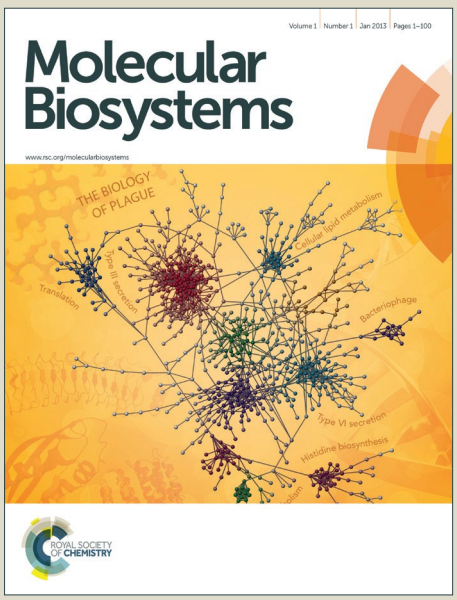
Molecular BioSystems

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

Cite this: DOI: 10.1039/c0xx00000x

www.rsc.org/xxxxxx

FULL PAPER

Collation and analyses of DNA-binding protein domain families from sequence and structural databanks

Sony Malhotra[a] and Ramanathan Sowdhamini[b]*

Received (in XXX, XXX) Xth XXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX

DOI: 10.1039/b000000x

DNA-protein interactions govern several high fidelity cellular processes like DNA-replication, transcription, DNA repair etc. Proteins that have an ability to recognise and bind DNA sequences can be classified either according to their DNA-binding motif or based on the sequence of the target nucleotides. We have collated the DNA-binding families by integrating information from both protein sequence family and structural databases. This resulted in a dataset of 1057 DNA-binding protein domain families. Their family properties (number of members, percent identity distribution and length of members) and domain architectures were examined. Further, sequence domain families were mapped to structures in the protein databank (PDB) and the protein domain structure classification database (SCOP). The DNA-binding families, with no structural information, were clustered together into potential superfamilies based on sequence associations. On the basis of functions attributed to DNA-binding protein folds, we observe that a majority of the DNA-binding proteins follow divergent evolution. This study can serve as a basis for annotation and distribution of DNA-binding proteins in genome(s) of interest. The entire collated set of DNA-binding protein domains is available for download as Hidden Markov Models.

Introduction

Proteins are known to perform a diverse variety of cellular functions to maintain the structural and functional integrity of the cell. They are comprised of independent folding units, which are known as domains.[1,2] There have been two highly accessed classifications of protein structural domains, namely SCOP (structural classification of proteins)[3] and the CATH protein structure classification (class architecture topology homologous family).[4,5] These two resources propose hierarchical classification systems of protein structural domains. On the other hand, protein family database (Pfam) classifies protein families on the basis of sequence features. Pfam database provides multiple sequence alignments and Hidden Markov Models (HMM) of protein sequence domain families. The related families, in terms of HMM profile similarities, are assembled into clans in Pfam.[6,7] These well-classified families can be used to study specific protein families, their functions, taxonomic distribution, domain architectures and to annotate available or newly sequenced genomes.

The specialised class of proteins with an ability to bind DNA, are known to govern many vital cellular functions like DNA replication, transcription, translation, DNA repair *etc.*[8,9] DNA-binding proteins are known to bind DNA partner through a number of structural motifs like helix-turn-helix, leucine zippers, Zinc-coordinating motif etc.[9,10] There are nine DNA-binding structural motifs reported in the literature and they have been studied extensively.[9,10] These proteins bind to their DNA targets in both sequence specific and non-specific manner.[8] Transcription factors and restriction enzymes are known to recognise specific nucleotide sequence, whereas chromatin binding proteins like histones recognise sugar-phosphate backbone and therefore bind DNA non-specifically.[11,12]

With the advancement in DNA-sequencing technology, there has been an increase in the availability of fully sequenced genomes. DNA-binding protein families constitute a majority of genomes in both eukaryotes[13,14] and prokaryotes.[15–17] The distribution of DNA-binding proteins is observed to vary across genomes and species-specific preferences can be also detected. Therefore, annotating DNA-binding proteins in newly sequenced or available genomes will help in understanding many important cellular functions and their regulation in the cell. There have been various attempts to invent repositories for transcription factors, so as to annotate transcription factor families in several genomes.[18–27]

There is a continuing need for a well-defined classification of existing DNA-binding proteins (DBP) as a starting point to accomplish searches for DBP in a given genome of interest. In 2000, Thornton and co-workers proposed a protein structure-centric classification of DNA-protein complexes.[9] This classification scheme listed the DNA-binding motifs employed

[a] *National Centre for Biological Sciences, Bellary Road, GKVK Campus, Bangalore, India. Tel: 91 80 2366 6250; E-mail: sonym@ncbs.res.in*
[b] *National Centre for Biological Sciences, Bellary Road, GKVK Campus, Bangalore, India. Fax: 91 80 2363 6462; Tel: 91 80 2366 6250; E-mail: mini@ncbs.res.in*

by the proteins as eight groups and then sub-classified them into 54 families that reflected their biological functions. We recently revisited this classification scheme[10] and proposed an additional DNA-binding motif (β-propeller) and about thrice the number of families (174). However, the number of available structures of DNA-protein complexes is much less than the sequence information available for the "DBPome".

Therefore, to cover the entire space of DBP families, sequence families of DNA-binding domains from Pfam were integrated with the existing structural families. To accomplish this, we mapped all the well-defined structural families of DBP to Pfam sequence domain families. Subsequently, Pfam was searched for DNA-binding function to identify new DNA-binding families, which was further verified with the help of GO annotations.[28] The complete set of DBP families was analysed for domain architectures, taxonomic distribution and functions. The fold space covered by these families indicates that a multitude of functions are performed using the same fold, thereby supporting a divergent mode of evolution. The DNA-binding protein families with no structure information (none of the members have a solved structure) must be significantly substantial and were waiting for attention. They were, therefore, clustered into putative sequence-based superfamilies using HMMScan.[29]

These data can be downloaded from http://caps.ncbs.res.in/DBPome. This will aid in determining the highly populated cluster(s), and proposing the targets for the structural genomics initiative (http://kb.psi-structuralgenomics.org/).

