroplast envelope membranes. In addition, the chloroplast envelope localization of five of the nine predicted transporter proteins was recently reported [19]. We also identified components of the chloroplast protein import complex (TOC and TIC components and the soluble factors [27, 28]), except for TIC20 and TIC40 (Figure 4 A). This could indicate that these two proteins are present at substoichiometric levels compared to the other TIC and TOC proteins or are, perhaps, not constitutive components of the import machinery.

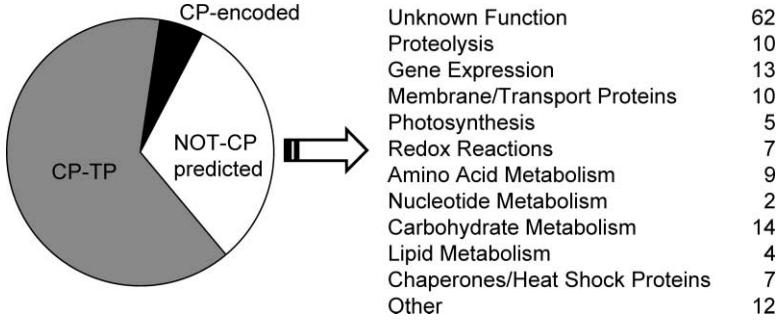
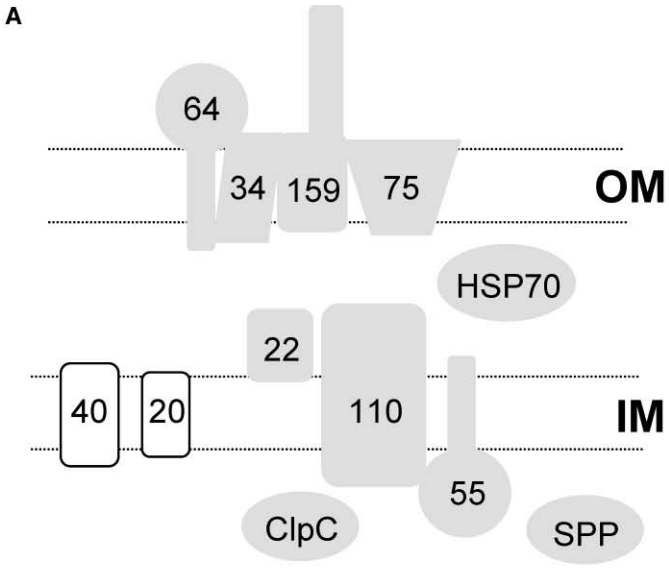


Figure 3. Functional Classification of Proteins Not Predicted to Localize to the Chloroplast (Not-CP Predicted).

Proteins identified from the MDC approach were classified by their putative function on the basis of information available in the MIPS (http://mips.gsf.de), TAIR (http://arabidopsis.org/tools/aracyc), and the KEGG (http://www.kegg.com) databases.



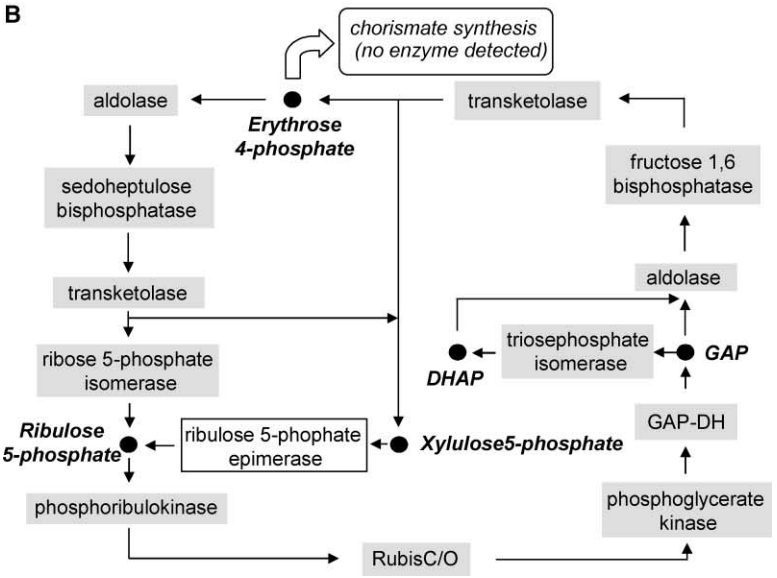


Figure 4. Identified Proteins from the Protein Import Machinery and the Calvin Cycle

