ISSN (Online): 2091-2609
DOI Prefix: 10.3126/ijasbt

International Journal of Applied Sciences and Biotechnology (IJASBT)

Research Article

In-Silico Structural and Functional Characterization of WsMYB44 Protein from Withania somnifera L. Dunal

Lakee Sharma, Bipin Maurya, Shashi Pandey Rai*

Laboratory of Morphogenesis, Department of Botany, Institute of Science, Banaras Hindu University (BHU), Varanasi-221005 (Uttar Pradesh), India

Article Information

Received: 14 November 2022
Revised version received: 19 December 2022
Accepted: 22 December 2022
Published: 30 December 2022

Cite this article as:
L. Sharma et al. (2022) Int. J. Appl. Sci. Biotechnol. Vol 10(4): 201-215. DOI: 10.3126/ijasbt.v10i4.50850

***Corresponding author**
Shashi Pandey Rai,
Laboratory of Morphogenesis, Department of Botany, Institute of Science, Banaras Hindu University, Varanasi-221005, Uttar Pradesh, India
Email: shashi.bhubotany@gmail.com

Peer reviewed under authority of IJASBT

Abstract

The plant based MYB transcription factors have been extensively studied for structural and functional characterization. MYB protein is known to perform several molecular and biological processes including regulation of gene expression. In the current study, we have identified and characterized a R2R3 type MYB TF (WsMYB44) pertaining to root specific library from *Withania somnifera*. The phylogenetic studies of the identified protein sequence revealed its recent origin from *Nicotiana sp*., a member of Solanaceae family. Motif scans and domain analysis depicted strong conservation near signature R2 and R3 DNA binding domains which proved the specificity of the WsMYB44 protein. The modeled 3-D structure of protein showed significant binding affinity with the MYB binding DNA element C/TAACG/TG. The synergistic act of the predicted protein was proved by studying protein-protein interaction networks. Furthermore, Gene Ontology terms predicted nuclear localization of protein which was found to be involved in several defense related signaling cascades.

Keywords: WsMYB44; SANT domains; phylogeny; DNA-protein interaction; *Withania somnifera*

Introduction

Sessile behaviour of the plant species become a major constraint in stress condition. Thus, during the course of evolution, plants have evolved a plethora of mechanisms to endure unpleasant surroundings. In this context, specific conserved proteins called transcription factors (TFs) play a crucial role in multitudinal ways to bypass stress toxicity towards enhanced vigour of the plant. There are many such TFs reported like DREB, WRKY, AP2-ERF, HSP, MYB, MYC, NAC, BHLH, WDR etc. These proteins are the products of trans-acting genes and binds with the promoter region of stress responsive genes in order to regulate them in the favour of plant (Rai *et al.*, 2019). Among all TF families, MYB protein is known to take part in both stress tolerance as well as in developmental aspects of all eukaryotes including plants. First plant based MYB protein was identified in maize which was a C1 type protein (Paz-Ares *et al.*. 1987). At present, this protein is thoroughly

identified and characterized in several model and non-model medicinal as well as crop plants. MYB protein family is one of the largest plant TFs which is involved in both pathogens induced defence machineries and also tackles with environment generated abiotic constraints by affecting the indigenous genes of stressed plant. MYB protein is identified by conserved DNA-binding domains (MYB/SANT domain) at the N-terminus while the C-terminus being variable in nature (Dubos *et a.l*, 2010). On the basis of presence of adjacent repeats in the form of DNA-binding domains, MYB proteins can be categorized in R1/2 (1R-MYB), R2R3-MYB (2R-MYB), R1R2R3 (3R-MYB), and R1R2R2R1/2 (4R-MYB). All types are present in plants but R2R3-MYB is the most abundant. Primary structure of typical R2R3-MYB protein can be represented by -W-(X_{19})-W(X_{19})-W-.....F/W-(X_{18})-W-(X_{18})-W- while secondary structure is denoted by presence of three helices in each imperfect repeat. However, in plants, one of the tryptophan residue of R3 repeat is often substituted by another aromatic amino acid i.e. phenylalanine or isoleucine (Ambawat *et al.*, 2013). Besides these functions, MYB TFs are also involved in cell cycle regulation, cellular proliferation, differentiation, reproduction, phenylpropanoid metabolism, formation of trichomes and essential oils, morphogenesis, and pigmentation (anthocyanin synthesis) of plants. Such function of MYB proteins is greatly dependent upon its capability to bind with the major grooves of several plant defence and secondary metabolite related genes. Majority of the plant R2R3 MYB TFs recognize nucleotide sequences such as ACC(A/T)ACC(A/C/T, ((T/C)AAC(G/T)G(A/C/T)(A/C/T)), ((T/C)AAC(G/T)G(A/C/T)(A/C/T)), (AGTTAGTTA), and ((C/T)ACC(A/T)A(A/C)C etc (Romero *et al*., 1998). Choi *et al*. (2017) reported the possible binding sites of *Panax ginseng* MYB as C/TAACG/TG. These nucleotide patches are totally different from the binding sites of animal MYB proteins. In addition, role of this protein in stress tolerance is also performed by its synergistic activities with other such TFs (Prouse and Campbell 2012).

In the present study, we have identified a MYB44 like TF from root specific library of *Withania somnifera* (from NCBI database). Among plethora of MYB TFs in plants, the role of MYB44 protein in abiotic stress tolerance has been well studied in Arabidopsis (Persak and Pitzschke 2014). MYB44 is a R2R3 type transcription factor known to be directly involved in the transcriptional reprogramming of genes which are involved in stress signalling cascades. Role of MYB in the development and morphogenesis of plant's underground part has been extensively studied by several workers (Feng *et al*., 2004; Mu *et al*., 2009). Roots are the special tissues of the plants involved in water and nutrient uptake. Roots of *Withania somnifera* is reported as repertoire of the phytoactive compound i.e. withanolide A (Misra *et al.*, 2008). *W. somnifera* is a solanaceous plant, known for its medicinal properties and phytoactive compounds like triterpenoid steroidal lactones (Withaferins and withanolide A), flavonoids, etc. According to Misra *et al.* (2008) aerial parts of the plant (leaves) is rich in Withaferin A while, plenty of Withanolide A resides in the roots. In medicinal plants, abiotic stress mediated responses and secondary metabolism are regulated by many TFs like MYB (Cao *et al.*, 2020). In spite of several research works and findings, *W. somnifera* is an orphan plant in the field of genome database. Due to unavailability of Whole Genome Sequence (WGS) we have retrieved a nucleotide sequence pertaining to MYB44 TF from Expressed Sequence Tags (EST) database of *W. somnifera*. Since there is lack of any previous knowledge in context to motifs, binding sites, active regions, sequence conservation, phylogenetic relationships, three dimensional structure, etc of MYB TF in *W. somnifera*, this report may provide a hypothetical mechanism about functioning of this protein in stress tolerance. In spite of this, a comprehensive and comparative study of structures of hypothetical modelled protein (WsMYB44) with structurally resolved templates validates the novel protein. The functionality of the predicted protein has also been elucidated with the help of Gene Ontology and its vocabularies.

Methodology Opted

Sequence Identification and Phylogenetic Analysis

The nucleotide sequence of *WsMYB44* gene was retrieved from root cDNA library of *Withania somnifera* through NCBI (Accession no. GR923779.1). The open reading frame (ORF) was determined by employing NCBI blastX online tool. Further, Pfam and BLAST-TAIR tools were employed for identification and nomenclature of the protein respectively (Bateman *et al*., 2004). All the homologous sequences having significant sequence identity were downloaded as a result of protein BLAST-P analysis and Multiple Sequence Alignment (MSA) was performed by ClustalW https://www.ebi.ac.uk/Tools/msa/clustalo (Larkin *et al*., 2007) server of T-Coffee. Finally, the phylogenetic tree was generated with MEGA 6 software (Tamura *et al*., 2013). During tree construction UPGMA method was selected and replications values at 1000 bootstraps were uploaded in MEGA tool.

Analysis of Conserved Domain, Consensus Motifs, and Active Sites

The functional and conserved DNA- binding domains (DBDs) i.e. R2R3-MYB domains were identified by employing PROSITE and SMART server (Hofmann *et al*., 1999; Ponting *et al*., 1999). The Conseq server of ConSurf online tool was used to detect the extent of sequence conservation at residual level, within the functional domains of predicted protein https://consurf.tau.ac.il/). The arrangement pattern and distribution of consensus motifs residing in WsMYB44 protein were scanned by using

MEME Suite 4.1.1.2 (Bailey *et al*., 2006). For motif analysis all the values were obtained at default parameters.

Active sites are the pockets of several amino acid residues within the protein where ligands bind by covalent bonds. Most appropriate target site in a protein can be in a range of 25.0 Å from the ligand coordinates. Thus, significant active sites in modelled protein were predicted by uploading the pdb file in SCFBio, IIT Delhi India (http://www.scfbio-iitd.res.in/dock/ActiveSite.jsp).

