Chapter 21

Computational Identification of Plant Transcription Factors and the Construction of the PlantTFDB Database

Kun He, An-Yuan Guo, Ge Gao, Qi-Hui Zhu, Xiao-Chuan Liu, He Zhang, Xin Chen, Xiaocheng Gu, and Jingchu Luo

Abstract

Transcription factors (TFs) play an important role in gene regulation. Computational identification and annotation of TFs at genome scale are the first step toward understanding the mechanism of gene expression and regulation. We started to construct the database of *Arabidopsis* TFs in 2005 and developed a pipeline for systematic identification of plant TFs from genomic and transcript sequences. In the following years, we built a database of plant TFs (PlantTFDB, http://planttfdb.cbi.pku.edu.cn) which contains putative TFs identified from 22 species including five model organisms and 17 economically important plants with available EST sequences. To provide comprehensive information for the putative TFs, we made extensive annotation at both the family and gene levels. A brief introduction and key references were presented for each family. Functional domain information and cross-references to various well-known public databases were available for each identified TF. In addition, we predicted putative orthologs of the TFs in other species. PlantTFDB has a simple interface to allow users to make text queries, or BLAST searches, and to download TF sequences for local analysis. We hope that PlantTFDB could provide the user community with a useful resource for studying the function and evolution of transcription factors.

Key words: Transcription factors, database construction, plant genome, HMMER search, Ortholog.

1. Introduction

Transcription factors (TFs) bind selectively to specific DNA sequences in order to turn on or off the transcription of their target genes. Eukaryotes have a much more sophisticated transcription regulation mechanism than prokaryotes. For example,

I. Ladunga (ed.), *Computational Biology of Transcription Factor Binding*, Methods in Molecular Biology 674,
DOI 10.1007/978-1-60761-854-6_21, © Springer Science+Business Media, LLC 2010

eukaryotic RNA polymerase complexes cannot turn on the gene transcription by themselves. Instead, TFs are needed to recognize and bind to the *cis*-regulatory elements that primarily reside in the promoter region of the target genes and to recruit the RNA polymerase complex for the initiation of transcription. The special sequence region in a TF that interacts directly with DNA is called the DNA-binding domain (DBD). Based on the sequence patterns and structural features of DBDs, TFs can be classified into diverse families (1).

Multicellular eukaryotes have to deal with cell differentiation with the aid of more sophisticated regulatory mechanism using a larger number of TFs (2). Larger genomes usually have higher numbers of TFs. It has been recently reported that a normal human somatic cell was turned into a fully functional stem cell by introducing four TFs into it (3). Therefore, deciphering the binding relationship and regulatory network is a key step to the understanding of the process of development and many other biological phenomena (4). Identification of TFs at the genome level and construction of knowledge databases for the TFs using computational approach have therefore become the fundamental step in studies on the regulation of gene expression.

TRANSFAC is one of the earliest attempts to build a TF knowledgebase with experimentally proven TFs, their binding sites, and regulated target genes (5). The January 2009 release contains 12,183 factors and 24,745 binding sites from various taxa including plants, mammals, fungi, and bacteria. It started to collect data identified by high-throughput technologies, such as chromatin immunoprecipitation on chip (ChIP-chip). Several databases of plant TFs have been developed after the completion of the *Arabidopsis* genome sequencing in 2000 as well as several other plant model organisms such as rice and poplar in the following years (**Table** 21.1).

AtTFDB is the first *Arabidopsis* TF databases hosted by the *Arabidopsis* gene regulatory information server (AGRIS) at the Ohio State University [*see* **Chapter 2** and ref. (6)]. It classified TFs into different families based on the type of the corresponding DBDs. Recently, AtTFDB further integrated information about the potential regulatory relationship among TFs which makes it a useful resource for regulatory network analysis of *Arabidopsis*. The *Arabidopsis* TF database (RARTF) hosted by the RIKEN BioResource Center in Japan applies PSI-BLAST and InterProScan to the identification of all putative TFs in *Arabidopsis* (7). Sequence information, InterPro domains, and links to public *Arabidopsis* genome databases such as TAIR, MIPS, and TIGR are provided for each predicted TF. Recently, a database of tobacco TFs (TOBFAC) has been built by the University of Virginia (8). TOBFAC contains about 2,500 putative tobacco TF genes predicted from the data source of both gene-space reads

Table 21.1
Databases of plant transcription factors

Name	Data source	Website and institution	References
TRANSFAC	Mainly *Arabidopsis*	http://www.gene-regulation.com/ BIOBASE, Germany	(5)
AtTFDB	*Arabidopsis*	http://arabidopsis.med.ohio-state.edu/AtTFDB/ The Ohio State University, USA	(6)
RARTF	*Arabidopsis*	http://rarge.psc.riken.jp/rartf/ RIKEN BioResource Center, Japan	(7)
TOBFAC	Tobacco	http://compsysbio.achs.virginia.edu/tobfac/ University of Virginia, USA	(8)
LegumeTFDB	*Glycine max*, *Lotus japonicus, Medicago truncatula*	http://legumetfdb.psc.riken.jp/ RIKEN BioResource Center, Japan	
PlnTFDB	Genome sequences (19 species)	http://plntfdb.bio.uni-potsdam.de/ University of Potsdam, Germany	(9)
PlantTFDB	Genome sequence (5 species) EST sequence (17 species)	http://planttfdb.cbi.pku.edu.cn/ Peking University, China	(10)

