7212–7219 Nucleic Acids Research, 2013, Vol. 41, No. 15 Published online 17 June 2013
doi:10.1093/nar/gkt505

Genome-wide survey of DNA-binding proteins in *Arabidopsis thaliana*: analysis of distribution and functions

Sony Malhotra and Ramanathan Sowdhamini*

National Centre for Biological Sciences (TIFR), UAS-GKVK Campus, Bellary Road, Bangalore 560 065, India

Received March 13, 2013; Revised May 11, 2013; Accepted May 15, 2013

ABSTRACT

The interaction of proteins with their respective DNA targets is known to control many high-fidelity cellular processes. Performing a comprehensive survey of the sequenced genomes for DNA-binding proteins (DBPs) will help in understanding their distribution and the associated functions in a particular genome. Availability of fully sequenced genome of *Arabidopsis thaliana* enables the review of distribution of DBPs in this model plant genome. We used profiles of both structure and sequence-based DNA-binding families, derived from PDB and PFam databases, to perform the survey. This resulted in 4471 proteins, identified as DNA-binding in *Arabidopsis* genome, which are distributed across 300 different PFam families. Apart from several plant-specific DNA-binding families, certain RING fingers and leucine zippers also had high representation. Our search protocol helped to assign DNA-binding property to several proteins that were previously marked as unknown, putative or hypothetical in function. The distribution of *Arabidopsis* genes having a role in plant DNA repair were particularly studied and noted for their functional mapping. The functions observed to be overrepresented in the plant genome harbour DNA-3-methyladenine glycosylase activity, alkylbase DNA *N*-glycosylase activity and DNA-(apurinic or apyrimidinic site) lyase activity, suggesting their role in specialized functions such as gene regulation and DNA repair.

INTRODUCTION

The cell is a complex machine, possessing various macromolecules like proteins, nucleic acids, lipids and so forth. These macromolecules interact with each other and help in maintaining the physiological functions of the cell. DNA–protein interaction is known to govern many high fidelity cellular processes like DNA transcription, replication and damage repair. Proteins are known to interact with DNA in both sequence-specific and non-specific manner. The interactions of the protein with the DNA partner can be either nucleotide sequence-specific or non-specific in nature. Transcription factors and restriction enzymes are known to recognize nucleotide bases, whereas other proteins like histones and chromatin-binding proteins are known to exhibit non-specific interactions with the DNA phosphate backbone and are known to bind independent of the nucleotide base sequence. The study of the nature of interactions between DNA and protein will provide insights into the mechanism of base-specific or non-specific recognition of DNA targets, stereochemical aspect of interaction and also the change in DNA shape on interaction with the protein partner.

DNA-binding proteins (DBPs) use various scaffolds that are used to recognize its respective DNA partner. Some examples of the DNA-binding motifs are helix-turn-helix, zinc coordinating, leucine zippers and so forth (1). There are many classification schemes that aim to group the DBPs, either in protein-centric or DNA-centric manner. DNA centric classification focuses on the DNA partner and classifies DBPs based on the properties of DNA. This was proposed by Prabakaran *et al*. (2), where they used different structural descriptors and defined different clusters for the protein–DNA complexes. The majority of the classification schemes proposed for protein–DNA complexes have been protein-centric in nature where properties of the protein partner are analysed in detail, and the groups are based on the type of the DNA-binding motif present in the protein (1). This classification was then revisited and expanded in 2010, where the number of groups increased to nine and the families of DBPs reported were 174 (3). A 'group' here refers to the set of families with similar DNA-binding motif present in the proteins, and the family reflects the biological function. This is the latest classification for DNA–protein complexes that covers ∼1000 at two-tiered

*To whom correspondence should be addressed. Tel: +91 80 23666250; Fax: +91 80 23636462; Email: mini@ncbs.res.in assembly viz. nine groups and 174 families referred as 'structure-based DNA-binding families'.

The structures of DNA–protein complexes do not cover the entire space for DBPs; therefore, it is necessary to include DBP domain families (sequence-based DNA-binding families). A set of such DNA-binding domain families can be extracted from PFam (4) (our unpublished data). Both structure and sequence-based families together are expected to cover the entire space of DBPs.

Here, we study the plant genome *Arabidopsis thaliana*, as it is one of the model organisms and its genome is fully sequenced. Many microarray studies are carried out on this plant; hence, the transcription factors differentially expressed under various environmental conditions can also be studied in detail. We use three different sequence search strategies to identify DBPs from the plant genome. We use both sequence and structure-based DBP families as references to perform functional genomics studies on the genome of *A. thaliana*. We aim to identify the DBPs in the plant genome and analyse their distributions in different families. Associating such properties to the gene products can be of immense importance and helps in assigning the functions. Genome-wide studies help in bridging the gap between sequence and structure information available for a particular genome (5).

