COMPUTATIONAL STRUCTURAL ANALYSIS AND MODELING OF PLANT PRR DORN1 ECTODOMAIN

By

Aorchie Siddiqui

Student ID: 15136017

A thesis submitted to the Department of Mathematics and Natural Sciences in partial fulfillment of the requirements for the degree of Bachelor of Science in Biotechnology

Mathematics and Natural Sciences
Brac University
June, 2019

Dedicated to my parents and grandfather,

Latifa Jahan, Dr. Arifur Rahman Siddiqui

and M.A Latif

Declaration of Authenticity

I, the undersigned, declare that the research work embodying the results reported in this thesis entitled "Computational structural analysis and modeling of plant PRR DORN1 ectodomain" is my original work, gathered and utilized for the sole purpose of fulfilling the objectives of this study. I confirm that the work has not been previously submitted to any other institution, in whole or in part, for a higher degree or diploma. I further declare that the thesis has been composed entirely by me under the supervision of Dr. Mahboob Hossain, Professor, Microbiology Programme, Department of Mathematics and Natural Sciences, BRAC University, Dhaka, except where stated otherwise by reference or acknowledgement.

Aorchie Siddiqui

Certified

Dr. Mahboob Hossain, Supervisor,
Professor, Microbiology Programme,
Department of Mathematics and Natural Sciences,
BRAC University,
Dhaka.

Acknowledgement:

First and foremost, I am grateful to my parents for always having faith in me, loving and supporting me in every aspect of my life and all my endeavors. They have been a constant source of inspiration and without their encouraging words this thesis would not have been possible.

My profound gratitude goes to my supervisor Dr. Mahboob Hossain, Professor, Microbiology Program, Department of Mathematics and Natural Sciences, BRAC University for accepting me as his thesis student and allowing me to finish my degree.

I extend my gratitude to M H M Mubassir, Lecturer, Biotechnology Program, Department of Mathematics and Natural Sciences for providing me insight and expertise into this work.

I would also like to take this opportunity to thank Eusra Mohammad, Lecturer (former), Nahreen Mirza and Salman Khan Promon for their wise words, guidance and encouragement in the short time I had with them.

My warmest gratitude goes to my close friend and biggest well-wisher, Md. Tanvir Rounak Anjum for believing in me, being my support system and comforting me during my bad times. Last but not least special thanks go to my friend Faria Marha for always lending a helping hand whenever I needed it.

Aorchie Siddiqui

Table of Contents:

List of Tables:

List of Figures:

Abstract

DORN1 also known as P2K1 is the first member of the plant-specific purinoreceptor subfamily termed P2K and belongs to the vast gene family of legume-type (L-type) lectin receptor kinases. Research has shown that DORN1 recognizes extracellular ATP (eATP) with high affinity, which likely plays a role as a central danger signaling molecule in many plant stress response pathways. However, a lack of information on the tertiary structure of the L-type lectin ligand binding domain of DORN1 hinders efforts to fully explain the recognition and signaling mechanisms of DORN1. To date there exist no experimentally determined structure of DORN1 in protein databases as a result using computational approaches; ab-initio and comparative modeling techniques provide an alternative and reliable method in predicting the structure and thus the function accordingly

Chapter 1: Introduction

1.1 Plant Innate Immune System

As sessile organisms plants are continuously exposed to various biotic and abiotic stresses in their immediate environment as a result they have evolved effective mechanisms to defend themselves from such stresses. Lacking specialized immune cells and an adaptive immune system, plants instead rely entirely on the ability of each cell to trigger innate immune responses that result in both local and systemic responses (Zipfel, 2014). Consequently enabling the plant to combat pathogens quickly and locally, and on an extended time and space scale (Zipfel, 2014). Plants employ a two-tiered innate immune system. The first tier uses transmembrane pattern recognition receptors (PRRs) that recognize both exogenous danger signals referred to as PAMPs or MAMPs and endogenous danger signals called DAMPs (Jones & Dangl, 2006). The second tier uses polymorphic NB-LRR proteins that recognize pathogen effector molecules and mainly act inside the cell (Jones & Dangl, 2006). This type of two-tiered immune system was illustrated by Jones and Dangl in 2006 through a zigzag model containing four phases [Figure 1.1]:

In the 1st phase different PRRs recognize different MAMPs, PAMPs and DAMPs resulting in PTI (Pattern Triggered Immunity). Some successful pathogens can evade PTI and release effectors to subvert it. These effectors interfere with PTI resulting in phase 2 ETS (Effector Triggered Immunity). In the 3rd phase these effectors are recognized by specific NB-LRR proteins either directly or indirectly, activating ETI. In the 4th and final phase pathogens are forced to diversify existing effectors or gain new effectors in order to suppress ETI as a result of natural selection.

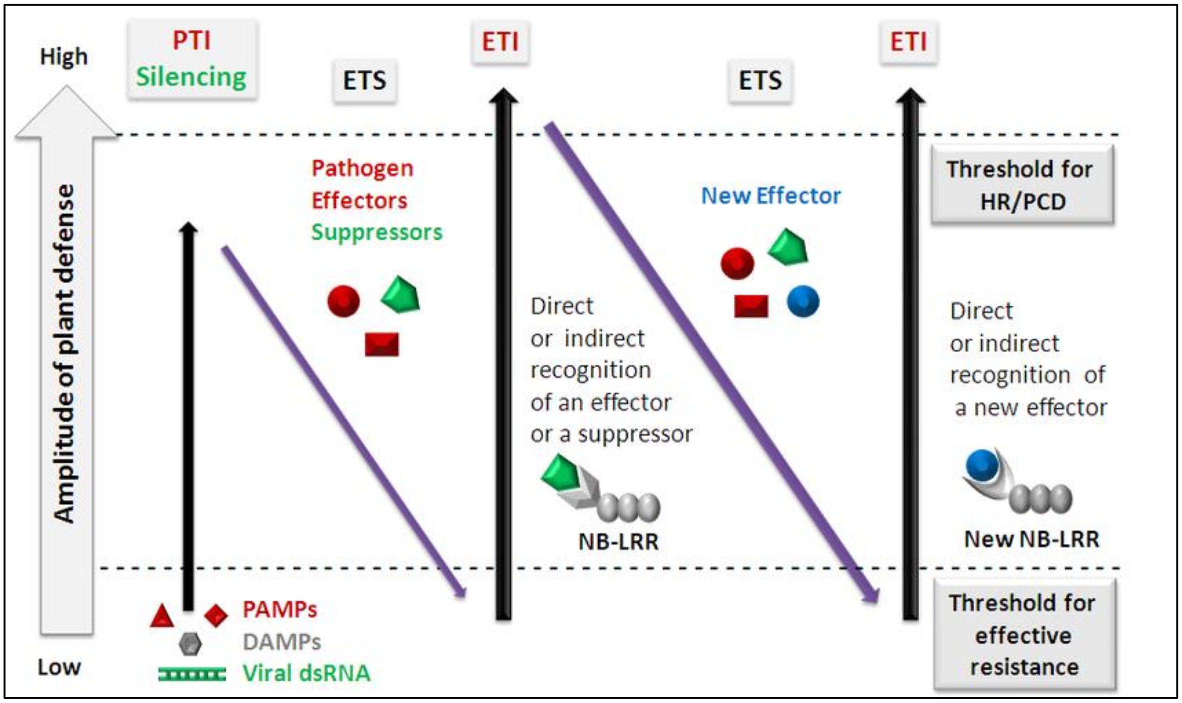


Figure 1.1 Zigzag model of plant immune system (Jones & Dangl, 2006)

1.2 Pattern Triggered Immunity

The first branch of the plant innate immune system is PTI, activated by cell surface localized PRRs that perceive conserved PAMPs, MAMPs (non-self) and host derived DAMPs (self) (Bigeard, Colcombet & Hirt, 2015). Currently, plant PRRs are of two types, one is receptor kinases which are comprised of a ligand binding ectodomain, a transmembrane domain, a juxtamembrane domain and a cytoplasmic kinase domain. The second is receptor like proteins which lack a cytoplasmic signaling domain (Ranf, 2017). To perceive chemically diverse ligands both RKs and RLPs combine with various extracellular domains such as leucine rich repeat (LRR), lysine-motif (LysM) or lectin domains. During recognition a co-receptor may or may not be recruited for full PTI activation (Ranf, 2017). Upon ligand binging certain PRRs homodimerize and form heterooligomeric complexes with other PRRs (Zipfel, 2014).

Due to the conserved nature of PAMPs and MAMP, PTI has the ability to prevent non-adapted pathogenic microbes from infecting the host, termed as non-host resistance (NHR) (Lee et al., 2017) and also contribute to basal immunity by restricting infection of adapted pathogens in susceptible hosts (Zipfel, 2014).

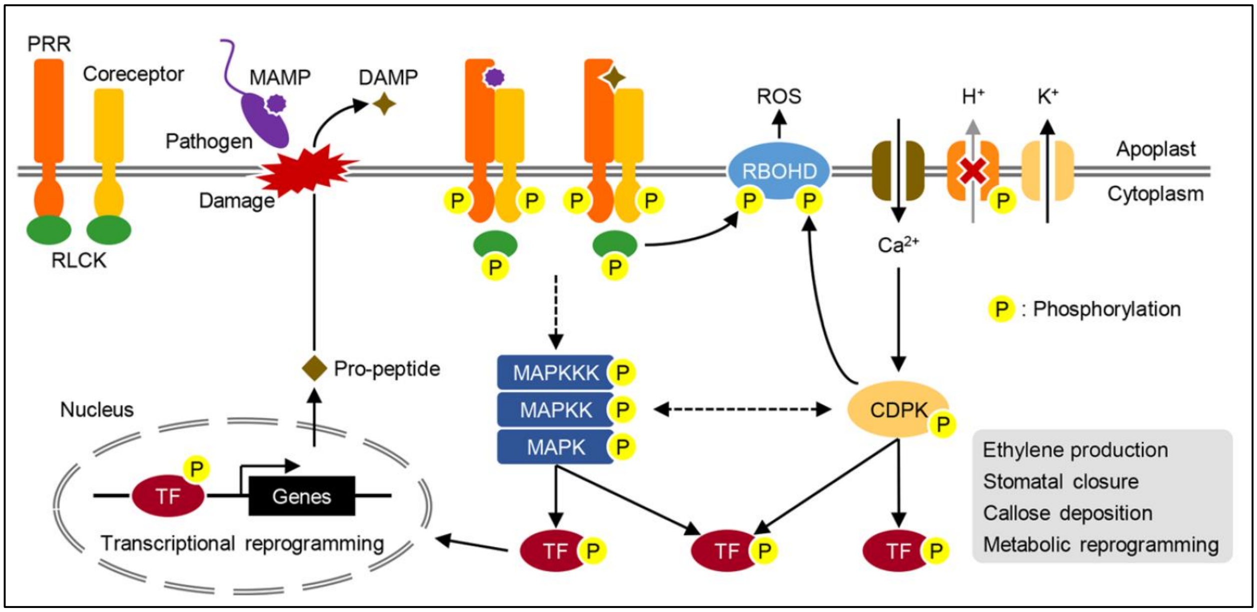


Figure 1.2 Overview of Pattern Triggered Immunity (Saijo, Loo & Yasuda, 2018)

1.3 MAMPs, PAMPs, and DAMPs:

The evolution of innate immune system in multicellular organisms required the development of cell surface receptors capable of recognizing / binding molecules whose chemical structure / pattern is usually preserved within different classes of foreign organisms but absent from "self" molecules (Choi & Klessig, 2016). Such conserved foreign (non-self) molecules are referred to as Microbe-Associated Molecular Patterns (MAMPs), also termed as Pathogen Associated Molecular Patterns (PAMPs). Examples of MAMPs include (Ranf, 2017): Bacterial cell surface and secreted compounds such as flagellin, peptidoglycans, lipopolysaccharide, and intracellular components such as elongation factor Tu (EF-Tu), proteins, DNA. Fungal cell wall chitin, enzymes xylanase, endopolygalacturonase. Oomycete elicitins, endoglucanase. Viral double stranded RNA. Nematode ascarosides. Glycoproteins of parasitic plant *Cuscuta spp.*

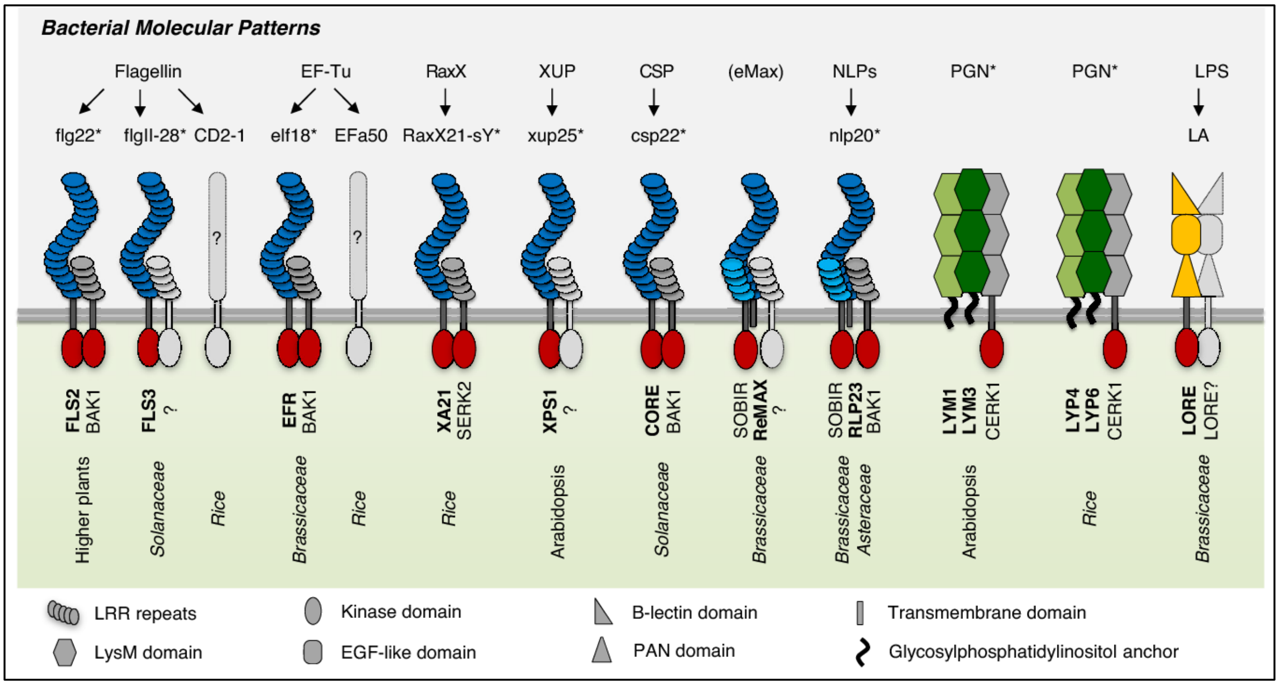


Figure 1.3 Bacterial MAMPs with their respective PRR (Ranf, 2017)

In addition to perceiving MAMPs plants are capable of recognizing components released from their usual location into the extracellular space due to cell/tissue damage. Such molecules are termed as Damage-Associated Molecular Patterns (DAMPs) (Choi & Klessig, 2016). Examples of DAMPs include oligogalacturonides a plant cell wall component, elicitor peptides and extracellular ATP (Ranf, 2017). Derived from microorganisms MAMPs are responsible for activating the plant innate immune system, while DAMPs are derived from host cell and both initiate and perpetuate innate immune responses (Choi & Klessig, 2016).

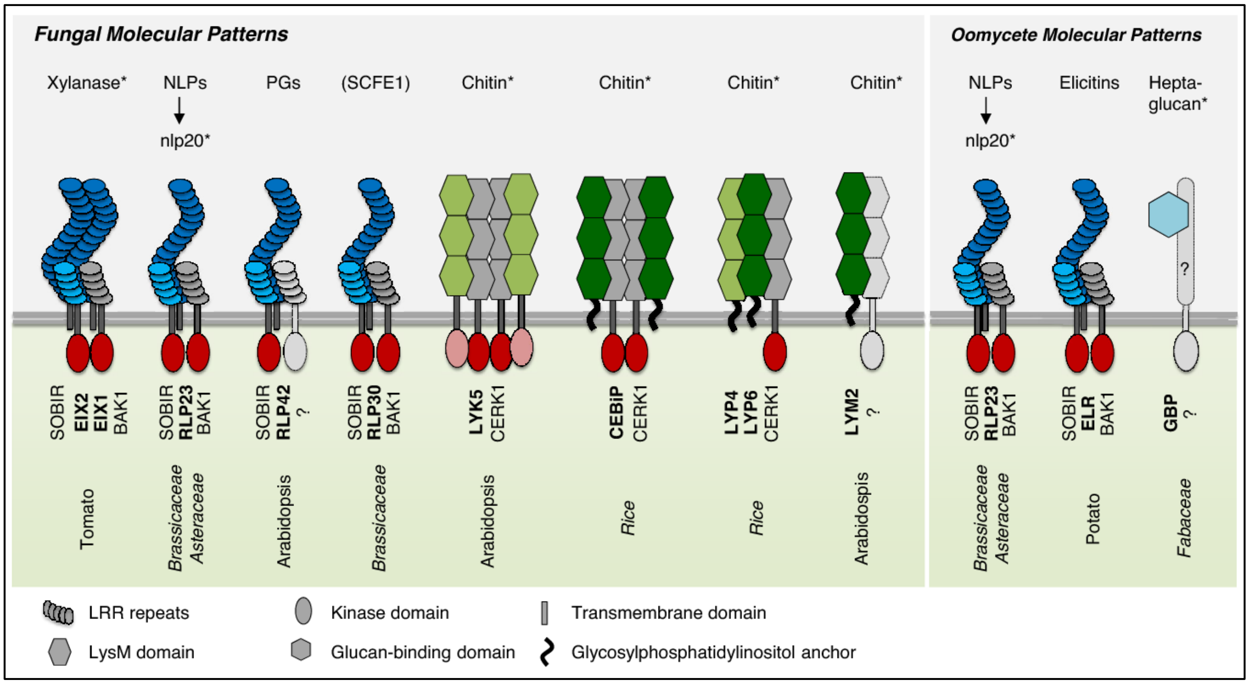


Figure 1.4 Fungal and Oomycete MAMPs with their respective PRR (Ranf, 2017)

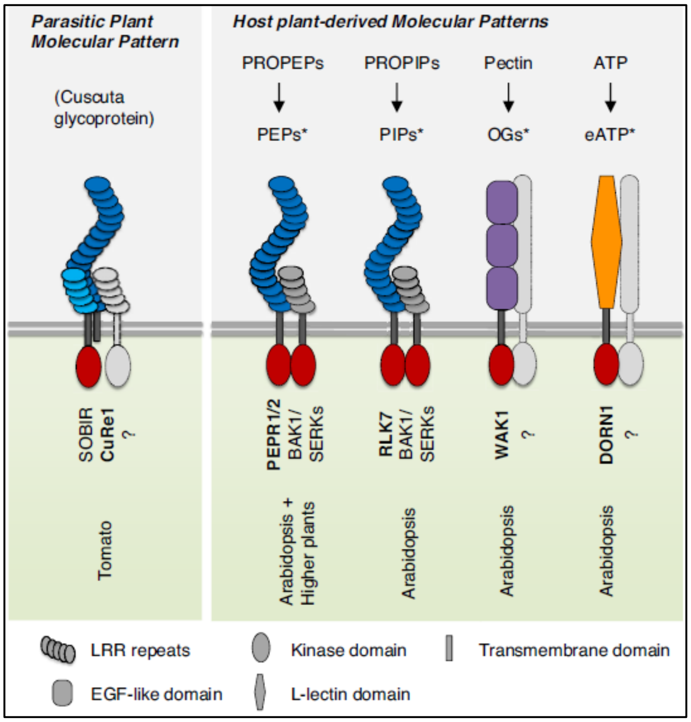


Figure 1.5 Parasitic plant MAMP and Host plant DAMPs and their respective PRRs (Ranf, 2017)

1.4 DORN1 :

DOes not Respond to Nucleotides 1(DORN1) is a PRR belonging to the lectin receptor kinase (LecRK) class of proteins originally described in *Arabidopsis* (Hervé et al., 1996). Based on their extracellular domains, plant RLKs can be grouped into over 21 structural classes (Shiu & Bleecker, 2001). The LecRKs are one of those classes in which the extracellular domain is associated with legume lectins (Barre, Hervé, Lescure & Rougé, 2002). Like other LecRKs, DORN1 is characterized by an extracellular legume (L-type) lectin domain that binds to eATP, a transmembrane domain and an intracellular serine/threonine kinase domain (Barre, Hervé, Lescure & Rougé, 2002).

