

Declaration

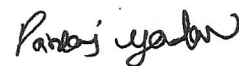
I hereby declare that the work presented in this Thesis titled *In silico studies of stress tolerance genes in non-desert and desert plants* submitted to the Indian Institute of Technology Jodhpur in partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy is a bonafide record of the research work carried out under the supervision of Dr. Pankaj Yadav and Dr. Ayan Sadhukhan. The contents of this thesis in full or in parts have not been submitted to and will not be submitted by me to any other Institute or University in India or abroad for the award of any degree or diploma.



Vedika Dhiman
(P19BB208)

Certificate

This is to certify that the thesis titled *In silico studies of stress tolerance genes in non-desert and desert plants*, submitted by *Vedika Dhiman (P19BB208)* to the Indian Institute of Technology Jodhpur for the award of the degree of *Doctor of Philosophy* is a bonafide record of the research work done by her under our supervision. To the best of our knowledge, the contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.



Dr. Pankaj Yadav
PhD. Thesis Supervisor



Dr. Ayan Sadhukhan
PhD. Co-supervisor

Abstract

The global population is expected to exceed 10 billion by 2050, posing a significant challenge to the agriculture sector in fulfilling increased demands of food production. This necessitates the adoption of sustainable strategies to ensure global food security. Among staple food crops, *Oryza sativa* (rice) is a cornerstone of global food production, serving as a dietary staple for nearly half of the world's population. However, its productivity is severely constrained by biotic and abiotic stresses. Biotic stress includes pathogenic bacteria, fungi, and other organisms, while abiotic stresses such as drought, and heat can result in yield losses of up to 90-100%. To mitigate these impacts, *O. sativa* possesses an intrinsic defense mechanism involving stress-responsive genes and proteins. Despite this, current interventions, including the use of pesticides and conventional breeding techniques, often provide only temporary relief. In response, the integration of computational tools with advanced breeding strategies has emerged as a promising avenue to enhance the resilience and productivity of *O. sativa* under stress conditions. This holistic approach has the potential to address the limitations of traditional methods and contribute to sustainable food production systems.

This thesis addresses the challenges posed by biotic and abiotic stresses in *O. sativa* by focusing on the identification, characterization, and analysis of stress-tolerant genes and proteins through computational approaches. The study has three primary objectives. First, it aims to identify novel disease-resistance proteins in *O. sativa Japonica* through *in silico* methods, targeting bacterial leaf blight (BB) and rice blast (RB) diseases, which are major biotic stressors affecting productivity of *O. sativa*. Second, it seeks to predict disease-resistance proteins in *O. sativa* and its related species by evaluating and comparing the performance of various deep learning (DL) and machine learning (ML) algorithms, thereby identifying the most effective computational frameworks for accurate protein prediction. Lastly, the research involves drought-responsive transcriptomic and genome-wide studies of *Prosopis cineraria*, a drought-resilient leguminous tree from the Indian Thar Desert, to uncover drought-tolerance genes that could potentially enhance drought resilience in *O. sativa*. Together, these objectives aim to contribute to sustainable strategies for improving the stress resilience of *O. sativa*.

In line with the objectives outlined above, five novel disease-resistance proteins were effective against BB and RB diseases in *O. sativa* were identified through a comprehensive *in silico* approach. These proteins were characterized using gene network construction, structural modeling, functional annotation, and phylogenetic analysis. The identified proteins exhibited key roles in disease-resistance mechanisms and showed evolutionary relationships with well-established resistance proteins. Furthermore, gene expression profiling revealed their differential expression under infections by *Xanthomonas oryzae* pv. *oryzae* (Xoo) and *Magnaporthe oryzae*, the causative agents of BB and RB diseases, respectively. These findings offer valuable insights for the development of disease-resistant varieties of *O. sativa*.

To extend the scope of disease-resistance protein identification in *O. sativa* and related species, various ML and DL models were employed. These models were rigorously trained, tested, evaluated, cross-validated, and compared for predictive performance. Among these, the DL-based Multi-Layer Perceptron model outperformed others, demonstrating precise prediction of disease-resistance proteins. The study thus provides computational frameworks and candidate proteins that can aid in breeding novel disease-resistant varieties of *O. sativa*.

In addition, drought-tolerance genes from *P. cineraria*, a resilient tree species from the Indian Thar Desert, were explored to identify robust orthologs that could enhance drought resilience in *O. sativa*. Transcriptomic analyses under drought, highlighted the APETALA2/Ethylene Responsive Factor (AP2/ERF) superfamily as key genes induced under drought, compelling us to conduct a genome-wide analysis of

this superfamily in *P. cineraria*. Comparative studies revealed significant copy number variations (CNVs) of *AP2/ERF* genes between drought-tolerant (DT) and drought-sensitive (DS) species, such as *O. sativa Japonica*, *Arabidopsis thaliana*, and *Pisum sativum*. Structural analyses of *P. cineraria* AP2/ERF proteins indicated stronger interactions between the DNA-binding AP2 domain of *P. cineraria* and the target *cis* element GCC box/dehydration-responsive element (DRE), as compared to orthologs from DS species. Additional investigations into amino acid variations and hydrogen bonding within AP2/ERF proteins among *O. sativa Japonica*, *O. sativa Indica*, and *P. cineraria* demonstrated substantial differences. Notably, DT *P. cineraria* AP2/ERFs contained polar amino acids at the variation positions and more hydrogen bonds compared to the orthologs from DS species *O. sativa* and *A. thaliana*, suggesting stronger DNA binding by AP2/ERF orthologs of DT species.

In conclusion, this thesis presents a comprehensive *in silico* investigation of stress-tolerant genes and proteins in *O. sativa*, providing breeders with novel genetic resources for developing stress-resilient varieties of *O. sativa*. Overall, these findings contribute to enhancing the productivity and resilience of *O. sativa*, thereby advancing global food security efforts.

Acknowledgments

A PhD is a journey, not just a thesis. Completing this PhD journey has been an incredible experience, and I owe my deepest gratitude to many people who have supported and supervised me throughout this endeavour.

