Indian Journal of Agricultural Sciences **91** (7): 1096–9, July 2021/Short Communication
https://doi.org/10.56093/ijas.v91i7.115140

Functional annotation of hypothetical proteins involved in multiple stress response in *Oryza sativa*

PRATHEEK H P[1], SHIVANI NAGAR[1], SANDEEP ADAVI[1], VISWANATHAN CHINNUSAMY[1] and RAKESH PANDEY[1]*

ICAR-Indian Agricultural Research Institute, New Delhi 110 012, India

Received: 12 January 2021; Accepted: 27 January 2021

Keywords: Abiotic stress, Hypothetical proteins, Multiple stress, Rice

Abiotic stress is one major global problem limiting crop growth and productivity (Wang *et al.* 2003). Stresses like drought, salinity, high light intensity and heavy metals are the primary causes of crop loss, particularly in cereals (Altman *et al.* 2003). Rice (*Oryza sativa* L.) is the world's most important staple food crop and will continue to be so in the coming decades. Understanding the connection between the initial stress response of plants and other downstream events to adjust to the altered conditions is one of the grand challenges in plant biology (Zhu *et al.* 2016). Intensive research over the last decade has gradually unraveled the mechanisms that underlie how plants cope with abiotic stresses (Singh *et al.* 2012)**,** but many aspects remain unresolved. The complete understanding of physiological, biochemical and molecular responses and identification of potential unknown stress-responsive pathways and genes in abiotic plant stress tolerance will contribute a better understanding of underlying molecular mechanisms (Chauhan *et al.* 2016). Discoveries of novel genes and pathways, analysis of expression patterns and the determination of function of genes during abiotic stress adaptation will provide the basis for effective engineering strategies to enhance abiotic stress tolerance of the crop plants (Onaga *et al.* 2016).

Computational approaches to this problem along with genomic approaches look more holistic and stress-related transcription factors are an essential candidate for crop improvements as they play regulatory roles in more significant way (Parreira *et al.* 2016). In this study, multi stress responsive genes from *Oryza* sp. were retrieved from STIFDB V2.0 (Naika *et al.* 2013 and Ambika *et al.* 2008) and analysis of those genes were carried out at Division of Plant Physiology, IARI, New Delhi during 2018–19 using different bioinformatics tools. Attempts have also been made to annotate multiple stress responsive hypothetical proteins from the database using multiple bioinformatics approaches. In addition to experimental results, our study provides a strong methodology and toolkit to annotate multiple stress responsive hypothetical proteins in plants. The new knowledge acquired through this research will help apply stress responsive determinants and in the engineering of plants with enhanced tolerance to abiotic stresses.

Retrieval of multi stress responsive genes from STIFDB2: The 1000 bp upstream sequence was retrieved from STIFDB2 (http://caps.ncbs.res.in/stifdb2/), the corresponding genes and protein sequences were obtained from RAP-DB (https://rapdb.dna.affrc.go.jp/)and NCBI (https://www.ncbi.nlm.nih.gov/), respectively. Further, the gene ontology analysis was carried out at Gene Ontology Consortium to know the ontology annotations of multi stress responsive genes.

Selection of candidate genes for annotation: Further, for the selected genes, NCBI/RAP-DB annotations were retrieved through a simple BLAST search. Fifty five proteins out of 66 were annotated, and 15 were un-annotated. The unannotated multi stress responsive genes were selected for further annotation and analysis (Fig 2).

Annotation of hypothetical proteins using the developed pipeline with multiple approaches: The 15 un-annotated protein sequences were submitted to various tools and searched against target databases to find the possible annotations. The procedure was as follows;

With the protein sequences, BLASTp search was made against the NCBI-nr database (https://blast.ncbi.nlm.nih.gov/Blast), and the resulting hits were considered based on the E-value, percent coverage and identity (%). For all the protein sequences PSI-BLST (https://www.ebi.ac.uk/Tools/sss/psiblast/) was performed at CAPS-IWS2(http://caps.ncbs.res.in/iws2/). PSI-BLAST iteratively search for remote homologs and the iteration cutoff was 3. PSI-BLAST helps more than BLASTp in finding remote homologs as it iteratively searches for homologs using a profile. The search was performed against the Swissprot database (https://

Present address: [1]ICAR-Indian Agricultural Research Institute, New Delhi. *Corresponding author email: r_pan_pdcsr@yahoo.co.in.

www.uniprot.org/statistics/Swiss-Prot) with expectation value of 10, e-value threshold for including the multipass model was 0.002. Other parameters were kept default while searching. Also, the PSI-BLAST was performed against the nr database at NCBI using NCB-PSI-BLAST.