and EST sequences. LegumeTFDB, a database of three legume plants (*Glycine max*, *Lotus japonicus, Medicago truncatula*) has been available online at the RIKEN BioResource Center, Japan. Interestingly, the number of predicted soybean (*Glycine max*) TFs listed in LegumeTFDB exceeds the number of predicted tobacco TFs by a factor of 2. Up to date, the two most comprehensive plant TF databases are the PlnTFDB developed by University of Potsdam, Germany (http://plntfdb.bio.uni-potsdam.de/) (9) and the PlantTFDB constructed by Peking University, China (http://planttfdb.cbi.pku.edu.cn/) (10). Both PlnTFDB and PlantTFDB attempt to collect plant TFs from all available data sources and provide comprehensive annotation at both the family and gene level.

In this chapter, we describe our computational approaches to the genome-wide identification of plant TFs, construction of the TF databases, and annotation of TF genes. In 2002, several research groups from China and the United States initiated a collaborative project for the genome-wide ORFeome cloning and analysis of *Arabidopsis* genes which encode TFs (11). As the only bioinformatics group involved in this project, we started to construct the database of *Arabidopsis* TFs (DATF) which became publicly available in 2005 (12). With the available genome sequences of two rice sub-species (*Oryza sativa*, ssp.

japonica and *Oryza sativa*, ssp. *indica*) and poplar (*Populus trichocarpa*), two databases of rice TFs (DRTF) (13) and poplar TFs (DPTF) (14) were further built up in the following years.

Based on the experiences obtained from the construction of these three plant TF databases, we have built an in silico pipeline for the systematic identification of the putative TFs from various plant genomes and developed a comprehensive plant TF database PlantTFDB. To provide a comprehensive data source for plant biologists, PlantTFDB contains TFs identified from 5 model organisms with whole genome sequences and 17 economically important plants with abundant transcripts (**Table** 21.2).

Table 21.2
TFs and ortholog numbers in PlantTFDB

Data source (version)	Name		Species	TFs[a]	TFs with orthologs
TAIR (v6)	*Arabidopsis*		*Arabidopsis thaliana*	2290	1346
JGI (v1.1)	Poplar		*Populus trichocarpa*	2576	2042
TIGR (v4.0)	Rice		*Oryza sativa* (ssp. indica) *Oryza sativa* (ssp. japonica)	2025 2384	1763 2124
JGI(v1.1)	Moss		*Physcomitrella patens*	1170	524
JGI(v3.0)	Green alga		*Chlamydomonas reinhardtii*	205	64
PlantGDB (v155a)	Crops	Barley	*Hordeum vulgare*	618	595
		Maize	*Zea mays*	764	734
		Sorghum	*Sorghum bicolor*	397	372
		Sugarcane	*Saccharum officinarum*	1177	1157
		Wheat	*Triticum aestivum*	1127	1074
	Fruits	Apple	*Malus x domestica*	1025	938
		Grape	*Vitis vinifera*	867	793
		Orange	*Citrus sinensis*	599	541
	Trees	Pine	*Pinus taeda*	950	644
		Spruce	*Picea glauca*	440	383
	Economic plants	Cotton	*Gossypium hirsutum*	1567	1430
		Potato	*Solanum tuberosum*	1340	1243
		Soybean	*Glycine max*	1891	1774
		Sunflower	*Helianthus annuus*	513	435
		Tomato	*Lycopersicon esculentum*	998	917
		Lotus	*Lotus japonicus*	457	434
		Medicago	*Medicago truncatula*	1022	914

[a]The TF numbers of *Arabidopsis* and rice *japonica* are the gene model numbers including alternative splicing.

TFs predicted from newly sequenced genomes are being added to PlantTFDB.

PlantTFDB attempts to provide comprehensive information for the identified TFs both at the family and at the gene level. A brief introduction can be found for each family. In addition to common sequence features derived from well-known domain database (15, 16) and Gene Ontology (http://www.geneontology.org/) (17), expression profiling data derived from UniGene and NCBI GEO repository are also available for each predicted TF. Moreover, automatically annotated homologs in related species can also be found for each TF. With a user-friendly Web interface, all sequences and annotation information are freely available online (http://planttfdb.cbi.pku.edu.cn/).

