In silico modelling and structural dynamics of PSY1R kinase domain and its interaction with SERK1 kinase

By
Noshin Nawer Ruhee
Student ID: 15136019

A DISSERTATION SUBMITTED TO BRAC UNIVERSITY IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF BACHELOR OF SCIENCE IN BIOTECHNOLOGY

Biotechnology Program
Department of Mathematics and Natural Science
September, 2021

DECLARATION

I hereby solemnly claim that the thesis project titled "*In silico* modelling and structural dynamics of PSY1R kinase domain and its interaction with SERK1 kinase." submitted by the undersigned student has been carried out under the supervision of MHM Mubassir, Lecturer, Biotechnology Program, Department of Mathematics and Natural Sciences, BRAC University.

The presented dissertation is an original research work carried out by myself and has not been submitted to any other institution for any degree or diploma. Any reference to work done by any other person or institution or any material obtained from other sources has been accordingly cited and referenced.

Noshin Nawer Ruhee
(Candidate)

Certified
(M H M Mubassir)
Supervisor
Lecturer
Biotechnology Program
Department of Mathematics and Natural Sciences
BRAC University, Dhaka

Acknowledgment:

My warmest gratitude to my parents, Md. Rezaul Karim and Mahmuda Khatun and my younger sister Subah Nawer Rethee for always believing in me and for their constant care throughout this journey of research and also in all other areas of life. They have always been my greatest support to lean on at the time of crisis and my cheerleader during my accomplishments.

My sincere thanks to my supervisor M H M Mubassir, Lecturer, Biotechnology Program, Department of Mathematics and Natural Sciences, BRAC University for giving me this opportunity to work under his supervision as his thesis student, for undertaking this task, and for assisting me in completing this research study.

I would like to sincerely thank Raghib Ishrak Alvy for his suggestions and counsel regarding this research work whenever it was needed.

I also want to thank my youngest aunt Mahbuba Khatun Lipi and my youngest uncle Md. Atiqur Raham for their constant motivation to keep pursuing my dreams.

Last but not least, I want to remember and acknowledge my friends Aorchie Siddique, Eera Ashrafy Anonna, and Anika Tasfia Rodoshi and thank them for their words of encouragement, support, advice, and every kind of help that is possible and for not letting my spirit down.

Noshin Nawer Ruhee

Abstract:
Due to plants' immobility, they are continuously faced with different challenges from their surroundings that also include a wide range of pathogens dealing in the environment. Hence by unraveling the mystery behind how plants tackle all these adversities and understanding the theory behind it would be a great step towards comprehending the mechanism that makes plant disease resistant. The two most important ways by which defenses are activated in plants are by structural interaction between the pathogen-associated molecular pattern (PAMP) known as the pattern-triggered immunity (PTI) and secondly via effectors known as effector-triggered immunity (ETI). In PTI, PRRs detecting PAMPs are employed along with the co-receptor proteins to combat the pathogens. PSY1R, a leucine-rich repeat receptor-like kinase (LRR-RLK) which has only been shown to act to regulate cell expansion and growth previously, has been recently found to have an important role in the immunity system of the plant by impacting in an antagonistic manner on the plant triggered immunity (PTI). This research aims to acquire a better understanding of the 3D structure of PSY1R receptor kinase and to predict its probable active sites for developing a clear perception of its impact on plant immunity by observing its interaction with receptor SERK1 kinase. Keeping this purpose in mind, modelling of these kinase proteins followed by docking and molecular dynamics (MD) simulation using GROMACS software suite have been performed. Different tools and servers have been used for the modelling of PSY1R and SERK1 kinase generating a number of 3D models. After undergoing verification processes, followed by observing the RMSD and radius of gyration graph at 5ns MD simulation the best model for PSY1R kinase and SERK1 kinase have been selected for docking. The docking result shows that PSY1R and SERK1 kinases interact on different levels with forming different kinds of bonds both before and after the MD simulation conducted at 10ns. The interactions between the two proteins were also analyzed using the protein interaction calculator (PIC) and it has been found that after the MD simulation the number of hydrogen bonds formed between them almost became half. Starting with 57 H-bonds before the simulation, whereas only 21 afterward. These changes are significant indicators of conformational changes that take place over the simulation period and are vital in understanding the early events of PTI by the receptor kinase protein.

Table of Contents:

Chapter 4 - Result and Discussion

List of tables:

List of Figures:

Chapter 1: Introduction

1.1 Significance of this study :

Plant defenses are enhanced through two main ways-first through antimicrobial compounds and secondly PTI. The first way invokes issues regarding biosafety whereas the second does not. This is why it is an important sector to study when considering improvements to a plant's defense mechanisms.

Several reports on plant growth and defense systems depict how sulfated peptides, acting as important signaling molecules, are employed by the plants in the growth and development processes along with stress responses. As plants continuously fight different environmental stresses, it is of prime importance to respond quickly to these stresses. But unfortunately, these stress management processes may start to compromise the overall plant's health and growth leading to stunted growth and even cell death.

PSY1R, the leucine-rich receptor-like kinase, regulates the expansion of cells and the activity of the plasma membrane by phosphorylation. Meanwhile, even though this mechanism has been discussed in detail, the recent studies showing the involvement of PSKR1 and PSY1R in the plant defense system have not been discussed properly, and presumably, there is yet a lot more to discover. The fact that they are found to modulate salicylate and jasmonate, two opposing pathways in an antagonistic manner acting against the biotrophic and necrotrophic pathogens has garnered some attention lately.

In some recent findings, a scenario has been insinuated where PSKα and/or PSY1 perception induced by PAMPS leads to the downregulation of SA-related responses that work against biotrophic pathogens. Again, an over-induction of this signaling pathway would result in compromising the fitness of the plants and would leave them vulnerable to necrotrophic pathogens.

Thus the modelling of PSY1R is imperative. Also, it is very important to see the interactions between this protein and its recruited co-receptor protein SERK1, in order to fully understand the mechanism of the first layer of defense in plants.

1.2 Literature review:

This section shows the overview of plant pattern triggered immunity mediated by the PSY1R kinase. Different methods and computational techniques are also briefly discussed here .

1.2.1 Plant immune system:

The two layers of defense systems working in favor of the plant immune system are PTI(pattern triggered immunity) and ETI (effector-triggered immunity). The defense mechanism processes were clearly illustrated by Jonathan in 2006 via a zigzag model.

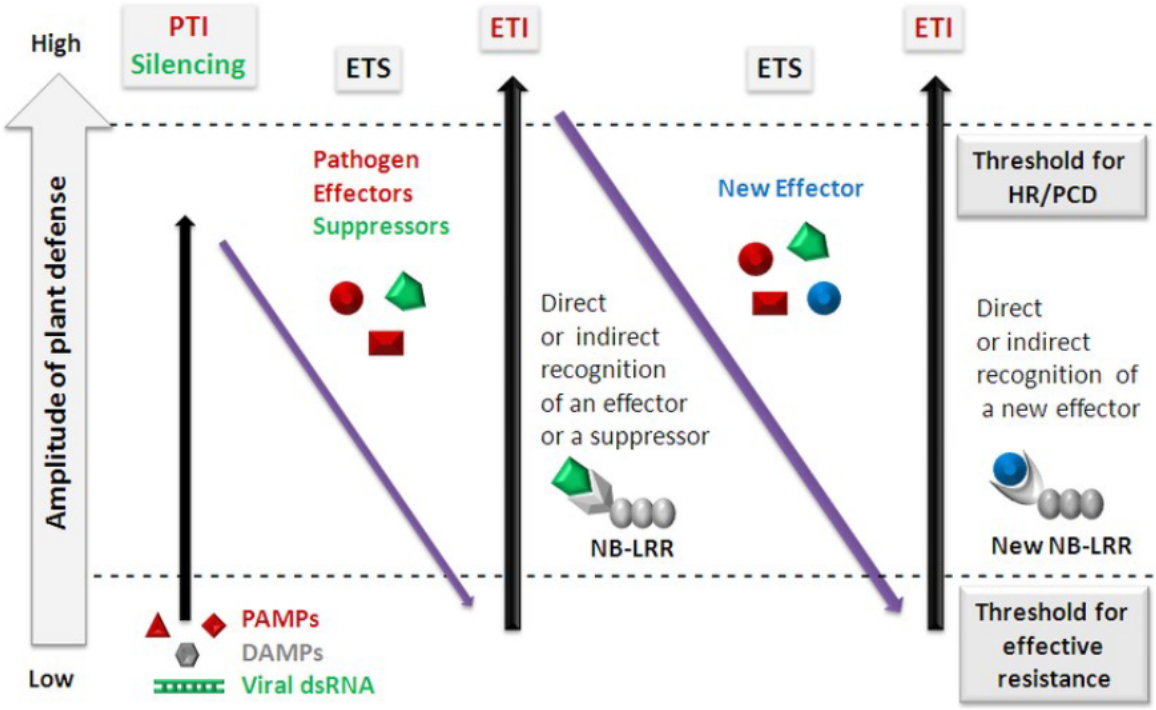

Figure 1.1: Plant immunity system (PTI and ETI) (Jones & Dangl, 2006).