After searching for proteins using the structure-based and sequence-based family references and performing stringent validations, we identified 1900 and 4303 proteins in the *A. thaliana* genome with potential DNA-binding properties, respectively. This consolidated set is called as 'At-Dbome' and comprises 4471 proteins. We then performed a detailed analysis of the proteins in At-Dbome for their distribution in different structure and sequence-based families. Several hypothetical proteins were assigned reliable functions. The sequence-based protein domain families in the *Arabidopsis* proteome with no hitherto structural data were also recognized. The list of all identified DBPs and the analysed data is accessible and available as Supplementary Data.

As plants are prone to DNA damage and get exposed to several biotic and abiotic stress conditions, we further focussed on the subset of DBPs that have DNA repair function. We also studied the distribution of DNA repair proteins in different PFam families. This study will provide insights into the nature of interactions between DNA and its protein partner. It will also help in understanding the distribution of DBP families present in the plant genome and their corresponding function.

MATERIALS AND METHODS

The plant genome *A. thaliana* was downloaded from TAIR (6) that encodes ~35 000 proteins. These proteins were used to perform the genome-wide survey for DBPs.

Compilation of DNA-binding families

For performing genome-wide searches, we first obtained structure-based DNA-binding families and the corresponding family representatives from the classification of protein–DNA complexes, as mentioned earlier (3)]. Additionally, the sequence-based DBP domain families were also identified from PFam database by mapping DNA–protein structural complexes to PFam, checking their family definitions and GO annotations (our unpublished data)

Search protocol and its validation

The searches for DBPs using structure-based families were performed using the three sensitive sequence search methods- PSI-BLAST (7), RPS-BLAST (8) and HMMscan of the HMMER3 suite (9).

(1) For performing PSI-BLAST searches, the representatives of structure-based families were used as queries to search the *Arabidopsis* proteome, with an E-value threshold of 10^{-5}.
(2) PSI-BLAST profiles were built for each structure-based DNA-binding family using both an alignment of all family members and the representative sequence as inputs to query against NR database with E-value threshold of 10^{-10}. These profiles were assembled as a database, and the *Arabidopsis* proteome was searched against this database using sequence-profile comparison method, namely, RPS-BLAST with E-value threshold of 10^{-3}.
(3) HMMs for each structure-based DNA-binding family were built using hmmbuild (HMMER3 suite) based on the alignment of all the family members. The *Arabidopsis* proteome was matched against these HMMs using HMMScan with an E-value threshold of 10^{-2}.

Figure 1 shows the overview of the search protocol adopted to perform comprehensive scan in the plant genome.

For the searches performed using sequence-based DNA-binding families, HMMScan was used with E-value threshold of 10^{-5}. The proteins obtained were validated by manually checking the family descriptions and its GO annotation as DNA binding.

Before performing the searches for DBP in *Arabidopsis* proteome, the search strategy was validated for three well-annotated and closely related plant protein families like serine carboxypeptidases and subtilisins. Tripathi and Sowdhamini (10) had earlier annotated 54 and 56 proteins as serine carboxypeptidases and subtilisins, respectively, in *Arabidopsis* proteome. Another well-annotated family selected for validating the search protocol was pectinesterase, which has 59 proteins from *Arabidopsis* identified in GeneFarm database (11). We selected their representatives using the same strategy of Jack-knifing as in Malhotra and Sowdhamini (3) and then carried out the searches to identify members of these families from the *Arabidopsis* proteome. The representative sequence, its PSSM-profile and the HMM model were used to perform searches in the *Arabidopsis* proteome using PSI-BLAST, RPS-BLAST and HMMScan, respectively.

Analysis of proteins identified as DNA binding

For the proteins that were obtained based on structure-based families, two-fold validations were performed.