3-D structure prediction and model validation of WsMYB44 protein

The protein sequence of WsMYB44 TF was used for homology modelling and similar resolved templates were obtained from SWISS-MODELLER https://swissmodel.expasy.org/ and I-TASSER https://zhanggroup.org/I-TASSER/ tools. Three structurally resolved templates having highest structural similarities were chosen which are *viz*, crystal structure of ternary protein-DNA complex3 (PDB ID: 1h8a.1), crystal structure of ternary protein-DNA complex2 (PDB ID: 1h89.1), and crystal structure of ternary protein-DNA complex1 (PDB ID: 1h88.1). By employing these templates, we have generated a 3-D model of N-terminus of predicted protein of *Withania somnifera*. The least energy model was refined by ModRefiner tool (Xu and Zhang, 2011). Topological features were analyzed by superimposition of refined model with each template with the help of Superpose version 1.0 http://superpose.wishartlab.com/ while alignment was done by MultAlin tool.

The reliability and accuracy level of the refined model was checked by SAVES server https://saves.mbi.ucla.edu/ which include several tools like ERRAT, Verify3D, PROVE, and PROCHECK (for Ramachandran Plot analysis). The plots in the form of backbone dihedral psi (ψ) and phi (ϕ) angels of predicted model as well as structurally resolved templates were generated simultaneously for comparative validation. RESPROX and ProSA https://prosa.services.came.sbg.ac.at/ servers were also used for atomic resolutions and Z-score respectively. The quantitative validation was performed by VADAR online tool http://vadar.wishartlab.com/analysis (Willard *et al*., 2003). Next, the final model qualifying all the validation levels was deposited in PMDB (protein modelling database) https://bioinformatics.cineca.it/PMDB/ (Castrignano *et al*., 2006).

Interaction Analysis of Wsmyb44 Protein with DNA and Protein

Interactome analysis of WsMYB44 protein was done by studying DNA-protein and protein-protein interactions. The Patchdock server was employed for protein and target DNA (C/TAACG/TG) interaction (Schneidman-Duhovny, 2005). The protein and DNA docked complex having least energy was selected and visualized by Discovery studio 3.0. The protein-protein interactomes of WsMYB44 protein were evaluated by STRING server (Search Tool for the Retrieval of Interacting Genes/ Proteins) http://string-db.org/ (Szklarczyk, 2015). The first and second shell interactors were studied at 0.40 confidence score.

Functional Annotation of Modelled Protein

Functional annotation (in molecular, biological, and cellular terms) of modelled protein was performed by COFACTOR server through enzyme commission, gene ontology, and ligand binding sites (Roy *et al*., 2012). Further structural classification of the modelled protein was predicted by CATH server and functional classifications of identified super-families were done with FunFHMMer http://www.cathdb.info/search/byfunfhmmer. The results were further analyzed using gene ontological vocabularies (Das *et al*., 2016). With the help of identified GO terms, a distribution test analysis was performed with using the REVIGO server http://revigo.irb.hr/ (Supek *et al*., 2011).

Analysis of Wsmyb44 Protein for Sub-Cellular Localization and Physicochemical Characteristics

For sub-cellular localization the protein was submitted to online server Cello2GO http://cello.life.nctu.edu.tw/cello2go/ (Yu *et al*., 2014). The physical characters (Isoelectric point, molecular weight, and Hydropathicity index) were estimated by ProtParam server of ExPaSy online tool http://www.expasy.org/tools.**To detect the presence of any trans-membrane domain, protein sequence was submitted to T**MHMM Server v. 2.0 http://www.cbs.dtu.dk/services/TMHMM/, and hydropathic behaviour of the protein was analyzed using ProtScale https://web.expasy.org/protscale/ and GRAVY tools http://bioinformatics.org/sms2/protein_gravy.html (Kyte *et al*., 1982).

Results

Database search and comparative phylogenetic analysis

Nucleotide sequence of *Withania somnifera* root library (Accession no. GR923779.1) was translated by ExPaSy translation tool which showed full length protein sequence for R2R3 type MYB transcription factor. Further, homologues sequences for the predicted protein was searched by BLAST alignment tool and listed for comparative phylogeny. The phylogenetic analysis revealed close similarity of WsMYB44 protein with its wild homologues *viz*, XP_019228483.1 (*Nicotiana attenuata*) and XP_009588514.1 (*Nicotiana tomentosiformis*) (Fig. 1a). Contrary to this, MYB44 protein sequences belonging to other family plants such as *Gossypium* sp. were grouped in a separate- out clade. These outcomes have clearly depicted that WsMYB44 protein was recently evolved from the wild homologues/paralogues of solanaceae family. Next, multiple sequence alignment of our protein with all the selected homologues proteins indicated the strong sequence conservation around N-terminus of the protein,

especially covering the R2 and R3 MYB DNA-binding domains (DBDs) (Fig. 1b).

For further validation of sequence conservation level, the amino acid sequence of query protein was submitted to ConSurf server which also revealed that significantly conserved residues to be residing in N-terminus while C-terminus was observed to be highly variable in nature (Fig. 2) which might indicate the DNA binding and regulatory role of N- terminus and C- terminus respectively. This fact also indicated least sequential disturbances in core residues of the WsMYB44 protein indicated its evolutionary significance during adaptation against several biotic and abiotic stresses.

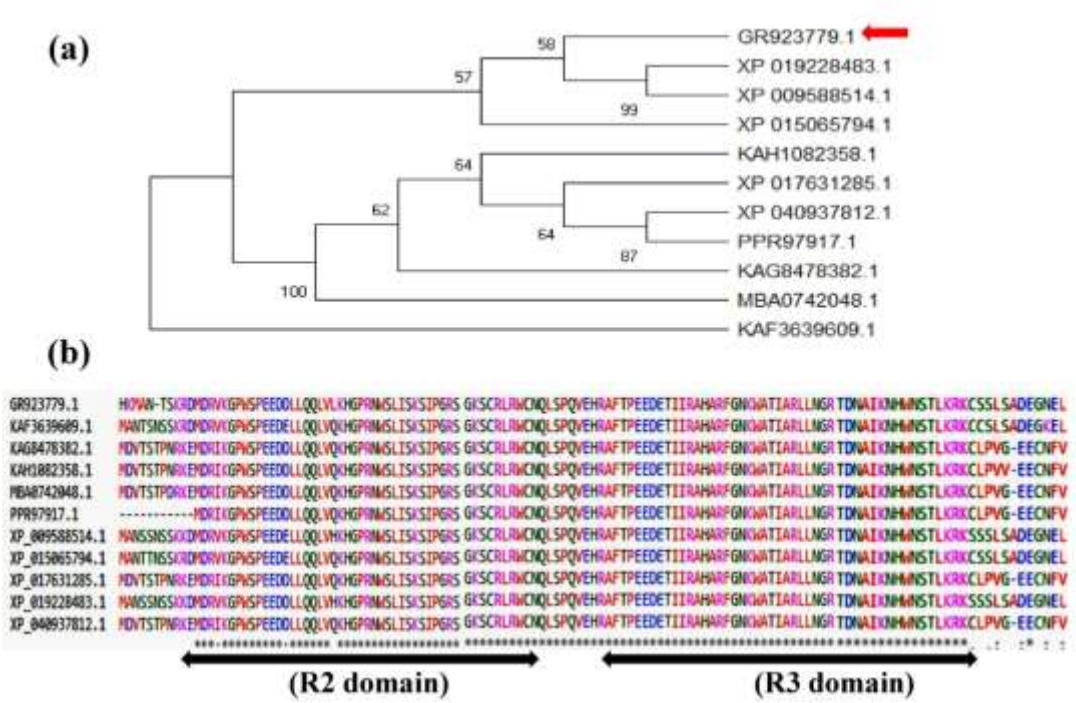

Fig. 1: Phylogenetic tree representing (a) evolutionary origin and relationships, and (b) Sequence alignment of N-terminal R2 and R3 MYB domains of *Withania somnifera* MYB44 (WsMYB44) protein

Fig. 2: Analysis of sequence conservation in WsMYB44 protein by ConSurf server. Red, green, and yellow color showing highly, average, and variably conserved nature of amino acid residues

Identification of conserved domains, motifs, and active sites in WsMYB44 protein

For prediction of conserved domains, various online tools such as InterProScan, SMART, and ExPaSy-PROSITE servers were employed. The outputs indicated that length of our protein to be 94 (8^{th} to 103^{rd}) amino acids that contains functional signature sequences in the form of R2 and R3 SANT domains. The R2 domain was indicated by 42 amino acids, fetching from Asp^{8} to Asn^{50} while R3 domain constituted by 43 (60^{th} to 103^{rd}) amino acids (Fig. 3). Both the functional domains were separated from each other by a small stretch of 9 amino acid residues. The overall signature sequence in conserved domain of WsMYB44 protein was denoted by -W-(X_{19})-W(X_{19})-W- (R2).....F/W-(X_{18})-W-(X_{18})-W- (R3) where first domain contained three regularly spaced tryptophan residues while first tryptophan residue of R3 domain was substituted by phenylalanine (Fig. 1b). This showed that mild adaptive mutations have occurs during the course of evolution.