This model shows how plant immunity is conferred in four phases where in the first phase, different pathogen-associated patterns (PAMPs) or microbes-associated molecular patterns (MAMPs) are recognized by different specified pattern recognition receptors (PRR) of the plant resulting in plant triggered immunity(PTI). The pathogens that deceivingly escaped the first phase of the immune system enter the second phase which is the effector susceptibility(ETS).
Pathogens manage to escape PTI by deploying the effectors which are then recognized by nucleotide-binding leucine-rich receptors (NB-LRRs) that activate ETI, the third phase of the

zigzag model. In the final phase, the pathogen by gaining new effectors can again work in order to suppress ETI.

1.2.2 Pattern Triggered Immunity:

Plant innate immune system which is the first branch immunity of plant 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 (RKs) which comprise a ligand-binding ectodomain, a transmembrane domain, juxtamembrane domain, and a cytoplasmic kinase domain. The second is receptor-like proteins (RLPs) 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 repeats.
(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 binding certain PRRs homodimerizer and form heterooligomeric complexes with other PRRs (Lee et al.,2017). 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), and also contribute to basal immunity by restricting infection of adapted pathogens in susceptible hosts (Lee et al.,2017).

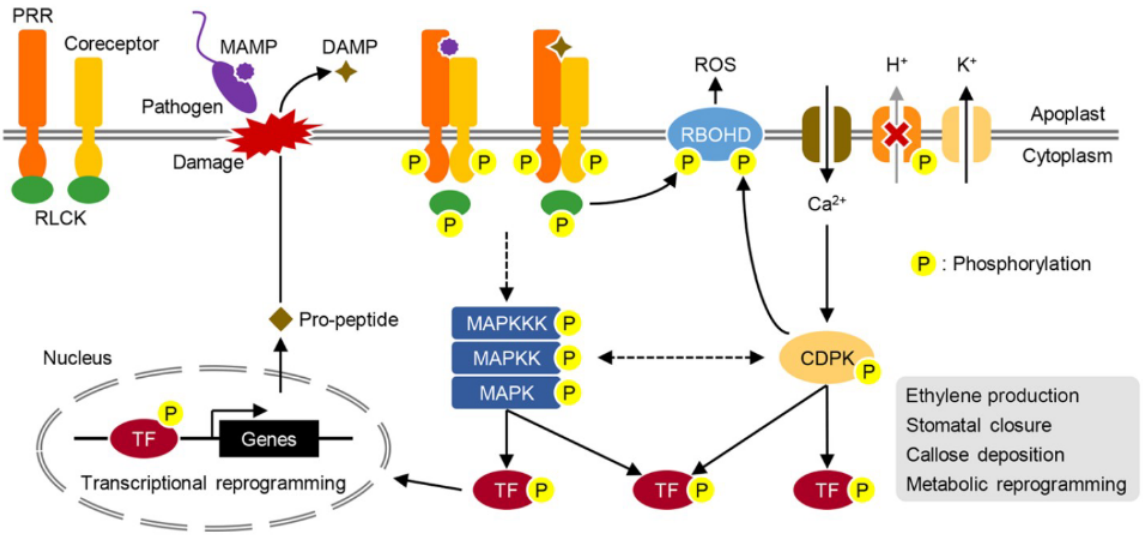

Figure 1.2: Pattern Triggered Immunity Overview (Saijo, Loo & Yasuda, 2018)

The evolution of the 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 toas 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.

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's innate immune system, while DAMPs are derived from host cells and both initiate and perpetuate innate immune responses (Choi & Klessig, 2016).

1.2.3 PSY1R kinase and Plant Immunity:

Plants being in an immobile state are left to defend against the pathogens prevailing in the environment just by themselves. With years of evolution, plants have developed a family of unique membrane receptor kinases for expanding the development of their organs and also to fight against pathogens by forming the primary line of defense in the plant system. In the last few years, the physiology of several receptor kinases was taken under inspection, ultimately concluding that members of this massive gene family are involved in plant growth, development, and also its adaptation to the environment. Recent research in understanding plants right from the molecular level includes mechanisms of lysin motif or leucine-rich repeats sensing ligands on a broad spectrum. With advances in this research sector, it has been found that PSKs which are first thought to be secreted cell proliferation promoting peptides have much more to add to the picture. Hence gaining the knowledge behind it would be a great step towards the process of making plants disease resistant (Mosher, S., & Kemmerling, B., 2013).

The receptor-like kinases PSKR1 and PSKR2 are critical for the binding of PSKα and also for physiological consequences of PSKα perception including root growth promotions and the generation of callus tissue (Mosher, S., & Kemmerling, B., 2013). PSKR mutations have been shown to induce early senescence in plants which is a salicylate(SA) associated response and

also shown to be impaired in wound healing which is a jasmonate (JA)- associated response suggesting a connection between phytosulfokine signaling and the SA/JA homeostasis (Mosher, S., & Kemmerling, B., 2013).

PSY1, an 18-amino acid sulfated and glycosylated peptide, perceived by a related receptor-like-kinase (RLK) mentioned as PSY1R, similarly to PSKα, is another peptide having similar physiological roles related to PSKα like cellular proliferation and differentiation. It was also identified as a secreted peptide, purified from plant cell culture media. It has been observed that even though mutations in the individual receptors has no morphological phenotypic effect, mutations in all of the three receptors- PSKR1, PSKR2, and PSY1R end in slight dwarfism because of the lessening number of the cells and reduction in their size (Mosher, S., & Kemmerling, B., 2013).

PSKα and PSY1 sulfation that is found to be critical for receptor binding and performance is mediated by TPST which is a singular Golgi-localized tyrosylprotein sulfotransferase. Any kind of depletion in TPST function for any reason effectuates dwarfism and premature senescence that further verifies the significance of sulfated peptide signaling in maintaining the balance in cellular homeostasis. There are instances of the involvement of the receptors, especially PSKR1 and PSY1R in PAMP responses (Mosher, S., & Kemmerling, B., 2013). The activation of PTI, i.e, PAMP-triggered immunity involves the assembly of reactive oxygen species, pathogenesis-related genes build-up 2-hydroxybenzoic acid accumulation, callose deposition, and seedling growth suppression. The Salicylate Associated response pathway which is associated with the biotrophic pathogens is anticipated to work in an opposing way to the Jasmonate Associated defense pathways working against the necrotrophic pathogens.

Mosher, S., & Kemmerling, B. (2013) in their study have shown that plants lacking PSKR1 and PSY1R have heightened resistance to the biotrophic pathogen Pto DC3000. Alternatively, it has also been found that psyr1, pskr1, and pskr2 mutant plants inoculation with *Alternaria brassicicola* which is a necrotrophic fungal pathogen, separately and also additively, has enhanced susceptibility when compared with wild-type plants that in turn suggests a partially redundant role for PSKα and PSY1 signaling.

The Molecular analysis revealed that upon Pto DC3000 inoculation, pskr1/pskr2/psy1r mutants accumulate elevated levels of salicylate and SA-responsive PR-gene transcripts. Somehow, PDF1.2 and OPR3 genes, which are the JA-responsive genes, were significantly repressed (Mosher, S., & Kemmerling, B., 2013). All these findings indicate a correlation of enhanced levels of SA signaling with increased resistance to the biotrophic pathogen Pto DC3000 and, antagonistically, decreased levels of JA signaling with resistance loss to the *Alternaria brassicicola*, a necrotrophic pathogen. Taken together, all these results portray a scenario where sulfated peptide signaling shifts the balance of defense signaling toward JA responses (Mosher, S., & Kemmerling, B., 2013).

In their research from the molecular analysis that has been revealed that upon Pto D3000 inoculation, boosted levels of salicylate and SA-responsive PR-gene transcripts have been recorded being accumulated by the pskr1/pskr2/psyr1 mutants. But on the other hand genes responsible for JA responses have been found significantly repressed.

These findings draw an analogy between enhanced levels of SA signaling with increased resistance to the biotrophic pathogen Pto DC3000 and, antagonistically, decreased levels of JA signaling with resistance loss to the *Alternaria brassicicola*, a necrotrophic pathogen, altogether referring to a scenario where sulfated peptide signaling shifts the balance of defense signaling toward JA responses in the wild type variants of the plant.

