D852–D856 *Nucleic Acids Research, 2007, Vol. 35, Database issue* *Published online 16 November 2006*
doi:10.1093/nar/gkl795

POGs/PlantRBP: a resource for comparative genomics in plants

Nigel S. Walker, Nicholas Stiffler and Alice Barkan*

Institute of Molecular Biology, University of Oregon, Eugene, OR 97403, USA

Received August 7, 2006; Revised September 19, 2006; Accepted October 2, 2006

ABSTRACT

POGs/PlantRBP (http://plantrbp.uoregon.edu/) is a relational database that integrates data from rice, Arabidopsis, and maize by placing the complete Arabidopsis and rice proteomes and available maize sequences into 'putative orthologous groups' (POGs). Annotation efforts will focus on predicted RNA binding proteins (RBPs): i.e. those with known RNA binding domains or otherwise implicated in RNA function. POGs form the heart of the database, and were assigned using a mutual-best-hit-strategy after performing BLAST comparisons of the predicted Arabidopsis and rice proteomes. Each POG entry includes orthologs in Arabidopsis and rice, annotated with domain organization, gene models, phylogenetic trees, and multiple intracellular targeting predictions. A graphical display maps maize sequences on to their most similar rice gene model. The database can be queried using any combination of gene name, accession, domain, and predicted intracellular location, or using BLAST. Useful features of the database include the ability to search for proteins with both a specified domain content and intracellular location, the concurrent display of mutual best hits and phylogenetic trees which facilitates evaluation of POG assignments, the association of maize sequences with POGs, and the display of targeting predictions and domain organization for all POG members, which reveals consistency, or lack thereof, of those predictions.

INTRODUCTION

Comparative analysis of orthologous genes is a powerful method for elucidating gene structure, function and evolution. Identification of orthologs on a gene-by-gene basis can be labor-intensive, and accessing the data available for orthologous genes typically requires visits to multiple species-specific databases. Thus, there is a need for resources that predict orthologous groups and bring together information about the orthologs in a manner that simplifies comparative analyses.

To facilitate cross-species comparisons among the major model plant species *Arabidopsis thaliana* (Arabidopsis), *Oryza sativa* (rice) and *Zea mays* (maize), we developed the database POGs/PlantRBP (http://plantrbp.uoregon.edu/). POGs/PlantRBP clusters proteins in the rice and Arabidopsis proteomes into putative orthologous groups (POGs) based on a mutual-best-hits strategy, with POG assignments subsequently evaluated by phylogenetic analysis. Each POG display page includes a graphical representation of the domain organization of POG members, the results of two targeting predictors for each of the nucleus, chloroplast, mitochondrion, and secretory system, and a phylogenetic tree from which users can navigate to related POGs. Maize genomic and cDNA sequences are associated with the POG containing the rice gene with which they are most similar. The design of the database is illustrated in Supplementary Figure 1.

A web interface allows searches that combine search terms for different feature types (e.g. predicted intracellular location and predicted domain content), and displays data in a form that aids comparisons among orthologs. Annotation efforts are focusing on 'RNA binding proteins' (RBPs), i.e. those proteins predicted to interact with RNA and/or influence RNA function. This protein class is more complex in plants than in metazoa (1–4) and includes several protein families that are largely specific to plants [e.g. the PPR (5) and CRM (6) families]. A prerequisite for a comprehensive understanding of the network of RNA–protein interactions in plants is a catalog of plant RBPs. POGs/PlantRBP provides a new tool for this purpose.

DATABASE CONSTRUCTION

POG assignment

Predicted rice and Arabidopis proteins were assigned to POGs in the following manner. The proteins predicted by all gene models in both species were compared to one another using WU-Blast 2.0 (http://blast.wustl.edu/) (7). For each of the ~90 000 gene models, the top 20 blast hits with an *E*-value <1e−5 were re-aligned using NEEDLE, an implementation of the Needleman–Wunsch algorithm (8) distributed with the EMBOSS package (9). These global

*To whom correspondence should be addressed. Tel: +1 541 346 5145; Fax: +1 541 346 5891; Email: abarkan@molbio.uoregon.edu alignments were used to cluster the gene-models into POGs via the mutual-best-hits strategy summarized in Supplementary Figure 2. This strategy yields both one-to-one and many-to-one orthology relationships. It also yields clusters in which no gene is in more than one cluster. All gene-models for a particular locus were placed in the same POG when at least one of the gene models met the mutual-best-hits criterion.

