ISSN 0006-2979, Biochemistry (Moscow), 2010, Vol. 75, No. 12, pp. 1470-1483. © Pleiades Publishing, Ltd., 2010.
Original Russian Text © N. B. Polyakov, D. K. Slizhikova, M. Yu. Izmalkova, N. I. Cherepanova, V. S. Kazakov, M. A. Rogova, N. A. Zhukova, D. G. Alexeev, N. A. Bazaleev, A. Yu. Skripnikov, V. M. Govorun, 2010, published in Biokhimiya, 2010, Vol. 75, No. 12, pp. 1695-1711.

Proteome Analysis of Chloroplasts from the Moss *Physcomitrella patens* (Hedw.) B.S.G.

N. B. Polyakov[1]*, D. K. Slizhikova[1], M. Yu. Izmalkova[1], N. I. Cherepanova[1], V. S. Kazakov[1], M. A. Rogova[2], N. A. Zhukova[2], D. G. Alexeev[1], N. A. Bazaleev[2], A. Yu. Skripnikov[1,3], and V. M. Govorun[1,2]

[1]*Shemyakin and Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, ul. Miklukho-Maklaya 16/10, 117997 Moscow, Russia; fax: (495) 336-0777; E-mail: polyakovnb@gmail.com*
[2]*Scientific Research Institute of Physicochemical Medicine, ul. Malaya Pirogovskaya 1a, 119435 Moscow, Russia; fax: (495) 246-4401; E-mail: govorun@hotmail.ru*
[3]*Biological Faculty, Lomonosov Moscow State University, 119991 Moscow, Russia; fax: (495) 939-1268; E-mail: a.skripnikov@gmail.com*

Received April 22, 2010
Revision received June 17, 2010

Abstract—Intact chloroplasts were prepared from protoplasts of the moss *Physcomitrella patens* according to an especially developed method. They were additionally separated into stroma and thylakoid fractions. The proteomes of intact plastids, stroma, and thylakoids were analyzed by 1D-electrophoresis under denaturing conditions followed by protein digestion and nano-LC-ESI-MS/MS of tryptic peptides from gel bands. A total of 624 unique proteins were identified, 434 of which were annotated as chloroplast resident proteins. The majority of proteins belonged to a photosynthetic group (21.3%) and to the group of proteins implicated in protein degradation, posttranslational modification, folding, and import (20.6%). Among proteins assigned to chloroplasts, the following groups are prominent combining proteins implicated in metabolism of: amino acids (6.9%), nucleotides (2.5%), lipids (2.2%), carbohydrates (2.4%), hormones (1.5%), isoprenoids (1.25%), vitamins and cofactors (1%), sulfur (1.25%), and nitrogen (1%); as well as proteins involved in the pentose-phosphate cycle (1.75%), tetrapyrrole synthesis (3.7%), and redox processes (3.6%). The data can be used in physiological and photobiological studies as well as in further studies of *P. patens* chloroplast proteome including structural and functional specifics of plant protein localization in organelles.

DOI: 10.1134/S0006297910120084

Key words: proteome, chloroplast, *Physcomitrella patens*, photosynthesis, subcellular localization

Chloroplasts of higher plants fulfill a series of significant functions associated with photosynthesis, carbon fixation, and assimilation of nitrogen and sulfur. They are implicated in synthesis of amino acids, nucleotides, lipids, fatty acids, and phytohormones, as well as secondary metabolites such as alkaloids and isoprenoids.

The chloroplast genome of the moss *Physcomitrella patens* only encodes 83 proteins [1], whereas other genes encoding proteins of this organelle are localized in the nucleus. The products of these genes are synthesized in cytosol and undergo posttranslational transport into chloroplasts using Tic/Toc translocases localized in the chloroplast outer membrane [2]. With the exception of several resident proteins of the chloroplast outer membrane, the sorting of proteins in chloroplasts depends on specific cleaved *N*-terminal presequence or leader (transit) peptide (cTP, chloroplast transit peptide) [3].

The determination of subcellular localization of proteins encoded by the nuclear genome is a challenge to biochemistry and cell biology that requires new-generation technologies, particularly various proteome investigation approaches. Chloroplast proteomes are most effectively and comprehensively studied using model

Abbreviations: ATP, ambiguous targeting predictor (calculation method for determination of proteins with dual localization); CID, collision-induced dissociation; DB, databases; DDA, data dependent acquisition; 1D-PAGE, one-dimensional polyacrylamide gel electrophoresis; DTT, dithiothreitol; LC-ESI-MS/MS, liquid chromatography electrospray ionization tandem mass spectrometry; SPI, scored peak intensity (ratio between intensity of significant peaks determining peptide structure and the overall intensity of peaks in MS/MS spectrum).

* To whom correspondence should be addressed.

objects, for example *Arabidopsis thaliana*, the genome databases of which are well-annotated and for which a significant pool of data exists concerning subcellular localization of distinct proteins [4-13]. A newer model is the moss *P. patens*, the importance of which in plant biology has increased dramatically after determination of the full nucleotide sequence of its genome. Bryophytes are interesting in terms of evolution since they were among the first terrestrial plants. During colonization, they have acquired and still retain unique mechanisms of resistance and ability for growth under extreme conditions of humidity, illumination, and temperature. Identification of biochemical mechanisms underlying resistance of bryophytes to adverse environmental factors is also interesting in the applied bioengineering aspect. Earlier, we developed a methodological platform for proteome analysis of *P. patens* gametophytes and led the way to moss organelle proteomics in the study on proteome of protoplasts isolated from the protonema [14]. Moss protoplasts were used in further studies for isolation of chloroplasts, mitochondria, and nuclei.

In this work we have performed a proteome analysis of moss chloroplasts using nondenaturing 1D-electrophoresis followed by determination of tryptic peptides by nano-LC-ESI-MS/MS and assigned subcellular localization to the identified proteins.

MATERIALS AND METHODS

Growth of moss protonema and isolation of protoplasts. Protonema of the moss *Physcomitrella patens* (Hedw.) B.S.G. strain Gransden was grown as described by Skripnikov and associates [14]. Protoplasts were isolated by the method described in the same work with slight modifications.

Isolation of chloroplasts from moss protoplasts. All isolation steps were conducted at 4°C. Protoplasts were resuspended in buffer A (50 mM HEPES-KOH, pH 7.5, 330 mM sorbitol, 2 mM EDTA, and 0.4 mM phenylmethylsulfonyl fluoride) and filtered through a double layer of Miracloth (Calbiochem-Behring Corp., USA). Degree of protoplast disintegration was estimated by light microscopy. The filtrate was centrifuged at 1200*g* for 3 min in 50-ml plastic tubes (Falcon) using a bucket rotor. The pellet was resuspended in a small volume of buffer A and fractionated by centrifugation on a centrifuge with a bucket rotor at 3800*g* for 10 min in a 40% + 85% Percoll (Sigma, USA) stepwise cocktail preformed in 15-ml plastic tubes (Falcon). Intact chloroplasts were collected from the border between 40 and 85% Percoll, washed with buffer A, resuspended in a small volume of buffer A, and fractionated again in the 40% + 85% Percoll stepwise cocktail under the same conditions. Intact chloroplasts were collected from the border between 40 and 85% Percoll, washed with buffer A, and centrifuged at 1200*g* for 3 min in 15-ml plastic tubes (Falcon) in a bucket rotor.

Plastids were sub-fractionated into thylakoids and stroma as described elsewhere [15].

Disintegration of specimen and extraction of proteins from plastids. All procedures were conducted at 4°C. The pellet of thylakoids was thoroughly resuspended in cold 80% acetone and centrifuged at 11,000*g* for 5 min in 2-ml Eppendorf tubes placed in a fixed-angle rotor. The pellet was washed two times with cold 80% acetone, each time followed by centrifugation under the same conditions, and dried in a SpeedVac vacuum concentrator.

1D-SDS-PAGE of proteins. Proteins (30% T, 2.67% C) were separated by one-dimensional electrophoresis of in a 5-20% polyacrylamide gel gradient with SDS by the standard method of Laemmli [16]. Protein bands were visualized by staining with Coomassie G-250.

Tryptic hydrolysis of proteins. Proteins separated by 1D-SDS-PAGE were hydrolyzed directly within gel fragments with trypsin using a slightly modified protocol of Shevchenko et al. [17].

Liquid chromatography electrospray ionization tandem mass spectrometry (LC-ESI-MS/MS). Tryptic peptides were analyzed on an Agilent HCT-Ultra mass spectrometer (Agilent Technologies, USA) equipped with an integrated Agilent Chip Cube™ chromatographic separation system and a quadrupole ion trap mass analyzer. This device was coupled with an Agilent 1200 sampler and nanochromatograph. Proteins were separated at a flow rate of 300 nl/min using a stepwise gradient of 0.1% formic acid in 5% acetonitrile (solvent A) and 0.1% formic acid in 90% acetonitrile (solvent B) as follows: 0% B, 2 min; 15% B, 5 min; 20% B, 20 min; 50% B, 55 min; 90% B, 60-65 min; and 100% B, 66 min.