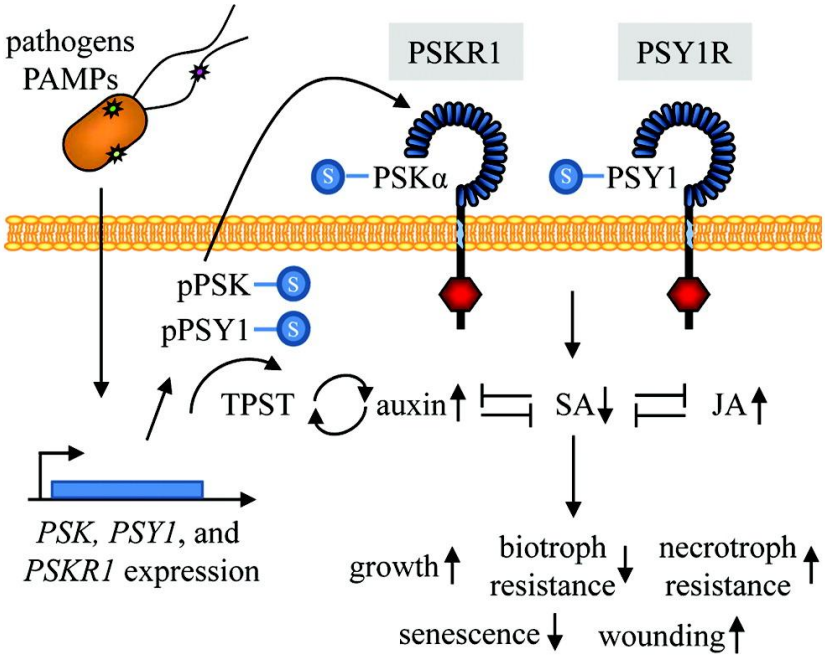

Figure 1.3: PSY1 and PSKR1 expression due to *Pseudomonas syringae* and PAMPS (Mosher, S., & Kemmerling, B. (2013))

This figure depicts the tyrosine-sulfated peptide signaling in plant defense clearly. *Pseudomonas syringae* pv tomato DC3000 infection and PAMPs that are depicted as stars on the bacterial cell induce *PSK* and *PSY1* expression. The preproproteins pPSK and pPSY1, are tyrosine-sulfated (-S) by TPST within the Golgi body which are later thought to be proteolytically processed by subtilisins. PSKα-S and PSY1-S peptides, being fully processed, finally bind to their respective receptors, PSKR1 and PSY1R. The extracellular leucine-rich repeat domain, the transmembrane domain, and the cytosolic domain of the receptors are presented in the color blue, light blue and red respectively.

From the figure, it can be seen how the activation of the receptors triggers the repression of SA signaling downwards which in consequence suppresses resistance towards the biotrophic pathogens and also impedes senescence. Again since the JA signaling is upregulated consecutively, resistance towards the necrotrophic pathogens and wounding responsiveness is also increased at a noticeable rate. Plant growth is also promoted by PSKα and PSY1, which is more or less relying on auxin. While auxin signaling invokes TPST expression, TPST activity elevates the upregulation of auxin signaling, which in turn implies that auxin and TPST pathways are also at interplay in the whole scenario of plant regulation and defense. PSKα and PSY1 signaling, initially thought to be just linked with plant growth and development, are surprisingly found to play a noticeable role in modulating stress responses like wound repair, the repression of stress-related gene induction and PAMP signaling. The recent studies have begun to underscore those areas played by these signaling molecules and there is a lot left yet to be explored.

1.2.4 PSY1R kinase and SERK1 kinase:

From the research done in the past it has been observed that BAKI and members of the SERK family participate as co-receptors for LRR-RLKs including BRI1, FLS2. In order to find out if SERK proteins can also help regulate activity of PSY1R by the means of phosphorylation , an experiment has been performed along with GST-tagged constructs of the intracellular domain of the SERK proteins kSERK and kBAK(Oehlenschlæger, C. et al, 2017). It was also tested if the inactive H6-kPSY1R K831A protein can undergo transphosphorylation via this process and for detecting transphosphorylation, pThr antibody had been used(Oehlenschlæger, C. et al, 2017). From the result, it was found that H_6-kPSY1R K831A can be phosphorylated by the intracellular domains of SERK1, SERK2, BAK1, and SERK4 but not by kSERK5, showing poor autophosphorylation occurrences(Oehlenschlæger, C. et al, 2017) .

For testing if kPSY1R had the ability for transphosphorylation of SERK protein, H_6-mBAK1 K317A was constructed which was an inactive H_6-tagged version of the intracellular domain of BAK1 (Oehlenschlæger, C. et al, 2017). Later in the result it was found that H_6-BAK1 K317A was transphosphorylated by kPSY1R. From all these results, it was concluded that one or more of the SERK proteins act as co-receptors for PSY1R through a trans-activation mechanism (Oehlenschlæger, C. et al, 2017).

Further investigation was conducted to find if PSY1R is able to form protein-complexes with SERK proteins with the help of BiFC assay (Oehlenschlæger, C. et al, 2017). It was found that in the plasma membrane PSY1R interacts with SERK5, BAK1, SERK1, and SERK2 to varying degrees whereas it hardly interacts with SERK4 under the BiFC assay (Oehlenschlæger, C. et al, 2017). Hence in this project SERK1(Somatic embryogenic receptor kinase1) has been used for performing docking with PSY1R kinase.

1.2.5 Protein Modeling:

Protein structure prediction computationally provides three-dimensional protein structures that are predicted by different *in-silico* techniques. This protein modeling relies on principles and structures from known protein structures obtained via NMR Spectroscopy, x-Ray crystallography, as well as from physical energy functions.

Approaches to predict a protein's tertiary structure:

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

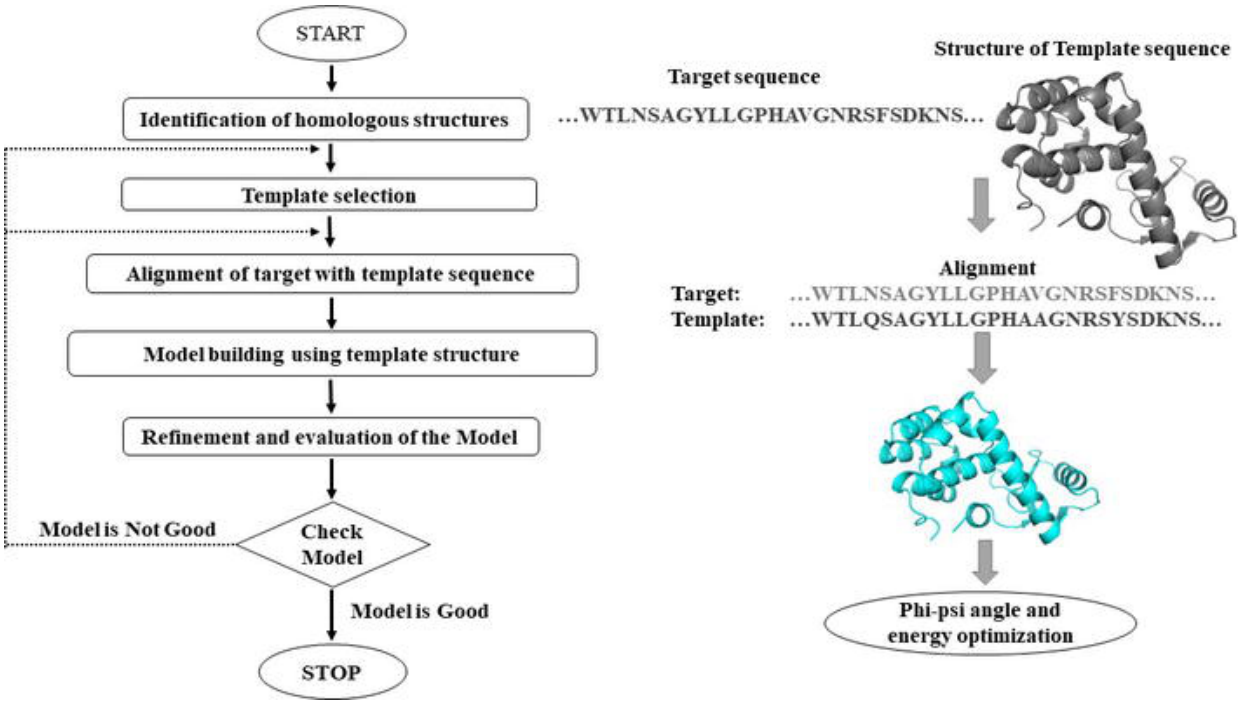

Figure 1.4 : Pipeline showing homology modelling steps (Kaczanowski & Zielenkiewicz, 2009).

Homology and threading modeling are relied upon the best templates selected, unlike ab-initio modeling. Homology modeling employs the idea that evolutionarily related proteins have matching portions of sequences and often similar structures (Kaczanowski & Zielenkiewicz, 2009). Protein tertiary structures are shown to be much more conserved among homologues sequences provided that the sequence identity does not fall below 20% (Chothia & Lesk, 1986). In homology modeling, a 3D protein model of a target sequence is generated by extrapolating enough experimental information from evolutionary-related protein structures serving as templates which can be one or more in number. The better the match, the more the accuracy.