Methods

Collation of DNA-binding families

DNA-binding protein families were obtained by employing a three-fold strategy.

(i) Vaquerizas *et al.*[18] in 2009 performed a study to perform a census of transcription factor families for human genome. Therefore, they collected all the transcription factor families from InterPro[30]. These families were mapped to Pfam families and were included in the dataset of DBP families.

(ii) Starting from our previous protein-centric classification of DNA-protein complexes, the sequences of interacting partner protein were subjected to HMMScan (HMMER3 suite[29]) at an E-value of 10^{-5} against the database constituting HMM of protein sequence domains family database (Pfam v26). The resulting Pfam families form a subset of DBP families.

(iii) Pfam database was searched for DNA-binding functions using keyword search. Further, the families were validated using GO annotation and Pfam abstract description for DNA-binding function.

Analyses of DNA-binding families:

DBP families identified from Pfam, were further analyzed for their family architecture. For this, we studied the distribution of members within DBP families, length of members and percent identity between them using CLUSTALW2.[31]

Studying their distribution in Pfam clans will help in identification of the relationships between these DBP families. Domain architectures associated with a Pfam domain family were extracted from Pfam. The taxonomic distribution across different kingdoms was obtained by mapping protein domain sequences in Pfam family to UniProt[32] and then their NCBI taxonomy[33,34] was obtained.

Structural mapping of DNA-binding families: SCOP mapping

For the DNA-binding sequence domain families identified in Pfam, we were interested in identifying the structural motif employed to bind DNA. These families were mapped to structures at two levels, firstly at SCOP level and then at PDB level.

SCOP (v1.75) domain sequences, filtered at 40% identity, were related to Pfam families by carrying out sequence-HMM against family-HMM models of Pfam DBP families, using an E-value threshold of 10^{-5}. The association of Pfam DBP families to SCOP was studied for their distribution across various hierarchies of SCOP, i.e. class, fold, superfamily and family.

Structural mapping of DNA-binding families: PDB mapping

The dataset of DBP domain families, collated using the above strategy, was further mapped to the structures in PDB. The sequence domain families with no known structural information were clustered together. The seed sequences of families with unknown structures were collected and all-against-all sequence-HMM comparisons were performed at E-value 10^{-4}. The clustering of these families was performed based on the reciprocal hit approach i.e. seed sequence of family-A recognizes HMM profile of family-B and vice versa, this places A and B in one cluster. Therefore, by validating these associations, we clustered families with unknown structures into new sequence-based superfamilies.

Tracing the mode of evolution

The GO molecular functions for SCOP folds pertaining to DBP-protein domain families were obtained from Superfamily database[35] using an information content threshold of 2.0. The distribution of these functions associated with different DNA-binding folds, was next analyzed.

Results and Discussion

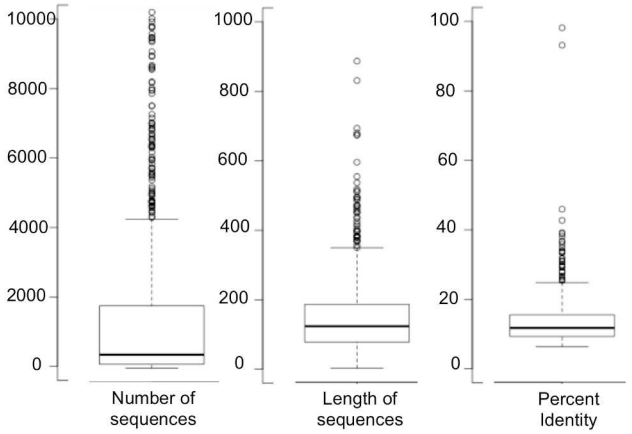
Set of DBP families and validation

A full set of DBP families was gathered using approaches (as described in Methods), which resulted in three subsets of DBP families. The first subset was derived from a census of human transcription factor families, where Vaquerizas *et al.*[18] identified 347 transcription factor families from InterPro. We identified 162 Pfam families corresponding to this dataset. Secondly, structural DBP families from Malhotra and Sowdhamini[10] were mapped to Pfam families using HMM-sequence comparisons (HMMScan). Lastly, keyword searches were performed in Pfam database for DNA-binding function and the results were validated using Pfam family definitions and GO annotations. This resulted in a merged dataset of DBP families using three approaches containing 1057 DBP families.

Analyses of DBP families in DBome

The set of DBP families were further investigated for their family properties and domain architectures. These analyses will provide useful insights to understand the distribution and functions of proteins in different DBP families.

Figure 1: **DBP family architecture:** The collated DBP families were



studied for their family architecture described using three features namely (a) number of family members, (b) length of family members and (c) the sequence identities between family members (for families with less than 5000 sequences).

Family features

DBP families were studied for their features. Three features, namely the number of members in the family, length of family members and percent identity among family members were quantified (Figure 1). We observed that the average number of members in DBP family is 1500. We also studied length distribution in DBP families and the average length was observed to be 170 amino acids.

The DBP families were examined for the extent of divergence of its members. The percentage identities between different family members were calculated using CLUSTALW2.[31] The families were observed to be very diverse in nature, as the average sequence identity was only 17% and ~33% of the families have sequence identity less than 10%. The examples of the diverse families include the TEA domain and DNA methylase (N6_N4_Mtase) with an average percent identity of 6.7% and 6.8%, respectively. TEA domain exhibits sequence-specific DNA-binding transcription factor activity[36] and methylases in bacteria confer protection to host DNA against restriction enzymes by methylating bases like adenine (N-6 adenine-specific DNA methylase) and cytosine (N-4 cytosine-specific DNA methylase) [37].