A phylogenetic approach was used to complement the mutual-best-hit method and aid evaluation of POG assignments (see Supplementary Figure 3). The top blast hits for each protein in a POG that met an *E*-value cutoff of $<1e-5$ (up to 20 such proteins) were examined, and those with >50% coverage (either hit/query or query/hit) were designated 'closely related'. The 'closely related' protein sets for all members of a POG were combined, and this combined set of proteins was used as input to MUSCLE version 3.6 (10) to produce a multiple alignment and corresponding guide tree. Where the tree topology supports the POG assignments, the POG is marked as 'well-supported'. POGs containing only proteins that are not 'closely-related' to any other protein (*E*-value < $1e-5$) in either species are also marked as 'well-supported'. Trees and multiple alignments were stored for later display.

POG annotation

Each predicted protein was analyzed for predicted intracellular localization using Predotar (11), TargetP (12), NucPred (http://www.sbc.su.se/~maccallr/nucpred/), and PredictNLS (13). This set of algorithms provides two independent predictions for targeting to each of the chloroplast (TargetP and Predotar), mitochondrion (TargetP and Predotar), nucleus (PredictNLS and NucPred) and secretory system (TargetP and Predotar). Searches can be performed for proteins predicted by either one or both of the 'redundant' algorithms to localize to a specific compartment. Each predicted protein was analyzed for domain content using InterproScan (14) version 3.3 (http://www.ebi.ac.uk/interpro/index.html) with Pfam (15) version 19 and SuperFamily (16) version 1.69 models. POGs were annotated as 'putative RNA binding' based on a hand-curated list of Interpro domains related to RNA metabolism (Supplementary Table 1), and on entries in the Arabidopsis Splicing Related Genes (ASRG) Database (1) (http://www.plantgdb.org/SRGD/ASRG/). Available maize genomic and expressed sequence tag (EST) assemblies and ESTs from the maize full-length cDNA project (see Source Sequences below) were associated with their closest rice counterpart using BLASTn against rice genomic DNA with an *E*-value cutoff of $< 1e-10$. BLASTn was chosen for this purpose because nucleotide similarity provides greater resolution than amino acid similarity for species as closely-related as maize and rice.

Source Sequences

Sequences incorporated into PlantRBP were derived from the following sources. Arabidopsis sequences are from version 6 of the Arabidopsis genome annotation and were downloaded from TAIR (ftp://ftp.arabidopsis.org/). Rice sequences (version 4) were downloaded from TIGR (ftp://ftp.tigr.org/). Maize genomic assemblies (MAGI version 4) were downloaded from (http://magi.plantgenomics.iastate.edu/), maize EST assemblies from PlantGDB (http://plantgdb.org/) and ESTs associated with the Arizona full-length cDNA project from (http://www.maizecDNA.org/download).

