Bioenerg. Res. (2009) 2:209–216
DOI 10.1007/s12155-009-9052-z

pDAWG: An Integrated Database for Plant Cell Wall Genes

**Fenglou Mao · Yanbin Yin · Fengfeng Zhou ·
Wen-Chi Chou · Chan Zhou · Huiling Chen · Ying Xu**

Published online: 13 October 2009

Abstract We have recently developed a database, pDAWG, focused on information related to plant cell walls. Currently, pDAWG contains seven complete plant genomes, 12 complete algal genomes, along with computed information for individual proteins encoded in these genomes of the following types: (a) carbohydrate active enzyme (CAZy) family information when applicable; (b) phylogenetic trees of cell wall-related CAZy family proteins; (c) protein structure models if available; (d) physical and predicted interactions among proteins; (e) subcellular localization; (f) Pfam domain information; and (g) homology-based functional prediction. A querying system with a graphical interface allows a user to quickly compose information of different sorts about individual genes/proteins and to display the composite information in an intuitive manner, facilitating comparative analyses and knowledge discovery about cell wall genes. pDAWG can be accessed at http://csbl1.bmb.uga.edu/pDAWG/.

Keywords Cell wall genes · Biological database · Bioinformatics · Biofuel

F. Mao · Y. Yin · F. Zhou · W.-C. Chou · C. Zhou · H. Chen ·
Y. Xu (✉)
Computational Systems Biology Lab,
Department of Biochemistry and Molecular Biology,
and Institute of Bioinformatics, University of Georgia,
120 Green St.,
Athens, GA 30602, USA
e-mail: xyn@bmb.uga.edu

F. Mao · Y. Yin · F. Zhou · W.-C. Chou · C. Zhou · H. Chen ·
Y. Xu
DOE BioEnergy Science Center (BESC),
Oak Ridge, TN, USA

Introduction

Considerable funding from federal and private agencies invested into biomass-based energy research in recent years has greatly accelerated the genome sequencing and omic data generation of biofuel-related plants and algae. As a result, at least seven plants and 12 algae have their genomes sequenced, and over 60 additional plants and algal genomes are in the pipeline of being sequenced and assembled. Computational analyses of these genomes and the associated omic data have led to vast amount of computed information about these organisms, which is generally reliable enough to be useful. The advent of the data generation capabilities has led to the development and deployment of a number of databases focused on plants such as PlantGDB [1], Phytozome [2], PlantTribes [3], GreenPhylDB [4], and Phytome [5]. In addition, a few databases specifically for plant cell wall syntheses and remodeling have also been developed. For example, the Cell Wall Navigator database at the UC Riverside [6] and the Cell Wall Gene Families database at Purdue University [7] classify cell wall-related enzymes, structural proteins, and signaling proteins into different families, and the Rice GT database is specifically designed for rice glycosyltransferase families [8].

While each of these databases provides some aspects of cell wall-related information, it has not been easy for a typical biofuel researcher to fully utilize all the information directly stored in individual databases or easily derivable through cross-referencing multiple databases. There are two general issues: (a) most of these databases are located at different locations on the Internet often using different input and output data formats, and (b) there has been lack of easy-to-use interfaces that link these databases to facilitate integrated analyses of cell wall genes, taking into consideration of multiple types of information. There is a clearly and rapidly increasing need for utilizing multiple types of information to improve the reliability of computationally predicted data and to cross-validate predictions, each having only weak supporting evidence. For example, to predict the function of a protein, one often has to consider not only sequence and structure information [9–13] but also phylogeny [14], gene expression [15], and other information [16].

To address this issue, we have recently developed a prototype database, pDAWG, containing multiple types of information all related to plant cell wall studies and equipped with search capabilities allowing easy information integration in an intuitive manner. Our goal was to ultimately develop a one-stop shop database containing all the key information derivable from genomes essential to studies of plant cell wall genes. Currently, the database contains seven complete plant and 12 complete algal genomes along with various types of functional, structural, and evolutionary information about plant cell wall genes, namely, (a) carbohydrate active enzyme (CAZy) family information when applicable; (b) Pfam domain information; (c) phylogenetic trees of cell wall-related CAZy proteins; (d) protein structure models if available; (e) physical and predicted interactions among proteins; (f) subcellular localizations; and (g) homology-based functional prediction. We intend to continue to expand this database by including both new experimental and computational data when they become available. By using the same gene identifier for different types of information in our database, one can easily assemble a composite view covering different aspects of individual cell wall genes as well as comparative views of different genes across multiple genomes, facilitating comparative analyses of cell wall genes.