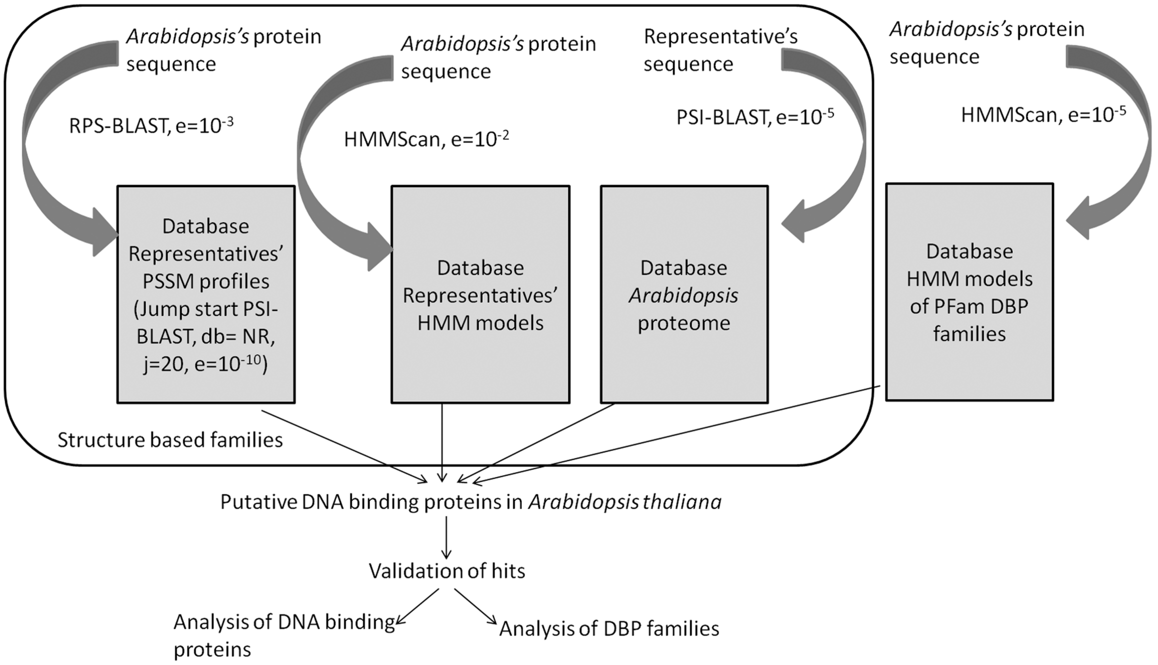
Arabidopsis's protein sequence
RPS-BLAST, e=10^{-3}
Database Representatives' PSSM profiles (Jump start PSI-BLAST, db= NR, j=20, e=10^{-10})
Arabidopsis's protein sequence
HMMScan, e=10^{-2}
Database Representatives' HMM models
Representative's sequence
PSI-BLAST, e=10^{-5}
Database Arabidopsis proteome
Arabidopsis's protein sequence
HMMScan, e=10^{-5}
Database HMM models of PFam DBP families
Structure based families
Putative DNA binding proteins in Arabidopsis thaliana
Validation of hits
Analysis of DNA binding proteins
Analysis of DBP families

Figure 1. Overview of the search protocol: three sensitive sequence search methods, namely, PSI-BLAST, RPS-BLAST and HMMScan were used to perform a comprehensive search for DBPs in *Arabidopsis* genome.

Those proteins that are identified by at least two of the three sequence search methods were marked as true positives. The proteins identified using only one of the methods were further tested for their ability to recognize DBP family, using HMMPfam.

Such validated proteins, obtained based on structure and sequence-based families, form part of the 'At-Dbome' and were analysed for their distribution in different families. Although studying the distribution of At-Dbome proteins in DNA-binding families, we identified the families that were overrepresented in this plant genome. For this, we calculated the normalized occurrence of a PFam family in At-Dbome as compared with PFam. The family was identified as overrepresented if its occurrence in the At-Dbome was at least 10 times its occurrence in PFam.

The DNA repair families were identified based on sequence-based family definitions. The proteins that were putative, hypothetical or unknown in nature, as recorded in TAIR, were further analysed and also validated by checking for the GO annotation for nucleic acid-binding function.

Analysis of DNA-binding families

The distribution of all the members of At-Dbome across various structure and sequence-based DNA-binding families was also studied.

For families in the At-Dbome, GO mapping (12) was obtained, and the annotations for molecular functions were extracted. The over- or underrepresented functions in *Arabidopsis* proteome were identified. For attributing significance to the annotation results, we performed similar genome-wide surveys for three other genomes *Saccharomyces cerevisiae*, *Caenorhabditis elegans* and *Drosophila melanogaster*. GO mapping was obtained for the PFam families that were identified as DNA binding. The odds of occurrence of a GO molecular function in a particular genome as well as in full set of sequence-based DNA-binding families were calculated to find over- and underrepresented functions. The log odds score was then used as a measure to assess the representation of a given molecular function. If a function was two-fold over- or underrepresented, it was considered to be significant.

The sequence-based families were further mapped to PDB to analyse whether they have a known structure(s) mapped to them. The PFam DNA-binding families from *A. thaliana,* with no available data on their 3D structures, were further clustered together (our unpublished data).

RESULTS

Validation of the search strategy

Using three different sequence comparison methods PSI-BLAST, RPS-BLAST and HMMScan, we performed searches in *Arabidopsis* proteome. The *Arabidopsis* genome was initially searched for the three families with closely related members—serine carboxypeptidase, subtilisin and pectinesterase. There are 56, 54 and 59 proteins, known from previous studies, in the *Arabidopsis* proteome annotated as serine carboxypeptidase, subtilisins and pectinesterase, respectively. All the three methods were able to identify the already annotated members and some additional members as well (Supplementary Table S1). In addition, for the family pectinesterase, all the three methods were able to identify all previously annotated proteins, except one that is now marked as pseudogene in *Arabidopsis* genome by TAIR. These well-annotated families were chosen for testing the search strategy. Further, at such closely related family-level relationships, question of divergent or convergent evolution does not arise.

At-Dbome: set of *Arabidopsis* DBPs and their distribution

In all, 174 structure-based DBP families and their 192 representatives were obtained from Malhotra and Sowdhamini (3). The 1219 sequence-based DBP families from PFam were also collected based on the family descriptions and mapping structural families to PFam domains (our unpublished data). PSSM profiles and HMM models were generated for all the families using thresholds as described in 'Materials and Methods' section.

