Different plant plastid types contain a distinct protein complement for specialized functions and metabolic activities. plprot was established as a plastid proteome database to provide information about the proteomes of chloroplasts, etioplasts and undifferentiated plastids. The current version of plprot features 2,043 protein entries and consists of two modules. Module one contains a BLAST search option and provides comparative information on the proteomes of different plastid types. The second module contains four searchable databases, three for each individual plastid type and one comprehensive composite database that provides the results of plastid proteome analyses from different laboratories. plprot is accessible at http://www.plprot.ethz.ch.

Plastids are typical plant cell organelles with essential cell functions. Their biosynthetic and metabolic activities include photosynthetic carbon fixation and synthesis of amino acids, fatty acids, starch, and secondary metabolites such as pigments. On the basis of their structure, pigment composition (color), metabolism and function, plastids are classified as elaioplasts that are found in the seed endosperm, chromoplasts in fruits and petals, amyloplasts in roots, etioplasts in dark-grown seedlings and chloroplasts in photosynthetically active tissues (reviewed in Neuhaus and Emes 2000). These different plastid types result from specific differentiation programs that are controlled by the nucleus (reviewed in Voitknecht and Westhoff 2001).

During evolution, most of the plastid genes were transferred to the nuclear genome (Martin and Herrmann 1998). A systematic search for transit peptides in plastid proteins and their cyanobacterial homologues suggested that many genes in the Arabidopsis genome originated from the cyanobacterial endosymbiont, but not all of their encoded proteins appear to be imported into the plastids because they lack predictable N-terminal transit peptides (Abdallah et al. 2000, Leister 2003). Although this type of computational analysis provides useful global information about the possible evolution of the plastid proteome, it is constrained by the type of data available for the development of prediction tools. TargetP is currently the most successful prediction program, with predictive values between 70 and 85% for known plastid proteins, suggesting that a significant number of plastid proteins will not be identified by this type of analysis (Emanuelsson et al. 2000, Kleffmann et al. 2004, Richly and Leister 2004, Baginsky et al. 2005, Nair and Rost 2005). Proteomics is a powerful complementary approach to identify proteins that localize to the plastid but do not contain a predictable transit peptide.

Results from several chloroplast proteome studies have been reported during the last few years, validating the need for proteomics approaches (reviewed in Baginsky and Gruissem 2004, van Wijk 2004). Current efforts are constrained by the dynamic range of chloroplast proteins, however, including the abundant photosynthetic proteins that dominate the proteome of mature and photosynthetically active chloroplasts. To reduce this constraint and to increase plastid proteome coverage, high-throughput protein identification strategies have been developed for different plastid types. Heterotrophic plastids do not contain highly abundant photosynthetic proteins and therefore facilitate identification of proteins for other metabolic activities and regulatory pathways. We have analysed the proteomes of Arabidopsis chloroplasts, rice etioplasts and undifferentiated plastids from cultured BY2 cells that resemble proplastids. From their proteomes, we were able to infer prevalent metabolic activities and specific regulatory processes (Kleffmann et al. 2004, Baginsky et al. 2004, Baginsky et al. 2005, Zychlinski et al. 2005). We have assembled all information from our plastid proteome studies together with publicly available information and relevant plastid-specific documentation in plprot (http://www.plprot.ethz.ch).

plprot provides information about the proteomes of different plastid types that were analysed in several large-scale proteomics experiments and currently features 2,043 protein entries. Each protein in the database has a unique plprot entry code that links the protein to the organism, specific plastid type and the publication in which the plastid protein identification was reported. The plprot-identifier has the general structure plp_xx_yyyyy. The xx-position identifies the organism (i.e. at-Arabidopsis thaliana, nt-Nicotiana tabacum and os-Oryza
The plprot database allows the user to compare different proteome studies of different plastid types and identify proteins that are common between the data sets. The plprot interface consists of two modules (Fig. 1). Module one contains a menu to select plastid type-specific information and tools for proteome data analysis. ‘Plastid-type information’ provides a short description of each plastid type, their isolation and purification protocols, and additional information that can be retrieved from the database such as protein solubility or correlations between protein and transcript levels. Further details on methodology, mass spectrometric protein identification and biological interpretation of the data are available in original publications that are accessible via the URL provided on each information sheet. The menu ‘tools for data integration’ allows the user to select the options ‘plastid type comparisons’.
Table 1  Description of the different plastid databases available via plprot and the information content therein