2. Materials

2.1. Sequence Data

Whole proteome sequences of five model organisms with completed genomes were downloaded from genome sequencing centers (**Table** 21.3). The *Arabidopsis* Information Resource (TAIR) maintains a database of genomic data for the model plant *Arabidopsis thaliana* (http://www.arabidopsis.org/). The genome sequence of the rice sub-species *japonica* was originally hosted at The Institute of Genome Research (TIGR) and moved to Michigan State University in 2007 (http://rice.plantbiology.msu.edu/), while another rice sub-species *indica*

Table 21.3
Data source of PlantTFDB

Species and data type	Website and institution[a]
Arabidopsis genome sequence	http://www.arabidopsis.org/ The Arabidopsis Information Resource (TAIR), USA
Rice (*japonica*) genome sequence	http://rice.plantbiology.msu.edu/ Michigan State University, USA
Rice (*indica*) genome sequence	http://rise.genomics.org.cn/ Beijing Genome Institute, China
Poplar, Moss, Green Algae genome sequence	http://www.jgi.doe.gov/ DOE Joint Genome Institute (JGI), USA
Plant unique transcripts (PUTs) of 17 plants assembled by PlantGDB	http://www.plantgdb.org/ Plant Genome Database (PlantGDB), USA

[a]Rice (*japonica*) was originally hosted at The Institute for Genomic Research (TIGR) and moved to Michigan State University in 2007.

was sequenced by the Beijing Genome Institute (BGI), China (http://rise.genomics.org.cn/). All other genome sequences are obtained from the Joint Genome Institute (JGI), the US Department of Energy (http://www.jgi.doe.gov/).

For the 17 plants whose genomic sequences were not available in 2007 when we started to construct PlantTFDB, we downloaded the assembled transcripts from the Plant Genome Database (PlantGDB, http://www.plantgdb.org/). By assembling mRNA and EST sequences available in public databanks, PlantGDB predicted a set of plant unique transcripts (PUTs) for each organism. Based on those transcripts, we further identified open reading frames and derived protein sequences using the framefinder program (http://www.ebi.ac.uk/~guy/estate/).

2.2. Software Tools

All data and information are stored in a MySQL relational database on a Linux server. MySQL is an open source database management system widely used for both small and large database applications (http://dev.mysql.com). Queries to the database are implemented in PHP scripts running in an Apache/PHP environment. Graphics are drawn using the PHP module of the GD graphics library. Three-dimensional structure illustration was created using Molscript (http://www.avatar.se/molscript/) (18).

The NCBI BLAST tool kit (19) is installed locally for sequence similarity search. The HMMER package (http://hmmer.janelia.org/) (20) is used for Hidden Markov Model (HMM) profile search. HMM profiles of known DNA-binding domains are obtained from the Pfam database (http://pfam.sanger.ac.uk/) (16). The ClustalW (21) program is used for multiple sequence alignment. The Phylip package (http://evolution.genetics.washington.edu/phylip.html) is implemented for the construction of the phylogeny trees. The InterProScan (15) program is employed to identify protein domains and assign Gene Ontology (GO) (17) terms to the putative TFs.

3. Methods

3.1. Classification of TF Families

Like most other homologous proteins, TFs are usually grouped into families based on the conserved sequences of their DBDs. DBDs are responsible for the recognition of the *cis*-regulatory elements in the promoter and other regulatory regions of target genes. As stated above, the DBD is used to determine whether or not a protein could be considered as a putative TF. However,

domain shuffling and horizontal gene transfer events have been abundant during evolution. Some of the TFs may contain more than one type of DBD so the classification of TFs into families is not always straightforward.

Richemann et al. (1) systematically compared families among *Arabidopsis* and other species and summarized the relationship between DBDs and TF families. Based on the convention they proposed, we classify all known *Arabidopsis* TFs into 64 families (**Fig.** 21.1). **Table** 21.4 lists the TF number of each family in *Arabidopsis* and the other four model plant species.

There are three different relationships between TF families and their DNA-binding domains: required, possible, or forbidden domain. A required domain means that a TF in a certain family must contain the corresponding domain. Using two TF families CCAAT-Dr1 and CCAAT-HAP3 as examples (bottom left in **Fig.** 21.1), the existence of a Dr1 or HAP3 domain in either family is required to classify this TF as a CCAAT-Dr1 or CCAAT-HAP3 family member. A possible domain is defined such that a TF from a certain family might contain this domain in addition to the required domain. For instance, a member of CCAAT-HAP3 family might contain a Dr1 domain besides the required HAP3 domain. Finally, a forbidden domain means that it should not be contained in the TF of a certain family. In the above example, a CCAAT-Dr1 family member should not contain a HAP3 domain, otherwise it will be classified as a CCAAT-HAP3 member. In summary, if a protein contains a HAP3 domain, it is classified into the CCAAT-HAP3 family no matter whether it contains a Dr1 domain or not. On the other hand, a protein containing only the Dr1 domain but not the HAP3 domain will be classified into the CCAAT-Dr1 family.

3.2. Prediction of TFs

We combine automated search with manual curation for the identification of *Arabidopsis* TFs (**Fig.** 21.2). A list of 64 TF families was obtained based on the fundamental work by Riechmann et al. (1) as well as from a literature survey. HMM Profiles (20) are statistical models of multiple sequence alignments and contain position-specific information about the occurrence probabilities of all possible residues for each column in the alignment. HMMER (http://hmmer.janelia.org/) is an implementation of profile HMMs for biological sequence analysis. HMMER can be used to construct profile according to multiple sequence alignments, and it can also use a given profile to search for sequences belonging to the same family with the given profile. Pfam (16) is a database of protein domains represented by multiple sequence alignments and HMM profiles built using HMMER. HMM profiles of 48 TF families can be found in Pfam and are used in HMMER search. For the remaining 16 families where HMM profiles were not available at the time when we started to construct