HMMER scan against Pfam database was performed with default parameters. Since, HMMER uses the sequence to HMM comparison approach, there are more possibilities of finding remote homologs. HHblits (https://toolkit.tuebingen.mpg.de/tools/hhblits), is a more sensitive remote homology search tool which finds hits using HMM to HMM based comparison. The search was done at MPI-Bioinformatics Toolkit (https://toolkit.tuebingen.mpg.de/) with an E-value inclusion threshold of 1e-3 with three iterations. The minimum probability of the hits in the list was 20% and the number of target sequence hits was restricted to 250.

TMHMM (http://www.cbs.dtu.dk/services/TMHMM/), PHOBIUS (https://phobius.sbc.su.se/), MEMSAT-SVM (https://bio.tools/memsat-svm), MEMPAC (https://bio.tools/) were used for the membrane protein and topology predictions with default parameters. Signal P search for finding sub cellular localization was made through InterProScan (https://www.ebi.ac.uk/interpro/search/sequence/). DeepLoc search (http://www.cbs.dtu.dk/services/DeepLoc/) was made using profiles and Cello2go search (http://cello.life.nctu.edu.tw/cello2go/) was performed at an E-value threshold of 0.001.

For sequence based fold prediction, CDD search (https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) was performed against NCBI database and Pfam (http://pfam.xfam.org/) search was performed against Pfam database with an E-value threshold of 0.01. Full length protein sequences were submitted to phyre2 for the fold prediction. UNIPROT database was searched using HHblits tool to find remote homologs of the protein. Databases such as PANTHER (http://www.pantherdb.org/), CATH-Gene3D (https://www.cathdb.info/), SUPERFAMILY (http://supfam.org/), PROSITE Profile (https://prosite.expasy.org/), SMART (http://smart.embl-heidelberg.de/) were searched for different types of homologs (sequence and structure) using InterPro-Scan. GenDiS provides a platform for cross genome comparison at the superfamily level. Query sequences were submitted to stand alone version of the tool.

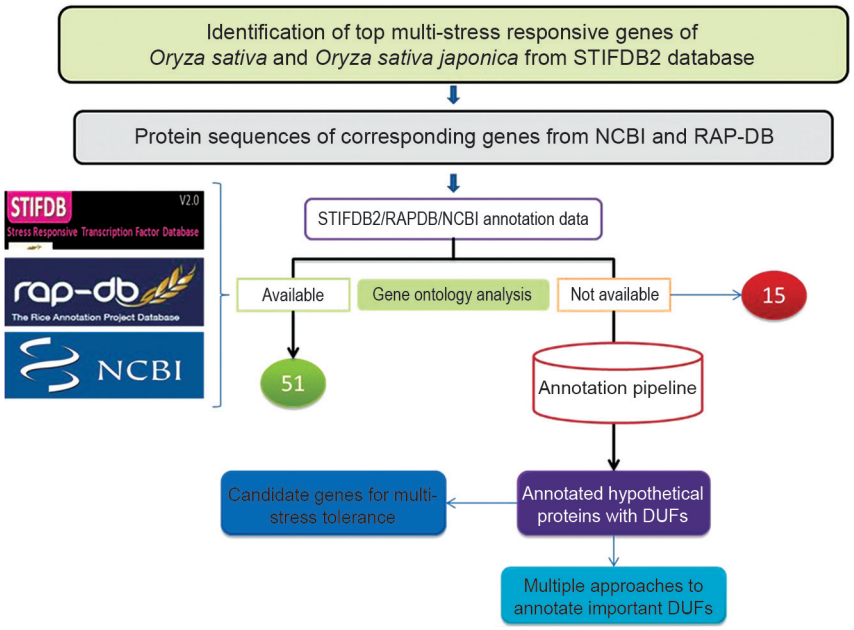

Fig 1 Representation of workflow for the project 66 multiple stress responsive proteins were identified and annotated using existing databases. Further, 15 unannotated proteins from the initial pipeline were subjected to an advanced annotation pipeline.