However, two of the families stand out in the percent identity plot (Figure 1, the two outliers represented as two isolated circles) and exhibit very high percent identity. These families were spermatozoal protamine and elongation factor SelB. Spermatazoal protamine family contains proteins, which help in sperm chromatin condensation during spermatogenesis[38] and elongation factor SelB possesses a winged helix DNA-binding motif with three-helical bundle fold and a small β-sheet wing.[39] Both of these families appear to be newly recognized families since they are small containing two-members each with very high sequence identity between them (98% for elongation factor and 93% for spermatozoal protamine family).

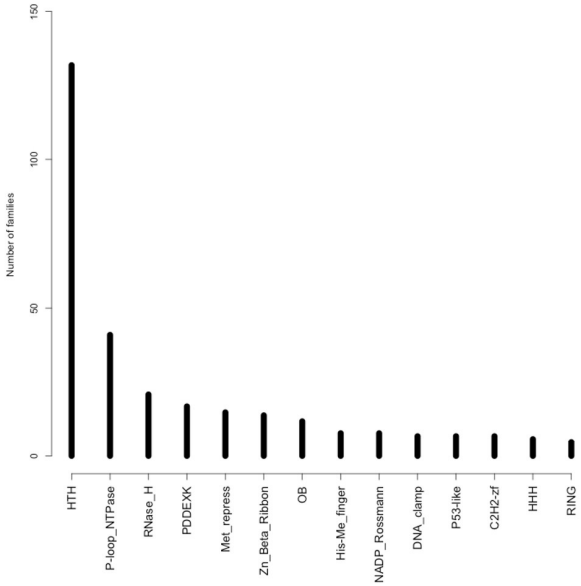


Figure 2: **Distribution of proteins in Pfam clans:** DBP families were mapped to Pfam clans and the three top-most populated clans were helix-turn-helix, P-loop containing nucleoside triphosphate hydrolase and Ribonuclease H-like.

Clan mapping and distribution

Pfam organizes similar protein domain families into clans, based on their HMM profile similarities. We studied DBP families for their clan distribution in Pfam. About 58% of the families do not map to any Pfam clan, exemplifying the diverse nature of DBP families. Only 446 families out of 1057 DBP families fall in 96 Pfam clans (Figure S1). The three top-most populated clans were helix-turn-helix, P-loop containing nucleoside triphosphate hydrolase and Ribonuclease H-like (Figure 2). We also calculated the normalized propensities of occurrence of DNA-binding families in each of the 96 Pfam clans (Figure S1). This highlights that 12 clans (TRD, P53-like, TBP-like, HUH, Homing_endonucl, MBD-like, PRD, LEF-8-like, DnaA_N, FadR_C, DNA_primase_lrg, bZIP) are purely DNA-binding ones i.e. all families in these clans are recorded to possess DNA-binding function.

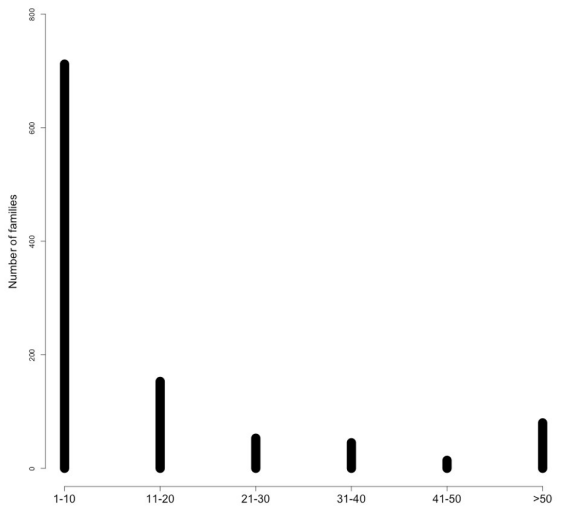


Figure 3: **Co-existing domains in DBP families:** The families were further analysed for their domain architectures. 57% of the families, possess single or less than 3 domains and approximately 83% of them have at-least one co-existing domain.

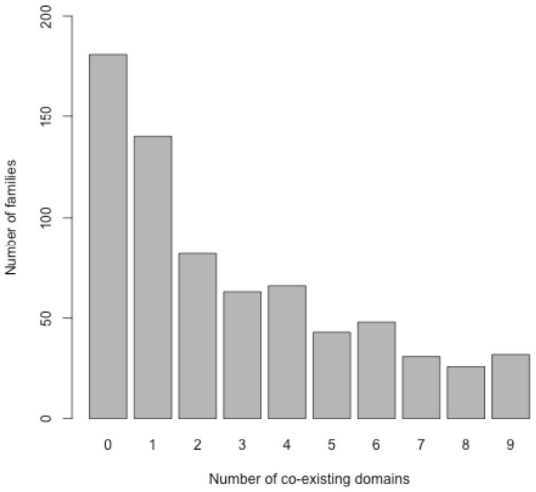


Figure 4: **Co-existing domains in DBP families:** 17% of DBP families were single domain families and possess only the DNA-binding domain.

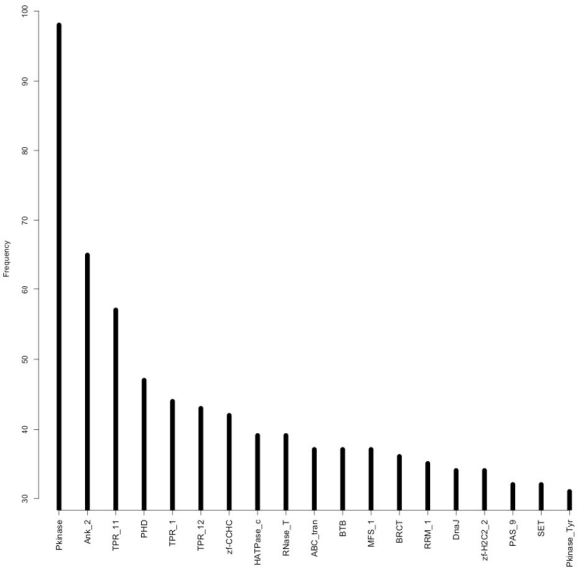


Figure 5: **Most frequent co-existing domain:** The domain architectures of DBP families were studied to identify the most common co-existing domain. Pkinase and Ankynin domain were observed to occur frequently.