USER INTERFACE

Searches

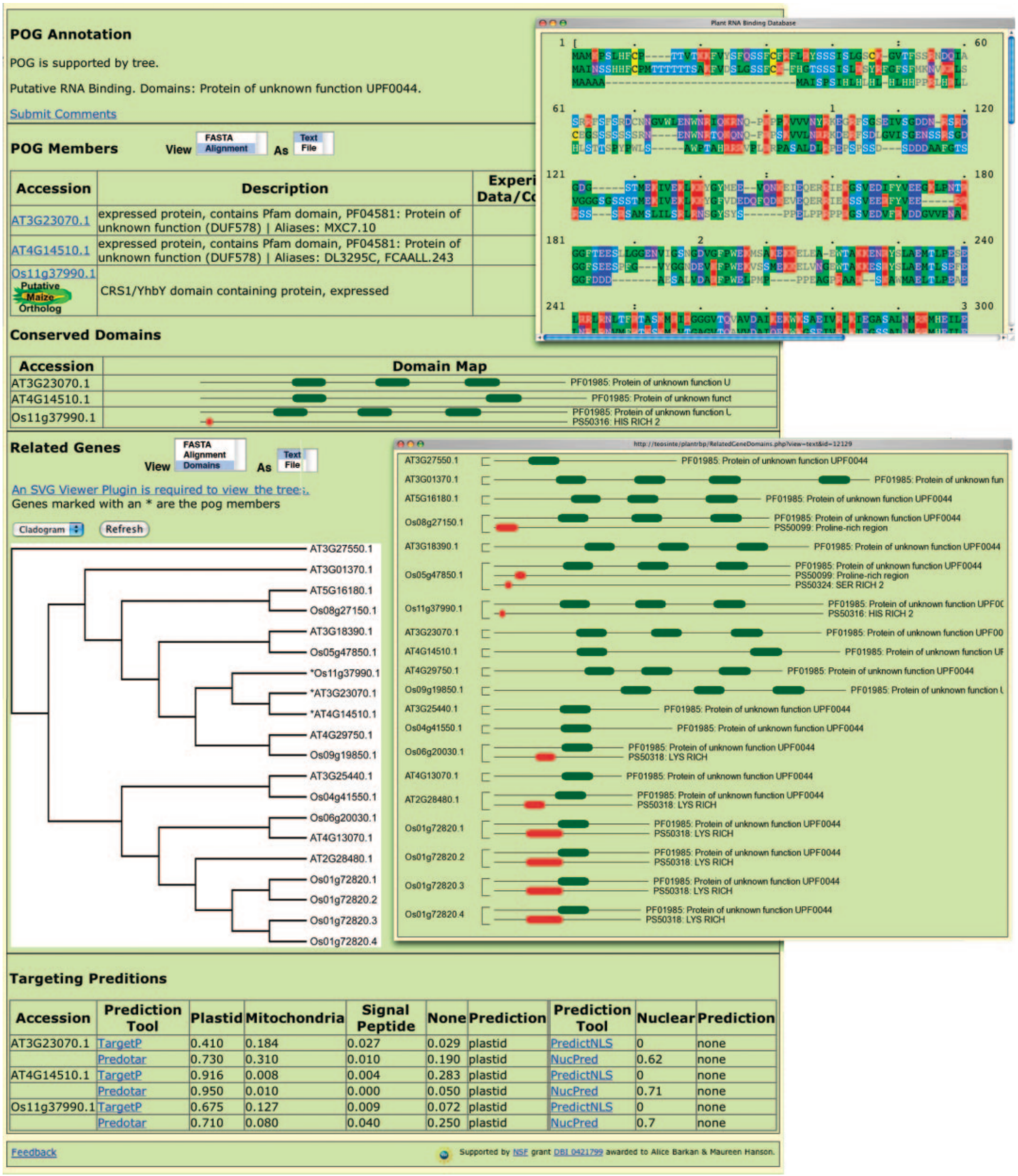
A search page includes fields for searching by gene (gene name, gene description, or gene identifier), domain, and predicted intracellular location. Multiple search criteria entered in the same or different field are linked with 'and', serving to narrow the search. For example, entering 'RRM PPR' under domain, and 'chloroplast' under targeting prediction returns POGs containing proteins with both an RRM and PPR domain that are predicted by either Predotar or TargetP to be targeted to chloroplasts. Searches can be further limited to return proteins predicted by two different algorithms to be targeted to either the chloroplast, mitochondrion, nucleus, or secretory system. A search can be broadened by adding a wild card symbol (*) to a truncated term. For example, typing 'At*' in the Gene field will limit the search to Arabidopsis proteins. A BLAST (7) interface is also available. The BLAST results are parsed and displayed using Bioperl (17). Each entry within the BLAST results is linked to the POG with which the gene is associated.

Search results are returned as a list of POGs that include any proteins that meet the search criteria. POGs in the list are annotated with domains shared by all POG members, protein name (where assigned), and an indication of whether the POG is 'well-supported' according to the criteria described above. Each POG on the list is linked to a more detailed 'POG View' annotation page.

