Pak. J. Bot., 45(SI): 309-319, January 2013.

IN SILICO PREDICTION OF REGULATORY ELEMENTS AND CORRESPONDING PROTEIN-DNA INTERACTIONS IN PLANT PROMOTERS

MADIHA TANVEER, HINA-UR-RAZAQ QURESHI, M. QAISER FATMI AND TAYYABA YASMIN*

Department of Bioscience, COMSATS Institute of Information Technology (CIIT) Islamabad, Campus, Pakistan.
**Corresponding author e-mail: tayyaba_ yasmin@comsats.edu.pk*

Abstract

The importance of *cis* or *trans* acting regulatory elements in gene regulation is quite obvious. Exploring these elements *in vivo* demands extensive experimentation and is time intensive. *In silico* methods of predicting these elements have been developed in this regard. In present study around 300 promoters belonging to monocots, dicots and algae were analysed through Consite tool for prediction of regulatory elements. Many putative regulatory elements of diverse functions were found in these promoters. In monocots, TATA-binding proteins (TBP), in dicots, hunchback and in Algae, aryl hydrocarbon receptor nuclear translocator (ARNT) were abundantly represented with 55, 33 and 86% respectively. It was observed that all three plant groups exhibited different families of transcription factors like basic helix-loop-helix (bHLH), basic helix-loop-helix leucine zipper (bHLH-ZIP), Forkhead, RUNT, HOMEO-ZIP, zinc finger (ZN-FINGER), REL, Nuclear receptor, MADS, bZIP and TATA-box. Moreover, selected transcription factors were explored through HADDOCK Webserver to predict possible interactions between their corresponding regulatory elements. It was observed that hydrogen bonds were mostly involved in these interactions. In addition, Lysine and Arginine were mainly found to be associated in establishing these interactions with thymine base.

Introduction

Transcription regulation is a vital phenomenon for the accurate execution of biological processes because it dictates the time and quantity of production of particular protein (Nakashima *et al*., 2000, Maston *et al*., 2006; Narang, 2008). The biological significance of transcription regulation is quite evident as alterations in the steps of transcription process can lead to diseases (Maston *et al*., 2006). The modulation of transcription is dependent not only upon *cis* acting regulatory elements which are located in the promoters of the coding genes but also on the *trans* acting regulatory proteins (Maston *et al*., 2006, Narusaka *et al*., 2003). The *cis*-acting transcriptional regulatory DNA elements are short DNA sequences that contain binding sites for *trans*-acting DNA-binding transcription factors (Rombauts *et al.,* 2003; Sandelin *et al*., 2004). *Trans*-acting DNA-binding factors are the factors which function either to enhance or repress transcription and also termed as transcription factors (Maston *et al*., 2006).

As far as transcription regulation in plants is concerned, this process plays a key role in plant development as well as generating response against environmental stresses. The genomic analysis of *Arabidopsis thaliana* and other plants have revealed that probably more than 3,000 genes were involved in transcription. Among these 3,000 genes more than one half were anticipated to code for transcription factors (Rombauts *et al.,* 2003). Moreover it is reported that in *Arabidopsis thaliana* about 15% of the genes on a specific chromosome have a role in transcription (Singh, 1998). Like other organisms, in plants, the determination of tissue specific behaviour and developmental stage activities are majorly dependent upon transcription regulation process (Grasser, 2007). In developmental stages, specifically during seed maturation phase, gene expression plays a crucial role by regulating the expression of seed storage proteins, gaining desiccation tolerance and entry into dormant state (Vicente-Carbajosa & Carbonero, 2005).

With the passage of time, coding genes in many organisms have been identified but the prediction of regulatory elements and the corresponding transcription factors is still a challenging issue (Guhathakurta & Stormo, 2007). The annotation of *cis*-regulatory regions to identify functional regulatory elements is an advance and significant area of bioinformatics (Satija *et al.,* 2008). Many computational approaches have been developed to cope with this dilemma, and these are based on variety of techniques such as phylogenetic foot printing, DNA sequence comparisons and position weight matrices etc. (Lenhard *et al*., 2003; Guhathakurta & Stormo, 2007). Currently, variety of databases are also available that contain data about transcription factors and their DNA binding sites for a number of organisms (Guhathakurta & Stormo, 2007).

The transcription factors interact with DNA mainly through Hydrogen bonds, ionic and Van der Waals interactions (Angarica *et al*., 2008; Luscombe *et al.,* 2001). Due to imperative role of Protein-DNA interactions, understanding of the binding specificity mechanisms in these interactions is equally important. Now a day the computational approaches that are being used for the exploration of protein-DNA interactions are based on statistical models. For instance, Random Forest method, Support vector machine etc (Angarica *et al*., 2008; Xie *et al*., 2009; Si *et al*., 2011). There are number of web servers available e.g., HADDOCK, AutoDock and many others which endow a platform for the identification of transcription factors that interact with a regulatory DNA sequence of interest (e.g. gene promoters) (Barrasa *et al*., 2007).

The present study was aimed to predict the putative regulatory elements and their distribution in 300 plant promoters which belonged to Monocots, Dicots and Algae by utilizing *in silico* methods. Further, the interactions between selected transcription factors and their corresponding regulatory elements were also predicted through docking studies. Many putative regulatory elements of diverse functions belonging to different families were observed in the selected promoters. Moreover, the hydrogen bond was found to be involved in the interactions of transcription factors with the corresponding DNA- elements selected for docking.

Materials and Methods

Data retrieval: Plant Promoter Sequences were retrieved from "PlantProm DB" available at http://mendel.cs.rhul.ac.uk/ and http://www.softberry.com/. This database contains proximal promoter sequences (-200 to +51) for RNA polymerase II from a range of plant species (Shahmuradov & Gammerman, 2002). The sequences represent three divisions of Plant species: 81 Monocots, 210 Dicots and 14 Algae and they were further subdivided as follow:

Plant Group	TATA Box Containing Promoters	TATA- Less Promoters
Monocots	47	34
Dicots	125	85
Algae	3	11

Regulatory elements identification: Consite was used for identification of regulatory elements (RE) and transcription factor binding sites (TFBS) in the promoter sequences. Consite is based on phylogenetic foot printing and is freely available at http:/www.phylofoot.org/consite (Sandelin *et al.*, 2004). Functions of transcription factors were explored through NCBI (http://www.ncbi.nlm.nih.gov /nucleotide), TRANSFAC (http://www.cbil.upenn.edu/cgi-bin/tess/tess) and UniProtKB (http://www.uniprot.org). Biological pathways of genes downstream of the promoters under study were also searched and investigated through UniProtKB.

