Wang et al. BMC Plant Biology 2010, 10:14
http://www.biomedcentral.com/1471-2229/10/14

BMC
Plant Biology

DATABASE Open Access

SoyDB: a knowledge database of soybean transcription factors

Zheng Wang[1], Marc Libault[2,3], Trupti Joshi[1,2], Babu Valliyodan[2,3], Henry T Nguyen[2,3], Dong Xu[1,2,4], Gary Stacey[2,3], Jianlin Cheng[1,2,4*]

Abstract

Background: Transcription factors play the crucial rule of regulating gene expression and influence almost all biological processes. Systematically identifying and annotating transcription factors can greatly aid further understanding their functions and mechanisms. In this article, we present SoyDB, a user friendly database containing comprehensive knowledge of soybean transcription factors.

Description: The soybean genome was recently sequenced by the Department of Energy-Joint Genome Institute (DOE-JGI) and is publicly available. Mining of this sequence identified 5,671 soybean genes as putative transcription factors. These genes were comprehensively annotated as an aid to the soybean research community. We developed SoyDB - a knowledge database for all the transcription factors in the soybean genome. The database contains protein sequences, predicted tertiary structures, putative DNA binding sites, domains, homologous templates in the Protein Data Bank (PDB), protein family classifications, multiple sequence alignments, consensus protein sequence motifs, web logo of each family, and web links to the soybean transcription factor database PlantTFDB, known EST sequences, and other general protein databases including Swiss-Prot, Gene Ontology, KEGG, EMBL, TAIR, InterPro, SMART, PROSITE, NCBI, and Pfam. The database can be accessed via an interactive and convenient web server, which supports full-text search, PSI-BLAST sequence search, database browsing by protein family, and automatic classification of a new protein sequence into one of 64 annotated transcription factor families by hidden Markov models.

Conclusions: A comprehensive soybean transcription factor database was constructed and made publicly accessible at http://casp.rnet.missouri.edu/soydb/.

Background

Soybean is a great source of protein, as it contains significant amounts of all the essential amino acids, including some that cannot be synthesized by the human body [1]. Soybean has been used as a food and a drug component in China for thousands of years [2] and over the past 60 years has become a leading crop in many nations around the world [3]. Because of its high value in the agricultural and food industry, soybean has received greater and greater research attention, both to improve soybean agronomic performances and as a model for basic biological studies. In early 2008, the Department of Energy-Joint Genome Institute (DOE-JGI) finished sequencing the soybean genome using a whole-genome shotgun approach [4], which makes soybean the most complex plant so far ever sequenced [5]. The homology-based gene prediction and annotation produced putative protein sequences [4,5], which makes it feasible to identify and annotate soybean transcription factors.

Transcription factors (TF) are proteins that bind to DNA sequences (*i.e.*, promoters) and regulate gene expression by one or more DNA binding domains. Virtually all biological processes are directly regulated or influenced by transcription factors [6]. For example, the transcription process in eukaryotes would not occur in the absence of a specific class of transcription factors named "general transcription factors" [7,8]. Studies have shown that transcription factors are closely involved in the process of cell development, such as cellular division, migration, and differentiation [9]. Transcription

* Correspondence: chengji@missouri.edu
[1]Computer Science Department, University of Missouri, Columbia, MO 65211, USA

factors of *Arabidopsis thaliana* have been well studied since its genome has been fully sequenced as a model specie [6,10-14]. This makes it possible to identify and study transcription factors of other newly sequenced species, such as soybean, by homology searching and comparative analysis.

Several databases for soybean genome analysis have been built and made publicly available, such as SoyGD [15], SoyBase [16], and SoyXpress [17]. These databases contain a variety of information, such as soybean genome sequences, bacterial artificial chromosome (BAC), expressed sequence tags (EST), and some useful tools including genome browsers, BLAST searching, and pathway searching. However, these databases only contain general annotations for the soybean genome, instead of knowledge specifically targeting the transcription factors. For example, none of them systematically organizes transcription factors into families or clearly points out the DNA binding domains. PlantTFDB [18] and DBD [19] are two existent transcription factor databases, which contain knowledge about transcription factors from multiple species. For each soybean transcription factor, PlantTFDB contains information including protein sequence, Gene Ontology annotation [20], putative binding domains found by InterProScan [21], and cross-links to external databases, including EMBL [22], UniProt [23], RefSeq [24], and TRANSFAC [25]. DBD contains the amino acid sequence of each transcription factor and external links to Ensembl [26], Pfam [27], and SUPERFAMILY [28]. Compared to PlantTFDB, DBD has less external database links, but DBD claims to contain the transcription factors of 927 completely sequenced genomes whereas PlantTFDB covers 22 species. The knowledge in these two databases is very useful; however, they were built based on a relatively older version of soybean sequence data and their annotations are still incomplete. The most important component they lack is the three dimensional structure for each transcription factor, because the visualization of the transcription factor, especially binding sites, can help further understanding the mechanism and functions of transcription factors, which is indispensible to structural genomics [29,30]. Furthermore, with the complete genome sequences of more and more species available, a computer system is needed that can automatically generate a knowledge database and publish it with a user-friendly interface.

To fill the gap, we developed SoyDB - a comprehensive and integrated database for soybean transcription factors. This database not only contains most of the content and features already existed in PlantTFDB and DBD, but also extends them by containing more comprehensive knowledge and links to more versatile external datasets. The annotations in SoyDB include predicted tertiary structures, protein domains, multiple sequence alignments, DNA binding sites, and web logos and consensus sequences for each family. The SoyDB database also contains links to the homologous EST sequences, and the same or homologous proteins in other databases including PlantTFDB, PDB [31], Swiss-Prot [32], TAIR [33], RefSeq [24], SMART [34], Pfam [27], KEGG [35], SPRINTS [36], EMBL [37], InterPro [38], PROSITE [39], and Gene Ontology [20].

Moreover, our system can automatically execute bioinformatics tools and generate annotations, link to other well-known protein databases, construct MySQL databases, and generate PHP scripts to build its website. This fully automated approach can be used to create a protein annotation database and website for any sequenced organism in the future.

Construction and Content

Database Overview

SoyDB contains the annotations of 5,671 putative transcription factors. These proteins were classified into 64 families (for details see the section “Transcription Factor Family Prediction Using SAM Hidden Markov Models”). Figure 1 illustrates the architecture of the SoyDB website. Users can access the main components from the home page: full-text search, PSI-BLAST sequence search, family classification by hidden Markov model, family browsing, protein browsing, family information, protein information, and FTP site.

Data Source

The soybean genome sequences and gene model predictions used in this study were acquired from the publicly available database Phytozome [5]. These sequences were generated by the preliminary GenomeScan [40], FgenesH [41], and PASA [42] gene annotations based on the Gm1.01 version of soybean assembly data [4].

Transcription Factor Identification

We used the standalone versions of InterProScan [21] to search all the soybean protein sequences against 11 databases integrated in InerPro [38]. These databases and their corresponding scanning methods include: PROSITE (*pfscan*) [39], PRINTS (*FingerPRINTScan*) [43], Pfam (*HMMPfam*) [27], ProDom (*ProDomBlast3i*) [44], SMART (*HMMSmart*) [34], TIGRFAMs (*HMMTigr*) [45], PIR SuperFamily (*HMMPIR*) [46], SUPERFAMILY (*superfamily*) [47], Gene3D (*gene3d*) [48], PANTHER (*HMMPanther*) [49], and HAMAP (*pfscan*) [50]. InterProScan systematically searches each of these databases using their corresponding scanning methods to find domains. The proteins predicted to contain TF related domain(s) were considered as putative transcription factors. Using the Plant Transcription Factor Database (PlnTFDB) [51] and the classification of *Medicago truncatula* TF genes [52] as references, we