Domain architecture (co-existing domains)

We studied DBP families for the domain architectures of the entire gene products that possess these domains. 57% of the families contain single or less than 3 domains. Majority of the families (83%) have an accompanying domain and only 17% families have single DNA-binding domain (Figure 3 and Figure 4). The DBP families with single domains were mapped to GO biological functions and majority of these families either perform regulatory functions like regulation of transcription, viral transcription or are involved in viral genome activities (like viral DNA genome packaging, replication, transcription or assembly).

We then analyzed these co-existing domains and plotted frequency distribution to identify the most frequent co-existing domain. The most frequently occurring DNA-binding domain was Helicase_C (Helicase conserved C-terminal domain), which is present in all helicases and helicase-related proteins like UvrD, DEAD, SNF_2 and topoisomerases. The most frequently observed co-existing domains, which are not DNA-binding in nature, were Pkinase and Ankyrin (Figure 5). Pkinase domain co-exists with DNA-binding domains like DNA ligase and DNA helicase. Some examples of DNA-binding families having Ankyrin domain are DNA ligase, heat shock factor (HSF) and UvrD.

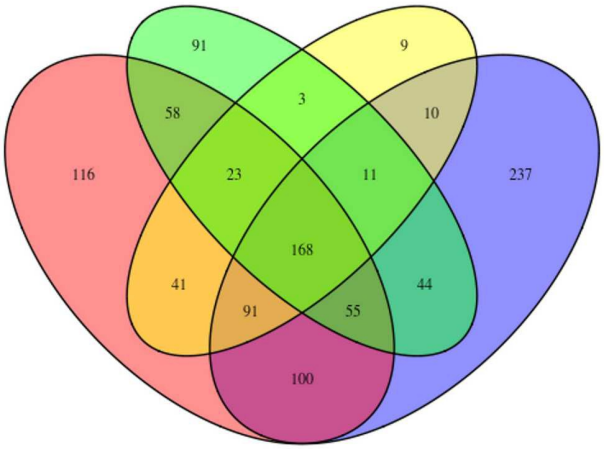


Figure 6: **Taxonomic distribution:** Distribution of DBP families across different domains of life. Blue, red, yellow and green- highlights Eukaryota, Bacteria, Archaea and Viruses respectively.

Taxonomic distribution

The DBP families were studied for their taxonomic distribution. Figure 6 highlights the distribution of DBP families across three domains of life (Bacteria, Archaea and Eukaryota) and viruses. 168 DBP families were distributed in all three domains and viruses. These common families were involved in generalized functions like DNA ligase activity, DNA primase activity, and polymerases.

Some families were observed to be specific in their distribution (*i.e.* bacterial, viral, archaeal or eukaryotic) (Table 1). There were nine families, which were distributed only in archaeal genomes (Table S1). The majority of these families were

involved in conferring stability to chromatin in order to survive in hostile environments. Viral specific families possess proteins like viral polymerases, DNA packaging proteins and viral helicases (Table S2).

We divided the 237 eukaryotic specific families into four further subclasses: metazoan, plant, fungi and other eukaryotes. There were 17 plant-specific families, majority of, which are plant-specific transcription factor families like HD-zip, Nozzle, NAM and leafy (Table 2). The plant-specific families exemplify that these families have specific regulatory functions, which evolved after the divergence of plants and animals. There were nine fungi-specific and 50 metazoan-specific families (Table S3 and S4).

Table 1: Taxonomic distribution of the DNA-binding families across three domains of life and viruses.

	Total families	Specific families			
Bacteria	652	116			
Archaea	356	9			
Eukaryota	716	*Metazoa*	*Plants*	*Fungi*	*Others*
		50	17	9	161
Viruses	453	91			

DNA-binding families were studied for their distributions across three domains of life and viruses. There were kingdom-specific families. Besides, 168 families were predicted to be present in Bacteria, Archaea, Eukaryota and Viruses.

Table 2: List of plant-specific DNA-binding families. Majority of these families are transcription regulators.

Pfam ID	Pfam Name	Function
PF02365	NAM	Development proteins
PF02362	B3	Transcription factor
PF03789	ELK	Domain in transcription factors
PF03004	Transposase_24	DNA transposition
PF08879	WRC	Zinc finger
PF06200	tify	Transcription factor
PF01698	FLO_LFY	Development proteins
PF03790	KNOX1	Transcription repressor
PF03791	KNOX2	Transcription repressor
PF08744	NOZZLE	Transcription factor
PF04640	PLATZ	Transcription factor
PF06640	P_C	Transcription regulator
PF04689	S1FA	Transcription regulator
PF04618	HD-ZIP_N	Transcription factor
PF13724	DNA_binding_2	Non-specific DNA binding
PF02701	zf-Dof	Zinc finger
PF03110	SBP	Transcription factor

Mapping DBP families to structures: SCOP and PDB

Pfam, as explained above, classifies proteins into families based on sequence domains. However, to obtain finer details of their function, we need to understand the overall fold of a given family. Therefore, we mapped the DBP sequence domain families to structures using two databases, namely SCOP and PDB.

We obtained sequences of SCOP members, which are <40% identical and performed sequence-HMM comparisons against a database of HMM profiles of 1057 DBP families. This resulted in mapping of ~50% (532) Pfam families to SCOP entries. We then studied the distribution of these families in SCOP classes, folds and superfamilies.

The most populated DNA-binding SCOP class was all-α which suggests that α-helix is used frequently by the proteins to mediate its interaction with the target DNA (Figure S2). SCOP fold level explains more about the structure adopted by the members of DBP family. The 532 DBP families were observed to belong to 185 SCOP folds. Further, it was noted that 30 SCOP folds have more than three families mapped (Figure 7) and the most populated SCOP fold was 3-helical bundle, which was followed by P-loop NTPases. This is in agreement with the Pfam clan distribution of these families (Figure 3) and it is also documented that majority of solved structures of DBP possess helix-turn-helix motif to bind DNA.[9,10] To understand the biological functions performed, we mapped DBP families to SCOP superfamilies. 232 SCOP superfamilies cover 532 Pfam DBP families (Table S5) and the most populated ones were winged helix DNA-binding domain, Homeodomain-like and P-loop containing nucleoside triphosphate hydrolase.