<table>
<thead>
<tr>
<th>Databases</th>
<th>No of entries</th>
<th>Fields (information)</th>
<th>Search options</th>
<th>BLAST searchable</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tobacco proplastid</td>
<td>168</td>
<td>plprot ID</td>
<td>All fields individually and wildcard searches (any of the words)</td>
<td>YES, via 'plprot database' and 'Complete plastid database'</td>
<td>Contains all protein identifications from a shotgun analysis of tobacco BY2 cell culture plastids</td>
</tr>
<tr>
<td>Rice etioplast</td>
<td>240</td>
<td>Gene ID, protein description, amino acid sequence, function, TargetP prediction, identified peptides, highest SEQUEST score, best homologue in the Arabidopsis genome, best homologue in Arabidopsis chloroplast</td>
<td>All fields individually and wildcard searches (any of the words)</td>
<td>YES, via 'plprot database' and 'Complete plastid database'</td>
<td>Contains all protein identifications from a shotgun analysis of rice etioplasts</td>
</tr>
<tr>
<td>Arabidopsis 690 chloroplast</td>
<td>690</td>
<td>plprot ID</td>
<td>All fields individually and wildcard searches (any of the words)</td>
<td>YES, via 'plprot database' and 'Complete plastid database'</td>
<td>Contains all protein identifications from a shotgun analysis of Arabidopsis chloroplasts together with transcript levels for all identified proteins</td>
</tr>
<tr>
<td>Complete plastid</td>
<td>2,043</td>
<td>plprot ID</td>
<td>All fields (also individually) except for homologues</td>
<td>YES, via 'Complete plastid database'</td>
<td>Assembly of all protein identifications from published shotgun analyses of higher plant plastids, including those above (proplastid, etioplast, chloroplast), partially redundant</td>
</tr>
</tbody>
</table>

The number of entries reflects the current status (September 2005).
The BLAST search can be performed on two different databases. One contains data exclusively from our own proteome studies that provide the foundation for plastid type comparisons (‘plprot’). The other database also includes publicly available information from different large-scale Arabidopsis chloroplast proteome studies as well as our own data on different plastid types (‘complete plastid proteome’). This database features the above-mentioned 2,043 protein entries. It is partially redundant because some proteins were identified in more than one study and from different organisms (see also references in plprot). The plprot entry code allows protein identifications from different proteome studies to be distinguished. The BLAST search results provide all relevant information about the identified proteins, including the plprot identifier, the organism, the literature reference and the BLAST e-value. With this search option, the user has access to a comprehensive compilation of all plastid proteins that have been identified in different proteome studies.

For the biological interpretation of the plastid proteome data, the user can select ‘plastid type comparison’, which is linked to a figure of a triangle that represents the different plastid types in each vertex (Fig. 2A). Since the direction of the homology search determines how many homologues can be identified, we have BLAST searched the data from each vertex of the triangle (i.e. each plastid type). Each field is clickable. The intersections between the plastid types display the proteins that are common to two or all three plastid types (Fig. 2A, B). The individual field of each vertex displays the proteins that were specifically identified in one plastid type only [i.e. etioplast (EP), chloroplast (CP) and proplastid (PP)]. The assignment of proteins to more than one plastid type was made on the basis of BLAST searches for homologues, for which their respective e-values are provided in the output table. Additional information (e.g. amino acid sequence, author and plprot accession number) can be accessed via the URL at the accession number. Activating the field in the intersection of, for example, EP, CP and PP will open a protein list that comprises those proteins that were identified in all three plastid types (Fig. 2B).

