The Plant-Specific Database. Classification of Arabidopsis Proteins Based on Their Phylogenetic Profile[1]

Rodrigo A. Gutiérrez[2]*, Matthew D. Larson, and Curtis Wilkerson

Department of Energy Plant Research Laboratory (R.A.G., C.W.), Department of Biochemistry and Molecular Biology (R.A.G.), and Department of Plant Biology and Genomics Technology Support Facility (M.D.L., C.W.), Michigan State University, East Lansing, Michigan 48824–1312

One of the main goals of the plant community is to determine the function of every Arabidopsis gene by the end of the year 2010 (Chory et al., 2000). Because many are not yet characterized and because they will not be studied in other model organisms, we believe Arabidopsis proteins that are unique to the plant lineage should be a priority for future studies. Furthermore, functional characterization of plant-specific proteins is likely to provide novel insights about plant biology. In an effort to aid the functional characterization of the Arabidopsis proteins that are specific to plants, we have created the Plant-Specific Database (PLASdb). PLASdb contains Arabidopsis proteins classified according to their pattern of sequence similarity to proteins in organisms from all forms of life. PLASdb identifies 3,848 Arabidopsis proteins as plant specific because they are found only in plant species. In addition, 4,816 other proteins are classified in various groups based on their specific patterns of conservation among other Eukarya, Bacteria, or Archaea.

The main goal of PLASdb is to stimulate further research on some of the least-studied proteins of plants. To this end, we have compiled and integrated information from public data sources (e.g. The Institute for Genomic Research [TIGR], Munich Information center for Protein Sequences [MIPS], The Arabidopsis Information Resource [TAIR]) with links to the original information and provided links to external databases (e.g. Salk Institute Genomic Analysis Laboratory [SIGNAL]). In addition, we have performed predictions of subcellular localization and transmembrane helices, analyzed gene expression in organs based on expressed sequence tag (EST) frequencies and microarray data, and grouped protein families based on sequence similarity clustering (BLASTCLUST). Web-based interfaces to several search engines allow gene-driven or exploratory modes of data access. Information in PLASdb can be quickly downloaded in text or Excel format for further analysis (http://genomics.msu.edu/plant_specific).

CONTENT OF THE PLASdb

PLASdb is a relational database of the nuclear-encoded Arabidopsis proteins (hereinafter proteins) classified according to their pattern of sequence similarity in the protein sets of the following organisms: *Homo sapiens*, *Rattus norvegicus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Mus musculus*, *Schizosaccharomyces pombe*, *Saccharomyces cerevisiae*, a combined set of 88 species of Bacteria, and a combined set of 16 species of Archaea. We determined the phylogenetic profile of each protein sequence as a vector of nine values that indicates the similarity of the Arabidopsis protein to the proteins in each of the other sets. A similar approach was taken in previous studies of protein function in *Escherichia coli* (Pellegrini et al., 1999; Peregrin-Alvarez et al., 2003). However, in contrast to these earlier studies, we applied a two-step filtering process to define the phylogenetic profile of each Arabidopsis protein. In the first filtering step, we defined the presence of a similar sequence in another protein set when the BLASTP E-value for the best match between the Arabidopsis protein query and one of the protein sets was 10^{-10} or less. We defined absence of a similar sequence when the BLASTP E-value was greater than 0.01. Proteins with intermediate E-values in the BLASTP comparisons against any of the protein sets were excluded from further consideration. Although almost one-half of the Arabidopsis proteins could not be assigned a phylogenetic profile with this criterion, this conservative criterion allowed us to focus on the proteins with the clearest pattern of conservation throughout the phylogeny (figure 3 in the PLASdb Web site).

Because the absence of Arabidopsis proteins in all protein sets utilized could occur trivially because of incorrect Arabidopsis gene predictions, we carried out the following second filtering step. We compared the

[1] This work was supported by the Department of Energy (R.A.G.; grant no. DE–FG02–91ER20021 to Dr. Pamela J. Green and Dr. Kenneth Keegstra) and by the National Science Foundation (R.A.G.; grant no. DBI–9943561 to Dr. Pamela J. Green, Dr. Kenneth Keegstra, and Dr. John B. Ohlrogge).