Following SCOP mapping, we mapped PDB structures to these sequence domain families to identify the families that have known structure. This will also help in identifying the DBP families for which there are no structures solved and hence these families must be taken up for structure determination in the near future. 700 DBP families were observed to have known structural information, whereas 357 families do not have any solved structures. We used a similar approach as followed by SUPFAM[40], to cluster sequence families into PNSF (potential new superfamilies), however, we implemented sequence-HMM comparison using HMMScan, which are reported to be more sensitive[29,41] than sequence-PSSM (Position Specific Scoring Matrix) searches performed by RPS-BLAST in SUPFAM[40]. We employed seed sequences of all 357 DBP families with no known structure information and searched it against a database of HMM models of 357 families within.

To perform clustering of these 357 families into sequence-based superfamilies, we analyzed the families whose seed sequence(s) identify non-self HMM model(s). The families were placed in a cluster by checking the associations. Sequences of 300 DBP families identified only self HMM model and 57 families identified non-self HMM model along with the self HMM model. This resulted in classifying 57 families into 16 putative superfamilies (Table 3). Some of these families within the cluster are known to belong to the same clan in Pfam. We studied the

functions associated with the families, which belong to the same clusters. The functions associated mostly fall into DNA replication, repair and recombination. Many families possessing transcription factors are also clustered (Table 3).

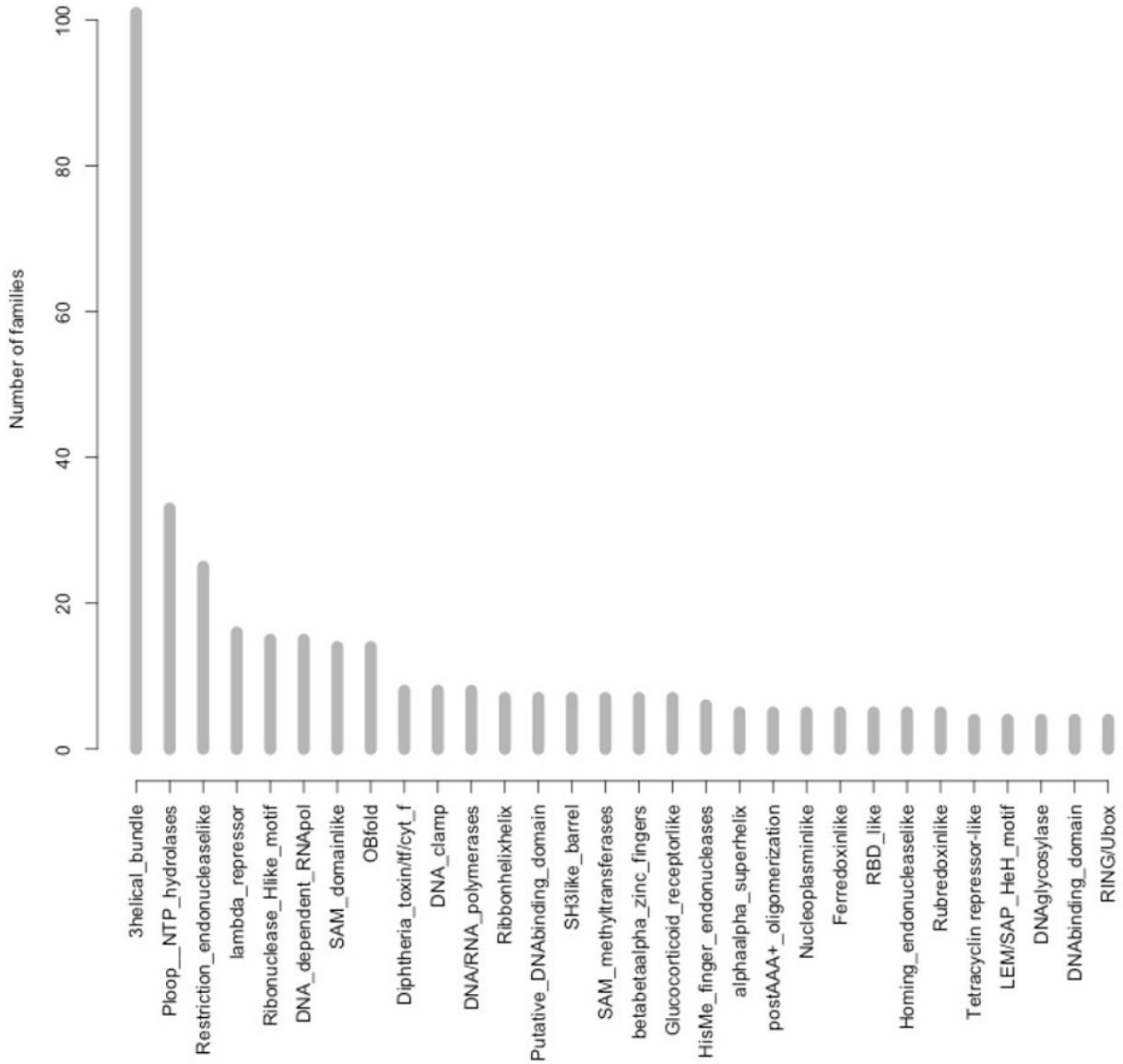


Figure 7: **Structural mapping:** Distribution of DBP families across different SCOP folds. Folds associated with atleast 3 DBP families are shown.