After performing searches in *Arabidopsis* proteome and carrying out further validations, we identified 1900 and 4303 DBPs (Supplementary Table S2) based on the structure and sequence-based family definitions, respectively, constituting the 'At-Dbome' data set. The 1732 proteins in the dataset were obtained based on both PDB and PFam definitions (Supplementary Table S2, marked with both the families). This gives rise to a consolidated set of 4471 genes in the 'At-Dbome' marked as DBPs (Supplementary Table S2) that are distributed among 300 DNA-binding families (Supplementary Table S3).

In the set of proteins in At-Dbome identified using structural families, 20% (376) were identified by all three sequence search methods. Forty-seven per cent of the proteins were identified as DBP by only one of the methods and were validated by performing HMMScan against the PFam HMM models to check whether they identify the DBP family (Table 1). In the At-Dbome, we studied the distribution of proteins in structural groups and families. They were observed to belong to the eight groups and 57 structural families (Figure 2). The three most populated groups were helix-turn-helix, enzymes and β-propeller.

At-Dbome proteins were observed to belong to 300 PFam families and 57 clans. The clans that were observed to be highly populated in At-Dbome were HTH and P-loop NTPase. These clans that were highly populated in At-Dbome contain families that are essential for DNA replication, transcription and DNA repair. P-loop NTPase clan harbours protein families like helicase, DNA polymerase and many DNA repair enzymes and HTH clan comprises mainly transcription factor families like E2F, Myb, HSF and so forth.

We also studied the distribution of At-Dbome sequences in 300 PFam families and identified overrepresented families. There were 22 PFam families that were observed to be highly overrepresented in At-Dbome as compared with PFam. Sixty-eight per cent of these families are reported to be plant specific in nature, which truly reflects its overrepresentation in At-Dbome (Table 2 and Supplementary Figure S1). Supplementary Figure S1 highlights families that are underrepresented or overrepresented 10-fold in At-Dbome.

Table 1. The number of proteins in *A. thaliana* genome that were identified as DNA-binding using three different methods

Set	Method identifying the protein	Number of proteins	Proteins belonging to same group and family
I	HMMScan, PSI-BLAST, RPS-BLAST	376	376
II	HMMScan, PSI-BLAST	463	463
	HMMScan, RPS-BLAST	600	598
	PSI-BLAST, RPS-BLAST	686	685
III	PSI-BLAST	271	258
	RPS-BLAST	453	290
	HMMScan	618	358

The proteins were divided into three sets I, II and III depending on the number of methods that identify them as DNA binding.

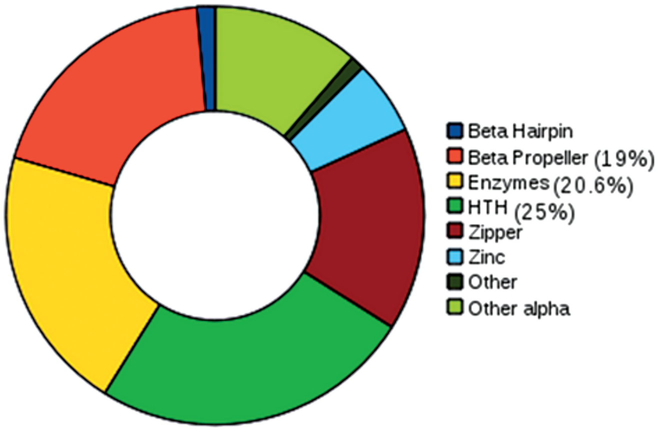

Figure 2. Distribution of proteins: the proteins in *Arabidopsis* genome that were identified DNA-binding on performing searches using the structure-based families were studied for their distribution in structural groups and families. The highest populated group was helix-turn-helix.

For the β-sheet group, we were not able to identify any DBP in the plant genome using structure-based families. The structural families included in group β-sheet were: TATA box binding, accessory gene regulator protein A, AP2 protein. The representatives for these three families, as suggested by Malhotra and Sowdhamini (3), were mapped to the PFam domains, and 146 proteins identified using HMM models of the PFam domains were included. These 146 proteins were further validated to be true positives by checking their gene description in TAIR. They were reported as proteins belonging to AP2/EREB family of transcription factors that are covered in β-sheet group.

Hypothetical proteins identified as DBPs in At-Dbome

The proteins that were marked as unannotated if their gene descriptions describe them as hypothetical, putative, unknown or predictive in function in TAIR were next examined. There are 6507 unannotated proteins in the *Arabidopsis* genome.