POG View

The POG View page (Figure 1) summarizes information about members of each POG in a format that facilitates comparative analyses. The page starts with a list of the component gene models linked to either TIGR (rice) or TAIR (Arabidopsis), and annotated with gene descriptions derived from TIGR (rice) or TAIR (Arabidopsis). Experimental data and literature citations for the RBP subset will be displayed on this page, after manual curation by POGs curators and with community input (see Future Work). The sequences and multiple alignments of POG members can be viewed or downloaded. The alignments are displayed using Mview (18). For each rice gene model, a pop-up window presents the putative maize ortholog(s) in a graphical alignment (Figure 2).

Below the table of gene models is a graphical representation of the conserved domains in each predicted protein. This display highlights inconsistent gene models and inconsistent domain predictions among orthologs. Each domain is linked to its Interpro page. A phylogenetic tree of rice and Arabidopsis proteins related to POG members (see criteria for 'related' proteins above) is displayed beneath the domain maps. Each gene model in the tree is linked to its respective POG, simplifying navigation to related POGs. The sequences, multiple alignment and domain maps of the related genes can be viewed or downloaded. Finally, a summary of the predictions from the four targeting prediction algorithms is shown in a table at the bottom of the page, in a manner that highlights the consistency, or lack thereof, of those predictions.


Plant RNA Binding Database
1
60
61
120
121
180
181
240
241
300
http://teosinte/plantrbp/RelatedGeneDomains.php?view=text&id=12129
FASTA
Alignment
Domains
Text
File
Cladogram
Refresh
Putative
Maize
Ortholog

POG Annotation

POG is supported by tree.

Putative RNA Binding. Domains: Protein of unknown function UPF0044.

Submit Comments

POG Members View FASTA Alignment As Text File

Accession	Description	Experi Data/Cc
AT3G23070.1	expressed protein, contains Pfam domain, PF04581: Protein of unknown function (DUF578) | Aliases: MXC7.10	
AT4G14510.1	expressed protein, contains Pfam domain, PF04581: Protein of unknown function (DUF578) | Aliases: DL3295C, FCAALL.243	
Os11g37990.1 Putative Maize Ortholog	CRS1/YhbY domain containing protein, expressed	

Conserved Domains

Accession	Domain Map
AT3G23070.1	PF01985: Protein of unknown function U
AT4G14510.1	PF01985: Protein of unknown funct
Os11g37990.1	PF01985: Protein of unknown function L PS50316: HIS RICH 2

Related Genes View FASTA Alignment Domains As Text File

An SVG Viewer Plugin is required to view the tree.

Genes marked with an * are the pog members

AT3G27550.1	PF01985: Protein of unknown function UPF0044
AT3G01370.1	PF01985: Protein of unknown fun
AT5G16180.1	PF01985: Protein of unknown function UPF0044
Os08g27150.1	PF01985: Protein of unknown function UPF0044 PS50099: Proline-rich region
AT3G18390.1	PF01985: Protein of unknown function UPF0044
Os05g47850.1	PF01985: Protein of unknown function UPF0044 PS50099: Proline-rich region PS50324: SER RICH 2
Os11g37990.1	PF01985: Protein of unknown function UPF0C PS50316: HIS RICH 2
AT3G23070.1	PF01985: Protein of unknown function UPF00
AT4G14510.1	PF01985: Protein of unknown function UF
AT4G29750.1	PF01985: Protein of unknown function UPF0044
Os09g19850.1	PF01985: Protein of unknown function L
AT3G25440.1	PF01985: Protein of unknown function UPF0044
Os04g41550.1	PF01985: Protein of unknown function UPF0044
Os06g20030.1	PF01985: Protein of unknown function UPF0044 PS50318: LYS RICH
AT4G13070.1	PF01985: Protein of unknown function UPF0044
AT2G28480.1	PF01985: Protein of unknown function UPF0044 PS50318: LYS RICH
Os01g72820.1	PF01985: Protein of unknown function UPF0044 PS50318: LYS RICH
Os01g72820.2	PF01985: Protein of unknown function UPF0044 PS50318: LYS RICH
Os01g72820.3	PF01985: Protein of unknown function UPF0044 PS50318: LYS RICH
Os01g72820.4	PF01985: Protein of unknown function UPF0044 PS50318: LYS RICH