Most significant DNA- binding regions in a transcription factor can be statistically obtained with the help of MEME motif scanning tool. It provides nearly exact consensus sequences, sharing commonalities amongst all the homologues and query protein which may help to understand the regulatory mechanism of respective TFs. Total 10 motifs (represented in the form of sequence logo) were identified that showed highest sequence conservation at MYB domain. The motif statistics were validated in the form of *E*- values (denotes frequency of occurrence of motifs) and *p*- values (provides probabilities of occurrences). In the presented study, we have identified three significant sequence logos for N- terminus, covering the R2 and R3 domains of WsMYB44 protein (Fig. 4a). Detailed study revealed that R2 domain was represented by motif no. 3 (MDRIKGPWSPEEDDLLQQLVQKHGPRNWSL; *p*-value 7.67 e-37) and motif no. 2 (SGKSCRLRWCNQLSPQVEHRAFTPEEDETI; *p*- value 2.87e-39) while R3 domain by motif no. 2 and motif no. 1 (IRAHARFGNKWATIARLLNGRTDNAIKNHW; *p*-value 3.64e-41). The respective locations and distributions of motifs on WsMYB44 protein as well as on all the homologues proteins were observed and represented in Fig. 4b. Contrary to this, the C-terminus was observed to be highly variable in nature and represented by motif no. 4 (EERGGGVMGFSAEFMAVMQEMIRVEVRNYM; *E*-value 5.2e-172) and motif no. 5 (NSTLKRKCLPVGEECNFVP; *E*- value 8.2e-091). Thus, motifs identified by MEME and MAST analysis have depicted that R2 and R3 DBDs reside in the N-terminus of WsMYB44 protein and are highly conserved in nature. While the C-terminus of the protein seemed to harbour variable sequences.

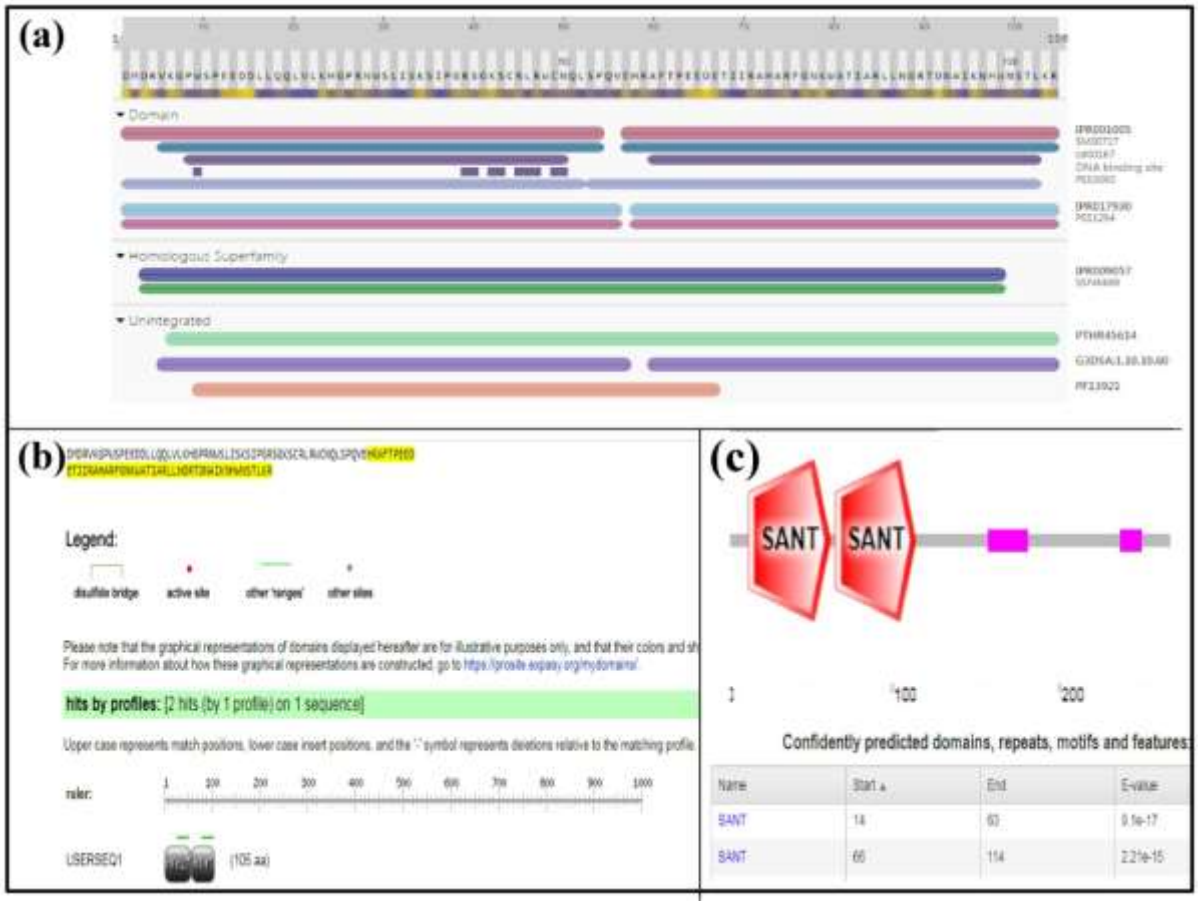

Fig. 3: Analysis of conserved domains in WsMYB44 protein by (a) InterproScan, (b) PROSITE, and (c) SMART servers

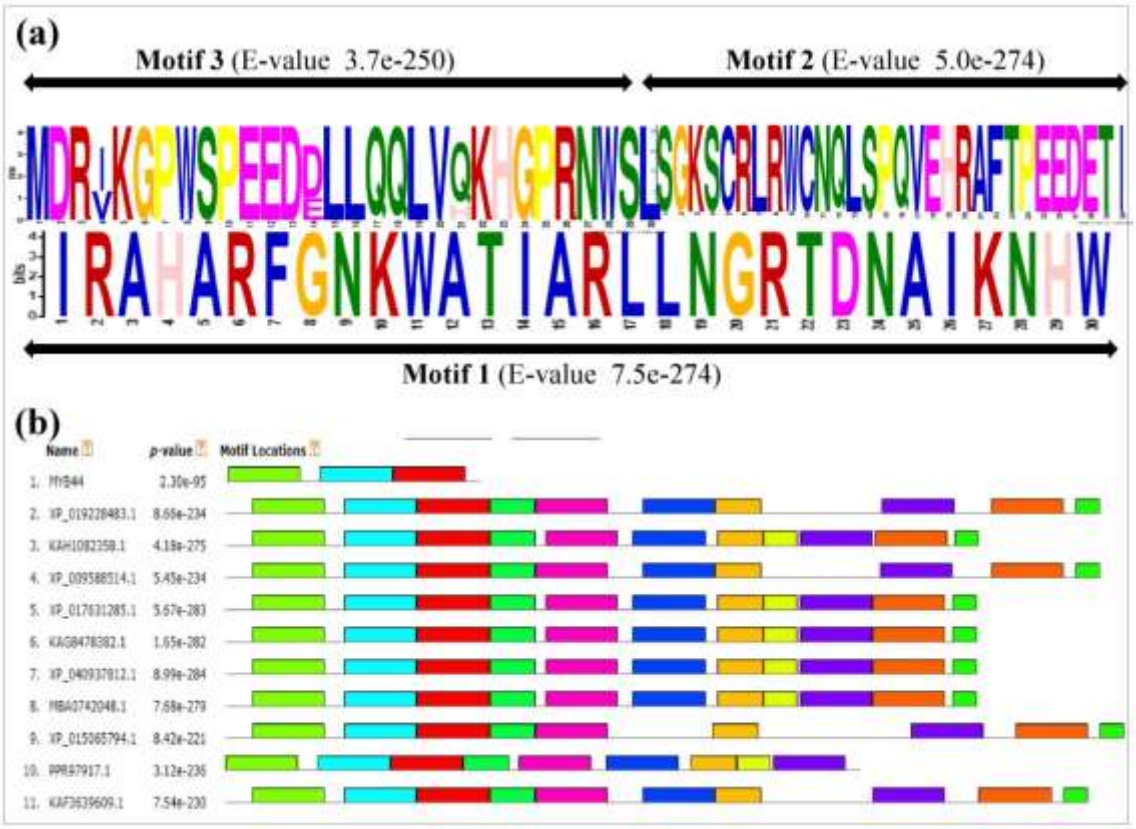

Fig. 4: Analysis of motif scan showing conserved motifs and their locations in WsMYB44 protein; (a) motifs covering R2 and R3 MYB domains, and (b) distribution pattern of the scanned motifs

Table 1: Comparative analysis of active pockets in predicted WsMYB44 protein and respective templates

S. no.	Cavity ID	Active sites (amino acid residues)	Cavity points			Cavity volume
1	Cavity_1	WKGSLRQNHACTDPVE	98.094	169.067	12.538	1057
2	Cavity_2	KRGVPSWLNQHACTDIE	91.982	167.861	16.578	1014
3	Cavity_3	RKDLQNWHACTPSFVEG	99.402	162.982	13.794	819
4	Cavity_4	KRVDGPISWELQNHACT	85.824	163.867	12.706	784
5	Cavity_5	RNSQKTHALDFVIEGP	96.407	153.740	21.962	549
6	Cavity_6	KVPGSRWELDQNHACT	90.312	158.034	9.844	459

Active sites are conserved regions on the surface of protein which provide attachment cavities for specific ligand. For prediction of active pockets in WsMYB44 protein, the modelled 3D-structure was submitted to online web server SCFBio. The output showed total six cavities, each with three cavity points, suitable for attachment of various ligands (Table 1). Most of these pockets were laid in MYB DBDs which again showed the regulatory behaviour of MYB TF in expression of various stress responsive genes.