Table 2. Overrepresented DNA-binding PFam families in At-Dbome

Pfam ID	Pfam Name	Normalized occurrence	Description
PF13724	DNA_binding_2	42.86	This domain, often found on ovate proteins, which is a plant Ku70 interacting protein involved in DNA double-strand break repair
PF08744	NOZZLE	25.89	NOZZLE is a transcription factor that plays a role in patterning the proximal–distal and adaxial–abaxial axes
PF04689	S1FA	24.53	S1FA is a DBP found in plants that specifically recognizes the negative promoter element S1F
PF04618	HD-ZIP_N	23.90	Homeodomain leucine zipper (HDZip) genes encode putative transcription factors that are unique to plants.
PF02362	B3	20.47	The B3 DNA-binding domain (DBD) is a highly conserved domain found exclusively in transcription factors, from higher plants
PF02365	NAM	19.24	NAM transcription factors are plant development proteins.
PF00097	zf-C3HC4	19.10	Zinc finger
PF06217	GAGA_bind	19.08	This family includes gbp a protein from soybean that binds to GAGA element dinucleotide repeat DNA
PF04640	PLATZ	16.72	Plant AT-rich sequence and zinc-binding proteins (PLATZ) are zinc-dependant DBPs. They bind to AT-rich sequences and functions in transcriptional repression
PF02701	zf-Dof	16.67	Zinc finger found in several DBPs of higher plants
PF07716	bZIP_2	16.36	Basic leucine zipper
PF06200	tify	15.58	The tify domain is a 36-amino acid domain only found among Embryophyta (land plants).found in a variety of plant transcription factors that contain GATA domains
PF02183	HALZ	15.30	Plant-specific leucine zipper that is always found associated with a homeobox
PF13921	Myb_DNA-bind_6	15.07	MYB like DNA-binding domain
PF14215	bHLH-MYC_N	13.24	MYB and MYC family regulate the biosynthesis of phenylpropanoids in several plant species
PF03859	CG-1	12.18	Sequence-specific DBP
PF03110	SBP	12.08	Plant-specific transcription factors
PF13639	zf-RING_2	11.91	RING finger domain
PF08879	WRC	11.76	WRC is named after the conserved Trp-Arg-Cys motif, it contains two distinctive features a putative nuclear localization signal and a zinc-finger motif (C3H). It is suggested that WRC functions in DNA-binding
PF07777	MFMR	11.17	Multifunctional mosaic region
PF02309	AUX_IAA	11.06	Plant-specific, repressors of auxin induces gene expression
PF02536	mTERF	10.53	Leucine zipper

DBPs' families in At-Dbome were analysed for their occurrence in *Arabidopsis* genome and in PFam. Normalized ratio of occurrence of a given DNA-binding family in At-Dbome was calculated. This table shows the 22 PFam families (with their description and normalized occurrences) that were observed to be overrepresented in the plant genome.

At-Dbome contained 142 such unannotated proteins in At-Dbome that are reported as hypothetical, putative or unknown function by TAIR (Supplementary Table S4). We checked the GO annotations of these proteins if these proteins were annotated as nucleic acid binding. Of the full set of 142 hypothetical proteins in At-Dbome, only 34 proteins were annotated for their molecular functions or biological process in GO database. Twenty per cent of these 142 hypothetical proteins were observed to be annotated as nucleic acid binding or DNA binding in GO (Supplementary Table S5).

DNA-repair proteins in At-Dbome

As plants are generally vulnerable to UV-induced stress, we next analysed At-Dbome for the distribution of DBPs with DNA repair function attributed to them. Three hundred DBP families in At-Dbome were further checked for their GO annotations to identify the families that are involved in the DNA repair function. This resulted in identification of 36 families possessing DNA repair activity. These 36 families are annotated in GO as DNA repair, damaged DNA-binding, double-strand break repair, base excision repair, nucleotide excision repair, DNA glycosylase activity, double-strand break repair via non-homologous end joining, DNA mis-repair or DNA recombination. One of the PFam family, WD40, is annotated as protein binding. However, in Malhotra and Sowdhamini (3), β-propeller was identified as a structural group where a DNA damage repair protein, DDB2, uses its seven-bladed propeller to bind the damaged DNA and then acts as a landing pad for other DNA repair proteins (13). There were 332 proteins belonging to WD40 family, in At-Dbome. However, we considered proteins that have at-least one other DNA-binding domain co-existing with WD40 to attribute the DNA repair function.

There were total of 226 proteins in At-Dbome that are involved in DNA repair mechanisms. We further studied the distribution of At-Dbome proteins in these 36 DNA repair families (Figure 3). The most populated family for DNA repair proteins in the At-Dbome was RuvB_N, which resolves the Holliday junctions during genetic recombination and DNA repair (14). This family belongs to PFam clan P-loop-NTPase that we observed as highly populated clan in the At-Dbome (as mentioned earlier).