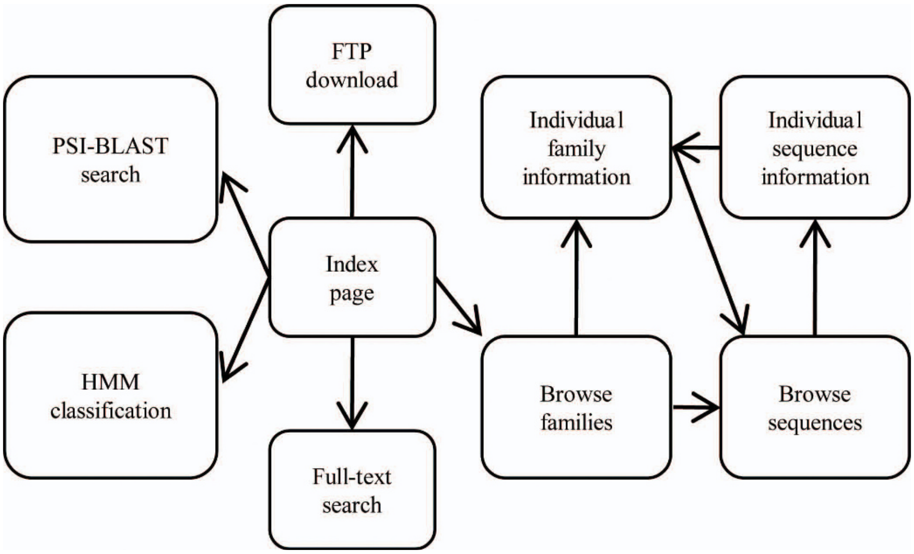


Figure 1 Architecture of SoyDB website. Main annotation components include family information page and sequence information page. Main functional components include full-text search, HMM family classification, PSI-BLAST sequential search, and FTP downloading.

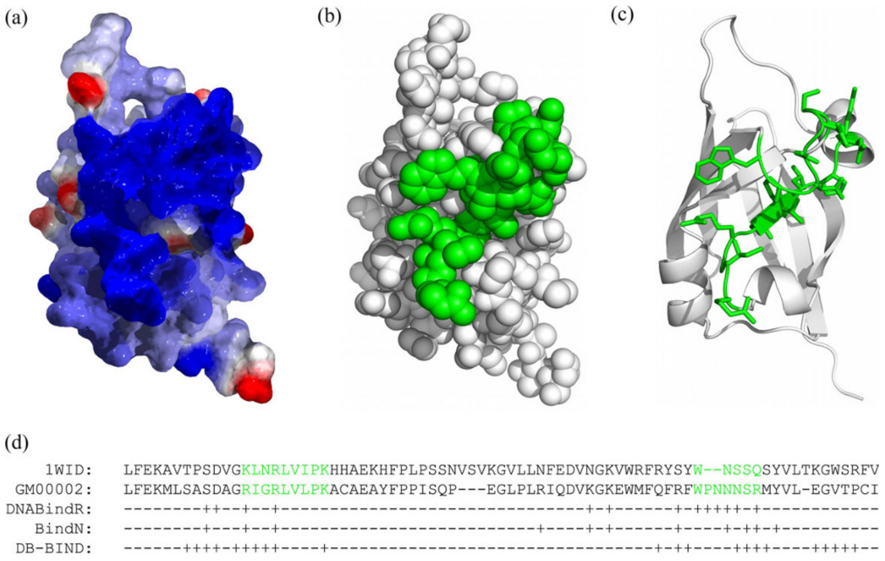
manually curated the list of putative transcription factors and eliminated any mistakenly identified sequences. In this way, we identified 5,671 putative TF sequences.

Transcription Factor Family Prediction Using SAM Hidden Markov Models

The transcription factors of *Arabidopsis thaliana* have been well studied and classified into 64 families [33]. This provides a model for us to classify soybean transcription factors. We used MUSCLE [53] to generate a multiple sequence alignment for each *Arabidopsis thaliana* TF family. The multiple sequence alignment was then input into SAM 3.5 [41] to build a hidden Markov model (HMM) for each family. Every soybean TF sequence was aligned with each of the 64 HMMs, which outputs an e-value. This e-value can be considered as a fitness score between a TF sequence and a hidden Markov model: lower e-value indicates better fitness. Finally, a transcription factor was classified into the family whose HMM yields the lowest e-value.

Annotations Using Bioinformatics Tools

The standalone versions of several bioinformatics tools were locally installed and executed to generate annotations for soybean transcription factors. An accurate protein structure prediction tool MULTICOM [54] was used to predict the tertiary structure of each transcription factor when homologous template proteins could be found in PDB. According to the official evaluations of the 8th community-wide Critical Assessment of Techniques for Protein Structure Prediction (CASP8) [55], MULTICOM was able to predict high-accuracy tertiary structures with an average GDT-TS score [56] 0.87 if suitable templates can be found. GDT-TS score ranges from 0 to 1 measuring the similarities between the predicted and experimental structures, whereas 1 indicates completely the same and 0 completely different. Figure 2 illustrates the predicted tertiary structure of a transcription factor in SoyDB with ID GM00002, and the electrostatic polarization of the predicted structure. The blue area in the electrostatic polarization shows residues positively charged, which is found to be highly identical to the green area in Figure 2(b), which is the putative DNA-binding sites identified by a pair-wise alignment between GM00002 and its template protein 1WID (Figure 2(d)). Since it has been studied and found that the DNA-binding area is positively charged if analyzed by electrostatic potentials [57], the highly identical area in Figure 2(a) and 2(b) strongly confirms that the predicted structure has the electrostatic properties of a transcription factor. This further confirms the qualities of MULTICOM predictions and the correctness of the predicted binding sites derived from the homology alignment. In SoyDB, a predicted tertiary structure is visualized by Jmol [58]. In order to clearly visualize the tertiary structure of the DNA-binding region, only the segments containing homologous DNA binding domains are visualized by Jmol. Users can view a TF structure from various perspectives in a three-dimensional way and perform many operations including selecting and highlighting interested regions, changing view styles and colors, and measuring atom distances and angles by right clicking on the Jmol console and selecting corresponding menus. Detailed instructions about Jmol menus can be found at Jmol website [58]. During the structure prediction process, MULTICOM generates the sequence



```
     1WID:   LFEKAVTPSDVGKLNRLVIPKHHAEKHFPLPSSNVSVKGVLLNFEDVNGKVWRFRYSYW--NSSQSYVLTKGWSRFV
  GM00002:   LFEKMLSASDAGRIGRLVLPKACAEAYFPPISQP---EGLPLRIQDVKGKEWMFQFRFWPNNNSRMYVL-EGVTPCI
DNABindR:   --------++--+--+-------------------------------+-+------+-+++++-+------------
    BindN:   --------+---+--+--------------------------+----+-+------+----+-++-+----------
  DB-BIND:   ------++++-+++++----+---------------------------------+-++----++++----+++++--
```

Figure 2 The predicted structure of a transcription factor in SoyDB. The electrostatic polarization (a) (blue, positive; red, negative), sphere (b) and ribbon (c) visualizations of MULTICOM predicted structure for GM00002. (d) a segment of the pair-wise alignment between 1WID (PDB template of GM00002) and GM00002, and, below, the DNA-binding site predictions from three independent tools: DNABindR [70], BindN [71], and DP-Bind [72] ("+" indicates predicted DNA-binding positions, "-" indicates gap or no prediction). The green regions in the sequence of 1WID are the DNA-binding regions identified by experimental methods [73]. The green regions in GM00002 sequence are the two DNA-binding regions derived from the alignment with 1WID. The predicted DNA-binding regions in GM00002 are illustrated in green in (b) and (c). (c) the side chains of the predicted binding regions. (a), (d), and (c) are in the same orientation. The electrostatic polarization (a) was computed and mapped to protein surface by Swiss-PDB viewer (deep view) [74], and the structures in sphere (b) and ribbon styles (c) were made with PyMol [75].

alignments between the transcription factor and its homologous templates using PSI-BLAST. These sequence alignments can be used to predict the binding sites of a transcription factor based on the experimentally determined binding sites of its template as shown in Figure 2.