Distribution of regulatory elements /transcription factors: For the distribution of regulatory elements/TFs according to genes following percentage formula was used:

$$\text{Percentage (\%)} = \text{Gene Counts/ Total number of genes} * 100$$

Gene count is showing the number of promoter sequences in which a particular transcription factor is present.

The distribution of each transcription factor was also calculated to know that how many transcription factors are present in a particular organism and in what proportion. Following formula was used:

$$\text{Percentage (\%)} = \frac{\text{No. of Transcription factors in an organism}}{\text{Total no. of Transcription factors}} \times 100$$

Modelling of regulatory elements and transcription factors: For modelling regulatory elements, online available software named "DNA Sequence to Structure" (url: http://www.scfbio-iitd.res.in/software/drugdesign/bdna.jsp) was utilized that provides DNA structure in PDB format. Similarly, structural data of TFs were collected through PDB (Protein Data Bank) accessible at http://www.rcsb.org/pdb/ (Berman *et al.*, 2000). After the filtration of data best PDB's were selected on the basis of amino acids length, missing residues, resolution, mutation and structure type. The best PDB's were explored for further analysis of protein-DNA docking.

Protein-DNA docking: The web-server of HADDOCK was used in present study for protein-DNA docking and is freely available at (http://haddock.chem.uu.nl/). It also serves as platform for protein-protein and protein-ligand docking. The web interface requires active residues (interacting residue) and structures of both bio-molecules (De Vries *et al*., 2010). Active residues of DNA were provided along with the structure files produced by Sequence to Structure tool. Active residues of TFs were identified with the help of another tool named "BindN".

HADDOCK Webserver evaluates models on the basis of HADDOCK score. The score is a combination of buried surface area, Van der Waals, Electrostatic, De-solvation and restraint violation energies. Cluster size demonstrates the number of best structures, and the structure with the lowest energy is favoured because lowest energy structures are considered to be good models (De Vries *et al*., 2010).

Visualization of molecular dynamics: In order to visualize the results generated by HADDOCK Webserver, a molecular visualization program VMD (http://www.ks.uiuc.edu/Development/) was used. VMD is 3D molecular graphics software that facilitates to visualize interactions between bio-molecules.

Results and Discussions

Transcription factors binding sites/regulatory elements (RE) in the promoter sequences ranging from 6-20 nucleotides were identified by Consite. It was observed that all three plant groups exhibited different families of transcription factors like helix-loop-helix (bHLH), bHLH-ZIP, Forkhead, RUNT, HOMEO-ZIP, ZN-FINGER, REL, Nuclear receptor, MADS, bZIP and TATA-box. Members of protein families such as bHLH, bHLH-ZIP, FORKHEAD and MADS strikingly exhibited sequential order of their appearance in the DNA sequence. Following types of transcription factors were identified during present study.

Forkhead: Forkhead proteins are a large family of functionally diverse TFs that have been implicated in a various cellular processes particularly regulation of development and immune response in animals. In plants, their functions are still to be uncovered. The forkhead domain has winged helix structure which is relatively conserved between proteins that belong to forkhead family (Coffer & Burgering, 2004). In the present study, binding sites of different TFs of forkead family like HFH-2, HFH-3 and HNF-3beta appeared one after the other in corresponding promoters in non-contiguous manner. In addition, their binding to the regulatory elements is context dependent and no consensus sequence was observed. This pattern was mainly observed in dicotyledons, less evident in monocots and was not found in algae. The genes representing this pattern are shown in Table 1. These genes were found to be involved in various metabolic processes. It can be assumed from the observations that, as these transcription factors belong to same protein family so they may have the same domains that are involved in the recognition of specific DNA motif

(Kidokoro *et al*., 2009). Another assumption could be that these factors function cooperatively. Although the functions of these factors in plants are not known yet but their occurrence in plants specifically in dicots is indicating their possible role in regulation of downstream genes. Moreover, this distribution seems to have some evolutionary relationship which could be explored further.

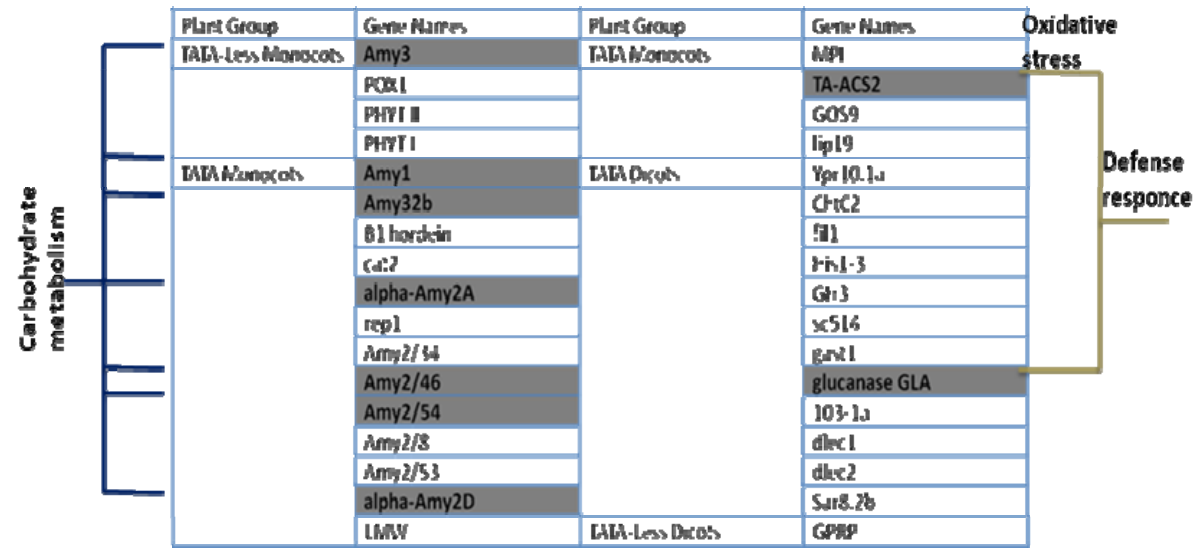
MADS-TATA-box: Transcription factors belonging to MADS-box family allows binding of DNA via N-terminal of the MADS-domain which is highly conserved among plants, fungi and animals (Verelst *et al*., 2007). MADS-box transcription factors are known to be major regulators of various plant developmental processes and also have a significant role in understanding the evolution of several gene families in higher plants (Parenicova *et al*., 2003). During present study, it has been observed that TBP and MEF2 which are members of TATA-box and MADS-box families respectively were present in a consecutive manner. Likewise, another important aspect of TBP-MEF2 pattern was observed that DNA recognition site of MEF2 lied within the recognition sequence of TBP i.e. MEF2 has 10 nucleotides long recognition site and TBP has 15 nucleotides long. The rationale of this feature is not exactly known in plants.