The pDWAG Database

Currently, the pDAWG database contains the following seven types of information about individual genes in addition to the 19 complete genomes. In addition to the seven types of data, we have also developed an interactive system allowing users to post their comments and suggestions to our database, which we will use to improve the data quality in our database. To ensure quality of the suggestions and comments, only users with approved logins can enter their feedbacks of using our system, while other users can view the comments.

Plant and Algal Genomes

Seven complete plant and 12 complete algal genomes are stored in the pDAWG database, along with their "standard" annotations downloaded from various databases, including annotated genes, gene structures, and protein functions among others. We included 12 algal genomes to facilitate comparative genome analyses related to plant cell wall genes. Among these algal genomes, the six green algae belong to "green plants" or "plantae," while the other six are either plant-like or fungi-like. Table 1 gives the names of the complete genomes included in our database along with the associated information such as the source from where the genome is downloaded. A subset of these genomes has been used in one of our previous papers [17]. We plan to update this portion of the database by adding any new complete plant and algal genomes on a regular basis. For example, in our planned next release of the pDAWG database, we will include genomes of papaya [18] and soybean.

CAZy Family and Pfam Information

We have predicted the members of the CAZy families [19] in the 19 genomes in pDAWG. The procedure we used is as follows. For each CAZy family, we tried to find the corresponding Pfam hidden Markov model (HMM) model. We found that among the 285 CAZy families, 142 have Pfam links provided by the CAZy families. Using RPS-BLAST [20] search against the NCBI CDD database [21], we linked additional 69 CAZy families to Pfam families. Overall, the 211 (142+69) CAZy families are represented by 185 Pfam HMMs (some CAZy families share the same Pfam HMM). Using these 185 HMMs, we searched the 19 complete genomes using HMMER [22] with E value cutoff 1e−2. Out of the 185 HMMs, 122 have hits in the 19 genomes, and 114 have at least two hits, which correspond to 133 CAZy families. For each of the 114 HMMs, we have stored the identified sequence in the FASTA format along with the gene's identifier. In addition, we have also tried to identify the Pfam family for each encoded protein in the 19 genomes by running pfam_scan.pl against Pfam 23.0, which returns one or more Pfam ID for most proteins.

This capability of pDAWG enables a user to easily retrieve the protein list of his/her interested CAZy families in one or multiple genomes. For example, by clicking *CAZy Families* on the pDAWG menu (see "Searching Capability in pDAWG"), a user can go to the page showing the first level of CAZy classification: GT, GH, CE, PL, or CBM; from there, the user can select, say GT, to go to the next page showing all the GT families; by clicking on a particular GT family, say GT2, one can go to the page of GT2 in all the 19 genomes; by clicking on one specific genome, the user can get to the page with all the proteins having the Pfam GT2 domains in the selected genome.

Table 1 19 complete genomes in pDAWG

sn.	Abbreviation	Clade	Species	Genome published/released
1	phr	Stramenopiles	*Phytophthora ramorum*	[19]
2	phs	Stramenopiles	*Phytophthora sojae*	[19]
3	tp	Diatom	*Thalassiosira pseudonana*	[20]
4	pht	Diatom	*Phaeodactylum tricornutum*	[21]
5	aa	Brown tide algae	*Aureococcus anophagefferens*	JGI
6	cm	Red algae	*Cyanidioschyzon merolae*	[22]
7	mpc	Green algae	*Micromonas pusilla CCMP1545*	[23]
8	mpr	Green algae	*Micromonas strain RCC299*	[23]
9	ol	Green algae	*Ostreococcus lucimarinus*	[24]
10	ot	Green algae	*Ostreococcus tauri*	[25]
11	cr	Green algae	*Chlamydomonas reinhardtii*	[26]
12	vc	Green algae	*Volvox carteri f. nagariensis*	JGI
13	pp	Moss	*Physcomitrella patens ssp patens*	[27]
14	sm	Spike moss	*Selaginella moellendorffii*	JGI
15	pt	Dicot	*Populus trichocarpa*	[28]
16	at	Dicot	*Arabidopsis thaliana*	[29]
17	vv	Dicot	*Vitis vinifera*	[30]
18	os	Monocot	*Oryza sativa*	[31, 32]
19	sb	Monocot	*Sorghum bicolor*	[33]