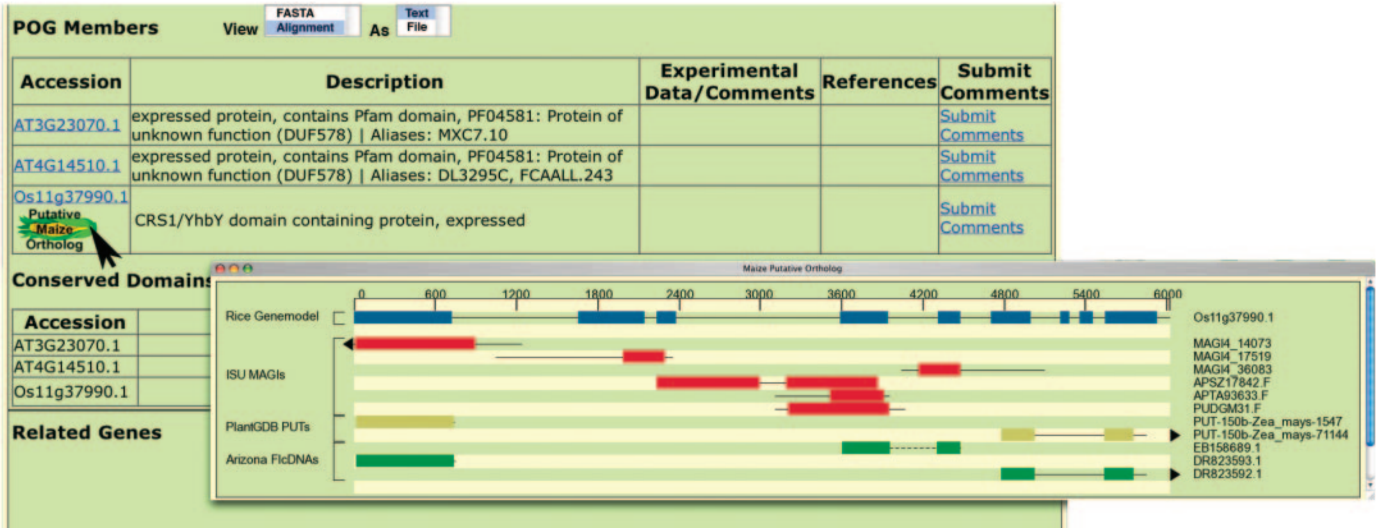
Targeting Preditions

Accession	Prediction Tool	Plastid	Mitochondria	Signal Peptide	None	Prediction	Prediction Tool	Nuclear	Prediction
AT3G23070.1	TargetP	0.410	0.184	0.027	0.029	plastid	PredictNLS	0	none
	Predotar	0.730	0.310	0.010	0.190	plastid	NucPred	0.62	none
AT4G14510.1	TargetP	0.916	0.008	0.004	0.283	plastid	PredictNLS	0	none
	Predotar	0.950	0.010	0.000	0.050	plastid	NucPred	0.71	none
Os11g37990.1	TargetP	0.675	0.127	0.009	0.072	plastid	PredictNLS	0	none
	Predotar	0.710	0.080	0.040	0.250	plastid	NucPred	0.7	none

Feedback

Supported by NSF grant DBI 0421799 awarded to Alice Barkan & Maureen Hanson.

Figure 1. The POG view page. The POG shown includes two Arabidopsis inparalogs and one rice protein, and is supported by the topology of the tree. The information displayed includes (from upper to lower panel) a list of gene models with the rice model linked to putative maize orthologs (see Figure 2), a representation of the conserved domains in POG members, a phylogenetic tree that includes POG members (indicated with an asterisk) and other closely-related proteins, and the results of four targeting prediction algorithms. Pop-up windows provide access to protein alignments for POG members (upper panel), the domain organization of all proteins in the tree (lower panel), and an alignment of all proteins in the tree (data not shown).