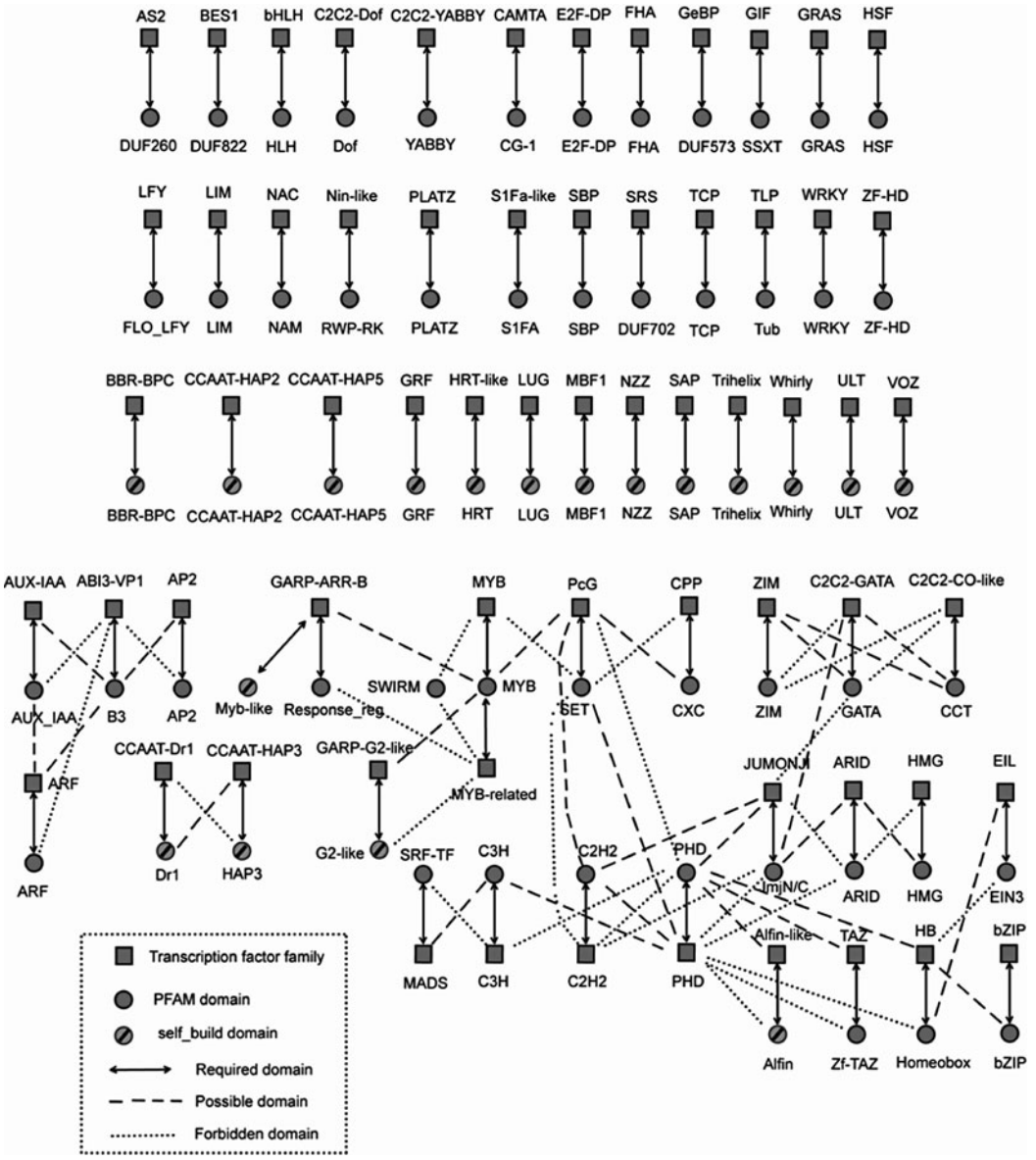

Fig. 21.1. Schemetic representation of the relationship between *Arabidopsis* TF families and their DNA-binding domains (DBDs). *Squares* denote TF families, *circles* show DBDs obtained from the Pfam database, *circles* with latches inside indicate the DBDs we constructed. A double-headed *arrow* connects the DBD which must be contained in the corresponding TF family, *dashed lines* link one or more DBDs which might exist within this TF family, *dotted lines* demonstrate one or more DBDs which should not be contained in this family.

the database of *Arabidopsis* TFs, seed sequences were retrieved from the public protein sequence databases such as GenPept and Swiss-Prot and were taken as query sequence for BLAST search against the protein sequences of the *Arabidopsis* genome. With more plant genome sequences available, we have accumulated more data for each of the above 16 families and built HMM