Phylogenetic Trees of Cell Wall Proteins

For each of the 114 Pfam HMMs representing 133 CAZy members with hits in the 19 genomes, we have built a phylogenetic tree among the homologous proteins in the 19 genomes by first building a multiple sequence alignment using MAFFT [23] and then running PhyML [24] to construct a maximum likelihood tree (100 bootstrap replicates were generated). It should be noted that this automatic alignment and tree building process could be improved by editing automatically or manually the multiple sequence alignments [25]. We plan to include an automatic alignment editing procedure by using Gblocks [26] in the next release of pDAWG. Each of the constructed trees gives the phylogeny of a plant CAZy family based on the conserved Pfam domains of the plant or algal proteins of that family. Currently, a small number of families such as the GT2 family are missing from our collection of trees since their family sizes are too large for the tree construction programs that we used. We anticipate that this problem will be solved in the very near future using other less computationally demanding tree construction programs.

By selecting *Phylogeny* in the pDAWG menu, a user can go to the page having links to all the CAZy families with phylogeny built. Clicking on an interested family will lead the user to a new page showing the maximum likelihood tree of the family; all the leave names are clickable linking to the page containing all the annotation information of the gene from other databases.

Protein Structure Models

We have attempted to predict the three-dimensional structure model for each of the proteins with no more than 500aa in length encoded in the 19 genomes in pDAWG using the MUSTER program [27], presumably the best threading-based protein structure prediction program in the public domain. In our prediction, we have used a template library consisting of all the PDB structures [28] sharing <70% sequence identity with any other protein in the library. If a template structure has multiple structural domains, we manually partition the structure into individual domains and include each domain structure in our template library in addition to the whole multi-domain structure. For each protein encoded in the 19 genomes, its threaded structures are ranked by the threading *Z* scores, and the best five structures with *Z* scores >7.5 (empirically determined cutoff for reliable predictions), if any, are input into the MODELLER program [29] to generate the full-length atomic structures, which are stored in the pDAWG database. The protein structure sub-database currently consists of 113,583 structures from 19 genomes, which we consider as generally reliable.

The predicted that structures can be visualized using any molecular visualization software, such as Rasmol and Pymol. The default viewer used in pDAWG development is Rasmol. The pDAWG PDB file can be opened by Rasmol directly from browser if the file extension .pdb is used. Conserved residues or functional sites on protein structures can be displayed on the structures. A user can also compare the structure similarity between two proteins using any structural alignment tool such as TM-align [30], which is in the process of being integrated into the pDAWG toolkit.

PPI Data

While we have constructed protein–protein interaction (PPIs) for proteins encoded in the 19 genomes, we currently have PPIs at a proteomic scale only for *Arabidopsis* and are in the process of adding the proteomic scale PPI data for Populus. The PPIs for *Arabidopsis* were generated by integrating four PPI sets. Two were constructed based on literature search and downloaded from the TAIR ftp site and from the IntAct database [31], respectively. The other two were downloaded from the computationally predicted interactome for *Arabidopsis* [32] and from the *Arabidopsis thaliana* Protein Interactome Database [33], respectively. The PPI data for the other 18 genomes were retrieved from the DOMINE database [34], which were all experimentally validated. Table 2 summarizes the amount of PPI data for each of the 19 genomes.