A predicted structure was parsed into domains by Protein Domain Parser (PDP) [59]. Since some transcription factors did not have homologous templates found in PDB, DOMAC [60], an accurate *ab initio* domain prediction tool, was also used to predict domains for each transcription factor.

The protein sequences in the same family were aligned by MUSCLE [53] and visualized by WebLogo [61]. A consensus sequence was derived from the multiple sequence alignment by selecting the most frequently appeared amino acid at each position. The multiple sequence alignments were also used to identify the conserved signatures (likely the DNA binding domains) for each family.

All of the bioinformatics tools incorporated to construct SoyDB can be used to automatically annotate other species in the future.

Links to External Databases and Datasets

In order to annotate the functions of soybean transcription factors, each TF protein sequence was searched against the soybean TF database PlantTFDB, NCBI known EST sequences, and other general protein databases by PSI-BLAST or TBLASTN. The external protein databases include Swiss-Prot [32], TAIR [33], RefSeq [24], SMART [34], Pfam [27], KEGG [35], SPRINTS [36], EMBL [37], InterPro [38], PROSITE [39], and Gene Ontology [20]. Web links to these databases were created when the same transcription factor or its homologous proteins were found in them; and for each database or EST dataset only the PSI-BLAST or PBLASTN hit with the smallest e-values was listed in SoyDB. To search the known EST sequences, PSI-BLAST was first used to build a position-specific score matrix for each transcription factor. TBLASTN was then used to search each protein sequence against three known EST datasets: EST human, EST mouse, and EST others. GenBank [62] web page of each EST hit was linked to SoyDB website. The gene expression of a subset of TF genes (about 1,000 TF genes) was recently published [63]. Transcription profile of all soybean TF genes in various conditions is under investigation.

These external links greatly expand the annotation scope of SoyDB providing related knowledge from various perspectives. SoyDB provides a systematic view of a transcription factor – from the features of the protein itself, to the biological pathway it locates in. The links to the external databases and datasets can be updated by a re-run of PSIBLAST and TBLASTN. Currently, these links are scheduled to be updated once every six months. This time interval can be changed if necessary.

Database and Website Implementation

The programs used to automatically annotate proteins were written in PERL. The relational database was built on MySQL with database schemas automatically generated by programs written in PERL. The website was implemented in PHP. The database and web site were automatically constructed by computer programs with little human intervention.

Utility and Discussion

Protein Information

This component contains the complete annotations for each transcription factor, including protein ID, protein name and description, tools used for TF identification, family ID, family name and description, amino acid sequence, homology domain prediction, *ab initio* domain prediction, PDB homologous templates, and predicted tertiary structure. This component can be reached by clicking the sequence ID, such as GM00001, or the Phytozome protein name, such as Glyma01g11670.1, at the "Protein Browsing" webpage (for details see the following "Protein Browsing" section). Figure 3 illustrates the protein information page. The sequence ID and family ID, such as GM00001 and GMF0001, are internal indices used by the SoyDB, and the sequence name is the standard soybean TF name used by the soybean genome database Phytozome [5]. We noticed the trend of unifying annotation formats within the soybean community. Therefore, the commonly used TF ID format, such as PTGm00009.1, is also compatible in SoyDB. Details are described in the "Full-Text Search" section below.

Family Information

This component contains the complete annotations for each TF family, including family ID, family name and description, number of sequences within the family, consensus sequence, consensus signatures (likely the DNA binding regions), web logo of the signature profile, and multiple sequence alignment of the protein sequences within the family. Figure 4 demonstrates a family information web page. This component can be reached from the "Family Browsing" web page.

Protein Browsing

The transcription factors within a family are listed in the order of sequence IDs. The list contains the thumbnail of tertiary structure, protein ID and name, family ID, and family name of each transcription factor (Figure 5). Users can further view the complete annotations by clicking its sequence ID or the Phytozome protein name. This component can be reached by clicking the number of sequences in the "Family Browsing" or the "Family Information" web page.

Family Browsing

A user can browse SoyDB from TF family perspective. The 64 TF families are listed in the order of family IDs. The family ID, family name, and the number of transcription factors within each family are listed. By clicking the family ID or name, users can view the complete annotations for a family, or further browse the sequences within a family by clicking the number of sequences. This component can be reached by clicking "Browse Database" in both the top and bottom menu bars from any SoyDB web pages. Additional file 1 (**Figure S1**) illustrates the web page showing a TF family list.

Full Text Search

This component allows users to search the entire SoyDB database by a query text, such as protein name or family name. Given input keywords, SoyDB searches all the fields in the database and returns matched transcription factors with links to their annotations. Users can also search SoyDB by the TF IDs used in PlantTFDB. The search component will return the homologous soybean TFs found in SoyDB.

PSI-BLAST Sequence Search

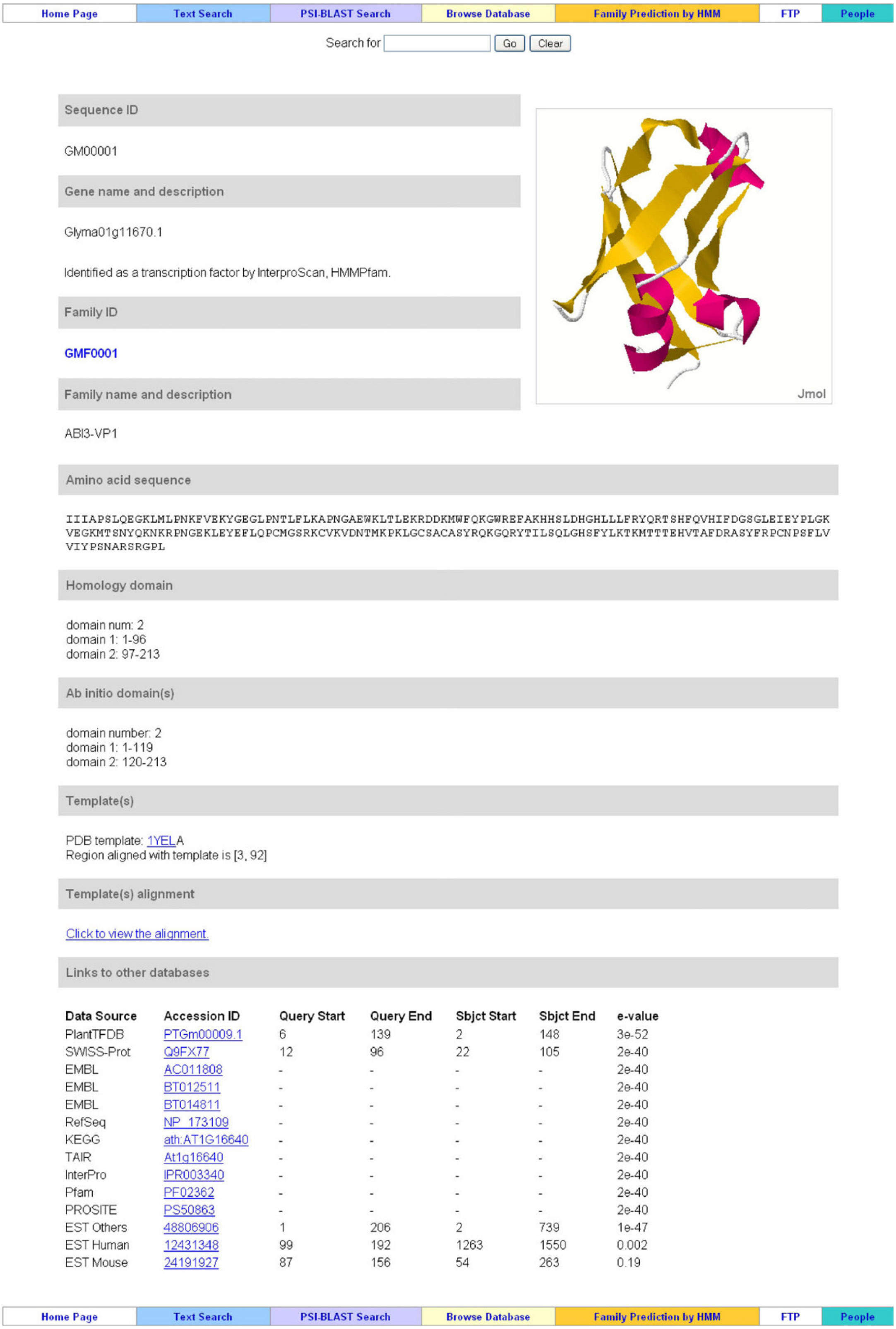
This component allows users to search a query sequence against every TF sequence stored in SoyDB. Users can submit a query sequence and adjust PSI-BLAST parameters from a web page as shown in Additional file 2 (**Figure S2**). After a PSI-BLAST search is performed, the significant hits, with links to their annotation web pages, are ranked based on the e-values generated by PSI-BLAST. Additional file 3 (**Figure S3**) illustrates a PSI-BLAST result web page.