[2] Present address: Department of Biology, 100 Washington Square East, 1009 Main Building, New York University, New York, NY 10003.

* Corresponding author; e-mail rg98@nyu.edu; fax 212–995–4204.

www.plantphysiol.org/cgi/doi/10.1104/pp.104.043687.

7,868 protein sequences that showed no detectable sequence similarity to proteins in other organisms against the Arabidopsis EST database and the EST databases of 13 other vascular plant species (see list of species in the PLASdb Web site). We considered as plant specific any of the 7,868 proteins identified previously that showed significant sequence similarity (E-value $\leq 10^{-10}$) against protein sequences in the Arabidopsis EST database and the databases of four other plant species. After excluding proteins encoded in retroelements or transposable elements, 3,848 Arabidopsis proteins were selected and classified as plant-specific proteins (figure 3B, PLASdb Web site). The current list of plant-specific genes is not exhaustive for several reasons: (1) the stringency of our criteria is too high; (2) not all correctly predicted genes will have an EST sequence due to a multiplicity of reasons; and (3) we restricted the study to protein-coding genes, and at least some noncoding RNAs are likely to be kingdom specific (MacIntosh et al., 2001). Although nonexhaustive, this approach allowed us to identify a set of sequences that are strong candidates for expressed proteins that are plant specific. For a detailed analysis of the Arabidopsis plant-specific proteins, see Gutiérrez et al. (2004).

In addition to the classification of Arabidopsis proteins based on their individual phylogenetic profiles, in PLASdb we have integrated information from multiple public databases and computer prediction programs. For each protein in PLASdb, the following fields of information are stored when available (summarized in Fig. 1): M_r and pI (TIGR); enzyme commission identifier from the Kyoto Encyclopedia of Genes and Genomes (Kanehisa et al., 2002); MIPS automatic functional category assignment (Frishman et al., 2003); prediction of subcellular localization by TargetP (Emanuelsson et al., 2000) and prediction of transmembrane helices by transmembrane hidden Markov model (Krogh et al., 2001); and protein domain analysis and SignalP analysis (TIGR). Each protein is associated with its corresponding gene, and the following gene information is also stored in PLASdb when available (Fig. 1): gene name and comment from the version 4.0 Arabidopsis genome of TIGR; GenBank accession number; gene family membership based on sequence similarity clustering (BLASTCLUST, described in the PLASdb Web site); gene family based on expert annotation (Anantharaman et al., 2002; Beisson et al., 2003; Rhee et al., 2003); external links to TIGR, MIPS, TAIR, and SIGNAL databases; gene expression in plant organs based on EST frequencies as described earlier (Beisson et al., 2003) and microarray experiments; and gene ontology assignments (TIGR). Details of the database implementation are available in the PLASdb Web site (http://genomics.msu.edu/plant_specific/implementation.html).

All the information associated with an Arabidopsis protein is displayed in an HTML page that we termed the Protein Properties page (see example page in Fig. 2A). This page can be accessed directly from the home page with a locus identifier (e.g. At2g39990) or from any of the HTML tables generated by the output engines (Fig. 1) as detailed below.