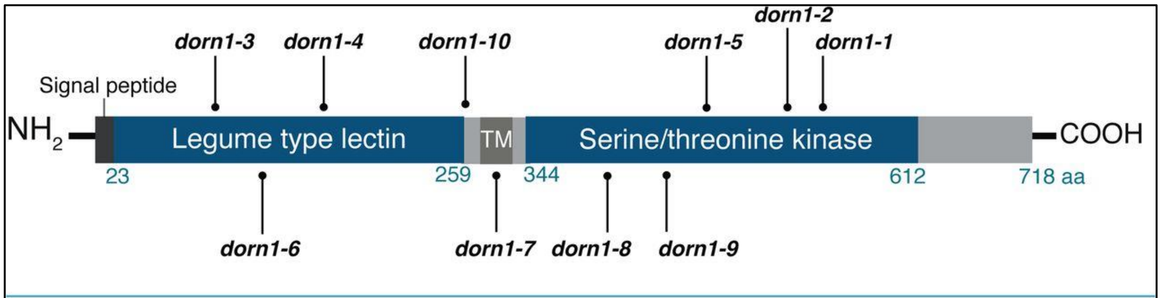
DORN1 was identified as the first plant receptor that binds to eATP, a DAMP, by genetic screening for ATP-insensitive mutants in *Arabidopsis* (Choi et al., 2014a). The genetic screening revealed that DORN1 is required for ATP- induced calcium influx and the activation of MPK, (Choi et al., 2014a) both involved in signature stress responses (Zhang, Du & Poovaiah, 2014) (Meng & Zhang, 2013) . In addition microarray analysis and gene-ontology enrichment test revealed that DORN1 was also required for upregulation in expression of stress responsive genes and signal transduction genes (Choi et al., 2014a).

With a dissociation constant of 45.7 nM, DORN1 ectodomain binds to ATP with relatively high affinity. It also shows a preference for purine nucleotides over pyrimidine nucleotides with ATP and ADP being the most active agonists followed by ITP, GTP and UTP. Pyrimidine nucleotides such as TTP and CTP showed no competition (Choi et al., 2014a).

The transmembrane domain of DORN1 plays an important role in the localization of the receptor to the plasma membrane as observed when transiently expressed from the native promoter in *Arabidopsis*, tobacco leaves (Bouwmeester et al., 2011) or stably expressed in potato (Bouwmeester et al., 2013). While a significance of the transmembrane domain beyond localization is unknown, a single nucleotide substitution, A306, within the transmembrane domain, in the dorn1-7 mutant disrupts the ATP-induced calcium response [Figure1.6].

The cytoplasmic domain of DORN1 is a typical serine/threonine kinase comprising 12 subdomains. In vitro kinase assays carried out using the purified kinase domain, demonstrated that the DORN1 has autophosphorylation and transphosphorylation activities (Choi et al., 2014a). In vitro phosphorylation assays showed that mutations in dorn1-1 (D572N) and dorn1-2 (D525N) resulted in the complete loss of kinase function and response to extracellular ATP (Choi et al., 2014a). These mutations resulted in the conversion of amino acid Asp572 and Asp525 to Asn, which are known to be vital for phosphorylation of the activation loop and stabilization of the catalytic loop of the kinase respectively (Choi et al., 2014a). Mutation in the conserved aspartate residue (Asp486) which interacts with Mg2+ associated with β and γ phosphates of ATP, in dorn1-5 results in reduced ATP responses. Furthermore mutation of the conserved arginine residue in mutant dorn1-9 (R467Q) compromises the ATP response indicating the significance of this residue for the proper functioning of DORN1.

In mammals purinergic receptors or purinoreceptors P2X and P2Y are involved eATP signal transduction (Ralevic & Burnstock, 1998; Abbracchio, 2006). However, no probable candidates for an ATP receptor were detected by genomic sequence-based surveys for canonical P2X and P2Y receptors in plants (Tanaka, Gilroy, Jones & Stacey, 2010). This is because DORN1 has a significantly different molecular structure than known animal ATP receptor. As a result Choi et al. proposed the term P2K (plant receptor kinase) for a new plant specific purinoreceptor subfamily of which DORN1 is the founding member.



Allele	Types of mutation	Base pair change	Amino acid change	Domain
dorn1-1	EMS	GAC to AAC	D572N	Kinase domain X
dorn1-2	EMS	GAT to AAT	D525N	Kinase domain IX, DxxWxG motif
dorn1-3	T-DNA (Salk_042209)	AA*G	K92	Extracellular domain
dorn1-4	T-DNA (Salk_024581)	GG*A	G189	Extracellular domain
dorn1-5	EMS	GAT to AAT	D486N	Kinase domain VII Activation loop
dorn1-6	EMS	GAT to AAT	D147N	Extracellular domain
dorn1-7	EMS	GCA to ACA	A306T	Transmembrane domain
dorn1-8	EMS	CGA to TGA	R411 to STOP	Kinase domain V
dorn1-9	EMS	CGA to CAA	R467Q	Kinase domain VIb
dorn1-10	EMS	CGA to TGA	R262 to STOP	Juxta-membrane region

Figure 1.6 Mutant alleles in DORN1 with loss-of-function in Arabidopsis (Choi et al., 2014b)

1.5 L-type lectin receptor kinase (LecRK):

The lectin-like receptor kinases (LecRKs) are a class of RKs that contain a lectin ectodomain. LecRKs were further classified into three categories, the G-, L-, and C-type lecRLKs, based on the type of lectin domain they contain (Bouwmeester & Govers, 2009) [Figure 1.7]. Legume-like or L-type lectin receptor kinases (LecRKs) contain a characteristic legume lectin ectodomain. According to Barre et al., 2002 the legume lectin domain consists of a flat six-stranded β-sheet (back face) and a curved seven-stranded β-sheet (front face) interconnected by turns and loops. The two beta sheets form what is referred to as the beta-sandwich fold, the hydrophobic core of the protein [Figure 1.8]. The domain also contains an additional loop of 17 residues rich in glycine and proline that connect beta strands 11 and 12.

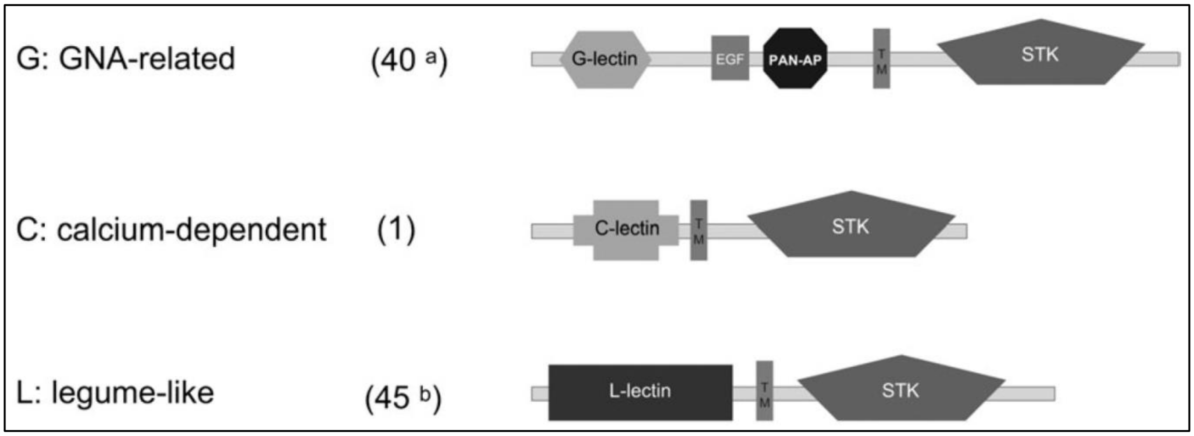


Figure 1.7: Organization and composition of domain of LecRKs (Bouwmeester & Govers, 2009)

Plant L-type lectins are ubiquitous in leguminous seeds as soluble proteins and are involved in binding monosaccharides (Lagarda-Diaz, Guzman-Partida & Vazquez-Moreno, 2017). Because of their sequence homology to legume lectins, it was hypothesized that L-type LecRKs could be involved in the recognition and transduction of saccharidic signals (André, Siebert, Nishiguchi, Tazaki & Gabius, 2005). However, molecular modeling of Arabidopsis L-type LecRKs revealed a poor conservation of the sugar binding residues and the amino acids involved in the Ca^{2+} and Mn^{2+} binding which are important for sugar binding, whereas it appeared that the hydrophobic binding site is better conserved. So L-type LecRKs are unlikely to bind to monosaccharides and instead bind to hydrophobic ligands and/or MAMPs/DAMPs (Barre et al., 2002; André et al., 2005).

Lastly, LecRKs have shown to play important roles in plant development and stress responses including defenses against pathogens and insect pests (Wang, Nsibo, Juhar, Govers & Bouwmeester, 2016; Chen et al., 2006), responses to abiotic stresses such as salt, drought, wounding, or extreme temperature (He, Zhang, Yan, Zhang & Chen, 2004; Vaid, Macovei & Tuteja, 2013), seed germination (Cheng et al., 2013), lateral root development (Deb, Sankaranarayanan, Wewala, Widdup & Samuel, 2014), pollen development (Wan et al., 2008) and hormone signaling (Deng et al., 2009).

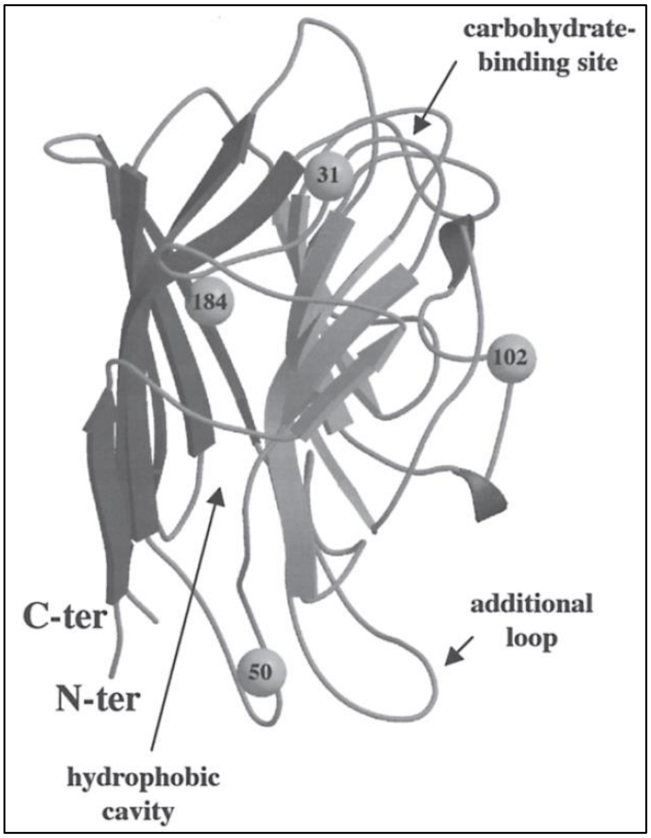


Figure 1.8 Ribbon diagram of a typical lectin domain of LecRK (Barre et al., 2002)

1.6 eATP as a DAMP:

Adenosine 5-triphosphate (ATP) serves as the universal energy source in all organisms. Inside the cell ATP is maintained at a very high concentration (mM). In plants eATP acts as a DAMP when ATP is released from the inside the cell into the extracellular matrix after wounding, pathogen-induced cell lysis as well as in response to various stress agents. These include MAMPs such as yeast extract (Wu & Wu, 2008), chitin (Kim, Sivaguru & Stacey, 2006), and plant stress hormone abcisic acid (ABA) (Dark, Demidchik, Richards, Shabala & Davies, 2011). Choi et al. showed that the expression of 332 genes was upregulated in *Arabidopsis* in response to the addition of ATP but none of these genes responded in dorn1 mutant plants. Among these 332 genes stress responsive genes as well as signal transduction genes were overrepresented. Exogenous application of ATP trigger signaling pathways characteristic of response to MAMPs such as elevation of cytoplasmic Ca^{2+} concentration (Tanaka, Swanson, Gilroy & Stacey, 2010), production of nitric oxide (NO) (Foresi, Laxalt, Tonon, Casalongue & Lamattina, 2007), reactive oxygen species (ROS) (Song, Steinebrunner, Wang, Stout & Roux, 2006), phosphatidic acid (Sueldo, Foresi, Casalongué, Lamattina & Laxalt, 2010), activation of mitogen-activated protein kinase (MPK) (Choi et al., 2014a). Comparison of the ATP induced gene lists to genes that were induced by wounding showed a 60% overlap. The majority (90%) of these overlapping genes responded very early to wounding (Choi et al., 2014a). As an example 112 genes out of these 332 genes were up-regulated only 15 minutes after wounding (Choi et al., 2014a). Moreover the expression level of DORN1 correlated with the expression level of those genes induced by ATP and wounding.

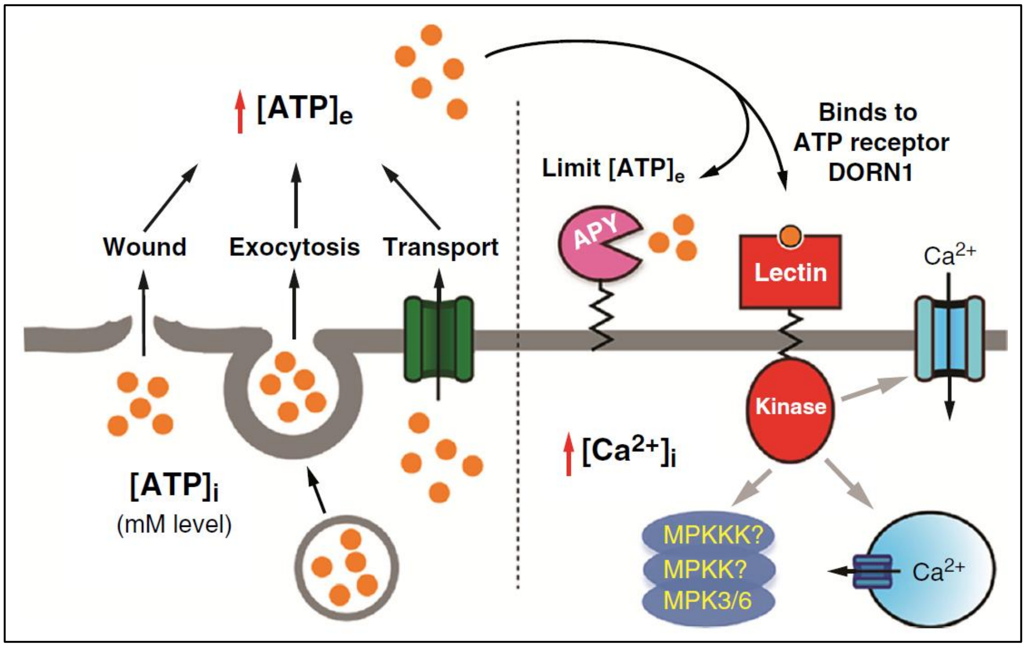


Figure 1.9 Overview of eATP signaling in plants** (**Cao, Tanaka, Nguyen & Stacey, 2014)

Studies have shown that the function of extracellular ATP in regulating plant immune responses seemed to be dose-dependent, either addition or depletion of eATP can trigger plant defense responses *(Choi et al., 2014b)*. Treatment with exogenous apyrase, ATP hydrolyzing enzymes, induced defense-related genes in plants (Chivasa et al., 2009a). Suppression of apyrases in Arabidopsis resulted in increased levels of eATP and elevated expression of stress related genes (Lim et al., 2014). Exposure to high levels of eATP triggered programmed cell death associated cellular responses in *Populouseuphratica* (SUN et al., 2011). On the other hand extreme depletion of eATP had shown to cause cell death in plant species such as *Arabidopsis*, bean, maize and tobacco (Chivasa, Ndimba, Simon, Lindsey & Slabas, 2005).

eATP action as a DAMP have also proved to be time-dependent. Long-term exposure to exogenous ATP has shown to suppress hypersensitive cell death caused by pathogenic mycotoxin (Chivasa, Ndimba, Simon, Lindsey & Slabas, 2005), reduce the production of salicylic acid, decrease pathogenic (PR) gene expression and increase resistance to bacterial pathogens and tobacco mosaic virus (Chivasa et al., 2009a, 2009b). Short-term (minutes to several hours) treatment with ATP induced defense related gene expression and ROS production in *Arabidopsis* (Song, Steinebrunner, Wang, Stout & Roux, 2006).

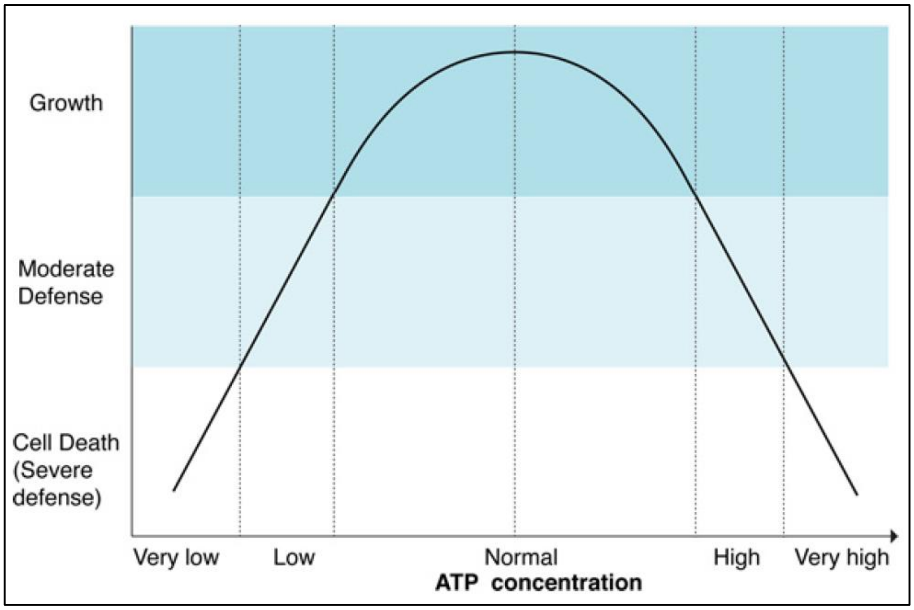


Figure 1.10 A model of the effect of eATP on plant immune system (Choi et al., 2014b)

Chapter 2: Computational Modeling Techniques

2.1 Protein Modeling Methods:

There are three major approaches to predicting a protein's tertiary structure:

1. Homology / comparative modeling
2. Threading / fold recognition
3. Ab-initio / de-novo modeling

Both Homology modeling and threading are template based modeling methods unlike ab-initio modeling.

Homology modeling exploits the fact that evolutionarily related proteins have similar sequences and often similar structures (Kaczanowski & Zielenkiewicz, 2009). It has been shown that protein tertiary structures are more conserved than their respective amino acid sequences among homologues given that sequence identity does not fall below 20% (Chothia & Lesk, 1986). Therefore by extrapolating experimental information from an evolutionary related protein structure that serves as a template, a 3D protein model of a target sequence is generated in comparative modeling. The prediction process consists of fold assignment, target-template alignment, model building, and model evaluation.

In template-based modeling the critical step is to identify correct template proteins from the PDB library which have similar folds as the query protein and to make correct alignment between the template structure and the query sequence. This process is known as fold recognition or threading. Threading is used to model targets that have the same fold as proteins with known structures, but lack homologues (Bowie, Luthy & Eisenberg, 1991). Protein threading is based on the observation that there are only a limited number of unique protein folds in nature (approximately 1300). While homology modeling uses only sequence homology between target and template for prediction, threading uses both sequence and structure information extracted from the target-template alignment for prediction. There are several threading algorithms in literature based on various approaches, for example the sequence profile–profile alignment, structural profile alignment, hidden Markov models, machine learning, and pair-wise potentials with optimal searching (Wu & Zhang, 2008).