Sl.No.	Hypothetical Protein Accession IDs	Annotation
1	XP_015626015	Protein of unknown function (DUF1645)
2	XP_015622949	Senescence regulator (Complex I intermediate-associated protein 30)
3	XP_015624026	Protein of unknown function (DUF1645)/ OsSGL
4	XP_015627047	Transcriptional regulator (ICP4) like
5	XP_015628852	Copper chaperone for superoxide dismutase
6	XP_015627952	Ubiquitin-like modifier-activating enzyme
7	XP_015629125	Glucocorticoid receptor-like (DNA-binding domain)
8	XP_015632561	Nucleotide-diphospho-sugar transferase
9	XP_015635059	Copper chaperone for superoxide dismutase
10	XP_015642997	Floral defensin-like protein
11	XP_015611148	F-box domain containing ligase like
12	XP_015610640	F-box domain containing ligase like
13	XP_015611382	Protein of unknown function (DUF4228) with probable nucleotide binding activity
14	XP_015611637	Ubiquilin like
15	XP_015641421	Lipocalin like

Confidence	Supporting approaches	No of proteins annotated
	1	03
	2	01
	3	02
	4	03
	5	03
	6	01
	7 or more	02

Legend

Fig 2 List of multiple stress responsive hypothetical proteins in rice with annotation, intensity of color represents the confidence of annotation, number of approaches supporting the annotation and proteins annotated are depicted.

PURE (in-house database) was run for all the sequences against non-redundant database with an E-value threshold of 0.01, CD-HIT Threshold Value 0.7, HMMPFAM - E-Value 0.01 and HMMSEARCH - Cut-off of 50. Other values were considered default.

All 15 protein sequences were submitted to STRING (https://string-db.org/) online server in FASTA format to predict protein-protein interactions. For analysis of DUF1645 RiceFREND tool (https://ricefrend.dna.affrc.go.jp/) was used which uses correlation expression data to interpret the interactions. Protein structure prediction was done by using I-Tasser (https://zhanglab.ccmb.med.umich.edu/I-TASSER/)) by submitting the sequence at online I-Tasser server. DomSerf (http://bioinf.cs.ucl.ac.uk/web_servers/psipred_server/psipred_help/) was used for predicting the domain structure and Bioserf (https://github.com/psipred/bioserf) was used for complete structure prediction. The sequences were submitted at PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred/) server for BioSerf analysis. Complete amino acid sequences of all the 15 proteins were submitted to 'ELM-Functional Sites in Protein' predictor server for prediction of functional motifs in proteins. Full length nucleotide sequences were submitted to PUFAS for automated annotation and also to predict the fold.

Gene ontology analysis suggests large group of protein of unknown function: The results obtained at gene ontology database suggested that 75% of proteins out of 66 submitted did not belong to any protein class, only 49% of the proteins molecular function was known and 67% of proteins cellular component was predicted. The molecular function analysis showed that 27% of the proteins were involved in catalytic activities followed by 13% of proteins with binding activity. Biological process analysis showed 19% of the proteins were involved in cellular processes followed by 18% of proteins which are involved in metabolic processes. Cellular component of 67% of the proteins submitted was unknown and rest of them belongs mostly to cell parts, i.e. 24% of the submitted proteins. This gave an important clue about the need of annotation of the hypothetical proteins.

Annotation of hypothetical proteins through multiple approaches: Annotation of hypothetical proteins using multiple approaches ranging from sequence comparison to phylogenetic analysis resulted in annotation of all hypothetical proteins except those which had DUFs (Domains of Unknown Function). Out of 15 proteins, two were annotated with high confidence and three proteins were annotated with medium confidence as Chaperones for copper oxide dismutase (Fig 2).

A case study for annotation pipeline with Os03g0152000: Sequence search methods suggested that the protein is made up of 348 amino acids and searches through Interproscan, CDD, Pfam showed the presence of two Heavy Metal Associated Domains (HMA). BLASTp analysis showed 95% sequence coverage and 65% identity with Heavy Metal Associated domain and PSI-BLAST analysis gave hits with Chaperone for copper transporting ATPase 1 at an e-value of 3e-04. Further the InterproScan predicted the presence of two HMA domains in the protein spanning from 20-73 and 134-194 amino acids. A scan with the PURE tool gave results which was supportive to Interpro results that the protein has two heavy metal binding domains. HMA domain is a conserved protein domain found in a number of heavy metal transport or detoxification proteins, therefore might play important role in multiple abiotic stress response. The soluble nature as well as nuclear localization of the proteins were predicted with Cello2Go (58.3% probability of nuclear localization) and Deeplock (0.69 nuclear localization score). Protein–protein interaction study using STRING database suggested it could be a protein of HMA domain and involved in ion transport, detoxification because it was co-expressing with most of the HMA domain containing proteins.