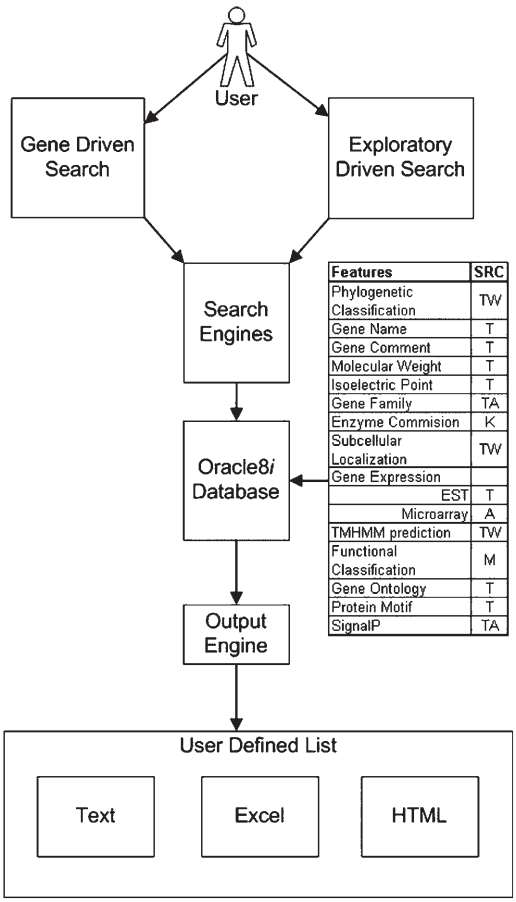


Figure 1. Schematic representation of the PLASdb implementation and content. The table on the side summarizes the features stored in PLASdb for each protein and its corresponding gene. Sources (SRC) for this information are as follows: K, Kyoto Encyclopedia of Genes and Genomes; M, MIPS; T, TIGR; TA, TAIR; TW, determined in this work. TAIR was the primary source of Arabidopsis gene families, but we have also incorporated lipid genes from the Lipid Gene Database (Beisson et al., 2003) and the Arabidopsis proteins involved in RNA metabolism as defined by Anantharaman et al. (2002).

FINDING AND DOWNLOADING INFORMATION

PLASdb was designed to allow easy and quick access to the information stored. We have implemented intuitive Web-based interfaces and facile ways to retrieve the data by the individual user for further analysis. We envision at least two scenarios that would motivate researchers to visit PLASdb.

Gene-Driven Scenario

In this scenario researchers would have a gene or list of genes of interest that they would like to study in

A

Protein Properties

Description	**Expressed Protein**
Comment	N/A
Classification	Plant Specific
TIGR ID	At1g08380
GB	AC006932
Gene Family based on BLASTCLUST	Family Size: 1
Expert gene family (data source: TAIR) (Super Gene Family / Gene Family)	N/A
MW (TIGR)	15143.51
pI (TIGR)	10.368
EC Analysis (KEGG)	N/A
External Links	[TIGR] [MIPS] [TAIR] [SIGnAL]

Subcellular Localization (TargetP Prediction)

Localization	Chloroplast Score	Mitochondrial Score	Secretory Score	Signal Length	Reliability Class
Chloroplast	.974	.029	.005	-1	1

Gene Expression in different plant organs

EST data

TC	Total # ESTs	Seeds	Roots	Leaves	Flowers	Siliques	Seedlings	Cell Suspension	Mixed Tissues
TC161032	63	**0** (0.00%)	**1** (0.00%)	**0** (0.00%)	**18** (0.24%)	**1** (0.01%)	**4** (0.06%)	**0** (0.00%)	**10** (0.02%)

AFGC microarray data (Log2 of normalized ratio)

Flower vs. Leaves	Stem vs. Ref.	Silique vs. Ref.	Root vs. Ref.	Flower vs. Ref.	Leaf vs. Ref.
.25375	-.25525	-.5355	-1.12975	-1.42325	-5.03425

Membrane/Soluble (TMHMM Prediction)

Location	Expected # of AA	Expected # of AA in first 60	Predicted Helices	Topology
membrane	42.39	.43	2	i61-80o110-132i

MIPS Automatic Functional Classification

Category Number	Description	Hitcode	P-value
No Functional Classification data available			

Gene Ontology

Accession	Description	Type	Code	Evidence	Reference	Synonym
No Gene Ontology data available						

B

Sort A to Z
Sort Z to A
Download This Dataset: [Text] [Excel]
Plant Specific list (number of genes returned is: **3848**)