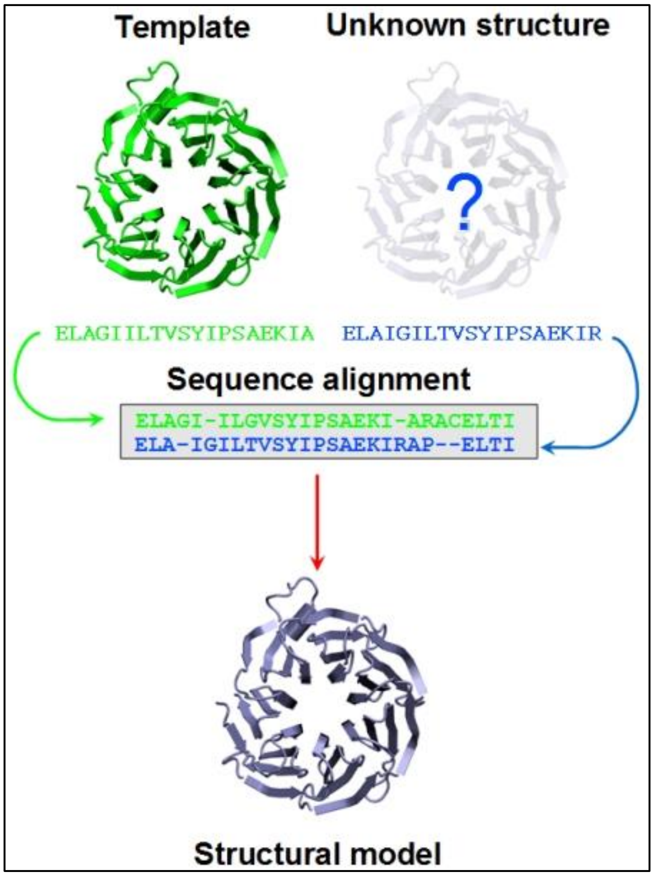


Figure 2.1 Principles of Homology modeling

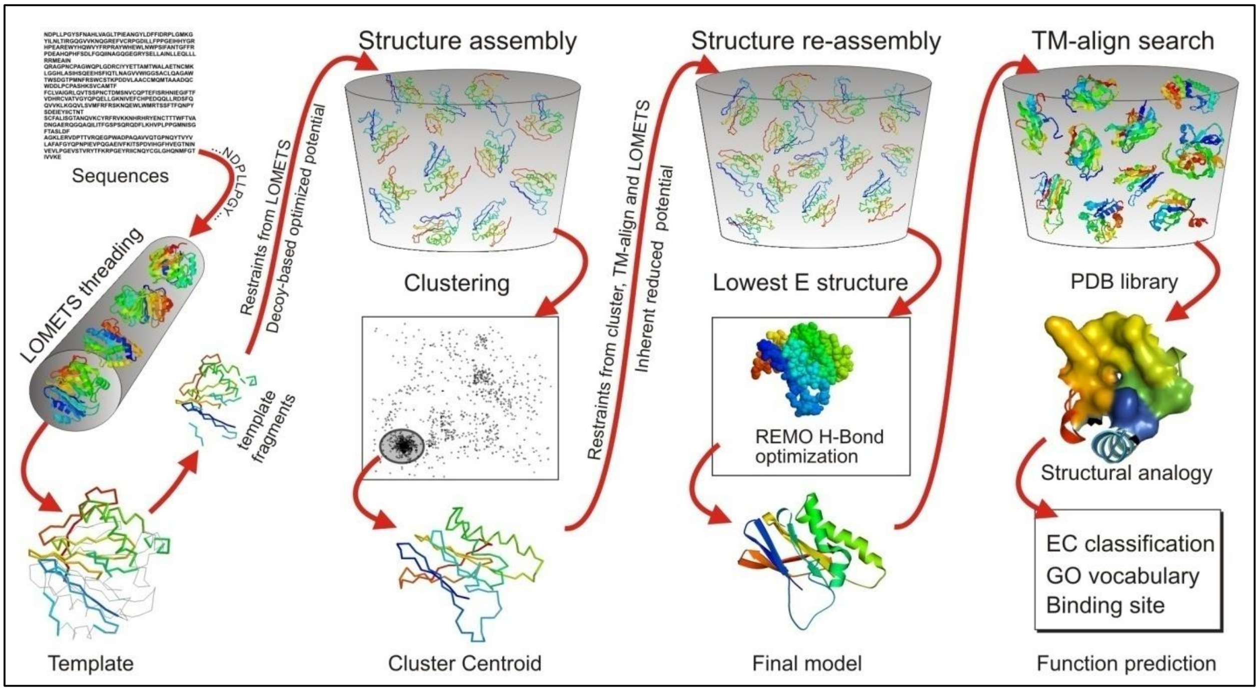


Figure 2.2 Steps of threading modeling method by I-TASSER tool

Ab initio modeling starts with the primary amino acid sequence which is searched for different conformations under the guidance of a designed energy function. This generates a number of possible conformations also called decoy structures, from which native-like conformations are selected based on their thermodynamic stability and energy state (Lee, Freddolino & Zhang, 2008). This approach is based on the thermodynamic hypothesis proposed by Anfinsen, according to which, under a given set of conditions, the native structure is the one that corresponds to the global free energy minimum.

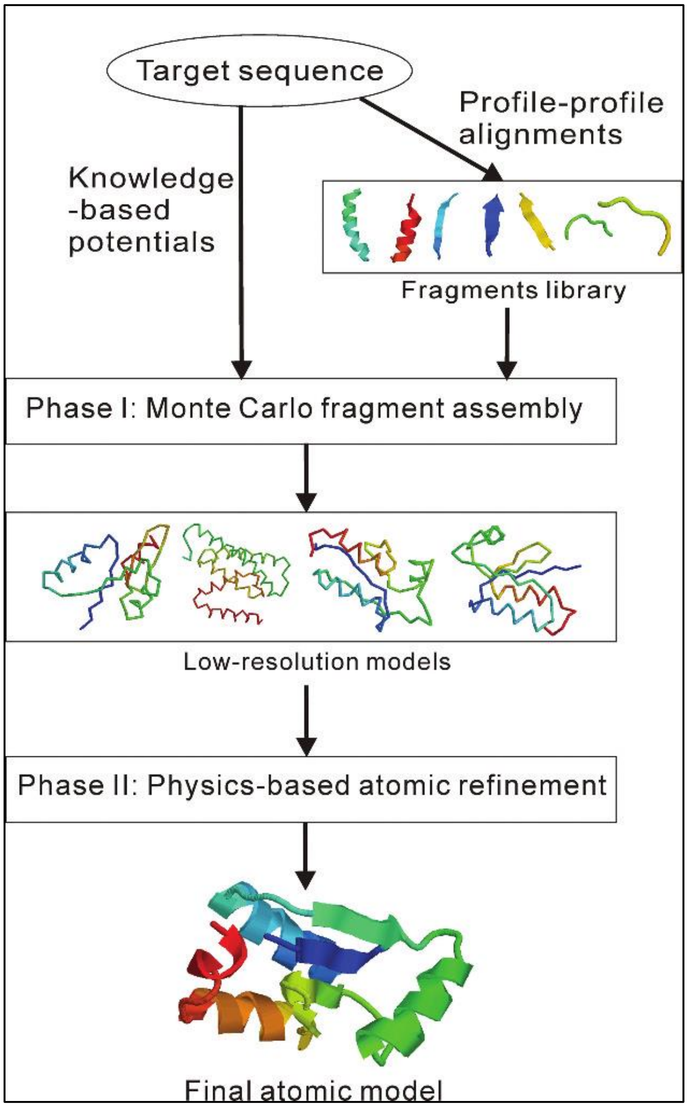

Target sequence
Profile-profile alignments
Knowledge -based potentials
Fragments library
Phase I: Monte Carlo fragment assembly
Low-resolution models
Phase II: Physics-based atomic refinement
Final atomic model

Figure 2.3 Flowchart of Ab-initio modeling steps of Rosetta tool (Lee, Freddolino & Zhang, 2008)

Table 1: Tools used in this study and their corresponding modeling technique

Tool	Method of Prediction
I-Tasser	Threading, Ab initio
Phyre2	Homology modeling
HHpred	Homology modeling
IntFOLD5	Threading
AIDA	Ab initio
FFAS03 Modeller	Homology modeling
Muster	Threading
(PS)2-v2	Threading
Raptor X	Threading
Spark X	Threading
Swiss Model	Homology modeling

2.2 Significance of using computational approaches:

In recent years, the number of known protein sequences has increased exponentially with the success of an expanding array of genome sequencing projects. However the functional characterization of a protein requires its 3D structure. As of 2019 there are 147 million protein sequences in UniProtKB/TrEMBL and only 140,393 structures in RSCB PDB. Such a wide gap is due to the inherently time-consuming and complicated nature of traditional structure determination techniques X-ray crystallography, NMR spectroscopy which limits their utility. The use of computational approaches can thus help to bridge the gap between the number of known sequences and the number of 3D models available. There are mainly two major classifications: single template modeling (STM) and multiple template modeling (MTM). STMs utilize only one template or reference structure whereas MTMs utilize more than one template or reference structure to compare the query amino acid sequence with, and build 3D models with respect to the query amino acid sequence.

2.3 Benefits of studying DORN1:

It has been shown that immune signaling pathways downstream of PRRs are conserved across plant families and thus allow for the functional transfer of PRRs between monocots and dicots (Holton, Nekrasov, Ronald & Zipfel, 2015). For e.g the expression of Arabidopsis eATP receptor DORN1/LecRK-I.9 in *N. benthamiana* and potato *(Solanum tuberosum*) enhances resistance to the oomycete *Phytophthora infestans* (Bouwmeester et al., 2013). Similarly, transfer of the rice PRR XA21 into sweet orange (*Citrus x sinensis*), tomato, and banana resulted in enhanced resistance to *X. axonopodis pv. citri*, *Ralstonia solanacearum*, and *Xanthomonas campestris pv. musacearum*, respectively (Mendes et al., 2010; Afroz et al., 2010; Tripathi, Lorenzen, Bahar, Ronald & Tripathi, 2014).

It now seems likely that eATP is a key missing signal in many plant stress responses that were previously attributed to a direct effect of the stress stimulus. Further investigation of the role of eATP requires knowledge of the tertiary structure of DORN1. This can not only reveal new details of how plants respond to stress, but offer a novel biotechnological approach to engineer disease-resistant crops leading to sustainable agriculture and increased crop yields, crucial goals in lieu of a rapidly changing climate and ever expanding global population.

Chapter 3: Methodology

Amino acid sequence (FASTA format) of DORN1 ectodomain was retrieved from UniProtKB database. UniProt allowed selection of only the legume lectin domain of the protein.

```
>sp|Q9LSR8|24-259

ETSFVYESFLDRQNLYLDKSAIVLPSGLLQLTNASEHQMGHAFHKKPIEFSSSGPLSFST

HFVCALVPKPGFEGGHGIVFVLSPSMDFTHAESTRYLGIFNASTNGSSSYHVLAVELDTI

WNPDFKDIDHNHVGIDVNSPISVAIASASYYSDMKGSNESINLLSGNPIQVWVDYEGTLL

NVSVAPLEVQKPTRPLLSHPINLTELFPNRSSLFAGFSAATGTAISDQYILWWSFS
```

The FASTA format was then inputted into the following tools:

- ProtPram for determination of amino acid composition and physiochemical properties
- InterPro for domain architecture analysis of DORN1
- PSIPRED for prediction of secondary structure
- ConSurf tool for prediction of conserved regions

In order to determine homologous protein sequences the FASTA format was inputted into NCBI-Blastp. The database chosen was the protein databank proteins (PDB). Based on max score the top 10 sequences were selected.

The tool MUSCLE was then utilized to carry our multiple sequence alignment of the top 10 sequences. In order to visualize the output from MUSCLE the MSA was copied and pasted in the BoxShade server with the output format set as postscript portrait.

This was followed by the construction of the maximum likelyhood phylogenetic tree using MEGA. The substitution model used was WAG model as it was determined to be the best suited model for maximum likelihood tree construction using protein sequences.

Single template modeling of query sequence was done using AIDA, FFAS03 Modeller, Muster, (PS)2-v2, Raptor X, Spark X and Swiss Model. For both FFAS03 and Swiss Model the templates based on which the models were to be built were manually chosen from a list of top templates provided by the tools.

Multiple template modeling of query sequence was done using I-Tasser, IntFOLD5, Phyre2 and HHpred Modeller. The templates for the generation of models were manually chosen only for HHPred.

Finally all the models generated were structurally validated using Veriify3D, ERRAT and Rampage Ramachandran Plot.

Chapter 4: Databases and Tools Utilized

4.1 Databases:

PDB:

The Protein Data Bank (PDB) was established as the 1st open access digital data resource in all of biology and medicine. Through an internet information portal and downloadable data archive, the PDB provides access to 3D structure data for large biological molecules (proteins, DNA, and RNA (Berman et al., 2000). The data, is typically obtained by X-ray crystallography, NMR spectroscopy, or cryo-electron microscopy, and submitted by biologists and biochemists from around the world. The primary information stored in the PDB archive consists of coordinate files for biological molecules. These files list the atoms in each protein, and their 3D location in space. These files are available in several formats (PDB, mmCIF, XML). Each structure published in PDB receives a four-character alphanumeric identifier, its PDB ID (Berman et al., 2000). Many other databases also utilize protein structures deposited in the PDB such as SCOP and CATH databases.

UniProt:

The Universal Protein Resource (UniProt)is a comprehensive resource for protein sequence and annotation data (The UniProt Consortium, 2018). The UniProt databases are the UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters (UniRef), and the UniProt Archive (UniParc) (The UniProt Consortium, 2018) [Figure 4.1]. UniProt is a collaboration between the European Bioinformatics Institute (EMBL-EBI), the SIB Swiss Institute of Bioinformatics and the Protein Information Resource (PIR). EMBL-EBI and SIB together produced Swiss-Prot and TrEMBL (Translated EMBL Nucleotide Sequence Data Library), while PIR produced the

Protein Sequence Database (PIR-PSD). In 2002 the three institutes decided to pool their resources and expertise and formed the UniProt consortium.

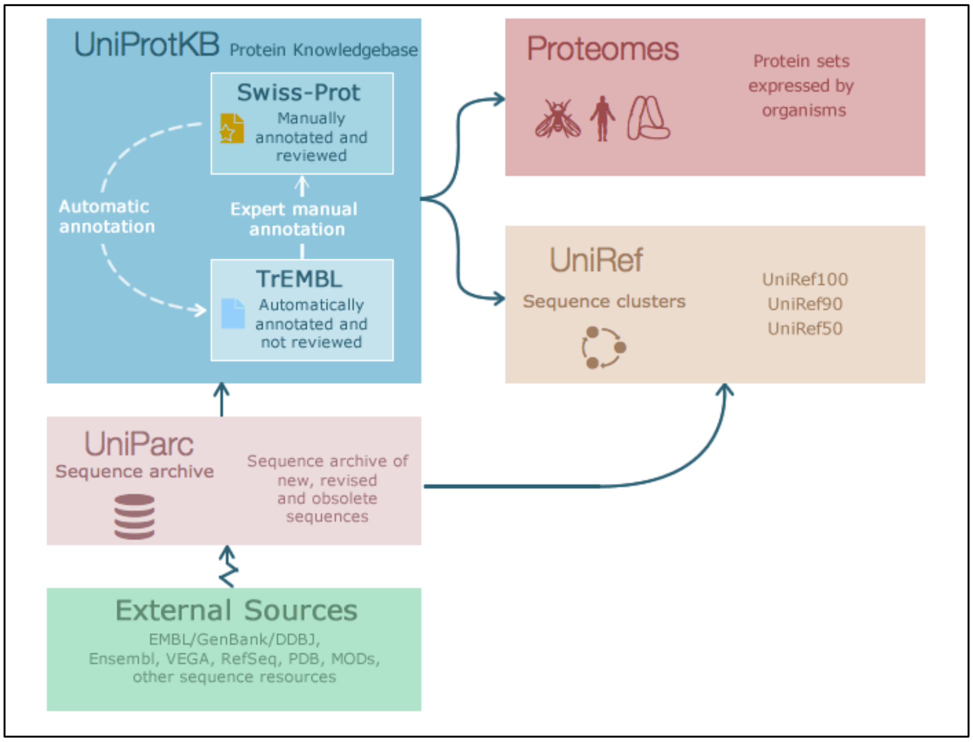

UniProtKB Protein Knowledgebase
Swiss-Prot
Manually annotated and reviewed
Automatic annotation
Expert manual annotation
TrEMBL
Automatically annotated and not reviewed
Proteomes
Protein sets expressed by organisms
UniRef
Sequence clusters
UniRef100
UniRef90
UniRef50
UniParc
Sequence archive
Sequence archive of new, revised and obsolete sequences
External Sources
EMBL/GenBank/DDBJ, Ensembl, VEGA, RefSeq, PDB, MODs, other sequence resources

Figure 4.1 Overview of Uniprot database

NCBI:

The National Center for Biotechnology Information (NCBI) supports a variety of databases for genes, genome, proteins and chemicals that are relevant to the medical and scientific communities and are an important resource for bioinformatics tools and services. NCBI also provides access to a wide variety of data analysis tools that allow users to manipulate, align, visualize and evaluate biological data. In addition NCBI's literature resources include the world's largest repository of medical and scientific abstracts, full-text articles, books and reports.

4.2 Tools and Servers:

NCBI BLAST:

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences (Altschul, Gish, Miller, Myers & Lipman, 1990). The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families. Different types of BLASTs are available according to the query sequence input, database being searched and what is being compared such as:

- Nucleotide-nucleotide BLAST (blastn),
- Protein-protein BLAST (blastp),
- Position-Specific Iterative BLAST (PSI-BLAST) (blastpgp)

ProtPram:

ProtParam is a tool which allows the computation of various physical and chemical parameters for a given protein stored in Swiss-Prot or TrEMBL or for a user entered protein sequence (Gasteiger et al., 2005).The computed parameters include the molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index and grand average of hydropathicity (GRAVY).

PSIPRED:

PSI-blast based secondary structure PREDiction (PSIPRED) is a protein structure prediction method. Incorporating two feed-forward neural networks, PSIPRED performs analysis on output obtained from PSI-BLAST (Position Specific Iterated - BLAST) (Jones, 1999). It can predict a protein's secondary structure (beta sheets, alpha helixes and coils) from the primary sequence. The idea of this method is to use evolutionary-related protein information to predict a new amino acid sequence's secondary structure. PSI-BLAST is used to identify related sequences and to create a position-specific scoring matrix. This matrix is processed by an artificial neural network (Chen, 2014) that was built and trained to predict the input sequence's secondary structure; in other words, it is a machine learning method.

ConSurf:

The ConSurf server is a bioinformatics tool for estimating the evolutionary conservation of amino/nucleic acid positions in a protein/DNA/RNA molecule based on the phylogenetic relations between homologous sequences (Landau et al., 2005; Ashkenazy et al., 2016). The degree to which an amino (or nucleic) acid position is evolutionarily conserved (i.e., its evolutionary rate) is strongly dependent on its structural and functional importance. Thus, conservation analysis of positions among members from the same family can often reveal the importance of each position for the protein (or nucleic acid)'s structure or function (Glaser et al., 2003). In ConSurf, the evolutionary rate is estimated based on the evolutionary relatedness between the protein (DNA/RNA) and its homologues and considering the similarity between amino (nucleic) acids as reflected in the substitutions matrix (Pupko, Bell, Mayrose, Glaser & Ben-Tal, 2002). One of the advantages of ConSurf in comparison to other methods is the accurate computation of the evolutionary rate by using either an empirical Bayesian method or a maximum likelihood (ML) method (Pupko, Bell, Mayrose, Glaser & Ben-Tal, 2002).