Identified proteins are depicted in gray and missing components are white. (A) Except for TIC 20 and 40, all known components of the chloroplast protein import machinery were identified. Abbreviations: OM, outer envelope membrane; IM, inner envelope membrane. (B) Important branch points in the schematic depiction of the Calvin cycle are highlighted by a black dot representing the designated metabolite (http://www.kegg.com). Erythrose 4-phosphate is a precursor for the synthesis of chorismate of which no enzymes were identified. Note that some metabolite fluxes were not depicted. Abbreviations: GAP, glyceraldehyde 3-phosphate; GAP-DH, glyceraldehyde 3-phosphate dehydrogenase; DHAP, dihydroxyacetone phosphate.

Assessment of Pathway Prevalence by Correlation of Transcript and Protein Abundance

The 690 identified proteins represent a cross-section of the full chloroplast proteome. Since MS/MS identification was biased toward abundant proteins (Figure 1), our results most likely reflect the identification of biochemical pathways in the chloroplast that are most active at the developmental stage and growth condition of the plants used in our experiments. Therefore we predicted that our proteomics approach could identify most, if not all, enzymes of active pathways. Less active or inactive pathways would result in a comparatively lower coverage of identified proteins. We tested our hypothesis by analyzing the proteins we identified for important plastid metabolic activities (Table 2 and Figure 4). Coverage of chloroplast functions was, e.g., 93% for Calvin cycle enzymes and 65% for nucleus-encoded photosynthetic electron transport proteins (Table 2). The identified proteins included all enzymes of the Calvin cycle except for a ribose epimerase, which is also expressed at very low RNA levels (Figure 4B). This enzyme catalyzes the conversion of xylulose 5-phosphate to ribulose 5-phosphate, which is important for the regeneration of ribulose 5-phosphate as the primary substrate for CO_2 fixation. The failure to detect ribose epimerase in our MDC approach suggests that ribulose 5-phosphate may be predominantly regenerated from sedoheptulose 1,7-bisphosphate by the concerted action of transke-

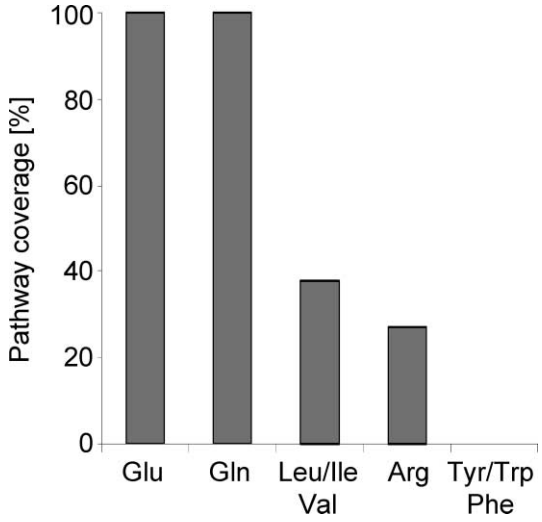


Figure 5. Assessment of Amino Acid Biosynthetic Pathway Prevalence

Depicted is the enzyme identification coverage for the glutamate (Glu), glutamine (Gln), arginine (Arg), branched chain (Leu, Ile, Val), and aromatic amino acid biosynthetic pathways. Pathway coverage is defined as the ratio between the number of identified enzymes and all enzymes in the pathway. Information about the enzymes active in each pathway was retrieved from the TAIR (http://arabidopsis.org/tools/aracyc) and the KEGG (http://www.kegg.com) databases. Only enzymes with a chloroplast transit peptide were taken into account.

tolase and ribose 5-phosphate isomerase, both of which were identified under our experimental conditions (Figure 4B).