Peptide masses were determined using the quadrupole ion trap at m/z ratios ranging from 300 through 2200 with the "trap optimization mass" of 900. Only positively charged ions were detected. Peaks were selected for following MS/MS analysis in the DDA mode. Three ions preceding maximum intensity were subjected to fragmentation, one after another by collision with helium atoms (CID mode). Only the ions with $z \geq 2$ and peak intensity above a threshold were analyzed in MS/MS experiments.

MS data processing. Peak lists were obtained using the Spectrum Mill Data Extractor software with the following parameters: combining scans of the same precursor ion at ±1.4 m/z and time slot ±15 sec. The signal-to-noise ratio for a precursor ion was 25 and maximal charge number was 7. Generated lists of ion masses (raw files) were submitted to the Spectrum Mill search engine (Rev A.03.03.084 SR4), as well as to Phenyx [18] (v.2.6, local version) and Mascot (local version 2.1.03). Peptide identities were searched in the UniProt database (http://uniprot.org) with *P. patens* taxon retrieval (the number of entries was 35,415). To exclude improper identities, the database was supplemented with sequences

of possible admixture proteins, such as trypsin, human keratins, and bovine serum albumin. The following parameters were used for the search: cleavage specificity – trypsin; maximum of missing cleavage sites – 2; error of ion mass determination ±2.7 Da (parental ion) and ±0.7 Da (daughter ions); possible protein modifications – methionine oxidation and cysteine carbamidomethylation.

The dataset of determined proteins was validated using built-in tools of the Spectrum Mill retrieval system in protein details mode with the following parameters depending on the charge number of parental ions (in the order: ***ion charge***, *peptide score threshold*, *SPI (scored peak intensity) threshold value*, *%*, *score difference between forward and reverse DBs for the best found peptide*, *minimal difference between score values of the first and second candidates)*: **1**,6,70,2,2; **2**,6,60,2,2; **2**,6,90,1,1; **3**,8,70,2,2; **4**,8,70,2,2; and **1**,6,90,1,1.

Apart from the above autovalidation procedure, the second validation step of distinct peptides was made in peptide mode with the following parameters: ***ion charge*** 1, 3, or 4; *peptide score threshold* – 13, *SPI* – 70%; ***ion charge*** 2, *peptide score threshold* – 11, *SPI* – 60%; the score difference between forward and reverse DBs for the best found peptide and minimal difference between score values of the first and second candidates was two or more.

We used three search engines to reveal possible semitryptic peptides (Spectrum Mill, Phenyx, and Mascot). The Spectrum Mill database contained 236 protein sequences determined from the single unique tryptic peptide (Table 1, list C; see Supplement on http://protein.bio.msu.ru/biokhimiya), and the same parameters were used for the search, but cleavage specificity; the following cleavage rule was applied: one of the terminal amino acid residues should be either K or R, with any another terminal residue. The peptide list generated by the Spectrum Mill search engine was validated as described above; the validity criterion for datasets generated by other search systems was exceeding of peptide score threshold (95%).

Subcellular and functional annotation of proteome (bioinformatics methods). *Protein sequence databases used in this work*. The main database we used in this work was UniProt (http://uniprot.org) with *P. patens* taxon retrieval (35,415 entries). The JGI (US Department of Energetics Joint Genome Institute) database (version 1.1, March 2007) containing the *P. patens* nuclear genome sequence (35,938 entries) was used as an accessory [19].

Proteins of the moss *P. patens* demonstrating high homology with *A. thaliana* protein sequences (database TAIR9, version from June 19, 2006, 33,410 entries) were identified using BLASTP (version 2.2.20, local 32-bit version) [20]. Proteins were considered as homologous at *e*-value < 0.001 and score > 50.

Amino acid sequences of proteins encoded in genomes of *Nostoc* sp. PCC 7120, *Synechocystis* sp. PCC 6803, and *Thermosynechocystis elongatus* BP-1 were acquired from the NCBI (US National Center for Biotechnology Information) (http://www.ncbi.nlm.nih.gov) database and combined into the reference database of cyanobacterial protein sequences containing 300,676 entries.

Amino acid sequences of cyanobacteria homologous to those of the moss *P. patens* were revealed as described above for *A. thaliana* with the same *e*-value and score values.

Computational methods for subcellular localization of P. patens proteins. The moss protein sequences acquired from the UniProt database as described above were analyzed using the programs TargetP [21], Predotar [22], and Proteome Analyst 3.0 [23].

RESULTS

The method for isolation of chloroplasts was developed on the basis of protocols of van Wijk et al. [15] and Kabeya and Sato [24] for isolation of *P. patens* plastids. The method developed in our laboratory is most appropriate because it allows preparation of specimens containing more than 60% chloroplast resident proteins.

We carried out four independent experiments for identification of proteins from intact chloroplasts, stroma, and thylakoids of *P. patens* protonema. In each

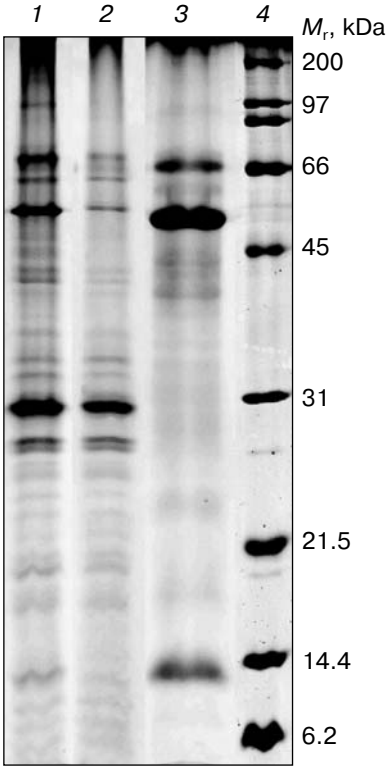

Fig. 1. 1D-electrophoresis of proteins from intact chloroplasts (*1*), thylakoids (*2*), and stroma (*3*); *4*) molecular mass markers.

experiment, intact plastids were separated into soluble (stroma) and membrane (thylakoid) fractions that were then subjected to 1D-electrophoresis under denaturing conditions in the presence of SDS (Fig. 1). The gel was cut into slices, about 1 mm in width, and treated with trypsin. Tryptic peptides were extracted and analyzed by LC-ESI-MS/MS. This approach revealed 624 unique proteins, 401 of which (64%) were identified by two and more peptides per protein using the Spectrum Mill search engine (Table 1, list B; see Supplement at http://protein.bio.msu.ru/biokhimiya). To increase identification reliability of proteins identified by one tryptic peptide validated using the Spectrum Mill program complex, we performed two additional experiments showing greater number of unique peptides belonging to the studied protein set.

First, we carried out an independent search for tryptic peptides in the UniProt database with *P. patens* taxon retrieval using the Phenyx and Mascot search engines. This approach identified 1059 additional unique peptides belonging to 356 proteins, 85 of which were identified by a single peptide in the same gel strips by the Spectrum Mill program (Table 1, lists C and D; see Supplement at http://protein.bio.msu.ru/biokhimiya).

Since Picotti et al. [25] showed earlier that protein digestion with trypsin always results in appearance of semi-tryptic peptides that give not so prominent peaks in the mass spectrum, but can substantially exceed the tryptic peptides in total amount in the reaction mixture, we made an additional search among those mass spectra that were not identified in the standard search for tryptic peptides using the Spectrum Mill system.

The only goal of the search was validation of proteins identified by a single tryptic peptide, so we created a database containing only similar protein identities (totally 236 protein sequences) in which the search for semi-tryptic peptides was carried out using three search engines: Spectrum Mill, Phenyx, and Mascot. Accounting for non-tryptic peptides for validation of single-peptide identities was already used [26].

As a result, additional semi-tryptic peptides in the same gel strips were found for 207 of 236 proteins identified by a single tryptic peptide using Spectrum Mill (Table 1, lists C and D; see Supplement at http://protein.bio.msu.ru/biokhimiya). Such low level of false-positive single-peptide identifications is probably due to high stringency of the automated peptide validation criterion in this search engine. Other search algorithms (Phenyx and Mascot) also allowed identification of unique semi-tryptic peptides of an additional 17 proteins. To estimate probability of false-positive identifications in the search for semi-tryptic peptides, we searched in a database containing randomly chosen 500 protein sequences from the UniProt database that were not identified in this work. Two or more certain semi-tryptic peptides were found for only 15 proteins, so the probability of false-positive identifications was 3% and proteins identified with the above-described algorithm surely fall within 95% confidence interval (data are shown in the Table 1, lists C and D; see Supplement at http://protein.bio.msu.ru/biokhimiya).