Table 1. Genes showing different pattern of HFH-2, HFH-3 and HNF-3 beta in various plant groups.

Plan group	Gene names	Plant group	Gene names
Tata-less Monocots	Zm-ER abpl	Tata dicots	Deficiens
	Alt*		Dhfr-Ts
Tat-Less dicots	gdcT		NTm19
	PsNOD6		Adh2
	Acyl-CoA binding protein 2		Cellulase*
	CER3		Glucanase GLA*
	CARSR12*		Pma3*
	Sedoheptulose-1,7 bisphosphatase, SBPase*		Ibc*
	33RNP		Nodulin-23
Tata Dicots	TMK1		ACP A1
	CRSD4H		CCA1
	NAT2		Hpr-A

For example, recent studies revealed that 107 genes in *Arabidopsis* genome are destined to encode for MADS proteins in which 84% of these proteins have unidentified functions, hence providing an area for research (Parenicova *et al*., 2003). Although, we know that MEF2 and TBP belong to different protein families, but MEF2 has unique feature that it preferentially binds to common consensus CTA $(A/T)_4$TAG (Verelst *et al*., 2007). Similar findings were observed during present study. Moreover, the consensus sequence for MEF2 was more or less similar to recognition site of TBP. The genes from different plant groups which expressed MEF2-TBP pattern were mostly involved in defence response and carbohydrate metabolism and are shown in Table 2.

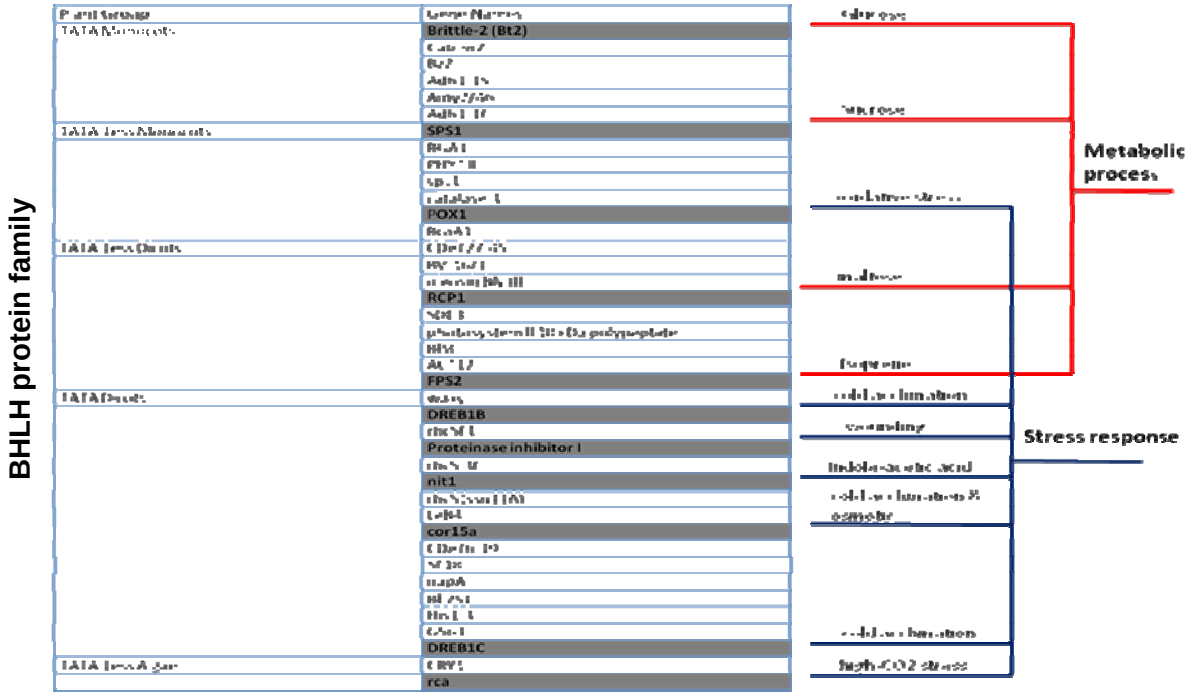
Table 2. Genes having contiguous appearance of MEF2 and TBP. This table is showing the abundant occurrence of these factors in TATA groups (TATA monocots and dicots).



bHLH: Basic helix-loop-helix family is a super family of transcription factors that bind to DNA in the form of dimers. bHLH domain also undergone modifications due to evolutionary multiple domain shuffling, gene duplications and gene deletion events. bHLHZ is one of the example of such modifications, that contain leucine zipper region which is nearby to carboxyl end of the bHLH domain and endows stability in dimerization (McFerrin & Atchley, 2011). Basic helix-loop-helix family domain is composed of two regions i.e., Basic (N-terminal) and HLH (C-terminal), however, HLH region is specifically concerned with formation of dimers. Some members of this family can form both homo and hetero dimers but in the case of heterodimers they particularly choose closely related members of family as dimerization partners (Toledo-Ortiz *et al*., 2003).

A new avenue has been explored through present study which indicates that transcription factors like ARNT, USF, n-MYC followed sequential order i.e., they were present in an order USF-ARNT-ARNT-USF-n-MYC-n-MYC. Even though in some cases it was observed that max and MYC-Max were present at the beginning of this pattern MYCMax-Max-USF-ARNT -ARNT-USF-n-MYC-n-MYC. However, MYCMax was not always present at start of this arrangement in all the promoters. It was also noticed that all above mentioned TFs recognized a specific DNA motif i.e. 5'-CACGTG-3. One of the expected reasons behind this could be that these transcription factors belong to same protein family i.e. basic helix-loop-helix (bHLH). This sequential pattern (MYCMax-Max-USF-ARNT -ARNT-USF-n-MYC-n-MYC) was present in many promoters of monocotyledonous plants. While, in Dicots and Algae few promoters exhibited this order. The genes from different plant groups which expressed observed sequential order of TFs were found to be involved in various metabolic processes and stress responses. These genes are shown in Table 3.

Table 3. Different genes which exhibit USF-ARNT---n-MYC arrangement of TFs are shown along with their plant groups i.e., monocot, algae and dicot.