First and foremost, I would like to express my sincere gratitude to my supervisor Dr. **Pankaj Yadav** and co-supervisor Dr. **Ayan Sadhukhan** for their invaluable guidance, endless support, continuous motivation, training, and teaching during the entire course of my PhD journey. Their valuable advice and insightful conversations made my research enjoyable. Their great passion, sincere commitment, optimism, and thirst for education were inspiring during my PhD tenure. I thank them for providing me with the opportunity to carry out my thesis work in the Statistical Multiomics Research Group and introducing me to the world of computational biology and bioinformatics and machine and deep learning in plant sciences. The thesis would not have been completed without their vision and technical support. Words cannot fully convey my gratitude for their invaluable efforts in enhancing my technical expertise and motivating me to grow into an independent, disciplined, and innovative researcher. They were a guide and guardian throughout my PhD and I feel proud to say that I am their student. I would also like to thank the members of my committee members, Dr. Rohan Diliprao Erande, Dr. Priyanka Singh, and Dr. Neha Jain, for their invaluable feedback and constructive criticism. Their insights have been instrumental in helping me refine my research questions and strengthen my arguments. I am also thankful to the staff members of the Department of Bioscience and Bioengineering for their logistic support on various occasions and for making all possible facilities available to me. I am indebted to IIT Jodhpur and the Ministry of Education for providing financial support that made my research possible. The resources, facilities, and opportunities provided by IIT Jodhpur were also instrumental in the success of my research.

I am thankful to my lab members, Jyoti Sharma, Pragya Kashyap, Rajveer Singh Shekhawat, Ashish Kumar Swain and Monika Sharma for creating such a pleasant atmosphere for research and for providing me with unconditional support on both academic and personal levels. I am truly grateful to my friends Jyoti Sharma and Pragya Kashyap for their unwavering support, both academically and personally. Your presence brings me strength and happiness in every up and down in my PhD journey.

I want to take this opportunity to express my gratitude to my beloved grandparents, parents and little sister for being the most substantial support not only during my PhD career but also throughout my life. I am deeply thankful to my late grandparents and father for always inspiring and supporting me to get out of my comfort zone and work honestly towards my goals. I sincerely acknowledge the hard work and sacrifices my mother made for the smooth completion of this thesis work. I am incredibly blessed to have a family which has always supported and blessed me in all my endeavours. I would also like to express my deepest gratitude to my soon-to-be fiancé, Shreshtha Saxena, whose unwavering love, support, and encouragement have been a constant source of strength throughout this journey. Thank you for standing by me through every challenge and reminding me of the light at the end of the tunnel. This accomplishment would not have been possible without your support and I am grateful to have you by my side in my PhD journey.

A significant part of my learning over the years has been shaped by the support and guidance of my friends and well-wishers, both within and beyond academics. I extend my heartfelt gratitude to all of them. I express all my gratitude towards Goddess Saraswati, without whose blessings whatever I have done or will do is not imaginable.

Vedikaa Dhiman

Contents

	page
Abstract	v
Acknowledgements	vii
Contents	ix
List of Figures	xi
List of Tables	xv
List of Symbols	xvii
List of Abbreviations	xix
1 Introduction	1
1.1 Agriculture and sustainable developmental goals	1
1.2 Role of <i>Oryza sativa</i> in achieving sustainable developmental goals	1
1.3 Functioning of immune system of <i>Oryza sativa</i> against biotic and abiotic stresses	3
1.4 Current trends in enhancing stress-tolerance in <i>Oryza sativa</i>	5
1.5 Thesis statement	7
1.6 Scope of thesis	8
2 Literature review	11
2.1 Integrative breeding strategies in plant stress tolerance	11
2.1.1 Conventional plant breeding approaches	11
2.1.2 Modern plant breeding approaches	13
2.2 Advanced computational approaches to uncovering stress tolerance mechanisms in <i>Oryza sativa</i>	16
2.2.1 <i>In silico</i> characterization of novel disease-resistance proteins in <i>Oryza sativa Japonica</i> against bacterial leaf blight and rice blast diseases	16
2.2.2 Deep learning-based identification of disease-resistance proteins in <i>Oryza sativa</i> and its related species	17
2.2.3 AP2/ERF transcription factor orthologs of the desert tree <i>Prosopis cineraria</i> show higher copy number and DNA-binding affinity than drought-sensitive species	17
2.3 Conclusion	18
3 Characterization of novel disease-resistance proteins in <i>Oryza sativa Japonica</i> against bacterial leaf blight and rice blast diseases	19
3.1 Introduction	19
3.2 Materials and Methods	20
3.2.1 Data collection	20
3.2.2 Gene network construction and functional analysis	20
3.2.3 Physicochemical analysis of unknown disease-resistance proteins	21
3.2.4 Structure prediction and stereochemical property analysis of unknown disease-resistance proteins	21
3.2.5 Subcellular localization analysis	22
3.2.6 Multiple sequence alignment and phylogenetic analysis	22
3.2.7 Functional enrichment analysis of disease-resistance proteins	23
3.2.8 Identification of functional interactors in disease-resistance proteins	23
3.2.9 <i>In silico</i> gene expression analysis	24
3.3 Results	25
3.3.1 Identification of unknown disease-resistance genes in <i>Oryza sativa Japonica</i> through gene network analysis	25
3.3.2 Physicochemical characteristics of unknown <i>Oryza sativa Japonica</i> disease-resistance proteins	26
3.3.3 Three-dimensional structure of unknown disease-resistance proteins in <i>Oryza sativa Japonica</i>	27
3.3.4 Stereochemical features of unknown disease-resistance proteins in <i>Oryza sativa Japonica</i>	28
3.3.5 Subcellular localization of unknown disease-resistance proteins in <i>Oryza sativa Japonica</i>	29
3.3.6 Phylogeny of unknown disease-resistance proteins in <i>Oryza sativa Japonica</i>	31
3.3.7 Predicted functions of unknown disease-resistance proteins in <i>Oryza sativa Japonica</i>	34
3.3.8 Predicted functional interactors of unknown disease-resistance proteins in <i>Oryza sativa Japonica</i>	35
3.3.9 Gene expression under biotic stress of unknown disease-resistance proteins in <i>Oryza sativa Japonica</i>	36
3.4 Discussion	37