Family Classification by Hidden Markov Model

This component classifies a query protein sequence into one of the 64 TF families. Additional file 4 (**Figure S4**) illustrates the web page for family classification. A submitted query sequence is aligned with each of the 64 hidden Markov models built based on the 64 *Arabidopsis thaliana* TF families. The query sequence is classified into a family whose hidden Markov model outputs the lowest e-value (correspondingly the highest alignment score or fitness score). More details about family classification can be found at the "Transcription Factor Family Prediction Using SAM Hidden Markov Models" section under "Construction and Content".

FTP Site

All of the information in SoyDB is available for users to download from an FTP site. For example, users can download all of the TF protein sequences in the FASTA format and the multiple sequence alignments for each family in plain text. This makes it possible for other websites to link to SoyDB by performing PSI-BLAST



Sequence ID

GM00001

Gene name and description

Glyma01g11670.1

Identified as a transcription factor by InterproScan, HMMPfam.

Family ID

GMF0001

Family name and description

ABI3-VP1

Amino acid sequence

```
IIIAPSLQEGKLMLPNKFVEKYGEGLPNTLFLKAPNGAEWKLTLEKRDDKMWFQKGWREFAKHHSLDHGHLLLFRYQRTSHFQVHIFDGSGLEIEYPLGK
VEGKMTSNYQKNKRPNGEKLEYEFLQPCMGSRKCVKVDNTMKPKLGCSACASYRQKGQRYTILSQLGHSFYLKTKMTTTEHVTAFDRASYFRPCNPSFLV
VIYPSNARSRGPL
```

Homology domain

domain num: 2
domain 1: 1-96
domain 2: 97-213

Ab initio domain(s)

domain number: 2
domain 1: 1-119
domain 2: 120-213

Template(s)

PDB template: 1YELA
Region aligned with template is [3, 92]

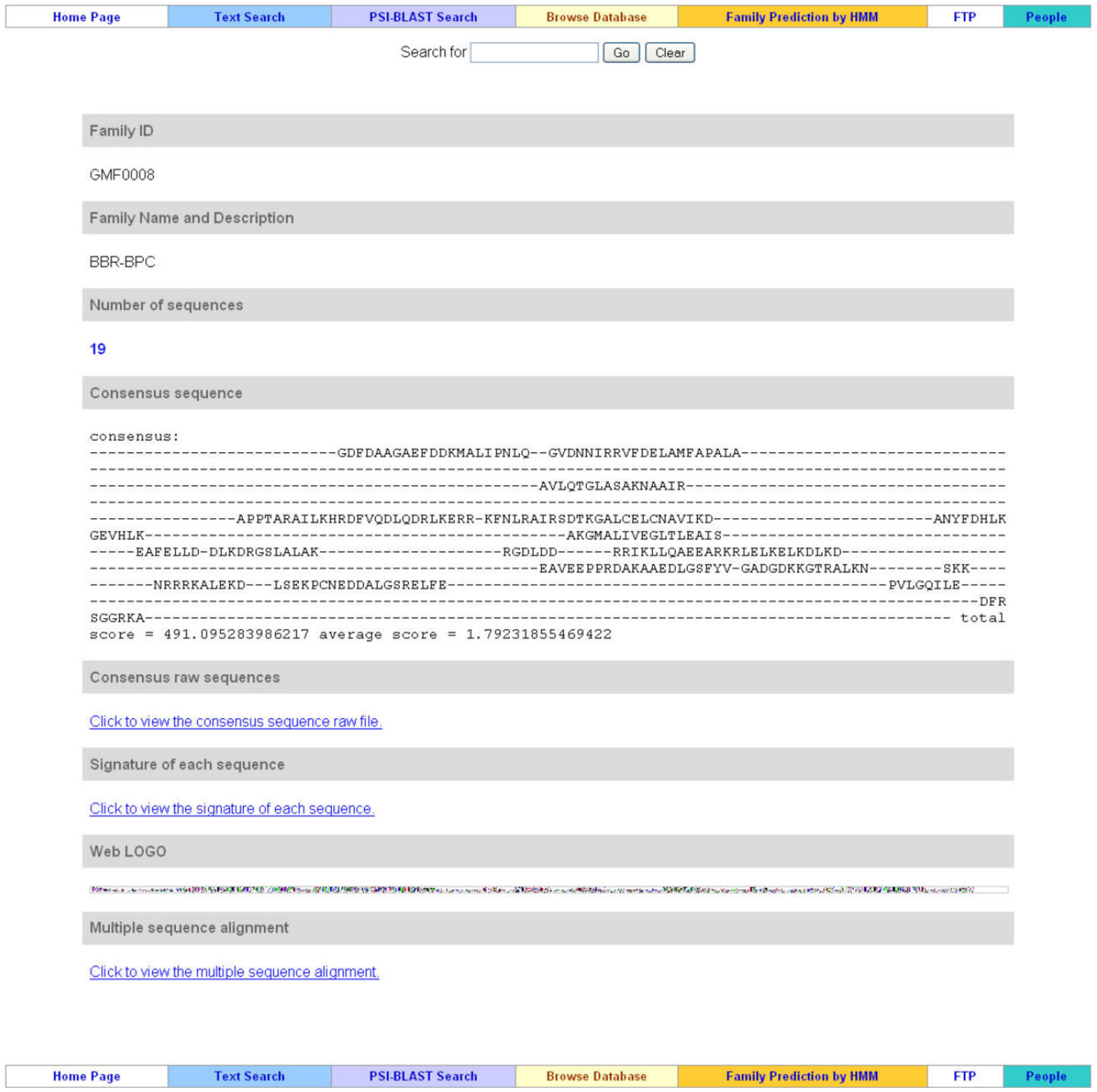
Template(s) alignment

Click to view the alignment.

Links to other databases

Data Source	Accession ID	Query Start	Query End	Sbjct Start	Sbjct End	e-value
PlantTFDB	PTGm00009.1	6	139	2	148	3e-52
SWISS-Prot	Q9FX77	12	96	22	105	2e-40
EMBL	AC011808	-	-	-	-	2e-40
EMBL	BT012511	-	-	-	-	2e-40
EMBL	BT014811	-	-	-	-	2e-40
RefSeq	NP_173109	-	-	-	-	2e-40
KEGG	ath:AT1G16640	-	-	-	-	2e-40
TAIR	At1g16640	-	-	-	-	2e-40
InterPro	IPR003340	-	-	-	-	2e-40
Pfam	PF02362	-	-	-	-	2e-40
PROSITE	PS50863	-	-	-	-	2e-40
EST Others	48806906	1	206	2	739	1e-47
EST Human	12431348	99	192	1263	1550	0.002
EST Mouse	24191927	87	156	54	263	0.19

Figure 3 Information page for a transcription factor. This example web page shows the knowledge for each transcription factor, which includes amino acid sequence, predicted tertiary structure, domain(s) found by homologous search and *ab initio* prediction, PDB template and alignment, and links to other protein databases and EST datasets.