Distribution of transcription factors: From all three plant groups 20 transcription factors were found to be abundant in present study (Fig. 1A, B, C). In dicotyledonous plants, hunchback was found to be most abundant with 33% of genes followed by TBP with 29%. The lowest percentage of transcription factors (p65, Chop-cEBP, Tal1beta-E47S, Myf, Pax6, RORalfa-2, bZIP911) in dicots was 1% (Fig. 1A). Generally, in plants, hunchback is involved in floral organ identity. In contrast to dicots, TBP was found to be highly abundant in monocots with 54%, despite the fact that dataset for monocot was small as compared to dicots. TBP plays significant role in determining the transcriptional level and selectivity of gene expression in plants. On the other hand, in algae ARNT showed highest percentage i.e. 86%. This proportion could not depict the actual situation in algae until the whole genome is not sequenced and publically available online. Generally, ARNT is involved in activation of the transcription of xenobiotic responsive elements and also have a role in metabolism and degradation of xenobiotics (Okay *et al.*, 2000). Table 4 is illustrating the variation in expression of TFs found within the same group and across other plant groups. The upstream stimulatory factor (USF) showed high percentage (43%) in algae whereas quite low (17%) in monocots, and 14% in dicots.

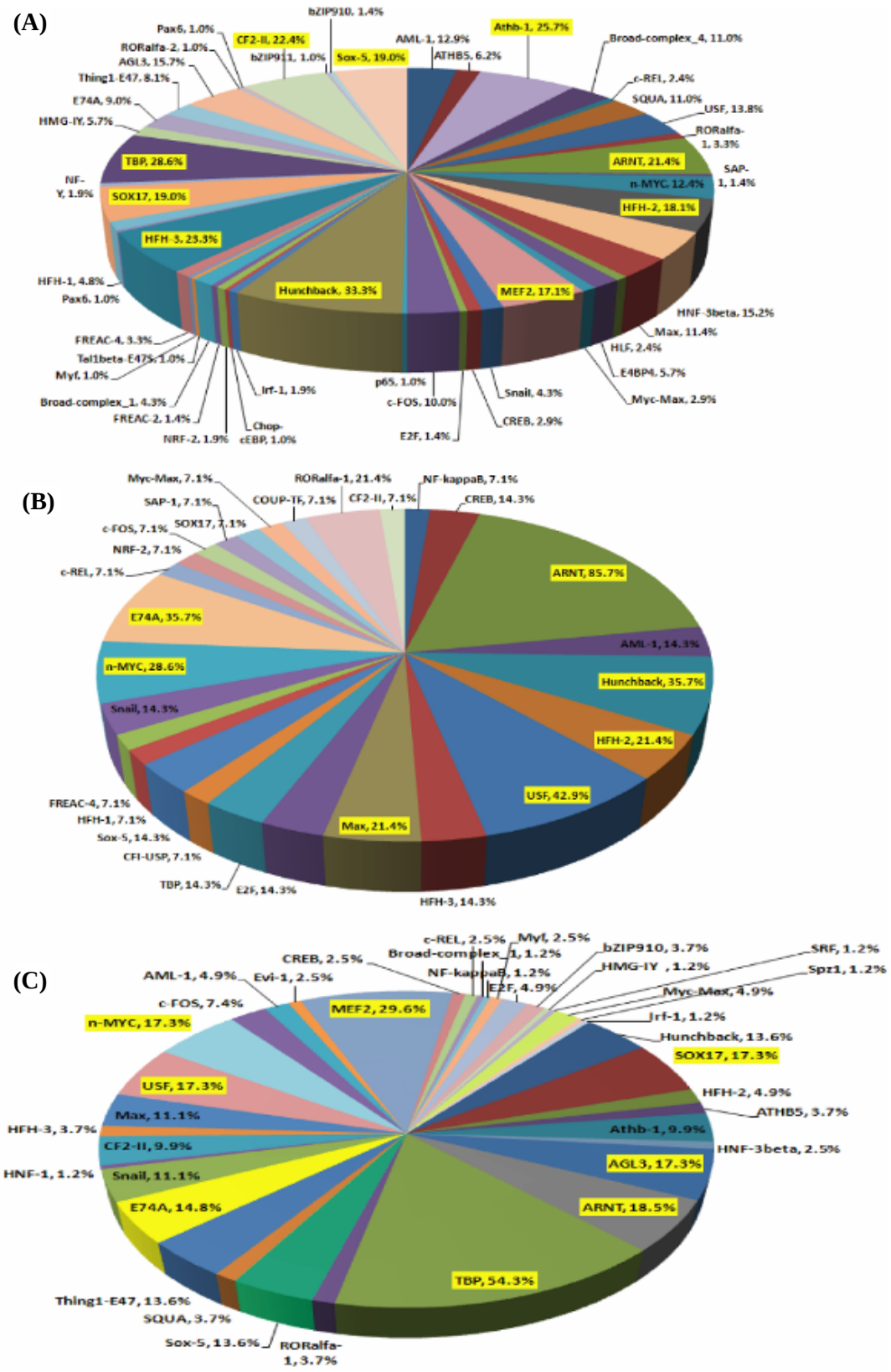


Fig. 1.(A) Percentages of TFs in dicots genes. Highlighted boxes show the highest expression (>17). (B) TFs in Algae. ARNT with maximum percentage of 86 and least expressed elements are with 7%. (C) TBP hit the 54% in Monocots which shows more abundance than other transcription factors.

Table 4. Diversity of TFs in different plant groups. MEF2, AGL3 and Athb-1 were not reported in algae.

TFs	Monocots	Dicots	Algae
MEF2	29.6%	17.1%	-
n-MYC	17.3%	12.4%	28.6%
USF	17.3%	13.8%	42.9%
TBP	54.3%	28.6%	14.3%
ARNT	18.5%	21.4%	85.7%
AGL3	17.3%	15.7%	-
SOX17	17.3%	19.0%	7.1%
Athb-1	9.9%	25.7%	-
CF2-II	9.9%	22.4%	7.1%
Sox-5	13.6%	19.0%	14.3%
HFH-3	3.7%	23.3%	14.3%
Hunchback	13.6%	33.3%	35.7%
HFH-2	4.9%	18.1%	21.4%
Max	11.1%	11.4%	21.4%
E74A	1.2%	9.0%	35.7%

When distribution of TFs within the group (monocots) was investigated, more or less similar results were observed as described previously. High percentage of TBP was found in *Triticum sativum*, *Oryza sativa*, *Zea mays*, *Hordeum vulgare*. The reason of high percentages could be that most of the monocot genes in our data sets were containing the TATA-box rich regions. However less percentage (14%) of TBP in *Hordeum vulgare* is due to the less representation of its genes in our data set.