Table 21.4
The number of predicted TF family members in five plant genomes

TF family	Description	At	Os	Pt	Pp	Cr	Function	TFBS
ABI3-VP1	Abscisic acid insensitive 3,viviParous1	60	52	108	37	1	ABA response	
Alfin(Zn)	Alfalfa protein	7	10	9	7	2	Salt response	
AP2-EREBP	APETALA2 and ethylene-responsive element binding proteins	146	165	212	156	12	Flower development, ethylene response, drought, cold response	GCC box
ARF	Auxin response factors	23	26	37	13		Auxin response	TGTCTC
ARID	AT-rich interaction domain	10	5	13	8	1	Chromatin remodeling	AT-rich
AS2	Asymmetric leaves2	42	36	57	29		Leaf venation establishment	
AUX-IAA	Auxin (indole-3-acetic acid)	29	29	33	2		Auxin response	
BBR-BPC	Barley B recombinant protein, basic pentacysteine	7	4	16			Ovule identity control	(GA/TC)8
BES1	Bri1-Ems-suppressor 1	8	6	12	6		Brassinosteroid regulation	
bHLH	Basic helix loop helix	127	151	148	98	3	Secondary metabolism, cell proliferation	G-box E-box
bZIP	Basic leucine zipper	72	84	85	38	9	Flower and leaf development, hormone response	ACGT

(continued)

Table 21.4 (continued)

TF family		Description	At	Os	Pt	Pp	Cr	Function	TFBS
C2C2(Zn)	CO-like	Constans	37	39	39	6	6	Meristem identity control	
	Dof	DNA binding with one finger	36	30	42	1	1	Carbohydrate metabolism, defense, germination, hormone response	
	GATA	GATA	26	21	32	11	11	Light response	[TA]GATA[GA]
	YABBY		5	8	13			Abaxial identity control	
C2H2(Zn)		Cys2-His2 motif	134	94	81	49	5	Development	
C3H(Zn)		Cys3-His motif	59	57	78	37	12	Development	
CAMTA		Calmodulin-binding transcription activators	6	6	7	1		Calcium pathway	
CCAAT	Dr1	DRAP1	2	1	2	1			CCAAT
	HAP2	Heme activator protein 2	10	11	11			Embryo and flower development, circadian rhythm control, light signaling	CCAAT
	HAP3	Heme activator protein 3	11	12	19	2		Flowering control	CCAAT
	HAP5	Heme activator protein 5	13	16	19	2			CCAAT
CPP(Zn)		Cell shape control protein phosphatase	8	11	13	6	2	Cell proliferation, leghemoglobin gene regulation	
E2F-DP		Electro acoustic 2 factor (E2F)-DRTF1 polypeptide (DP)	8	8	10	11	3	Cell cycle regulation	
EIL		Ethylene insensitive 3 like	6	9	6	2		Ethylene response	
FHA		ForkHead associated	16	16	19	15	11	Cell cycle regulation	

(continued)

Table 21.4 (continued)

TF family		Description	At	Os	Pt	Pp	Cr	Function	TFBS
GARP	ARR-B	*Arabidopsis* response regulator-B	10	8	15	1		Cytokinin signal transduction	
	G2-like	GLK2-like	43	46	67	4		chloroplast development	
GeBP		Glabrous1 enhancer binding protein	21	15	7			Leaf cell development	
GIF		GRF-interacting factor	3	3	5	4	1	Leaf growth, pattern formation	
GRAS		Acronym for three genes: Gai, RgA, Scr	33	55	96	39		Meristem development, gibberellin response	
GRF		Growth regulating factor	9	12	9	2		Leaf and cotyledon growth	
HB		Homeo Box	87	82	106	40	1	Development	CAATNATTG
HMG		High mobility group	11	9	12	9	8		AT-hook
HRT-like(Zn)		Hordeum repressor of transcription (HRT)	2	1	1	3		Hormone response	
HSF		Heat shock transcription factor	23	25	31	8	2	Heat shock response	
JUMONJI			17	15	20	10	7	Flowering control	
LFY		LeaFY	1	1	1	2		Flower development	
LIM(Zn)		Acronym for three genes: Lin11, Isl-1, Mec-3	13	10	21	11		Lignin synthesis	Pal-box
LUG		LeUniG	2	6	6			Flower development	
MADS		Acronym for four genes: MCM1, AGAMOUS, DEFICIENS, SRF	104	63	111	22	1	Flower development	CC[A/T]6GG
MBF1		Multiprotein bridging factor 1	3	2	3	3	1		
MYB		MYeloBlastosis viral oncogene homolog	150	129	216	64	14	Cell proliferation, secondary metabolism, defense, ABA response	GGTTTAG

(continued)

Table 21.4 (continued)

TF family	Description	At	Os	Pt	Pp	Cr	Function	TFBS
MYB-related		49	60	84	31	7	Circadian rhythm control, cell proliferation	
NAC	Acronym for three genes: NAM, ATAF1, CUC2	107	130	172	32		Meristem development, hormone response, defense	
Nin-like	Nodule INception like	14	13	18	9	8	Root nodule development	
NZZ	Nozzle	1	1	2	3		Flower development	
PcG	PolyComb group	34	33	45	31	21	Seed development	
PHD(Zn)	Plant homeo domain	56	63	86	68	13	Light response, auxin signaling, leaf polarity	
PLATZ(Zn)	Plant AT-rich sequence and zinc-binding protein	10	16	20	13	4	Transcription repression	
S1Fa-like		3	2	2	1			
SAP	Sterile apetala	1	0	1			Flower development	
SBP(Zn)	Squamosa promoter binding protein	16	20	29	14	21	Flower and fruit development	TNCGTACAA
SRS(Zn)	Shi-related sequence	10	5	10	2		Gibberellin response	
TAZ(Zn)	Transcriptional adaptor zinc-binding domain	9	6	7	5	2	Calcium binding, stress response	
TCP	Acronym for three genes: TB1, CYC, PCFs	23	21	34	6		Flower development, cell division	GGNCCC
TLP	Tubby-like proteins	11	14	11	6	3	ABA pathway	
Trihelix		26	20	47	28		Light response	box II
ULT	Ultrapetal	2	2	3			Apical meristem development	

