Journal of Computational Intelligence in Bioinformatics.
ISSN: 0973-385X Volume 4 Number 1 (2011) pp. 63—70

http://www.ripublication.com/jcib.htm

In Silico Characterization of DOF (DNA Binding with One Finger) Transcription Factor Genes in Different Crops

Dinesh Yadav[1*], Hariom Kushwaha[2], Shubhra Gupta[1], Sunayana Kashyap[1] and Vinay Kumar Singh[3]

[1]Department of Biotechnology,
D.D.U. Gorakhpur University, Gorakhpur, (U.P.), India
[2]Institute of Microbial Technology, Sector 39-A, Chandigarh, India
[3]School of Biotechnology, BHU, Varanasi (U.P.), India
**Corresponding Author E-mail: dinesh_yad@rediffmail.com*

Abstract

Plants need a large number of transcription factors (TF) for proper and strict transcriptional regulation in response to developmental stages and environmental changes. Some classes of plant transcription factors have DNA-binding domains similar to animal transcription factors, whereas other classes of transcription factors appear to have specifically evolved in plants. A family of TFs putatively specific to plants is the Dof family having a proteins of 200-400 amino acids long with a conserved DNA binding Dof domain of 50-52 amino acid residues structured as a Cys2/Cys2 zinc finger recognizing a *cis*-regulatory element with the common core sequence 5´-AAAG-3´. There exists a great diversity in terms of number of *Dof* genes in plants. Various *in silico* analysis have predicted 30, 36, 24, 31 and 54 *Dof* genes in rice, Arabidopsis, barley, wheat and maize respectively. The diversity of Dof transcription factors is associated with its multifarious roles played in plants. Dof transcription factors have been reported to be associated with regulation of vital processes in plants such as photosynthetic carbon assimilation, light regulated gene expression, accumulation of seed storage proteins, germination, dormancy, response to phytohormones, flowering time, and guard cell-specifc gene expression. There is a need to characterize plant specific transcription factor genes from different crops for the better understanding of gene regulation in response to change in the environment so that effective strategy could be made to create transgenic crops for desired traits. In our lab efforts are being made to decipher the diversity of Dof transcription factor gene(s) in

C4 crop-*Sorghum bicolor* (L) Moench and two of the economically important vegetable crops *Solanum tuberosum* and *Lycopersicon esculentum* using various molecular biology and *in silico* tools.

Introduction

Regulation of gene expression is central to a myriad of biological processes at the molecular level and is controlled mainly at different levels namely chromatin conformation, transcriptional, post- transcriptional, translational, post-translational modification, protein localization and protein turnover. As regulation can occur at many different stages of gene expression, but it is particularly important during transcription.

Gene regulation to a significant extent is controlled by transcription factors (TFs). Most TFs are modular proteins consisting of a DNA-binding domain that interacts with cis-regulatory elements of promoters of its target genes and a protein-protein interaction domain that facilitates oligomerization between TFs or other regulators (Wray *et al*., 2003). DNA-binding domains are, in general, highly conserved and have been used to classify the TFs into families. Sequence divergence in the DNA-binding domains of related TFs may lead to differences in affinities to a set of cis-regulatory elements. Together with the propensity for TFs to homodimerize and/or heterodimerize, the large TF repertoire in a eukaryote genome provides a wide range of combinatorial relationships for transcriptional regulation. TFs usually form gene families that vary considerably in size among organisms (Riechmann *et al*., 2000; Wray *et al*., 2003). The reasons behind such differences are not known, although it is suggested that organismal complexity correlates with an increase in the absolute number and the proportion of TFs in a proteome (Levine and Tjian, 2003).

Transcription factors are important regulators of gene expression comprising of at least four discrete domains, DNA-binding domain, nuclear localization signals (NLS), transcription activation domain, and oligomerization site, which operate together to regulate many physiological and biochemical processes by modulating the rate of transcription initiation of target genes (Du *et al*., 2009).

The Transcriptional regulation of a gene is often governed by number of conserved sequence. The promoter is one of major player in gene regulation by affecting the efficiency of transcription by binding of RNA polymerase to different domains of conserved sequences. These sequence elements are termed as “cis-elements” as they are on the same DNA strand as the coding region of the gene. Some of the sequence elements like TATA box, CAAT box, GC box etc. are crucial in regulation of plant gene expression. Besides the cis-elements sequences, other conserved sequences are also found in promoter region, which are involved with binding of specific proteins, termed as “transcription factors. These transcription factors (or trans-acting factors) are involved in linking various signals (both external and internal) to gene expression and are responsible for determining the level, place and timing of expression.