POG Members View FASTA Alignment As Text File

Accession	Description	Experimental Data/Comments	References	Submit Comments
AT3G23070.1	expressed protein, contains Pfam domain, PF04581: Protein of unknown function (DUF578) | Aliases: MXC7.10			Submit Comments
AT4G14510.1	expressed protein, contains Pfam domain, PF04581: Protein of unknown function (DUF578) | Aliases: DL3295C, FCAALL.243			Submit Comments
Os11g37990.1 Putative Maize Ortholog	CRS1/YhbY domain containing protein, expressed			Submit Comments

Conserved Domains

Accession
AT3G23070.1
AT4G14510.1
Os11g37990.1

Related Genes

Figure 2. Pop-up window showing putative maize orthologs. Rice loci in each POG are linked to a graphical representation of putative maize orthologs. Maize genomic assemblies (MAGI's at http://magi.plantgenomics.iastate.edu/), EST assemblies (PlantGDB PUTs at http://plantgdb.org/), and ESTs from the maize full-length cDNA project (www.maizecdna.org) are displayed below the TIGR gene model for their best rice hit in BLASTn searches against TIGR's rice pseudochromosomes, version 4.

CONCLUSIONS AND FUTURE WORK

The efficient exploitation of information gleaned with different model organisms, each with unique experimental attributes, requires the identification of orthologous gene sets, representations of data that facilitate comparisons among orthologs, and user-friendly interfaces that aid searches for orthologous groups with specified properties. Previously, two databases have been described that aim to achieve the first of these purposes for model plant species: OrthologID (19) (http://nypg.bio.nyu.edu/orthologid/) uses a rigorous phylogenetic method to predict orthologs in the fully-sequenced genomes of Arabidopsis, rice, and poplar, and displays phylogenetic data for each orthologous group; the Genome Cluster Database (GCD) (http://bioinfo.ucr.edu/projects/GCD) (20) clusters proteins in rice and Arabidopsis based on BLAST similarity and domain organization but does not aim to distinguish orthologs from paralogs. Genes in GCD are linked to protein alignments and expression profiling data. POGs/PlantRBP is distinct from these and other available resources in its use of both mutual-best-hit and phylogenetic analysis to assign orthologous groups, in the types of data that are displayed concurrently for orthologous proteins (domain organization, targeting predictions, literature citations, phylogenetic trees, links to putative maize orthologs), and in the types of searches that are easily accomplished (e.g. searches for orthologous groups containing proteins with both a specified domain content and predicted intracellular location). In addition, this database has a unique emphasis on the annotation of 'RNA binding' proteins, a protein class that is particularly complex in plants (1–5).

Future work will focus on three areas:

(i) An immediate priority is to enhance the annotations of POGs for predicted RBPs. In addition to the annotations available for all POGs (e.g. domains and targeting predictions), annotations for RBPs will include literature citations, mutant phenotypes, and established intracellular locations. Where appropriate, POGs will be linked to subcellular proteome databases [e.g. mitochondria (21) and plastids (22)] and to smaller databases that focus on RBP subsets [e.g. the ASRG Database (1)]. In addition, gene models will be displayed simultaneously for all POG members, to highlight differences in gene prediction. An interface will be provided for community input.

(ii) High quality POG assignments are essential for comparative analyses. Currently, the mutual-best-hit and phylogenetic approaches agree for ∼65% of the POGs; these POGs are designated as 'well-supported'. A priority for the future is to increase the number of well-supported POGs by modifying the strategies used for defining mutual-best-hits and by using more rigorous methods for building phylogenetic trees. Those POGs encoding RBPs and for which the POG is not supported by the tree will be manually curated to resolve the discrepancy. Users with an interest in a POG that is not supported by the tree can assess orthology by downloading the sequences of closely-related proteins from the POG View page, generating and editing global protein alignments and using the edited alignments to calculate trees that include bootstrap values. Users who take this step will be encouraged to share their results through the community annotation interface.

(iii) New maize genome and cDNA sequence data are deposited in public repositories on a frequent basis. To maintain an up-to-date representation of putative maize orthologs, we will develop a pipeline to automate the incorporation of new maize genome and EST sequence data. In addition, the database will be updated with recalculated POGs, protein domains, and targeting predictions when new versions of the rice and Arabidopsis gene sets are released.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank John Robbins, Kenny Watkins, and Rosalind Williams-Carrier for helpful input on database functionalities. This work was supported by grant DBI-0421799 from the National Science Foundation. Funding to pay the Open Access publication charges for this article was provided by grant DBI-0421799 from the National Science Foundation.