Unlike monocots, in algae particularly *Chlamydomonas reinhardtii* and *Chlorococcum littorale* have elevated percentage of ARNT i.e., 19.In dicotyledons, *Petunia integrifolia, Betula pendula Roth, Solanum tuberosum, Nicotiana tabacum*, showed high percentages of CF2-II. On the other hand, hunchback showed high percentage in *Arabidopsis thaliana, Solanum tuberosum and Pisum sativum*. *Arabidopsis thaliana* also showed higher percentages of other TFs like ARNT, Athb-1, USF, HFH-2 and HFH-3. TBP was found to be less abundant in dicots but *Nicotiana tabacum* exhibited 20% which is the highest percentage of TBP within any dicot plant. The abundant TFs in different plant groups along with their percentages are shown in Table 5.

Table 5. Abundant TFs in different plant groups.

TFs	Organisms	Percentage	Plant group
TBP	*Triticum sativum*	31%	Monocots
	Oryza sativa	31%	
	Zea mays	26%	
	Hordeumvulgare	14%	
CF2-II	*Petunia integrifolia*	20%	Dicots
	Betulapendula Roth	14%	
	Solanumtuberosum	37%	
	Nicotianatabacum	22%	
Hunchback	*Solanumtuberosum*	31%	
	Pisumsativum	33%	
	Arabidopsis thaliana	104%	
ARNT		55%	
Athb-1		47%	
USF	*Arabidopsis thaliana*	45%	
HFH-2		39%	
HFH-3		37%	
ARNT	*Chlamydomonasreinhardtii*	19%	Algae
	Chlorococcumlittorale	19%	

Interaction of transcription factors with regulatory elements (Docking): Due to the imperative role of protein-DNA interactions, it is necessary to identify these interactions as they play a fundamental role in understanding the molecular mechanisms of gene regulation (Si *et al*., 2011; Wang *et al*., 2009). In order to study protein-DNA interactions in present study, criteria were set for the filtration of protein structures gathered from PDB. The criteria were dependent on the number of mutations, protein length, missing residues in the structure and also upon resolution of PDB structure. Consequently, only 6 out of 20 TFs which were found abundant in three plant groups were selected to dock. Protein-DNA complexes were modelled through HADDOCK Webserver and the study was preferentially focused on hydrogen bonding between protein and DNA residues. As prior studies indicated that two-third of the interactions between residues of protein and DNA are hydrogen bonds (Angarica *et al*., 2008). The details of results generated by HADDOCK Webserver are shown in the Table 6.

Table 7. Possible hydrogen bonds atomic interactions in HNF-3Beta with DNA sequence AACTATTTGCTC.

Protein residue	Protein atoms	Aminoacid position	DNA residue	DNA atoms	Nucleotide position	Distance in A^o
HIS	HE2	169	A	H61	18	2.33
LYS	HZ3	119	T	H3’	6	3.18
ARG	HH11	158	A	O2P	12	2.02
TYR	HH	124	T	O2P	7	2.00
ARG	HE	162	T	O1P	8	3.14
ARG	HE	162	T	O2P	8	2.14
ARG	HH22	210	A	H2	2	2.03
ARG	HH12	168	A	H2’	14	3.23
ASN	HD22	165	G	N7	9	3.03
LYS	HZ2	216	T	O1P	4	3.24
TYR	HN	124	T	O2P	6	1.94
SER	HG	123	A	O1P	5	2.43
ARG	HH21	211	A	O2P	5	2.13
SER	HG	166	T	H71	7	2.63

Table 6. HADDOCK Webserver Results. Rows are representing the information about models of Protein-DNA complexes and their corresponding energies. For each TF, binding affinities with three DNA binding sites are shown. HADDOCK score is combination of buried surface area, Van der Waals, Electrostatic, De-solvation and restraint violation energies.

Protein/DNA	HADDOCK score	Cluster size	RMSD (A^o)	Van der Waals energy (Kcal/ Mol)	Electrosta-tic energy (Kcal/ Mol)	Desolvation energy (Kcal/ Mol)	Restrains violation energy (Kcal/ Mol)	Buried surface area
AML-1-**CTTGCGGTT**	-96.3	18	2.0	-67.0	-642.2	33.0	661.0	1721.3
AML-1-**TTTGTGGTT**	-137.9	17	1.5	-68.7	-696.5	32.4	376.7	1763.3
AML-1-**CCTGTGGTC**	-59.8	8	9.3	-71.2	-630.8	34.5	1029.8	1792.5
c-Fos- **ATGATTCA**	-154.1	17	1.5	-52.4	-666.3	28.4	30.8	1506.0
c-Fos- **CTGACTCA**	-130.1	25	2.1	-57.8	-724.8	35.9	367.8	1668.2
c-Fos- **GTGATTAA**	-143.3	21	7.5	-55.9	-622.2	32.4	46.1	1567.5
Hnf-3beta- **CAATTTTTATTT**	-196.4	73	3.9	-98.4	-864.6	17.5	573.9	2411.8
Hnf-3beta- **AAATATTTATTT**	-199.3	47	1.0	-104.3	-892.4	20.9	626.0	2408.0
Hnf-3beta- **AACTATTTGCTC**	-233.3	65	2.4	-101.2	-860.8	17.3	227.6	2466.2
MEF2- **CTAATTATAG**	-126.7	18	2.2	-78.1	-688.3	32.3	568.5	2056.5
MEF2- **TTATATATAG**	-175.5	8	3.7	-76.4	-793.0	30.6	288.7	2061.8
Sox17- **AACCACAAA**	-58.2	7	1.1	-76.4	-664.1	33.8	1171.4	2247.2
Sox17-**TTCATTGTC**	-98.3	8	3.6	-87.2	-653.7	32.6	870.1	2293.2
Sox17-**CACAATGCT**	-133.5	4	1.8	-91.8	-623.7	25.2	578.0	2393.8
HFH-2- **GTATGTTTGTAT**	-142.7	38	1.4	-80.8	-461.6	0.9	295.1	1896.0
HFH-2- **TCTTGTTTGTTT**	-130.7	7	2.0	-72.0	-442.9	1.0	288.7	1789.2
HFH-2-**ATATATTTTTTT**	-131.5	8	1.1	-61.5	-467.3	-5.7	291.3	1849.4