4	Deep learning based prediction of disease-resistance proteins in <i>Oryza sativa</i>	41
4.1	Introduction	41
4.2	Materials and Methods	42
4.2.1	Data collection	42
4.2.2	Data preprocessing	43
4.2.3	Feature extraction	43
4.2.4	Feature selection	46
4.2.5	Modeling of disease-resistance protein on the internal dataset	47
4.2.6	Evaluation of the DL-based model with ML models on an external dataset	50
4.2.7	Comparison of performance of the MLP model with well-established models	50
4.3	Results	51
4.3.1	Data optimization for model performance	51
4.3.2	Protein feature extraction and selection	52
4.3.3	Evaluation of the ML and DL model performance on the internal dataset	53
4.3.4	Evaluation of the ML and DL model performance on the external dataset	55
4.3.5	Comparison of the DL-based MLP model performance with existing models to predict disease-resistance proteins	56
4.3.6	Discussion	57
4.3.7	Conclusion	59
5	Genome-wide transcriptomic studies of the desert tree <i>Prosopis cineraria</i> to uncover drought tolerance genes	61
5.1	Introduction	61
5.2	Materials and methods	62
5.2.1	Reference-guided transcriptome analysis of <i>Prosopis cineraria</i> under drought stress	62
5.2.2	<i>De novo</i> transcriptome assembly of the Indian cultivar of <i>Prosopis cineraria</i>	63
5.2.3	Network analysis of <i>Prosopis cineraria</i> genes differentially expressed under drought	63
5.2.4	Phylogenetic classification of AP2/ERF proteins in <i>Prosopis cineraria</i> and <i>Oryza sativa Japonica</i> and <i>Indica</i>	64
5.2.5	Real-time quantitative PCR analysis of <i>Prosopis cineraria</i> AP2/ERF genes under drought	64
5.2.6	Analysis of the conserved motifs and gene structures in <i>Prosopis cineraria</i> AP2/ERF proteins	64
5.2.7	Promoter and <i>cis</i> -acting element analysis of AP2/ERF superfamily genes in <i>Prosopis cineraria</i>	65
5.2.8	Polymorphism analysis in AP2/ERF genes between Indian and Arabian cultivars of <i>Prosopis cineraria</i>	65
5.2.9	Copy-number variation analysis of the AP2/ERF superfamily between drought-tolerant and sensitive species	65
5.2.10	Modeling of apo and DNA-bound structures of the AP2/ERF and DREB proteins from drought-tolerant and sensitive species	65
5.2.11	Analysis of hydropathy and electrostatic properties of AP2/ERF proteins	66
5.2.12	All-atom molecular dynamics simulations	66
5.2.13	Hydrogen bond analysis	66
5.2.14	Binding energy calculation	66
5.2.15	3D structure and <i>in silico</i> DNA binding analysis of AP2/ERF proteins of <i>Prosopis cineraria</i> , <i>Oryza sativa Japonica</i> and <i>Indica</i>	67
5.3	Results	67
5.3.1	Stress-tolerance genes are highly induced in <i>Prosopis cineraria</i> under drought	67
5.3.2	Classification of <i>Prosopis cineraria</i> AP2/ERF genes	70
5.3.3	<i>Prosopis cineraria</i> AP2/ERF genes display different expression patterns under drought stress	70
5.3.4	<i>Prosopis cineraria</i> AP2/ERF promoters contain <i>cis</i> -elements for abiotic stress response	71
5.3.5	Gene structure and conserved motifs of <i>Prosopis cineraria</i> AP2/ERF proteins	73
5.3.6	AP2/ERF gene polymorphisms between Indian and Arabian cultivars of <i>Prosopis cineraria</i>	74
5.3.7	Drought-tolerant legumes contain more copies of AP2/ERF genes than drought-sensitive plants	81
5.3.8	The DNA-binding domains of <i>Prosopis cineraria</i> and <i>Arabidopsis thaliana</i> ERFs adopt the same fold but have chemical dissimilarity at a few positions	81
5.3.9	Differences in hydrogen bonding profile and charge distribution at the N- and C-terminal tails between <i>Prosopis cineraria</i> and <i>Arabidopsis thaliana</i> ERFs might be crucial in their differential DNA search and binding	83
5.3.10	DREB proteins have similar structures in drought-tolerant and drought-sensitive species	84
5.3.11	Classification of AP2/ERF genes in <i>Oryza sativa Japonica</i> and <i>Indica</i>	84
5.3.12	Structural differences between the AP2/ERF proteins of <i>Prosopis cineraria</i> , <i>Oryza sativa Japonica</i> and <i>Indica</i>	86
5.4	Discussion	90
6	Conclusions and Future work	95
6.1	Conclusions	95
6.2	Scope of Future Work	96
	Bibliography	99

Appendices	113
A Supplementary information for Chapter 3	115
B Supplementary information for Chapter 4	125
C Supplementary information for Chapter 5	131
Publications	135

List of Figures

Figure	Title	Page
1.1	A schematic diagram showing different types of stress and their impact on the overall health of <i>Oryza sativa</i>. The right-hand side panel depicts biotic stresses such as bacteria, fungi, viruses, etc. They affect <i>O. sativa</i> by declining its yield, growth, and photosynthetic response, and increasing cellular degradation and reactive oxygen species (ROS). In addition, the right-hand side panel highlights abiotic stresses such as heat, cold, and heavy metals including lead (Pb) and mercury (Hg), salinity by sodium ions (Na ⁺) and much more. The abiotic stress negatively impacts <i>O. sativa</i> by enhancing DNA damage, protein misfolding and declining CO ₂ and plant biomass.	3
1.2	Immune system of <i>Oryza sativa</i> consists of various physical barriers and induces biochemical, and molecular responses against biotic and abiotic stresses. The front line of defense occurs in the cell wall, stomata, waxy cuticle layer and lignins. The recognition of pathogens by MAMPs in <i>O. sativa</i> triggers immune responses as physical barriers. Further, at biochemical and molecular levels, PTI and ETI are activated which induces ROS, phytoalexin (low molecular weight antimicrobial compounds) and secondary metabolites such as carotenoids, terpenes, etc. In addition, the activation of signal transduction pathways including MAPK, protein kinases, TFs, etc., stress-tolerant and defense-related genes such as Xa21, AP2/ERF and phytohormones like ethylene, salicylic acid and so on were further triggered by various stresses in <i>O. sativa</i> . The immune responses of <i>O. sativa</i> not only helps in enhancing tolerance against stresses but also aids in maintaining high productivity.	5
1.3	Immune responses of <i>Prosopis cineraria</i> against abiotic stresses. <i>Prosopis cineraria</i> shows different stress tolerance responses against abiotic stress which include physical barriers, and biochemical and molecular responses. In DT <i>P. cineraria</i> , the primary defense is governed by various physical barriers such as cell wall rigidity, stomatal closure, thick cuticle layer and changes in roots and shoots. Moreover, the biochemical responses include the production of reactive oxygen species (ROS), enhancement in proline and phytohormones such as abscisic acid, ethylene and much more. Further, the activation of different signal transduction pathways includes MAPK and calcium signalling, stress-tolerant genes and transcription factors (TFs) such as Dehydration-Responsive Element-Binding Protein (DREB), Ethylene Responsive Factor (ERF) subfamily, etc. and osmoregulation.	7
1.4	A schematic representation of the thesis structure organized in chapters.	9
3.1	Gene-gene interaction network of hypothetically conserved and known disease-resistance genes against bacterial blight and rice blast diseases in <i>O. sativa</i>. (a) A gene-gene interaction network of bacterial blight disease-resistance genes in <i>O. sativa</i> is depicted. The network was constructed through the Expasy STRING tool and analyzed in the Cytoscape software (see "Materials and methods" section). (b) A gene-gene interaction network of rice blast disease-resistance genes in <i>O. sativa</i> is shown. The nodes indicate genes, and branches indicate the interactions in the network. The nodes were colored differentially based on their function. Yellow, black, green, purple, red, and grey colored nodes indicate their involvement in disease resistance, defense mechanism, disease and defense mechanism, stress-related, and less annotation, respectively.	25