(continued)

Table 21.4 (continued)

TF family	Description	At	Os	Pt	Pp	Cr	Function	TFBS
VOZ(Zn)	Vascular plant transcription factors with one zinc finger	2	2	4	2		Pollen development	GCGTNx7ACGC
Whirly		2	1	2		1	Stress response	TGACAnnnnTGTCA
WRKY(Zn)	WRKY sequence motif	72	98	104	37	1	Defense	TGAC
ZF-HD(Zn)	Zinc finger homeo domain	16	15	25	8		Flower development	
ZIM(Zn)	Zinc finger motif	18	18	22	16		Leaf development	

Names with underlines are plant-specific families; (Zn): zinc finger protein; At: *Arabidopsis thaliana*; OS: *Oryza sativa japonica*; Pt: *Populus trichocarpa*; Cr: Pp: *Physcomitrella patens*

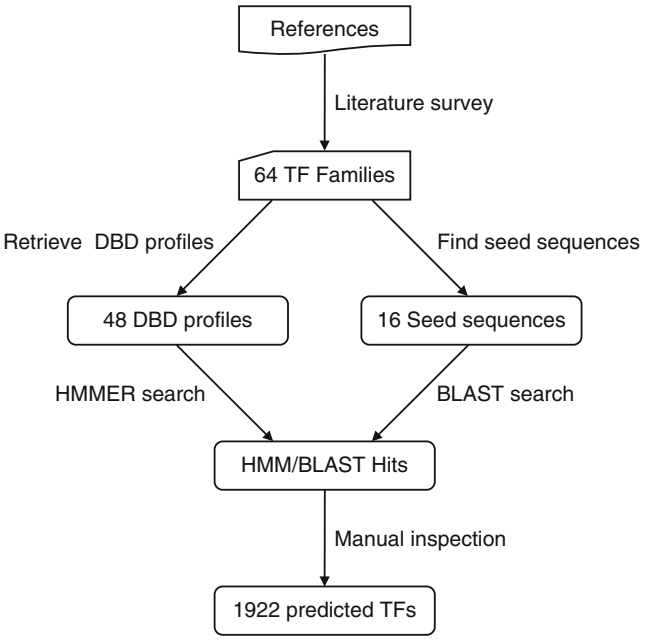

Fig. 21.2. Flowchart of the computational approach for genome-wide identification of transcription factors from *Arabidopsis thaliana*. Literature survey through published references is the first step to find all TFs characterized by experimental studies. A list of 64 TF families was constructed. HMM profiles of 48 DBDs were retrieved from the Pfam database and used in HMMER search. Seed sequences of DBDs for 16 families without HMM profiles were retrieved from the public protein sequence databases for BLAST search. Both HMMER and BLAST search results were manually checked to remove false positives. A total of 1,922 putative TFs were predicted from the *Arabidopsis* genome.

profiles for each DBD which can be used in the prediction of TFs in other species with either whole genome sequence or EST data.

3.3. Annotation

To provide sufficient information about the putative TFs, we made various annotations at both the family and gene levels. For each TF family, PlantTFDB gives a brief introduction including the potential function, the three-dimensional structure of the DBD, the characterization of the *cis*-regulatory element bound by the DBD of the family. For individual TFs, PlantTFDB shows general information such as database identifier, gene name, DNA sequence of both genomic and coding region, and protein sequence. The database of *Arabidopsis* TFs has the most comprehensive annotations benefiting from the rich published results of genetic and functional investigations of this model organism (**Fig.** 21.3). In addition to the general information, DATF includes the unique information as to whether a TF has been cloned (11) which can be browsed and searched in the "clone information" field. BLAST search was performed against well-known public databases and cross-references are linked to various public databases. Putative functional domains are identified