Structural Modelling and Model Validation of Wsmyb44 Protein with Identical Templates

For *in-silico* modelling of WsMYB44 protein a three-dimensional (3-D) model was generated on the basis of available structurally resolved templates. Since only N-terminus contains conserved MYB domains thus for structural modelling, amino acid sequences from variable C- terminus was discarded. Next, three structurally resolved templates *viz;* 1h8a, 1h88, and 1h89 were selected on the basis of percent identity, sequence completeness, and coverage score from I-TASSER tool. The model was predicted with the aid of SWISS-MODEL server and visualized by Discovery Studio (DS) Client 3.0. Out of three, model with least energy and good spatial parameters was finalized for further refinement by uploading the generated pdb file into Modrefiner server (Fig. 5a). The refinement tool finally generated most resolved model highlighting R2 and R3 MYB domains which was more reliable than unrefined one in terms of RMSD, ERRAT score, and electrostatic energy (Fig. 5b and c). At last, the finalized model was submitted in PMDB protein repository database and a PMDB identifier with unique ID (PM0084234) was assigned.

For validation, evaluation and reliability check several qualitative and quantitative parameters of the predicted model were calculated and compared with all the three structurally resolved templates (Fig. 6). For qualitative assessment of WsMYB44 protein Ramachandran plots were evaluated which were generated with the help of RAMPAGE web server and also validated by PROCHECK analysis of PdbSum tool. Outcomes of these servers

concluded that 95% of residues of our model were detected in most favoured regions while 0.0% (no any residue) was located in disallowed and outlier region (Fig. 6a; Table 2). These values were superior to all the three resolved templates, which showed reliability of the predicted model. ProSA and QMEAN (Qualitative Model Energy Analysis) were analyzed for minimal structural difference in the form of Z-score and geometrical aspects respectively which revealed very close resemblance of model with templates (Fig. 6b and 6c).

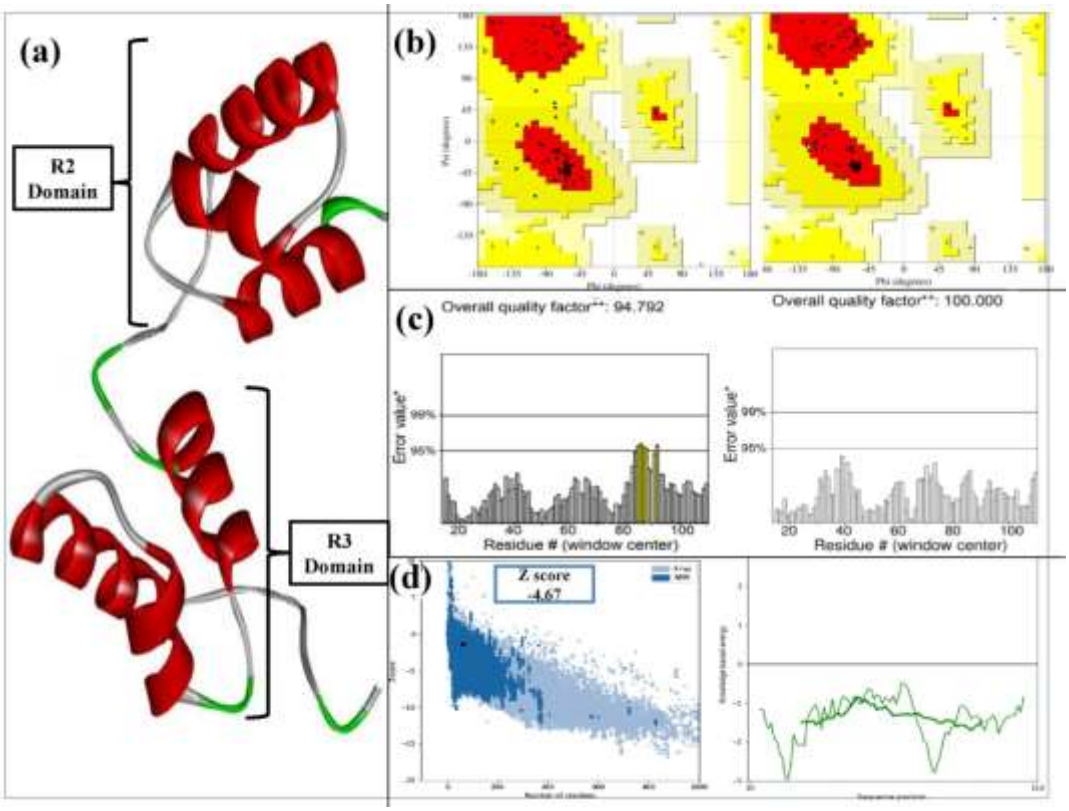

Fig. 5: (a) Prediction of 3-D model for WsMYB44 protein, (b) Ramachandran plot showing reliability of unrefined and refined models, (c) ERRAT scoring for unrefined and refined model, and (d) Z-score of refined model

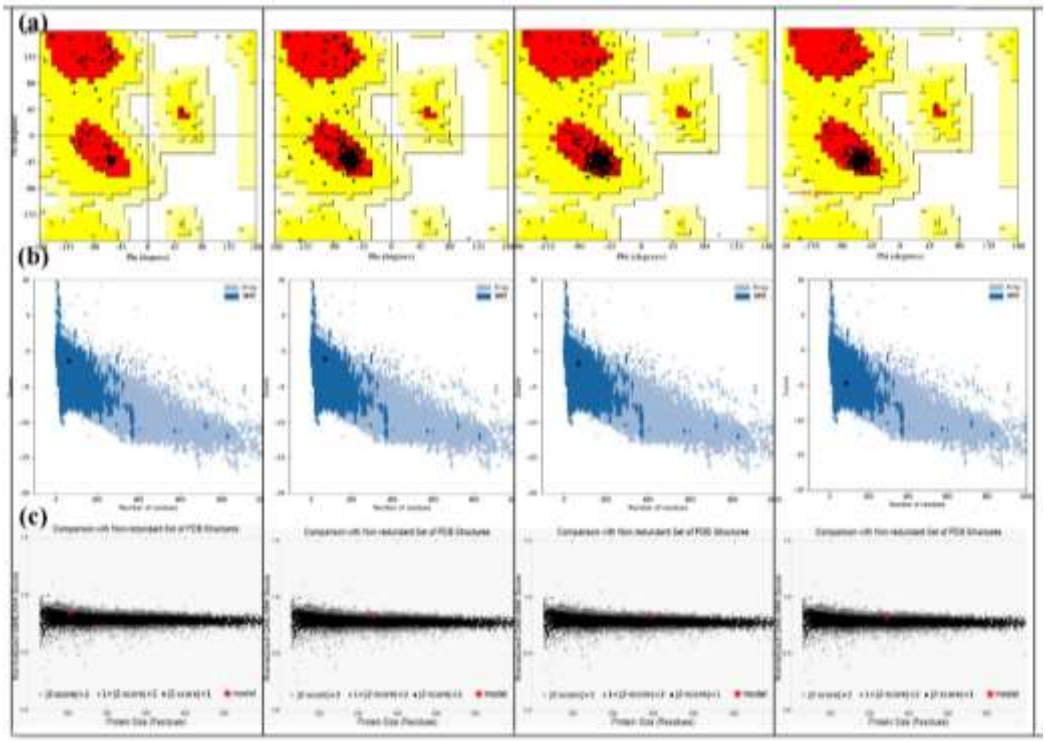

Fig. 6: Comparative analysis of predicted model and resolved templates for qualitative parameters using; (a) PROCHECK, (b) ProSA, and (c) QMEAN scoring

Table 2: Qualitative and quantitative analysis of modelled WsMYB44 protein

S. no.	Protein name	QMEAN score		Z-score	Overall quality (ERRAT score)	ResProx	Ramachandran plot analysis scores (%)		
		QMEAN4	QMEANDisCo				Favoured	Additionally allowed	Outlier region
1	WsMYB44	0.96	0.75±0.08	-4.67	100.00	1.82	94.5	5.5	0.00
2	1h8a (Crystal structure of ternary protein-DNA complex3)	1.33	0.72±0.05	-1.12	100.00	2.07	92.8	7.2	0.00
3	1h88 (Crystal structure of ternary protein-DNA complex1)	1.42	0.72±0.05	-1.75	97.75	3.03	89.7	10.3	0.00
4	1h89 (Crystal structure of ternary protein-DNA complex2)	0.93	0.72±0.05	-1.36	95.85	2.96	93.3	6.2	0.4