The PPI data in pDAWG can be accessed in two ways: through *Genomes* or through *Protein–Protein Interaction* listed in the main menu of pDAWG. Through the *Genomes* option, one can select a specific genome by clicking on the name of the genome; a list of genes will be displayed, and then clicking on the link “PPI” under “Other” column, then the user will be able to see a list of interacted proteins. By clicking on the menu *Protein–Protein Interaction* and then clicking on the genome name, the user will be able to see the displayed interaction map using the SVG viewer plug-in, which requires the installation of this plug-in on the user’s browser.

Protein Subcellular Localization

We have predicted the subcellular locations for each protein in the pDAWG database using three prediction programs, PSORT[35], Predotar [36], and SubLoc [37]. We also provided subcellular localization for *A. thaliana* and *Oryza sativa* by using their species-specific predictor AtSubP [38] and RSLpred [39], respectively. We assign a subcellular location prediction to a protein only if at least two programs provide the prediction for the protein. Currently, 98.6% of proteins have assigned subcellular localization information across the 19 genomes. Figure 1 shows the protein distributions of 11 subcellular localizations across 19 genomes. The localization information gives a user the spatial information about where proteins and specific biological pathways may function. A user can browse the subcellular localization information through the *Subcellular Localization* or *Genomes* menu.

EST Data

We have downloaded all the available expressed sequence tag (EST) data along with the associated experimental

Table 2 PPI data for the 19 genomes in pDAWG

sn.	Abbreviation	Species	Number of PPI	Number of proteins having PPI data
1	phr	*Phytophthora ramorum*	5,929	633
2	phs	*Phytophthora sojae*	8,894	877
3	tp	*Thalassiosira pseudonana*	2,313	388
4	pht	*Phaeodactylum tricornutum*	1,761	319
5	aa	*Aureococcus anophagefferens*	11,767	764
6	cm	*Cyanidioschyzon merolae*	839	225
7	mpc	*Micromonas pusilla CCMP1545*	2,238	413
8	mpr	*Micromonas strain RCC299*	1,967	393
9	ol	*Ostreococcus lucimarinus*	1,693	317
10	ot	*Ostreococcus tauri*	1,392	297
11	cr	*Chlamydomonas reinhardtii*	4,342	601
12	vc	*Volvox carteri f. nagariensis*	4,326	671
13	pp	*Physcomitrella patens ssp patens*	14,950	1,129
14	sm	*Selaginella moellendorffii*	25,290	1,560
15	pt	*Populus trichocarpa*	36121	2,316
16	at	*Arabidopsis thaliana*	45,059	13,346
17	vv	*Vitis vinifera*	15,547	1,316
18	os	*Oryza sativa*	40,508	2,141
19	sb	*Sorghum bicolor*	18,654	1,334

conditions from PlantGDB [1] for all the seven land plants and three green algae in pDAWG so a user can easily find out if a gene might be expressed under some conditions. These ESTs are actually longer UniGenes assembled by PlantGDB and thus should be in a better quality than raw EST sequences. Other EST data will be added to pDAWG once they become available. The EST data can be accessed through the gene information page for each gene where one can submit the protein sequence to do BLAST search against the EST database; from the resulting BLAST result page, the user can click the gi number of the hit EST linking to NCBI to find out its expression location and so on.

Searching Capability in pDAWG

In addition, the intuitive browsing capability in pDAWG provides a search capability for a user to directly get the information he/she may be interested through two search options: (a) search by specifications and (b) search by sequence.

Search by Specification

pDAWG provides a “Query Builder” which can be accessed from the menu *Search for Genes* to facilitate a user to search its database. The Query Builder currently provides 16 categories of specifications such as *gene name*, *species name*, *Pfam domain*, *CAZy family*. Under each of these categories, a user can input a name, value, or even an expression (or leaving it blank). The user can link the specified conditions using logic operations like “AND” or “OR”. For example, a user can compose a query accomplishing the following: *find all proteins in CAZY GT8 family in Arabidopsis located in Golgi* through the Query Builder. pDAWG converts such queries into SQL queries, searches its database, and then returns the search results on a result page. The detailed information about constructing a query through Query Builder is provided in the *Tutorial* page of pDAWG.