Evaluation of docking results: The transcription factor HNF-3BETA when binds to the regulatory element 5'-AACTATTTGCTC-3' exhibited three actual hydrogen bonds (SER91: HN… 1.93A^{o} ... G15: OP1, LYS119: HZ1 … 1.89A^{o} … T6: OP1 and SER172: HG … 1.97A^{o} … C16: OP2) (Fig. 2). Structural information of another protein-DNA complex was analyzed to check the number of actual and possible hydrogen bonds in C-Fos-5'-ATGATTCA-3'. Only one hydrogen bond was exhibited in this complex between ARG157: HH21 and A11: OP1 with the distance 1.91A^{o} (Fig. 3). The same procedures were repeated with rest of the protein-DNA complexes i.e. SOX17 complexes with 5'-CACAATGCT-3' form only one actual hydrogen bond (LYS114: HZ1 …1.91A^{o}…A2:OP1), MEF2-5'-TTATATATAG-3' exhibited only one hydrogen bond (ARG15: HH11…1.90A^{o}…T12: OP2). The other protein-DNA complexes i.e. HFH2-5'-TCTTGTTTGTTT-3'and AML1 -5'-CTTGCGGTT-3' did not show any hydrogen bond. The structural representations of all complexes are shown in (Figs. 4, 5, 6 & 7) and details of possible hydrogen bonds are provided in their corresponding tables (Tables 7, 8, 9, 10, 11 & 12). It was observed during the study that Lysine and Arginine had high probability of interactions with thymine and adenine. In all the docked protein-DNA complexes the lowest distance range between the residues of protein and DNA was 1.60A^{o}- 1.99A^{o}. However, due to geometrical constraints they cannot be visualized in rigid docking. In contrast to rigid docking such short distances between protein and DNA residues surely establish bonds in nature. Those protein-DNA residues which have distance less than 3.5A^{o} are considered to be feasible for hydrogen bond formation (Si *et al*., 2011). The term "possible hydrogen bonds" refers to those bonds that are in much closer proximity (<3.5A^{o}) but unable to establish H-bonds because of rigid docking and geometrical constraints that cause hindrance in binding of protein-DNA residues. However, in order to study the formation of hydrogen bonds between protein-DNA, a molecular dynamics simulation should be performed. Hydrogen bond formation studies are significantly important because these hydrogen bonds confer specificity and stability in the protein-DNA complexes (Coulocheri *et al*., 2007). It can be concluded from present study that the distribution pattern of transcription factors is quite diverse in the promoters of monocots, dicots and algae. The selected docked TFs and their corresponding regulatory DNA regions exhibited many potential hydrogen bonds. In addition, Lysine and Arginine were found to be associated in establishing these interactions with thymine.

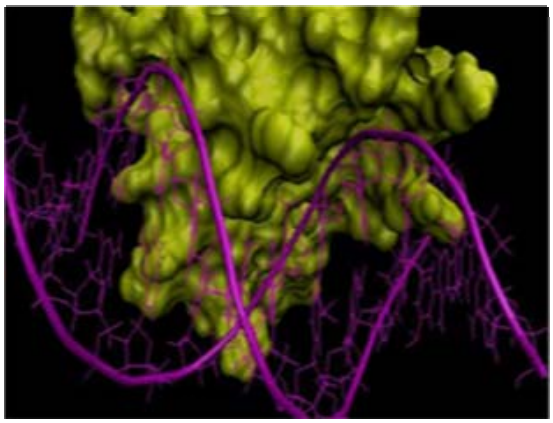


Fig. 2. Interaction of HNF-3Beta regulatory protein with DNA sequence AACTATTTGCTC.

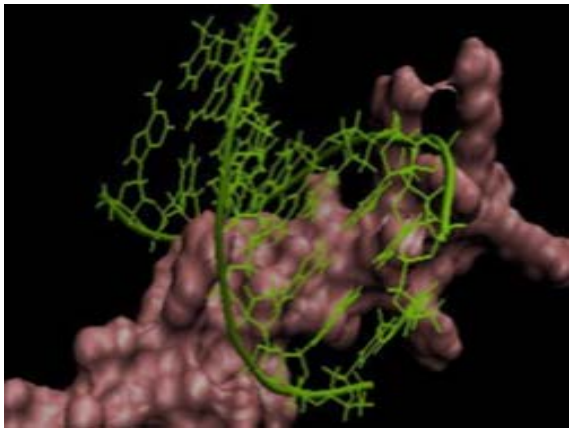


Fig. 3. Complex of C-Fos protein with DNA sequence ATGATTCA.

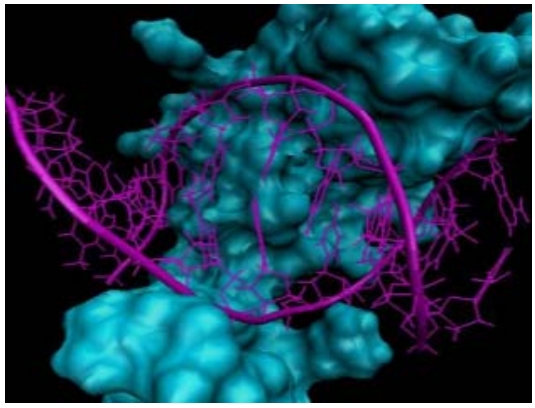


Fig. 4. Complex of SOX17 with DNA of sequence CACAATGCT.

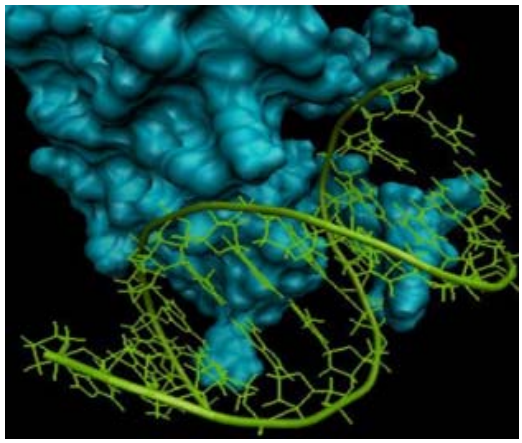


Fig. 5. Complex of MEF2 with regulatory element having sequence TTATATATAG.

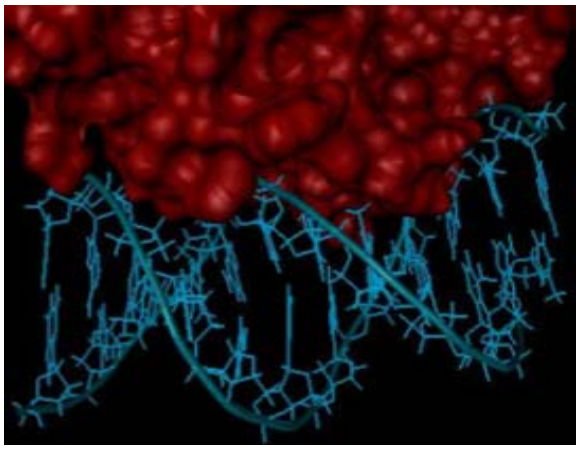


Fig. 6. HFH-2 protein complex with TCTTGTTTGTTT DNA sequence.