Plants exhibits number of unique biological processes like photosynthesis, nitrogen fixation, the reproductive process, development and responses to

environmental signals and hence assumed to have some transcription factors that are unique to plants only besides common TFs found in eukaryotes. The term "Zinc finger" represents the sequence motifs in which cysteines and/ or histidines coordinate a zinc atom (s) to form local peptide structures that are required for their specific functions. The Zinc-finger motifs, often classified based on the arrangement of the zinc-binding amino acids, are frequently observed in many transcription factors performing a critical roles in interactions with other molecules involved with gene expression. Some of the commonly found Zinc-finger transcription factors reported in plants are TFIIIA Type, SUPERMAN, WRKY Family,GATA1-Like,RING-finger Type, PHD-finger Type, LIM-Family and DOF-Family etc (Takatsuji,1998;Yadav *et al*., 2008)

The DOF (DNA-binding with One Finger) family represents one of the important class of plant specific transcription factor associated with multifarious roles exclusive to plants and has been extensively reviewed (Takatsuji 1998; Liu et. al 1999; Riechman and Ratcliffe 2000; Yanagisawa 2002, 2004; Lijavetzky et al. 2003; Yadav et al., 2008; Kushwaha and Yadav 2010). Dof proteins are typically composed of 200-400 amino acids with a well conserved DNA binding Dof domain of 52 amino acid residues structured as a Cys2/Cys2 Zn^{+} finger recognizing a *cis* regulatory element with the common core sequence 5´-AAAG-3´ (Yanagisawa and Schmidt 1999; Yanagisawa 2002; Umemura et al. 2004).

There exists great diversity in terms of number of *Dof* genes in different crops. The number of Dof genes predicted in rice, barley, wheat, maize and sorghum is 30, 24, 31, 54 and 28 respectively using various bioinformatics tools (Lijavetzky et al. 2003; Moreno-Risueno et al. 2007; Lindsay et al. 2009; Libault 2009 and Kushwaha et al., 2010). The phylogenetic relationships between rice and Arabidopsis Dof proteins revealed the presence of four major clusters of orthologous genes (MCOGs) (Lijavetzky et al. 2003). The origin and evolution of the Dof transcription factor family based on phylogenetic analysis of Dof sequences across the representative organisms belonging to green unicellular algae to vascular plants has been reported (Moreno-Risueno et al. 2007).

In silico tools used for characterizing the Dof gene/proteins

Various bioinformatics tools could be utilized for characterizing the Dof transcription factor gene family of different crops provided the whole genome sequence information or at least ESTs are available. The recent advances in genome sequencing of crops like pigeonpea (*Cajanu cajan),* tomato *(Lycopersium esculentum),* wheat *(Triticum vulgare)* could be utilized for *in silico* prediction of diversity of *Dof* genes. In order to characterize the Dof transcription factor gene family in a particular crop involves following steps

Database searches for the identification of Dof family members

The Dof sequences of crops needs to be retrieved from available GenBank. The nucleotide sequences of conserved Dof domain is generally used to search the

potential Dof domain homologs hit in the whole genome sequence or the ESTs of that particular crop through BLASTN, TBLASTN and discontinuous MEGABLAST (Altschul et al. 1990, 1997) at the NCBI database (http://www.ncbi.nlm.nih.gov) (Pruitt et al. 2007). Further the annotated sequences could be subjected to bioinformatics softwares namely FGENESH (Solovyev et al. 2006) and GENESCAN (Stormo, 2000) for prediction of full length genes with putative CDS and protein sequences. The putative Dof protein sequences can be subjected to protein functional analysis using PFAM version 24.0 (Finn et al. 2006), PROSITE version 20.58 (Castro et al. 2006), INTERPROSCAN version 24 (Quevillon et al. 2005), and MOTIFSCAN (Falquet et al. 2002). The Dof protein sequences can also be analyzed for nuclear localization signal (NLS) by subjecting it to NUCPRED server (Barmeier et al. 2007).

Mapping of *Dof* genes on chromosome and its intron/exon gene structure prediction

Each of the predicted *Dof* genes could be was positioned on chromosome of a particular crop by the BLASTN search with NCBI genomes (chromosome) database. The resulting position of predicted *Dof* genes on the available chromosome can be manually marked on bar. The gene structure of predicted *Dof* genes can be carried out using FGENESH software and intron/exon structures can be manually designed (Figure-3).