3.2	<i>A comparative analysis of the 3D models of the disease-resistance proteins against bacterial blight and rice blast diseases predicted by the SWISS-MODEL and refined by the Galaxy Refine2 server and ModRefiner in O. sativa. a-d 3D models of the proteins BB.1 and BB.2 against bacterial blight disease in O. sativa. The models were predicted and refined by the SWISS-MODEL and Galaxy Refine2 servers, respectively. The template sequences for each protein were selected based on their GMQE score and sequence identity (see "Materials and methods" section). e-j 3D models of the proteins RB.1, RB.2, and RB.3 against rice blast disease predicted by the SWISS-MODEL server.</i>	29
3.3	<i>Subcellular localization of the disease-resistance proteins in the cell of O. sativa against bacterial blight and rice blast diseases. a, b Subcellular localization of proteins BB.1 and BB.2 against bacterial blight disease in O. sativa. DeepLoc 1.0, a DNN-based model, was executed to identify the subcellular localization through BB.1 and BB.2 amino acid sequence information. c-e The subcellular localization of the proteins RB.1, RB.2, and RB.3 against rice blast disease, identified through DeepLoc 1.0, are shown. The red line in the prediction of the subcellular localization indicates the existence of RB.1, RB.2, and RB.3 (a, b) at various locations in the cell of O. sativa, such as cell membrane, mitochondria, nucleus, etc., along with a likelihood score at each location.</i>	30
3.4	<i>Multiple Sequence Alignment and phylogenetic tree analysis of the disease-resistance proteins against bacterial blight and rice blast diseases. a, b Schematic representation of the phylogenetic trees for the proteins BB.1 and BB.2 against bacterial blight disease. First, the MSA was performed using the ggmsa package of the R software (version 4.2.2). Second, the phylogenetic analysis of the MSA results was performed using the NJ method available in the phylogram package of the R software. The trees were constructed through bootstrapping for 100 cycles to get more statistically confident node distances (see "Materials and methods" section). c-e Phylogenetic analysis of the proteins RB.1, RB.2, and RB.3 against rice blast disease. The MSA and phylogeny analysis were performed by the packages of the R program (a, b), and the dendrogram of all the disease-resistance proteins in O. sativa was in the form of unrooted trees.</i>	33
4.1	Schematic representation of proposed machine learning and deep learning workflow. <i>It includes proteomics data collection and its pre-processing to ensure data quality. This step involves filtering datasets, normalizing features, and addressing missing or redundant values. Further, relevant features were selected for model building. Finally, the model performance was cross-validated based on various performance metrics in post-analysis. .</i>	52
4.2	Dimensionality reduction and visualization on both internal and external datasets : (A) and (B) display PCA and t-SNE plots for the internal dataset, while (C) and (D) for the external dataset, respectively. The blue and orange dots correspond to class labels "0" and "1," where "0" represents non-disease-resistance proteins and "1" represents disease-resistance proteins. PCA applied to the 19,503 extracted protein features and showed a clear representation of variance distribution within the dataset. Additionally, t-SNE was employed to project the datasets into a three-dimensional space, enabling a more detailed visualization of the separation between disease-resistance and non-disease-resistance proteins based on the selected features. The t-SNE plots confirmed clustering patterns, indicating the ability of the selected features to discriminate between the two classes effectively.	53
4.3	A comparative evaluation of the performance of ML and DL models on the internal dataset. <i>The bar chart provides a visual representation of each model's performance during K-fold cross-validation. The results indicate that the DL-based MLP model outperformed other models, achieving the highest scores across multiple metrics, including accuracy, ROC-AUC, F1-score, precision, and recall. The x-axis represents the model names, while the y-axis corresponds to the range of scores attained by the models.</i>	55

- 4.4 **Comparative assessment of the DL-based MLP model with other ML models on the external dataset.** Comparison of the performance metrics, i.e., accuracy, ROC, F1-score, precision, and recall for each model predicting disease-resistance proteins from the new dataset. The graphical representation of the performance of the models shows the out-performance of the DL-based MLP model against other models with positive scores even when tested on an entirely new set of proteins. 56
- 5.1 ***Prosopis cineraria* genes differentially expressed under drought.** Roots of one-month-old *P. cineraria* seedlings were immersed for 6 hours in $\frac{1}{4}$ Hoagland's solution (control) and $\frac{1}{4}$ Hoagland's solution with 5% PEG-6000 (drought stress). The transcriptome of the fifth compound leaf was analyzed from three control and three drought-stressed plants. **(A)** Volcano plot showing differentially expressed genes in *P. cineraria*. Red dots indicate 2434 up-regulated genes, and green dots 1377 down-regulated genes with false discovery rate (FDR)-adjusted P -value < 0.05 and three biological replicates. Protein-protein interaction (PPI) network of *P. cineraria* **(B)** upregulated genes with $\log_2$ fold-change > 1 and FDR < 0.05 , **(C)** down-regulated genes with $\log_2$ fold-change -1 and FDR < 0.05 under drought stress. The protein IDs were converted to their corresponding *A. thaliana* orthologs (see Methods). The pink node color indicates up-regulated genes, and the green color indicates down-regulated genes. The colored circles around the nodes represent different enriched pathways to which these nodes are linked. A key to the node size representing the degree of the nodes and a color key of the nodes representing the $\log_2$ (fold-change) value of genes are shown at the bottom. The color chart represents enriched GO biological process categories and annotation identities. 68
- 5.2 **Classification of *Prosopis cineraria* AP2/ERF superfamily genes.** The amino acid sequences of 60 differentially expressed genes of the AP2/ERF superfamily of *P. cineraria* under drought stress were mined from the NCBI database. NCBI PSI-BLAST and HH-blits searches were performed exhaustively on these 60 proteins, and unique sequences containing AP2 domains were filtered. MSA was performed using Clustal Omega on the 255 unique hits obtained from the NCBI BLASTP search, together with 159 *A. thaliana* AP2/ERF superfamily proteins. The phylogenetic tree was constructed using the maximum likelihood method with the Jones-Taylor-Thornton model with 1000 bootstrap replicates in the MEGA11 application. The groups I-X were divided based on the classification of Nakano et al., whereas groups A1-A6 and B1-B6 followed the classification of Sakuma et al. [Sakuma et al., 2006] [Nakano et al., 2006]. Red and blue arrows indicate *P. cineraria* genes up-regulated and down-regulated under drought, respectively. *A. thaliana* proteins in each group are shown in red or white text. 71