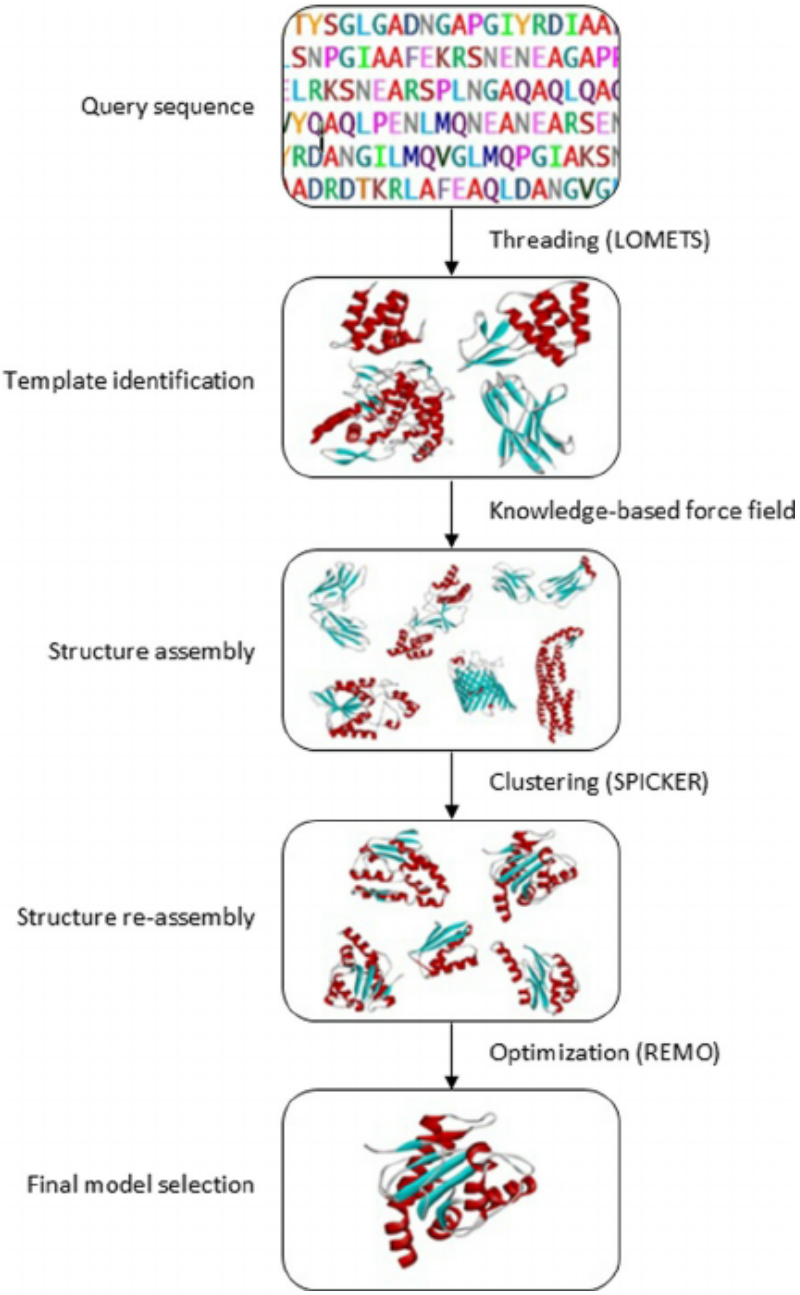

Figure 1.5: Pipeline for threading method of protein modeling (Bowie, Luthy & Eisenberg, 1991).

For threading prediction, the process primarily consists of the assignment of the fold, alignment of target-template, building the model, and evaluating it. In template-based modeling, the critical step is searching and selecting the correct templates from the PDB bank containing identical folds as the query protein is given and to make proper alignment with as few gaps as possible between the template structure and the query sequence. This process is known as fold

recognition or threading. The threading method has been used for modeling targets that have similar folds as proteins with known structures but in short of homologous sequences (Bowie, Luthy & Eisenberg, 1991). It is based on the fact 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 that are based on a wide range of approaches, as for example the sequence profile-profile alignment, structural profile alignment, hidden Markov models, machine learning, and pairwise potentials with optimal searching. (Wu & Zhang, 2008).

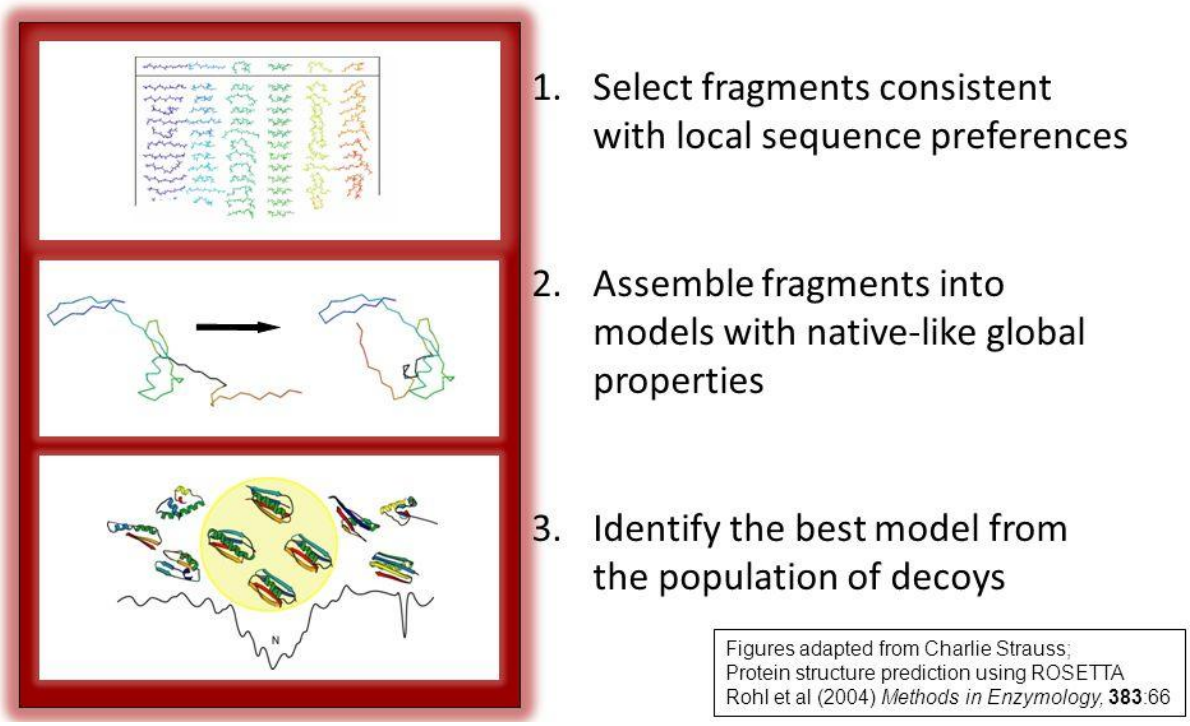

Figure 1.6: Ab initio method of protein modeling (Lee, Freddolino & Zhang, 2008).

Ab initio modeling commences with the primary amino acid sequence that is being searched for various conformations with a designed energy function taken into consideration. Later on, probable conformations known as decoy structures are produced from which native-like conformations are selected based on their thermodynamic stability and energy state (Lee, Freddolino & Zhang, 2008). This approach is grounded on the thermodynamic hypothesis which was proposed by Anfinsen. Accordingly, under a given set of conditions, the structure that corresponds to the global free energy minimum is the native structure.

1.2.6 Molecular dynamics simulation:

All classical simulation methods have the root of empirical sets of parameters known as force fields. These force fields calculate interactions and evaluate the potential energy of the system considering it as a point-like atomic coordinate function. It comprises equations set in order to calculate the potential energy and particle coordinate forces and also parameters set to be utilized in the equations.

Bonded interactions and non bonded interactions are the two classes the potential functions are divided by almost all common force fields. Angle-bending, stretching of covalent bonds, torsion potentials when rotating around bonds and out-of-plane "improper torsion" potentials can be fixed throughout a simulation. These belong to the bonded interactions. The remaining are the nonbonded interactions between atoms that are close in space containing Lennard–Jones repulsion and dispersion also known as the Coulomb electrostatic. These are typically computed by neighbor lists updated periodically. Meanwhile, during the energy minimization step in the process, the steepest descent algorithm simply moves each atom a brief distance in direction of decreasing energy (the negative gradient of energy), and molecular dynamics is performed by integrating Newton's equations of motion.

The tiniest chemical sample that can be imagined can be way too large to be incorporated completely during a simulation. For avoiding surface artifacts biomolecular simulations normally use periodic boundary conditions. The size of the box is taken proportionately as being quite large will hinder significant interactions of the molecules with their periodic copies. Apart from this, errors including cutoffs and rounding can give rise to small drifts in energy resulting in the entire system heating up during the simulation. In spite of having a theoretically perfect setup, we might face different errors and problems since molecular dynamics simulation is generally commenced with an imperfect molecular structure.

There are different parameters throughout the simulation process that are needed to be taken care of. For example, as the potential energy of this structure decreases during the simulation, the kinetic energy (i.e., temperature) would increase if the entire system energy has to remain constant. In order to regulate this, the system is generally coupled to a thermostat that scales velocities to take care of the temperature. Similarly, the entire pressure within the system is often adjusted through scaling the simulation box size, either isotropically or separately in x/y/z dimensions. Another most intensive part of simulations is the computation of the nonbonded interactions since millions of pairs have to be evaluated for each time step. By extending the time step the simulation performance can be improved, when the desired result is not met.