Coverage of enzymes from amino acid synthesis pathways was less complete (only 23%, Table 2). Since glutamate and glutamine synthesis occur in the light while synthesis of aromatic amino acids is downregulated in the light [29], we investigated pathway coverage (i.e., the ratio between the number of identified enzymes and all enzymes in the pathway) for these metabolic pathways in more detail (Figure 5). All enzymes from the glutamate and glutamine synthesis pathways (100% pathway coverage) were identified suggesting that they are most active during photosynthesis (Figure 5). Several enzymes from the branched chain amino acid and arginine synthetic pathways were identified, but pathway coverage was less complete (below 50%, Figure 5). No enzyme was identified for the synthesis of aromatic amino acids. Aromatic amino acids are precursors for a large variety of secondary metabolites, many of which are involved in stress responses. Their synthesis is induced during stress and is usually not detectable under normal growth conditions [29].

Pathway prevalence as suggested by pathway coverage is also well reflected at the RNA expression levels, which are highest for enzymes in the glutamate, glutamine, and branched-chain amino acid synthesis pathways and lowest for the aromatic amino acid synthesis pathways. For a general assessment of the correlation between transcript and relative protein abundance, we calculated the Spearman rank correlation for important chloroplast metabolic functions (Table 2, last column). Collectively, we found a positive correlation between transcript abundance and protein detection for the amino acid synthesis pathways and the Calvin cycle (Table 2), resulting in a Spearman rank correlation of 0.72 and 0.66, respectively (Table 2 and Figure S2). The positive correlation suggests that in photosynthetically active chloroplasts, these metabolic pathways are regulated primarily at the transcriptional level. This is in contrast to the tetrapyrrole synthesis pathway, for which we could not find a strong correlation between transcript levels and relative protein abundance (SpC, −0.25). Only a weak correlation was found for nucleus-encoded proteins of photosystems I and II (SpC, 0.42) (Table 2 and Figure S2). Since tetrapyrrole synthesis is tightly coordinated with the synthesis of photosystem light-harvesting proteins, our results support a view that this coordination is largely controlled by posttranscriptional processes, which are not fully understood at present (reviewed in [30]). Similarly, our proteome analysis identified only 25% of the known proteins involved in the regulation of chloroplast gene expression (Table 2). Together, these results may reflect the growth of the plants at a light condition (100 μE) that reduces photodamage-induced repair of photosystem II (reviewed in [31]), and the time of harvest (2 hr after the onset of illumination), which is most likely before the chloroplasts have reached their full photosynthetic potential.

Discussion

Prediction of Pathway Prevalence from Combined MS/MS Shotgun Proteomics and RNA Expression Analysis

The functional categorization of the proteins identified in our MS/MS shotgun chloroplast proteome analysis suggests that we have found enzymes and other proteins for nearly all known chloroplast photosynthetic complexes as well as for metabolic and regulatory pathways (Figure 4 and Tables S3–S5). For most of the metabolic pathways, however, we identified only a limited number of their enzymes, although the activity of the pathways has been demonstrated in different plastid types. In combination with RNA expression profiling, the information compiled in the list of identified proteins and their distribution over metabolic pathways therefore provides useful information about the prevalence of specific metabolic pathways in chloroplasts under specific experimental conditions (Table 2). We believe that combining RNA expression profiling and MS/MS shotgun proteomics can contribute to new insights into regulatory levels of gene expression on a large scale, as we have demonstrated for five essential chloroplast functions (Table 2). Analysis of amino acid synthesis in yeast already showed that metabolic pathways were regulated in a concerted fashion at the transcriptional level during adaptation from rich to minimal growth media [32]. Our results suggest that expression of amino acid synthesis and Calvin cycle enzymes is largely regulated at the transcriptional level as well under our experimental conditions (Table 2 and Figure S2). The correlation between transcript levels and protein detection does

not extend to all chloroplast functions, as was the case for tetrapyrrole synthesis and photosystem complexes. Our results are consistent, therefore, with reports that suggest a complex regulation of enzymes in the tetrapyrrole pathway that involves both transcriptional and post-transcriptional mechanisms ([33], reviewed in [30]).