Locus	Description	Subcellular Localization (TargetP Prediction)	Membrane/Soluble (TMHMM Prediction)	Automatic MIPS Functional Assignment (Top Hit only)	# ESTs
At1g07090.1	Expressed Protein	No Definition	Soluble	Unclassified Proteins	6
At1g07795.1	Expressed Protein	No Definition	Soluble	Not Defined	1
At1g08380.1	Expressed Protein	Chloroplast	Membrane	Not Defined	63
At1g08390.1	Expressed Protein	No Definition	Soluble	Not Defined	3
At1g08480.1	Expressed Protein	No Definition	Soluble	Not Defined	8
At1g09155.1	Expressed Protein	Secretory	Soluble	Other Protein Fate Related Activities	3
At1g09310.1	Expressed Protein	No Definition	Soluble	Unclassified Proteins	1
At1g09520.1	Expressed Protein	Chloroplast	Soluble	Not Defined	7
At1g09950.1	Expressed Protein	Secretory	Soluble	Transcriptional Control	5
At1g10020.1	Expressed Protein	No Definition	Soluble	Unclassified Proteins	10

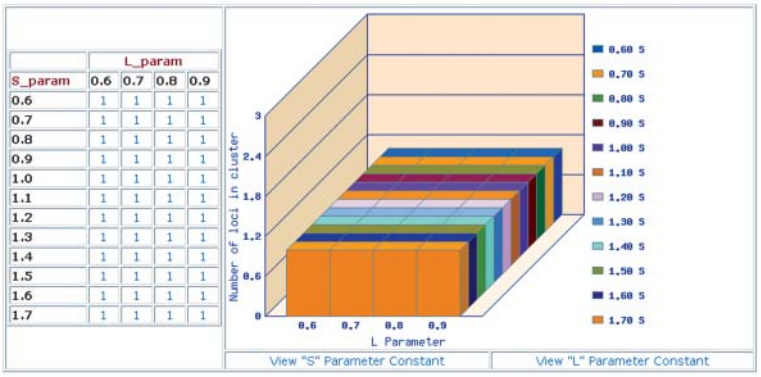
Figure 2. Sample HTML pages obtained by querying PLASdb. A, Protein Properties page for locus At1g08380. Only the first portion of the page is displayed in the figure. B, Example of the table obtained by using the Locus Search, Advanced Search, or Gene Family Search engines. Loci numbers in the left-most column are hyperlinked to the corresponding Protein Properties page. C, BLASTCLUST analysis for gene At1g08380. D, BLASTCLUST analysis for gene At1g74420. Gene families in PLASdb are defined using the BLASTCLUST parameters L = 0.6 and S = 0.8. The graph allows quick evaluation of the gene family size as a function of L and S. In addition, the table at left allows the user to generate clusters at various L and S values for further analysis.

PLASdb. The database can be queried for individual or multiple Arabidopsis loci identifiers. Searching by genes will retrieve the entry(ies) found in the database and a list of the following properties: locus number, gene name, prediction of subcellular localization, prediction of transmembrane helices, the top automatic MIPS functional category assignment, and the total number of ESTs for the corresponding gene (TIGR Arabidopsis Gene Index). Individual Protein Properties page(s) can be accessed from this table by following the hyperlink on each of the loci numbers in the left-most column (Fig. 2B). For users interested in groups of functionally related genes, it is also possible to search PLASdb by gene families, as defined by experts in the field. We have incorporated the gene family information compiled and maintained by TAIR (Rhee et al., 2003), the information from the Lipid Gene Database (Beisson et al., 2003), and a list of proteins involved in RNA metabolism defined by Anantharaman et al. (2002) into a simple table that directly queries the PLASdb (http://genomics.msu.edu/cgi-bin/plant_specific/family_search.cgi). Searching by gene families will produce a table with the same information as searching by genes, with an additional column indicating the expert gene family annotation.