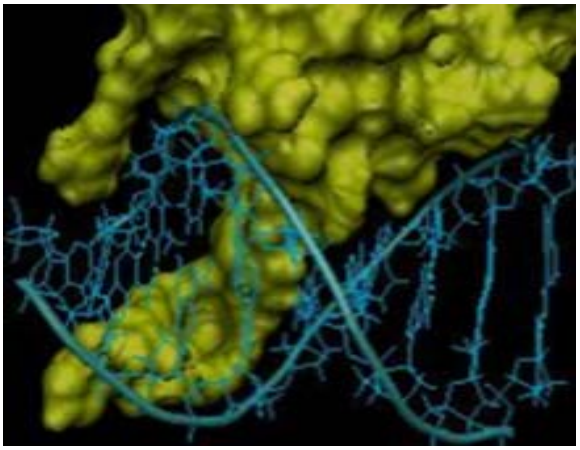


Fig. 7. Complex of AML-1 with DNA having sequence CTTGCGGTT.

Table 8. Atomic interactions in possible hydrogen bonding of C-Fos with ATGATTCA DNA sequence.

Protein residue	Protein atoms	Amino acid position	DNA residue	DNA Atoms	Nucleotide position	Distance in A^o
ARG	HH12	158	T	C7	13	2.82
ARG	HH21	158	T	H71	2	2.39
ARG	HH11	158	A	O2P	12	2.02
ARG	CG	142	T	H3’	2	2.26
ARG	HH22	135	T	O1P	2	2.53
ARG	HH21	155	G	O2P	13	2.26
LYS	HZ2	153	G	O1P	10	2.11
ARG	HE	157	G	H3’	10	2.42
ARG	HH22	159	T	H3’	2	1.85
ARG	HE	159	T	O1P	2	2.13
ALA	CB	150	T	H2’’	9	3.34
CYS	CB	154	G	H2’	10	3.03
ARG	HH22	155	G	H2’	3	2.48
ASN	HD22	147	C	H41	7	2.08

Table 9. Atomic interactions in possible hydrogen bonding of Sox17 with CACAATGCT DNA sequence.

Protein residue	Protein atoms	Amino acid position	DNA residue	DNA atoms	Nucleotide position	Distance in A^o
LYS	HZ2	80	A	H3’	13	1.79
ARG	HH21	69	A	H4’	13	2.04
ASN	HD22	95	A	H61	13	3.16
ARG	NE	69	T	H5**	14	2.11
ALA	CB	74	C	H5**	1	3.18
LYS	HZ2	100	A	OP1	5	2.11
ARG	HH12	69	C	H4’	8	2.22
ARG	HN	70	T	O1P	15	2.29
TYR	HH	137	G	O2P	16	2.10
ARG	HE	69	T	H5**	14	1.35
LYS	HZ2	84	C	H3*	12	2.51
MET	CE	72	T	O2P	14	3.33
LYS	NZ	80	A	H3*	14	2.34
LYS	HZ2	84	C	H3’	12	2.51
ASN	HD22	73	C	H42	3	1.98
ARG	HH22	70	G	O2P	16	1.99

Table 10. Possible hydrogen bonds atomic interactions in MEF2 with DNA sequence TTATATATAG.

Protein residue	Protein atoms	Amino acid position	DNA residue	DNA atoms	Nucleotide position	Distance in A^o
ARG	HH21	24	A	O1P	3	1.94
THR	HG1	20	T	O1P	4	2.05
ARG	HE	24	A	H3’	3	2.89
LYS	HZ1	23	T	O2P	4	2.24
LYS	HZ3	23	T	H2’	4	1.82
LYS	HZ3	31	T	O1P	2	2.73
LYS	HZ2	30	T	O1P	16	2.17
THR	HG1	22	T	O2P	14	2.05
LYS	HZ1	30	T	O1P	16	1.97
LYS	HZ2	53	A	O1P	13	1.97
ARG	HH22	15	C	O1P	11	2.04
LYS	HZ2	4	A	O1P	5	1.99
ARG	HH21	3	T	H5’	18	3.13
LYS	HZ3	5	A	H1’	20	2.49
GLY	HT1	2	A	H2	19	2.03
LYS	HZ2	25	T	H3’	14	1.60
THR	CG2	22	T	H72	14	3.0

Table 11. Possible hydrogen bonds atomic interactions in HFH-2 with binding site TCTTGTTTGTTT.

Protein residue	Protein atoms	Aminoacid position	DNA residue	DNA atoms	Nucleotide position	Distance in A^o
LYS	HZ2	3	C	O2P	24	2.00
LYS	HZ3	3	A	H3’	23	2.24
SER	HG	56	A	H3’	15	2.37
ASN	HD22	49	T	H73	8	2.68
LYS	HZ1	63	T	H3’	14	2.23
SER	HG	50	T	H3’	7	3.06
ILE	HN	9	G	O1P	5	2.28
TYR	HH	40	T	H5**	6	2.52
LYS	HZ3	43	T	H3’	22	2.39
TYR	O	6	G	H4’	5	2.47
ARG	HH11	52	T	H73	14	3.23
HIS	O	53	C	H41	16	3.02
TYR	HN	8	G	H3’	5	1.96
SER	HG	56	C	O2P	16	1.99

Table 12. Representing the picture of potentially forming hydrogen bonds in complex of AML-1 with DNA having sequence CTTGCGGTT.

Protein residue	Protein atoms	Aminoacid position	DNA residue	DNA atoms	Nucleotide position	Distance in A^o
ARG	HH22	118	A	O1P	11	2.46
ARG	NH2	118	A	P	11	3.38
THR	OG1	84	C	P	12	3.28
ARG	CG	142	C	C5’	3	3.47
ARG	HH22	135	A	C5	11	3.43
ARG	NH2	80	C	N4	13	3.00
ARG	NH1	139	G	C3’	4	3.23
LYS	NZ	83	C	O2P	13	2.83
THR	HN	84	C	O2P	12	2.07
ARG	HH12	135	A	H3’	11	2.60
ARG	HH12	118	A	O2P	11	1.95
ARG	HH22	174	G	O2P	14	1.98
LYS	NZ	83	C	H3’	12	2.94
VAL	CG1	170	C	H5	5	2.45
VAL	O	170	C	H41	5	1.93
ARG	HH22	139	C	O2P	5	1.97

References

Angarica, V.E., A.G. Pérez, A.T. Vasconcelos, J. Collado-Vides and B. Contreras-Moreira. 2008. Prediction of TF target sites based on atomistic models of protein-DNA complexes. *BMC Bioinformatics*, 1: 436.