MDC-MS/MS Shotgun Proteomics and Structure Predictions Facilitate Functional Annotation of Chloroplast Proteins

The analysis of the chloroplast proteome reported here has produced one of the most comprehensive protein lists for a cell organelle available to date. Our strategy focused on the identification of as many *Arabidopsis* chloroplast proteins as possible without fractionating the organelle into specific subproteomes. A comparison of our strategy with recent reports of proteins identified from isolated thylakoid lumen fractions [17, 18] and envelope membranes [19, 20] shows that our MDC-coupled MS/MS shotgun proteomics is equally sensitive and results in a similar coverage of identified protein. Two of these reports suggested, however, that specific enrichment and extraction of hydrophobic membrane proteins from chloroplast envelope membranes were necessary to significantly improve the identification of membrane proteins and novel transporters in the chloroplast envelope [19, 20]. Although our strategy did not identify all of the published envelope proteins, we found several new proteins in the envelope fraction that are potential transporters located in the outer envelope membrane. Among these are two ABC transporters (At1g15210 and At1g59870, Table S1) that have also been predicted in silico to localize to the envelope membrane [26]. Together, our MDC-coupled MS/MS shotgun protein identification strategy significantly expanded the number of proteins identified from chloroplasts. In combination with bioinformatics analysis and structure prediction tools (see Results and Supplemental Experimental Procedures), we were able to develop a more informative annotation for "hypothetical" proteins in relevant databases. Although our proposed annotations are still preliminary, they demonstrate that extraction of protein sequences by MDC-coupled MS/MS shotgun proteomics coupled with structure- and function-prediction tools is a useful and efficient approach to facilitate and improve functional annotations.

Chloroplast Proteome Analysis Offers New Insights into Organelle Evolution

Organelle proteome analysis offers the opportunity to define the protein complement of entire cell compartments and to improve prediction tools for protein localization. The identification of a large number of proteins establishes an important basis to gain new insights into protein sorting and import, as well as the apparent complexity of the proteome. For example, after subtracting from our list of proteins all proteins identified from the envelope membrane fraction (some envelope proteins do not contain a predictable transit peptide), plastid-encoded proteins, and all potential contaminants (Table S2A), only 67% of the remaining proteins are predicted to localize to the chloroplast by available prediction tools [3] (http://www.cbs.dtu.dk/services/TargetP/). With the current prediction sensitivity of TargetP, we now estimate that the *Arabidopsis* plastid proteome is comprised of approximately 3800 proteins. This number represents a theoretical proteome, which includes proteins of all plastid differentiation states (e.g., etioplast, amyloplast, chromoplast, etc.). The chloroplast proteome would comprise only a fraction of these 3800 proteins that reflect chloroplast-specific functions. A comparison of the chloroplast proteins reported here with the most abundant proteins from undifferentiated tobacco BY-2 plastids revealed only approximately 50% overlap, which supports this view (S.B., A. Siddique, and W.G., in preparation). Clearly, the complete plastid proteome can only be revealed through analysis of all plastid differentiation states under different environmental and experimental conditions. Given the differentiation capacity of higher plant plastids, it is currently difficult to estimate the percentage that our protein identification represents of the chloroplast proteome. In theory, our MDC-MS/MS shotgun proteomics approach should allow for the detection of proteins that are present at 42 copies per chloroplast, assuming a detection limit of 20 fmol with 2.9×10^8 chloroplasts in each experiment. Proteins present at a lower copy number would be difficult to detect, although the threshold for each protein depends on several parameters, including intrinsic peptide characteristics and complexity of the protein sample.