MUSCLE:

MUltiple Sequence Comparison by Log-Expectation (MUSCLE) is computer tool for multiple sequence alignment of protein and nucleotide sequences. MUSCLE is claimed to achieve both better average accuracy and better speed than ClustalW2 or T-Coffee, depending on the chosen options. Multiple alignments of protein sequences are important in many applications, including phylogenetic tree estimation, secondary structure prediction and critical residue identification.

BoxShade:

Boxshade is a program for creating visually pleasing images of protein or DNA multiple sequence alignments. The program does no alignment by itself, it has to take as input a file preprocessed by a multiple alignment program or a multiple file editor. In the standard BOXSHADE output, identical and similar residues in the multiple-alignment chart are represented by different colors or shadings. There are some options concerning the kind of shading to be applied, sequence numbering, consensus output and so on.

MEGA X:

Molecular Evolutionary Genetics Analysis (MEGA) is computer program package for conducting statistical analysis of molecular data, estimating evolutionary distances and for constructing phylogenetic trees (Kumar, Stecher, Li, Knyaz & Tamura, 2018). MEGA X can be used to construct maximum likelihood tree, neighbor-joining tree, UPGMA tree, maximum parsimony tree and minimum evolution tree. It allows sequence alignment by both MUSCLE and CLUSTALW.

AIDA:

Ab initio domain assembly (AIDA) server is a tool capable of identifying individual domains in multidomain proteins and then assembling their 3D structures and predicting their relative spatial arrangements guided by the ab-initio folding potential (Xu, Jaroszewski, Li & Godzik, 2015). The AIDA server allows the assembly of domains inserted into other domains (discontinuous domains) by fixing the relative positions of the domains. The server also supports structure assembly from sequence only. In order to do this the sequence domains are iteratively split and aligned with the PDB template found by the FFAS-3D fold recognition program (Xu, Jaroszewski, Li & Godzik, 2015). Additionally, AIDA supports restraint-guided simulation, allowing the user to specify inter-domain distance restraints that guide the AIDA energy minimization (Xu, Jaroszewski, Li & Godzik, 2015)

FFAS03:

The FFAS03 server provides an interface to the profile-profile alignment and fold recognition algorithm FFAS (Rychlewski, Li, Jaroszewski & Godzik, 2008). A profile-profile alignment utilizes information present in sequences of homologous proteins to amplify the sequence conservation pattern defining the protein family. This method allows detection of remote homologies beyond the reach of other sequence comparison methods. The FFAS03 server accepts a user supplied protein sequence and automatically generates a profile, which is then compared with several sets of sequence profiles of proteins from PDB, COG, PFAM and SCOP (Jaroszewski, Rychlewski, Li, Li & Godzik, 2005). In addition to homologs detected in profile database(s) by the FFAS method, the FFAS03 server also displays homologs collected by PSI-BLAST using PDB-BLAST protocol and templates detected with BLAST method

Modeller:

Modeller is a computer program used for homology modeling to produce models of protein tertiary structures (Webb & Sali, 2016). It implements a method inspired by protein NMR (nuclear magnetic resonance spectroscopy) called satisfaction of spatial restraints, by which a set of geometrical criteria are utilized to create a probability density function for the location of each atom in the protein (Šali & Blundell, 1993). The method relies as input on a sequence alignment between the query amino acid sequence which is to be modeled and a template protein for which structure has been resolved.

Muster:

MUSTER (MUlti-Sources ThreadER) is a protein threading algorithm that identifies template structures from the PDB library using sequence profile–profile alignment method, PPA. It generates sequence-template alignments by combining various sequence and structure information into single-body terms which can be conveniently used in dynamic programming search: (1) sequence profiles; (2) secondary structures; (3) structure fragment profiles; (4) solvent accessibility; (5) dihedral torsion angles; (6) hydrophobic scoring matrix (Wu & Zhang, 2008). The output of the MUSTER server include:

- Top five template proteins and the query-template alignments
- Full-length models built by MODELLER

(PS)2-v2:

(PS)2-v2 is an automatic homology modeling server. The method uses a new substitution matrix, S2A2, that combines both sequence and secondary structure information for the detection of homologous proteins with remote similarity and the target-template alignment (Chen, Hwang & Yang, 2009). The final three dimensional structure is built using the modeling package MODELLER (Chen, Hwang & Yang, 2006). After generating a predicted model, the programs ProQ and ProQres are used to evaluate the quality of this model based on the LGscore and MaxSub scores(Chen, Hwang & Yang, 2009). Finally, the predicted model is displayed by AstexViewer and automatically sent to users.

Raptor X:

RaptorX is a protein structure prediction server developed by Xu group, excelling at predicting 3D structures for protein sequences without close homologs in the Protein Data Bank (PDB) (Källberg et al., 2012). Given an input sequence, RaptorX predicts its secondary and tertiary structures, contacts, solvent accessibility, disordered regions and binding sites (Källberg et al., 2012). RaptorX also assigns some confidence scores to indicate the quality of a predicted 3D model: P-value for the relative global quality, GDT (global distance test) and uGDT (un-normalized GDT) for the absolute global quality, and modeling error at each residue (Källberg et al., 2012). RaptorX excels at the alignment of hard targets, which have less than 30% sequence identity with solved structures in PDB.

Spark X:

Is a fold recognition server which combines single fold recognition method SPARKS which is based on weighted matching of multiple profiles with SPINEX techniques for improving prediction of secondary structure, backbone torsion angle and solvent accessible surface area and 3D model generation (Yang, Faraggi, Zhao & Zhou, 2011). It recognizes a match in the same family, same superfamily and same fold according to the SCOP definition within top-N templates.

Swiss Model:

SWISS-MODEL is a fully automated protein structure homology-modeling server, accessible via the ExPASy web server, or from the program DeepView (Swiss Pdb-Viewer). Through its website SWISS-MODEL support several inputs including only query amino acid sequence, target-template alignment file and user defined template. For a given query protein sequence, a library of experimental protein structures is searched to identify suitable templates. Based on a sequence alignment between the target protein and the template structure, a three-dimensional model for the target protein is generated (Waterhouse et al., 2018). Model quality assessment tools are used to estimate the confidence of the resulting models (Waterhouse et al., 2018).

I-Tasser:

I-TASSER (Iterative Threading ASSEmbly Refinement) is a hierarchical approach to protein structure and function prediction (Yang & Zhang, 2015). It first identifies structural templates from the PDB by multiple threading approach LOMETS, with full-length atomic models constructed by iterative template-based fragment assembly simulations (Yang & Zhang, 2015). Function insights including ligand-binding sites, enzyme commission number, and gene ontology terms of the target are then derived by threading the 3D models through protein function database BioLiP (Yang & Zhang, 2015). I-TASSER (as 'Zhang-Server') was ranked as the No 1 server for protein structure prediction in community-wide CASP experiments.

IntFOLD5:

IntFOLD is an independent, integrated web server for protein structure and function prediction (McGuffin et al., 2019). The only input required is the amino acid sequence for the target protein. The server provides the following (McGuffin et al., 2019):

- Tertiary structure prediction/3D modeling
- 3D model quality estimates - with the option to refine/fix errors
- Intrinsic disorder prediction
- Domain prediction
- Prediction of protein-ligand binding residues

Phyre2:

Phyre2 is a suite of tools available on the web to predict and analyse protein structure, function and mutations (Kelley, Mezulis, Yates, Wass & Sternberg, 2015). Phyre2 servers use techniques of homology modeling to predict the three-dimensional structure of a protein sequence (Kelley, Mezulis, Yates, Wass & Sternberg, 2015). The Phyre2 server which serves as a replacement for the original Phyre server provides extra functionality over Phyre, a more advanced interface, fully updated fold library and uses the HHpred / HHsearch package for homology detection. Additionally it provides the following functionalities:

- Batch processing
- One to one threading
- Multi-template modeling

HHPRED:

Part of the HH-suite, HHpred is a server for protein structure prediction that uses homology information from HH-suite. The HH-suite searches for sequences using hidden Markov models (HMMs) (Soding, Biegert & Lupas, 2005). It allows searching a wide range of databases, such as the PDB, SCOP, Pfam, SMART, COGs and CDD. It accepts a single query sequence or a multiple alignment as input. It returns the search results in a format similar to that of PSI-BLAST. Specific parameters can be included during search such as local or global alignment and scoring secondary structure similarity. HHpred can produce pairwise query-template alignments, multiple alignments of the query with a set of templates selected from the search results, as well as 3D structural models generated by the MODELLER software from the aforementioned alignments (Soding, Biegert & Lupas, 2005).

UCSF Chimera:

UCSF Chimera is a highly extensible program for interactive visualization and analysis of molecular structures and related data, including density maps, supramolecular assemblies, sequence alignments, docking results, trajectories, and conformational ensembles (Pettersen et al., 2004). High-quality images and animations can be generated. Chimera was developed by the Resource for Biocomputing, Visualization, and Informatics (RBVI), supported in part by the National Institutes of Health.

Verify3D:

Verify3D is an online model assessment tool that determines the compatibility of a three dimensional atomic model with its own amino acid sequence by assigning a structural class based on its location and environment (alpha, beta, loop, polar, nonpolar etc.) and comparing the results to good structures (Lüthy, Bowie & Eisenberg, 1992).

ERRAT:

ERRAT is a web tool that analyzes the statistics of non-bonded interactions between different atom types and plots the value of the error function versus position of a 9-residue sliding window, calculated by a comparison with statistics from highly refined structures (Colovos & Yeates, 1993).

Rampage Ramachandran plot:

Rampage is a web tool that generates Ramachandran plots for any three dimensional model. It accepts pdb file as input. Ramachandran plot provides an easy way to visualize the distribution of torsion angles in a protein structure. Additionally it provides an overview of allowed and disallowed regions of torsion angle values, which serves as an important factor in the evaluation of the quality of three-dimensional structure of proteins.

Chapter 5: Results and Discussion

5.1 Analysis of physiochemical properties:

Analysis of various physical and chemical properties of DORN1 ectodomain was performed using ProtParam, which revealed that the DORN1 ectodomain consists of 236 AA and has a molecular weight of 25859.01 kDa. The amino acid composition of the protein showed that the most abundant amino acid is serine which accounts for 14% of the protein's primary structure whereas cysteine is the least common amino acid and makes up 0.4% of its primary structure. Containing a hydroxymethyl group in its side chain, serine is classified a polar amino acid. Serines are common in protein functional centers as the hydroxyl group is fairly reactive, being able to form hydrogen bonds with a variety of polar substrates. Cysteine residues are capable of formation of disulfide bonds which plays a role in stability and folding of the structure. The very low amounts of cysteine residues indicate that the protein gains its stability from other interactions, such a hydrophobic interactions, ionic bonds and hydrogen bonds as chances of disulfide bond formation are very low. In fact 45% of the legume lectin domain of DORN1 is composed of hydrophobic (non-polar) amino acids involved in hydrophobic interactions (van der Waals dispersion forces). Despite being relatively weak interactions, small stabilizing interactions can add up to make an important contribution to the overall stability of a protein. On the other hand charged amino acids make up 18.3% of the lectin domain and are responsible for formation of salt bridges, important for the stabilization of protein three-dimensional structure.

The isoelectric point (pI) was seen to be as 5.20 consistent with the fact that the total number of acidic (negatively charged) residues is greater than the total number of basic (positively charged) residues. The aliphatic index of a protein is defined as the relative volume occupied by aliphatic side chains (alanine, valine, isoleucine, and leucine). It may be regarded as a positive factor for the increase of thermostability of proteins. The aliphatic index of DORN1 ectodomain was 88.81

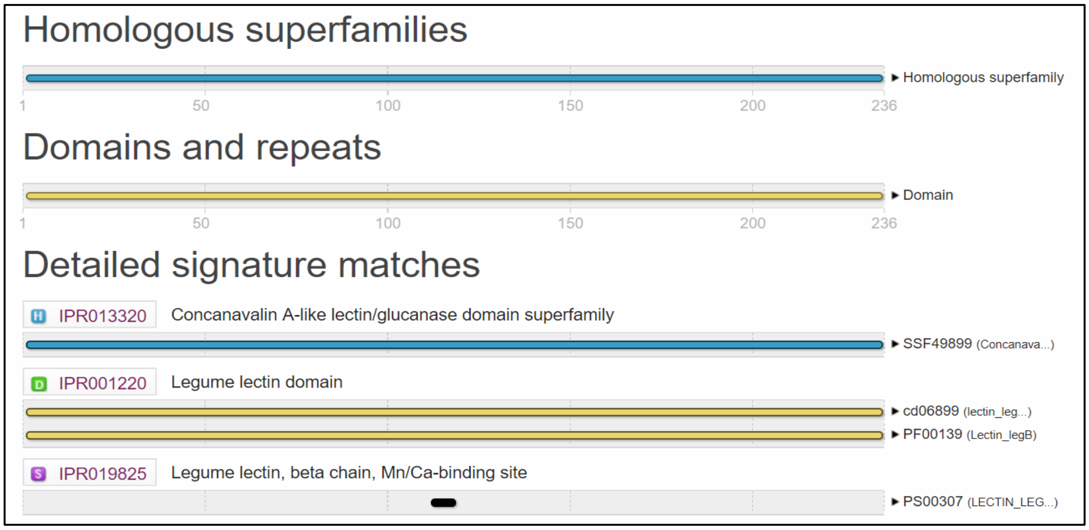
indicating the stability of the protein in a wide range of temperatures. The GRAVY value for a peptide or protein is calculated as the sum of hydropathy values of all the amino acids, divided by the number of residues in the sequence. A GRAVY value of 0.003 tells us that the protein is very much hydrophilic in nature.

Table 2: Results of ProtPram tool

<table>
<tr><td colspan="3"><b>Number of amino acids:</b> 236</td></tr>
<tr><td colspan="3"><b>Theoretical pI:</b> 5.20</td></tr>
<tr><td colspan="3"><b>Molecular weight:</b> 25859.01 kDa</td></tr>
<tr><td colspan="3"><b>Total number of negatively charged residues (Asp + Glu):</b> 22</td></tr>
<tr><td colspan="3"><b>Total number of positively charged residues (Arg + Lys):</b> 11</td></tr>
<tr><td colspan="3"><b>Aliphatic index:</b> 88.81</td></tr>
<tr><td colspan="3"><b>Grand average of hydropathicity (GRAVY)</b>: 0.003</td></tr>
<tr><td colspan="3"></td></tr>
<tr><td colspan="3"><b>Amino acid composition:</b></td></tr>
<tr><td>Residue</td><td>Number</td><td>Molecule %</td></tr>
<tr><td>Ala</td><td>15</td><td>6.4</td></tr>
<tr><td>Arg</td><td>4</td><td>1.7</td></tr>
<tr><td>Asn</td><td>13</td><td>5.5</td></tr>
<tr><td>Asp</td><td>11</td><td>4.7</td></tr>
<tr><td>Cys</td><td>1</td><td>0.4</td></tr>
<tr><td>Gln</td><td>6</td><td>2.5</td></tr>
<tr><td>Glu</td><td>11</td><td>4.7</td></tr>
<tr><td>Gly</td><td>15</td><td>6.4</td></tr>
<tr><td>His</td><td>10</td><td>4.2</td></tr>
<tr><td>Ile</td><td>14</td><td>5.9</td></tr>
<tr><td>Leu</td><td>24</td><td>10.2</td></tr>
<tr><td>Lys</td><td>7</td><td>3.0</td></tr>
<tr><td>Met</td><td>3</td><td>1.3</td></tr>
<tr><td>Phe</td><td>15</td><td>6.4</td></tr>
<tr><td>Pro</td><td>14</td><td>5.9</td></tr>
<tr><td>Ser</td><td>33</td><td>14.0</td></tr>
<tr><td>Thr</td><td>12</td><td>5.1</td></tr>
<tr><td>Trp</td><td>4</td><td>1.7</td></tr>
<tr><td>Tyr</td><td>8</td><td>3.4</td></tr>
<tr><td>Val</td><td>16</td><td>6.8</td></tr>
</table>

5.2 Domain architecture analysis:

Domain architecture analysis by InterPro showed that DORN1 ectodomain belongs to homologous concanavalin A-like lectin or glucanase domain superfamily. The lectins/glucanases are a diverse group of proteins found in a wide range of species from prokaryotes to humans. The different family members all contain a concanavalin A-like domain, which consists of a sandwich of 12-14 beta strands in two sheets with a complex topology ("Concanavalin A-like lectins/glucanases superfamily", 2019).



Homologous superfamilies

Domains and repeats

Detailed signature matches

IPR013320 Concanavalin A-like lectin/glucanase domain superfamily

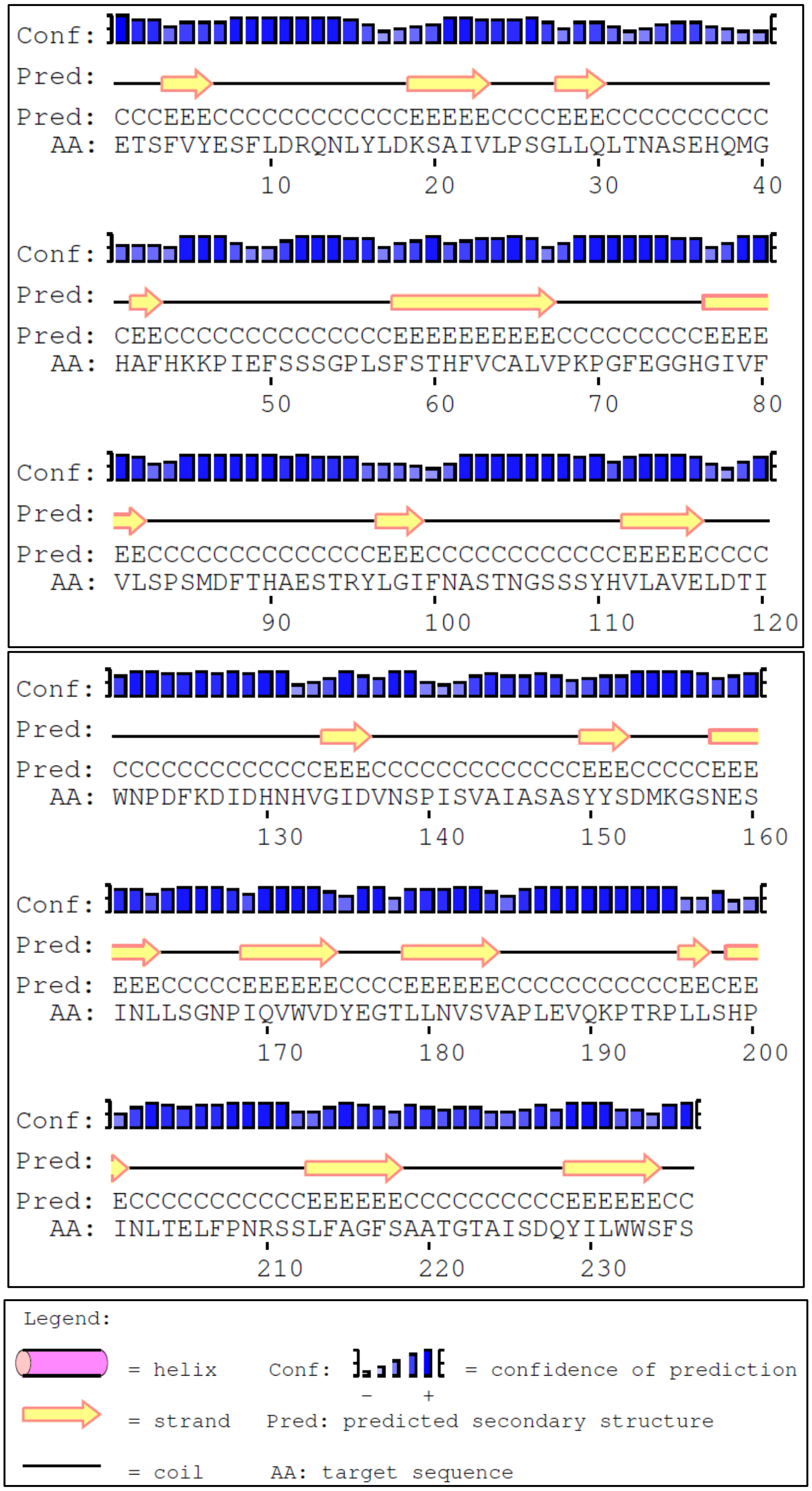
IPR001220 Legume lectin domain

IPR019825 Legume lectin, beta chain, Mn/Ca-binding site

Figure 5.1: Domain architecture analysis by Interpro

5.3: Secondary structure prediction:

The secondary structure predicted using PSIPRED showed that the DORN1 ectodomain is mainly composed of beta sheets (strands) and coils with the absence of any alpha helix. According to PSIPRED there are 17 beta strands and 18 coils. While all L-type lectins consist of beta-strands and coils, some also contain helixes such as lentil lectin which contains 2.