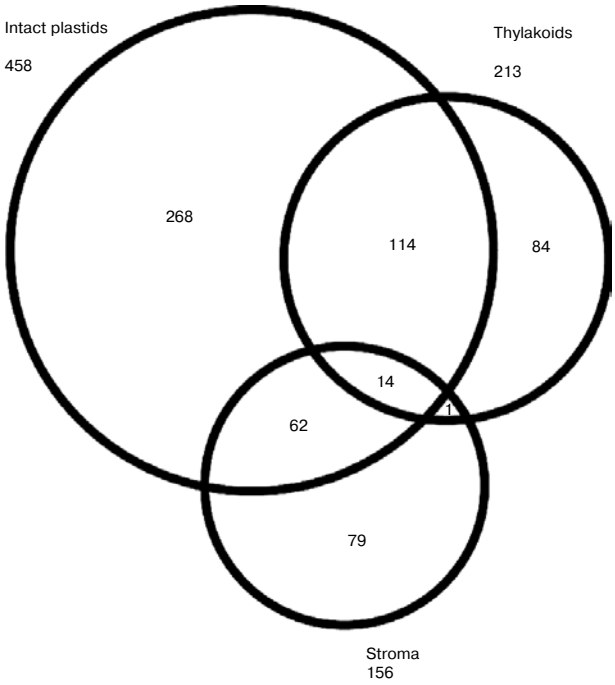

Fig. 2. Distribution of identified proteins from different fractions of chloroplasts (Venn diagram).

The distribution of proteins from different fractions of chloroplasts is shown in Fig. 2. One can see that our strategy of chloroplast subdivision into stroma and thylakoids has allowed not only identification of a series of proteins that we could not identify from analysis of intact chloroplasts, but also acquisition of data on their suborganelle localization.

Physical and chemical characteristics of proteins identified in this work. The proteins identified in fractions of intact chloroplasts, thylakoids, and stroma were characterized by indices of isoelectric point and molecular mass. Also, the hydrophobicity index was calculated for each subproteome by the GRAVY (Grand average of hydropathy) algorithm as described earlier [27].

Molecular masses of proteins from different plastid fractions range from 10 through 380 kDa with the majority of proteins ranging from 20 through 60 kDa. The p*I* values vary from 4 to 12, and their distribution is distinctly bimodal (data not shown). This is typical of proteomes of all species and subproteomes of organelles, as recently shown by Kiraga et al. [28], and mode intensity often correlates with both subcellular localization of proteins and the size of the proteome. It is worth noting that the isoelectric point and molecular mass values, as well as hydrophobicity index, were calculated for

unprocessed proteins regardless of their possible post-translational modifications, although a great number of plastid proteins contain the import signal (presequence) that splits off during the import of nuclear genome-encoding proteins into chloroplasts, and many resident proteins of this organelle are modified at the posttranslational level. The GRAVY protein hydrophobicity index varied from 1.1 through 0.8. The portion of highly hydrophobic proteins (GRAVY index > 0) was typically 20-30% depending on the analyzed fraction with the predominance of such proteins in membrane fraction of thylakoids (the protein distribution versus GRAVY index is not shown).

Subcellular localization of distinct proteins. The UniProt database used as basic (see details in "Materials and Methods") contains 35,415 protein sequences with only 2257 sequences being well-assigned; other sequences are predicted proteins with gene numbers in the JGI genome database. The latter was used as an accessory.

First, we compared protein sequences acquired from the JGI genome database using the BlastP algorithm [20] and from TAIR9 database containing *A. thaliana* protein sequences (see details in "Materials and Methods"). The general scheme of annotation is shown in Fig. 3. Chosen *e*-value and score parameters allowed identification of 19,178 proteins (53%) with sufficient homology between *P. patens* and *A. thaliana*.

Annotation of this protein set in the context of their possible subcellular localization was performed using the data from the following databases: 1) PPDB (The Plant Proteome Database, http://ppdb.tc.cornell.edu/) [29] particularly containing the data on identification of chloroplast proteomes (which were initially obtained in the Laboratory of Proteomics and Biology of Chloroplasts, Cornell University, USA) and functions and subcellular localization of plant proteins; 2) SUBA (SUB-cellular location database for *Arabidopsis* proteins, http://suba.plantenergy.uwa.edu.au) [30] containing the data of proteomic and fluorescence studies of *A. thaliana* proteins; 3) eSLDB (Eukaryotic Subcellular Localization Database, http://gpcr2.biocomp.unibo.it/esldb/index.htm) [31] containing the data on subcellular localization of proteomes of five eukaryotes including *A. thaliana*.

To additionally verify subcellular localization of distinct proteins in the moss chloroplast proteome, we used in this work – apart from the data acquired from the data-

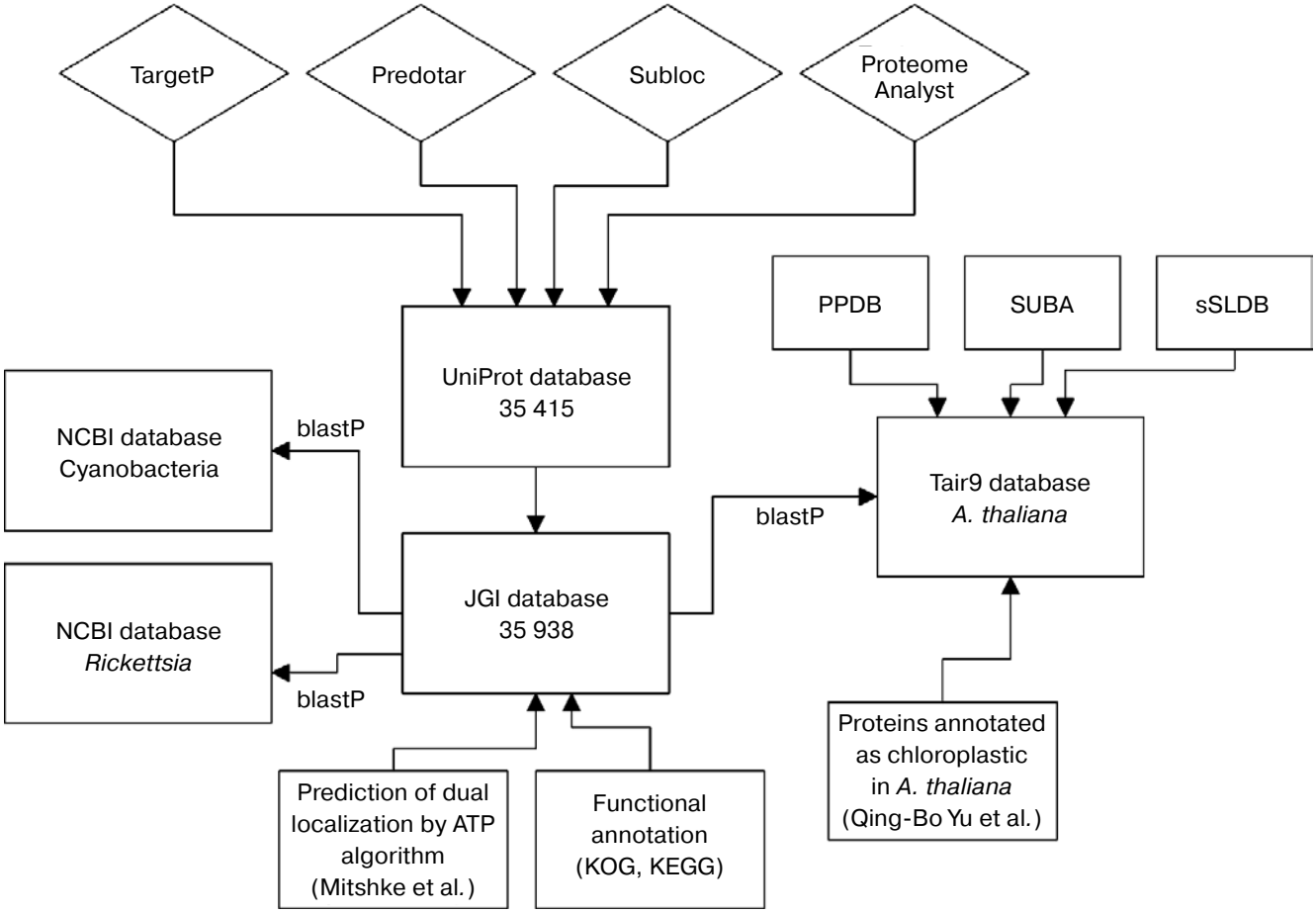

Fig. 3. General scheme of *P. patens* proteome annotation.

bases described above – a protein set identified in constructing the metabolic net of *A. thaliana* chloroplasts [32]. The authors of this work collected information from different computational algorithms allowing prediction of protein localization in the cell, data on homology with cyanobacteria, information from databases (Plprot [33], PPDB), as well as results of a series of studies on identification of chloroplast proteome, and, using a naive Bayes classifier, identified 1808 chloroplast resident proteins and 5784 proteins classified as putative chloroplast proteins.

Since many recent studies have shown possible dual localization of a series of plant proteins [34-37] containing the same leader peptide, a new computation algorithm has been developed allowing identification of such proteins [38]. We used *P. patens* as a model organism for information on predicted proteins with dual localization in mitochondria and chloroplasts and for annotation of plastid proteome proteins identified in this work. We used three prediction services – TargetP, Predotar, and Proteome analyst 3.0 – for subcellular localization of *P. patens* proteins with no homology with protein identities with *Arabidopsis*, as well as for additional verification of localization of *A. thaliana* protein homologs.