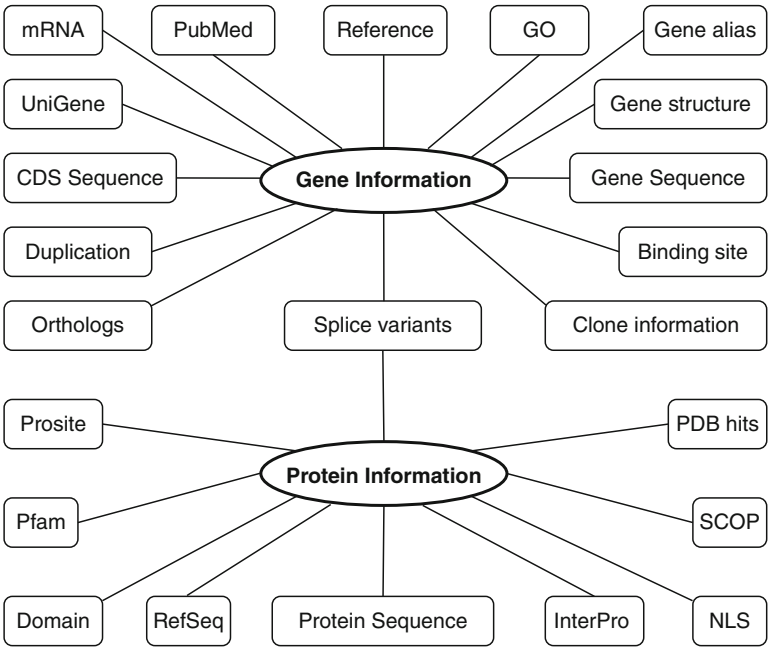

Fig. 21.3. Schematic demonstration of the annotation for individual TF genes in the PlantTFDB. Two ellipses show the key fields of "Gene Information" and "Protein Information" linked by "Splice Variants." *Round squares* with lines connected to either "Gene Information" or "Protein Information" show the annotation retrieved from various resources or cross-references to related databases.

and annotated by InterProScan, and Gene Ontology annotations are further extracted. In addition, the expression profiles collected from UniGene EST/cDNA information are also available.

3.4. Ortholog Identification

Recent genomic studies have already discovered several TF families that are specifically expanded in the plant kingdom. The ortholog information among different species is predicted using the BLAST score ratio (BSR), which is widely adopted by ENSEMBL (http://www.ensembl.org/) and other studies. An all-against-all BLASTP search with a strict cutoff *E*-value $<1 \times 10^{-20}$ was performed, and the BSR value was calculated for each hit. After comparing results at different BSR values, we chose the BSR value 0.4 as the cutoff and we retrieved the top sequences in a species with the largest BSR value as the putative ortholog(s). Using the bZIP family TF HY5 as an example, we identified orthologs of HY5 in 13 species (**Table** 21.5).

HY5 is an important activator for photomorphogenesis, which is essential for most of the land plants. As expected, all four land plants with completed genome data have HY5 orthologs. It was reported that the closest HY5 homologs in green algae has no COP complex interaction site detected, suggesting that HY5 might not be essential for this aquatic plant. The fact that HY5 orthologs have not been found in green algae could also be

Table 21.5
The orthologs of AtHY5 among different plant species

Species	PlantTFDB ID	Score ratio	Coverage	Identity	*E*-value
Physcomitrella patens	213225	0.43	0.82	0.58	4×10^{-35}
Populus trichocarpa	fgenesh4_pm.C_LG_XVIII000127	0.78	0.99	0.78	2×10^{-68}
Oryza sativa (japonica)	LOC_Os02g10860.1	0.58	0.98	0.69	7×10^{-49}
Oryza sativa (indica)	OsIBCD000496	0.53	0.95	0.65	2×10^{-44}
Citrus sinensis	PTCs00574.1	0.77	0.98	0.81	1×10^{-68}
Glycine max	PTGm01858.1	0.59	0.93	0.65	6×10^{-51}
Helianthus annuus	PTHa00512.1	0.54	0.71	0.74	3×10^{-46}
Lycopersicon esculentum	PTLe00971.1	0.69	0.94	0.77	4×10^{-60}
Lotus japonicus	PTLj00452.1	0.59	0.92	0.64	6×10^{-51}
Malus x domestica	PTMx01004.1	0.76	0.98	0.78	2×10^{-67}
Picea glauca	PTPg00432.1	0.43	0.55	0.76	3×10^{-35}
Solanum tuberosum	PTSt01300.1	0.64	0.86	0.79	7×10^{-56}
Vitis vinifera	PTVv00842.1	0.59	0.8	0.76	4×10^{-51}

due to our overly stringent criteria and the sequence difference between green alga and other plants. For the species from which HY5 orthologs were not detected, either genome sequence or more EST data should be used in the future.

4. Notes

1. Data source and database updating
   With the rapid progress of next-generation sequencing technology, more and more plant genomes have been already or are being sequenced. Sequence data of both genomic DNA and mRNA transcripts of different plant lineages become a rich source for computational identification of plant TFs at the genome level. We shall update PlantTFDB with the new release of sequence data of the five model organisms, as well as other newly sequenced genomes.[1]

[1] The PlantTFDB was updated to version 2.0 in July 2010, with predicted TFs from more species, and a new interface.