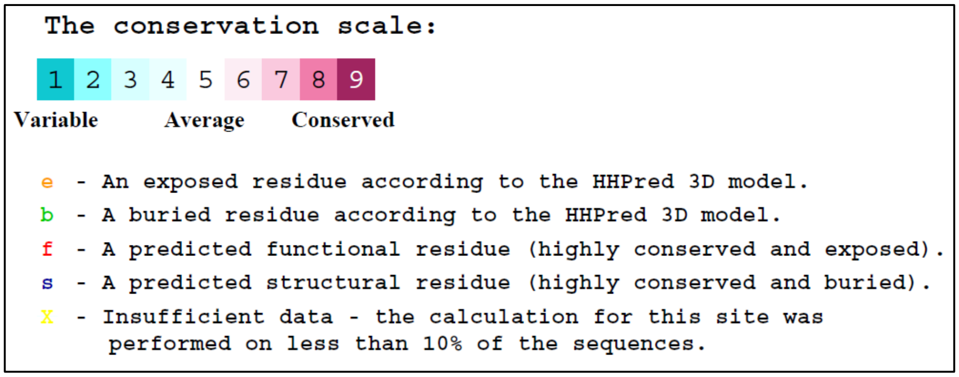
```
Conf:
Pred:
Pred: CCCEEECCCCCCCCCCCCEEEEECCCCEEECCCCCCCCCC
  AA: ETSFVYESFLDRQNLYLDKSAIVLPSGLLQLTNASEHQMG
                10        20        30        40

Conf:
Pred:
Pred: CEECCCCCCCCCCCCCCEEEEEEEEEECCCCCCCCCEEEE
  AA: HAFHKKPIEFSSSGPLSFSTHFVCALVPKPGFEGGHGIVF
                50        60        70        80

Conf:
Pred:
Pred: EECCCCCCCCCCCCCCEEECCCCCCCCCCCCEEEEECCCC
  AA: VLSPSMDFTHAESTRYLGIFNASTNGSSSYHVLAVELDTI
                90        100       110       120

Conf:
Pred:
Pred: CCCCCCCCCCCCCEEECCCCCCCCCCCCCEEECCCCCEEE
  AA: WNPDFKDIDHNHVGIDVNSPISVAIASASYYSDMKGSNES
                130       140       150       160

Conf:
Pred:
Pred: EEECCCCCEEEEEECCCCEEEEEECCCCCCCCCCCEECEE
  AA: INLLSGNPIQVWVDYEGTLLNVSVAPLEVQKPTRPLLSHP
                170       180       190       200

Conf:
Pred:
Pred: ECCCCCCCCCCCEEEEEECCCCCCCCCCEEEEEECC
  AA: INLTELFPNRSSLFAGFSAATGTAISDQYILWWSFS
                210       220       230
```

Legend:
= helix
Conf:
= confidence of prediction
-
+
= strand
Pred: predicted secondary structure
= coil
AA: target sequence

Figure 5.2: Secondary structure predicted by PSIPRED

5.4 Prediction of conserved regions:

The ConSurf tool was utilized to estimate the evolutionary conservation of residues in DORN1 ectodomain. According to the consurf results majority of the residues that placed high on the consuf conservation scale were buried residues, while most of the exposed residues of the protein placed low on the conservation scale indicating they are variable residues. Overall there is an even distribution of buried and exposed residues with 20 predicted functional residues and 24 predicted structural residues. Both functional and structural residues are considered to be highly conserved except while functional residues are usually exposed structural residues are buried. This is because the buried residues usually form hydrophobic cores to maintain the structural integrity of proteins while the exposed residues are hydrophilic and closely related to protein functions.



The conservation scale:

e - An exposed residue according to the HHPred 3D model.
b - A buried residue according to the HHPred 3D model.
f - A predicted functional residue (highly conserved and exposed).
s - A predicted structural residue (highly conserved and buried).
X - Insufficient data - the calculation for this site was performed on less than 10% of the sequences.

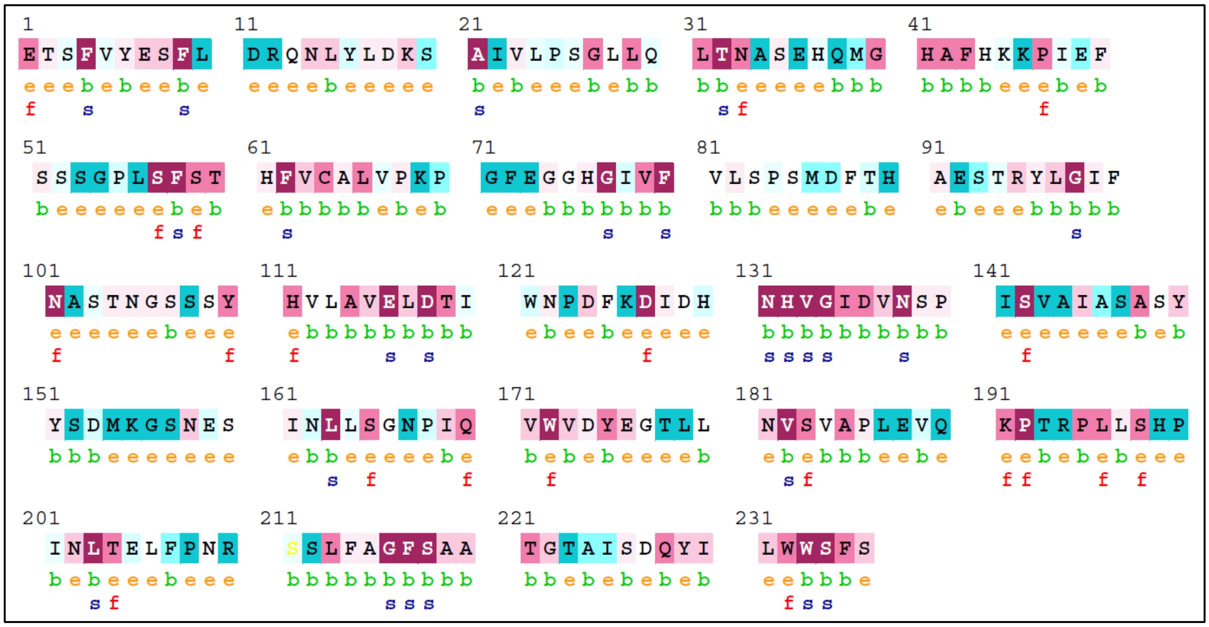


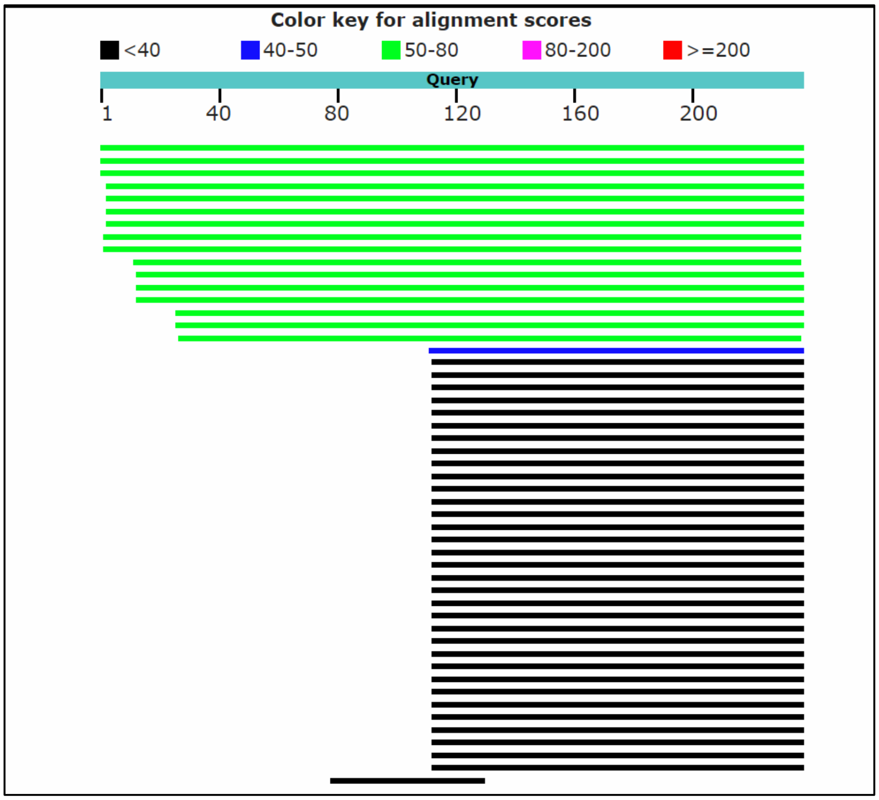
Figure: 5.3 Prediction of conserved regions by Consurf

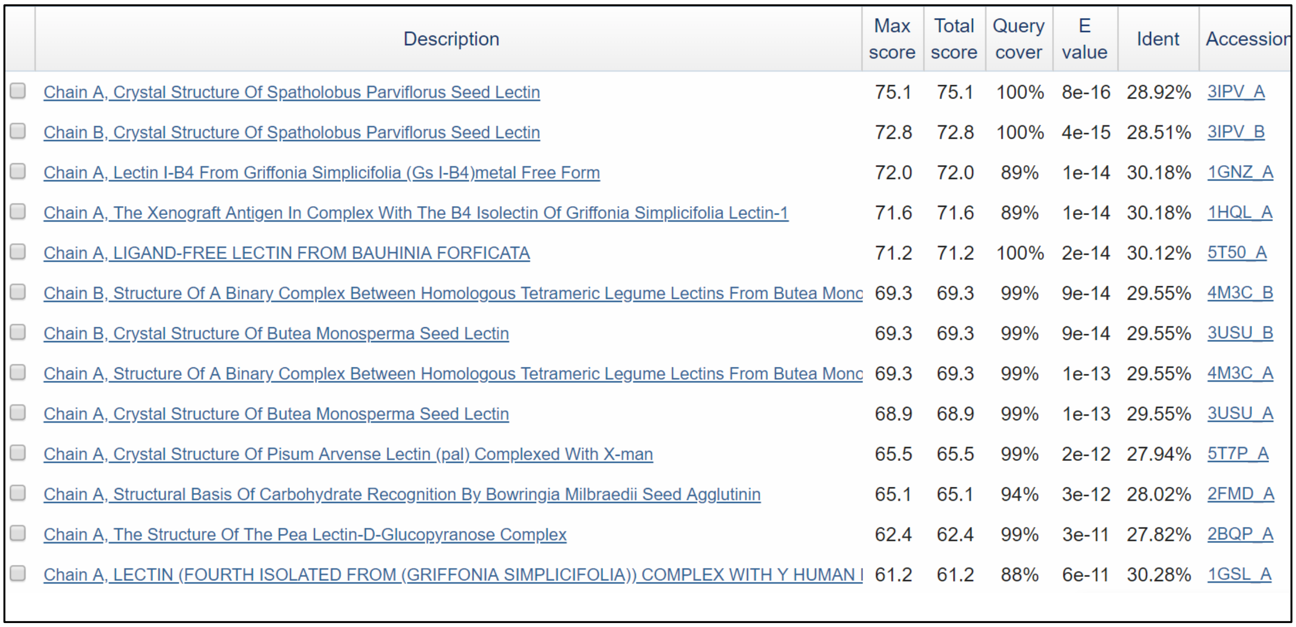
5.5 Blast Results:

The target protein sequence was subjected to blast using P-Suite (protein-protein BLAST) of the BLAST software against protein data bank (pdb) database. Upon completion of search, a blast report was presented split into three sections: a graphical summary, a list of sequences producing significant alignments, and the corresponding alignments. The graphical summary shows the alignments (as colored boxes) of protein sequences that matched the query sequence. The color keys represent the score (S) of the alignment, with red indicating the highest score and black indicating the lowest score. The higher alignment scores the more significant hit.

The summary table shows all the sequences in the database that showed significant match to the query sequence. The results were sorted in descending order in terms of increasing Expect value (E-value) and decreasing Max score and total score. The E-value is the number of alignments expected by chance with the same score. A number close to zero means that the hit has to be significant and not due to chance. Normally, E < .05 is required to be considered significant. The Max score is the blast score from part of the subject sequence that aligns best to the query, while the total score is the sum of the blast scores from each region where the query and subject sequence align. If two sequences align in multiple places then the total score is higher than the max score.

From the list of top 51 blast hits, the top 10 sequences were selected and renamed for convenience. The FASTA format of each of the top sequence was retrieved from the NCBI protein database.





	Description	Max score	Total score	Query cover	E value	Ident	Accession
□	Chain A, Crystal Structure Of Spatholobus Parviflorus Seed Lectin	75.1	75.1	100%	8e-16	28.92%	3IPV_A
□	Chain B, Crystal Structure Of Spatholobus Parviflorus Seed Lectin	72.8	72.8	100%	4e-15	28.51%	3IPV_B
□	Chain A, Lectin I-B4 From Griffonia Simplicifolia (Gs I-B4)metal Free Form	72.0	72.0	89%	1e-14	30.18%	1GNZ_A
□	Chain A, The Xenograft Antigen In Complex With The B4 Isolectin Of Griffonia Simplicifolia Lectin-1	71.6	71.6	89%	1e-14	30.18%	1HQL_A
□	Chain A, LIGAND-FREE LECTIN FROM BAUHINIA FORFICATA	71.2	71.2	100%	2e-14	30.12%	5T50_A
□	Chain B, Structure Of A Binary Complex Between Homologous Tetrameric Legume Lectins From Butea Monc	69.3	69.3	99%	9e-14	29.55%	4M3C_B
□	Chain B, Crystal Structure Of Butea Monosperma Seed Lectin	69.3	69.3	99%	9e-14	29.55%	3USU_B
□	Chain A, Structure Of A Binary Complex Between Homologous Tetrameric Legume Lectins From Butea Monc	69.3	69.3	99%	1e-13	29.55%	4M3C_A
□	Chain A, Crystal Structure Of Butea Monosperma Seed Lectin	68.9	68.9	99%	1e-13	29.55%	3USU_A
□	Chain A, Crystal Structure Of Pisum Arvense Lectin (pal) Complexed With X-man	65.5	65.5	99%	2e-12	27.94%	5T7P_A
□	Chain A, Structural Basis Of Carbohydrate Recognition By Bowringia Milbraedii Seed Agglutinin	65.1	65.1	94%	3e-12	28.02%	2FMD_A
□	Chain A, The Structure Of The Pea Lectin-D-Glucopyranose Complex	62.4	62.4	99%	3e-11	27.82%	2BQP_A
□	Chain A, LECTIN (FOURTH ISOLATED FROM (GRIFFONIA SIMPLICIFOLIA)) COMPLEX WITH Y HUMAN I	61.2	61.2	88%	6e-11	30.28%	1GSL_A

Figure 5.4 Results of DORN1 from BLAST

Table 3: List of Top 10 sequences obtained from BLAST results

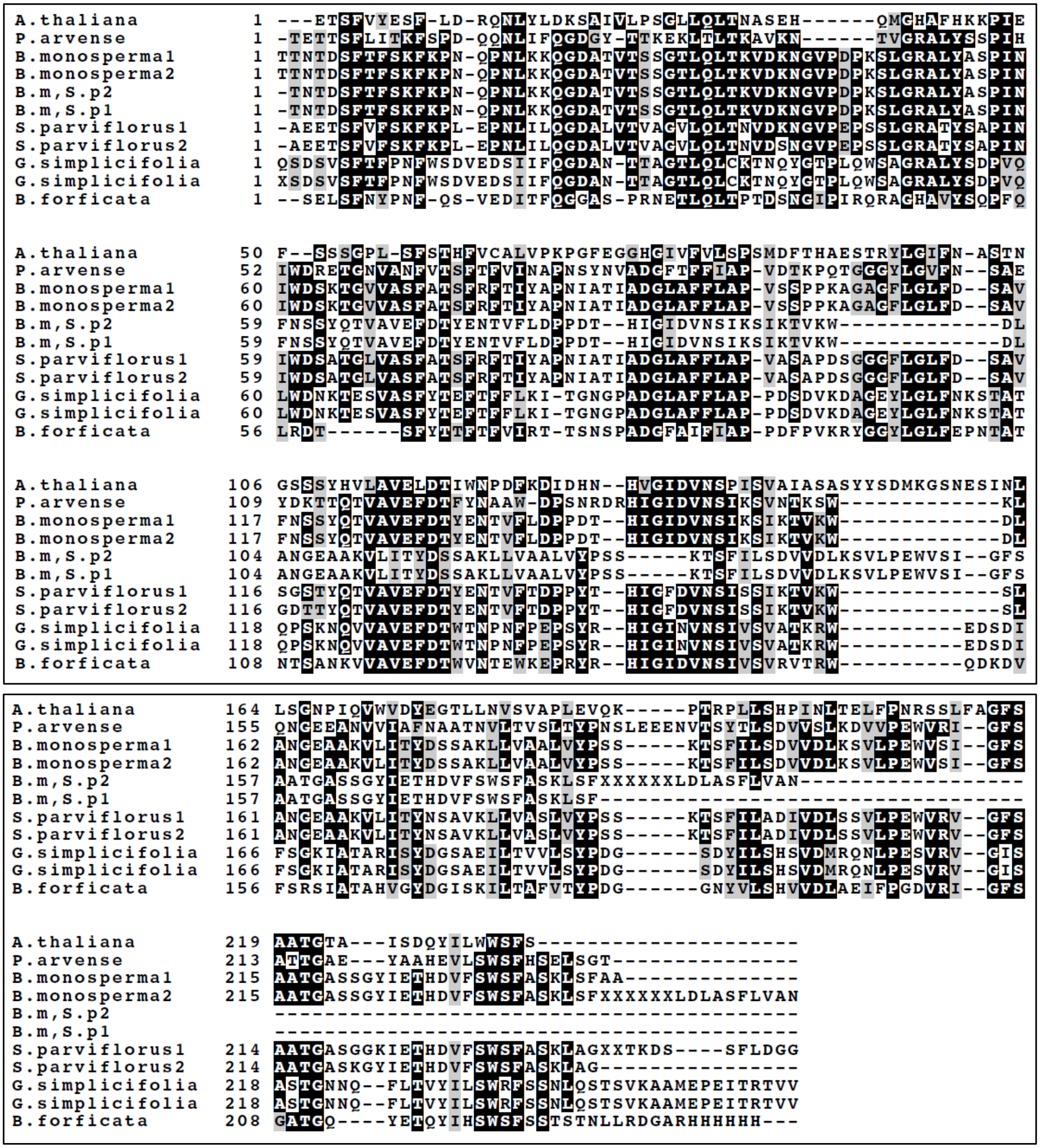
Serial No.	PDB ID	Max Score	E-value	Query Cover (%)	Ident (%)	Template Short Identity
1	3IPV_A	75.1	8e-16	100%	28.92%	Chain A, Crystal Structure Of Spatholobus Parviflorus Seed Lectin
2	3IPV_B	72.8	4e-15	100%	28.51%	Chain B, Crystal Structure Of Spatholobus Parviflorus Seed Lectin
3	1GNZ_A	72.0	1e-14	89%	30.18%	Chain A, Lectin I-B4 From GriffoniaSimplicifolia (Gs I-B4)metal Free Form
4	1HQL_A	71.6	1e-14	89%	30.18%	Chain A, The Xenograft Antigen In Complex With The B4 Isolectin Of Griffonia Simplicifolia Lectin-1
5	5T50_A	71.2	2e-14	100%	30.12%	Chain A, LIGAND-FREE LECTIN FROM BAUHINIA FORFICATA
6	4M3C_B	69.3	9e-14	99%	29.55%	Chain B, Structure Of A Binary Complex Between Homologous Tetrameric Legume Lectins From Butea Monosperma And Spatholobus Parviflorus Seeds
7	3USU_B	69.3	9e-14	99%	29.55%	Chain B, Crystal Structure Of Butea Monosperma Seed Lectin
8	4M3C_A	69.3	1e-13	99%	29.55%	Chain A, Structure Of A Binary Complex Between Homologous Tetrameric Legume Lectins From Butea Monosperma And Spatholobus Parviflorus Seeds
9	3USU_A	68.9	1e-13	99%	29.55%	Chain A, Crystal Structure Of Butea Monosperma Seed Lectin
10	5T7P_A	65.5	2e-12	99%	27.94%	Chain A, Crystal Structure Of Pisum Arvense Lectin (pal) Complexed With X-man