The data on possible localization of proteins acquired from the sources described above are presented in Table 2 (see Supplement at http://protein.bio.msu.ru/biokhimiya). These data were manually checked, and a definite decision on localization of each protein was made if the data of the several independent methods or literature data were unequivocally indicative of its localization in a distinct cell compartment. The diagram (Fig. 4) shows that the great majority of proteins (400 or 63%) in the plastid proteome identified in this work are chloroplast resident proteins, and 1.7% are proteins with dual localization predicted using the above-mentioned ambiguous targeting predictor (ATP) algorithm (the score value for these proteins exceeded 0.7). Plastid proteins with possible dual localization and putative chloroplast proteins (score values 0.5 and 2.4%, respectively) were assigned to separate groups. Decision on separate grouping of these proteins was made according to the data of annotation if the literature data confirmed localization of a protein in mitochondria or chloroplasts. A significant

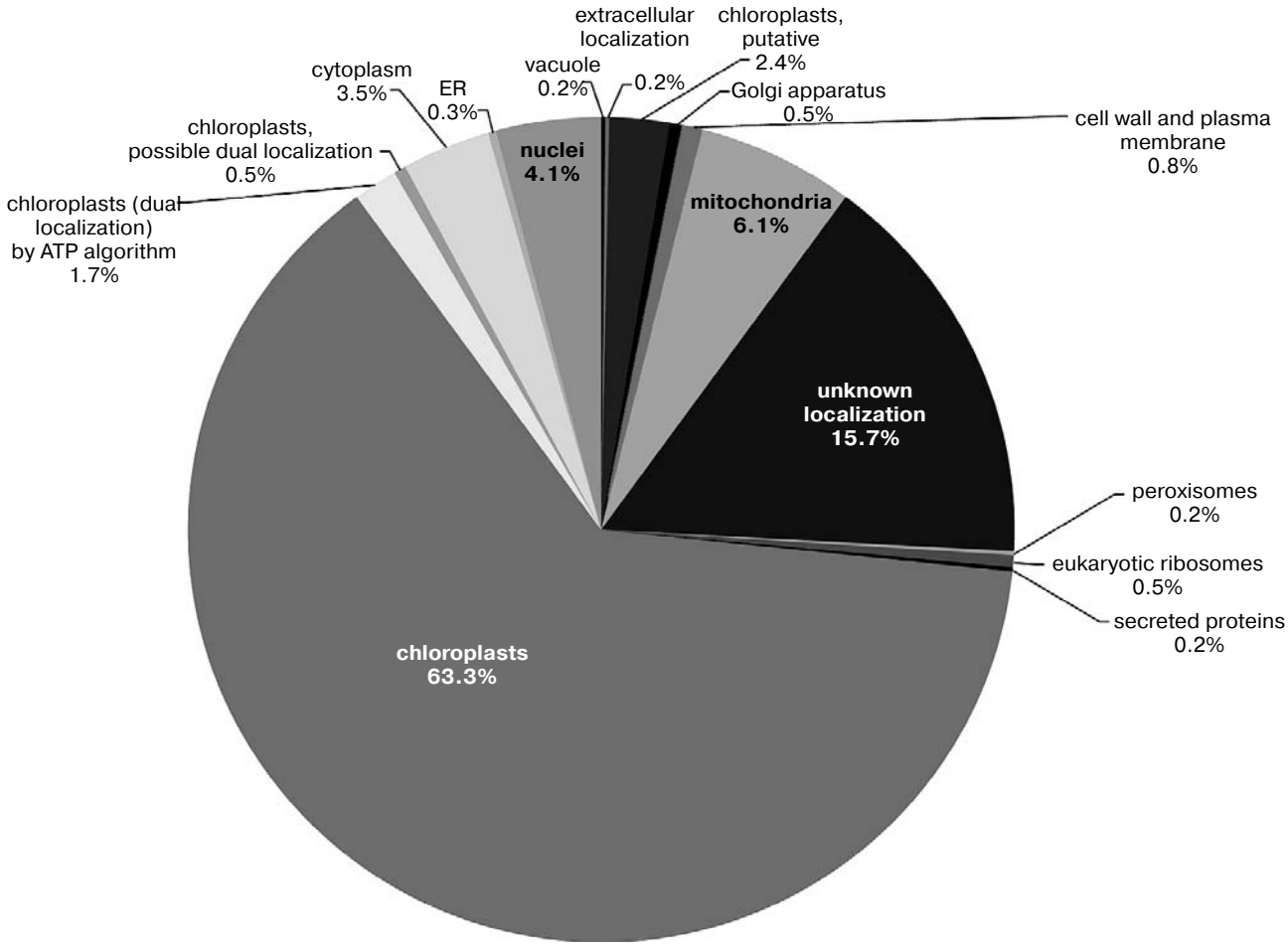

Fig. 4. Distribution of 624 identified proteins in accordance with their subcellular localization.

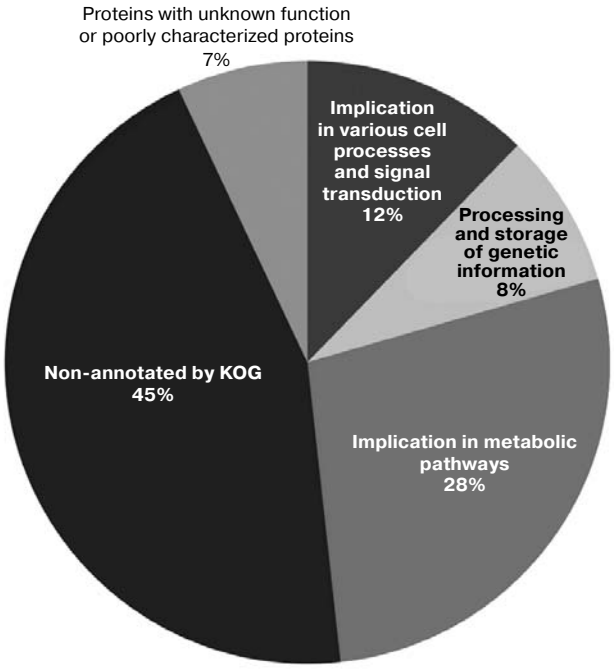

Fig. 5. Distribution of *P. patens* chloroplast resident proteins by KOG classes.

group (15.7%) contained proteins whose localization was not certainly determined because of either poor or contradictory data.

Functional annotation of identified chloroplast resident proteins. All 434 proteins (Fig. 4) annotated as either chloroplast resident proteins or proteins with dual localization or putative chloroplast proteins were classified on the basis of KOG (EuKaryotic Orthologous Groups, http://genome.jgi-psf.org/Chlre3/tutorial/kog.html) criteria [39] into four classes and 21 subclasses (Fig. 5). Thus, 55% of the total number of identified proteins was classified by KOG. Twelve percent are implicated in various cell processes and signal transduction pathways, 28% are associated with metabolism, 8% are implicated in processing and storage of genetic information, and 7% are proteins with unknown function or poorly characterized by KOG.

Among the proteins annotated in accordance with this classification (Fig. 6), 13.8% are proteins implicated in protein metabolism and 11.9% are ribosomal proteins and proteins associated with translation. The proteins implicated in metabolic processes are divided into five classes: proteins of energy metabolism (6.7%); proteins associated with metabolism and transport of amino acids (13.4%), sugars (11.5%), inorganic ions (4.7%), and