Functional annotation of *Arabidopsis* DBP families

The proteins in At-Dbome were observed to belong to 300 sequence-based families. These families were further

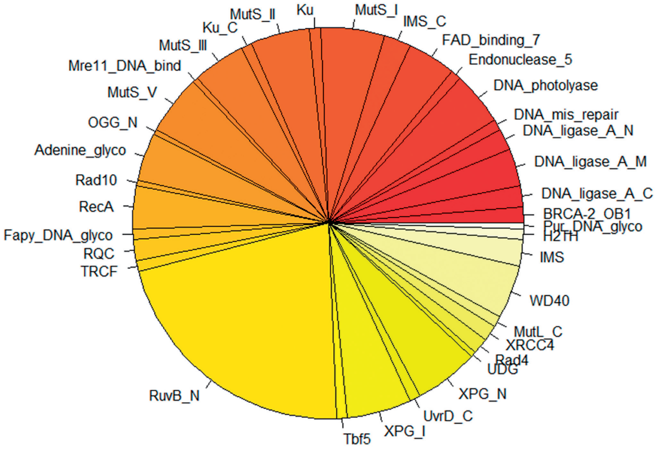

Figure 3. Distribution of DNA repair proteins: the proteins in At-Dbome were further analysed for their involvement in DNA repair processes. This is the subset of proteins having DNA repair function and their distribution in different families. The most populated family was observed to be RuvB_N.

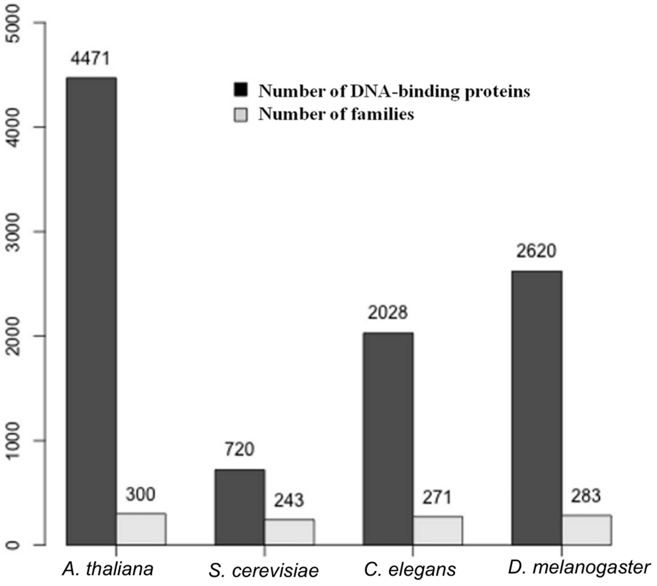

Figure 4. Genome-wide survey in four genomes: using the similar search protocol and using sequence-based families; genome-wide survey was performed in three other genomes, namely, *C. elegans*, *D. melanogaster* and *S. cerevisiae*. The number of proteins identified as DBPs in *A. thaliana, S. cerevisiae, C. elegans* and *D. melanogaster* were 4471, 720, 2028 and 2620, respectively.

analysed for their functional annotations, and the molecular functions were derived using GO mapping.

As described in 'Materials and Methods' section, genome-wide survey was performed in three other genomes *S. cerevisiae*, *C. elegans* and *D. melanogaster* using the 1219 DNA-binding families in PFam. This resulted in the identification of 720, 2028 and 2620 DBPs in the *S. cerevisiae*, *C. elegans* and *D. melanogaster* genomes, respectively (Figure 4). These proteins were studied for their distribution in PFam-based families, and each of this family was further manually validated for their DNA-binding function.

There were 44 common GO functions observed for all the four genomes. Supplementary Figure S2 depicts these function and their log(odds) ratio in the four genomes studied here. Further, we studied in detail the functions that were overrepresented in the individual genomes and also in all the four genomes.

The function annotated and overrepresented only in the plant proteome was DNA-3-methyladenine glycosylase activity, alkylbase DNA *N*-glycosylase activity and DNA-(apurinic or apyrimidinic site) lyase activity (Figure 5). The alkylbase DNA *N*-glycosylase activity is associated with the removal of transcriptional repression caused because of cytosine methylation. Methylation is known as a mechanism for regulating gene expression in plants and animals. Both animal and plants are known to have similar methylating machinery counterparts, whereas there are different mechanisms known for demethylation. This process of relieving transcription repression with the help of the glycosylases is known only in plants (15); therefore, we see this function as overrepresented in the plant proteome. These functions that were observed to be overrepresented only in plant proteome as compared with the other genomes are associated with DNA repair processes; this highlights the overrepresentation of DNA repair functions in plant genome. Likewise, there were, four, one and seven functions observed to be overrepresented only in fly, *C. elegans* and yeast genome (Supplementary Table S6).

There were two functions that were observed to be underrepresented in plant genome- 3′–5′ exonuclease activity and transposase activity (Figure 5). We studied the families corresponding to these in At-Dbome, and they correspond to bacterial enzymes. Therefore, we see them underrepresented in plant genome. The 3′–5′ exonuclease activity maps to PFam family DNA_pol_A_exo1, which is a bacterial domain involved in proof reading activity of bacterial DNA polymerase (16). This domain is present in bacterial species and underrepresented in plant genomes in PFam as well. Another underrepresented GO function, transposase activity maps to DDE_Tnp_1 that is a family of transposases that was originally identified in bacteriophages and is reported to be present in bacterial species (17).

We also analysed the GO molecular functions that were overrepresented in all the four genomes (Supplementary Figure S1). There were 17 functions that were observed to be present in all the genomes to maintain the homeostasis in the cell like DNA-clamp loader, DNA ligase, nucleic acid binding, transcription factor binding, DNA topoisomerase activity, DNA polymerase activity and so forth. These activities are important for the normal functioning of the cell and involve processes like DNA replication and transcription.