Table 4: List of Top 10 sequences obtained from BLAST results with matching altered names

Serial No.	PDB ID	Organism	Matching Altered Name
1	3IPV_A	Spatholobus Parviflorus	S.parviflorus 1
2	3IPV_B	Spatholobus Parviflorus	S.parviflorus 2
3	1GNZ_A	Griffonia Simplicifolia	G.simplicifolia 1
4	1HQL_A	Griffonia Simplicifolia	G.simplicifolia 2
5	5T50_A	Bauhinia Forficata	B.forficata
6	4M3C_B	Butea Monosperma, Spatholobus Parviflorus	B.M,S.P 1
7	3USU_B	Butea Monosperma	B.monosperma 1
8	4M3C_A	Butea Monosperma, Spatholobus Parviflorus	B.M,S.P 2
9	3USU_A	Butea Monosperma	B.monosperma 2
10	5T7P_A	Pisum Arvense	P.arvense

5.6 Multiple Sequence Alignment:

Multiple sequence alignment of the selected protein sequences was performed using MUSCLE. To convert the results into a publishable format BoxShade server was used. The output file generated was downloaded where identical or similar amino acid sequences are shaded black and grey respectively and gap regions are indicated by “ – “ . From the MSA it could be seen that while there are identical residue regions among the selected protein sequences and DORN1 sequence, there are also gap regions. Moreover exposed residues are more frequently observed in gap regions than buried residues. Identical residues indicate conserved regions among query and selected sequences, while gap regions can be interpreted as insertion or deletion mutations. In multiple sequence alignments regions of variable length that are bounded by conserved residue stretches on either side typically correspond to surface loops in proteins (Blouin, Butt & Roger, 2004).



```
A.thaliana          1 ---ETSFVYESF-LD-RQNLYLDKSAIVLPSGLLQLTNASEH------QMGHAFHKKPIE
P.arvense           1 -TETTSFLITKFSPD-QQNLIFQGDGY-TTKEKLTLTKAVKN------TVGRALYSSPIH
B.monosperma1       1 TTNTDSFTFSKFKPN-QPNLKKQGDATVTSSGTLQLTKVDKNGVPDPKSLGRALYASPIN
B.monosperma2       1 TTNTDSFTFSKFKPN-QPNLKKQGDATVTSSGTLQLTKVDKNGVPDPKSLGRALYASPIN
B.m,S.p2            1 -TNTDSFTFSKFKPN-QPNLKKQGDATVTSSGTLQLTKVDKNGVPDPKSLGRALYASPIN
B.m,S.p1            1 -TNTDSFTFSKFKPN-QPNLKKQGDATVTSSGTLQLTKVDKNGVPDPKSLGRALYASPIN
S.parviflorus1      1 -AEETSFVFSKFKPL-EPNLILQGDALVTVAGVLQLTNVDKNGVPEPSSLGRATYSAPIN
S.parviflorus2      1 -AEETSFVFSKFKPL-EPNLILQGDALVTVAGVLQLTNVDSNGVPEPSSLGRATYSAPIN
G.simplicifolia     1 QSDSVSFTFPNFWSDVEDSIIFQGDAN-TTAGTLQLCKTNQYGTPLQWSAGRALYSDPVQ
G.simplicifolia     1 XSDSVSFTFPNFWSDVEDSIIFQGDAN-TTAGTLQLCKTNQYGTPLQWSAGRALYSDPVQ
B.forficata         1 --SELSFNYPNF-QS-VEDITFQGGAS-PRNETLQLTPTDSNGIPIRQRAGHAVYSQPFQ

A.thaliana         50 F--SSSGPL-SFSTHFVCALVPKPGFEGGHGIVFVLSPSMDFTHAESTRYLGIFN-ASTN
P.arvense          52 IWDRETGNVANFVTSFTFVINAPNSYNVADGFTFFIAP-VDTKPQTGGGYLGVFN--SAE
B.monosperma1      60 IWDSKTGVVASFATSFRFTIYAPNIATIADGLAFFLAP-VSSPPKAGAGFLGLFD--SAV
B.monosperma2      60 IWDSKTGVVASFATSFRFTIYAPNIATIADGLAFFLAP-VSSPPKAGAGFLGLFD--SAV
B.m,S.p2           59 FNSSYQTVAVEFDTYENTVFLDPPDT--HIGIDVNSIKSIKTVKW-------------DL
B.m,S.p1           59 FNSSYQTVAVEFDTYENTVFLDPPDT--HIGIDVNSIKSIKTVKW-------------DL
S.parviflorus1     59 IWDSATGLVASFATSFRFTIYAPNIATIADGLAFFLAP-VASAPDSGGGFLGLFD--SAV
S.parviflorus2     59 IWDSATGLVASFATSFRFTIYAPNIATIADGLAFFLAP-VASAPDSGGGFLGLFD--SAV
G.simplicifolia    60 LWDNKTESVASFYTEFTFFLKI-TGNGPADGLAFFLAP-PDSDVKDAGEYLGLFNKSTAT
G.simplicifolia    60 LWDNKTESVASFYTEFTFFLKI-TGNGPADGLAFFLAP-PDSDVKDAGEYLGLFNKSTAT
B.forficata        56 LRDT------SFYTTFTFVIRT-TSNSPADGFAIFIAP-PDFPVKRYGGYLGLFEPNTAT

A.thaliana        106 GSSSYHVLAVELDTIWNPDFKDIDHN--HVGIDVNSPISVAIASASYYSDMKGSNESINL
P.arvense         109 YDKTTQTVAVEFDTFYNAAW-DPSNRDRHIGIDVNSIKSVNTKSW-------------KL
B.monosperma1     117 FNSSYQTVAVEFDTYENTVFLDPPDT--HIGIDVNSIKSIKTVKW-------------DL
B.monosperma2     117 FNSSYQTVAVEFDTYENTVFLDPPDT--HIGIDVNSIKSIKTVKW-------------DL
B.m,S.p2          104 ANGEAAKVLITYDSSAKLLVAALVYPSS-----KTSFILSDVVDLKSVLPEWVSI--GFS
B.m,S.p1          104 ANGEAAKVLITYDSSAKLLVAALVYPSS-----KTSFILSDVVDLKSVLPEWVSI--GFS
S.parviflorus1    116 SGSTYQTVAVEFDTYENTVFTDPPYT--HIGFDVNSISSIKTVKW-------------SL
S.parviflorus2    116 GDTTYQTVAVEFDTYENTVFTDPPYT--HIGFDVNSISSIKTVKW-------------SL
G.simplicifolia   118 QPSKNQVVAVEFDTWTNPNFPEPSYR--HIGINVNSIVSVATKRW----------EDSDI
G.simplicifolia   118 QPSKNQVVAVEFDTWTNPNFPEPSYR--HIGINVNSIVSVATKRW----------EDSDI
B.forficata       108 NTSANKVVAVEFDTWVNTEWKEPRYR--HIGIDVNSIVSVRVTRW----------QDKDV
```

```
A.thaliana        164 LSGNPIQVWVDYEGTLLNVSVAPLEVQK-----PTRPLLSHPINLTELFPNRSSLFAGFS
P.arvense         155 QNGEEANVVIAFNAATNVLTVSLTYPNSLEEENVTSYTLSDVVSLKDVVPEWVRI--GFS
B.monosperma1     162 ANGEAAKVLITYDSSAKLLVAALVYPSS-----KTSFILSDVVDLKSVLPEWVSI--GFS
B.monosperma2     162 ANGEAAKVLITYDSSAKLLVAALVYPSS-----KTSFILSDVVDLKSVLPEWVSI--GFS
B.m,S.p2          157 AATGASSGYIETHDVFSWSFASKLSFXXXXXXLDLASFLVAN------------------
B.m,S.p1          157 AATGASSGYIETHDVFSWSFASKLSF----------------------------------
S.parviflorus1    161 ANGEAAKVLITYNSAVKLLVASLVYPSS-----KTSFILADIVDLSSVLPEWVRV--GFS
S.parviflorus2    161 ANGEAAKVLITYNSAVKLLVASLVYPSS-----KTSFILADIVDLSSVLPEWVRV--GFS
G.simplicifolia   166 FSGKIATARISYDGSAEILTVVLSYPDG------SDYILSHSVDMRQNLPESVRV--GIS
G.simplicifolia   166 FSGKIATARISYDGSAEILTVVLSYPDG------SDYILSHSVDMRQNLPESVRV--GIS
B.forficata       156 FSRSIATAHVGYDGISKILTAFVTYPDG------GNYVLSHVVDLAEIFPGDVRI--GFS

A.thaliana        219 AATGTA---ISDQYILWWSFS---------------------
P.arvense         213 ATTGAE---YAAHEVLSWSFHSELSGT---------------
B.monosperma1     215 AATGASSGYIETHDVFSWSFASKLSFAA--------------
B.monosperma2     215 AATGASSGYIETHDVFSWSFASKLSFXXXXXXLDLASFLVAN
B.m,S.p2              ------------------------------------------
B.m,S.p1              ------------------------------------------
S.parviflorus1    214 AATGASGGKIETHDVFSWSFASKLAGXXTKDS----SFLDGG
S.parviflorus2    214 AATGASKGYIETHDVFSWSFASKLAG----------------
G.simplicifolia   218 ASTGNNQ--FLTVYILSWRFSSNLQSTSVKAAMEPEITRTVV
G.simplicifolia   218 ASTGNNQ--FLTVYILSWRFSSNLQSTSVKAAMEPEITRTVV
B.forficata       208 GATGQ----YETQYIHSWSFSSTSTNLLRDGARHHHHHH---
```

Figure 5.5 Results of multiple sequence alignment of DORN1 by MUSCLE edited by BoxShade server

Many residues in a given protein form regions of regular structure, in α-helices and β-sheets. The protein segments that join these secondary structure elements together, which do not have easily observable regular patterns in their structure, are referred to as loops (Regad, Martin, Nuel & Camproux, 2010). Such loops are typically found on protein surfaces and are subjected to more insertions, deletions and substitutions than the more conserved core regions (Blouin, Butt & Roger, 2004). Meaning surface loops are prone to rapid evolution unlike core regions. This implies that, for a homologous set of proteins, the loops are the regions that vary the most between structures.

5.7 Phylogenetic tree generation:

A phylogenetic tree is a display of information about the inferred evolutionary relationships between a set of sequences. In a phylogenetic tree, the species of interest are found at the tips of lines referred to as the tree's branches. Each branch point (also called an internal node) represents a splitting of lineage (speciation). Each node represents the most recent common ancestor of all the groups descended from that branch point. Two descendants that split from the same node are called sister groups. Moving forward from nodes to tips represent moving forward in time. Additionally the branch lengths indicate the amount of evolutionary change. The greater the branch length, the greater the amount of genetic change. Lastly the vertical dimension in a phylogenetic tree is meaningless as branches can be swapped at any internal nodes.

Based on the results obtained Arabidopsis thaliana’s DORN1 has undergone the greatest amount of character change compared to the other groups as it has largest branch length. Due to this, DORN1 is also distantly related to all the other sequences. In the tree obtained, some branches show zero branch length as these sister groups are actually different chains of the same protein so show no evolutionary change relative to their most recent common ancestor which is the entire protein itself. The percentage beside each node is referred to as data coverage and acts as an indicator of the confidence of the node. The higher the data coverage the greater the statistical confidence of the node.

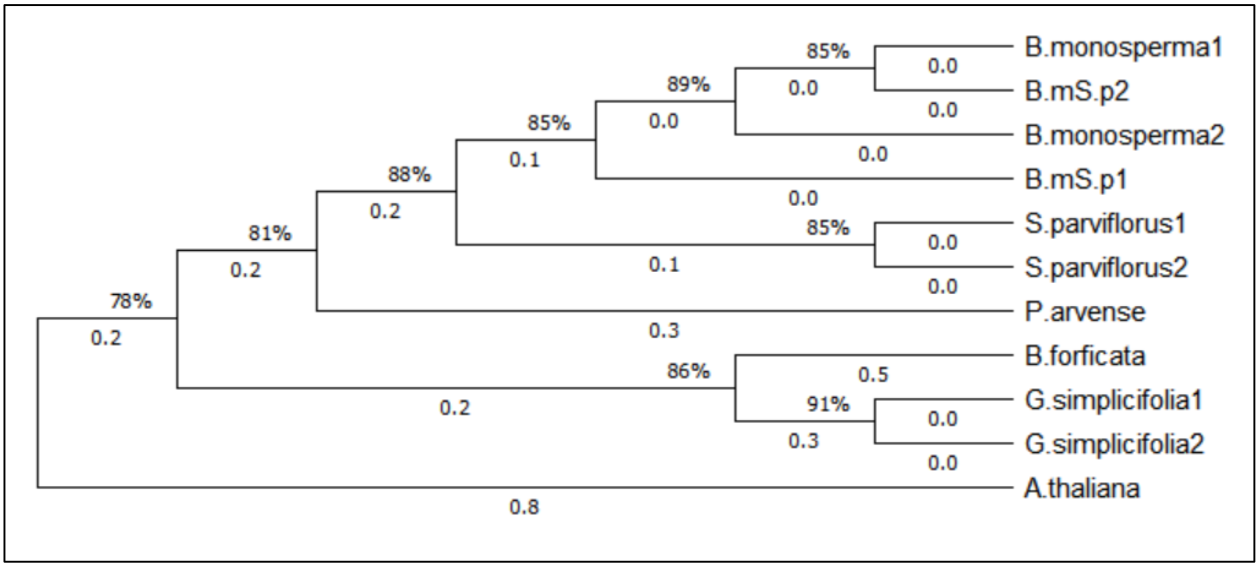


Figure 5.6 Maximum likelyhood tree of top 10 sequences obtained from BLAST results

5.8 Single template modeling:

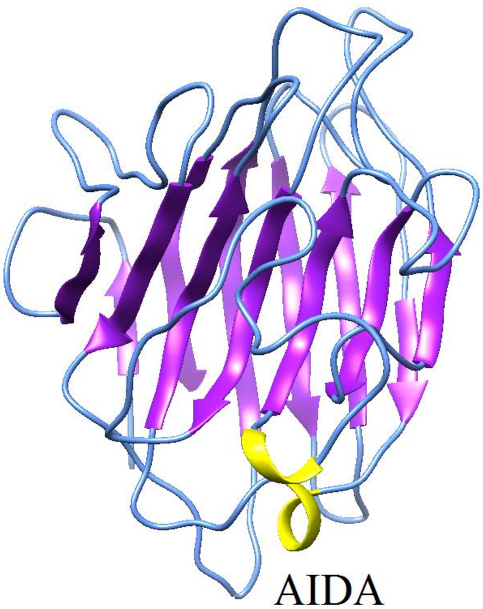
A total of seven tools were utilized to perform single template modeling of the DORN1 extracellular domain. Both FFAS03 server and Swiss Model provided a list of templates from which to choose for modeling. The top-ranked template in FFAS03 was chosen based on the FFAS score where lower FFAS score indicate higher confidence of prediction. In Swiss Model the 1st template in the list was chosen as the list is ordered from top ranked to lowest ranked based on parameters such as coverage, sequence identity, Global Model Quality

Estimation(GMQE) and Quaternary Structure Quality Estimate (QSQE). SPARKX on the other hand generated 10 models based on the top 10 best matches in target-template alingments. The models were ordered according to their Z-score and the model with highest Z-score (model 1) was chosen.

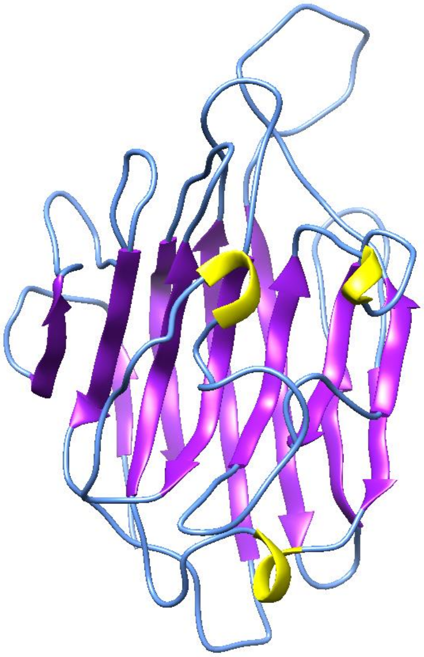
All tools were able to correctly construct the beta-sandwich fold that is characteristic of the L-type lectin domain. However the model generated by SPARKX contains 15 beta-strands instead of the typical 13 strands found in L-type lectins. This result is more in line with the PSIPRED results, which predicted that there are 17 beta-strands in the DORN1 ectodomain. In addition there are inconsistencies in the number of helixes present in the seven different models. FFAS03 and SPARKX model contains 3 helix where the rest of the models 2 helix. Lastly, none of the models contained the 17 amino acids loop between beta strand 11 and 12 typical of L-type lectins.