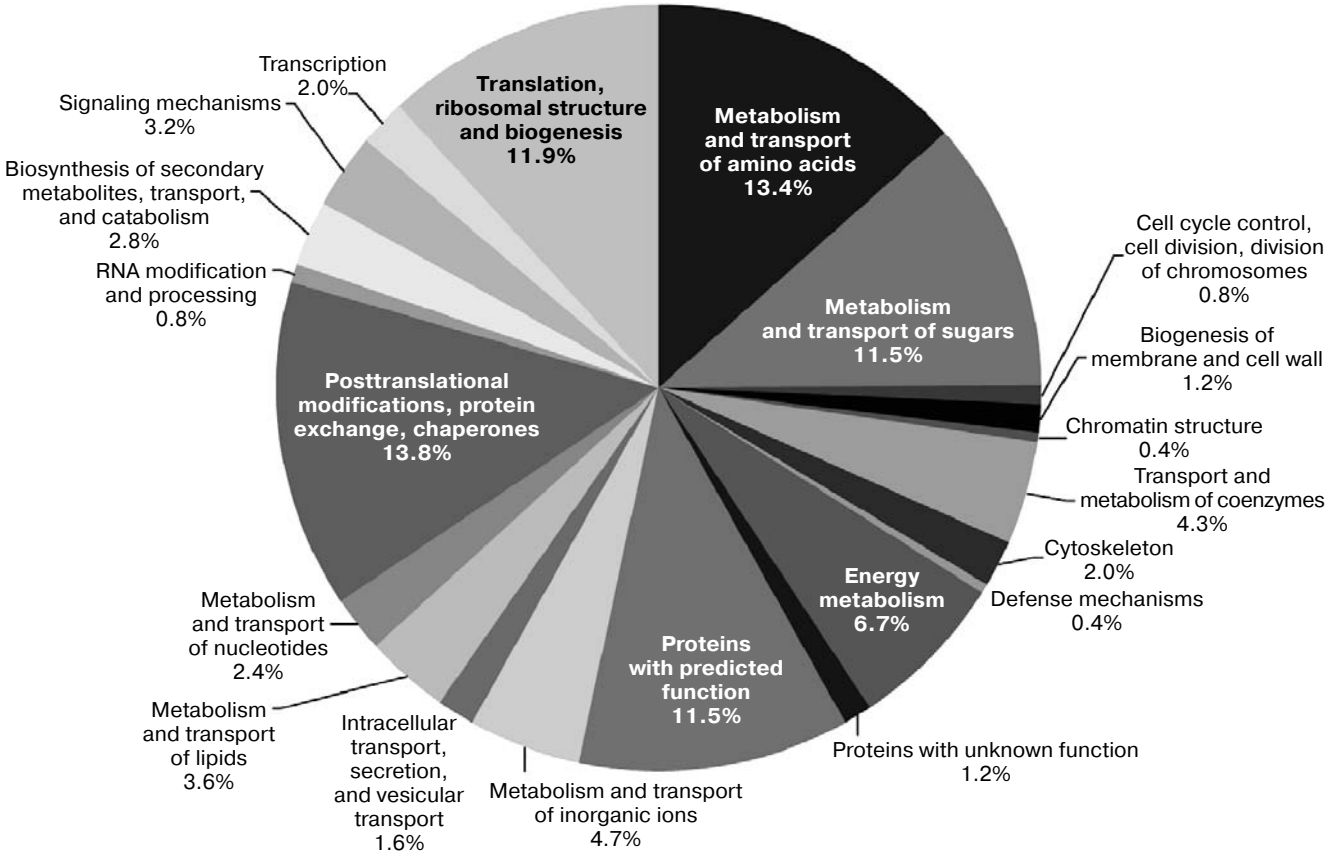

Fig. 6. Distribution of *P. patens* chloroplast resident proteins by KOG subclasses.

coenzymes (4.3%). Proteins with predicted function were classified into a separate subgroup (11.5%), and the proteins with unknown function (1.2%) have also been classified into another subgroup. Other groups represent proteins implicated in transcription (2%), metabolism and transport of lipids (3.6%), secondary metabolites (3%), nucleotides (2.4%), as well as proteins involved in RNA modification and processing (0.8%) and providing putative defense mechanisms (0.4%).

KEGG classification. Since the above classification of distinct chloroplast proteins by clusterization of orthologous eukaryotic genes has allowed annotation of only 55% of the proteins identified in this work, we classified the same protein set using the KEGG (Kyoto Encyclopedia of Genes and Genomes; http://www.genome.jp/kegg/) database.

Among 434 proteins localized in chloroplasts according to the data of this work, 159 unique proteins were attributed to metabolic pathways by KEGG. Preponderant are proteins implicated in various metabolic pathways (22%), photosynthesis (6%), biosynthesis of growth hormones (5%), phenylpropanoids (4.5%), and alkaloids (5%), as well as ribosomal proteins (11%). This classification compared to KOG allowed expansion of proteins by their implication in various pathways of biosynthesis and metabolism of amino acids (Fig. 7), sugars, terpenoids and steroids, biosynthesis of growth hormones, and in some other metabolic pathways, but this method allowed classification of only 36% of identified plastid proteins, which is insufficient for comprehensive functional characterization of the proteome.

Classification by MapMan. The set of proteins (Fig. 8) annotated in this work as either chloroplast resident proteins or proteins with dual localization or putative chloroplast proteins was classified by the MapMan criteria proposed by Thimm and associates [40]. Among the proteins annotated according to this classification, 21.3% of the proteins are implicated in photosynthesis, 20.6% are associated with translation, implicated in ubiquitination, posttranslational modification, folding, and import; some ribosomal proteins are also placed to this group. Identified proteins implicated in metabolic processes were divided into ten classes: proteins implicated in metabolism and transport of amino acids (6.9%), nucleotides (2.5%), lipids (2.2%), carbohydrates (2.4%), and hormones (1.5%), secondary metabolism of isoprenoids (1.25%), vitamins and cofactors (1%), anabolism of sulfur (1.25%), metabolism of nitrogen (1%), and C1-metabolism (0.2%).

Some proteins are implicated in the pentose-phosphate cycle (1.75%), glycolysis (1.5%), gluconeogenesis (0.5%), and the tricarboxylic acid cycle (1.25%).

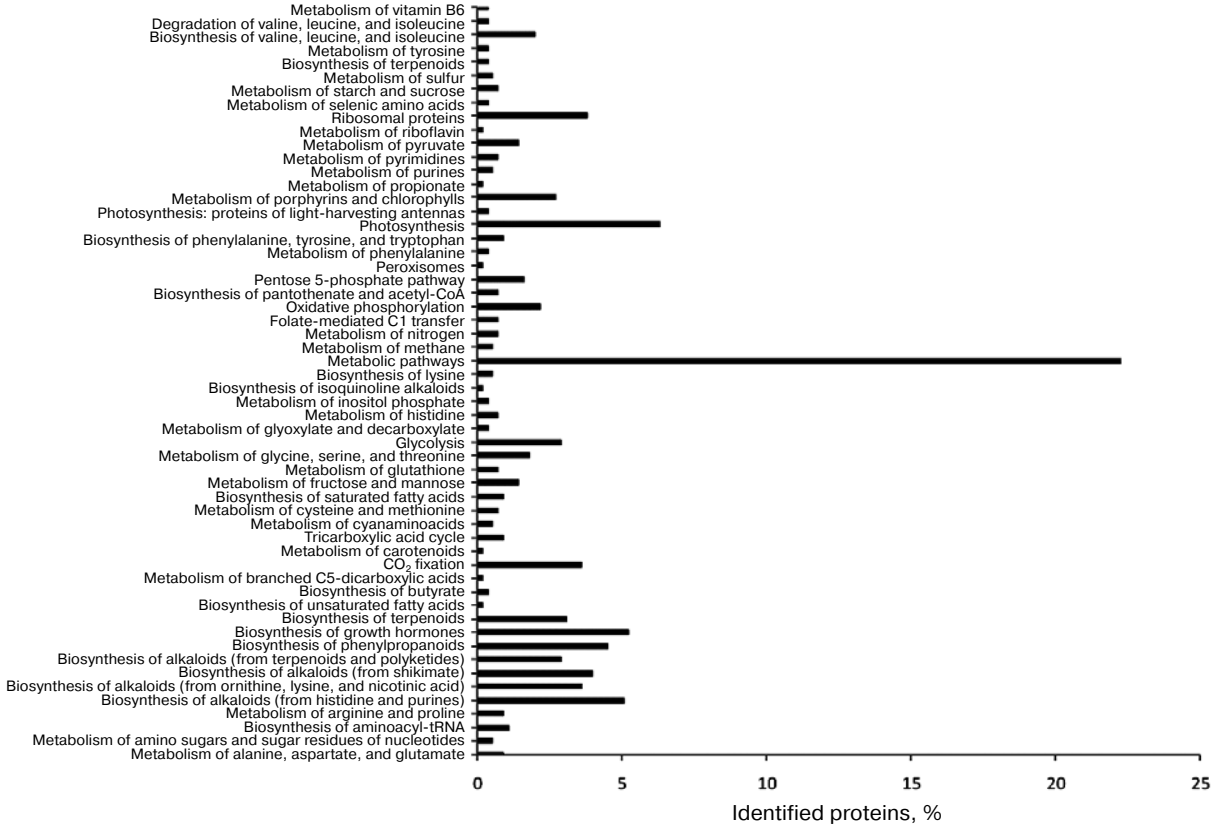

Fig. 7. Distribution of *P. patens* chloroplast resident proteins by KEGG subclasses.