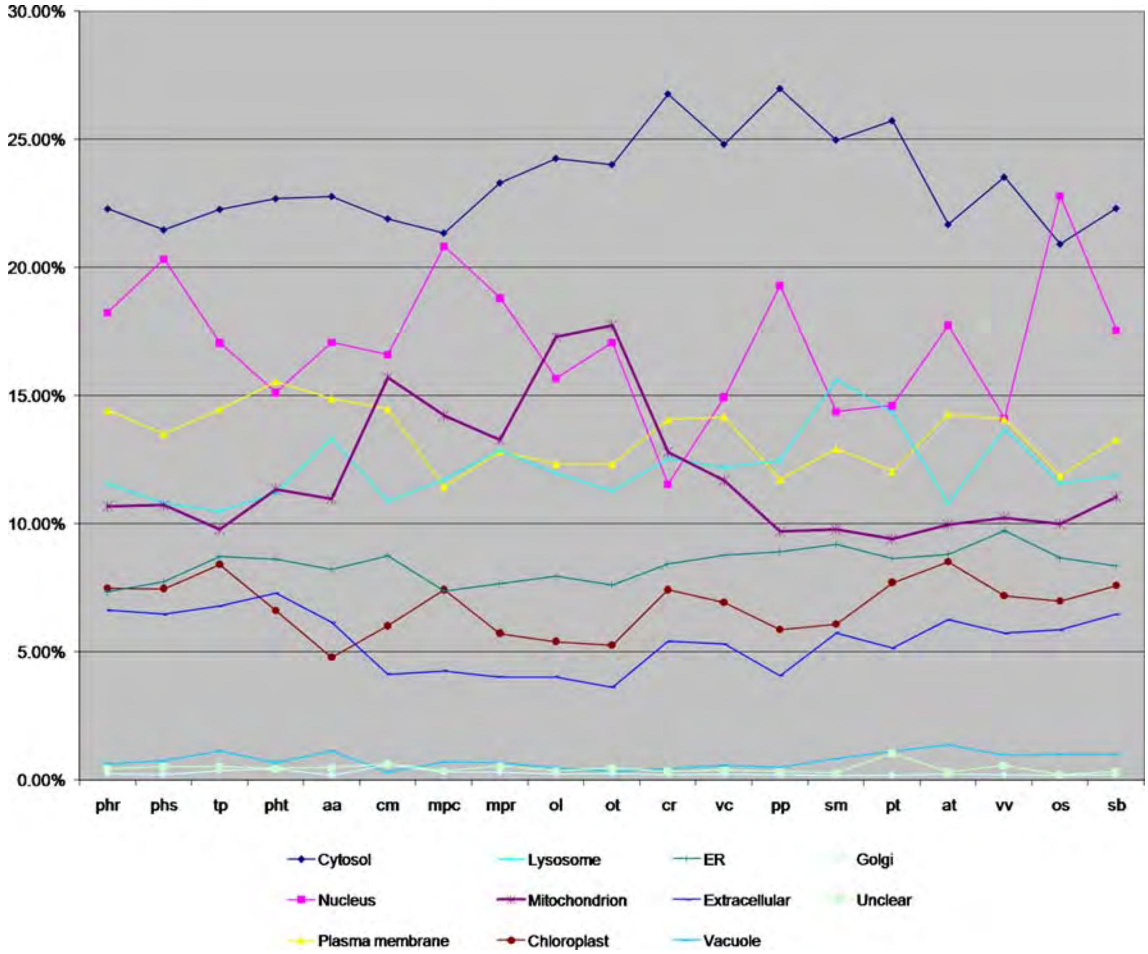

Fig. 1 Distributions of subcellular localizations across 19 genomes predicted by PSORT. The x-axis represents the genome axis and the y-axis represents the relative abundance of proteins in each of 11 major locations (color-coded)

Search by Sequence

pDAWG also provides a search interface for a user to find available information in pDAWG about a specific gene. To do this, a user will need to click the menu *BLAST* and enter the sequence of the gene and then select a database from a collection of user-specified databases, including the protein sequence database for all pDAWG genes, the protein sequence databases for individual organisms, and the EST databases and then click Submit to run BLAST. For detailed information about this search capability, we refer the reader to the *Tutorial* page of the pDAWG.

pDAWG Interface and Implementation

All functions provided by pDAWG can be accessed from the menu bar on top of each pDAWG page. The menu items (other than *Tutorial*) are in two categories: browsing menu items and searching menu items. The first six menu items *Genomes*, *CAZy families, Subcellular Localization, Phylogeny*, and *Protein–Protein Interaction* can lead the user to browse the corresponding database. The menu items *Search for Genes* and *BLAST* provide the capability to search the database.