From an evolutionary perspective, the complete chloroplast proteome could also provide new insights into the extent of gene transfer between the endosymbiont (organelle) and host cell. A GenBank search with all identified proteins not predicted by TargetP to localize to the chloroplast revealed already that 7.4% of their genes have homologs in *Synechocystis*, but not in yeast, suggesting that during evolution these genes were transferred from the cyanobacterial endosymbiont genome to the nuclear genome of the plant cell. This percentage is significantly higher than would be expected from a random search of cyanobacterial gene homologs in the *Arabidopsis* genome (approximately 4%) and reinforces our view that many more proteins without a predicted transit peptide are imported into the chloroplast, possibly via novel import mechanisms.

Conclusions

We report here the most comprehensive list of proteins identified from chloroplasts to date. A full description of the chloroplast proteome is important to understand chloroplast biogenesis and metabolism. The identification of proteins that cannot be predicted in silico to localize to the chloroplast supports the view that proteomics with cell organelles is perhaps the only reliable way to provide information about protein sorting and pathway compartmentalization on a large scale. When combined with metabolite profiling, the proteome and RNA expression analysis reported here can be developed into a valuable systems approach strategy to gain new insights into the metabolic and regulatory networks that control plastid and ultimately plant cellular functions.

Experimental Procedures

Chloroplast Isolation

Chloroplasts were isolated from seven-week-old short day-grown (8 hr light/16 hr dark) *Arabidopsis thaliana* plants and purified by three consecutive steps of Percoll density gradient centrifugation. The purity of the chloroplast preparations was established by fluorescence microscopy, immunological detection of marker proteins, and diagnostic enzyme activity assays (fumarase, catalase).

Protein Fractionation

Chloroplast proteins were fractionated according to their different physico-chemical properties by using a combination of serial solubilization followed by ion exchange or Cibacron Blue Sepharose chromatography (Figure S1). Chloroplast envelope membranes were enriched as previously described [21]. All protein fractions were subjected to SDS-PAGE; for each protein fraction the gel was cut into ten fractions, and each gel segment was subjected to "in gel tryptic digest" as described [22]. Tryptic peptides were further fractionated by reversed phase chromatography coupled online to an ion trap mass spectrometer (LCQ DecaXP).

Mass Spectrometry and Bioinformatics Analyses

Scans were defined as data-dependent scans with one full scan followed by four MS/MS scans of the highest intensity ions. Peptides were identified from a combined database containing all *Arabidopsis*, chloroplast, and mitochondrial proteins by using the SEQUEST search program followed by a visual examination of MS/MS spectra. (For a detailed description of the analysis and the interpretation of MS/MS data see the Supplemental Experimental Procedures.) Identified proteins were subjected to bioinformatics analyses as described in the Supplemental Experimental Procedures by using software tools available at http://mips.gsf.de/, http://www.tigr.org/, http://www.ncbi.nlm.nih.gov/, http://www.cbs.dtu.dk/services/TargetP/, http://www.sanger.ac.uk/Software/Pfam/index.shtml/, http://phylogenomics.berkeley.edu/resources/, http://scop.berkeley.edu/, http://www.rcsb.org/pdb/, http://www.kegg.com/, and http://arabidopsis.org/tools/aracyc/.

GeneChip RNA Expression Analysis

For mRNA-expression analysis total RNA was extracted in triplicates from 100 mg leaf material (same growth conditions and developmental age as for proteome analysis) by using Trizol (Life Technologies, USA) and further purified by using Qiagen RNeasy kit (Qiagen, Germany) according to the manufacturer instructions. Probe preparation, hybridization to the full genome *Arabidopsis* Affymetrix GeneChip, washing, staining, and scanning were carried out according to the manufacturer instructions (Affymetrix, Santa Clara). Raw data were processed with Affymetrix MicroarraySuite 5.0.