- 5.3 **Expression and promoter cis-element analysis of *Prosopis cineraria* AP2/ERF genes.** (A) Time-dependent expression analysis of *P. cineraria* AP2/ERF genes under drought stress is shown. *P. cineraria* seedlings, grown for one month in soil, were uprooted carefully, washed with distilled water, and submerged in $\frac{1}{4}$ strength Hoagland's media for control samples and $\frac{1}{4}$ Hoagland's media supplemented with 5% PEG 6000 for stressed samples. The fifth leaf from the cotyledon for both control and stressed samples was harvested at 0 h, 1 h, 6 h, 12 h, and 24 h and snap-frozen in liquid nitrogen. Total RNA was isolated and converted to cDNA for real-time PCR. The expression levels of five upregulated genes, PcDREB1E-like (gene ID: 129311911), PcAIL5 (129320898), PcDREB2F (129320117), PcDREB1A-like (129312011), and PcERF109 (129293970) in the transcriptome analysis were validated by quantitative real-time PCR, by the standard curve method. PcActin (129287134) was the housekeeping gene to normalize sample expression levels. The relative fold changes in expression levels under PEG versus control at different time points are shown in panel A. The bars indicate an average of three biological replicates with standard errors. Asterisks indicate significant differences between relative expression levels at different time points ($P < 0.05$; Student's t-test). (B) Promoter cis-elements of *P. cineraria* 15 AP2/ERF genes. The promoter sequence up to 2.4 kb of a representative member from each group of the phylogenetic tree (see Figure 5.2) having the highest fold change of expression is shown. The cis-elements were identified by the PlantCare database and visualized by the TBtools. The scale at the bottom shows the promoter positions in base pairs. Colored boxes on the right indicate different cis-elements in the promoters. 72
- 5.4 **Conserved motifs and gene structures of *Prosopis cineraria* AP2/ERF superfamily proteins.** The boxes show an arrangement of protein motifs and the gene structures according to the phylogenetically classified I-Xb, AP2, RAV and Soloist of *P. cineraria* (see Figure 5.2). *P. cineraria* AP2/ERF mRNA and amino acid sequences were mined from NCBI. A motif search tool identified conserved AP2 and dehydrin motifs <https://www.genome.jp/tools/motif/>. PEST motifs were identified using epestfind <https://emboss.bioinformatics.nl/cgi-bin/emboss/epestfind>. Blue, red, cyan and orange boxes represent the AP2 domain, PEST, dehydrin and B3 motifs. The intron/exon arrangement-based gene structure of the AP2/ERF superfamily, which was analyzed in TBtools software, is next to the conserved motifs. The yellow and green boxes show exons and untranslated regions (UTR) in each mRNA sequence. 74
- 5.5 **Copy number variation of AP2/ERF proteins between drought-tolerant (DT) and drought-sensitive (DS) species.** Ten DT and 10 DS diploid plant species were selected to analyze copy number variations in the AP2/ERF superfamily proteins. The study was conducted through an exhaustive NCBI BLASTP search and copy numbers reported in the literature (see Methods). The vertical axis indicates the number of unique AP2/ERF proteins in each species. Among the diploid DT species, *P. cineraria* has the second-highest copy number of AP2/ERF proteins. 82

5.6 **Structure and multiple sequence alignment of DNA binding domains of ERF proteins from *Prosopis cineraria* and *Arabidopsis thaliana*.** (A) Crystal structure of DNA binding domain (DBD) of ERF protein from PDB (1GCC.pdb). (B) Modelled structures of DBD of ERF proteins from *A. thaliana* group VII (upper panel in blue cartoon) and *P. cineraria* group VII (lower panel in green cartoon). (C) MSA of homologous ERF sequences from group VII of *A. thaliana* and *P. cineraria*. (D) Modelled structures of DBD of ERF proteins from *A. thaliana* group VIIa (upper panel in blue cartoon) and *P. cineraria* group VII (lower panel in green cartoon). In both (B) and (D), the figures show the superimposition of ERF cartoon structure (shades of blue for *A. thaliana* and shades of green for *P. cineraria*) with 1GCC along with their root mean square deviation (RMSD). The electrostatic surface of the proteins is shown on the right, where blue and red represent positively and negatively charged surfaces, respectively. (E) MSA of ERF homologous sequences from group VIIa of *A. thaliana* and *P. cineraria*. In both (C) and (E), only a representative section from the DBD corresponding to the α -helix and the β -sheet region is shown. A difference in hydrogen bond donor position between the two species is highlighted with a black box and triangle. Similar triangles also show other specific hydrogen bonding positions (blue for *A. thaliana*, green for *P. cineraria*, and maroon when bonding was observed in both species). 83

5.7 **DNA binding interactions of ERF proteins from *Prosopis cineraria* and *Arabidopsis thaliana* determined from all-atom molecular dynamics (MD) simulations.** (A) Hydropathy (top) and Electrostatics (bottom) profiles for the full-length ERF proteins of group VII and VIIa from *A. thaliana* and *P. cineraria* showing their local hydrophobicity and charge along the amino acid sequence. For each residue position, the running average score for a window of five residues is plotted (see Method). In Group VII, peak (i) indicates positively charged residues in the DBD of *A. thaliana*, peak (ii) indicates the negatively charged region in the N-terminus of *A. thaliana* and peak (iii) shows a negatively charged region in the C-terminus of *P. cineraria*. In Group VIIa, peaks (i) and peak (ii) indicate positively charged residues for *A. thaliana* and *P. cineraria* in the DBD region, while peaks (iii) and (iv) show a negatively charged region in the C-terminus of *P. cineraria*. (B) Stable hydrogen bonds formed by the ERF DBD of *A. thaliana* and *P. cineraria* with the DNA in MD simulations. Protein-DNA complexes are shown in the cartoon, the hydrogen bond-forming residues in the sticks, and aromatic residues in red. The average number of stable hydrogen bonds in each case is shown at the bottom. Hydrogen bonds in each structure are calculated using the HBPLUS program. (C) Binding energy of DBD-DNA complexes calculated from MD trajectories using MM/GBSA method in AMBER package (see Method). Average values from 3 independent simulations are plotted in each category. (D) Residue-wise binding energy contributions of the DBD-DNA complexes. Only the residues making significant binding energy contributions (also highlighted in the MSA of Figure 5.6) are shown. 85