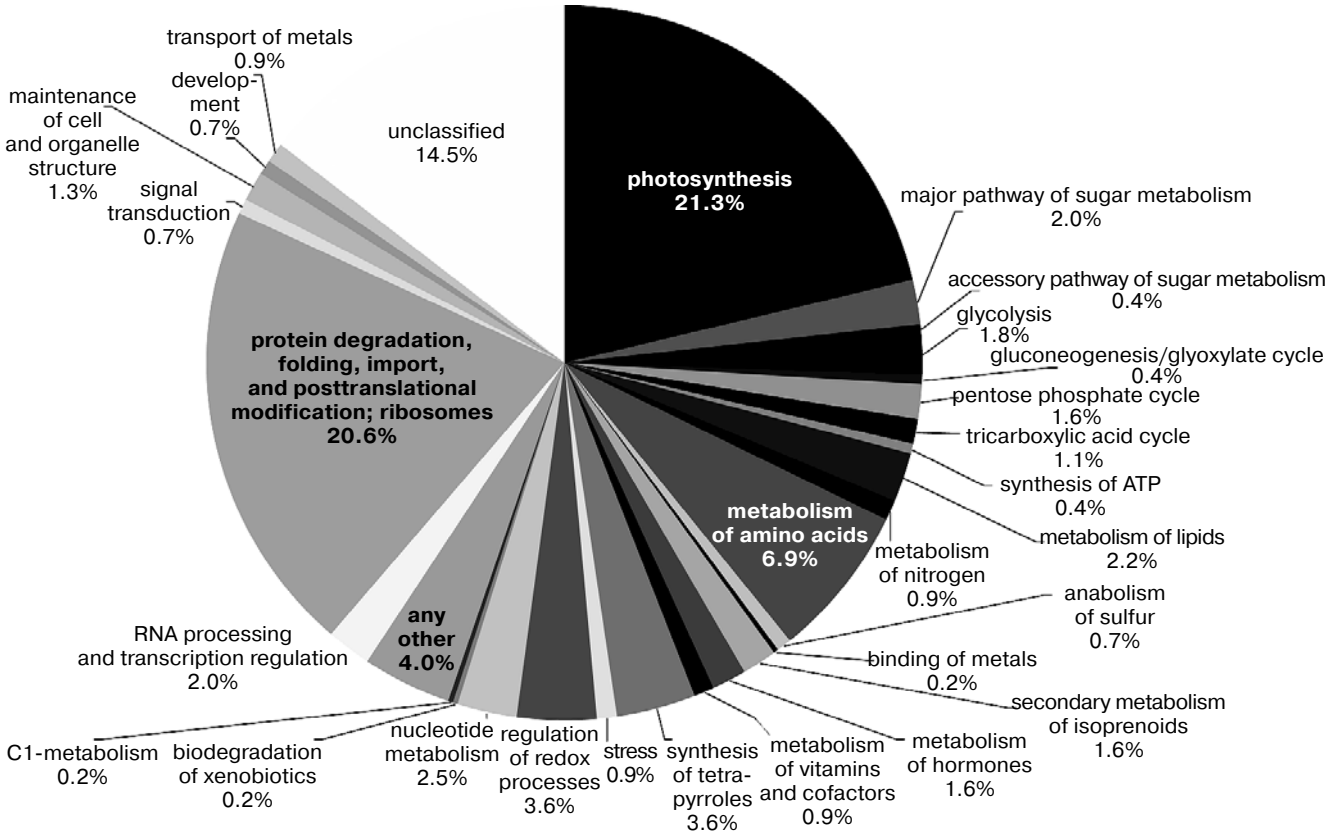

Fig. 8. Distribution of *P. patens* chloroplast resident proteins by MapMan subclasses.

The proteins that contain several detected functional domains but cannot be attributed to any of the declared classes were classified into a separate subgroup "Any others" (4%). Another subgroup combines the proteins remaining unclassified by the MapMan algorithm (13.5%) (see detailed information on these proteins in Table 3 (see Supplement at http://protein.bio.msu.ru/biokhimiya). We also subdivided into separate groups the abiotic stress proteins (1.5%) and proteins likely implicated in development. Other groups are represented by proteins associated with synthesis of tetrapyrroles (3.7%), RNA processing and transcription regulation (3.5%), and implicated in metal transport and binding (5 and 0.25%, respectively), signal transduction (0.75%), biodegradation of xenobiotics (0.25%), and maintenance of lipid bilayer structure and ATP synthesis (0.25%).

DISCUSSION

Large-scale studies of chloroplast protein composition (study of proteome and subproteomes of these organelles) in line with their transcription profiling are necessary for comprehensive characterization of their functions and biogenesis as well as implication in various plant metabolic pathways. These studies reveal earlier unknown protein–protein interactions, possible post-translational modifications, and refine protein localization within the studied organelle.

Determination of subcellular localization of proteins when studying organelle proteomes is a very important but difficult and not always trivial task. At present, the methods for data acquisition on subcellular protein localization can be classified as follows: 1) data obtained from studies on solution of a particular problem, for instance when studying a distinct protein – the data is accumulated and stored in public databases such as UniProt [41], Gene Ontology [42], Swiss-Prot, etc., and it is this data that is more significant at present since they represent a result of multiple independent studies, but they are relatively few; 2) information obtained in large-scale experiments using express cloning of open reading frames with their subsequent conjugation with various epitopes [43] or GFP-protein [44]. Localization of the produced gene products is further determined using fluorescence microscopy. The bottleneck of these methods is that the native protein structure is subjected to substantial modification during the experiment, which can result in erroneous localization. Besides, these methods are laborious and time-consuming, and interpretation of the fluores-

cence microscopy data is largely subjective. However, these methods allow localization of proteins with unknown function or poorly characterized cellular proteins; 3) the approach employed in the present work and associated with subcellular fractioning and subsequent proteome analysis of the resulting fractions. However, so-called protein impurities are always co-isolated during separation of organelles, and it is often hard to determine whether these proteins are resident or their presence in a proteome is a consequence of the destruction of other cellular compartments during isolation or insufficient purity of the specimen.

Therefore, when localization or function of a protein is unknown, the final decision on the protein assignment to a distinct organelle proteome requires additional studies involving methods other than proteomics. Even if proteins with proved localization are found in other compartments they should not be regarded as "admixtures", because multiple localizations cannot be excluded.

However, the accumulation of facts on localization of distinct proteins in the cell has enabled development of various bioinformatics methods that can predict protein localization by computational methods on the basis of its sequence.

In particular, in the present work we have used software such as TargetP [45, 46], Predotar [22], and Proteome Analyst 3.0 [23] for the confirmation of experimentally determined protein subcellular localization.

One can see (Fig. 9) that each software predicts localization of different (from 35 to 60%) number of identified proteins, and all three together predict a small number of proteins. This can result from the difference in prediction algorithms since Proteome Analyst uses data on localization accessible from public databases, while Predotar and TargetP use algorithms of "artificial intelligence" for recognition of chloroplast, mitochondrial, and secretory signal peptides within a target protein on the basis of its primary structure analysis.

It should be noted that the computational systems described above are based on determination of protein signal sequences and do not operate with nonclassical protein import via secretory pathway or when a protein has dual localization (see details in reviews [47, 48]). Besides, some reports are indicative of the absence of signal sequences in many chloroplast outer membrane proteins.

Thus, proteins related to photosynthesis predominate in the plastid proteome determined in the present work (Fig. 8). We studied some metabolic pathways involved in this process and only identified 12 of 14 key enzymes of the Calvin cycle (Fig. 10). In particular, we failed to identify ribulose 5-phosphate epimerase, the enzyme catalyzing conversion of xylulose 5-phosphate from ribulose 5-phosphate and sustaining regeneration of the major substrate of this metabolic pathway. It is worth noting that this enzyme is absent in the *A. thaliana* proteome identified by Kleiffmann et al. [49]. The authors attribute this fact to low expression level of this enzyme and suppose that ribulose 5-phosphate is preferably regenerated through sedoheptulose 1,7-bisphosphate by transketolase and ribose 5-phosphate isomerase.

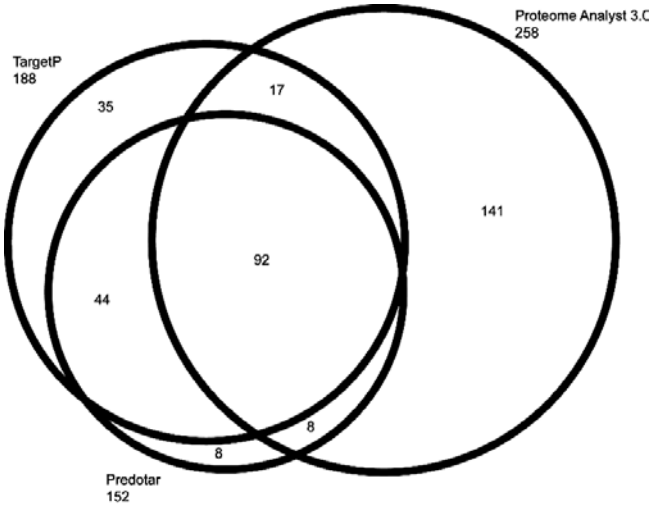

Fig. 9. Distributions of proteins identified in this work and predicted *in silico* by different computational systems as plastid resident proteins.

We identified with certainty more than 60% of the subunits comprising photosystem I and photosystem II multisubunit protein complexes, cytochrome b_6/f complex, and ATP-synthase playing an important role in light-dependent photosynthetic reactions in thylakoid membranes (Table 4; see Supplement at http://protein.bio.msu.ru/biokhimiya). Subunits that we failed to identify in the proteome are very hydrophobic with two or more transmembrane domains or have a limited number of trypsin cleavage sites (for instance, *a* and *c* subunits of chloroplast ATP-synthase). Besides, a substantial number of unidentified proteins represents minor components of photosynthetic complexes, whose detection by the methods used in the present work is complicated by their low molecular weight (below 10 kDa) and high hydrophobicity.

We also found 30 proteins among plastid resident proteins that are implicated in different amino acid biosynthetic pathways with the most comprehensively characterized ones being those of valine, leucine, and isoleucine (coverage percentage is 82% by KEGG, Fig. 11).