Mode of evolution

The distribution of GO molecular functions across different SCOP DNA-binding folds was studied. We obtained molecular functions for 42 of the SCOP DNA-binding folds with reliable information content of 2.0 and above. Only six of the SCOP folds (SAM-domain like, Bacillus chorismate mutase like, β and β prime subunit of DNA-dependent RNA polymerase, ATP grasp, Resolvase-like and DCoH-like) were mapped to a single molecular function. Majority of the folds (86%) were observed to perform more than one molecular function (Figure 8). Four of these folds are superfolds[42] and are known to occur in many proteins. There have been previous studies where families within these superfolds are known to follow divergent evolution.[43,44] This supports the notion that majority of DNA-binding proteins may follow divergent evolution.

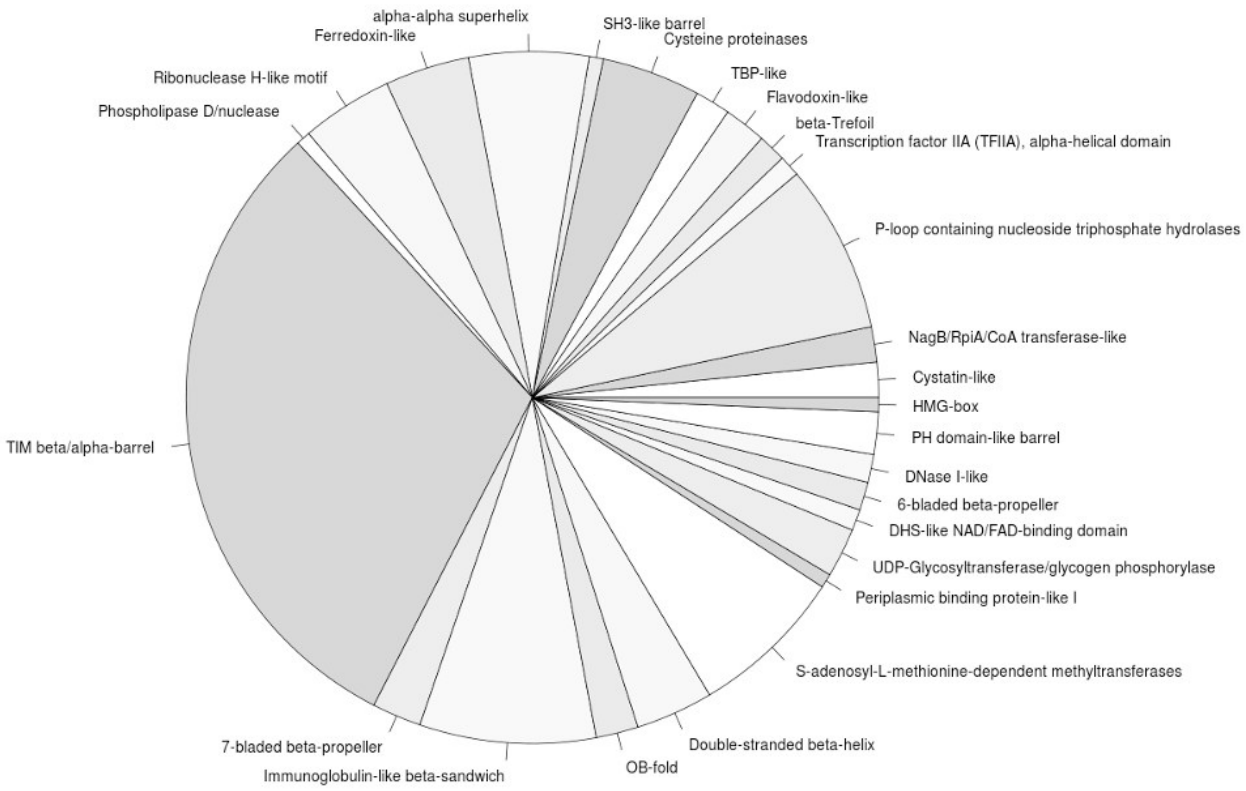


Figure 8: **Functions to folds:** Distribution of GO Molecular function terms associated with different SCOP folds mapped to DNA-binding function.

Conclusions

With the advancement in sequencing technologies[45], there are number of genomes being sequenced. The annotation of the sequenced genome helps in understanding various biological functions performed by a genome of interest. Here, we present collation and computational analyses of an important class of proteins *i.e.* DNA-binding proteins. DNA-binding proteins are known to govern many cellular activities like DNA replication, transcription, DNA repair. The proteins with DNA-binding function can be grouped at both structure and sequence level. There are attempts in the field to group DBP based on the structure of the DNA-binding motif in the protein.[9] However, at the sequence level, classifications are derived only for a subset of DBP, namely the transcription factors. We identified 1057 sequence-based DBP families and studied various family features like number of proteins, their length and distribution of their sequence identities. On an average, these families are highly populated and can be very diverse as the percent identities within a family are very low. This is supported by the fact that DBP are known to form a major portion of protein coding genes in all kingdoms of life (Bacteria, Archaea, Viruses and Eukaryota) and performs diverse functions. We also studied the taxonomic distribution of DBP families in Bacteria, Archaea, Viruses and Eukaryota. Most of the DBP families specific for archaea provide stability to the chromatin, as archaea are known to survive in adverse environmental conditions. The fungi and metazoan specific families are involved in DNA repair, telomere capping, DNA transposition and regulation of transcription. These functions are also observed in other domains of life; however, these families may be involved in specific regulatory pathways unique for these organisms. The plant-specific families were mainly transcription factor families unique to plants. Understanding the functional roles of these kingdom-specific proteins will be interesting and help in elucidating the pathways involved.

DBP-families identified from Pfam, were studied for their distribution in Pfam clans. Due to their diverse nature, more than half of the families (~55%) were not mapped to any of the clan. For the grouped families, the most populated clans were helix-turn-helix, P-loop containing nucleoside triphosphate hydrolase and Ribonuclease H-like. As mentioned, DBP perform variety of cellular function. Therefore, we examined the co-existing domains with the DNA-binding domain. The DBP families with single domain proteins were observed to perform regulatory functions like regulation of transcription and are involved in viral genome activities like viral DNA genome packaging, replication, transcription, and assembly.