Table 3: Quantitative assessment of modelled WsMYB44 protein by VADAR scoring

S. no.	Protein name	Secondary structure			Hydrogen bonds		
		Helix	Beta	Coil	Mean H-bond distance	Mean H-bond energy	Residues with H-bond
1.	WsMYB44	64 (60%)	0 (0%)	41 (39%)	2.3 ± 0.4	-1.6 ± 1.1	88 (83%)
2.	1h89 (Crystal structure of ternary protein- DNA complex3)	174 (72%)	0 (0%)	66 (27%)	2.2 ± 0.3	-1.7 ± 1.0	214 (89%)
3.	1h89 (Crystal structure of ternary protein- DNA complex1)	206 (70%)	0 (0%)	87 (29%)	2.1 ± 0.3	-1.9 ± 1.1	257 (87%)
4.	1h89 (Crystal structure of ternary protein- DNA complex2)	182 (74%)	0 (0%)	61 (25%)	2.1 ± 0.3	-1.9 ± 1.0	221 (90%)

Detailed QMEAN and Z-scores are listed in Table 2. Furthermore, overall quality scoring in the form of Atomic bond interaction was performed by ERRAT webpage of SAVES server which depicted 100% score of our model that again showed more stability of our model than the selected templates. With RESPROX server, atomic resolution of the predicted model was also found to be superior to the templates.

Quantitative parameters of the computed model and templates were compared by VADAR tool which analyzed total accessible topology, stereo quality index, and residual volume. The outcomes suggested that our protein existed as helix type 64 (60%) and coiled type 41 (39%) secondary structure. However, in templates these values were significantly different (Table 3). The observed H- bond energy and distance were nearly similar for model and templates while, residues in H- bond were lesser (78%) and close to expected value (75%) than templates (87-90% with expected value of 75%). This result again attested the reliability of WsMYB44 protein over templates.

Analysis of Superimposition with Templates and Predicted Model

Structured proteins are three times more conserved than amino acid sequences. Thus, in spite of sequence similarity, templates and predicted model was superimposed on each other to analyze their topological features in detail (Fig. 7a). The outcomes of superimposition from Superpose server revealed less structural similarity of 1h8a and 1h88 with predicted WsMYB44 domain as compared to 1h89 template. These findings were further attested by observing local and global RMSD values at α carbon of model and templates. Lesser the RMSD, higher the structural similarity and values for 1h8a, 1h88, and 1h89 when superimposed on predicted model were 0.88 A°, 0.88 A° and 0.86 A° alpha carbon respectively, which depicted closer resemblance of 1h89 and model than rest two templates. Greater RMSD revealed synonymous and non-synonymous amino acid substitution which may have resulted in sequence dissimilarity at some extent. This fact was clearly proved by sequence alignment of our protein and templates by MultAlin server (Fig. 7b).

DNA-Protein and Protein-Protein Interaction Analysis

For DNA-protein interaction analysis, the most reliable model was chosen and docked with the MYB binding DNA element C/TAACG/TG. The respective DNA element was modeled by DNA sequence to structure conversion tool SCFBio, IIT Delhi (Arnott *et al*., 1974) (http://www.scfbio-iitd.res.in/software/drugdesign/bdna.jsp). Docking was performed by PatchDock server and the docked complex having lowest binding energy i.e. most negative score was considered to be stable complex. The interaction analysis revealed that WsMYB44 protein binds with TAACGT nucleotide target sequence through conserved SANT (R2 and R3) motifs (Fig. 8a). The key residues involved in interaction were Asn[43], Tyr[44], Phe[54], etc which were found to reside in conserved MYB domain.

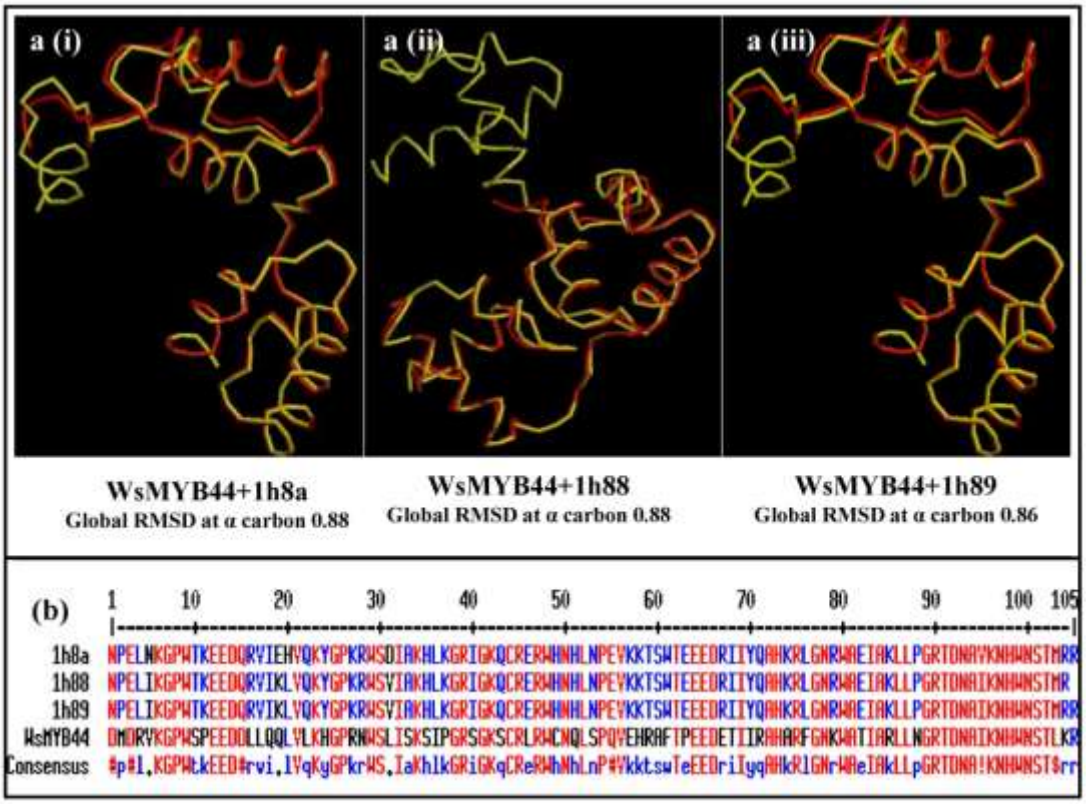

Fig. 7: (a) Results of superimposition showing structural conservation between WsMYB44 protein with respective templates (1h8a, 1h88, and 1h89) and (b) MultAlin results showing sequence similarities within predicted protein and templates

Since every transcription factor regulate the gene expression in synergistic way by interacting with other such TFs thus, we have analyzed protein-protein interaction networks of WsMYB44 protein by STRING online server. Interaction analysis at medium confidence interval by selecting total 10 interactomes revealed that WsMYB44 TF significantly interacted with *ARF19* (AUXIN RESPONSE FACTOR 19), CIPK9 (calcium sensor-interacting protein kinase), CDSP32 (CHLOROPLASTIC DROUGHT-INDUCED STRESS PROTEIN-32), RHYA1 (E3 ubiquitin-protein ligase) etc from its first shell (Fig. 8b). These results depicted that WsMYB44 protein involved in several defence, morphogenetic, and hormone signalling pathways of the *Withania somnifera*.

Analysis of Gene Ontology Enrichment and Subcellular Localization of Wsmyb44 Protein

For analysis of gene ontologies, WsMYB44 protein sequence was submitted in COFACTOR and CATH servers in order to predict its cellular, molecular, and biological functions and location in the cell. The obtained CATH vocabularies were further analyzed through ReviGO analysis by plotting scattered plots in the form of log clusters (Fig. 9).