Family ID

GMF0008

Family Name and Description

BBR-BPC

Number of sequences

19

Consensus sequence

```
consensus:
---------------------------GDFDAAGAEFDDKMALIPNLQ--GVDNNIRRVFDELAMFAPALA-----------------------------
----------------------------------------------------------------------------------------------------
-------------------------------------------------AVLQTGLASAKNAAIR-----------------------------------
----------------------------------------------------------------------------------------------------
----------------APPTARAILKHRDFVQDLQDRLKERR-KFNLRAIRSDTKGALCELCNAVIKD------------------------ANYFDHLK
GEVHLK----------------------------------------------AKGMALIVEGLTLEAIS-------------------------------
-----EAFELLD-DLKDRGSLALAK--------------------RGDLDD------RRIKLLQAEEARKRLELKELKDLKD------------------
-------------------------------------------------EAVEEPPRDAKAAEDLGSFYV-GADGDKKGTRALKN--------SKK----
-------NRRRKALEKD---LSEKPCNEDDALGSRELFE------------------------------------------------PVLGQILE-----
-------------------------------------------------------------------------------------------------DFR
SGGRKA---------------------------------------------------------------------------------------- total
score = 491.095283986217 average score = 1.79231855469422
```

Consensus raw sequences

Click to view the consensus sequence raw file.

Signature of each sequence

Click to view the signature of each sequence.

Web LOGO

Multiple sequence alignment

Click to view the multiple sequence alignment.

Figure 4 Family information page. This example web page shows the knowledge for each TF family, which includes number of sequences in the family, consensus sequence of the family, signature of sequences in the family, web logo, and multiple sequence alignment of the sequences in the family.

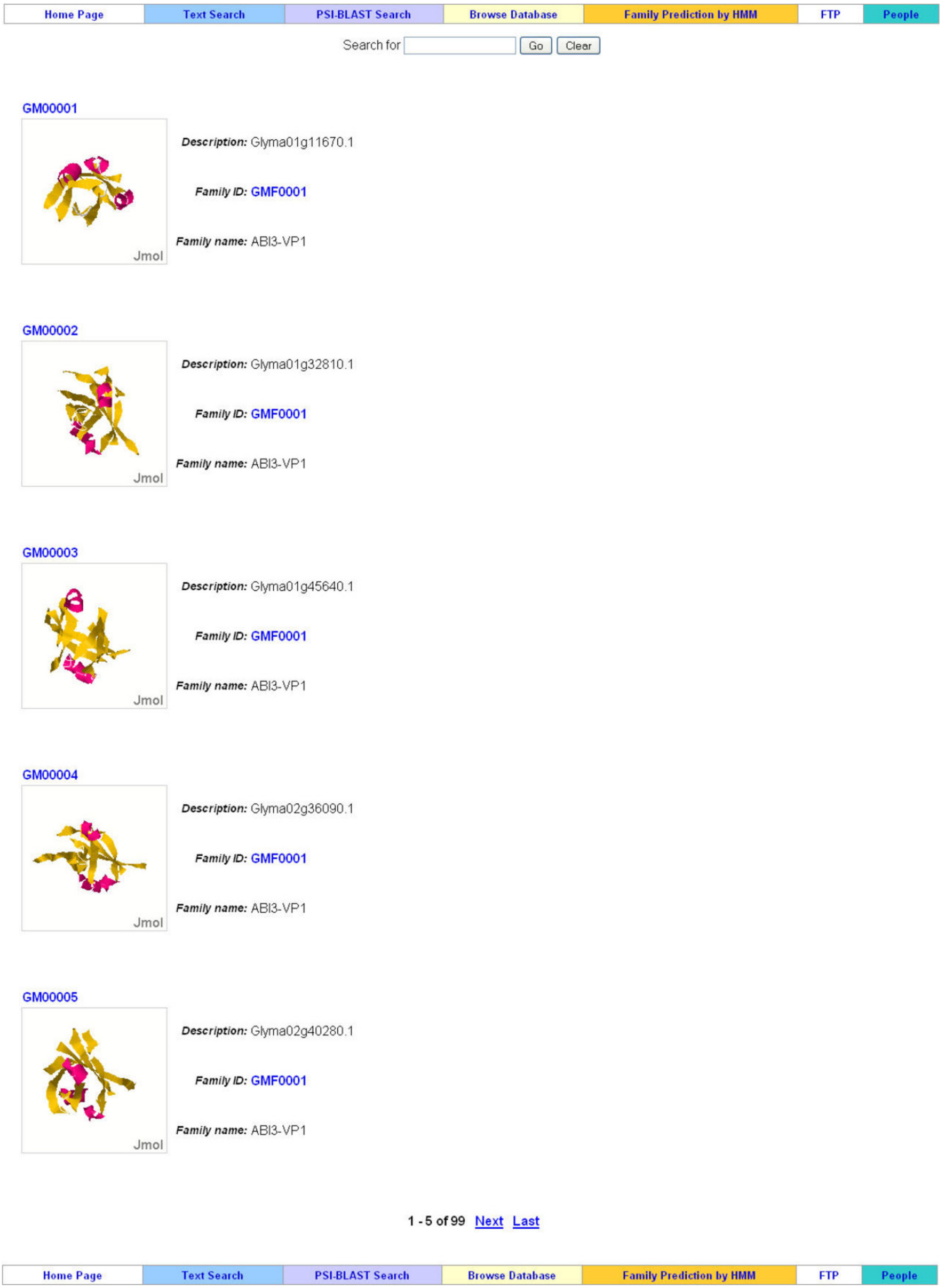
searches on SoyDB sequences, similarly as SoyDB links with other external databases.

Comparisons and Overlapping between SoyDB and PlantTFDB

In total, SoyDB has 5,671 transcription factors - 4,306 of them (75.9%) have hits found in PlantTFDB identified by PSI-BLAST with an e-value threshold 10^{-3}. PlantTFDB has 1,891 soybean transcription factors (based on the FASTA file downloaded from PlantTFDB FTP site), and 1,805 of them (95.5%) have hits found from SoyDB based on a PSI-BLAST search with an e-value threshold 10^{-3}.

Comparisons of Soybean Transcription Factor Family Distributions with Other Plants

The collection and analyses in SoyDB allows us to perform comparisons of TF family distributions across the plant kingdom. The large number of soybean TF genes (5,671) described in this study is likely due to the two soybean whole genome duplication events; one estimated to have occurred 40-50 million years ago (mya)



GM00001

Description: Glyma01g11670.1

Family ID: GMF0001

Family name: ABI3-VP1

GM00002

Description: Glyma01g32810.1

Family ID: GMF0001

Family name: ABI3-VP1

GM00003

Description: Glyma01g45640.1

Family ID: GMF0001

Family name: ABI3-VP1

GM00004

Description: Glyma02g36090.1

Family ID: GMF0001

Family name: ABI3-VP1

GM00005

Description: Glyma02g40280.1

Family ID: GMF0001

Family name: ABI3-VP1

1 - 5 of 99 Next Last

Figure 5 Transcription factor browsing page. This page lists the transcription factors in a TF family. The tertiary structure of each sequence is displayed in an interactive way, *i.e.*, users can zoom in/out and rotate the structure. Users can further view sequence annotations by clicking the TF IDs, and view family annotations by clicking family names.