Table 5: Templates used by Single Template Modeling tools to model DORN1 ectodomain

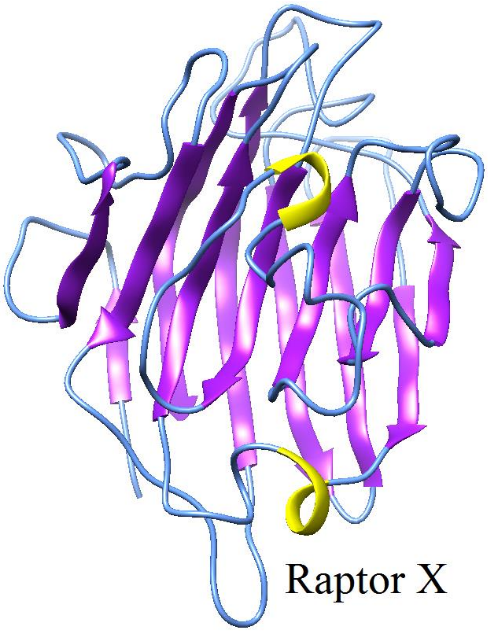
Tool	Template Used (PDB ID)
AIDA	3usu_A
FFAS03	2fmd_A
Muster	2bqp_A
RaptorX	5t7p_A
SparkX	2bqp_A
Swiss Model	3wcs_A
(PS)2-v2	1fat_B



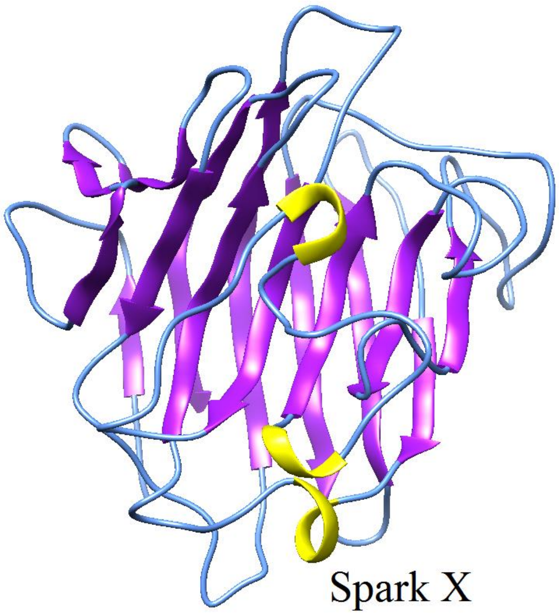
AIDA



FFAS03



Raptor X



Spark X

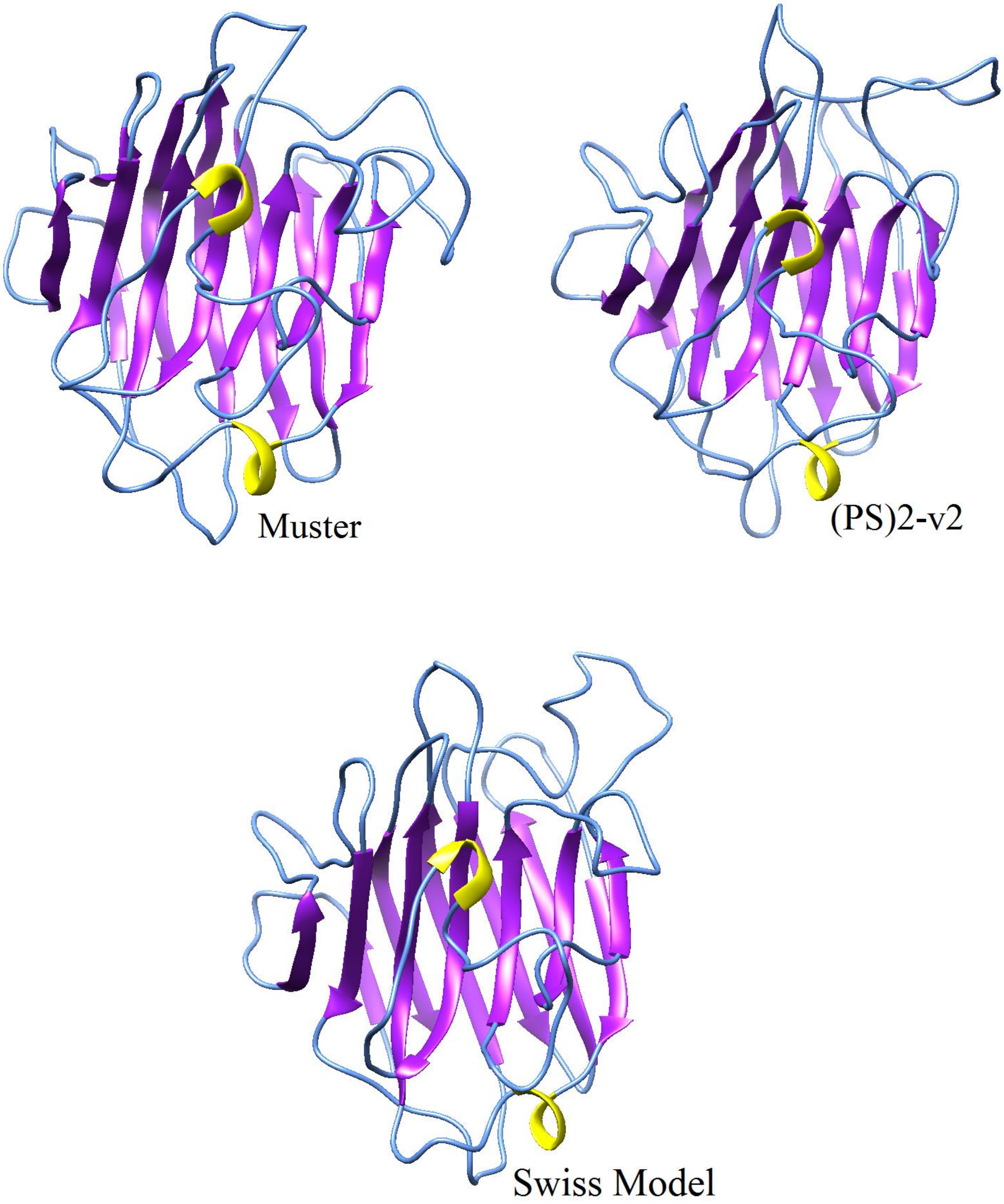


Figure 5.7: Models of DORN1 ectodomain generated using Single Template Modeling tools

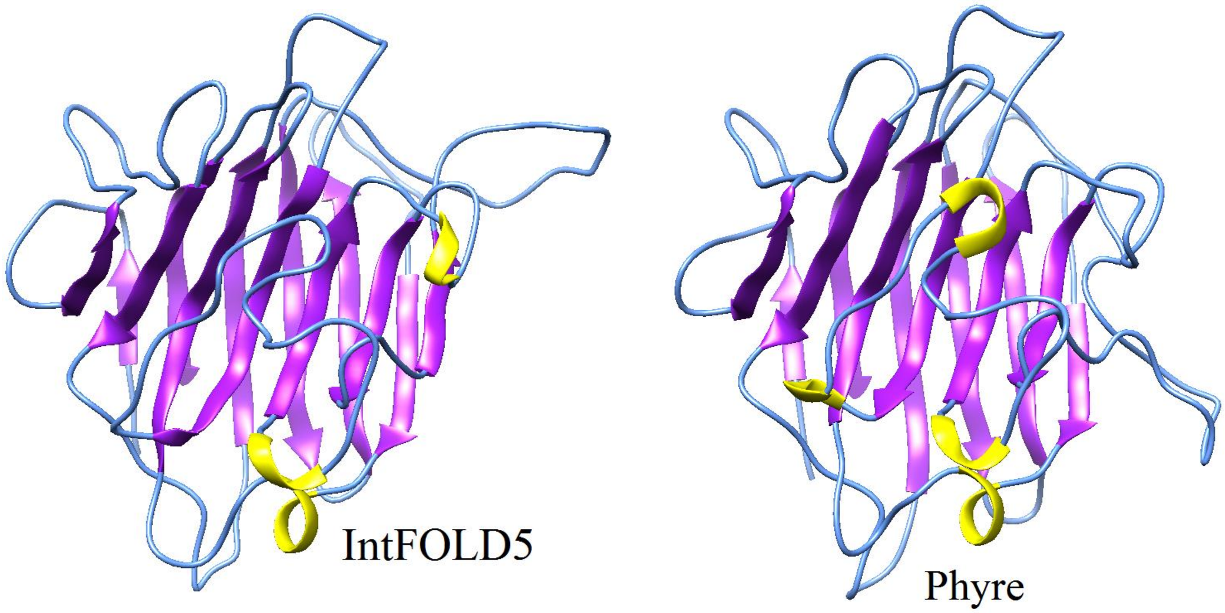
5.9 Multiple template modeling:

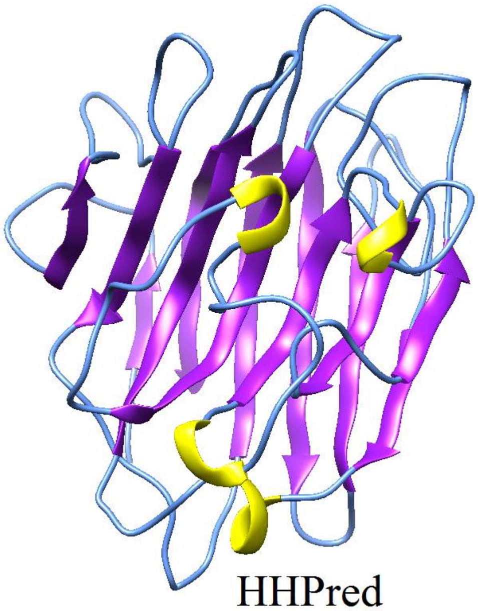
Multiple template modeling was carried out using four tools HHPRED, INTFOLD5, I-Tasser and PHYRE2. HHPRED produced a hit list of best-matched templates that was ordered in decreasing probability and increasing E-value. The top 5 templates were chosen for modeling. IntFOLD generated 5 models and ordered them according to increasing P value and decreasing global model quality score. P-value is a probability metric representing the likelihood that a match occurred by chance. The closer the P-value is to zero the more significant the match. Thus the first model with the lowest P value and the highest global model quality score was picked. I-Tasser also generated 5 models and the confidence of each model is quantitatively measured by C-score that is calculated based on the significance of threading template alignments and the convergence parameters of structure assembly simulations. C-score is typically in the range of [-5,2], where C-score of high value signifies a model with higher confidence. However since the top 5 models are ranked by cluster size, not C-score so it is possible that lower ranked models can have higher C-score than higher ranked models. Although the first model has a higher C-score and a better quality in most cases, it is not unusual that the lower-rank models have a better quality than the higher-rank models. Consequently all 5 I-Tasser models were picked for further model validation.

Among the 8 models generated by multiple template modeling there are inconsistencies in the number of helixes as well as beta strands. There are also uncharacteristic structural features in 2 of the I-tasser models. Both models 2 and 4 have 12 strands, as opposed to the typical 13, and form a five stranded and seven stranded beta sheet. On top of which the arrangement of the individual strands on the five stranded beta sheet leaves a portion of the seven stranded beta sheet either exposed to the solvent or interacting with loop regions as opposed to interacting with the five stranded beta-sheet. Therefore, potentially altering the stability of the structure. IntFOLD5, Phyre2, HHPred and I-Tasser(model 1,3 and 5) were all able to correctly construct the typical beta-sandwich fold. Despite having 12 beta strands, the Phyre2 model has two six stranded beta sheet, leaving no portion of either beta-sheet exposed to the solvent. The number of beta strands in I-Tasser models 2, 4, 5 and Phyre is in line with that found in members of concanavalin A-like lectin or glucanase domain superfamily. Similar to STM no loop of 17 residues was observed in any of the models generated.

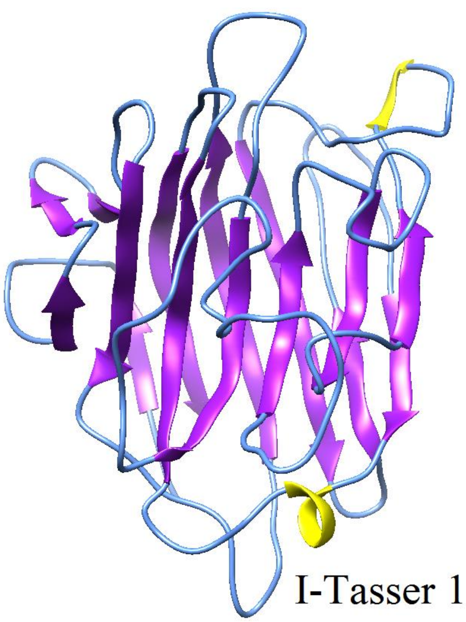
Table 6: Templates used by Multiple Template Modeling tools to model DORN1 ectodomain

Tool	Template Used (PDB ID)
IntFOLD5	1gzc_A, 1fny_A, 3wcs_A, 3ipv_A
Phyre	3ipv_C, 5kxc_B, 1lu1_A, 1fx5_A, 1fny_A, 1dbn_A
HHPred	3ipv_A, 1fny_A, 1v6i_D, 1dbn_A, 1sbf_A
I-Tasser	2bqp_A, 5t50_A, 3ipv, 1jxn, 1ciw_A, 1dbn, 2fmd_A, 1dzq_B, 3ipv_A

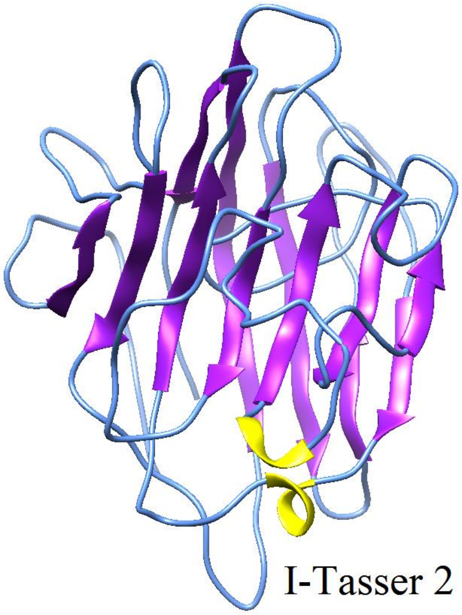




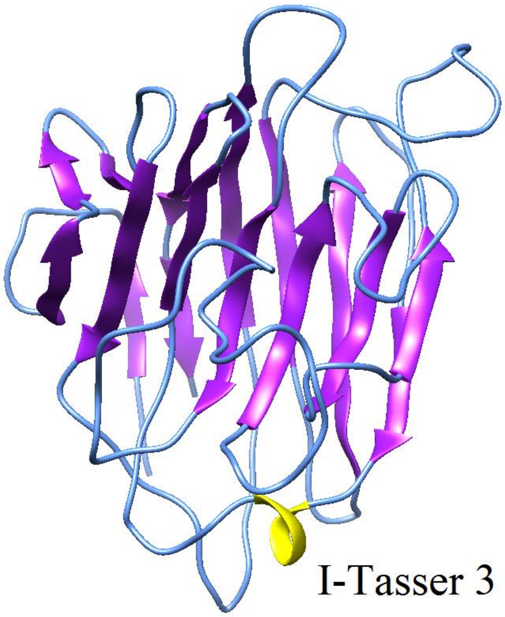
HHPred



I-Tasser 1



I-Tasser 2



I-Tasser 3

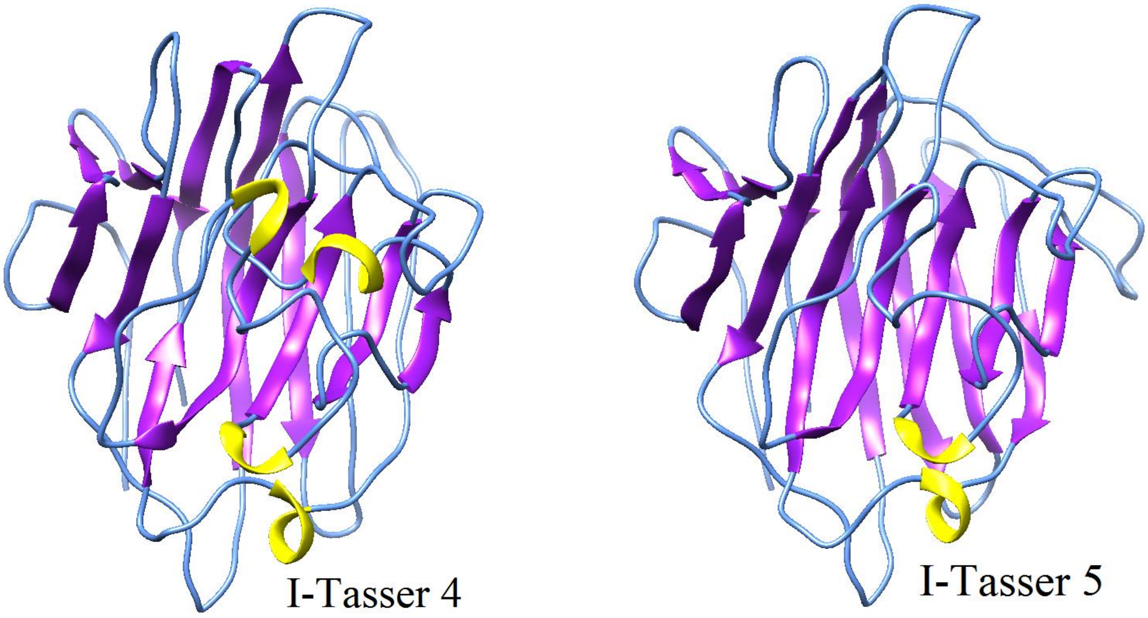


Figure 5.8 Models of DORN1 ectodomain generated using Multiple Template Modeling tools

5.10 Evaluation and validation of models:

All 15 models were subjected to evaluation and model validation by different structural validation tools. Verify 3D was used to analyze the compatibility of the model with its own amino acid sequence, ERRAT was used to analyze the non-bonded interactions between the amino acids and gain a better understanding of the overall quality of the model, and finally the Ramachandran distribution plot indicated the stability of the model of the protein based on rotations of a polypeptide that are allowed and disallowed due to steric hindrance.

Table 7: Validation scores obtained for DORN1 ectodomain modeled by Single Template Modeling tools

Modeling Tool	**Verify 3D (%)**	**ERRAT**	**Ramachandran Plot (%)**		
			FR	**AR**	**OR**
AIDA	93.64	46.9298	93.2	4.3	2.6
FFAS03 Modeller	83.05	43.4978	93.2	4.7	2.1
Muster	73.31	44.2982	91.9	5.6	2.6
(PS)2-v2	88.98	32.8889	91.5	4.7	3.8
Raptor X	84.75	50	91.9	5.6	2.6
Spark X	76.27	40.7895	94.4	4.3	1.3
Swiss Model	82.55	81.1659	87.6	9.4	3.0

It was observed that among the single template models generated, only Swiss Model and Raptor X succeeded in meeting the ERRAT cut off score (>50) while the rest were unsuccessful. Swiss model had the highest ERRAT score 81.1659 while (PS)2-v2 had the lowest 32.8889. Besides Muster and SparkX all other models were able to meet the Verify3D cut off value (>80%) and passed the test, while Muster and SparkX scoring 73.31% and 76.27% respectively failed. AIDA scored the highest in Verify3D, 93.64% of the amino acids are compatible with their 3D structure. Although none of the models acquired the ideal scores in the Ramachandran Plot (Favored region ~ 98% and Allowed region~2%) using the RAMPAGE server, SparkX showed the most promise with 94.4% in the favored region, followed closely by AIDA and FFAS03 with 93.2% in the favored region.

Table 8: Validation scores obtained for DORN1 ectodomain modeled by Multiple Template Modeling tools

Modeling Tool	**Verify 3D (%)**	**ERRAT**	**Ramachandran Plot (%)**		
			FR	**AR**	**OR**
Phyre2	86.44	40.9692	87.6	9.0	3.4
IntFOLD	81.36	40.5286	93.2	4.7	2.1
HHpred	84.75	41.2281	91.9	6.0	2.1
I-Tasser (Model 1)	97.03	87.7193	76.9	15	8.1
I-Tasser (Model 2)	79.66	70.1754	76.9	14.5	8.5
I-Tasser (Model 3)	93.22	72.807	75.6	15.8	8.5
I-Tasser (Model 4)	77.97	63.1579	69.7	22.2	8.1
I-Tasser (Model 5)	87.29	85.0877	70.5	20.5	9.0

Among the multiple template models, all five I-Tasser models crossed the ERRAT cut off score but HHPRED, Phyre2 and IntFOLD were unsuccessful. Except I-tasser model 2 and 4 all the other models scored over 80% in Verify3D and thus passed, with I-Tasser model 1 scoring the highest 97.03%. This showed that the amino acids have good compatibility with their own 3D structure. In Ramachandran distribution plot only HHPRED and IntFOLD have most of their residues >90% in the favored regions. In contrast all five I-Tasser models have less than 80% of their residues in the favored region, making them the models with the lowest values out of all 15 models generated.

After quantitative structural validation and visual analysis was done, the models with the best overall scores were short listed. It was observed that, overall the single template modeling tools performed better than the multiple template modeling tools. However no particular modeling method outperformed the others as the top 5 models included a mix of ab-initio, homology modeling and threading.

Table 9: Top 5 models generated

Modeling Tool	Verify 3D (%)	ERRAT(%)	Ramachandran Plot (%)		
			FR	AR	OR
AIDA	93.64	46.9298	93.2	4.3	2.6
FFAS03	83.05	43.4978	93.2	4.7	2.1
SPARKX	76.27	40.7895	94.4	4.3	1.3
IntFOLD	81.36	40.5286	93.2	4.7	2.1
HHPred	84.75	41.2281	91.9	6.0	2.1

Chapter 6: Conclusion

Conclusion:

The main objective of this study was to gain a better understanding of the first plant-specific eATP receptor, DORN1 legume lectin ectodomain using computational approaches. The aim behind choosing such a protein was to work on a comparatively novel macromolecule whose structure is yet to be determined. The study revealed that like other legume lectins the hydrophobic core of DORN1 ectodomain comprised of the beta-sandwich fold consisting of individual beta strands connected by loops however, unlike other legume lectins DORN1 did not contain an additional 17 amino acid loop between beta strands 11 & 12. The protein was shown to be acidic with much of its stability coming from hydrogen bonds, van der waals forces and ionic bonds. Compared to its homologous relatives, DORN1 was observed to have undergone a greater amount of evolutionary change which could explain its significant affinity for eATP. While the exact amino acids responsible for eATP binding is yet to be determined, the models generated in this study, along with the information of the physiochemical properties determined could be utilized for further investigation in docking experiments to reveal the potential ligand binding sites.