1.2.7 Protein-protein docking:

Molecular Docking is referred to as the process of predicting the orientation and binding of one molecule with another one in order to build a stable complex. Molecules include proteins, nucleic acids, carbohydrates, etc. The receptor molecule is the receiving molecule most probably a protein with which the partner molecule binds. The partner molecule can be a ligand or even another protein. When the receptor protein molecule binds with a ligand, it is known as protein-ligand docking whereas when the receptor protein binds with another protein it is called protein-protein docking. In the case of protein models in low-resolution structures, the steric clash is the major artifact prevailing as a result of the unnatural overlap of any two nonbonding atoms in those structures. Therefore, before performing the docking procedure the selected input structures should be subjected to energy minimization. Protein-protein docking is very simple to conduct using different web servers and docking tools since it does not demand a lot of time or serious high configuring resources relative to the molecular dynamics simulation processes. The end result is the 3D structure of a protein-protein complex in some very plausible orientation and binding formation. Protein-protein interactions are very significant in biological science to understand the method of action of the receptor protein in a specific pathway or just its metabolic functions. Even though every detail and characteristic of protein dynamics cannot be addressed by protein-protein docking, especially the changes in the global conformation while forming protein complex, here in this context of in-silico approaches, it is a complementary method to the experimental procedures to elucidate and to understand better the 3D structures of the protein complexes.

Chapter 2: Tools and Databases

2.1 A brief description of the Databases and Tools used in this project :

1. PDB:

The Protein Data bank provides access to digital data resources that include different types of biomolecules such as proteins, DNA and RNA. (Berman et al., 2000) With the help of an internet information portal, the contents can be downloaded and used for analyses and other research works. These data are generally obtained with authenticity by various experimental procedures including X-ray crystallography, NMR spectroscopy, or cryo-electron microscopy, and are deposited by researchers and scientists from around the world. The vital information that is stored in the PDB archive is in the form of coordinates for the molecules. Data about the atoms in each protein and their location with the 3D structure are contained in this file in several formats like PDB, mmCIF, XML. The structure gets an alphanumeric identifier of four characters which is called the PDB ID when it is published in the database. One can look for that structure by searching by its PDB ID (Berman et al.,2000). Databases such as SCOPand CATH also use data by deriving from the PDB bank.

2. NCBI:

The National Center for Biotechnology Information (NCBI) provides a variety of data for genes, genomes, proteins, and chemicals that are relevant to the medical and scientific communities hence is a significant resource for research tools and services. A wide variety of data analysis tools that are capable of manipulating, aligning, visualizing and evaluating biological data are also provided and accessible by NCBI. Alongside this, the world's largest repository of full-text articles, scientific abstracts, books, and reports are also included in the NCBI's literature resources. The NCBI database is one of the largest data repositories for biological science and has been providing useful information and resources to the research community since 1988 with an average of 2.5 million users from around the world accessing through them.
(Fiorini, N. & Lipman, 2017).

3.UniProt:

Standing for Universal Protein Resource, UniProt is a comprehensive resource for protein sequences and annotation data (The UniProt Consortium, 2018). The UniProtKnowledgebase (UniProtKB), the UniProt Reference Clusters (UniRef), and the UniProt Archive(UniParc) make up The UniProt databases (The UniProt Consortium, 2018). With the team effort and participation of the European Bioinformatics Institute (EMBL-EBI), SIB Swiss Institute of Bioinformatics and Protein Information Resource (PIR) UniProt was founded in 2003. Since then Uniprot has been working as an open-source database for protein structures with very

detailed and in-depth information available that is being used for different research purposes from time to time.

4. NCBI BLAST:

In order to look for regions of local similarity and to find homologous sequences the Blast Local Alignment Search Tool i.e., BLAST is used (Altschul, Gish, Miller, Myers & Lipman, 1990). In this process nucleotide or protein sequences after being compared to sequence databases, the statistical significance of matches are calculated. BLAST offers a significant role to deduce both functional and evolutionary relationships among the sequences and helps identification of the members belonging to the same gene families. Several types of BLASTs processes are available in the NCBI BLAST according to the type of the input sequences such as Nucleotide-nucleotide BLAST (blastn), Protein-protein BLAST (blastp), Position-Specific Iterative BLAST (PSI-BLAST). The database in comparison to what is being searched can apply through the selection boxes. For example, selecting the PDB database for protein-protein blasts in order to find homologous sequences of the query amino acid sequences that also have 3D structures deposited.

5. PROSITE :

PROSITE is a bioinformatics tool and database for protein formed in 1988 by Swiss bioinformatician Amos Bairoch which is an annotated collection of descriptors of the protein motifs and is employed for distinguishing protein families and their domains. PROSITE also identifies probable functions of proteins that are newly discovered and analyzes already identified proteins for any undetermined activity. Profiles or patterns obtained from multiple sequence alignments of homologous sequences are utilized as the motif descriptors in PROSITE. PROSITE patterns being motifs of short sequence, limited to short specific regions with a high rate of sequence familiarity, specialize in predicting protein functions whereas PROSITE profiles are score matrices that are position-specific which cover the entire domain are useful for anticipating the constructive characteristics of the proteins, both comprising features that are complementary to each other.

6.ProtPram:

The quantitative and qualitative estimation of different physical and chemical characteristics of a protein deposited in different databases or of a user-entered protein sequence can be performed using ProtPram (Gasteiger et al., 2005). The calculated and evaluated parameters that are presented afterward include the molecular weight, atomic composition, extinction coefficient, pI, amino acid composition, estimated half-life, instability index, aliphatic index and grand average of hydropathicity (GRAVY). 7. Pepstats:

Pepstats is a bioinformatic server that calculates statistics of protein properties. Pepstats analyses thoroughly one or more protein sequences and provide an output file with various statistics on the protein properties including the total number of residues, residue charge, isoelectric point,

The average weight of residue, molecular weight, etc. It is a useful tool for getting a whole account of the physicochemical properties of the protein under inspection.

8. PSIPRED:

PSIPRED (PSI-blast-based secondary structure PREDiction) is a secondary structure prediction tool for proteins that executes analysis on the results derived from PSI-BLAST which is the Position-Specific Iterated -BLAST (Jones, 1999). PSIPRED functions by relying on the details of the proteins that are evolutionarily related to each other which makes it possible to predict the beta-sheets, alpha helices, and coils of a new amino acid sequence inserted in the query box. PSI-BLAST, by identifying the evolutionarily related sequences, forms a position-specific scoring matrix that is in turn managed by an artificial neural network that was built to speculate the input sequence's secondary structure (Chen, 2014).

9. InterPro:

InterPro served as a database for families, domains and functional sites for proteins. InterPro applies an identifiable feature from a known protein to a new protein with just primary sequence and characterizes the functions of the query protein (Hunter and Apweiler, 2009). The database of InterPro has many other databases integrated from where it derives information regarding the query protein. The other databases include Gene3D, PANTHER, Pfam, PIRSF, PRINTS, ProDom, PROSITE, SMART, SUPERFAMILY and TIGRFAMs (Hunter and Apweiler, 2009). InterPro output comprises investigated signatures and protein groups that are significant matches to the new protein.

10.ConSurf:

The ConSurf server performs the estimation of the evolutionary conservation of positions of amino/nucleic acid in a protein/DNA/RNA sequence on the basis of the phylogenetic connection among the sequences that are homologous (Landau et al., 2005; Ashkenazy et al., 2016). The degree or rate to which a nucleic or amino acid position is conserved evolutionarily, which is its evolutionary rate, is strongly dependent on its structural and functional significance. Hence it is possible to find out the significance of a specific position in the amino or nucleic acid sequence in terms of structure and function, by conservation analyses done on the equivalent positions of the members belonging to the same family (Glaser et al., 2003). By being able to use an empirical Bayesian method or maximum likelihood (ML) process, ConSurf provides a higher accuracy rate while computing the evolutionary relatedness in a sequence in contrast to other methods available (Pupko, Bell, Mayrose, Glaser & Ben-Tal, 2002).

11. MUSCLE:

MUltiple Sequence Comparison by Log-Expectation (MUSCLE) is being used for aligning multiple sequences of protein and nucleotide. MUSCLE supposedly can achieve a higher rate of exactness, precision and quickness than any other multiple alignment program or servers

available, like ClustalW2 or T-Coffee. Multiple alignments of protein sequences have much significance while performing analyses like the prediction of secondary structure, construction of the phylogenetic tree, identification of critical residue and even tertiary structure prediction for proteins.

12.BoxShade:

BoxShade is a program for creating visually pleasing images of protein or DNA multiple sequence alignments. There is no alignment done by this program, it accepts an input file that is already processed by other alignment programs. Different colors and shadings represent a group of matching or identical residues in the multiple-alignment chart in standard BOXSHADE output. Different formats can be used for the output. There are also various options available concerning the kind of shading to be applied, sequence numbering and so on that can be availed by the user conveniently.

13. MEGA X:

Molecular Evolutionary Genetics Analysis (MEGA) is a computer program package in order to compute 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 a UPGMA tree, maximum parsimony tree, maximum likelihood tree(ML), neighbor-joining tree and minimum evolution tree. It allows sequence alignment by both MUSCLE and CLUSTALW (Kumar, Stecher, Li, Knyaz & Tamura, 2018).