The BLAST search results are the basis for a comprehensive comparison of different plastids and allow a core proteome that is common to all plastid types to be defined. Most of the proteins that constitute the core proteome are involved in carbohydrate and amino acid metabolism (Fig. 2B). Examples of these enzymes include glutamate synthase, cysteine synthase, transketolase, phosphoglycerate kinase and glycer-aldehyde 3-phosphate dehydrogenase. Amino acid and carbohydrate metabolism are housekeeping functions that are essential for the plant cell. It is therefore not surprising that these functions are common to all different plastid types analysed to date. More surprisingly, a large number of proteins found in all plastid types are proteases and chaperones. This suggests that high protein turnover is a common feature of plastids. The most abundant proteins in this group are the Rubisco-binding protein CPN60 and the chloroplast heat shock protein 70.

The second plprot module contains the databases for the three plastid types and a composite database that contains all proteins identified from large-scale proteome analyses (‘complete plastid’, 2,043 entries altogether). Since each database contains specific information, the databases can be searched individually to display all information (summarized in Table 1). For example, the BY2 database contains information about the specific serial extraction step from which the protein was identified, i.e. osmotic shock, 8 M urea wash or detergent solubilization. This information is useful to assess the solubility of proteins and their membrane attachment in vivo. For most of the identified Arabidopsis chloroplast proteins, we provide RNA expression levels determined by Affymetrix ATH1 GeneChip® analysis (Kleffmann et al. 2004). The transcript level of identified proteins was useful to estimate the depth of our proteome analysis and the abundance of a protein in the chloroplast (Baginsky et al. 2005). These features, together with the manual annotation of protein functions that includes different plastid developmental stages, make plprot a unique protein database that reports plastid proteome dynamics resulting from specific differentiation processes. In this respect, plprot differs considerably from PPDB, a chloroplast protein database that is available at the Boyce Thompson Institute (Friso et al. 2004).

plprot will be updated regularly as new proteins are being identified from different plastid types. We are currently developing a two-dimensional PAGE-based proteome map of rice etioplasts that will be integrated with plprot to complement the rice etioplast shotgun proteomics data. In the longer term, plprot will be expanded to display proteome dynamics during light-induced differentiation of etioplasts into chloroplasts. We are convinced that plprot will become a useful tool and valuable resource for plant scientists who are interested in basic cell biology and organelle differentiation processes.

Materials and Methods

plprot was constructed as a user-friendly online tool for the comparison of different plastid-type proteomes. It consists of an MySQL relational database and a Web server application programmed in the PHP (PHP Hypertext Preprocessor) scripting language. With plprot, we offer a WWW-BLAST server that was constructed from the software available at http://www.ncbi.nlm.nih.gov/Ftp/ (see also descriptions therein). Each database entry is connected to a unique plp identifier that carries the prefix ‘plp’ followed by a two-digit code for the plant species and a unique five-digit number. The database design allows an easy and almost infinite upgrade with new data sets and related data. It consists of a basic table, which contains the TIGR IDs or the NCBI gene ID, the plp identifier, the organism, the names of the authors who published the protein identification, the protein annotation and the amino acid sequence. All additional information is integrated into supplementary tables and is uniquely linked to the basic table via the plp identifier. The database works as a ‘data warehouse’ contain-
ing experimental, annotation and comparative data of proteins identified from different plastid types.

The comparative proteome data for different plastid types were obtained by a BLAST search of “Complete plastid” (Table 1) against itself, considering an e-value of e-20 as significant homology (Altschul et al. 1997). For the plastid type comparison that is provided in plprot, we only used the protein identifications from our laboratory (Kleffmann et al. 2004, Baginsky et al. 2004, Baginsky et al. 2005, Zychlinski et al. 2005). This is necessary since a valid comparison of proteomes requires that the data were obtained from the same mass spectrometric (MS) set-up to minimize experimental or technical variation. All homologues in these sets of proteins were assorted to the plastid type and are provided in the plprot database (‘plastid type comparison’). All homologues in studies from other laboratories are provided in the module ‘Complete Plastid’.

The MS protein identification is described in the publications that were used to construct plprot. In general, all data were derived from shotgun proteomics analysis using MS and MS/MS scans for peptide amino acid sequence assignments. The protein annotation was retrieved from the protein databases that were downloaded from the TIGR server (http://www.tigr.org) for Arabidopsis thaliana and Oryza sativa, and from the NCBI server (http://www.ncbi.nlm.nih.gov) for Nicotiana tabacum.

References


(Received November 15, 2005; Accepted January 3, 2006)