DISCUSSION

In 2000, Thornton and co-workers (1) identified a limited number of DNA-binding motifs, for example,

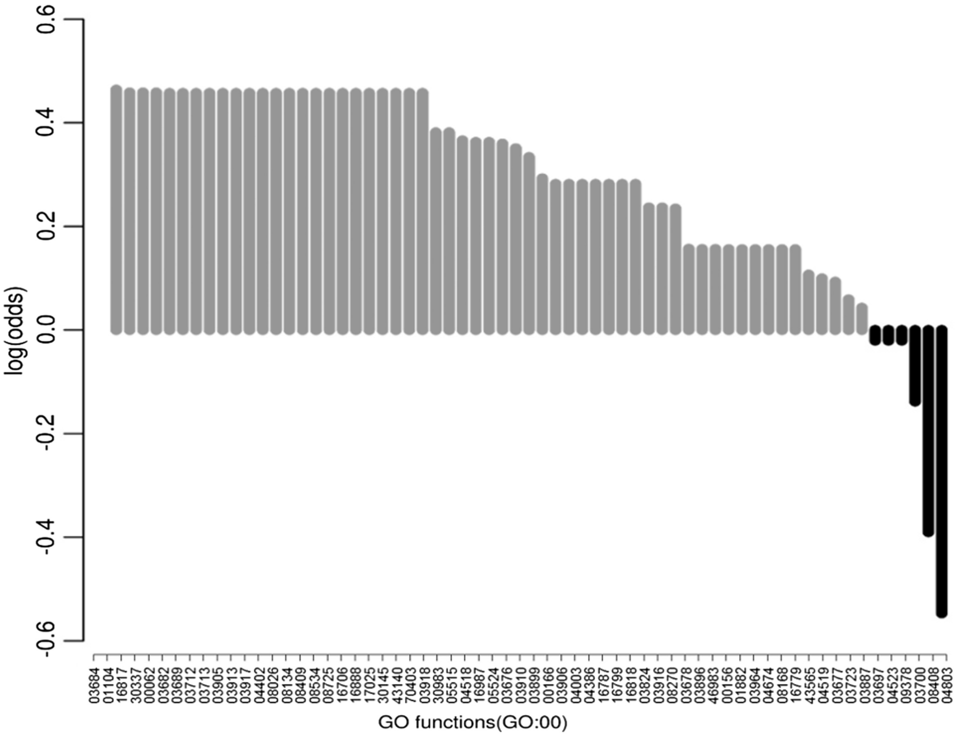

Figure 5. Log odds score for GO molecular functions: the proteins in At-Dbome were mapped to their PFam families. The GO mapping for these families was performed, and we studied the over and underrepresented functions in *Arabidopsis* genome. The functions that were observed to be overrepresented in plant genome were involved in DNA repair mechanisms.

helix-turn-helix, zinc coordinating, zipper type, β-sheet and β-hairpin/ribbon. However, the functions of these proteins with similar motifs are diverse in nature (which is highlighted by the number of families within these DNA-binding motif-based groups). We revisited this classification in 2010 (3), with nearly five times the number of DNA–protein complexes. We identified a new type of DNA-binding motif (β-propeller group), but the families within these motif-based groups increased by three times (174 versus 54). Therefore, the structural motif recognizing the DNA target remains conserved, whereas the newer functions evolve over time. Therefore, majority of DBP follow divergent evolution.

The genome-wide scan for the presence of DBPs was performed in the plant proteome *A. thaliana* starting from structure-based and sequence-based families, for genes with a putative DNA-binding property. For performing this study, we used both structural and sequence-based families. Structure-based families consists of the DBPs that have their structures solved as a complex and deposited in PDB, whereas sequence-based families correspond to the families gathered from protein domain family database (PFam).

To perform the genome-wide searches in the plant proteome, a search strategy was used that was validated for three well-known families in *Arabidopsis*. The search protocol was observed to identify the already annotated members for those three families. After carefully validating the search protocol, searches were performed in the plant proteome for DBPs, and then the proteins identified were further assessed, as described in 'Materials and Methods' section. Together, these identified DBPs, using both the PDB and PFam family definitions, constitute At-Dbome.

At-Dbome was observed to possess two subsets of identified proteins, 1900 proteins and 4303 proteins that were identified as DBP using PDB and PFam definitions, respectively. Their distribution was studied in the structural and sequence families. They were observed to be distributed in 57 different structural families and 300 sequence-based families. The 1732 proteins were common in these two sets. Therefore, At-Dbome has a consolidated set of 4471 identified DBPs of *A. thaliana* genome that are distributed in 300 PFam families.

Two kinds of analyses were carried out on At-Dbome, first, at the level of the proteins that were identified as DBP, and second the families where the proteins belong to were analysed in detail.