14. IntFOLD5:

IntFOLD is an independent, integrated web server in order to predict the structure and functions of proteins (McGuffin et al., 2019). The only input required is the sequence of the amino acid of the targeted protein. The tasks that can be performed by IntFold5 include 3D model quality estimation along with options to fix errors, refinement of the structure, tertiary structure prediction, prediction of intrinsic disorder, prediction of protein-ligand binding sites, prediction of the domain. The output by this server can be later observed and downloaded for further analyses (McGuffin et al., 2019).

15. LOMETS:

LOMETS (Local Meta-Threading Server) is designed to provide protein structure prediction based on template and function annotation based on structure (Sitao Wu and Yang Zhang, 2007) This server follows the threading process of protein modelling by integrating multiple deep learning-based threading methods. LOMETS has an integrated system to look for templates for 3D modelling of protein that lacks homologous templates. The results from the LOMETS server includes a prediction of solvent accessibility, prediction of contact map and distance map by using DeepPotential, prediction of secondary structure, presenting top ten templates from around

a hundred templates for 3D modeling, full-length models of the query sequence based on the templates selected (Sitao Wu and Yang Zhang, 2007).

16. AIDA:

AIDA, Ab initio domain assembly server identifies individual and specific domain in multi-domain proteins and then assembles their tertiary structures and thus predicts their relative spatial positions and orientation based on the ab-initio folding potential (Xu, Jaroszewski, Li & Godzik, 2015). By fixing the relative positions of the domains, the AIDA server encourages the assembly and arrangement of domains positioned into other domains (discontinuous domains). Structure assembly from sequence only is also supported by this server. For this, the sequence domains are iteratively split then aligned with the PDB template presented 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).

17. FFAS03:

The FFAS03 server is an interface to the profile-profile alignment and folds recognition algorithm FFAS (Rychlewski, Li, Jaroszewski & Godzik, 2008). After accepting an input protein sequence, a profile about the protein sequence is automatically generated, which is later compared with different other sets of protein sequence profiles taken from PDB, COG, PFAM and SCOP(Jaroszewski, Rychlewski, Li, Li & Godzik, 2005). Along with the homologous sequences found in profile databases by the FFAS method, the FFAS03 server also gathers and presents homologous sequences collected by PSI-BLAST using PDB-BLAST protocol and templates from BLASTp that in turn serve as probable templates for the modeling.

18. Robetta:

Robetta server is another tool for structure prediction and analysis of unknown protein sequences. Sequences are parsed into putative domains, later by applying either homology modeling or with de novo structure prediction methods, structural models are generated. The template that is used for comparative modelling is found by confident match to an already published protein structure with the help of FFAS03, PSI-BLAST, BLAST or 3D-Jury. HHSEARCH, SPARKS, and Raptor detect and align the structure from which the comparative models are built. The loop regions being assembled from fragments are optimized in order to fit the structures of the templates that are aligned. De novo protocol of rosetta creates the de novo models which is a fully automated procedure.

19. Raptor X:

RaptorX is another server designed for the purpose of structure prediction that is created by the Xu group, especially for protein sequences that do not have close homologous sequences in the Protein Data Bank (Källberg et al., 2012). With an input sequence, one can have a prediction

output on solvent accessibility, secondary structure, disordered regions, contacts, binding sites and tertiary structures (Källberg et al.,2012). The confidence scores include P-value, GDT(Global Distance Test) and uGDT (un-normalized GDT) that help us understand the standard and accuracy of a predicted 3D Model. For alignments of difficult targeted regions or sequences having below 30% sequence identity with solved structures in the Protein Data Bank, RaptorX can be considered reliable to a great extent.

20. Swiss Model:

SWISS-MODEL is a protein structure homology modeling server that supports several inputs including only query amino acid sequence, target-template alignment file along with the user-defined template. The protein tertiary structure library is checked thoroughly for the input protein sequence to find out appropriate templates. Relying on the alignment between the input protein sequence and the sequence of the template structure, a 3D model is produced that can be viewed using the program DeepView (Waterhouse et al., 2018). The quality of the produced models is estimated using the quality assessment tools integrated within the program (Waterhouse et al., 2018).

21. (PS)2-v2:

(PS)2-v2 is an automated homology modeling server using the matrix, S2A2 which is a substitution. By combining required information from the sequence and secondary structure it tracks down homologous proteins with distant relatedness and forms alignment between the target and the template (Chen, Hwang & Yang, 2009). Later, with the help of the modelling package of MODELLER, the tertiary structure is produced (Chen, Hwang & Yang, 2006). The integrated programs ProQ and ProQres evaluate the generated models afterward on the basis of LGscores and MaxSub scores (Chen, Hwang & Yang, 2006). The model can be viewed with the Ast Viewer and the result is sent to the user via email.

22. I-TASSER:

Standing Iterative Threading ASSEmbly Refinement, I-TASSER has a threading-based principle with steps arranged hierarchically to protein structure and function prediction (Yang & Zhang, 2015). This is done by first identifying templates from the PDB by a multiple threading approach and then finally building the models on the basis of iterative template-based fragment assembly simulations (Yang & Zhang, 2015). In community-wide CASP experiments, I-TASSER being termed as the 'Zhang-Server' has been ranked first for the purpose of protein structure prediction.

23. SPARKS-X:

This is a fold recognition server that combines single fold recognition method SPARKS with SPINEX techniques incorporated for improving secondary structure prediction, backbone torsion angle and solvent accessible surface area and tertiary model generation (Yang, Faraggi, Zhao & Zhou, 2011). It recognizes a match from the same family, same superfamily and same fold

following the SCOP definition within top-nTemplates. SPARKS-X is considered as one of the best single-method fold recognition techniques available out there (Yang, Faraggi, Zhao & Zhou, 2011).

24. CEthreader:

CEthreader (Contact Eigenvector-based threader) is a protein structure prediction algorithm that is template-based, guided by contact maps, that converts contact maps predicted by ResPRE into a set of Eigenvectors of a single body with the help of the Eigen Decomposition technique (Zheng and Yang, 2019). Based on the contact Eigenvector, dynamic programming is subsequently performed. Accuracy is improved in template detection compared to traditional profile-based threading through the combination of contacts, secondary structures, and profiles (Zheng and Yang, 2019).

25. Phyre2:

Phyre2 suite is to predict and analyze protein structure based on homology modeling technique, function and mutations (Kelley, Mezulis, Yates, Wass & Sternberg, 2015). The Phyre2 server includes an advanced interface, a fully updated fold library and utilizes the HHpred / HHsearch package for the detection of the homologous sequences to be used as the templates. Additionally, it provides many other functionalities such as batch processing, one-to-one threading, and multi-template modeling (Kelley, Mezulis, Yates, Wass & Sternberg, 2015).

26. M4T:

M4T stands for Multiple Mapping Method with Multiple Templates and serves as a fully automated comparative protein structure modeling server. Two major modules are integrated into M4T, one of which is Multiple Templates (MT) and the other is Multiple Mapping Method (MMM) (Narcis and Carlos, 2007). Both modules function in a congruous method to produce a 3D model of the input sequence by the user. At first, a PSI-BLAST search is carried out with the input sequence from where the templates are selected in the MT module. Later on, sequence-to-structure alignments are performed by the MMM module and finally, the protein model is constructed by Modeller (Narcis and Carlos, 2007).

27. PRIMO:

PRotein Interactive MOdeling (PRIMO) is a server for modeling protein. In spite of having a detailed orientated process in modelling, it also allows freedom for the user to make certain decisions (Rowan Hatherley and David K. Brown, 2016). There is a default parameter already set for each step that can be changed by the user according to one's requirements with additional external input. The process in the modelling has three steps that include identifying and selecting the template, alignment of the target sequence and template, and finally modelling and evaluating the models (Rowan Hatherley and David K. Brown, 2016).

28. HHPRED:

HHPRED conducts protein structure prediction by using homology information from HH-suite that searches for sequences using hidden Markov models(HMMs) (Soding, Biegert & Lupas, 2005). A wide range of databases that include PDB, Pfam, SMART, etc is searched thoroughly to look for homologous sequences. As an input HHPRED accepts a single query sequence or a MSA file and the output result is presented in a similar way to that of PSI-BLAST results. Specific parameters can be included during the search including local alignment, global alignment, scoring, secondary structure relatedness, etc. HHPRED by building a MSA for the query sequence searches for templates that are homologues and rank them. Multiple alignments are again formed between the query sequence and the top selected templates, which are later reranked. From the top-ranked target-template alignments, templates are selected which are finally forwarded to MODELLER that generates and presents the model with the highest score (Soding, Biegert & Lupas, 2005) can be produced by HHpred.

29. MODELLER:

MODELLER is used for homology modeling to generate protein 3D models (Webb & Sali, 2016) by implementing a method encouraged by NMR (nuclear magnetic resonance spectroscopy). MODELLER needs an input sequence of the query amino acid to be modeled and single or multiple templates. The templates are the coordinates file of the 3D structures of the proteins that are homologous to the query amino acid sequence. Modelling is dependent on the sequence alignment between the query and the templates.