Exploratory Scenario

A second scenario can be envisioned in which users would browse the database in an exploratory mode to discover new aspects of plant biology. A researcher would have a process or general question in mind and would use PLASdb as a way to guide the generation

C BLASTCLUST result for At1g08380



D BLASTCLUST result for At1g74420

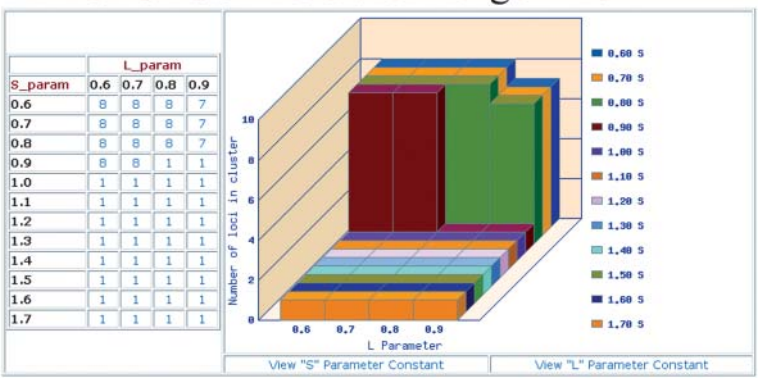


Figure 2. (*continued*).

of a new hypothesis. To illustrate this scenario, we present two examples.

Case 1: Plant-Specific Proteins Involved in Photosynthesis

A researcher is studying aspects of photosynthesis in plants and would like to identify potential candidates for future reverse-genetic strategies that might lead to findings that are of significance to plants but not to other photosynthetic organisms. As a starting point, a researcher may want to focus on proteins that are found only in plant species but not in photosynthetic bacteria. PLASdb can quickly generate a list of Arabidopsis proteins that lack similarity to protein sequences in cyanobacteria or other organisms. In addition, PLASdb allows narrowing the list further, for example, to proteins that may be involved in processes related to energy pathways and that are localized in the chloroplast. Using the Advanced Search tool, the user can quickly generate a list of proteins with the properties of interest and download a table with these proteins to a personal computer. It is possible to learn more about the individual proteins by visiting the individual Protein Properties pages. Thus, in a relatively short period of time, candidates for future studies can be identified and evaluated.

Case 2: Using PLASdb as a Guide to Study the Function of Unknown Arabidopsis Proteins

In this second example, we illustrate the use of the database to identify Arabidopsis lines with T-DNA insertions in genes that encode plant-specific proteins of unknown function. In addition, to facilitate potential future studies, the user would like to concentrate the effort on genes that are highly expressed and belong to small gene families. The user can quickly retrieve a table with Arabidopsis proteins that are plant specific and sort this table by the number of ESTs present in the TIGR Arabidopsis Gene Index (The TIGR Gene Index Databases, 2003) for the corresponding gene. The user can now focus on the Description column and visually identify proteins annotated as Expressed proteins only (example of such table is shown in Fig. 2B). The size of the gene family can be evaluated by looking at the BLASTCLUST results reported in the Protein Properties page. In PLASdb we have arbitrarily defined a gene family as those proteins that clustered by BLASTCLUST using the parameters L (length) = 0.6 and S (similarity) = 0.8. Results for At1g08380, a one-member gene family, are shown in Figure 2C. Another gene, which belongs to a larger gene family, is presented for comparison (Fig.

2D). Because the number of genes included in a gene family depends on the criteria for inclusion (Fig. 2D), a graphical representation of the gene family size as a function of L and S is provided for the user. In addition, clusters at various combinations of L and S can be readily generated by the user for further analysis (Fig. 2, C and D). Additional information presented in the Protein Properties page can help the final selection. For example, it is possible to evaluate the expression of the corresponding genes in plant organs based on EST frequency in libraries from specific organs or based on publicly available microarray experiments (Fig. 2A). Finally, the user can follow the link to the SIGNAL database (Alonso et al., 2003) to investigate whether potential knockout lines are available for the selected genes. From the SIGNAL database and then through TAIR, seeds for the putative knockout line can be requested.

See the HOW-TO in the PLASdb Web site for further details and step-by-step examples of how to use the database.