2. Family name and classification
   There is no standard nomenclature for the name of TF families. Some of the families were named after the biological processes in which they are involved, and others by the three-dimensional structures of their DBDs. For example, the family ARF is referring to a group of auxin responsive factors, while bHLH is given to a group of TFs with conserved DBDs that form basic helix loop helix structures. For the former group, it is reasonable to expect most if not all of the members are involved in auxin responses-related processes.
   Although distinct functions might be seen in different TF families, the family classification implemented in PlantTFDB should not be taken as the reflection of their biological functions. Functions of the TFs from the same family could be dramatically different. On the other hand, TFs from different families can recognize similar or even identical binding sites and be involved in the same biological process. For example, many bZIP family TFs share similar binding motifs with the bHLH family TFs, and MYB12 was found to cooperate with HY5 to regulate the expression of genes from the anthocyanin pathway.
3. Prediction
   During the construction procedure of DATF when the HMM profiles for some families were not available, we used DBDs rather than the full-length protein sequence as seed sequences in BLAST search since members of the same TF family may share sequence conservation only at the DBD regions. On the other hand, non-TF proteins which do not contain DBDs may share sequence similarity with TFs in other regions in the flanking region of DBDs. We built the HMM profiles for those families and used them to predict TFs in other genomes. The best score ratio method used for ortholog prediction may result in both false positives and false negatives. On the other hand, phylogenetic methods reported for small-scale analysis are computationally too expensive for large TF families with dozens or even hundreds of members. New approaches such as comparative genomics are being investigated and hopefully can be successfully implemented in the future.
4. Annotation
   The gene regulatory system is so complex that no single current available technology is sufficient to decipher the mechanism behind the complex networks. The simple model of one-to-one relationship between TFs and their binding sites may not reflect the real world of the regulatory network. Not only the same binding motif could be recognized by different TFs from the same or even different families, but

also the same TF could bind to more than one known *cis*-regulatory element. The nonlinear relationship between TFs and their target genes is the basis of co-regulation underlying the very complex life processes. However, most of the current TF databases are using relational database systems such as MySQL. More complex schema or object-oriented database systems are required to handle more sophisticated regulatory network information.

In conclusion, the information provided by PlantTFDB for the predicted TFs should not be taken as a unique reference. Rather, it may serve as a starting point for further biological investigations using experimental approaches of genetic and molecular biology.

References

1. Riechmann, J.L., Heard, J., Martin, G., Reuber, L. et al. (2000) *Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes. *Science* 290, 2105–2110.
2. Badis, G., Berger, M.F., Philippakis, A.A. et al. (2009) Diversity and complexity in DNA recognition by transcription factors. *Science* 324, 1720–1723.
3. Yu, J.Y., Vodyanik, M.A., Smuga-Otto, K. et al. (2007) Induced pluripotent stem cell lines derived from human somatic cells. *Science* 318, 1917–1920.
4. Qu, L.J., and Zhu, Y.X. (2006) Transcription factor families in *Arabidopsis*: major progress and outstanding issues for future research. *Curr Opin Plant Biol* 9, 544–549.
5. Wingender, E., Dietze, P., Karas, H. et al. (1996) TRANSFAC: a database of transcription factors and their DNA binding sites. *Nucleic Acids Res* 24, 238–241.
6. Davuluri, R.V., Sun, H., Palaniswamy, S.K. et al. (2003) AGRIS: *Arabidopsis* gene regulatory information server, an information resource of *Arabidopsis* cis-regulatory elements and transcription factors. *BMC Bioinformatics* 23, 25.
7. Iida, K., Seki, M., Sakurai, T., Satou, M. et al. (2005) RARTF: database and tools for complete sets of *Arabidopsis* transcription factors. *DNA Res* 12, 247–256.
8. Rushton, P.J., Bokowiec, M.T., Laudeman, T.W. et al. (2008) TOBFAC: the database of tobacco transcription factors. *BMC Bioinformatics* 9, 53.
9. Riaño-Pachón, D.M., Ruzicic, S., Dreyer, I. et al. (2007) PlnTFDB: an integrative plant transcription factor database. *BMC Bioinformatics* 8, 42.
10. Guo, A.Y., Chen, X., Gao, G. et al. (2008) PlantTFDB: a comprehensive plant transcription factor database. *Nucleic Acids Res* 36, D966–D969.
11. Gong, W., Shen, Y.P., Ma, LG. et al. (2004) Genome-wide ORFeome cloning and analysis of *Arabidopsis* transcription factor genes. *Plant Physiol* 135,773–782.
12. Guo, A., He, K., Liu, D. et al. (2005) DATF: a database of *Arabidopsis* transcription factors. *Bioinformatics* 21, 2568–2569.
13. Gao, G., Zhong, Y., Guo, A., Zhu, Q. et al. (2006) DRTF: a database of rice transcription factors. *Bioinformatics* 22, 1286–1287.
14. Zhu, Q.H., Guo, A.Y., Gao, G. et al. (2007) DPTF: a database of poplar transcription factors. *Bioinformatics* 23, 1307–1308.
15. Quevillon, E., Silventoinen, V., Pillai, S. et al. (2005) InterProScan: protein domains identifier *Nucleic Acids Res* 33, W116–W120.
16. Finn, R.D., Tate, J., Mistry, J. et al. (2008) The Pfam protein families database. *Nucleic Acids Res* 36, D281–D288.
17. The Gene Ontology Consortium. (2000) Gene ontology: tool for the unification of biology. *Nat Genet* 25, 25–29.
18. Kraulis, P.J. (1991) MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J Appl Cryst* 24, 946–950.
19. Altschul, S.F., Madden, T.L., Schäffer, A.A. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25, 3389–3402.
20. Durbin, R., Eddy, S., Krogh, A. et al. (1998) *Biological sequence analysis: probabilistic models of proteins and nucleic acids.* Cambridge University Press, Cambridge, MA.
21. Larkin, M.A., Blackshields, G., Brown, N.P. et al. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* 23, 2947–2948.