The pDAWG database is implemented on a Fedora 8 linux computer, the web server is apache 2.2.8, the database is MySQL 5.0.45, and PHP is used as the script language to generate dynamic webpages.

Application Examples

We use the following example to showcase how the pDAWG data can help identify putatively novel cell wall-related genes. Our query is an *A. thaliana* gene: AT5G20260 of the GT47 family, which has been shown to be involved in both xylan and pectin biosynthesis [40, 41]. A user can go to the search page, type in "AT5G20260," and click "Submit." The search result is shown in a table where the first column is the gene ID and the last column is the relevant information retrieved from the database, such as the PPI, the subcellular localization, the CAZy family to which the query gene belongs. If clicking on the "PPI" link in the column, the user will be directed to a new page with a table listing all the protein–protein interaction pairs involving the query gene. In this case, this gene has two interaction partners: AT5G25310, also a GT47 gene, and AT3G55830, a GT64 gene. Interestingly, the latter is implicated to play a role in pectin synthesis [42],

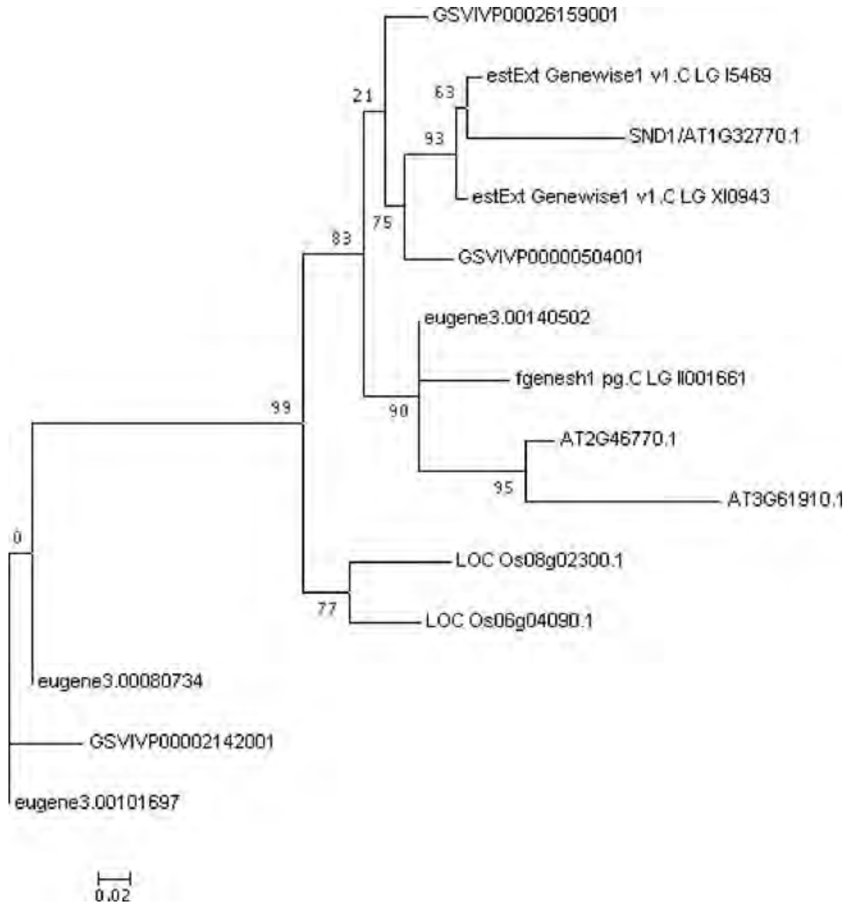

Fig. 2 Phylogenetic tree generated from top 14 BLAST hits in pDAWG for Arabidopsis SND1 gene

and our PPI data indicate an interaction between this gene and a GT47 gene, supporting the hypothesis that this GT64 gene may be a novel cell wall-related gene. Nevertheless, further experimental evidences are clearly needed to confirm this.