Dof protein alignment and phylogenetic analysis

The identified Dof proteins can be aligned using ClustalX2.0.10 (Thompson et al. 1998) (Figure-1). The phylogenetic tree could be constructed by neighbor joining method (Saitou and Nei 1987).

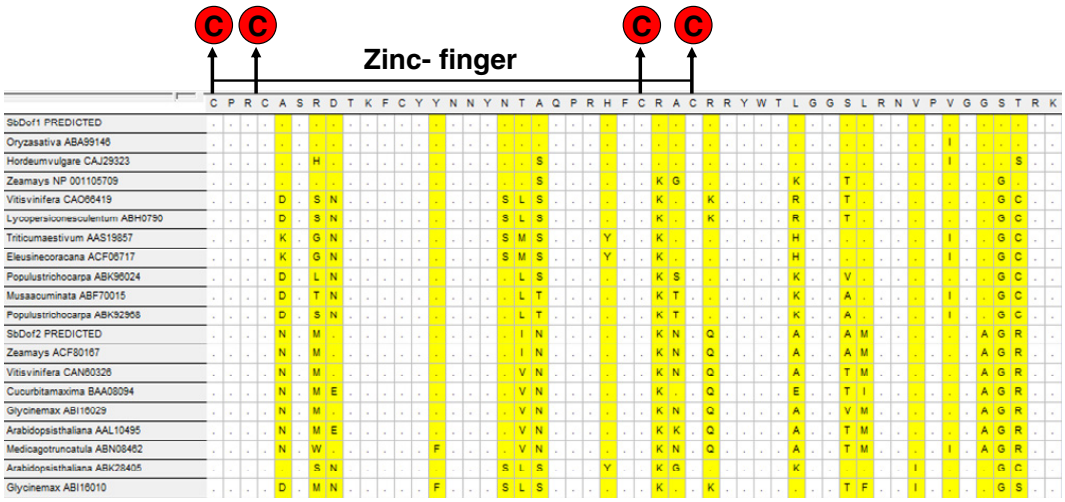


Figure 1: Variations in amino acid residues of Dof domain region of different *Dof* genes predicted for *S. bicolor* subjected to multiple sequence alignment. The yellow portion shows variable sites.

Identification of Conserved motif, Motif scan, and *Cis*- regulatory element analysis

Further the deduced protein sequences of the predicted *Dof* genes of a particular crop along with the available Dof protein of rice, Arabidopsis, sorghum whose whole genome sequence information is available could be analyzed by means of the MEME (Multiple EM for Motif Elicitation) program software version 4.4.0 (Bailey et al. 1998, 2006) for motif analysis (Figure-2). For promoter analysis 500 bp to 1000 bp upstream sequences from the initiation codon of the putative *Dof* genes need to be retrieved and then subjected to search for CARE program (http://bioinformatics.psb.ugent.be/webtools/plantcare/html/) of PlantCARE databases (Lescot et al. 2002) for identification of *Cis*-regulatory elements.

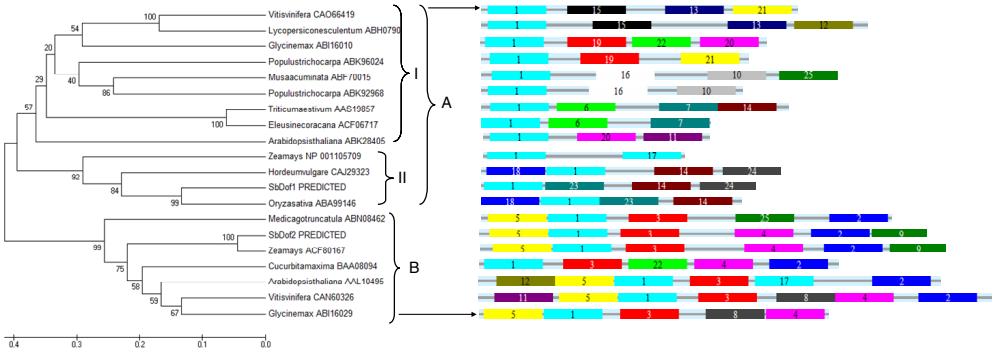
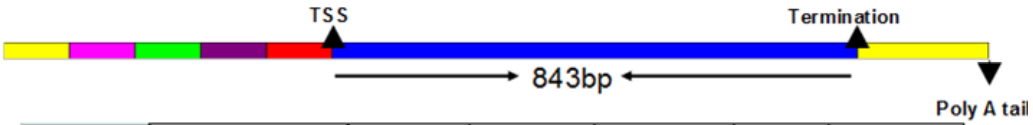


Figure 2: Phylogenetic tree constructed based on DOF protein sequences and schematic distribution of respective conserved motifs identified by means of MEME software.