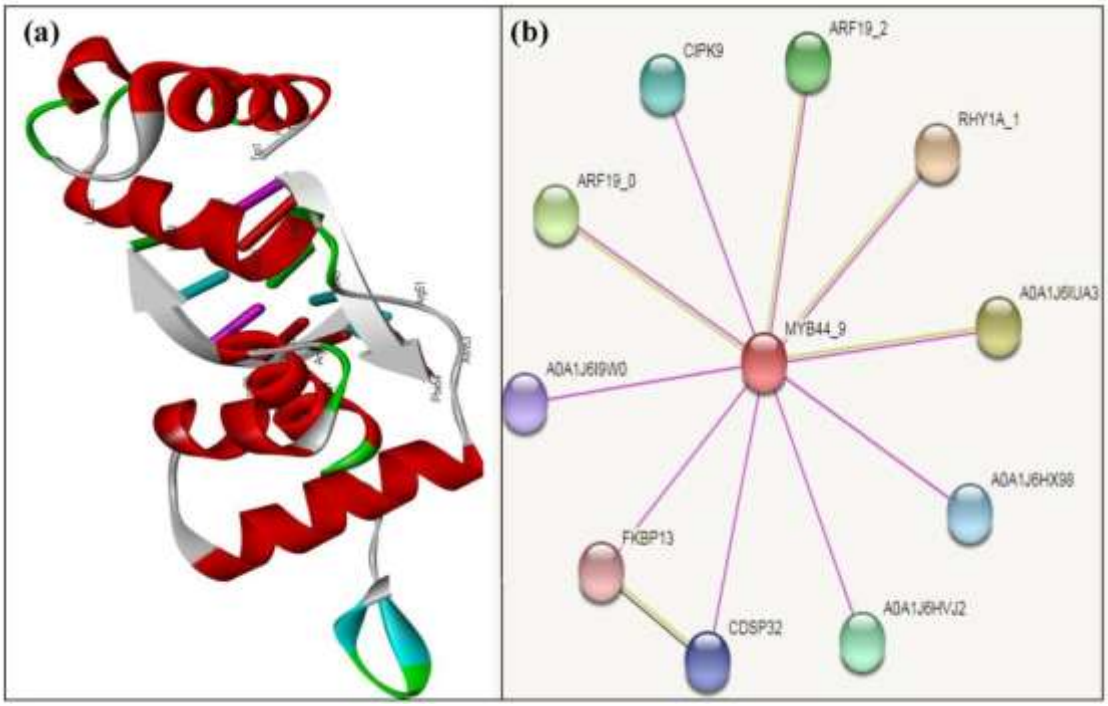

Fig. 8: Interaction analysis of WsMYB44 protein; (a) showing docked complex between protein and MYB binding element and (b) showing protein- protein interectomes of WsMYB44 protein

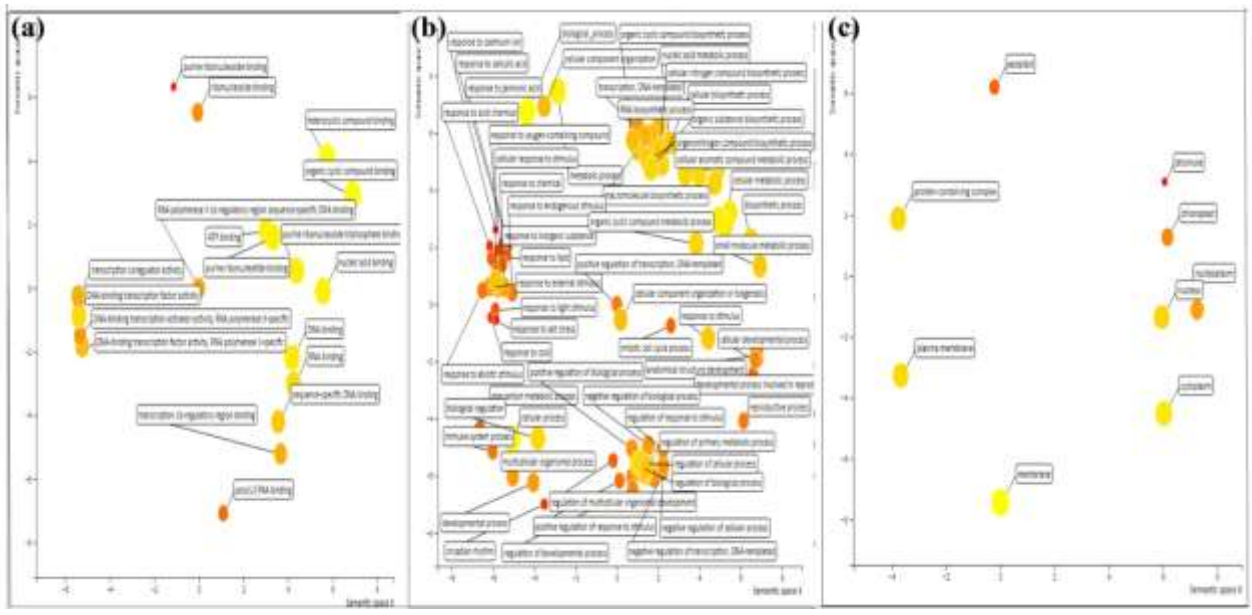

Fig. 9: Analysis of Gene Ontology with ReviGO tool. Functional GO terms having role in; (a) molecular, (b) biological, and (c) cellular components for WsMYB44 protein

The results showed involvement of our protein in various biological functions such as metabolic processes (GO:0008152), response to abiotic stress (GO:0050896), regulation of biological processes (GO:0050789), response to chemicals (GO:0042221) etc. Likewise, molecular functions were denoted by GO terms like transcription factor activities (GO:0003700), nucleic acid binding activities (GO:0003676), purine ribonucleoside binding activities (GO:0032555 and GO:0032550) etc. While cellular functionalities were represented by nucleus (GO:0005634) and cytoplasmic part (GO:004444). Additionally, Cello2GO server also depicted the biological role of predicted protein in response to stress and other biosynthetic processes (Fig. 10). This webserver also provided information about Subcellular localization of WsMYB44 protein in nucleus (69.1% probability with 3.45 score).

Comparative Analysis of Physicochemical Properties of Wsmyb44 Protein

Various physicochemical attributes of the modelled protein was evaluated by employing ExPaSy ProtParam tool (Table 4). The length of our protein was found to be slightly less than the templates and percentage contribution of total negative amino acids were also comparatively less. Likewise, molecular weight was least for WsMYB44 protein (12182.8 Da) while highest for the 1h88 (18308.9 Da). Next, theoretical pI was calculated in order to get information regarding nature of protein. In our observation we found that our protein and all the respective templates showed pI value more than 7.0 which revealed their basic nature. The predicted WsMYB44 protein and all the studied templates showed a negative GRAVY (Hydropathicity index) value which revealed the hydrophilic nature of the proteins (Table 4). Instability index depicted that protein and templates were unstable in nature.

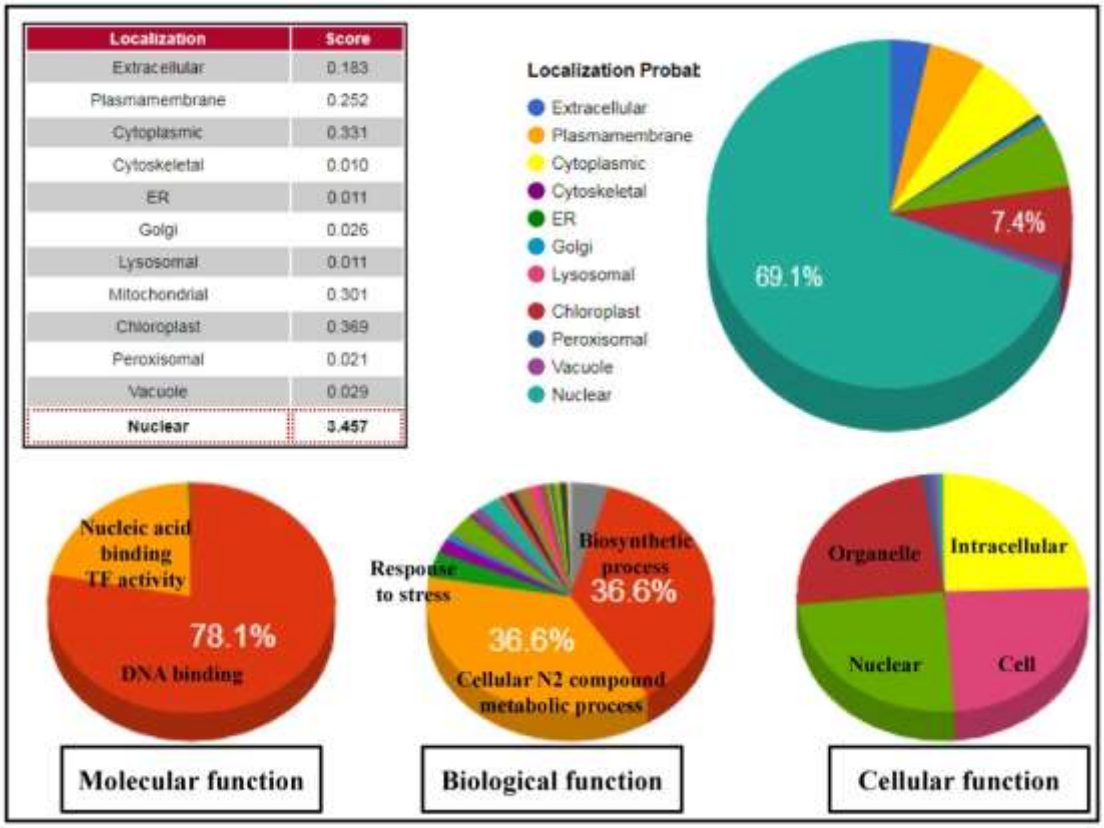

Fig. 10: Prediction of Subcellular localization and functional annotation of WsMYB44 protein employing Cello2GO web server

Table 4: Assessment of physicochemical properties of predicted WsMYB44 protein by ProtParam analysis

S. no.	Protein/ template	Amino acid residues			Molecular wt. (Dalton)	Theoretical pI	Hydropathicity index	Instability index
		Total	-Ve charge	+Ve charge				
1	WsMYB44	105	12	18	12182.8	9.91	-0.91	45.34
2	1h8a	105	14	23	12656.3	10.00	-1.43	61.09
3	1h88	152	20	33	18308.9	9.93	-1.24	42.13
4	1h89	115	12	26	13936.07	10.32	-1.22	51.65

Discussion

Transcription factors are the principal protein moiety which can regulate the expression of genes associated with several plant biological and molecular processes such as defence, response to external stimuli in the form of biotic and abiotic factors, primary and secondary metabolism, growth, development, reproduction, etc (Nath V *et al*., 2019). In *Withania somnifera*, there are plenty of information regarding identification and characterization of several TFs like AP2-ERF, WRKY, and MYC etc. However, to our best knowledge there is no record which can explain structural and functional insight of MYB TF in *W. somnifera* by applying computational approaches. Thus, this study may help to investigate identification phylogenetic assessment, binding domains and conserved motifs, physicochemical properties, and structural functionalities of MYB protein from *W. somnifera* (WsMYB44) by *in-silico* tools.