<u>I) Modelling by MODELLER:</u>

MODELLER follows the homology principle of modelling by offering two kinds of modelling - basic modelling and advanced modelling.

Basic Modeling is modeling a sequence to a template that has a higher percentage of similarity which is also the single template modelling. In the modeller, basic modelling is of five steps:

1. Searching for 3D structures (PDB files) related to the query sequence.
2. Selecting a template (the best one considering all other factors).
3. Aligning query sequence with the template.
4. Building the model.
5. Evaluating the models generated.

Advanced modelling is done by tasking multiple templates and the loop is modelled using *the ab-initio* method. The steps include:

1. Multiple template selection.
2. Multiple sequence alignment.
3. Building the models.
4. Model evaluation and ranking them in order.
5. Loop refinement.

For each step, there is a dedicated python script that can be modified according to the user's requirements that include different criteria. The python script is run one by one in the window's/linux's command prompt and the process is completed.

II) Model selection:

The models produced are evaluated on the basis of different scoring systems, but the most important criterion that is taken into consideration is the DOPE score. DOPE is the Discrete Optimized Potential Energy which is an atomic distance-dependent statistical potential that has been derived from a sample of native structures by utilizing probability theory, without depending on any adjustable parameters. (Shen MY, Sali A., 2006) MODELLER facilitates the use of DOPE in a wide range of functions including model assessment, loop modeling, etc. The DOPE score has been allowed to be used for both assessments of given structures as well as their refinement with other optimization methods. DOPE, along with other scoring functionals already integrated, provides us a clear idea of the 3D protein models generated by MODELLER. In the case of the DOPE score the lower the value, the better the model. Hence models that have the lowest DOPE scores are selected for further evaluation.

29. Verify3D:

Verify3D is an online model assessment tool utilized to compare the three-dimensional model with its amino acid sequence. In its mechanism, the results are compared to the very fine experimental structures (Lüthy, Bowie & Eisenberg, 1992).

30. ERRAT:

ERRAT is a web tool that can analyze the statistics of non-bonded interactions between different atom types. (Colovos & Yeates, 1993).

31. PROCHECK:

The PROCHECK suite of programs analyses protein structure stereochemistry and presents a detailed report comprising a comprehensive listing of residues. By providing an overview of the structure quality compared with refined structures, it also gives an idea of the area that needs special attention and further inspection This program is utilized for assessing the quality of the existing structures and also of those being modelled on the basis of the known structures.

32. Molprobity:

MolProbity, serving as a structure validation server, determines the quality of proteins and nucleic acids at both global and local levels. MolProbity can perform some of the corrections locally and present the results in the chart and graphical forms thus guiding to manually rebuild the molecule. By presenting a detailed report of all-atom contact analysis of steric problems it also provides an updated diagnostics of dihedral-angle. Results are presented in multiple forms

such as overall numeric scores, as PDB and graphics files, local problems charts or lists, 3D kinemage graphics which are manipulatable and can be viewed by the online KiNG viewer.

33. GROMACS:

GROningen MAchine for Chemical Simulations(GROMACS) is used for performing molecular dynamics in virtuality for systems of biomolecules, proteins, lipids, and nucleic acids, with particles ranging from hundreds to millions. This package can conduct virtual simulations even with the structures that have complex, intertwined bonded interactions and also for non-biological systems, such as polymers. By computing the nonbonded interactions dominating the simulations with great speed, GROMACS serves a greater purpose in the bioinformatics and structural biology research field by revealing physiological changes in a protein at the atomic level.

34. Xmgrace:

Xmgrace is a package for plotting purposes that is available with the Linux operating system which is a plotting tool in 2D. It allows users to manipulate plots, determines all kinds of plot parameters, modifies the presentation of the figure, and saves the figure in the format of choice.

35. Vim text editor:

Vim is a versatile text editor by being able to support different formats of text including a wide range of programming languages that have been created to manipulate and change any kind of text. Included as “vi” with most of the UNIX systems it is continuously being worked on for more improvements.

36 . Chiron:

Chiron, a protein-energy minimization server, conducts rapid energy minimization of proteins employing a quantitative approach of identification of steric clashes in proteins (Ramachandran, S., Kota, 2011). Chiron is found to be capable of resolving clashes from the homology models within 1 Å of the initial structure without drifting away from the native structure state (Ramachandran, S., Kota, 2011). The input structure that can be a PDB file of protein structure or of protein structures with ligands, first being examined for any missing atoms or side chains are reconstructed using rotamer libraries of Dunbrack (Ramachandran, S., Kota, 2011). After computing the clash score Chiron determines if the structure needs any minimization. Clash-score greater than 0.02 undergoes minimization under constrained discrete molecular dynamics (DMD) simulations (Ramachandran, S., Kota, 2011).

37. ClusPro:

ClusPro is the first web-based automated program to perform a docking with protein structures computationally. 3D-structure coordinates file of two proteins are uploaded by the user or the PDB codes of the needed structures are entered, which are then downloaded from the PDB server

(Comeau et al, 2004; Kozakov et al, 2006; Kozakov et a, 2013; Kozakov et al, 2017). Billions of probable complexes are computed using a docking algorithm from where the complexes are sorted out by filtering the ones with the good electrostatic and solvation free energies, taken for further clustering (Comeau et al, 2004; Kozakov et al, 2017). A shortlist containing around thirty complexes that are ranked according to their clustering properties are presented as output, which is also automatically sent to the user via email.

38. UCSF Chimera:

USCF Chimera has been used for visualization and analysis of molecular structures and their docking results analyzing the interaction between the proteins, trajectories, etc (Pettersen et al., 2004). Pictures and animations of finer quality can be produced that can be used for research papers, presentations and various other purposes.

39. PyMOL:

PyMOL is a python-based molecular graphics tool used for the visualization of 3D structures of biomolecules and their trajectories. It has the ability to edit molecules, ray tracing, and even making movies covering areas for visualization. Protein-ligand interactions, protein-protein interactions, molecular simulations and different structural analyses can be done through PyMol. Being based on python and having many python plugin tools by side, Pymol has enhanced utilities for studies of the biomolecular structures and drug designing.

40. VMD:

VMD, which stands for Visual Molecular Dynamics can be used for visualization, editing and analysis of biomolecules and systems. The molecular simulated trajectory files can be presented using VMD. VMD provides a Tk console that can be utilized to perform different operations on the PDB file including mutating the residues, changing the chain name, changing the residue name and atom number and many other manipulations successfully without any stress.

41. PIC:

Protein Interactions Calculator (PIC) calculates different categories of intra protein (within a protein) and interprotein interactions (among proteins in a docked complex). These include ionic interactions, hydrogen bonds, aromatic-aromatic interactions, disulfide bonds, interactions between hydrophobic residues, aromatic-sulfur interactions and cation-pi interactions. The server takes a PDB file of a single protein or protein-protein complex as an input. Hydrogen bonds on different aspects like protein-protein main chain bond, protein-protein side-chain side-chain bond and protein-protein main chain side-chain bond are calculated while computing the protein-protein interaction in a docking complex. These computed bonds that are specified with chain name, residue number and name can later be observed by manually inputting the specifications using a molecular visualization tool.

Table 2.1: Tools used and their functions at a glance

<table>
<tr><th>No.</th><th>Functions</th><th>Software/Server/Database used</th></tr>
<tr><td>1</td><td>Sequence Retrieval.</td><td>NCBI, UniProt</td></tr>
<tr><td>2.</td><td>Characterizing protein domains over the entire length of the receptor sequence.</td><td>PROSITE</td></tr>
<tr><td>3.</td><td>Determination of amino acid composition and physicochemical properties</td><td>ProtParam, Pepstats</td></tr>
<tr><td>4.</td><td>Domain architecture analysis</td><td>InterPro</td></tr>
<tr><td>5.</td><td>Prediction of secondary structure</td><td>PSIPRED</td></tr>
<tr><td>6.</td><td>Prediction of conserved regions.</td><td>ConSurf</td></tr>
<tr><td>7.</td><td>Determination of homologous and similar sequences</td><td>Blastp, PDB(database)</td></tr>
<tr><td>8.</td><td>Multiple sequence alignment</td><td>MUSCLE</td></tr>
<tr><td>9.</td><td>Presentable view of multiple sequence alignment</td><td>BoxShade</td></tr>
<tr><td>10.</td><td>Phylogenetic tree construction</td><td>MEGA</td></tr>
<tr><td>1.</td><td>Protein Modelling<table><tr><td>Single Template modelling</td><td>AIDA, FFAS03, (PS)2-v2, CEthreader, SPARKS-X, Swiss Model, RaptorX, MODELLER</td></tr><tr><td>Multiple Template modelling</td><td>I-TASSER, M4T, PRIMO, IntFOLD5, Phyre2, Robetta, HHpred, MODELLER , LOMETS</td></tr></table></td><td></td></tr>
<tr><td>12.</td><td>Ramachandran distribution plot.</td><td>PROCHECK, Molprobity</td></tr>
<tr><td>13</td><td>Non-bonded interactions statistics between different atom types</td><td>ERRAT</td></tr>
<tr><td>14.</td><td>The compatibility between an atomic 3D protein model and its amino acid sequence.</td><td>Verify 3D</td></tr>
<tr><td>15.</td><td>Molecular Dynamics Simulation</td><td>GROMACS</td></tr>
</table>

16.	Graph Production	Xmgrace
17.	Energy Minimization	Chiron
18.	Protein-Protein Docking	ClusPro
19.	Docking interaction analysis	PIC (protein interaction calculator)
20.	Editing PDB file	VMD
21.	Viewing and analyzing PDB files	UCSF Chimera, PyMol

Chapter 3: Methodology

3.1 Analysis using the sequence:

The protein sequence has been retrieved from the database in the FASTA format and has been used to perform different other analyses. The study was done for domain architecture analysis and its characterization, determination of amino acid composition and physicochemical properties, prediction of secondary structure, prediction of conserved regions.