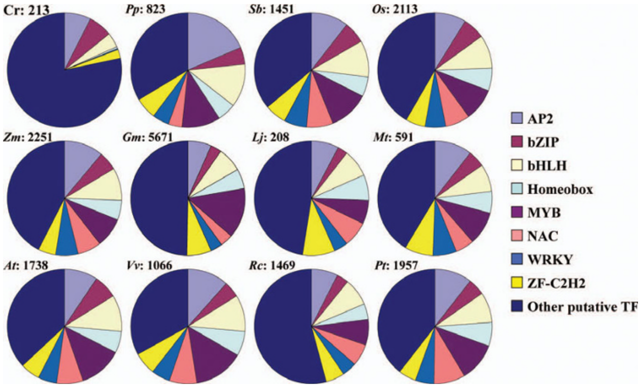


Figure 6 Distributions of transcription factor families across major plant species. Phytozome and DBD databases were mined to identify transcription factor genes in soybean (Gm: Glycine max) and in the 11 remaining plant species, respectively (Cr: Chlamydomonas reinhardtii; Pp: Physcomitrella patens; Sb: Sorghum bicolor; Os: Oryza sativa; Zm: Zea mays; Lj: Lotus japonicus; Mt: Medicago truncatula; At: Arabidopsis thaliana; Vv: Vinis vinifera; Rc: Ricinus communis; Pt: Populus trichocarpa). After being classified based on their family membership, nine major TF families are represented for each plant species. Numbers next to the plant name abbreviation are the total number of TF genes available in DBD. Details are available in Table 1.

and the most recent one approximately 10-15 million years ago [64,65]. By comparing the total number of genes in different organisms, it was found that the increase of plant gene number is related to multicellularity and ploidy. For example, compared to the unicellular eukaryote *Chlamydomonas reinhardtii* where 15,143 genes were predicted [66], larger numbers of protein-encoding genes were reported in multicellular plant organisms, *e.g.*, *Physcomitrella patens* (35,938; [67]), *Arabidopsis thaliana* (32,944; TAIR [33]), and the tetraploid *Glycine max* (66,153, Phytozome). We hypothesize that TF gene number also follows the same trend as land plants, which have a larger number of TF genes compared to algae. To perform comparisons of plant TF genes and their distributions across TF gene families, we mined the last updated DBD database [19]

Table 1 Distributions of transcription factor families across major plant species

	At	Zm	Os	Gm	Lj	Mt	Sb	Pt	Pp	Cr	Vv	Rc
AP2	162	251	186	381	16	63	153	207	153	16	124	111
bZIP	116	119	130	176	6	27	89	90	38	14	48	50
bHLH	183	207	203	393	16	47	148	172	102	9	110	112
homeobox	105	121	132	319	15	35	81	129	44	1	74	66
MYB	212	192	193	791	14	56	165	210	89	0	151	105
NAC/NAM	132	149	146	208	15	31	111	174	32	0	81	92
WRKY	89	141	123	197	8	39	92	103	37	1	59	58
ZF-C2H2	98	114	117	395	19	49	88	101	51	5	67	78
Other TF	641	957	883	2823	99	244	524	771	277	167	352	797
Total	1738	2251	2113	5671	208	591	1451	1957	823	213	1066	1469

%	At	Zm	Os	Gm	Lj	Mt	Sb	Pt	Pp	Cr	Vv	Rc
AP2	9%	11%	9%	7%	8%	11%	11%	11%	19%	8%	12%	8%
bZIP	7%	5%	6%	3%	3%	5%	6%	5%	5%	7%	5%	3%
bHLH	11%	9%	10%	7%	8%	8%	10%	9%	12%	4%	10%	8%
homeobox	6%	5%	6%	6%	7%	6%	6%	7%	5%	0%	7%	4%
MYB	12%	9%	9%	14%	7%	9%	11%	11%	11%	0%	14%	7%
NAC/NAM	8%	7%	7%	4%	7%	5%	8%	9%	4%	0%	8%	6%
WRKY	5%	6%	6%	3%	4%	7%	6%	5%	4%	0%	6%	4%
ZF-C2H2	6%	5%	6%	7%	9%	8%	6%	5%	6%	2%	6%	5%
Other TF	37%	43%	42%	50%	48%	41%	36%	39%	34%	78%	33%	54%
Total	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

for 11 plant species (*C. reinhardtii, P. patens, Oryza sativa, Zea mays, Sorghum bicolor, Lotus japonicum, Medicago truncatula, A. thaliana, Vinis vinifera, Ricinus communis*, and *Populus trichocarpa*). These species were then compared with the soybean TF genes stored in our SoyDB database.

Our analysis showed that the unicellular *C. reinhardtii* has the lowest number of TF genes compared to multicellular land plants (the exceptions are *L. japonicus* and *M. truncatula* where only a partial genome sequence is available). This trend also reflects the differences of total gene number in the organisms shown in Figure 6. For example, it is interesting to note that homeobox, MYB, NAC, and WRKY TF genes in *C. reinhardtii* lack or have very low representations compared to the 11 other plant models (Table 1). Previous studies defined a role for homeobox [68] and WRKY genes [13] in plant development. Therefore, the occurrence of these genes only in multicellular plants may reflect their special roles in development. In addition, a close relationship between TF gene number and total gene number [69] is observed when comparing the TF gene numbers of *G. max* and *A. thaliana* with their total gene numbers (*i.e.*, *G. max* encodes 66,153 protein-coding genes including 5,683 TF genes; *A. thaliana* encodes 32,944 protein-coding genes and 1,738 TF genes). Thus, the family distribution of soybean TF genes is similar to other land plant species, except for *P. patens* (*e.g.*, AP2 represents 7% of total TF genes in soybean vs. 8-12% for other land plants; bZIP: 3% vs. 3-7%; bHLH: 7% vs. 8-11%; homeobox: 6% vs. 4-7%; MYB: 14% vs. 7-14%; NAC: 4% vs. 4-9%; WRKY: 3% vs. 4-7%; ZF-C2H2: 7% vs. 5-9%) (Figure 6 and Table 1).

Collectively, these results suggest that soybean TF genes were not lost following soybean genome duplication, and may have evolved for specialized functions in plant development or response to the environment.

Future Development Plan

In the future, we plan to link to more soybean database, such as SoyBase, and add a human expert discussion section for each transcription factor where biologists can register, log in, and make comments on any annotation items. Also, we plan to link the protein name, such as Glyma01g11670.1, listed in each protein information page to its entry in Phytozome. By doing this, SoyDB can be linked with other soybean genome annotations. Furthermore, we may identify the binding regions on the soybean DNA sequences, which can further help biologists target the regulated regions on soybean genome.

Conclusions

SoyDB is a comprehensive database for soybean transcription factors. It integrates bioinformatics tools and various external databases to provide rich annotations, which can be browsed and retrieved through convenient web interfaces. The automated process generates annotations and creates database and website, and can be used to annotate other sequenced species.

Availability and Requirements

SoyDB is freely available at http://casp.rnet.missouri.edu/soydb/ for academic use. Based on our test, SoyDB is fully functional with three web browsers: Mozilla Firefox, Internet Explorer, and Safari, and four operating systems: Windows XP, Windows Vista, Linux (Red Hat), and Mac OS. The only system requirement for SoyDB is that JAVA runtime environment (JRE) needs to be installed and set fully functional in order to make Jmol work.

Additional file 1: Figure S1 The SoyDB web page showing a list of transcription factor families. TF families are shown with their family ID, family name, and number of sequences within the family. Click on the family ID can further view the detailed information about the family as shown in Figure 4, and click on the number of sequences can open the webpage showing all the transcription factors within the family, as shown in Figure 5.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2229-10-14-S1.PDF]

Additional file 2: Figure S2 The PSI-BLAST search web page. Users can paste or type in a query amino acid sequence and specify PSI-BLAST parameters on the web page. Click on the "Run" button will execute PSI-BLAST.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2229-10-14-S2.PDF]

Additional file 3: Figure S3 The result web page of PSI-BLAST search. PSI-BLAST result page shows the hit TF sequence ID, and the PSI-BLAST score and E-value. The hits are listed in a decreasing order of the PSI-BLAST score. Click on the sequence ID can open the web page showing detailed TF information, as shown in Figure 3.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2229-10-14-S3.PDF]

Additional file 4: Figure S4 The HMM family classification web page. Users can paste or type in a query amino acid sequence. Click on the "Predict" button will execute family classification by HMM.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2229-10-14-S4.PDF]

Acknowledgements
This work was partially funded by a University of Missouri (MU) start-up grant, a MU research board grant, and a MU research council grant to JC. The worked conducted by ML and GS was funded by a grant from the National Science Foundation (Plant Genome Program, #DBI-0421620). Other funding includes grants from the Missouri Soybean Merchandising Council to HN, DX, and GS.

Author details
[1]Computer Science Department, University of Missouri, Columbia, MO 65211, USA. [2]Christopher S Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA. [3]Division of Plant Sciences, National Center for Soybean Biotechnology, Christopher S Bond Life Sciences Center, University

of Missouri, Columbia, MO 65211, USA. [4]Informatics Institute, University of Missouri, Columbia, MO 65211, USA.