5.8	Classification of AP2/ERF superfamily genes of <i>Oryza sativa Japonica</i> and <i>Indica</i>. The amino acid sequences of 255 AP2/ERF proteins identified in <i>P. cineraria</i> were utilized for the NCBI PSI-BLAST search against the proteomes of <i>O. sativa Japonica</i> and <i>Indica</i> . NCBI PSI-BLAST search was performed exhaustively on these 255 AP2/ERF proteins and the unique sequences in <i>O. sativa Japonica</i> and <i>Indica</i> containing AP2 domains were filtered. MSA was performed using Clustal Omega on 162 unique hits obtained from the PSI-BLAST search against <i>O. sativa Japonica</i> and <i>Indica</i> , together with 255 <i>P. cineraria</i> AP2/ERF proteins. The phylogenetic tree was constructed using the maximum likelihood method with the Jones-Taylor-Thornton model with 1000 bootstrap replicates in the MEGA11 application. The rooted phylogenetic tree was visualized and studied to further classify the AP2/ERF proteins based on their evolutionary relationship in the iTOL tool. The groups I-X were divided based on the classification of [Nakano et al., 2006] whereas groups A1-A6 and B1-B6 followed the classification of [Sakuma et al., 2006]. Black, red and white coloured text indicate AP2/ERF protein members from <i>O. sativa Japonica</i> , <i>O. sativa Indica</i> and <i>P. cineraria</i>	86
5.9	Structure and multiple sequence alignment of DNA binding domains of ERF proteins from <i>Prosopis cineraria</i> and <i>Oryza sativa Japonica</i> and <i>Indica</i>. A-F Modelled structures of DBD of AP2/ERF proteins from <i>O. sativa Japonica</i> and <i>Indica</i> , and <i>P. cineraria</i> groups II, IV, V, VI and X. The red, green and blue boxes are for <i>Japonica</i> and <i>Indica</i> and <i>P. cineraria</i> AP2/ERF proteins, respectively, having amino acid variations. The A-F consists of the MSA of homologous ERF sequences from groups II, IV, V, VI and X. Further, hydrogen binding between AP2/ERF proteins with GCC-box is depicted with red, green and blue arrows for <i>O. sativa Japonica</i> , <i>O. sativa Indica</i> and <i>P. cineraria</i> AP2/ERF proteins, respectively, Similarly, the amino acid variation positions are highlighted with black arrows.	88
A.1	Analysis of the stereochemical properties of the 3D models of the disease-resistance proteins against bacterial blight and rice blast diseases in the rice plant. (a-b) Ramachandran plots of the 3D models of the disease-resistance proteins against bacterial blight disease. The stereochemical properties of the proteins BB.1 and BB.2 were identified using the Ramachandran plot server and PROCHECK tool. The Ramachandran plots describe the peptide backbone torsion angles, phi and psi (ϕ and ψ). (c-e) The stereochemical properties of the proteins RB.1, RB.2, and RB.3 were examined through the Ramachandran plot server and PROCHECK tool. In the plots, the most favored region, additionally allowed regions, generously allowed regions, and disallowed regions are indicated as red, yellow, light yellow, and white, respectively.	116
A.2	Structural alignment of the 3D models of the disease-resistance proteins against bacterial blight and rice blast diseases in the rice plant predicted by the SWISS-MODEL with the I-TASSER and AlphaFold. (a-d) Alignment of the 3D models predicted by the SWISS-MODEL with I-TASSER and AlphaFold for the proteins BB.1 and BB.2 is shown. The alignment was performed by the 'align' function in the PyMOL, which superimposes the two protein structures on each other based on the homology between them. (e-j) Structural alignment of the 3D models of the proteins RB.1, RB.2, and RB.3 predicted by the SWISS-MODEL with I-TASSER and AlphaFold is depicted. The alignment of disease-resistance proteins against bacterial blight and rice blast diseases in the rice plant indicates that SWISS-MODEL was the best-suited method compared to the I-TASSER and AlphaFold.	117

A.3	Structural alignment between the predicted 3D and refined models for the disease-resistance proteins against the bacterial blight and rice blast diseases. (a-b) Alignment of the proteins BB.1 and BB.2 against bacterial blight disease performed by the PyMOL is shown. The alignment between 3D and refined models of BB.1 and BB.2 was performed by the 'align' function in the PyMOL, which superimposes the two protein structures on each other based on the homology between them. (c-e) PyMOL was utilized to perform the structural alignment between the 3D and refined models of the proteins RB.1, RB.2, and RB.3. The alignment indicates the 3D models before and after the refinement executed by the Galaxy Refine2 server and ModRefiner.	120
A.4	Protein-protein interactions between the identified disease-resistance and known disease-resistance proteins against bacterial blight and rice blast diseases in the rice plant. (a-b) Protein-protein interactions between proteins BB.1 and BB.2 and known disease-resistance proteins against bacterial blight disease. The interactions were constructed and visualized on the amino acid sequences of BB.1 and BB.2 in the Expasy STRING database. (c-e) A protein-protein network of the proteins RB.1, RB.2, and RB.3 with known disease-resistance proteins was constructed by utilizing the Expasy STRING database. The query proteins (a-e) were labelled with light pink, whereas other colors indicated the association of the query proteins with the known disease-resistance proteins. Purple, dark blue, green, dark green, light violet, light blue, and black colors indicate different types of associations between query and the known disease-resistance proteins, viz., experimentally-determined, gene co-occurrence, text mining, gene neighborhood, protein homology from curated databases, and co-expression.	121
B.1	Confusion matrix of each model on the internal dataset. (A) to (F) Confusion Matrix of ML and DL models, including LR, RF, SVM, NB, KNN and MLP, respectively. The matrix was generated through the TP, TN, FP, and FN values of each model. It shows the highest TP values in the DL-based MLP model compared to other ML models (see "Results" section).	127
B.2	Confusion matrix of the DL-based MLP model and ML models for the external dataset. (A) to (F) A comparison between TP, TN, FP, and FN values in the confusion matrix for each of the LR, RF, SVM, NB, KNN, and MLP models. The confusion matrix was constructed for the external dataset provided to the models for evaluating their performance.	127
B.3	AUC curves for the external dataset for all the ML and DL models. (A) to (F) A comparison between AUC scores for all the models shows that the panel (F) i.e., the DL-based MLP model has the best performance. The x-axis shows the False Positive Rate (FPR) and True Positive Rate (TPR) which can be found highest in the DL-based MLP model.	129
C.1	Phylogenetic tree of AP2/ERFs from <i>Vigna unguiculata</i> and <i>Prosopis cineraria</i>. The amino acid sequences of 60 differentially expressed genes of the AP2/ERF superfamily of <i>P. cineraria</i> under drought stress (see Table S1) were mined from the NCBI database. NCBI PSI-BLAST and HHblits searches were performed exhaustively on these 60 proteins, and unique sequences with AP2 domain filtered. Multiple Sequence Alignment (MSA) was performed using Clustal Omega on the 255 unique hits obtained from the NCBI BLASTP search, together with 290 <i>V. unguiculata</i> AP2/ERF superfamily proteins. The phylogenetic tree was constructed using the maximum likelihood method with the Jones-Taylor-Thornton model with 1000 bootstrap replicates in the MEGA11 application. The groups I-X were divided based on the classification of Nakano et al. (2006), whereas groups A1-A6 and B1-B6 followed the classification of Sakuma et al. (2002). <i>V. unguiculata</i> proteins are shown in red or white font.	131