(i) Proteins identified as DBPs:
   (a) They were carefully validated using the three different sensitive sequence search methods.
   (b) The family-wise distribution for the identified and validated DBPs was studied. Also the distribution of DBPs in At-Dbome was studied for different clans.
   (c) The proteins that are marked as unknown, hypothetical or putative in function by TAIR and were identified as DBP by our search protocol were studied in detail.

(d) The subset of DBPs that have a role in plant DNA repair was studied in detail, and their distribution in different PFam families was studied.

(ii) Families possessing the identified DBPs:

The families were analysed in detail for their molecular function to study the over and underrepresented functions in the plant proteome. There were four functions that were observed to be overrepresented in the four genomes analysed for the presence of DBPs. These functions are involved in maintaining the cell homeostasis.

The overrepresented functions observed only in the plant proteome were involved in mechanism for relieving transcription repression, which is known only in plants and in various DNA repair processes.

The present work is a comprehensive study of DBPs in *Arabidopsis* genome. Identification of the DBPs in a particular genome helps us in studying their distribution and analysing the functions performed by them. In this study, both the sequence and structure-based families of DBPs were used to identify DBPs from *Arabidopsis* genome using only sequence information. This will enable us to have a better understanding of the previously uncharacterized DBPs encoded in the plant genome.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–6 and Supplementary Figures 1–2.

ACKNOWLEDGEMENTS

Both authors acknowledge NCBS for infrastructural support.

FUNDING

Department of Biotechnology (to S.M.). Funding for open access charge: National Centre for Biological Sciences (NCBS).

Conflict of interest statement. None declared.

REFERENCES

1. Luscombe,N.M., Austin,S.E., Berman,H.M. and Thornton,J.M. (2000) An overview of the structures of protein-DNA complexes. *Genome Biol.*, **1**, REVIEWS001.
2. Prabakaran,P., Siebers,J.G., Ahmad,S., Gromiha,M.M., Singarayan,M.G. and Sarai,A. (2006) Classification of protein-DNA complexes based on structural descriptors. *Structure*, **14**, 1355–1367.
3. Malhotra,S. and Sowdhamini,R. (2012) Re-visiting protein-centric two-tier classification of existing DNA-protein complexes. *BMC Bioinformatics*, **13**, 165.
4. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al.* (2009) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
5. May,A.C., Johnson,M.S., Rufino,S.D., Wako,H., Zhu,Z.Y., Sowdhamini,R., Srinivasan,N., Rodionov,M.A. and Blundell,T.L. (1994) The recognition of protein structure and function from sequence: adding value to genome data. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **344**, 373–381.
6. Lamesch,P., Berardini,T.Z., Li,D., Swarbreck,D., Wilks,C., Sasidharan,R., Muller,R., Dreher,K., Alexander,D.L., Garcia-Hernandez,M. *et al.* (2011) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.*, **40**, D1202–D1210.
7. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
8. Marchler-Bauer,A., Panchenko,A.R., Shoemaker,B.A., Thiessen,P.A., Geer,L.Y. and Bryant,S.H. (2002) CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.*, **30**, 281–283.
9. Eddy,S.R. (2011) Accelerated Profile HMM Searches. *PLoS Comput. Biol.*, **7**, e1002195.
10. Tripathi,L.P. and Sowdhamini,R. (2006) Cross genome comparisons of serine proteases in Arabidopsis and rice. *BMC Genomics*, **7**, 200.
11. Aubourg,S., Brunaud,V., Bruyère,C., Cock,M., Cooke,R., Cottet,A., Couloux,A., Déhais,P., Deléage,G., Duclert,A. *et al.* (2005) GeneFarm, structural and functional annotation of Arabidopsis gene and protein families by a network of experts. *Nucleic Acids Res.*, **33**, D641–D646.
12. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.*, **25**, 25–29.
13. Scrima,A., Konícková,R., Czyzewski,B.K., Kawasaki,Y., Jeffrey,P.D., Groisman,R., Nakatani,Y., Iwai,S., Pavletich,N.P. and Thomä,N.H. (2008) Structural basis of UV DNA-damage recognition by the DDB1-DDB2 complex. *Cell*, **135**, 1213–1223.
14. Dickman,M.J., Ingleston,S.M., Sedelnikova,S.E., Rafferty,J.B., Lloyd,R.G., Grasby,J.A. and Hornby,D.P. (2002) The RuvABC resolvasome. *Eur. J. Biochem.*, **269**, 5492–5501.
15. Chan,S.W., Henderson,I.R. and Jacobsen,S.E. (2005) Gardening the genome: DNA methylation in *Arabidopsis thaliana*. *Nat. Rev. Genet.*, **6**, 351–360.
16. Moser,M.J., Holley,W.R., Chatterjee,A. and Mian,I.S. (1997) The proofreading domain of Escherichia coli DNA polymerase I and other DNA and/or RNA exonuclease domains. *Nucleic Acids Res.*, **25**, 5110–5118.
17. Ferrante,A.A. and Lessie,T.G. (1991) Nucleotide sequence of IS402 from *Pseudomonas cepacia*. *Gene*, **102**, 143–144.