A major portion of DBP families (83%), were observed to possess accompanying domain(s). The most frequently observed co-existing domains were Pkinase and Ankyrin domain. The Ankyrin domain is known to mediate many protein-protein interactions. However they are reported to be present in proteins

with diverse functions like transcription factors, toxins and various enzymes.[46]

Table 3: DBP families with no structural information were clustered into 16 clusters.

Cluster	Families in cluster	Size of cluster (Number of families)	Functions within the cluster
1	**AFT, FAR1**	2	Transcription factors with WRKY-like fold
2	Bro-N, **HTH_17**, **HTH_22, HTH_10**, DDE_4_2, **Terminase_5**, Phage_Cox, PyocinActivator	8	Regulators of DNA replication and/or transcription
3	**Cytomega_UL84, Herpes_UL82_83**	2	Viral DNA replication
4	**DDE_4, DDE_Tnp_1_2, DDE_Tnp_1_3, DDE_Tnp_1_6**, Whib	5	Transposases
5	**DDE_Tnp_IS1, DDE_Tnp_IS66, DDE_Tnp_ISL3**, HTH_21, HTH_33	5	DNA transposition
6	DUF3071, Phage_lambda_P	2	Phage DNA replication
7	HTH_Tnp_IS66, **Zn_ribbon_recom**, HTH_Tnp_ISL3, **Ogr_Delta**, **zf-C4_Topoisom**, zf-Dof, zf-Mss51, **zf-GRF**, **A2L_zn_ribbon**, Chordopox_RPO7, **OrfB_Zn_ribbon**, YL1_C, **Zn_ribbon_2**	13	DNA recombination, topoisomerases and associated with transcription factors
8	K-box, HALZ	2	Leucine zippers and coiled coils (Transcription factors)
9	N6-adenineMlase, EcoRI_methylase	2	DNA methylases
10	**SfsA, NERD, RmuC**	3	Transcription factor and DNA recombination
11	Rep_trans, Phage_CRI	2	DNA replication initiation
12	Phage_GPL, KfrA_N	2	DNA packaging and partioning
13	Phage_rep_O, RepL	2	DNA replication
14	**RPA, TrfA**	2	DNA replication, repair and recombination
15	TMF_DNA_bd, CtIP_N, TrbI_Ftype	3	Cell cycle regulation and transcription factor
16	zf-C3Hc3H, WRC	2	Chromatin remodeling

DBP families with no structural information were clustered into 16 clusters. These are potential targets for structural genomics initiative. The families belonging to the same Pfam clan are marked in bold.

The sequence-based DBP families were studied for their structural features. Only 50% of the protein domain families have a representative in SCOP classification whereas 66% of the families have atleast a structure deposited in the PDB. As reported earlier in 2000 and 2012,[9,10] most of the DNA-binding proteins with solved structures employ α-helix to recognize their target DNA. We observed that most populated SCOP class is all-α and three-helical bundle is the most populated fold. Some of the families with no structure information[47] were clustered together into 16 sequence-based potential superfamilies. These clusters can serve as targets for structure genomics initiative. This may help in understanding the fold adopted by these families and hence the underlying mechanism of their function.

As the structures are slow-evolving as compared to protein sequences[48], SCOP folds mapped to DBP families were annotated for their functions using GO database.[28] This was performed to study the nature of evolution of DBP families. The families that have diverged from a common ancestor will possess the features of this ancestor in terms of function, structure and sequence.[43,47] A single SCOP fold was observed to perform multiple biological functions. This is further supported by an earlier study performed by Thornton and coworkers in 2000,[9] where they report limited number of DNA-binding motifs in DBP (such as helix-turn-helix, Zinc-coordinating, β-sheet, zipper type, β-hairpin/ribbon), associated with 54 different functions. We revisited this classification in 2012,[10] and observed the same trend with nine groups (DNA-binding motifs), and a three-fold increase in the number of underlying families (174 families). This implies that majority of these families would have evolved to perform variety of functions but retained common fold (divergent evolution).

The set of sequence-based protein domain DBP families can be used to annotate a sequenced genome for DNA-binding proteins. The entire set of such families is inscribed as mathematical profiles (Hidden Markov Models) and is available for download from http://caps.ncbs.res.in/DBPome. Interesting species-specific preferences were obtained in a genome-wide survey for DBPs in the model plant genome *Arabidopsis thaliana*.[49] Such genome-wide studies will help us in understanding distribution and functions of DBP families in a genome or phyla of interest.

References

D. B. Wetlaufer, *Proc. Natl. Acad. Sci. U. S. A.*, 1973, **70**, 697–701.

J. S. Richardson, *Adv. Protein Chem.*, 1981, **34**, 167–339.

A. G. Murzin, S. E. Brenner, T. Hubbard and C. Chothia, *J. Mol. Biol.*, 1995, **247**, 536–540.

C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells and J. M. Thornton, *Structure*, 1997, **5**, 1093–1109.

F. M. G. Pearl, C. F. Bennett, J. E. Bray, A. P. Harrison, N. Martin, A. Shepherd, I. Sillitoe, J. Thornton and C. A. Orengo, *Nucleic Acids Res.*, 2003, **31**, 452–455.

A. Bateman, E. Birney, R. Durbin, S. R. Eddy, K. L. Howe and E. L. Sonnhammer, *Nucleic Acids Res.*, 2000, **28**, 263–266.

R. D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. L. L. Sonnhammer, S. R. Eddy and A. Bateman, *Nucleic Acids Res.*, 2009, **38**, D211–D222.