We now use another example to illustrate how to use pDAWG and its search capability. The example is to find the Poplar ortholog of the *Arabidopsis* SND1 protein, a key transcription factor controlling the secondary cell wall synthesis [43]. One way to do this through pDAWG is to find its plant homologs first and then build a phylogeny and then predict the homolog that is phylogenetically closest to the *Arabidopsis* SND1 to be its ortholog.

Step 1. Go to the pDAWG BLAST page; copy and paste the protein sequence of SND1 into the text box in the page; choose the database *All pDAWG genomes*. Select 1e−6 in the *e* value box as the BLAST *E* value cutoff; and click the Submit button. In this case, over 500 hits are found and only the top 500 hits are shown. This number is too large for the current phylogeny reconstruction program. To reduce the number, the user can consider only the close homologs by selecting 1e−70 as the *E* value cutoff and repeat the above procedure. Now, 14 hits are returned.

Step 2. Clicking on the gene name of each hit will show the detailed information about that gene. Clicking the first clickable "here" on the top of the table will show the Blast output file including the sequence alignment. Clicking the second clickable "here" will save a file. In this case, the box asks whether to save the 14 hits in one FASTA file; click "save."

Step 3. Click "this website" to link to the webpage "www.phylogeny.fr," a phylogenetic tree construction web site for non-specialists [44]. The system will run a predefined workflow consisting of a multiple tree construction tools using default parameters if the user clicks on "one click." A dialog webpage will appear on the screen; click on "browse" and then choose the FASTA file that was just saved; enter an e-mail address and then click Submit. It will return a tree in minutes.

Figure 2 shows a tree constructed using the above procedure. In this case, the most likely Poplar ortholog of the query *Arabidopsis* SND1 is the gene estExt_Genewise1_v1.C_LG_I5469. By going through the above procedure, phylogenetic trees can be generated for all homologous proteins of the query in 19 plant and algal genomes.

Acknowledgment This work is supported in part by the BioEnergy Science Center (BESC) grant from the Office of Biological and Environmental Research in the DOE Office of Science and National Science Foundation (DBI-0354771, ITR-IIS-0407204, DBI-0542119).

References

1. Duvick J et al (2008) PlantGDB: a resource for comparative plant genomics. Nucl Acids Res 36(suppl_1):D959–D965
2. JGI (2009) Phytozome: a tool for green plant comparative genomics. Available from http://www.phytozome.net/
3. Wall PK et al (2008) PlantTribes: a gene and gene family resource for comparative genomics in plants. Nucleic Acids Res 36 (Database issue):D970–D976
4. Conte MG et al (2008) GreenPhylDB: a database for plant comparative genomics. Nucleic Acids Res 36(Database issue): D991–D998
5. Hartmann S et al (2006) Phytome: a platform for plant comparative genomics. Nucleic Acids Res 34(Database issue):D724–D730
6. Girke T et al (2004) The Cell Wall Navigator database. A systems-based approach to organism-unrestricted mining of protein families involved in cell wall metabolism. Plant Physiol 136 (2):3003–3008 discussion 3001
7. Yong W et al (2005) Genomics of plant cell wall biogenesis. Planta 221(6):747–51
8. Cao P-J et al (2008) Construction of a rice glycosyltransferase phylogenomic database and identification of rice-diverged glycosyltransferases. Molecular Plant 1(5):858–877
9. Punta M, Ofran Y (2008) The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function. PLoS Comput Biol 4(10):e1000160
10. Minshull J et al (2005) Predicting enzyme function from protein sequence. Curr Opin Chem Biol 9(2):202–209
11. Watson JD, Laskowski RA, Thornton JM (2005) Predicting protein function from sequence and structural data. Curr Opin Struct Biol 15(3):275–284
12. Lee D, Redfern O, Orengo C (2007) Predicting protein function from sequence and structure. Nat Rev Mol Cell Biol 8(12):995–1005
13. Mazumder R, Vasudevan S (2008) Structure-guided comparative analysis of proteins: principles, tools, and applications for predicting function. PLoS Comput Biol 4(9):e1000151
14. Sjolander K (2004) Phylogenomic inference of protein molecular function: advances and challenges. Bioinformatics 20(2):170–179
15. Brown DM et al (2005) Identification of novel genes in *Arabidopsis* involved in secondary cell wall formation using expression profiling and reverse genetics. Plant Cell 17(8):2281–2295
16. Nariai N, Kolaczyk ESD, Kasif S (2007) Probabilistic protein function prediction from heterogeneous genome-wide data. PLoS One 2(3):e337
17. Yin Y, Huang J, Xu Y (2009) The cellulose synthase superfamily in fully sequenced plants and algae. BMC Plant Biol 9(1):99
18. Ming R et al (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature 452 (7190):991–996
19. Cantarel BL et al (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. Nucleic Acids Res 37(Database issue):D233–D238
20. Altschul SF et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402
21. Marchler-Bauer A et al (2009) CDD: specific functional annotation with the Conserved Domain Database. Nucleic Acids Res 37 (Database issue):D205–D210
22. Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14(9):755–763
23. Katoh K et al (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33(2):511–518
24. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52(5):696–704

25. Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 56(4):564–577
26. Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17(4):540–552
27. Wu S, Zhang Y (2008) MUSTER: Improving protein sequence profile–profile alignments by using multiple sources of structure information. Proteins 72(2):547–556
28. Berman HM et al (2000) The protein data bank. Nucleic Acids Res 28(1):235–242
29. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234(3):779–815
30. Zhang Y, Skolnick J (2005) TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 33(7):2302–2309
31. Kerrien S et al (2007) IntAct—open source resource for molecular interaction data. Nucleic Acids Res 35(Database issue):D561–D565
32. Geisler-Lee J et al (2007) A predicted interactome for *Arabidopsis*. Plant Physiol 145(2):317–329
33. Cui J et al (2007) AtPID: *Arabidopsis thaliana* protein interactome database—an integrative platform for plant systems biology. Nucleic Acids Res 36(Database issue):D999–D1008
34. Raghavachari B et al (2008) DOMINE: a database of protein domain interactions. Nucleic Acids Res 36(Database issue):D656–D661
35. Nakai K, Horton P (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem Sci 24(1):34–35
36. Small I et al (2004) Predotar: a tool for rapidly screening proteomes for N-terminal targeting sequences. PROTEOMICS 4 (6):1581–1590
37. Chen H, Huang N, Sun Z (2006) SubLoc: a server/client suite for protein subcellular location based on SOAP. Bioinformatics 22 (3):376–377
38. Kaundal R, Zhao PX (2009) AtSubP: the *Arabidopsis* subcellular localization prediction server. Available from http://bioinfo3.noble.org/AtSubP/index.html
39. Kaundal R, Raghava GPS (2009) RSLpred: an integrative system for predicting subcellular localization of rice proteins combining compositional and evolutionary information. PROTEOMICS 9 (9):2324–2342
40. York WS, O'Neill MA (2008) Biochemical control of xylan biosynthesis—which end is up? Curr Opin Plant Biol 11(3):258–265
41. Mohnen D (2008) Pectin structure and biosynthesis. Curr Opin Plant Biol 11(3):266–277
42. Singh SK et al (2005) Cell adhesion in *Arabidopsis thaliana* is mediated by ECTOPICALLY PARTING CELLS 1—a glycosyltransferase (GT64) related to the animal exostosins. Plant J 43 (3):384–397
43. Zhong R, Demura T, Ye ZH (2006) SND1, a NAC domain transcription factor, is a key regulator of secondary wall synthesis in fibers of Arabidopsis. Plant Cell 18(11):3158–3170
44. Dereeper A et al (2008) Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res 36(Web Server issue):W465–W469