Supplemental Data

Supplemental Data including Experimental Procedures, two figures, and five tables are available at http://www.current-biology.com/cgi/content/full/14/5/354/DC1/.

Acknowledgments

We would like to thank Matthias Hirsch-Hoffmann for constructing the database web interface and the Functional Genomics Center Zurich for technical support of the project. The research was supported by funds from the ETH Zurich and SEP Life Sciences to W.G. A.v.Z is supported by a fellowship from the VELUX Foundation, and research in the laboratory of K.S. at University of California, Berkeley, is supported by NSF award 0238311.

Received: November 30, 2003
Revised: January 19, 2004
Accepted: January 19, 2004
Published: March 9, 2004

References

1. Martin, W., and Herrmann, R.G. (1998). Gene transfer from organelles to the nucleus: how much, what happens, and why? Plant Physiol. *118*, 9–17.
2. 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.
3. 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.
4. Abdallah, F., Salamini, F., and Leister, D. (2000). A prediction of the size and evolutionary origin of the proteome of chloroplasts of Arabidopsis. Trends Plant Sci. *5*, 141–142.
5. Martin, W., Rujan, T., Richly, E., Hansen, A., Cornelsen, S., Lins, T., Leister, D., Stoebe, B., Hasegawa, M., and Penny, D. (2002). Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc. Natl. Acad. Sci. USA *99*, 12246–12251.
6. Leister, D. (2003). Chloroplast research in the genomic age. Trends Genet. *19*, 47–56.
7. Pfanner, N., Hoeben, P., Tropschug, M., and Neupert, W. (1987). The carboxyl-terminal two-thirds of the ADP/ATP carrier polypeptide contains sufficient information to direct translocation into mitochondria. J. Biol. Chem. *262*, 14851–14854.
8. Folsch, H., Guiard, B., Neupert, W., and Stuart, R.A. (1996). Internal targeting signal of the BCS1 protein: a novel mechanism of import into mitochondria. EMBO J. *15*, 479–487.
9. Miras, S., Salvi, D., Ferro, M., Grunwald, D., Garin, J., Joyard, J., and Rolland, N. (2002). Non-canonical transit peptide for import into the chloroplast. J. Biol. Chem. *277*, 47770–47778.
10. Kruft, V., Eubel, H., Jansch, L., Werhahn, W., and Braun, H.-P. (2001). Proteomic approach to identify novel mitochondrial proteins in Arabidopsis. Plant Physiol. *127*, 1694–1710.
11. Millar, A.H., Sweetlove, L.J., Giege, P., and Leaver, C.-J. (2001). Analysis of the Arabidopsis mitochondrial proteome. Plant Physiol. *127*, 1711–1727.
12. Bardel, J., Louwagie, M., Jaquinod, M., Jourdain, A., Luche, S., Rabilloud, T., Macherel, D., Garin, J., and Bourguignon, J. (2002). A survey of the plant mitochondrial proteome in relation to development. Proteomics *2*, 880–898.
13. Heazlewood, J.L., Howell, K.A., Whelan, J., and Millar, A.H. (2003). Towards an analysis of the rice mitochondrial proteome. Plant Physiol. *132*, 230–242.
14. Fukao, Y., Hayashi, M., and Nishimura, M. (2002). Proteomic analysis of leaf peroxisomal proteins in greening cotyledons of *Arabidopsis thaliana*. Plant Cell Physiol. *43*, 689–696.
15. Andon, N.-L., Hollingworth, S., Koller, A., Greenland, A.-J., Yates, J.-R., 3rd, and Haynes, P.-A. (2002). Proteomic characterization of wheat amyloplasts using identification of proteins by tandem mass spectrometry. Proteomics *2*, 1156–1168.
16. 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.
17. Peltier, J.B., Emanuelsson, O., Kalume, D.E., Ytterberg, J., Friso, G., Rudella, A., Liberles, D.A., Soderberg, L., Roepstorff, P., von Heijne, G., et al. (2002). Central functions of the lumenal and peripheral thylakoid proteome of Arabidopsis determined by experimentation and genome-wide prediction. Plant Cell *14*, 211–236.
18. Schubert, M., Petersson, U.-A., Haas, B.-J., Funk, C., Schroder, W.-P., and Kieselbach, T. (2002). Proteome map of the chloroplast lumen of *Arabidopsis thaliana*. J. Biol. Chem. *277*, 8354–8365.
19. Ferro, M., Salvi, D., Riviere-Rolland, H., Vermat, T., Seigneurin-Berny, D., Grunwald, D., Garin, J., Joyard, J., and Rolland, N. (2002). Integral membrane proteins of the chloroplast envelope: identification and subcellular localization of new transporters. Proc. Natl. Acad. Sci. USA *99*, 11487–11492.