In the present study, we have identified a R2R3 type MYB protein TF (WsMYB44) from root tissue of *W. somnifera* and analyzed its phylogenetic relationship with homologous and paralogous sequences belonging to Solanaceae and other plant families. The bootstrap phylogenetic assessment depicted close relatedness of WsMYB44 protein from MYB44 TF belonging to *Nicotiana* species. Corroborating these findings, Tripathi *et al*., (2020) also reported close phylogenetic resemblances of AP2-ERF domain of *W. somnifera* from other members of Solanaceae family including *Nicotiana* sp. These outcomes depicted that this novel protein may have originated from its wild homolog sharing common ancestors, during the course of evolution. These outcomes were further validated by performing MSA of amino acid sequences of our protein and all homologs, which revealed the strong sequence conservation and high similarity around R2 and R3 domains of the MYB TF.

Every transcription factor has its unique identity in the form of conserved signature domains or amino acid patches which mainly act as DNA- binding domains. The MYB TF is a well-studied regulatory protein which is ubiquitous in all existing eukaryotes. This protein is known to harbour conserved SANT domains. In R2R3 type MYB proteins number of these domains are two. In this study we have also observed R2R3 type protein from *W. somnifera* in which the conserved domains were separated by unique patches of amino acids. R2 domain contained three helices each defined by a tryptophan residue while in R3 domain, first trp residue was replaced by another aromatic amino acid that is phenylalanine. Attesting this outcome, Ambawat *et al*., (2013) reported that this replacement might have taken place during the stress adaptation strategies in the plant. They have also reported variable nature of C- terminus of the protein which was also observed in the predicted protein. Like signature domains, conserved motifs also reflect the sequence conservation and protein identity. We have observed a set of conserved motifs residing in the N-terminus domain of the MYB protein while the C- terminus possessed motifs with relatively low sequence conservation. This observation can be attested according to Wang *et al*. (2015).

Protein in its tertiary nature (three dimensional structures) is believed to be more functional than primary and secondary conformation (Bajaj and Blundell, 1984). The functionality of the protein in its native nature can easily be studied for DNA-protein and protein-protein interactions. In this context, a 3D- model was predicted for WsMYB44 protein and validated by comparing it with already existing structurally resolved templates. Nearly in all parameters including Ramachandran plotting, VADAR values, ProSA Z-score, ERRAT scoring, ResProx calculations, PROCHECK, and QMEAN values, the predicted model was found to be superior over its structurally similar templates. The topological attributes of a protein signifies its functional characteristics by reveling the attachment sites and mode of action (Illergård *et al*. 2009). In this regard, we have performed superimposition of the predicted model over each of the structurally resolved template to reveal functional and structural constraints along with other topological details. Thus, structural proximity between crystal structure of the templates (1h8a, 1h88, and 1h89) and 3-D model of WsMYB44 protein was calculated and deciphered on the basis of RMSD (local and global) values. Results depicted closest structural similarity of our predicted protein with 1h89 as compared to rest of the two templates. These outcomes were devoid of any parameters which are based on sequence similarity which indicated the notion that during evolution proteins have shifted their sequential features but maintained the structural parameters (Panchenko and Madej 2005).

Gene regulatory role of the transcription factors can be studied by its interactions with major/minor grooves from nucleotide sequence of the target gene. The binding DNA elements for MYB transcription factor has been reported earlier by Choi *et al*. (2017) i.e. the core motif C/TAACG/TG, present on the major groove of target gene's promoter region. The interaction analysis was performed after docking the predicted protein with respective DNA element i.e. TAACTG and docked complex showed significant interaction which involved several amino acid residues, belonging to the R2 and R3 DNA binding domains. These results depicted the role of conserved domains of MYB TF in gene regulation and defence related signalling pathway cascades (Prouse and Campbell, 2012). Like interaction of TFs from promoter sequences of target gene, these proteins also perform their action after binding with several proteins including enzymes and other TFs (Alves *et al*., 2014). In this context, several interactors for WsMYB44 protein were hypothetically predicted by using STRING protein- protein interaction server. In this study, we have reported interactomes such as *ARF19* (auxin

related), CIPK9 (protein kinase), CDSP32 (draught related), and other proteins related to abiotic stress tolerance and ubiquitination. Similar to this, Chen *et al*. (2022) also observed these interacting entities from PmMYB7 protein of *Pinus massoniana*. The protein-protein interaction networks of particular TF can be well documented by studying its Gene Ontology (GO) vocabularies (Nguyen *et al*., 2011). In this regard, several GO terms were deciphered for cellular, biological, and molecular functionalities of WsMYB44 protein. For molecular functions the most significant term i.e. (GO:0003676) was observed which is dedicated to nucleic acid binding activities. Likewise, for biological functions, most significant ontologies were metabolic processes (GO:0008152), response to abiotic stress (GO:0050896), regulation of biological processes (GO:0050789), response to chemicals (GO:0042221) etc. Based on gene ontology, Subcellular localization of WsMYB44 (performed by Cello2GO) revealed that this protein mostly resides in the nuclear region and perform DNA binding activities therefrom. For instance, gene ontology and Subcellular location analysis were also linked from each other for TFs like WRKY3, WRKY4, and DREB1 of tomato plant (Aamir *et al*., 2017; Rai *et al*., 2019).

The molecular attributes of a protein (transcription factor) can be determined at fine tune level by studying its amino acid composition that is ultimately responsible for intrinsic reactivity of particular protein moiety. To fulfil this fact, we have studied several physicochemical properties such as molecular weight, pI values, hydropathic and hydrophilic index etc. of WsMYB44 protein along with crystal structures of templates. Almost in all parameters, the predicted model showed similarity with structurally resolved templates. This indicated the reliability of the modelled protein.

Conclusion

In-silico approaches very effective and convenient tool for computational based structural and functional characterization of transcription factors which can provide a platform for studying transcriptional reprogramming of genes involved in plant defence and morphogenesis. The present study was performed to study root derived MYB transcription factor from *Withania somnifera*. Phylogenetic and alignment results showed strong sequence conservation at R2 and R3 MYB domains, which were significantly involved in DNA and protein interaction with the help of active pockets. Additionally, predicted 3-D model showed good reliability score at every aspect, as compared to resolved templates. The Gene Ontology analysis revealed the active participation of predicted WsMYB44 TF in diverse cellular, molecular, and biological functions including nucleotide binding activities, role in abiotic stress tolerance, response to external stimulus, biosynthetic processes etc. Further, physicochemical properties of the protein revealed the basic nature of the WsMYB44 TF. Since there is dearth of knowledge regarding *in-silico* characterization of MYB protein in *W. somnifera*, this work may open new insights for deep study of TF regulated gene tailoring in medicinal plants.

Authors' Contribution

Lakee Sharma and Shashi Pandey-Rai have designed and conceptualized the whole manuscript. Bipin Maurya has analyzed the phylogenetic data.

Declaration of Competing Interest

All authors declare no any conflict of interest

Acknowledgement

All authors are thankful to CAS and DST-FIST facilities of Botany department, and IOE- incentive grant, Banaras Hindu University, Varanasi for providing all the necessary support. Lakee Sharma is highly grateful to CSIR, New Delhi for financial support as SRF fellow.

References

Aamir M, Singh VK, Meena M, Upadhyay RS, Gupta VK and Singh, S (2017) Structural and functional insights into WRKY3 and WRKY4 transcription factors to unravel the WRKY–DNA (W-Box) complex interaction in tomato (Solanum lycopersicum L.). A computational approach. *Front Plant Sci* **8**: 819. DOI: 10.3389/fpls.2017.00819

Alves, MS, Dadalto SP, Gonçalves AB, De Souza GB, Barros VA and Fietto LG (2014) Transcription factor functional protein-protein interactions in plant defense responses. *Proteomes* **2(1)**: 85-106. DOI: 10.3390/proteomes2010085

Ambawat S, Sharma P, Yadav NR and Yadav RC (2013) MYB transcription factor genes as regulators for plant responses: an overview. *Physiol Mol Biol Pla* **19(3)**: 307-321. DOI: 10.1007/s12298-013-0179-1

Arnott S, Chandrasekaran R and Marttila CM (1974) Structures for polyinosinic acid and polyguanylic acid. *Biochem J* **141(2)**:537-43. DOI: 10.1042/bj1410537.