C.2	Structures of DNA-binding domain of ERF from <i>Arabidopsis thaliana</i> and <i>Prosopis cineraria</i>. (A, B) Cartoon representation of the DNA binding ERF (1GCC, NMR structure, in grey) and AP2/ERF domain (5WX9, x-ray structure, in grey) from <i>Arabidopsis</i> in complex with its target DNA fragment (in orange). Residues involved in specific interactions with DNA are shown in blue (positively charged) and red (aromatic) sticks. .	132
C.3	Multiple sequence alignments of AP2/ERF superfamily proteins of <i>Prosopis cineraria</i> and <i>Arabidopsis thaliana</i>. The alignments were conducted by CLUSTALW in BioEdit software, separately for each group of the phylogenetic tree (5.2). The black boxes indicate AP2 domains.	132
C.4	Structural comparison of DREB1 proteins from drought tolerant and sensitive species. Modeled full-length structure of dehydration-responsive element binding protein 1 (DREB1) from drought tolerant (top panel) and drought sensitive (bottom panel) species using AlphaFold2. The same representation schemes as that of Figure 1 are used. For each species, the left pane shows only the DNA-binding domain (in black cartoon) superposed on 1GCC.pdb (grey cartoon) whereas the right panel shows the electrostatic surface of the full-length protein in which blue and red correspond to the positive and negatively charged surfaces, respectively. DNA is shown in an orange cartoon. (B) Sequence alignments of DREB1 proteins from four drought-sensitive and tolerant species are shown.	133
C.5	Structural comparison of DREB2 proteins from drought tolerant and sensitive species. (A) Modeled full-length structure of dehydration-responsive element binding protein 2 (DREB2) from drought tolerant (top panel) and drought sensitive (bottom panel) species using AlphaFold2. The same representation schemes as that of Figure 1 are used. For each species, the left pane shows only the DNA-binding domain (in black cartoon) superposed on 1GCC.pdb (grey cartoon) whereas the right panel shows the electrostatic surface of the full-length protein in which blue and red correspond to the positive and negatively charged surfaces, respectively. DNA is shown in an orange cartoon. (B) Sequence alignments of DREB2 proteins from four drought-sensitive and tolerant species are shown.	134

List of Tables

Table	Title	Page
2.1	Show the drought tolerance genes (or proteins) incorporated in <i>O. sativa</i> from drought-tolerant species.	18
3.1	Shows physiochemical features of five predicted disease-resistance proteins.	27
3.2	Domain analysis of the predicted disease-resistance proteins.	31
3.3	Functional annotation of the predicted disease-resistance proteins.	34
3.4	KEGG pathway analysis of the predicted disease resistance proteins.	35
3.5	Gene expression analysis of the predicted disease-resistance proteins under pathogen infection in <i>Oryza sativa Japonica</i>	37
4.1	Performance matrix comparison of different models on the internal dataset through the 10-fold cross-validation method.	54
4.2	Performance matrix comparison of different models on the external dataset by the 10-fold cross-validation method.	54
4.3	Comparison of performance metrics between MLP model and other existing models on internal and external datasets.	57
5.1	Enriched Gene Ontology and KEGG Pathways for 2434 up-regulated genes and 1377 down-regulated genes in <i>Prosopis cineraria</i> under drought stress identified using the g:Profiler software. The frequency of occurrence of a particular pathway in the set of 2434 genes (input number) versus the frequency in the <i>P. cineraria</i> genome (background number) determined the enrichment. Enrichment categories with <i>P</i> -value <0.05 are presented (Benjamini-Hochberg FDR).	69
5.2	Polymorphism analysis in AP2/ERF genes between Indian and Arabian cultivars of <i>Prosopis cineraria</i>. DNA polymorphisms in AP2/ERF superfamily genes, differentially expressed under drought, between an Arabian and an Indian cultivar of <i>P. cineraria</i> are presented. The gene ID and positions refer to the UAE cultivar PC_DT_omini genome reported in NCBI (GCF_029017545.1). * Indicates a STOP codon and "del" denotes a deletion. Note that the gene ID pairs 129292617 and 129298057, 129295493 and 129318720, and 129301613 and 129300577 share 99.92%, 99.7%, and 100% sequence similarities.	75
5.3	Overview of the hydrogen binding positions and their counts in AP2/ERF proteins with GCC-box in <i>Prosopis cineraria</i> and <i>Oryza sativa Japonica</i> and <i>Indica</i>	89
A.1	Overview of different databases and tools used in the study.	117
A.2	The secondary structure profile of the five unknown disease-resistant proteins.	119
A.3	The stereochemical analysis of the five disease-resistant proteins using different tools.	120
A.4	Comparison between SWISS-MODEL, I-TASSER and AlphaFold predicting five disease-resistant proteins.	122
A.5	Shows CYPRED predicted cysteine residues playing important in disulphide bonding.	122
A.6	Information on pathway analysis of the disease-resistance proteins.	122
A.7	Comparison between the pI DDT and RMSD values of the SWISS-MODEL, AlphaFold and I-TASSER for the five disease-resistance proteins.	123

B.2	Summary of the predicted disease-resistance proteins by the DL-based MLP model on the external dataset.	125
B.1	List of various descriptors computed by <i>protr</i> package.	128