A significant group of plastid proteome proteins is represented by enzymes possessing protease activity and participating in various processes of protein degradation (Table 5; see Supplement at http://protein.bio.msu.ru/biokhimiya). Virtually all proteins belonging to this group are annotated in the used database as "predicted", but they possess relatively high homology with *Arabidopsis*

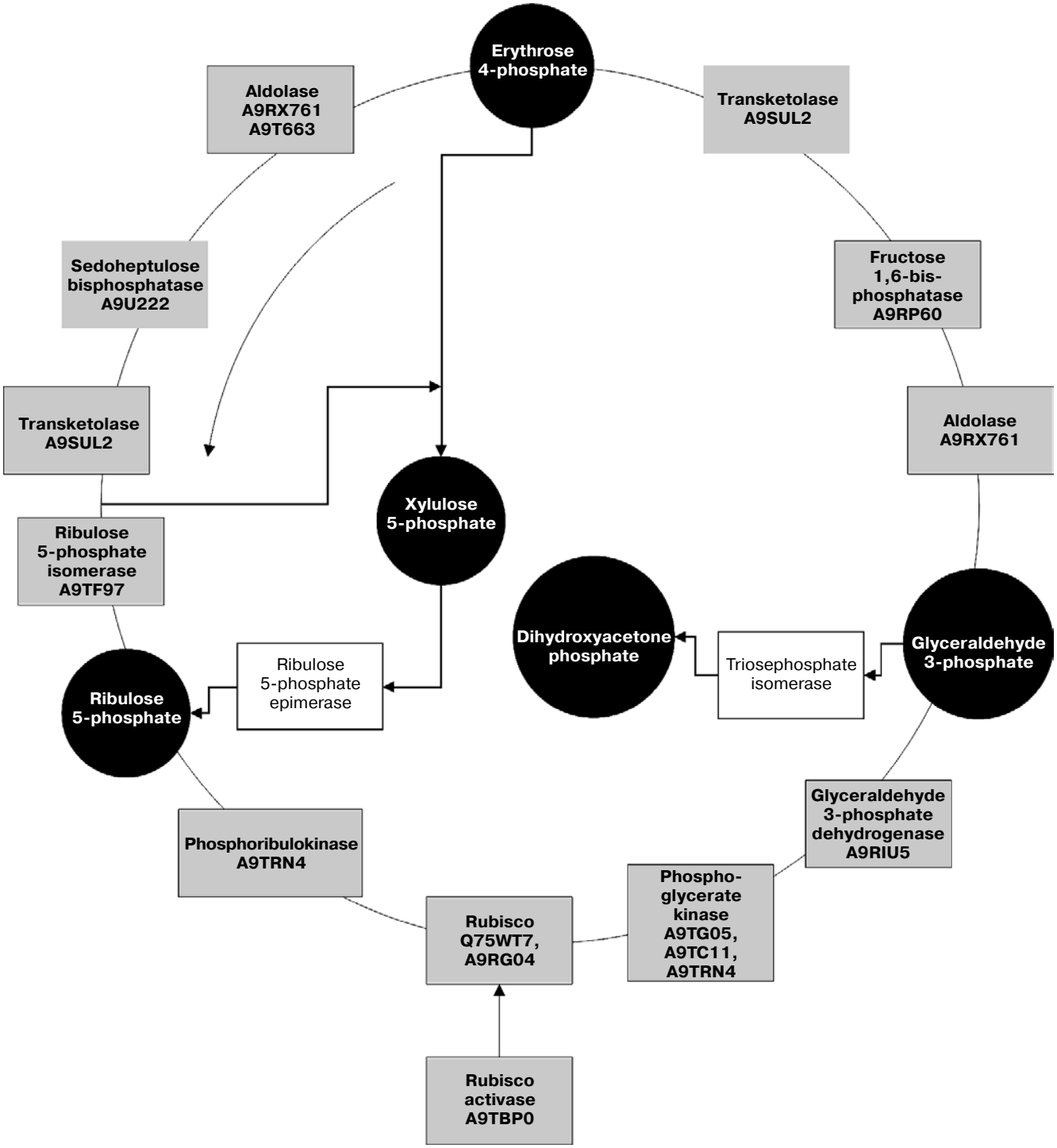

Fig. 10. Calvin cycle. Proteins identified in the present work are marked with gray squares, and some substrates of the cycle are marked with black circles. In gray squares, identification number of protein in the UniProt database is given in line with the name of the enzyme.

proteins. Among them we have revealed ATP-binding (ClpC and ClpD) and proteolytic (ClpP1, ClpP5, and ClpP6) subunits of the ATP-dependent Clp-protease complex of chloroplasts, whose main function is removal of misfolded and damaged enzymes and regulatory stromal proteins [50], maintenance of proper stoichiometry of multisubunit complexes, particularly, Rubisco [51], and fine regulation of activity of various metabolic pathways in plastids via inactivation or complete degradation of proteins composing them. Three proteins related to various metal-dependent FtsH-proteases have been identified among proteases found in thylakoid membranes.

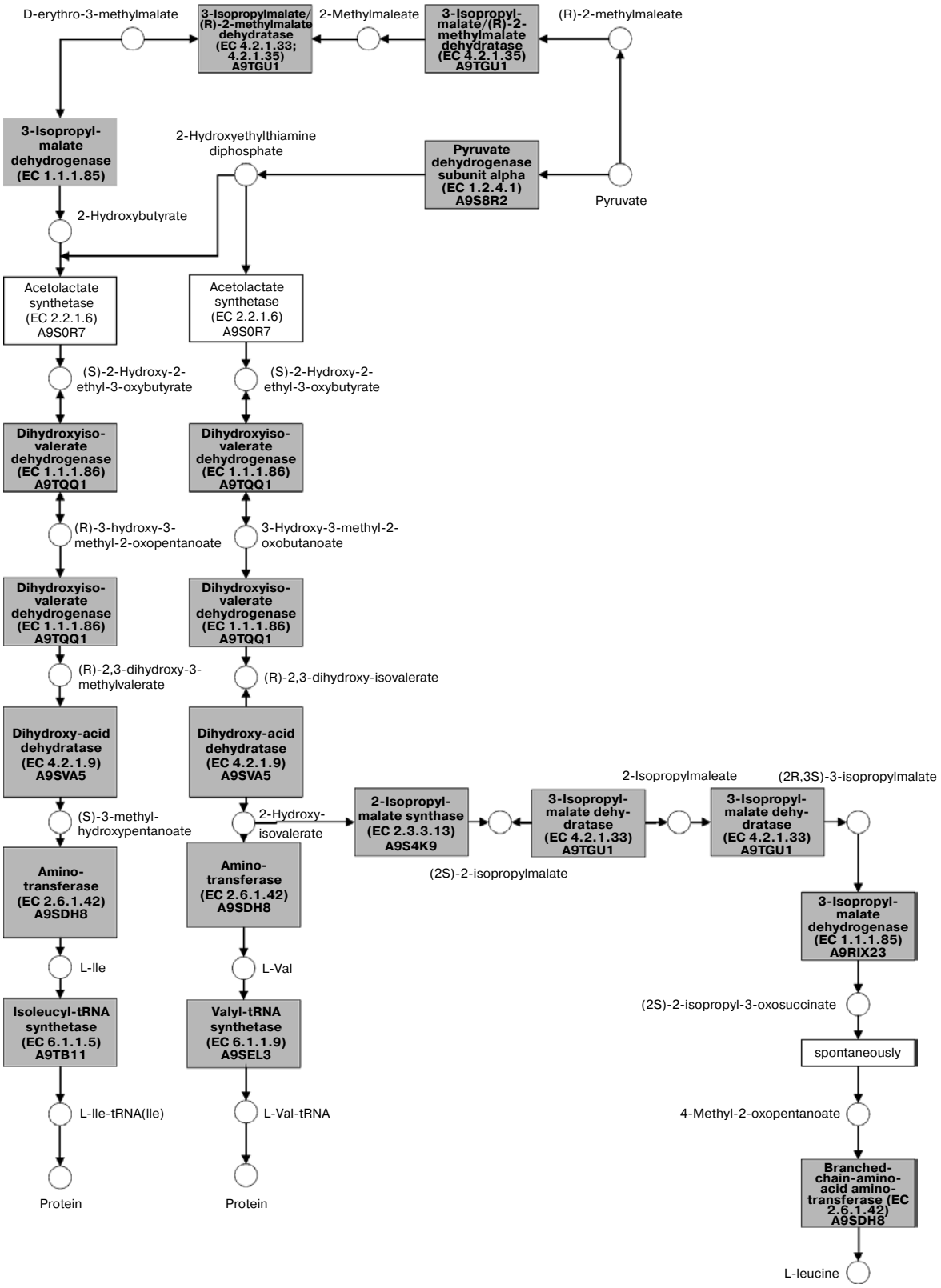

Fig. 11. Biosynthesis of valine, leucine, and isoleucine. The diagram is drawn on the basis of KEGG database report (http://kegg.jp/kegg/kegg2.html). The proteins identified in this work are marked by gray rectangles, and some substrates are marked by dark circles. Within the gray rectangles protein names and accession numbers in the UniProt database are given.

For instance, 82% identity was found between the At5g42270 protein and the predicted one numbered A9RHM7, which is the FtsH2-protease. This protein is implicated in proteolytic degradation of damaged D1 protein of plastid photosystem II, preventing cell death under extreme illumination. A9RJ05 and A9SCI8 proteins also possess high homology with *A. thaliana* proteins AT4G23940 and AT3G16290, respectively, which are annotated as putative FtsH-proteases.