Bailey TL, Williams N, Misleh C and Li WW (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res* **34**(suppl_2): W369-73. DOI: 10.1093/nar/gkl198

Bajaj M and Blundell T (1984) Evolution and the tertiary structure of proteins. *Annu Rev Biophys Bio* **13(1)**: 453-92. DOI: 10.1146/annurev.bb.13.060184.002321

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL and Studholme DJ (2004) The Pfam protein families database. *Nucleic Acids Res* **32**(suppl_1): D138-41. DOI: 10.1093/nar/gkt1223

Cao Y, Li K, Li Y, Zhao X and Wang L (2020) MYB transcription factors as regulators of secondary metabolism in plants. *Biology* **9(3)**:61. DOI: 10.3390/biology9030061

Castrignano T, De Meo PD, Cozzetto D, Talamo IG and Tramontano A (2006) The PMDB protein model database. *Nucleic Acids Res* **34**(suppl_1): D306-9. DOI: 10.1093/nar/gkj105

Chen P, Li R, Zhu L, Hao Q, Yao S, Liu J and Ji K (2022) Characterization and Interaction Analysis of the Secondary Cell Wall Synthesis-Related Transcription Factor PmMYB7 in Pinus massoniana Lamb. *Int J Mol Sci* **23(4)**: 2079. DOI: 10.3390/ijms23042079.

Choi JY, Abbai R, Kim YJ, Silva J, Rahimi S, Myagmarjav D, Chung IS, Kwon WS and Yang DC (2017) Molecular characterization of MYB transcription factor genes from Panax ginseng. *Russ J Plant Physl+* **64(3):** 398-409. DOI: 10.1134/S1021443717030050

Das S and Orengo CA (2016) Protein function annotation using protein domain family resources. *Methods* **93**:24-34. DOI: 10.1016/j.ymeth.2015.09.029

Dubos C, Stracke R, Grotewold E, Weisshaar B, Martin C and Lepiniec L (2010) MYB transcription factors in Arabidopsis. *Trends Plant Sci* **15(10)**: 573-581. DOI: 10.1016/j.tplants.2010.06.005

Feng C, Andreasson E, Maslak A, Mock HP, Mattsson O and Mundy J (2004) Arabidopsis MYB68 in development and responses to environmental cues. *Plant Sci* **167(5)**: 1099-1107. DOI: 10.1016/j.plantsci.2004.06.014

Hofmann K, Bucher P, Falquet L and Bairoch A (1999) The PROSITE database, its status in 2002. *Nucleic Acids Res* **27(1)**: 215-219. DOI: 10.1093/nar/30.1.235

Illergård K, Ardell DH and Elofsson A (2009) Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. *Proteins: Struct. Funct. Bioinform.* **77(3)**: 499-508. DOI: 10.1002/prot.22458

Kyte J and Doolittle RF (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157(1)**:105-32. DOI: 10.1016/0022-2836(82)90515-0

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R and Thompson JD (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* **23(21)**: 2947-2948. DOI: 10.1093/bioinformatics/btm404

Misra L, Mishra P, Pandey A, Sangwan RS, Sangwan NS and Tuli R (2008) Withanolides from Withania somnifera roots. *Phytochemistry* **69(4)**:1000-1004. DOI: 10.1016/j.phytochem.2007.10.024

Mu RL, Cao YR, Liu YF, Lei G, Zou HF, Liao Y, Wang HW, Zhang WK, Ma B, Du JZ and Yuan M (2009) An R2R3-type transcription factor gene AtMYB59 regulates root growth and cell cycle progression in Arabidopsis. *Cell Res* **(11)**:1291-1304. DOI: 10.1038/cr.2009.83

Nguyen CD, Gardiner KJ and Cios KJ (2011) Protein annotation from protein interaction networks and gene ontology. *J Biomed Inform* **44(5)**: 824-829. DOI: 10.1016/j.jbi.2011.04.010.

Panchenko AR and Madej T (2005) Structural similarity of loops in protein families: toward the understanding of protein evolution. *BMC Evol Biol* **5(1)**: 1-8. DOI: 10.1186/1471-2148-5-10

Paz-Ares J, Ghosal D, Wienand U, Peterson PA and Saedler H (1987) The regulatory c1 locus of Zea mays encodes a protein with homology to myb proto-oncogene products and with structural similarities to transcriptional activators. *The EMBO journal* **6(12)**: 3553-5358. DOI: 10.1002/j.1460-2075.1987.tb02684.x

Persak H and Pitzschke A (2014) Dominant repression by Arabidopsis transcription factor MYB44 causes oxidative damage and hypersensitivity to abiotic stress. *Int J Mol Sci* **15(2):** 2517-2537. DOI: 10.3390/ijms15022517

Ponting CP, Schultz J, Milpetz F and Bork P (1999) SMART: identification and annotation of domains from signalling and extracellular protein sequences. *Nucleic Acids Res* **27(1)**: 229-232. DOI: 10.1093/nar/27.1.229

Prouse MB and Campbell MM (2012) The interaction between MYB proteins and their target DNA binding sites. *(BBA)-Gene Regul Mech* **1819(1)**:67-77. DOI: 10.1016/j.bbagrm.2011.10.010

Rai KK, Rai N and Rai SP (2019) Prediction and validation of DREB transcription factors for salt tolerance in Solanum lycopersicum L.: An integrated experimental and computational approach. *Environ Exp Bot* **165**: 1-8. DOI: 10.1016/j.envexpbot.2019.05.015

Romero I, Fuertes A, Benito MJ, Malpica JM, Leyva A and Paz-Ares J (1998) More than 80R2R3-MYB regulatory genes in the genome of Arabidopsis thaliana. *The Plant journal: for cell and molecular biology* **14(3)**: 273-284. DOI: 10.1046/j.1365-313x.1998.00113.x

Roy A, Yang J and Zhang Y (2012) COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Res* **40(W1)**: W471-7. DOI: 10.1093/nar/gks372

Schneidman-Duhovny D, Inbar Y, Nussinov R and Wolfson HJ (2005) PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res* **33(suppl_2)**: W363-7. DOI: 10.1093/nar/gki481

Sukumari Nath V, Kumar Mishra A, Kumar A, Matoušek J and Jakše J (2019) Revisiting the role of transcription factors in coordinating the defense response against citrus bark cracking viroid infection in commercial hop (Humulus Lupulus L.). *Viruses*. **11(5)**:419. DOI: 10.3390/v11050419

Supek F, Bošnjak M, Škunca N and Šmuc T (2011) REVIGO Summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* **6:e21800**. DOI: 10.1371/journal.pone.0021800

Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP and Kuhn M (2015) STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* **43(D1)**:D447-52. DOI: 10.1093/nar/gku1003

Tamura K, Stecher G, Peterson D, Filipski A and Kumar S (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol*. **30(12)**:2725-9. DOI: 10.1093/molbev/mst197

Tripathi S, Srivastava Y, Sangwan RS and Sangwan NS (2020) In silico mining and functional analysis of AP2/ERF gene in Withania somnifera. *Sci Rep-UK* **10(1)**:1-2. DOI: 10.1038/s41598-020-60090-7

Wang F, Suo Y, Wei H, Li M, Xie C, Wang L, Chen X and Zhang Z (2015) Identification and characterization of 40 isolated Rehmannia glutinosa MYB family genes and their expression profiles in response to shading and continuous cropping. *Int J Mol Sci* **16(7)**:15009-30. DOI: 10.3390/ijms160715009

Willard L, Ranjan A, Zhang H, Monzavi H, Boyko RF and Sykes BD and Wishart DS (2003) VADAR: a web server for quantitative evaluation of protein structure quality. *Nucleic Acids Res* **31(13)**:3316-9. DOI: 10.1093/nar/gkg565

Xu D and Zhang Y (2011) Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. *Biophys J* **101(10)**:2525-34. DOI: 10.1016/j.bpj.2011.10.024

Yu CS, Cheng CW, Su WC, Chang KC, Huang SW, Hwang JK and Lu CH (2014) CELLO2GO: a web server for protein subCELlular LOcalization prediction with functional gene ontology annotation. *PloS ONE* **9(6)**:e99368. DOI: 10.1371/journal.pone.0099368

Zheng XW, Yi DX, Shao LH and Cong LI (2017) In silico genome-wide identification, phylogeny and expression analysis of the R2R3-MYB gene family in Medicago truncatula. *Journal Integr Agr* 16(7):1576-91. **DOI:** 10.1016/S2095-3119(16)61521-6