The deduced protein sequences of Dof genes can also be subjected to motif scan. Motif scan integrated with PeroxiBase profiles, PROSITE patterns, PROSITE profiles, HAMAP profiles, Pfam HMMs (local models) and Pfam HMMs (global models) databases will provide information about the presence of different amino acid which is abundantly distributed and also conserved for Dof proteins.



-1 to -200	5UTR Py-rich stretch	ATC-motif	GCN4_motif	TA-rich region	TATA-box	TGA-element
-200 to -400	MNF1	Sp1				
-400 to -600	RY-element					
-600 to -800	CCAAT-box	G-box				
-800 to -1000	CAAT-box	Skn-1_motif	circadian			

Figure 3: Putative gene structure of one of the predicted *Dof* gene of sorghum with distribution of commonly observed *cis* regulatory elements based on *in silico* analysis of 1000 bp upstream region.

Protein sequence and two dimensional structure analysis

Protein sequence analysis and two dimensional structure prediction of Dof proteins can be carried out by PROTEAN program (Plasterer, 1996), one of seven modules in the *LASERGENE* suite which analyzes and predicts protein characteristics and motifs from primary sequence data. The translated protein of Dof genes could be taken for structural classification and prediction and properties like molecular weight, pI value can be determined. Further two dimensional proteins structural analyses can also reveal the presence of alpha, beta, turn and coil region based on Garnier-Robsen method (Garnier et al., 1978). The resulting hydrophobicity plot using Kyte-Doolittle method (Kyte and Doolittle, 1982) can be analyzed for various structural features typical to Dof like proteins (Figure-4).

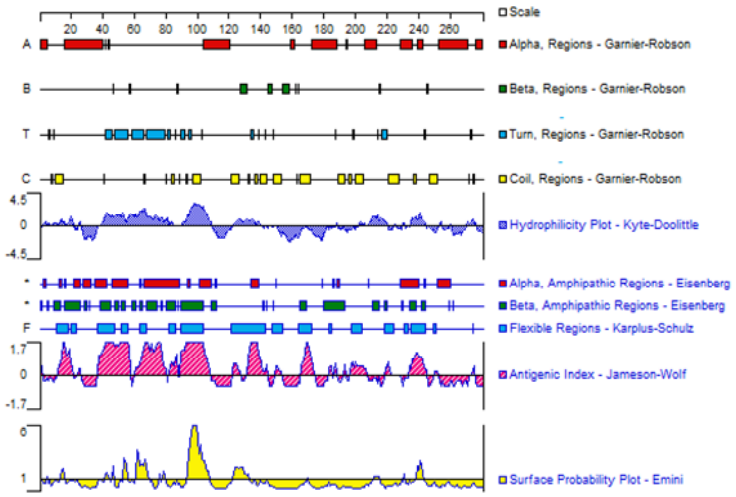


Figure 4: Graphical representation of structural properties of Dof protein.

***In silico* investigation of Dof transcription factor gene family of cereals and millets**

Attempts have been made to use various bioinformatics tools for characterizing Dof gene family in cereals, millets and recently in sorghum crop. Sequence analysis of PCR amplified 13Dof domains of cereals (rice, wheat, sorghum, barley, oat, maize, barnyard millet, proso millet, little millet, kodo and foxtail millet) and 5 Dof genes (rice, wheat, barley, maize and finger millet) revealed its identity to Dof like proteins (Kushwaha et al., 2008). Further *in silico* investigation of the cloned *Dof* genes of finger millet, barley, wheat and maize revealed its identity to PBF Dof based on the presence of motifs related to regulation of endosperm specific seed storage protein genes (Kushwaha et al., 2008).