Authors' contributions

JC conceived the idea of automating the process of protein annotation and database construction, and coordinated the project. JC and ZW designed the database and formulated the content of the annotations. JC developed or executed the tools that generate structures, templates, domains, DNA binding sites, and multiple sequence alignments. ZW implemented the computer system that automatically gathers data, constructs database, builds web site, and links to external databases and data sources. ZW, TJ, ML and JC performed the soybean TF family classification. ZW, JC, ML, and TJ wrote the first draft of the manuscript. ML analyzed and compared the distributions of transcription factors in multiple species. TJ and ML identified the list of transcription factors. GS, DX, HN, and BV contributed some annotation ideas. JC, GS, DX, and HN provided financial support for the project. All edited and approved the manuscript.

Received: 11 August 2009
Accepted: 18 January 2010 Published: 18 January 2010

References

1. Henkel J: **Soy: health claims for soy protein, questions about other components.** *FDA consumer* 2000, **34.**
2. Han B-Z, Rombouts FM, Nout MJR: **A Chinese fermented soybean food.** *International Journal of Food Microbiology* 2001, **65(1-2)**:1-10.
3. Carpenter J, Gianessi L: **Agricultural biotechnology: updated benefit estimates.** Washington, USA National Centre for Food and Agricultural Policy (NCFAP) 2001.
4. Schmutz J, Cannon S, Schlueter J, Ma J, Hyten D, Song Q, Mitros T, Nelson W, May G, Gill N, *et al*: **Genome sequence of the paleopolyploid soybean (Glycine max (L.) Merr.).** *Nature* .
5. **Phytozome.** http://www.phytozome.net/soybean.
6. Jakoby M, Weisshaar B, Droge-Laser W, Vicente-Carbajosa J, Tiedemann J, Kroj T, Parcy F: **bZIP transcription factors in Arabidopsis.** *Trends in Plant Science* 2002, **7(3)**:106-111.
7. Reese J: **Basal transcription factors.** *Current opinion in genetics & development* 2003, **13(2)**:114-118.
8. Weinzierl R: **Mechanisms of gene expression: structure, function and evolution of the basal transcriptional machinery.** London, UK Imperial College Press 1999.
9. Corrinne C: **Transcription factors and mammalian development.** *Current Topics in Developmental Biology* Academic Press, IncPedersen RA 1992, **27**:351.
10. Liu Q, Kasuga M, Sakuma Y, Abe H, Miura S, Yamaguchi-Shinozaki K, Shinozaki K: **Two transcription factors, DREB1 and DREB2, with an EREBP/AP2 DNA binding domain separate two cellular signal transduction pathways in drought-and low-temperature-responsive gene expression, respectively, in Arabidopsis.** *The Plant Cell Online* 1998, **10(8)**:1391-1406.
11. Riechmann J, Heard J, Martin G, Reuber L, Z C, Keddie J, Adam L, Pineda O, Ratcliffe O, Samaha R: **Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes.** *Science* 2000, **290(5499)**:2105-2110.
12. Sakuma Y, Maruyama K, Osakabe Y, Qin F, Seki M, Shinozaki K, Yamaguchi-Shinozaki K: **Functional analysis of an Arabidopsis transcription factor, DREB2A, involved in drought-responsive gene expression.** *The Plant Cell Online* 2006, **18(5)**:1292-1309.
13. Johnson C, Kolevski B, Smyth D: **TRANSPARENT TESTA GLABRA2, a trichome and seed coat development gene of Arabidopsis, encodes a WRKY transcription factor.** *The Plant Cell* 2002, **14(6)**:1359-1375.
14. Ulker B, Somssich I: **WRKY transcription factors: from DNA binding towards biological function.** *Current Opinion in Plant Biology* 2004, **7(5)**:491-498.
15. Shultz J, Kurunam D, Shopinski K, Iqbal M, Kazi S, Zobrist K, Bashir R, Yaegashi S, Lavu N, Afzai A, *et al*: **The Soybean Genome Database (SoyGD): a browser for display of duplicated, polyploid, regions and sequence tagged sites on the integrated physical and genetic maps of Glycine max.** *Nucleic Acids Research* 2006, , **34 Database:** D758-D765.
16. **SoyBase and the soybean breeder's toolbox.** http://soybase.org/.
17. Cheng K, Stromvik M: **SoyXpress: a database for exploring the soybean transcriptome.** *BMC genomics* 2008, **9(1)**:368.
18. Guo A, Chen X, Gao G, Zhang H, Zhu Q, Liu X, Zhong Y, Gu X, He K, Luo J: **PlantTFDB: a comprehensive plant transcription factor database.** *Nucleic Acids Research* 2008, , **36 Database:** D966-D969.
19. Wilson D, Charoensawan V, Kummerfeld S, Teichmann S: **DBD taxonomically broad transcription factor predictions: new content and functionality.** *Nucleic Acids Research* 2007, , **36 Database:** D88-D92.
20. Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, Dolinski K, Dwight S, Eppig J: **Gene ontology: tool for the unification of biology.** *Nature Genetics* 2000, **25(1)**:25-29.
21. Zdobnov E, Apweiler R: **InterProScan-an integration platform for the signature-recognition methods in InterPro.** *Bioinformatics* 2001, **17(9)**:847-848.
22. Stoesser G, Tuli M, Lopez R, Sterk P: **The EMBL nucleotide sequence database.** *Nucleic Acids Research* 1999, **27(1)**:18-24.
23. Apweiler R, Bairoch A, Wu C, Barker W, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, *et al*: **UniProt: the universal protein knowledgebase.** *Nucleic Acids Research* 2004, , **32 Database:** D115-D119.
24. Pruitt K, Tatusova T, Maglott D: **NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Research* 2006, , **00 Database:** D1-D5.
25. Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, Meinhardt T, Prub M, Reuter I, Schacherer F: **TRANSFAC: an integrated system for gene expression regulation.** *Nucleic Acids Research* 2000, **28(1)**:316-319.
26. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T: **The Ensembl genome database project.** *Nucleic Acids Research* 2002, **30(1)**:38-41.
27. Bateman A, Coin L, Durbin R, Finn R, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer E: **The Pfam protein families database.** *Nucleic Acids Research* 2004, **32(1)**:276-280.
28. Madera M, Vogel C, Kummerfeld S, Chothia C, Gough J: **The SUPERFAMILY database in 2004: additions and improvements.** *Nucleic Acids Research* 2004, , **32 Database:** D235-D239.
29. Burley S, Almo S, Bonanno J, Capel M, Chance M, Gaasterland T, Lin D, Sali A, Studier F, Swaminathan S: **Structural genomics: beyond the human genome project.** *Nature Genetics* 1999, **23**:151-158.
30. Burley S: **An overview of structural genomics.** *Nature Structural & Molecular Biology* 2000, **7**:932-934.
31. Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov I, Bourne P: **The protein data bank.** *Nucleic Acids Research* 2000, **28(1)**:235-242.
32. Boeckmann B, Bairoch A, Apweiler R, Blatter M, Estreicher A, Gasteiger E, Martin M, Michoud K, O'Donovan C, Phan I: **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.** *Nucleic Acids Research* 2003, **31(1)**:365-370.
33. Rhee S, Beavis W, Berardini T, Chen G, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M, *et al*: **The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community.** *Nucleic Acids Research* 2003, , **31:** 224-228.
34. Letunic I, Copley R, Pils B, Pinkert S, Schultz J, Bork P: **SMART 5: domains in the context of genomes and networks.** *Nucleic Acids Research* 2006, , **34 Database:** D257-D260.
35. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T: **KEGG for linking genomes to life and the environment.** *Nucleic Acids Research* 2008, , **36 Database:** D480-D484.
36. Attwood T, Blythe M, Flower D, Gaulton A, Mabey J, Maudling N, McGregor L, Mitchell A, Moulton G, Paine K: **PRINTS and PRINTS-S shed light on protein ancestry.** *Nucleic Acids Research* 2002, **30(1)**:239-241.
37. Angiuoli S, Gussman A, Klimke W, Cochrane G, Field D, Garrity G, Kodira C, Kyrpides N, Madupu R, Markowitz V: **Toward an online repository of standard operating procedures (SOPs) for (Meta) genomic annotation.** *OMICS: A Journal of Integrative Biology* 2008, **12(2)**:137-141.
38. Mulder N, Apweiler R, Attwood T, Bairoch A, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P: **InterPro: An integrated documentation resource for protein families, domains and functional sites.** *Briefings in Bioinformatics* 2002, **3(3)**:225-235.
39. Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux P, Pagni M, Sigrist C: **The PROSITE database.** *Nucleic Acids Research* 2006, , **34 Database:** D227-D230.
40. Yeh R, Lim L, Burge C: **Computational inference of homologous gene structures in the human genome.** *Genome research* 2001, **11(5)**:803-816.
41. **FgenesH.** http://linux1.softberry.com/berry.phtml.