Conflict of interest statement. None declared.

REFERENCES

1. Wang,B.B. and Brendel,V. (2004) The ASRG database: identification and survey of *Arabidopsis thaliana* genes involved in pre-mRNA splicing. *Genome Biol*., **5**, R102.
2. Belostotsky,D. (2003) Unexpected complexity of Poly(A)-binding protein gene families in flowering plants: three conserved lineages that are at least 200 million years old and possible auto- and cross-regulation. *Genetics*, **163**, 311–319.
3. Lorkovic,Z. and Barta,A. (2002) Genome analysis: RNA recognition motif (RRM) and K homology (KH) domain RNA-binding proteins from the flowering plant *Arabidopsis thaliana*. *Nucleic Acids Res*., **30**, 623–635.
4. Vermel,M., Guermann,B., Delage,L., Grienenberger,J., Marechal-Drouard,L. and Gualberto,J. (2002) A family of RRM-type RNA-binding proteins specific to plant mitochondria. *Proc. Natl Acad. Sci. USA*, **99**, 5866–5871.
5. Lurin,C., Andres,C., Aubourg,S., Bellaoui,M., Bitton,F., Bruyere,C., Caboche,M., Debast,C., Gualberto,J., Hoffmann,B. *et al*. (2004) Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. *Plant Cell*, **16**, 2089–2103.
6. Barkan,A., Klipcan,L., Ostersetzer,O., Kawamura,T., Asakura,Y. and Watkins,K. (2007) The CRM domain: an RNA binding module derived from an ancient ribosome-associated protein. *RNA*, in press.
7. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.
8. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol*., **48**, 443–453.
9. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*., **16**, 276–277.
10. Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
11. Small,I., Peeters,N., Legeai,F. and Lurin,C. (2004) Predotar: a tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics*, **4**, 1581–1590.
12. Emanuelsson,O. and Heijne,G.V. (2001) Prediction of organellar targeting signals. *Biochim. Biophys. Acta*, **1541**, 114–119.
13. Cokol,M., Nair,R. and Rost,B. (2000) Finding nuclear localization signals. *EMBO Rep*., **1**, 411–415.
14. Quevillon,E., Silventoinen,V., Pillai,S., Harte,N., Mulder,N., Apweiler,R. and Lopez,R. (2005) InterProScan: protein domains identifier. *Nucleic Acids Res*., **33**, W116–W120.
15. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al*. (2004) The Pfam protein families database. *Nucleic Acids Res*., **32**, D138–141.
16. Gough,J., Karplus,K., Hughey,R. and Chothia,C. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol*., **313**, 903–919.
17. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G., Korf,I., Lapp,H. *et al*. (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res*., **12**, 1611–1618.
18. Brown,N.P., Leroy,C. and Sander,C. (1998) MView: a web-compatible database search or multiple alignment viewer. *Bioinformatics*, **14**, 380–381.
19. Chiu,J.C., Lee,E.K., Egan,M.G., Sarkar,I.N., Coruzzi,G.M. and DeSalle,R. (2006) OrthologID: automation of genome-scale ortholog identification within a parsimony framework. *Bioinformatics*, **22**, 699–707.
20. Horan,K., Lauricha,J., Bailey-Serres,J., Raikhel,N. and Girke,T. (2005) Genome cluster database. A sequence family analysis platform for Arabidopsis and rice. *Plant Physiol*., **138**, 47–54.
21. Heazlewood,J.L. and Millar,A.H. (2005) AMPDB: the Arabidopsis Mitochondrial Protein Database. *Nucleic Acids Res*., **33**, D605–D610.
22. Sun,Q., Emanuelsson,O. and van Wijk,K.J. (2004) Analysis of curated and predicted plastid subproteomes of Arabidopsis. Subcellular compartmentalization leads to distinctive proteome properties. *Plant Physiol*., **135**, 723–734.