Based on published whole genome shotgun sequence of *Sorghum bicolor* (L.) Moench (Paterson, et al. 2009), the whole set of Dof genes have been characterized for multiple sequence alignment, gene structures, phylogeny, chromosome location and cis regulatory elements analysis. *In silico* analysis revealed existence of 28 copies of C2C2-like *Dof* genes in *S. bicolor* genome. Multiple sequence alignment of these

*Sb*Dof proteins showed well conserved four cysteine residue and phylogenetic analysis resulted in to four subgroups constituting six clusters. Further analysis of intron/exon gene structures revealed majority of the predicted *Dof* genes to be intronless as observed in case of rice and Arabidopsis. The *cis*-regulatory element analysis of the predicted *Dof* genes revealed the major putative functions as regulation of genes associated with seed storage proteins, abiotic and biotic stress, photoperiod, growth hormone and meristem (Kushwaha et al., 2010).

Conclusion

The international efforts for sequencing plant genomes will results in the accumulation of vast sequences information which needs to be annotated for putative functions. The existing diversity of transcription factor like *Dof* gene family as studied in rice, Arabidopsis, sorghum whose whole genome sequence information is available need to be further analyzed in other crops. The bioinformatics tools could be utilized for deciphering the diversity of *Dof* genes in a particular crop as soon as its whole genome sequence information is made available. The analysis of such TFs in plants might provide some clue to the diverse gene regulation observed in response to various stresses. Targeting a transcription factor like Dof might be a powerful tool for crop improvement based on the fact that a single transcription factor regulates expression of multiple related genes. Further the advances in the science of genomics, transcriptomics and proteomics with the recent availability of whole genome sequence information of many crops has shifted the 'gene-centric' to 'genome centric' approach for analyzing transcription factor gene family.

References

[1] Altschul, S. F. et al., 1990. *Journal of Molecular Biology* 215:403-410.
[2] Altschul, S. F. et al., 1997. *Nucleic Acids Research* 25: 3389-3402.
[3] Bailey, T. L. et al., 2006. *Nucleic Acids Research* 34: W369–W373.
[4] Bailey, T. L. et.al. 1998. *Bioinformatics* 14: 48-54
[5] Brameier, M. et al., 2007. *Bioinformatics* 23(9): 1159-1160.
[6] Castro, E.D., et.al., 2006. *Nucleic Acids Research* 34: W362-W365.
[7] Du H., 2009, Biochemistry (Moscow) 74: 1-11
[8] Falquet, L., et.al., 2002. *Nucleic Acids Research* 30:235-238.
[9] Finn, R.D. et. al., 2006. *Nucleic Acids Research* 34: D247–D251
[10] Kushwaha, H. et al. 2010, *Current topics on Bioprocesses in Food Industry, volume III Asiatech, New Delhi*, 150-168.
[11] Kushwaha, H. et. al., 2008. *Online Journal of Bioinformatics* 9(2):130-143.
[12] Lescot M, et. al., 2002. *Nucleic Acids Research* 30(1): 325-327.
[13] Levin M. 2003, *Nature* 424; 147-151
[14] Libault M, et.al., 2009. *Plant physiology* 151: 991–1001,
[15] Lijavetzky D, et. al. 2003. *BMC Evol Biol* 3:17
[16] Lindsay M. et.al., 2009. *Funct inter genomics* 9:485-498

[17] Liu L. et. al., 1999., *Eur. J. Biochem*. 262:247-257
[18] Miguel Angel Moreno-Risueno, et al., 2007 *Mol Genet Genomics.* 277:379-390
[19] Paterson et al 2009, *Nature*, 457: 551-556
[20] Pruitt K.D. et. al., 2007 . *Nucleic Acids Research* 35:D61–D65.
[21] Quevillon E. et. al., 2005. *Nucleic Acids Research* 33: W116–W120.
[22] Riechmann L. J. et.al. 2000. *Current opinion in Plant Biology* 3: 423-434
[23] Saitou N. et. al., 1987. *Mol Biol Evol* 4:406–425
[24] Solovyev V., et.al. 2006. *Genome Biology* 7(1): S10
[25] Stormo G.D. 2000. *Genome Research* 10:394-397
[26] Takatsuji H. 1998. *Cell. Mol. Life Sci.*54:582-596
[27] Thompson J.D., et al. 1997 . *Nucl Acids Res* 25:4876–4882
[28] Umemura Y., et.al. 2004 *Plant J* 37:741–749
[29] Wray G.A., 2003, *Mol Biol Evol* 20: 1377-1419
[30] Yadav. D., et.al., 2008. *Ecosystem diversity and carbon Sequesteration; climate Change Daya publishing house*, 187-197.
[31] Yanagisawa S. 2002. *Trends Plant Sci* 7:555–560
[32] Yanagisawa S. 2004. *Plant Cell Physiol* 45(4): 386–391.