42. Haas B, Delcher A, Mount S, Wortman J, Smith R Jr, Hannick L, Maiti R, Ronning C, Rusch D, Town C: **Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies.** *Nucleic Acids Research* 2003, **31(19)**:5654-5666.
43. Attwood T, Croning M, Flower D, Lewis A, Mabey J, Scordis P, Selley J, Wright W: **PRINTS-S: the database formerly known as PRINTS.** *Nucleic Acids Research* 2000, **28(1)**:225-227.
44. Corpet F, Gouzy J, Kahn D: **Recent improvements of the ProDom database of protein domain families.** *Nucleic Acids Research* 1999, **27(1)**:263-267.
45. Haft D, Loftus B, Richardson D, Yang F, Eisen J, Paulsen I, White O: **TIGRFAMs: a protein family resource for the functional identification of proteins.** *Nucleic Acids Research* 2001, **29(1)**:41-43.
46. Wu C, Huang H, Yeh L, Barker W: **Protein family classification and functional annotation.** *Computational Biology and Chemistry* 2003, **27(1)**:37-47.
47. Gough J, Karplus K, Hughey R, Chothia C: **Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure.** *Journal of Molecular Biology* 2001, **313(4)**:903-919.
48. Buchan D, Shepherd A, Lee D, Pearl F, Rison S, Thornton J, Orengo C: **Gene3D: structural assignment for whole genes and genomes using the CATH domain structure database.** *Genome Research* 2002, **12(3)**:503-514.
49. Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, Rabkin S, Guo N, Muruganujan A, Doremieux O, Campbell M: **The PANTHER database of protein families, subfamilies, functions and pathways.** *Nucleic Acids Research* 2005, , **33 Database:** D284-D288.
50. Lima T, Auchincloss A, Coudert E, Keller G, Michoud K, Rivoire C, Bulliard V, de Castro E, Lachaize C, Baratin D: **HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot.** *Nucleic Acids Research* 2009, , **37 Database:** D471-D478.
51. Riano-Pachon D, Ruzicic S, Dreyer I, Mueller-Roeber B: **PlnTFDB: an integrative plant transcription factor database.** *BMC Bioinformatics* 2007, **8(1)**:42.
52. Kakar K, Wandrey M, Czechowski T, Gaertner T, Scheible W, Stitt M, Torres-Jerez I, Xiao Y, Redman J, Wu H: **A community resource for high-throughput quantitative RT-PCR analysis of transcription factor gene expression in Medicago truncatula.** *Plant Methods* 2008, **4(1)**:18.
53. Edgar R: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Research* 2004, **32(5)**:1792-1797.
54. Cheng J: **A multi-template combination algorithm for protein comparative modeling.** *BMC Structural Biology* 2008, **8(1)**:18.
55. **CASP8 Group Performance.** http://www.predictioncenter.org/casp8/results.cgi.
56. Zemla A: **LGA: a method for finding 3D similarities in protein structures.** *Nucleic Acids Research* 2003, **31(13)**:3370-3374.
57. Jones S, Shanahan H, Berman H, Thornton J: **Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins.** *Nucleic Acids Research* 2003, **31(24)**:7189-7198.
58. **Jmol: an open-source Java viewer for chemical structures in 3D.** http://jmol.sourceforge.net/.
59. Alexandrov N, Shindyalov I: **PDP: protein domain parser.** Oxford Univ Press 2003, **19**:429-430.
60. Cheng J: **DOMAC: an accurate, hybrid protein domain prediction server.** *Nucleic Acids Research* 2007, , **35 Web Server:** W354-W356.
61. Crooks G, Hon G, Chandonia J, Brenner S: **WebLogo: a sequence logo generator.** *Genome Research* 2004, **14(6)**:1188-1190.
62. Benson D, Boguski M, Lipman D, Ostell J, Ouellette B, Rapp B, Wheeler D: **GenBank.** *Nucleic Acids Research* 1999, **27(1)**:12-17.
63. Libault M, Joshi T, Takahashi K, Hurley-Sommer A, Puricelli K, Blake S, Xu D, Nguyen H, Stacey G: **Large-scale analysis of putative soybean regulatory gene expression identifies a Myb gene involved in soybean nodule development.** *Plant Physiology* 2009, **151**:1207-1220.
64. Schlueter J, Lin J, Schlueter S, Vasylenko-Sanders I, Deshpande S, Yi J, O'Bleness M, Roe B, Nelson R, Scheffler B: **Gene duplication and paleopolyploidy in soybean and the implications for whole genome sequencing.** *BMC genomics* 2007, **8(1)**:330.
65. Schlueter J, Dixon P, Granger C, Grant D, Clark L, Doyle J, Shoemaker R: **Mining EST databases to resolve evolutionary events in major crop species.** *Genome* 2004, **47(5)**:868-876.
66. Merchant S, Prochnik S, Vallon O, Harris E, Karpowicz S, Witman G, Terry A, Salamov A, Fritz-Laylin L, Marechal-Drouard L: **The chlamydomonas genome reveals the evolution of key animal and plant functions.** *Science* 2007, **318(5848)**:245-250.
67. Rensing S, Lang D, Zimmer A, Terry A, Salamov A, Shapiro H, Nishiyama T, Perroud P, Lindquist E, Kamisugi Y: **The physcomitrella genome reveals evolutionary insights into the conquest of land by plants.** *Science* 2008, **319(5859)**:64-69.
68. Scofield S, Murray J: **KNOX gene function in plant stem cell niches.** *Plant Molecular Biology* 2006, **60(6)**:929-946.
69. Libault M, Joshi T, Benedito V, Xu D, Udvardi M, Stacey G: **Legume transcription factor genes; what makes legumes so special?.** *Plant Physiology* 2009.
70. Yan C, Terribilini M, Wu F, Jernigan R, Dobbs D, Honavar V: **Predicting DNA-binding sites of proteins from amino acid sequence.** *BMC Bioinformatics* 2006, **7(1)**:262.
71. Wang L, Brown S: **BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences.** *Nucleic Acids Research* 2006, , **34 Web Server:** W243-W248.
72. Hwang S, Gou Z, Kuznetsov I: **DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins.** *Bioinformatics* 2007, **23(5)**:634-636.
73. Yamasaki K, Kigawa T, Inoue M, Tateno M, Yamasaki T, Yabuki T, Aoki M, Seki E, Matsuda T, Tomo Y: **Solution structure of the B3 DNA binding domain of the Arabidopsis cold-responsive transcription factor RAV1.** *The Plant Cell Online* 2004, **16(12)**:3448-3459.
74. Kaplan W, Littlejohn T: **Swiss-PDB viewer (deep view).** *Briefings in Bioinformatics* 2001, **2(2)**:195-197.
75. **The PyMOL molecular graphics system.** http://pymol.sourceforge.net/.

doi:10.1186/1471-2229-10-14
Cite this article as: Wang *et al.*: SoyDB: a knowledge database of soybean transcription factors. *BMC Plant Biology* 2010 **10**:14.