S. C. Harrison, *Nature*, 1991, **353**, 715–719.

9 N. M. Luscombe, S. E. Austin, H. M. Berman and J. M. Thornton, *Genome Biol.*, 2000, **1**, REVIEWS001.
10 S. Malhotra and R. Sowdhamini, *BMC Bioinformatics*, 2012, **13**, 165.
11 Y. Kao-Huang, A. Revzin, A. P. Butler, P. O'Conner, D. W. Noble and P. H. Von Hippel, *Proc. Natl. Acad. Sci. U. S. A.*, 1977, **74**, 4228–4232.
12 S. E. Halford and J. F. Marko, *Nucleic Acids Res.*, 2004, **32**, 3040–3052.
13 R. M. R. Coulson and C. A. Ouzounis, *Nucleic Acids Res.*, 2003, **31**, 653–660.
14 V. Charoensawan, D. Wilson and S. A. Teichmann, *Nucleic Acids Res.*, 2010, **38**, 7364–7377.
15 J. Collado-Vides, B. Magasanik and J. D. Gralla, *Microbiol. Rev.*, 1991, **55**, 371–394.
16 E. Pérez-Rueda and J. Collado-Vides, *Nucleic Acids Res.*, 2000, **28**, 1838–1847.
17 E. Pérez-Rueda, J. Collado-Vides and L. Segovia, *Comput. Biol. Chem.*, 2004, **28**, 341–350.
18 J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann and N. M. Luscombe, *Nat. Rev. Genet.*, 2009, **10**, 252–263.
19 J. L. Riechmann, J. Heard, G. Martin, L. Reuber, C. -Z, Jiang, J. Keddie, L. Adam, O. Pineda, O. J. Ratcliffe, R. R. Samaha, R. Creelman, M. Pilgrim, P. Broun, J. Z. Zhang, D. Ghandehari, B. K. Sherman and G.-L. Yu, *Science*, 2000, **290**, 2105–2110.
20 M. Martínez-Bueno, A. J. Molina-Henares, E. Pareja, J. L. Ramos and R. Tobes, *Bioinformatics*, 2004, **20**, 2787–2791.
21 J. Wu, F. Zhao, S. Wang, G. Deng, J. Wang, J. Bai, J. Lu, J. Qu and Q. Bao, *BMC Genomics*, 2007, **8**, 104.
22 D. Wilson, V. Charoensawan, S. K. Kummerfeld and S. A. Teichmann, *Nucleic Acids Res.*, 2008, **36**, D88–D92.
23 N. Sierro, Y. Makita, M. de Hoon and K. Nakai, *Nucleic Acids Res.*, 2008, **36**, D93–D96.
24 Y. Minezaki, K. Homma and K. Nishikawa, *DNA Res.*, 2006, **12**, 269–280.
25 S. Moreno-Campuzano, S. C. Janga and E. Pérez-Rueda, *BMC Genomics*, 2006, **7**, 147.
26 E. Portales-Casamar, S. Thongjuea, A. T. Kwon, D. Arenillas, X. Zhao, E. Valen, D. Yusuf, B. Lenhard, W. W. Wasserman and A. Sandelin, *Nucleic Acids Res.*, 2010, **38**, D105–D110.
27 V. Matys, O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. E. Kel and E. Wingender, *Nucleic Acids Res.*, 2006, **34**, D108–D110.
28 M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock, *Nat. Genet.*, 2000, **25**, 25–29.
29 S. R. Eddy, *PLoS Comput. Biol.*, 2011, **7**, e1002195.
30 S. Hunter, P. Jones, A. Mitchell, R. Apweiler, T. K. Attwood, A. Bateman, T. Bernard, D. Binns, P. Bork, S. Burge, E. de Castro, P. Coggill, M. Corbett, U. Das, L. Daugherty, L. Duquenne, R. D. Finn, M. Fraser, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, C. McMenamin, H. Mi, P. Mutowo-Muellenet, N. Mulder, D. Natale, C. Orengo, S. Pesseat, M. Punta, A. F. Quinn, C. Rivoire, A. Sangrador-Vegas, J. D. Selengut, C. J. A. Sigrist, M. Scheremetjew, J. Tate, M. Thimmajanarthanan, P. D. Thomas, C. H. Wu, C. Yeats and S.-Y. Yong, *Nucleic Acids Res.*, 2012, **40**, D306–D312.
31 M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson and D. G. Higgins, *Bioinformatics*, 2007, **23**, 2947–2948.
32 The UniProt Consortium, *Nucleic Acids Res.*, 2011, **40**, D71–D75.
33 D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell and E. W. Sayers, *Nucleic Acids Res.*, 2009, **37**, D26–31.
34 E. W. Sayers, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, V. Miller, I. Mizrachi, J. Ostell, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, E. Yaschenko and J. Ye, *Nucleic Acids Res.*, 2009, **37**, D5–15.
35 J. Gough and C. Chothia, *Nucleic Acids Res.*, 2002, **30**, 268–272.
36 T. R. Bürglin, *Cell*, 1991, **66**, 11–12.
37 A. Timinskas, V. Butkus and A. Janulaitis, *Gene*, 1995, **157**, 3–11.
38 J.-P. Dadoune, *Microsc. Res. Tech.*, 2003, **61**, 56–75.
39 M. Selmer and X.-D. Su, *EMBO J.*, 2002, **21**, 4145–4153.
40 S. B. Pandit, D. Gosar, S. Abhiman, S. Sujatha, S. S. Dixit, N. S. Mhatre, R. Sowdhamini and N. Srinivasan, *Nucleic Acids Res.*, 2002, **30**, 289–293.
41 S. R. Eddy, *Bioinforma. Oxf. Engl.*, 1998, **14**, 755–763.
42 C. A. Orengo, D. T. Jones and J. M. Thornton, *Nature*, 1994, **372**, 631–634.
43 D. L. Theobald and D. S. Wuttke, *J. Mol. Biol.*, 2005, **354**, 722–737.
44 N. Nagano, C. A. Orengo and J. M. Thornton, *J. Mol. Biol.*, 2002, **321**, 741–765.
45 W. W. Soon, M. Hariharan and M. P. Snyder, *Mol. Syst. Biol.*, 2013, **9**.
46 P. Bork, *Proteins*, 1993, **17**, 363–374.
47 E. Zuckerkandl and L. Pauling, *Evol. Genes Proteins*, 1965, **97**, 97–166.
48 N. V. Grishin, *J. Struct. Biol.*, 2001, **134**, 167–185.
49 S. Malhotra and R. Sowdhamini, *Nucleic Acids Res.*, 2013, **41**, 7212–7219.