FINAL REMARKS

We believe the study of Arabidopsis proteins that are unique to the plant lineage should be a priority for future functional studies. Many of these proteins have unknown functions and are not likely to be studied in other model organisms, such as yeast. Thus, their characterization represents a challenge for the plant community. Here, we describe a new resource to facilitate the functional characterization of the Arabidopsis proteins that are specific to plants. The PLASdb identifies 3,848 Arabidopsis proteins as plant specific. In addition, 4,816 other proteins are classified in various groups based on their specific patterns of conservation among other eukaryotes, bacteria, or archaea. PLASdb contains extensive information compiled from multiple public data sources (e.g. annotation information, expression in organs based on microarray data), and generated with predictive algorithms (e.g. protein families, subcellular localization). We hope this new resource stimulates further research by identifying and providing quick and easy access to available information about the Arabidopsis plant-specific proteins.

ACKNOWLEDGMENTS

We thank Dr. Pamela J. Green, Dr. Kenneth Keegstra, and Dr. John B. Ohlrogge for support and valuable comments throughout this project. We thank Dr. John B. Ohlrogge for critical reading of this manuscript. We thank Dr. Robert Halgren for expert bioinformatics assistance. We thank Dr. Vivek Anantharaman and Eugene V. Koonin for providing the list of Arabidopsis proteins involved in RNA metabolism. We thank Karen Bird for editorial assistance.

Received March 28, 2004; returned for revision April 8, 2004; accepted April 8, 2004.

LITERATURE CITED

Alonso JM, Stepanova AN, Leisse TJ, Kim CJ, Chen H, Shinn P, Stevenson DK, Zimmerman J, Barajas P, Cheuk R, et al (2003) Genome-wide insertional mutagenesis of Arabidopsis thaliana. Science **301:** 653–657

Anantharaman V, Koonin EV, Aravind L (2002) Comparative genomics and evolution of proteins involved in RNA metabolism. Nucleic Acids Res **30:** 1427–1464

Beisson F, Koo AJ, Ruuska S, Schwender J, Pollard M, Thelen JJ, Paddock T, Salas JJ, Savage L, Milcamps A, et al (2003) Arabidopsis genes involved in acyl lipid metabolism. A 2003 census of the candidates, a study of the distribution of expressed sequence tags in organs, and a web-based database. Plant Physiol **132:** 681–697

Chory J, Ecker JR, Briggs S, Caboche M, Coruzzi GM, Cook D, Dangl J, Grant S, Guerinot ML, Henikoff S, et al (2000) National Science Foundation-Sponsored Workshop Report: ''The 2010 Project'' functional genomics and the virtual plant. A blueprint for understanding how plants are built and how to improve them. Plant Physiol **123:** 423–426

Emanuelsson O, Nielsen H, Brunak S, von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol **300:** 1005–1016

Frishman D, Mokrejs M, Kosykh D, Kastenmuller G, Kolesov G, Zubrzycki I, Gruber C, Geier B, Kaps A, Albermann K, et al (2003) The PEDANT genome database. Nucleic Acids Res **31:** 207–211

Gutiérrez RA, Green PJ, Keegstra K, Ohlrogge JB (2004) Phylogenetic profiling of the Arabidopsis thaliana proteome: What proteins distinguish plants from other organisms? Genome Biol **5:** R53

Kanehisa M, Goto S, Kawashima S, Nakaya A (2002) The KEGG databases at GenomeNet. Nucleic Acids Res **30:** 42–46

Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol **305:** 567–580

MacIntosh GC, Wilkerson C, Green PJ (2001) Identification and analysis of Arabidopsis expressed sequence tags characteristic of non-coding RNAs. Plant Physiol **127:** 765–766

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci USA **96:** 4285–4288

Peregrin-Alvarez JM, Tsoka S, Ouzounis CA (2003) The phylogenetic extent of metabolic enzymes and pathways. Genome Res **13:** 422–427

Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M, et al (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. Nucleic Acids Res **31:** 224–228

The TIGR Gene Index Databases (2003). The Institute for Genomic Research. http://www.tigr.org