List of Abbreviations

AAC	Amino acid composition
AAindex	Amino acid properties
ABA	Absciscic acid
ABRE	Absciscic Acid-Responsive Element
AFLP	Amplified Fragment Length Polymorphisms
AP2	APETALA2
AP2/ERF	APETALA2/Ethylene Responsive Factor
API	Application programming interface
ARE	Anaerobic Response Element
ASM	Allele-specific marker
AUC	Area Under Curve
Avr	Avirulence
BB	Bacterial leaf blight
BC1	Backcross hybrids
BUSCO	Benchmarking Universal Single-Copy Orthologue
bwa	Burrows-Wheeler Alignment
C	Cysteine
CDD	Conserved Domain Database
CNVs	Copy number variations
CRISPR-Cas9	Clustered Regularly Interspaced Short Palindromic Repeats
CRT	C-repeat element
CTD	Composition-transition-distribution
D	Aspartic acid
DAMPs	Danger-associated molecular patterns
DBD	DNA binding domain
DDA	Data-dependent acquisition
DeepLRR	Deep Leucine-rich-repeats
DEGs	Differentially expressed genes
DenseNet	Densely Convolution Neural Network
DIA	Data Independent Acquisition
DL	Deep learning
DNN	Deep Neural Network
DRAGO3	Disease Resistance Analysis and Gene Orthology
DRE	Dehydration-responsive element
DREB2	Dehydration-Responsive Element-Binding Protein2
DS	Drought-sensitive
DT	Drought-tolerant
ENA	European Nucleotide Archive
ER	Endoplasmic reticulum
ERE	Ethylene Response Element
ERF	Ethylene-Responsive Factor
ETI	Effector-triggered immunity
F1	Filial 1

Faster R-CNN	Faster Region-Based Convolutional Neural Network
FN	False negative
FP	False positive
FPR	False positive rate
FTP	File Transfer Protocol
GATK	Genome Analysis Toolkit
GEO	NCBI's Gene Expression Omnibus
GMQE	Global Model Quality Estimate
GO	Gene Ontology
GS	genomic selection
GWAS	Genome Wide Association Studies
H	Gistidine
Hd1	HEADING DATE 1
Hd3a	Heading date 3a
Hg	Mercury
HMM	Hidden Markov models
HTP	High Throughput Phenotyping
INSDC	International Nucleotide Sequence Database Collaboration
I-TASSER	Iterative Threading ASSEmbly Refinement
iTOL	Interactive Tree of Life
K	lysine
KEGG	Kyoto Encyclopedia of Genes and Genomes
KNN	K-nearest neighbour
LEA	Late embryogenesis abundant
LR	Logistic regression
LRR	Leucine-rich repeats
LRRfinder	Leucine-rich-repeats finder
LRRpredictor	Leucine-rich-repeats predictor
LRRsearch	Leucine-rich-repeats search
<i>M. oryzae</i>	<i>Magnaporthe oryzae</i>
MAMPs	Microbe-associated molecular patterns
MAPKs	Mitogen activated protein kinases
MAS	Marker Assisted Selection
MD	Molecular dynamic
MDA	Malondialdehyde
MDGs	Millennium Development Goals
MFT	MOTHER of FT and TFL1-like
MIAME	Minimum Information About a Microarray Experiment
ML	machine learning
MLP	Multilayer Perceptron
MNPs	Multi-nucleotide polymorphisms
MSA	Multiple Sequence Alignment
MYB	Myeloblastosis
NB	Naïve Bayes
NBS	Nucleotide-binding sites
NCED1	9-cis-epoxycarotenoid dioxygenase 1
NJ	Neighbour joining
NJ-MCL	Neighbor-joining maximum composite likelihood
NLRs	Nucleotide-binding domain and leucine-rich repeat-containing proteins
NRD	Negative regulatory domain
OsG1	<i>O. sativa</i> GIGANTEA

PAM	Protospacer adjacent motif
PANNZER	Protein Annotation with Z-score
Pb	Lead
PCA	Principal Component Analysis
PCM	Proteochemometric modeling
Pfam	Protein families database
pI	Isoelectric point
PLS-DA	Partial least squares discriminant analysis
PPI	Protein-protein interactions
PPV	Positive predictive value
PR	Pathogenesis-related
PRGdb	Plant Resistance Genes Database
PRIDE	Proteomics IDentification
Prosite	Protein database
PRRs	Pattern recognition receptors
PseAAC	Pseudo-amino acid composition
PSI-BLAST	Position-Specific Iterated BLAST
PSIPRED	Based Secondary Structure Prediction
PSSM	Position-Specific Scoring Matrix
PTI	Pattern-triggered immunity
QSO	Quasi-sequence-order
QTL	Quantitative Trait Loci
RAPD	Random Amplified Polymorphic DNA _v
RAV	RELATED TO ABI3 AND VP1
RB	Rice blast
RF PDR	Random Forest for Plant Disease
RF	Random Forest
RFLP	Restriction Fragment Length Polymorphism
RFT1	Rice flowering locus T1
RLKs	Receptor-like kinases
RLPs	Receptor-like proteins
RMSD	Root Mean Square Deviation
ROC	Receiver Operating Curve
ROS	Reactive oxygen species
RRG	Predictor Resistance Gene predictor
RRK XA21	Rice Receptor Kinase XA21
SANS	Suffix array neighborhood search
SAR	Systemic acquired resistance
SASA	Solvent-accessible surface area
SDGs	Sustainable Development Goals
SET	Su(var)3–9, Enhancer-of-zeste and Trithorax
sgRNA	single-guide RNA
SMART	Simple Modular Architecture Research Tool
SNPs	Single nucleotide polymorphisms
SOAP	Simple Object Access Protocol
SOCN	Sequence-order-coupling number
SPRY	SE141a and RYanodine receptor
SSRs	Simple sequence repeats
STAR	Spliced Transcripts Alignment to a Reference
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
SVM	Support vector machine

TFs	Transcription factors
TM	Transmembrane
TN	True negative
TP	True positive
TPR	True positive rate
UTRs	5'-untranslated regions
VCF	Variant call format
W	Tryptophan
Webin-CLI	Webin command line submission interface
WHIM	Weighted holistic invariant molecular
Xoo	<i>Xanthomonas oryzae</i> pv. <i>Oryzae</i>
Y	Tyrosine

List of Symbols

$f(r)$	Fraction of each amino acid type
N_r	Number of the amino acid type r
r	Amino acid type r
$\bar{P}$	Average property of the twenty amino acids
C_r	Composition
n_r	Number of amino acid type r in the encoded sequence
n, N	Sequence length
T_{rs}	Transition
n_{rs}	Numbers of dipeptides encoded as “rs”
n_{sr}	numbers of dipeptides encoded as “sr” in the sequence
f_i	i^{th} value of frequency of the triads of the protein sequences
f_r	Normalized occurrence for amino acid type i
w	Weighting factor
τ_d	d -th rank sequence-order-coupling number
$d_{i,i+d}$	Distance between the two amino acids at position i and $i+d$
$H_1(i)$	Original hydrophobicity value
x	Input where each tree casts a unit vote for the most popular class
Θ_k	Independent identically distributed random vector
P_r	Autocorrelation function