High homology with *A. thaliana* proteins has been found in 374 of 434 plastid resident proteins (86%) in this study. More than 90% of them were identified earlier in one and more proteome studies fulfilled on intact chloroplasts isolated from various plant tissues and on particular sub-compartments of these organelles (see Table 2, Supplement, for the references in "Proteomics publications from PPDB" column and "Location according to proteomics investigation (SUBA database) as "a unique identification number" of PMID publication in PubMed (http://www.ncbi.nlm.nih.gov) database with short annotation about the studied sub-compartment or protein localization). It should be noted that 318 of 374 proteins are identified in the study of Zybailov et al. [52] representing the most comprehensive analysis of chloroplasts to date.

Thus, a method for chloroplast isolation from protoplasts of the moss *P. patens* has been developed in the present study, which is optimum for proteome study at the organelle level. We identified 624 proteins using 1D gel-electrophoresis and mass-spectrometry, and 434 of them are annotated as chloroplast resident proteins that are functionally adequate for photosynthetic organelles of moss protoplasts. These data will be used for further studies of proteins composing the chloroplast proteome and possessing unknown functions, as well as proteins whose localization is unclear or not unambiguously determined. This study can be used as a proteome platform in physiological and photobiological experiments in which chloroplasts are studied as important functional and signal components of green plants.

REFERENCES

1. Sugiura, C., Kobayashi, Y., Aoki, S., Sugita, C., and Sugita, M. (2003) *Nucleic Acids Res.*, **31**, 5324-5331.
2. Jarvis, P., and Soll, J. (2002) *Biochim. Biophys. Acta*, **1590**, 177-189.
3. Bruce, B. D. (2000) *Trends Cell. Biol.*, **10**, 440-447.
4. Ferro, M., Salvi, D., Brugiere, S., Miras, S., Kowalski, S., Louwagie, M., Garin, J., et al. (2003) *Mol. Cell. Proteom.*, **2**, 325-345.
5. Ferro, M., Salvi, D., Riviere-Rolland, H., Vermat, T., Seigneurin-Berny, D., Grunwald, D., et al. (2002) *Proc. Natl. Acad. Sci. USA*, **99**, 11487-11492.
6. Kieselbach, T., Hagman, B., Andersson, B., and Schroder, W. P. (1998) *J. Biol. Chem.*, **273**, 6710-6716.
7. Nakabayashi, K., Ito, M., Kiyosue, T., Shinozaki, K., and Watanabe, A. (1999) *Plant Cell. Physiol.*, **40**, 504-514.
8. Peltier, J. B., Emanuelsson, O., Kalume, D. E., Ytterberg, J., Friso, G., Rudella, A., et al. (2002) *Plant Cell*, **14**, 211-236.
9. Peltier, J. B., Friso, G., Kalume, D. E., Roepstorff, P. F., Nilsson, F., Adamska, I., and van Wijk, K. J. (2000) *Plant Cell*, **12**, 319-341.
10. Gomez, S. M., Nishio, J. N., Faull, K. F., and Whitelegge, J. P. (2002) *Mol. Cell. Proteom.*, **1**, 46-59.
11. Schubert, M., Petersson, U. A., Haas, B. J., Funk, C., Schroder, W. P., and Kieselbach, T. (2002) *J. Biol. Chem.*, **277**, 8354-8365.
12. Balmer, Y., Koller, A., del Val, G., Manieri, W., Schurmann, P., and Buchanan, B. B. (2003) *Proc. Natl. Acad. Sci. USA*, **100**, 370-375.
13. Schleiff, E., Eichacker, L. A., Eckart, K., Becker, T., Mirus, O., Stahl, T., and Soll, J. (2003) *Protein Sci.*, **12**, 748-759.
14. Skripnikov, A. Y., Polyakov, N. B., Tolcheva, E. V., Velikodvorskaya, V. V., Dolgov, S. V., et al. (2009) *Biochemistry (Moscow)*, **74**, 480-490.
15. Van Wijk, K. J., Peltier, J. B., and Giacomelli, L. (2007) *Meth. Mol. Biol.*, **355**, 43-48.
16. Laemmli, U. K. (1970) *Nature*, **227**, 680-685.
17. Shevchenko, A., Tomas, H., Havlis, J., Olsen, J. V., and Mann, M. (2006) *Nature Protocols*, **1**, 2856-2860.
18. Colinge, J., Masselot, A., Cusin, I., Mahe, E., Niknejad, A., Argoud-Puy, G., et al. (2004) *Proteomics*, **4**, 1977-1984.
19. Rensing, S. A., Lang, D., Zimmer, A. D., Terry, A., Salamov, A., et al. (2008) *Science*, **319**, 64-69.
20. 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.
21. Emanuelsson, O., Brunak, S., von Heijne, G., and Nielsen, H. (2007) *Natl. Protoc.*, **2**, 953-971.
22. Small, I., Peeters, N., Legeai, F., and Lurin, C. (2004) *Proteomics*, **4**, 1581-1590.
23. Szafron, D., Lu, P., Greiner, R., Wishart, D. S., Poulin, B., Eisner, R., et al. (2004) *Nucleic Acids Res.*, **32**, 65-71.
24. Kabeya, Y., and Sato, N. (2005) *Plant Physiol.*, **138**, 369-382.
25. Picotti, P., Aebersold, R., and Domon, B. (2007) *Mol. Cell. Proteom.*, **6**, 1589-1598.
26. Rohrbough, J. G., Breci, L., Merchant, N., Miller, S., and Haynes, P. A. (2006) *J. Biomol. Tech.*, **17**, 327-332.
27. Kyte, J., and Doolittle, R. F. (1982) *J. Mol. Biol.*, **157**, 105-132.
28. Kiraga, J., Mackiewicz, K., Mackiewicz, D., Kowalczuk, M., Biecek, P., Polak, N., et al. (2007) *BMC Genomics*, **8**, doi:10.1186/1471-2164-8-163.
29. Sun, Q., Zybailov, B., Majeran, W., Friso, G., Olinares, P. D., and van Wijk, K. J. (2009) *Nucleic Acids Res.*, **37**, 969-974.
30. Heazlewood, J. L., Verboom, R. E., Tonti-Filippini, J., Small, I., and Millar, A. H. (2007) *Nucleic Acids Res.*, **35**, 213-218.
31. Pierleoni, A., Martelli, P. L., Fariselli, P., and Casadio, R. (2007) *Nucleic Acids Res.*, **35**, 208-212.
32. Yu, Q. B., Li, G., Wang, J. C., Sun, P. C., Wang, C., Wang, H. L., et al. (2008) *Cell. Res.*, **18**, 1007-1119.
33. Kleffmann, T., Hirsch-Hoffmann, M., Gruissem, W., and Baginsky, S. (2006) *Plant Cell. Physiol.*, **47**, 432-436.

34. Peeters, N., and Small, I. (2001) *Biochim. Biophys. Acta*, **1541**, 54-63.
35. Duchene, A. M., Peeters, N., Dietrich, A., Cosset, A., Small, I. D., and Wintz, H. (2001) *J. Biol. Chem.*, **276**, 15275-15283.
36. Silva-Filho, M. C. (2003) *Curr. Opin. Plant. Biol.*, **6**, 589-595.
37. Mackenzie, S. A. (2005) *Trends Cell Biol.*, **15**, 548-554.
38. Mitschke, J., Fuss, J., Blum, T., Hoglund, A., Reski, R., Kohlbacher, O., and Rensing, S. A. (2009) *New Phytol.*, **183**, 224-235.
39. Tatusov, R. L., Koonin, E. V., and Lipman, D. J. (1997) *Science*, **278**, 631-637.
40. Thimm, O., Blasing, O., Gibon, Y., Nagel, A., Meyer, S., Kruger, P., et al. (2004) *Plant J.*, **37**, 914-939.
41. Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., et al. (2005) *Nucleic Acids Res.*, **33**, D154-159.
42. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000) *Nat. Genet.*, **25**, 25-29.
43. Kumar, A. S., Agarwal, S., Heyman, J. A., Matson, S., Heidtman, M., Piccirillo, S., et al. (2002) *Genes Dev.*, **16**, 707-719.
44. Huh, W. K., Falvo, J. V., Gerke, L. C., Carroll, A. S., Howson, R. W., Weissman, J. S., and O'Shea, E. K. (2003) *Nature*, **425**, 686-691.
45. Emanuelsson, O., and von Heijne, G. (2001) *Biochim. Biophys. Acta*, **1541**, 114-119.
46. Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G. (2000) *J. Mol. Biol.*, **300**, 1005-1016.
47. Millar, A. H., Whelan, J., and Small, I. (2006) *Curr. Opin. Plant Biol.*, **9**, 610-615.
48. Radhamony, R. N., and Theg, S. M. (2006) *Trends Cell. Biol.*, **16**, 385-387.
49. Kleffmann, T., Russenberger, D., von Zychlinski, A., Christopher, W., Sjolander, W., Gruissem, W., and Baginsky, S. (2004) *Curr. Biol.*, **14**, 354-362.
50. Halperin, T., and Adam, Z. (1996) *Plant Mol. Biol.*, **30**, 925-933.
51. Schmidt, G. W., and Mishkind, M. L. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 2632-2636.
52. Zybailov, B., Rutschow, H., Friso, G., Rudella, A., Emanuelsson, O., Sun, Q., and van Wijk, K. J. (2008) *PLoS One*, **3**, e1994.