Phyre2 fold recognition method gave hits for an antiviral protein (resistance protein pikp-1) with 99% confidence and 59% identity and the query protein was aligning with the HMA domain of the protein. The 70 residues (59% of query sequence) were modelled with 99.9% confidence by the single highest scoring template. DomTHREADER gave hit with an e-value of 3e-06 as superoxide dismutase copper chaperone and the results were also supported by the results of pGenThreader with high confidence with an e-value of 5e-04 as copper-exporting P-type ATPase A which belongs to CATH superfamily 3.30.70.100. So, it was well predicted that the query protein could be a heavy metal associated chaperone kind of protein. Further a query with PUFAS- the automated server for annotation of unknown proteins gave the results as heavy metal associated domain containing protein and also the fold prediction by PUFAS suggested that it has Copper Transporting P-type ATPase fold.

There has been a lot of data available on stress responsive genes on public databases, but it is the need of the hour to annotate the proteins with unknown functions to get a complete understanding of the genes involved in stress and also in multiple stress responses. First part of this study gave a very good validation for the annotation pipeline developed for annotations of any hypothetical proteins. Secondly, we tried to annotate one of the hypothetical proteins. With various approaches, it is predicted that the query protein might be a copper chaperone for superoxide dismutase. Copper chaperone for superoxide dismutase specifically delivers Cu to copper/zinc superoxide dismutase and may activate copper/zinc superoxide dismutase through direct insertion of the Cu cofactor. Though the computational study provides the idea about the particular protein, validation of these results with wet lab experiment will further certify the results.

SUMMARY

Plants being sessile in nature face multiple stresses during its growth and development. Understanding the dynamic tolerance mechanism of plants to changing environment is the need of the hour and developing climate smart crops is still a nightmare. Functional characterization of relevant genes is a prerequisite when identifying

candidates for crop improvement, in this study we identified 66 multiple stress responsive genes in rice that could serve as potential candidates for breeding for stress tolerance. The 15 hypothetical genes rendering multiple stress tolerance were functionally annotated using a rich set of tools ranging from sequence similarity search to fold recognition methods. It is predicted that the query protein might be a copper chaperone for superoxide dismutase. A novel approach to annotate hypothetical proteins with different remote homology search and structure-based threading approaches was designed. In addition to identified multiple stress responsive proteins in rice (*Oryza sativa* L.), our study also provides a robust toolkit to annotate other hypothetical proteins in plants.

ACKNOWLEDGEMENTS

Authors acknowledge the support provided by ICAR-Indian Agricultural Research Institute, New Delhi and ICAR New Delhi for fellowship.

REFERENCES

Altman A 2003. From plant tissue culture to biotechnology: scientific revolutions, abiotic stress tolerance, and forestry. *In Vitro Cellular & Developmental Biology-Plant* **39**(2): 75–84.

Ambika S, Varghese S M, Shameer K, Udayakumar M and Sowdhamini R. 2008. STIF: hidden Markov model-based search algorithm for the recognition of binding sites of stress upregulated transcription factors and genes in Arabidopsis thaliana. *Bioinformation* **2**(10): 431–37.

Chauhan N and Gupta K 2016. Mechanisms of abiotic stress responses and tolerance in plants physiological intervention. *Research & Reviews: A Journal of Life Sciences* **6**(1): 19–26.

Naika M, Shameer K, Mathew O K, Gowda R and Sowdhamini R. 2013. STIFDB2: an updated version of plant stress-responsive transcription factor database with additional stress signals, stress-responsive transcription factor binding sites and stress-responsive genes in Arabidopsis and rice. *Plant and Cell Physiology* **54**(2): e8-e8.

Onaga G and Wydra K 2016. Advances in plant tolerance to abiotic stresses. *Plant Genomics* 229–72.

Parreira J R, Branco D, Almeida A M, Czubacka A, Agacka-Mołdoch M, Paiva J A and de Sousa Araújo S. 2016. Systems biology approaches to improve drought stress tolerance in plants: state of the art and future challenges. *Drought Stress Tolerance in Plants* **2**: 433–71. Springer, Cham.

Pathak P S 2000. Agro forestry: A tool for arresting land degradation. *Indian Farming* **49**(11): 15–19.

Singh C M, Binod K, Suhel M and Kunj C. 2012. Effect of drought stress in rice: a review on morphological and physiological characteristics. *Trends in Biosciences* **5**(4): 261–65.

Wang W, Vinocur B and Altman A 2003. Plant responses to drought, salinity and extreme temperatures: towards genetic engineering for stress tolerance. *Planta* **218**(1): 1–14.

Zhu J K. 2016. Abiotic stress signaling and responses in plants. *Cell* **167**(2): 313–24.