20. Ferro, M., Salvi, D., Brugiere, S., Miras, S., Kowalski, S., Louwagie, M., Garin, J., Joyard, J., and Rolland, N. (2003). Proteomics of the chloroplast envelope membranes from *Arabidopsis thaliana*. Mol. Cell. Proteomics *5*, 325–345.
21. Douce, R., and Joyard, J. (1982). Purification of the chloroplast envelope. In Methods in Chloroplast Molecular Biology, M. Edelmann, R.B. Hallick, and N.H. Chua, eds. (Amsterdam: Elsevier Biomedical Press), pp. 239–256.
22. Shevchenko, A., Wilm, M., Vorm, O., and Mann, M. (1996). Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. Anal. Chem. *68*, 850–858.
23. Peeters, N., and Small, I. (2001). Dual targeting to mitochondria and chloroplasts. Biochim. Biophys. Acta *1541*, 54–61.
24. Krömer, S. (1995). Respiration during photosynthesis. Annu. Rev. Plant Physiol. Plant Mol. Biol. *46*, 45–70.
25. Krogh, A., Brown, M., Mian, I.-S., Sjölander, K., and Haussler, D. (1994). Hidden Markov models in computational biology. Applications to protein modeling. J. Mol. Biol. *235*, 1501–1531.
26. Koo, A.-J., and Ohlrogge, J.-B. (2002). The predicted candidates of Arabidopsis plastid inner envelope membrane proteins and their expression profiles. Plant Physiol. *130*, 823–836.
27. Schnell, D.-J. (2000). Functions and origins of the chloroplast protein-import machinery. Essays Biochem. *36*, 47–59.
28. Jarvis, P., and Soll, J. (2001). Toc, Tic, and chloroplast protein import. Biochim. Biophys. Acta *1590*, 177–189.
29. Last, R., and Coruzzi, G. (2000). Amino acids. In Biochemistry and Molecular Biology Of Plants, B. Buchanan, W. Gruissem, and R. Jones, eds. (Rockville Maryland: American Society of Plant Physiologists) pp. 358–410.
30. Papenbrock, J., and Grimm, B. (2001). Regulatory network of tetrapyrrole biosynthesis–studies of intracellular signalling involved in metabolic and developmental control of plastids. Planta *213*, 667–681.
31. Aro, E.-M., Virgin, I., and Andersson, B. (1993). Photoinhibition of photosystem II. Inactivation, protein damage and turnover. Biochim. Biophys. Acta *1143*, 113–134.
32. Washburn, M.-P., Koller, A., Oshiro, G., Ulaszek, R.-R., Plouffe, D., Deciu, C., Winzeler, E., and Yates, J.-R. (2003). Protein pathway and complex clustering of correlated mRNA and protein expression analyses in *Saccharomyces cerevisiae*. Proc. Natl. Acad. Sci. USA *100*, 3107–3112.
33. Franklin, K.-A., Linley, P.-J., Montgomery, B.-L., Lagarias, J.-C., Thomas, B., Jackson, S.-D., and Terry, M.-J. (2003). Misregulation of tetrapyrrole biosynthesis in transgenic tobacco seedlings expressing mammalian biliverdin reductase. Plant J. *35*, 717–728.