Barrasa, M.I., P. Vaglio, F. Cavasino, L. Jacotot and A. JmWalhout. 2007. EDGEdb: a transcription factor-DNA Interaction database for the analysis of C. elegans differential gene expression. *BMC Genomics*, 1: 21.

Berman, H.M., J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov and P.E. Bourne. 2000. The Protein Data Bank. *Nucleic Acids Res.*, 1: 235-242.

Coffer, P.J. and B.M.T. Burgering. 2004. Forkhead-box transcription factors and their role in the immune system. *Nat. Rev. Immunol*., 4(11): 889-899.

Coulocheri, S.A., D.G. Pigis, K.A. Papavassiliou and A.G. Papavassiliou. 2007. Hydrogen bonds in protein-DNA complexes: where geometry meets plasticity. *Biochimie*, 89(11): pp. 1291-1303.

De Vries, S.J., M.V. Dijk and A.M.J. JBonvin. 2010. The HADDOCK web server for data-driven biomolecular docking. *Nature Protocols*, 5: 883-897.Grasser, K.D. (Ed.). Regulation of Transcription in Plants: Volume 29. Blackwell Publishing.

Guhathakurta, D. and G.D. Stormo. 2007. Finding regulatory elements in DNA sequence. In: *Bioinformatics: Methods Express*. (Ed.): P.H. Dear. Scion Publishing Ltd. Bloxham, UK, pp. 117-140.

Kidokoro, S., K. Nakashima, Z.K. Shinwari, K. Shinozaki and K. Yamaguchi-Shinozaki. 2009. The Phytochrome-Interacting Factor PIF7 Negatively Regulates *DREB1* Expression under Circadian Control in *Arabidopsis. Plant Physiology*, 151(4): 2046-2057.

Lenhard, B., A. Sandelin, L. Mendoza, P. Engström, N. Jareborg and W.W. Wasserman. 2003. Identification of conserved regulatory elements by comparative genome analysis .*Journal of Biology*, 2: 13.

Luscombe, N.M., R.A. Laskowski and J.M. Thornton. 2001. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res.*, 29(13): pp. 2860-2874.

Maston, G.A., S.K. Evans and M.R. Green. 2006. Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet*., 1: 29-59.

McFerrin, L., G. and W.R. Atchley. 2011. Evolution of the Max and Mlx Networks in Animals. *Genome biology and evolution*, 3(206): 915-937.

Nakashima, K., Z.K. Shinwari, S. Miura, Y. Sakuma, M. Seki, K. Yamaguchi-Shinozaki and K. Shinozaki. 2000. Structural organization, expression and promoter activity of an Arabidopsis gene family encoding DRE/CRT binding proteins involved in dehydration- and high salinity-responsive gene expression. *Plant Molecular Biology*, 42 (4): 657-665.

Narusaka, Y., K. Nakashima, Z.K. Shinwari, Y. Sakuma, T. Furihata, H. Abe, M. Narusaka, K. Shinozaki and K.Y. Shinozaki. 2003. Interaction between two cis-acting elements, ABRE and DRE, in ABA-dependent expression of Arabidopsis rd29A gene in response to dehydration and high salinity stresses. *The Plant Journal*, 34(2): 137-149.

Narang, V. 2008. Gene regulatory element prediction with bayesian networks. *PhD Thesis.* National University of Singapore. pp 5-8

Okay, O.S., P. Donkin, L.D. Peters and D.R. Livingstone. 2000. The role of algae (Isochrysisgalbana) enrichment on the bioaccumulation of benzo[a]pyrene and its effects on the blue mussel Mytilusedulis. *Environmental Pollution*, 110(1): 103-113.

Parenicová, L., S. de Folter, M. Kieffer, D.S. Horner, C. Favalli, J. Busscher, H.E. Cook, R.M. Ingram, M.M. Kater, B. Davies, G.C. Angenent and L. Colombo. 2003. Molecular and Phylogenetic Analyses of the Complete MADS-Box Transcription Factor Family in Arabidopsis: New Openings to the MADS World. *Plant Cell*, 15: 1538-1551.

Rombauts, S., K. Florquin, M. Lescot, K. Marchal, P. Rouze and Y.V. De Peer. 2003. Computational approaches to identify promoters and cis-regulatory elements in plant genomes. *Plant Physiol.*, 132(3): 1162-1176.

Sandelin, A., W. Alkema, P. Engström, W.W. Wasserman and B. Lenhard. 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, D91-D94.

Satija, R., L. Pachter and J. Hein. 2008. Combining statistical alignment and phylogenetic foot printing to detect regulatory elements. *Bioinformatics*, 10: 1236-1242.

Shahmuradov, I.A., A.J. Gammerman, J.M. Hancock, P.M. Bramley and V.V. Solovyev. 2003. PlantProm: a database of plant promoter sequences. *Nucleic Acids Res.*, 31(1): 114-117.

Si, J., Z. Zhang, B. Lin, M. Schroeder and B. Huang. 2011. Meta DB Site: A meta approach to improve protein DNA-binding sites prediction. *BMC Systems Biology*, 1: S7.

Singh, K.B. 1998. Transcriptional Regulation in Plants: The Importance of Combinatorial Control. *Plant Physiol.*, 4: 1111-1120.

Toledo-ortiz, G., E. Huq and P.H. Quail. 2003. The Arabidopsis Basic / Helix-Loop-Helix Transcription Factor Family. *Society*, 15: 1749-1770.

Verelst, W., H. Saedler and T. Münster. 2007. MIKC* MADS-protein complexes bind motifs enriched in the proximal region of late pollen-specific Arabidopsis promoters. *Plant Physiol.*, 1: 447-460.

Vicente-Carbajosa, J. and P. Carbonero. 2005. Seed maturation: developing an intrusive phase to accomplish a quiescent state. *The International Journal of Developmental Biology*, 49(5-6): 645-651.

Wang, Liangjiang, M.Q. Yang and J.Y. Yang. 2009. Prediction of DNA-binding residues from protein sequence information using random forests. *BMC Genomics*, 10(1).

Xie, Z., S. Hu, S. Blackshaw, H. Zhu and J. Qian. 2010. hPDI: a database of experimental human protein-DNA interactions. *Bioinformatics*, 2: 287-289.

(Received for publication 1 September 2012)