Bibliography

Bibliography:

Zipfel, C. (2014). Plant pattern-recognition receptors. Trends in Immunology , 35 (7), 345-351. doi: 10.1016/j.it.2014.05.004

Jones, J. & Dangl, J. (2006). The plant immune system. Nature, 444(7117), 323-329. doi:10.1038/nature05286

Bigeard, J., Colcombet, J. and Hirt, H. (2015). Signaling Mechanisms in Pattern-Triggered Immunity (PTI). Molecular Plant, 8(4), 521-539. doi: 10.1016/j.molp.2014.12.022

Ranf, S. (2017). Sensing of molecular patterns through cell surface immune receptors. Current Opinion in Plant Biology, 38, 68-77. doi: 10.1016/j.pbi.2017.04.011

Lee, H., Lee, H., Seo, E., Lee, J., Kim, S., Oh, S., Choi, E., Choi, E., Lee, S. and Choi, D. (2017). Current Understandings of Plant Nonhost Resistance. Molecular Plant-Microbe Interactions, 30(1), 5-15. doi: 10.1094/MPMI-10-16-0213-CR

Saijo, Y., Loo, E. and Yasuda, S. (2018). Pattern recognition receptors and signaling in plant-microbe interactions. The Plant Journal, 93(4), 592-613. doi: 10.1111/tpj.13808

Choi, H. & Klessig, D. (2016). DAMPs, MAMPs, and NAMPs in plant innate immunity. BMC Plant Biology, 16(1). doi: 10.1186/s12870-016-0921-2

Hervé, C., Dabos, P., Galaud, J., Rougé, P. & Lescure, B. (1996). Characterization of an Arabidopsis thaliana Gene that Defines a New Class of Putative Plant Receptor Kinases with an Extracellular Lectin-like Domain. Journal of Molecular Biology, 258(5), 778-788. doi: 10.1006/jmbi.1996.0286

Shiu, S. & Bleecker, A. (2001). Receptor-like kinases from Arabidopsis form a monophyletic gene family related to animal receptor kinases. Proceedings of the National Academy of Sciences, 98(19), 10763-10768. doi: 10.1073/pnas.181141598

Barre, A., Hervé, C., Lescure, B. & Rougé, P. (2002). Lectin Receptor Kinases in Plants. Critical Reviews in Plant Sciences, 21(4), 379-399. doi: 10.1080/0735-260291044287

Choi, J., Tanaka, K., Cao, Y., Qi, Y., Qiu, J., Liang, Y., Lee, S. and Stacey, G. (2014a). Identification of a Plant Receptor for Extracellular ATP. Science, 343(6168), 290-294. doi: 10.1126/science.343.6168.290.

Zhang, L., Du, L. and Poovaiah, B. (2014). Calcium signaling and biotic defense responses in plants. Plant Signaling & Behavior, 9(11), e973818. doi: 10.4161/15592324.2014.973818

Meng, X. & Zhang, S. (2013). MAPK Cascades in Plant Disease Resistance Signaling. Annual Review of Phytopathology, 51(1), 245-266. doi: 10.1146/annurev-phyto-082712-102314

Choi, J., Tanaka, K., Liang, Y., Cao, Y., Lee, S., & Stacey, G. (2014b). Extracellular ATP, a danger signal, is recognized by DORN1 in Arabidopsis. Biochemical Journal, 463(3), 429-437. doi: 10.1042/BJ20140666

Ralevic,V. & Burnstock,G. (1998). Receptors for Purines and Pyrimidines. Pharmacological Reviews, 50(3), 413-492.

Abbracchio, M. P., Burnstock, G., Boeynaems, J. M., Barnard, E. A., Boyer, J. L., Kennedy, C., … Weisman, G. A. (2006). International Union of Pharmacology LVIII: update on the P2Y G protein-coupled nucleotide receptors: from molecular mechanisms and pathophysiology to therapy. Pharmacological reviews, 58(3), 281–341. doi: 10.1124/pr.58.3.3

Bouwmeester, K., & Govers, F. (2009). Arabidopsis L-type lectin receptor kinases: phylogeny, classification, and expression profiles. Journal Of Experimental Botany, 60(15), 4383-4396. doi: 10.1093/jxb/erp277

Lagarda-Diaz, I., Guzman-Partida, A. M., & Vazquez-Moreno, L. (2017). Legume Lectins: Proteins with Diverse Applications. International journal of molecular sciences, 18(6), 1242. doi:10.3390/ijms18061242

André, S., Siebert, H., Nishiguchi, M., Tazaki, K., & Gabius, H. (2005). Evidence for lectin activity of a plant receptor-like protein kinase by application of neoglycoproteins and bioinformatic algorithms. Biochimica Et Biophysica Acta (BBA) - General Subjects, 1725(2), 222-232. doi: 10.1016/j.bbagen.2005.04.004

Wang, Y., Nsibo, D., Juhar, H., Govers, F., & Bouwmeester, K. (2016). Ectopic expression of Arabidopsis L-type lectin receptor kinase genes LecRK-I.9 and LecRK-IX.1 in Nicotiana benthamiana confers Phytophthora resistance. Plant Cell Reports, 35(4), 845-855. doi: 10.1007/s00299-015-1926-2

Chen, X., Shang, J., Chen, D., Lei, C., Zou, Y., & Zhai, W. et al. (2006). A B-lectin receptor kinase gene conferring rice blast resistance. The Plant Journal, 46(5), 794-804. doi: 10.1111/j.1365-313x.2006.02739.x

He, X., Zhang, Z., Yan, D., Zhang, J., & Chen, S. (2004). A salt-responsive receptor-like kinase gene regulated by the ethylene signaling pathway encodes a plasma membrane serine/threonine kinase. Theoretical And Applied Genetics, 109(2), 377-383. doi: 10.1007/s00122-004-1641-9

Vaid, N., Macovei, A., & Tuteja, N. (2013). Knights in Action: Lectin Receptor-Like Kinases in Plant Development and Stress Responses. Molecular Plant, 6(5), 1405-1418. doi: 10.1093/mp/sst033

Deb, S., Sankaranarayanan, S., Wewala, G., Widdup, E., & Samuel, M. (2014). The S-Domain Receptor Kinase Arabidopsis Receptor Kinase2 and the U Box/Armadillo Repeat-Containing E3 Ubiquitin Ligase9 Module Mediates Lateral Root Development under Phosphate Starvation in Arabidopsis. PLANT PHYSIOLOGY, 165(4), 1647-1656. doi: 10.1104/pp.114.244376

Wan, J., Patel, A., Mathieu, M., Kim, S., Xu, D., & Stacey, G. (2008). A lectin receptor-like kinase is required for pollen development in Arabidopsis. Plant Molecular Biology, 67(5), 469-482. doi: 10.1007/s11103-008-9332-6

Deng, K., Wang, Q., Zeng, J., Guo, X., Zhao, X., Tang, D., & Liu, X. (2009). A Lectin Receptor Kinase Positively Regulates ABA Response During Seed Germination and Is Involved in Salt and Osmotic Stress Response. Journal Of Plant Biology, 52(6), 493-500. doi: 10.1007/s12374-009-9063-5

Wu, S., & Wu, J. (2008). Extracellular ATP-induced NO production and its dependence on membrane Ca2+ flux in Salvia miltiorrhiza hairy roots. Journal Of Experimental Botany, 59(14), 4007-4016. doi: 10.1093/jxb/ern242

Kim, S. Y., Sivaguru, M., & Stacey, G. (2006). Extracellular ATP in plants. Visualization, localization, and analysis of physiological significance in growth and signaling. Plant physiology, 142(3), 984–992. doi:10.1104/pp.106.085670

Dark, A., Demidchik, V., Richards, S. L., Shabala, S., & Davies, J. M. (2011). Release of extracellular purines from plant roots and effect on ion fluxes. Plant signaling & behavior, 6(11), 1855–1857. doi:10.4161/psb.6.11.17014

Tanaka, K., Swanson, S. J., Gilroy, S., & Stacey, G. (2010). Extracellular nucleotides elicit cytosolic free calcium oscillations in Arabidopsis. Plant physiology, 154(2), 705–719. doi:10.1104/pp.110.162503

Foresi, N. P., Laxalt, A. M., Tonón, C. V., Casalongué, C. A., & Lamattina, L. (2007). Extracellular ATP induces nitric oxide production in tomato cell suspensions. Plant physiology, 145(3), 589–592. doi:10.1104/pp.107.106518

Song, C. J., Steinebrunner, I., Wang, X., Stout, S. C., & Roux, S. J. (2006). Extracellular ATP induces the accumulation of superoxide via NADPH oxidases in Arabidopsis. Plant physiology, 140(4), 1222–1232. doi:10.1104/pp.105.073072

Sueldo, D., Foresi, N., Casalongué, C., Lamattina, L., & Laxalt, A. (2010). Phosphatidic acid formation is required for extracellular ATP-mediated nitric oxide production in suspension-cultured tomato cells. New Phytologist, 185(4), 909-916. doi: 10.1111/j.1469-8137.2009.03165.x

Cao, Y., Tanaka, K., Nguyen, C., & Stacey, G. (2014). Extracellular ATP is a central signaling molecule in plant stress responses. Current Opinion In Plant Biology, 20, 82-87. doi: 10.1016/j.pbi.2014.04.009

Chivasa, S., Murphy, A., Hamilton, J., Lindsey, K., Carr, J., & Slabas, A. (2009a). Extracellular ATP is a regulator of pathogen defence in plants. The Plant Journal, 60(3), 436-448. doi: 10.1111/j.1365-313x.2009.03968.x

Lim, M. H., Wu, J., Yao, J., Gallardo, I. F., Dugger, J. W., Webb, L. J., … Roux, S. J. (2014). Apyrase suppression raises extracellular ATP levels and induces gene expression and cell wall changes characteristic of stress responses. Plant physiology, 164(4), 2054–2067. doi:10.1104/pp.113.233429

SUN, J., ZHANG, C., DENG, S., LU, C., SHEN, X., & ZHOU, X. et al. (2011). An ATP signalling pathway in plant cells: extracellular ATP triggers programmed cell death in Populus euphratica. Plant, Cell & Environment, 35(5), 893-916. doi: 10.1111/j.1365-3040.2011.02461.x

Chivasa, S., Ndimba, B. K., Simon, W. J., Lindsey, K., & Slabas, A. R. (2005). Extracellular ATP functions as an endogenous external metabolite regulating plant cell viability. The Plant cell, 17(11), 3019–3034. doi:10.1105/tpc.105.036806

Chivasa, S., Tomé, D. F., Murphy, A. M., Hamilton, J. M., Lindsey, K., & Carr, J. P. (2009b). Extracellular ATP: a modulator of cell death and pathogen defense in plants. Plant signaling & behavior, 4(11), 1078–1080. doi:10.4161/psb.4.11.9784

Bouwmeester, K., de Sain, M., Weide, R., Gouget, A., Klamer, S., Canut, H., & Govers, F. (2011). The Lectin Receptor Kinase LecRK-I.9 Is a Novel Phytophthora Resistance Component and a Potential Host Target for a RXLR Effector. Plos Pathogens, 7(3), e1001327. doi: 10.1371/journal.ppat.1001327

Kaczanowski, S., & Zielenkiewicz, P. (2009). Why similar protein sequences encode similar three-dimensional structures?. Theoretical Chemistry Accounts, 125(3-6), 643-650. doi: 10.1007/s00214-009-0656-3

Bowie, J., Luthy, R., & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. Science, 253(5016), 164-170. doi: 10.1126/science.1853201

Wu, S., & Zhang, Y. (2008). MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. Proteins: Structure, Function, And Bioinformatics, 72(2), 547-556. doi: 10.1002/prot.21945

Lee, J., Freddolino, P.L., & Zhang, Y.A. (2008). Chapter 1 Ab Initio Protein Structure Prediction.

Holton, N., Nekrasov, V., Ronald, P., & Zipfel, C. (2015). The Phylogenetically-Related Pattern Recognition Receptors EFR and XA21 Recruit Similar Immune Signaling Components in Monocots and Dicots. PLOS Pathogens, 11(1), e1004602. doi: 10.1371/journal.ppat.1004602

Bouwmeester, K., Han, M., Blanco-Portales, R., Song, W., Weide, R., & Guo, L. et al. (2013). The Arabidopsis lectin receptor kinase LecRK-I.9 enhances resistance to Phytophthora infestans in Solanaceous plants. Plant Biotechnology Journal, 12(1), 10-16. doi: 10.1111/pbi.12111

Mendes, B., Cardoso, S., Boscariol-Camargo, R., Cruz, R., MourÃ£o Filho, F., & Bergamin Filho, A. (2010). Reduction in susceptibility toXanthomonas axonopodispv.citriin transgenicCitrus sinensisexpressing the riceXa21gene. Plant Pathology, 59(1), 68-75. doi: 10.1111/j.1365-3059.2009.02148.x

Afroz, A., Chaudhry, Z., Rashid, U., Ali, G., Nazir, F., Iqbal, J., & Khan, M. (2010). Enhanced resistance against bacterial wilt in transgenic tomato (Lycopersicon esculentum) lines expressing the Xa21 gene. Plant Cell, Tissue And Organ Culture (PCTOC), 104(2), 227-237. doi: 10.1007/s11240-010-9825-2

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., … Bourne, P. E. (2000). The Protein Data Bank. Nucleic acids research, 28(1), 235–242. doi:10.1093/nar/28.1.235

The UniProt Consortium. (2018). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Research, 47(D1), D506-D515. doi: 10.1093/nar/gky1049

Altschul, S., Gish, W., Miller, W., Myers, E., & Lipman, D. (1990). Basic local alignment search tool. Journal Of Molecular Biology, 215(3), 403-410. doi: 10.1016/s0022-2836(05)80360-2

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M., Appel, R., & Bairoch, A. (2005). Protein Identification and Analysis Tools on the ExPASy Server. The Proteomics Protocols Handbook, 571-607. doi: 10.1385/1-59259-890-0:571

Jones, D. (1999). Protein secondary structure prediction based on position-specific scoring matrices 1 1Edited by G. Von Heijne. Journal Of Molecular Biology, 292(2), 195-202. doi: 10.1006/jmbi.1999.3091

Chen, Y. (2014). Bioinformatics Technologies. Berlin: Springer Berlin.

Landau, M., Mayrose, I., Rosenberg, Y., Glaser, F., Martz, E., Pupko, T., & Ben-Tal, N. (2005). ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. Nucleic acids research, 33(Web Server issue), W299–W302. doi:10.1093/nar/gki370

Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T., & Ben-Tal, N. (2016). ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic Acids Research, 44(W1), W344-W350. doi: 10.1093/nar/gkw408

Glaser, F., Pupko, T., Paz, I., Bell, R., Bechor-Shental, D., Martz, E., & Ben-Tal, N. (2003). ConSurf: Identification of Functional Regions in Proteins by Surface-Mapping of Phylogenetic Information. Bioinformatics, 19(1), 163-164. doi: 10.1093/bioinformatics/19.1.163

Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T., & Ben-Tal, N. (2016). ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic acids research, 44(W1), W344–W350. doi:10.1093/nar/gkw408

Pupko, T., Bell, R., Mayrose, I., Glaser, F., & Ben-Tal, N. (2002). Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics, 18(Suppl 1), S71-S77. doi: 10.1093/bioinformatics/18.suppl_1.s71

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Molecular Biology And Evolution, 35(6), 1547-1549. doi: 10.1093/molbev/msy096

Xu, D., Jaroszewski, L., Li, Z., & Godzik, A. (2015). AIDA: ab initio domain assembly for automated multi-domain protein structure prediction and domain-domain interaction prediction. Bioinformatics (Oxford, England), 31(13), 2098–2105. doi:10.1093/bioinformatics/btv092

Rychlewski, L., Jaroszewski, L., Li, W., & Godzik, A. (2000). Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein science : a publication of the Protein Society, 9(2), 232–241. doi:10.1110/ps.9.2.232

Jaroszewski, L., Rychlewski, L., Li, Z., Li, W., & Godzik, A. (2005). FFAS03: a server for profile--profile sequence alignments. Nucleic acids research, 33(Web Server issue), W284–W288. doi:10.1093/nar/gki418

Webb, B., & Sali, A. (2016). Comparative Protein Structure Modeling Using MODELLER. Current protocols in bioinformatics, 54, 5.6.1–5.6.37. doi:10.1002/cpbi.3

Šali, A., & Blundell, T. (1993). Comparative Protein Modeling by Satisfaction of Spatial Restraints. Journal Of Molecular Biology, 234(3), 779-815. doi: 10.1006/jmbi.1993.1626

Wu, S., & Zhang, Y. (2008). MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. Proteins, 72(2), 547–556. doi:10.1002/prot.21945

Chen, C., Hwang, J., & Yang, J. (2009). (PS)2-v2: template-based protein structure prediction server. BMC Bioinformatics, 10(1), 366. doi: 10.1186/1471-2105-10-366

Chen, C. C., Hwang, J. K., & Yang, J. M. (2006). (PS)2: protein structure prediction server. Nucleic acids research, 34(Web Server issue), W152–W157. doi:10.1093/nar/gkl187

Källberg, M., Wang, H., Wang, S., Peng, J., Wang, Z., Lu, H., & Xu, J. (2012). Template-based protein structure modeling using the RaptorX web server. Nature protocols, 7(8), 1511–1522. doi:10.1038/nprot.2012.085

Yang, Y., Faraggi, E., Zhao, H., & Zhou, Y. (2011). Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. Bioinformatics (Oxford, England), 27(15), 2076–2082. doi:10.1093/bioinformatics/btr350

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., … Schwede, T. (2018). SWISS-MODEL: homology modeling of protein structures and complexes. Nucleic acids research, 46(W1), W296–W303. doi:10.1093/nar/gky427

Yang, J., & Zhang, Y. (2015). Protein Structure and Function Prediction Using I-TASSER. Current protocols in bioinformatics, 52, 5.8.1–5.8.15. doi:10.1002/0471250953.bi0508s52

McGuffin, L., Adiyaman, R., Maghrabi, A., Shuid, A., Brackenridge, D., Nealon, J., & Philomina, L. (2019). IntFOLD: an integrated web resource for high performance protein structure and function prediction. Nucleic Acids Research. doi: 10.1093/nar/gkz322

Kelley, L., Mezulis, S., Yates, C., Wass, M., & Sternberg, M. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. Nature Protocols, 10(6), 845-858. doi: 10.1038/nprot.2015.053

Pettersen, E., Goddard, T., Huang, C., Couch, G., Greenblatt, D., Meng, E., & Ferrin, T. (2004). UCSF Chimera?A visualization system for exploratory research and analysis. Journal Of Computational Chemistry, 25(13), 1605-1612. doi: 10.1002/jcc.20084

Lüthy, R., Bowie, J., & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. Nature, 356(6364), 83-85. doi: 10.1038/356083a0

Colovos, C., & Yeates, T. (1993). Verification of protein structures: Patterns of nonbonded atomic interactions. Protein Science, 2(9), 1511-1519. doi: 10.1002/pro.5560020916

Concanavalin A-like lectins/glucanases superfamily. (2019). Retrieved from http://supfam.org/SUPERFAMILY/cgi-bin/scop.cgi?sunid=49899

Blouin, C., Butt, D., & Roger, A. J. (2004). Rapid evolution in conformational space: a study of loop regions in a ubiquitous GTP binding domain. Protein science : a publication of the Protein Society, 13(3), 608–616. doi:10.1110/ps.03299804

Regad, L., Martin, J., Nuel, G., & Camproux, A. C. (2010). Mining protein loops using a structural alphabet and statistical exceptionality. *BMC bioinformatics*, *11*, 75. doi:10.1186/1471-2105-11-75