For determining homologous protein sequences the FASTA format was inputted into NCBI Blastp. The database chosen for this action was the protein data bank (PDB). On the basis of the max score, top sequences were selected for multiple sequence alignment, for construction of the maximum likelihood phylogenetic tree, and for template selection protein in order to perform protein modelling in the next step. For a better understanding of the protein structure and formation, these pre-modelling analyses bear great importance as the data produced give us insights regarding the nature of the protein we are working on.

3.2 Protein Modelling:

For single template modelling of the PSY1R kinase domain, a total of six tools which include FFASO3, Swiss Model, SPARKS-X, RaptoxX(old server), PS2-(V2) and CEthreader were utilized. 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 a lower FFAS score indicates higher confidence of the prediction. In the Swiss Model, the 1st two templates in the list were 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). SPARKS-X on the other hand generated 10 models based on the top 10 best matches in target-template alignments. The models were ordered according to their Z-score. RaptorX, PS2-(V2), SPARKS-X, CEthreader used the threading method for modelling whereas FFAS03 and Swiss Model used the homology modelling method.

Multiple template modeling was carried out using seven tools HHPRED, IntFOLD, I-TASSER, M4T, LOMETS, AIDA 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 less the difference between P-value and zero i.e. the closer it is to zero the more significant the match is. I-TASSER generated 1 model best model instead of 5 models as in regular cases. LOMETS and I-TASSER used the threading method, AIDA used the ab initio method whereas HHPRED, IntFOLD, PHYRE2 and M4T used

the homology modelling method. Robetta and RaptorX server(new) produced 5 models each where the template selection was automated. Robetta followed both homology and ab initio modelling principle whereas RaptorX server followed threading. In total for PSY1R kinase, 51 models have been produced using a variety of template combinations.

For the modelling SERK1 kinase the tool MODELLER was used; modelling was performed with both single template and multiple templates and have been considered observing the DOPE score. The top three templates were chosen from the blastp for the purpose of modelling. In total for SERK1 kinase, 15 models were produced and 4 models with the lowest DOPE score were selected for evaluation.

3.3 Model validation:

The models produced after modelling were evaluated on the basis of the statistics of non-bonded interactions between different atom types. Whether the atomic 3D model is compatible with its primary amino acid sequences is also taken while inspecting the quality of the model. The angle of rotation of the residues and whether or not they have been positioned in an allowed or disallowed region have been observed by analyzing the Ramachandran distribution plot. Ramachandran Z score, a statistical metric, was also taken into consideration from which we can understand the normality of a model when it is compared to a high-resolution structure as the reference sets. (Hooft *et al*., 1997). It was suggested by the original paper that when a protein model has a Z-score of −4 and lower, that indicates a very unstable, erroneous, and inexact structure. The value considered appropriate for normal structures is $-4 < x < 2$. For testing the structural integrity of the 3D structure of the generated proteins, a wide range of bioinformatics tools have been employed and later the output was analyzed and compared. The tools used were-Verify3D, ERRAT, PROCHECK, and Molprobity server.

3.4 Molecular Dynamic Simulations:

The steps included in the molecular simulation in this project are -

(i) choosing and obtaining a protein structure,
(ii) preparing it for simulation i.e. a topology file generation,
(iii) creating a box for simulation,
(iv) adding water as a solvent,
(v) performing energy minimization on the structure,
(vi) equilibrating the structure in temperature and pressure
(vii) performing the production simulation, and
(vii) analyzing the produced trajectory data.

To reproduce it, a Linux machine with a molecular dynamics package installed is required.
The molecular viewer PyMOL and Unix graph program Grace were installed and GROMACS was used.

i) Obtaining the structure for simulation :
The models with the best scores from the previous step of evaluation are selected to perform molecular simulations one by one for further analysis and for the selection of the final structure to inspect.

ii) A topology file generation :
Topology is the file that keeps an account of all atoms and interactions and records the change of velocity in each step in addition to the coordinates. This file is obtained by the program pdb2gmx from the PDB structure in the scene. Missing hydrogen atoms can also be added if by any chance the 3D structure is faulty. For generating a topology file parameters were called within a field of force which is mentioned at the very start of the file, indicating that each one of the subsequent parameters has been derived from this aforementioned field.

Developed by Prof. William L. Jorgensen, OPLS which stands for Optimized Potential for Liquid Simulations is the set of force fields that have been used for the single model simulation purpose of PSY1R kinase and also in the case of SERK1 kinase. OPLS is mostly considered for condensed phase simulations. The latest version is called OPLS-AA/M.

Another general-purpose molecular dynamics package for the study of biomolecular systems is GROMOS (GROningen MOlecular Simulation) which also incorporates its own force field. Its force field supports proteins, nucleotides, sugars, etc., and can be used for both chemical and physical systems including solutions of biomolecules. Supported by GROMACS, the GROMOS force field includes 43a1, 43a2, 45a3, 53a5, 53a6, and 54a7, with all the required parameters provided in the distribution. These above-mentioned force fields are united atom force fields, that is. without explicit aliphatic (non-polar) hydrogens. In this project, for all the protein-protein complexes after docking GROMOS 54a7 force field has been utilized.

The topology file contains the name [“PROTEIN_A”] as the name of the molecule that supports the very fact that the protein has been identified as chain A in its PDB file. The next section is defined by the information regarding atoms within the protein which covers the atom number (nr), type of atom (type), a number of the amino acid residue (resnr), name of the amino acid residue (residue), atom name, charge group number(cgnr), charge - “qtot”, a current total charge on the molecule, etc. are included in The subsequent sections include bonds, pairs, angles, and dihedrals which are quite straightforward in its meaning.

The rest of the topology file is involved in defining a couple of other topologies that are also necessary. This includes the position restraints “ posre.itp” that defines a force constant for

keeping the atoms in position during equilibration. Then, there is also another section in the topology file that defines other molecules and provides a system-level description. The solvent that has been used during this case is SPC/E (extended simple point charge model) water, chosen by passing "-water spce" to pdb2gmx. SPC, TIP3P, and TIP4P are the other choices available for water. In order to successfully formulate a topology file, the right order of the listed molecules must be maintained and it should exactly match the order of the molecules within the coordinate file, here which is the .gro file.

Failing to satisfy these concrete requirements, the grompp may show fatal errors including mismatched names, molecules not being found, or a number of others. Once examined the contents of a topology file, the procedure is sustained.

iii) Defining the box and solvate:

A simulation in water requires a larger box than the default box taken from the PDB crystal cell but not too large. The simulation was done in an easy aqueous system. Proteins and other molecules can also be simulated in various other solvents, as long as the right parameters are available for all involved molecule types. The simulation box was defined by using the editconf module and by using the solvate module it was filled with the solvent, i.e water. A simple cubic box is used as the unit cell.

- gmx editconf -f PROTEIN.gro -o PROTEIN_BOX1.gro -c -d 1.0 -bt cubic

Hereby using this command the protein molecule has been centered within the box (-c) that is defined as a cube (-bt cubic). It has been placed at a minimum distance of 1.0 nm from the box edge (-d 1.0). The best possible option would be positioning the solute within the center of a cube which can be availed by the space (-d) flag that automatically centers the protein within the box. The new conformation is recorded and written to the file BOX1.gro.

The space to the edge of the box is a crucial parameter as periodic boundary conditions have been used; it is imperative to satisfy the minimum image convention. One thing needs to be remembered that a protein must not see its periodic image or the forces would tend to be specious. By integrating a distance of 1.0 nm in the solute box, it interprets that between any two periodic images of the protein there is a distance minimum of 2.0 nm. For any cutoff scheme commonly utilized in simulations, this distance is going to be sufficient. For the simulation of the protein-protein complexes, 2.0 nm was used. The next step is to feature water within the box to solvate the protein and in GROMACS the particular parameters for this processing, this step is stored within the topology and field. Solvation is done using the solvate module:

- gmx solvate -cp PROTEIN_BOX1.gro -cs spc216.gro -o PROTEIN_solvated.gro -p topol.top

The output of the previous editconf step contains the protein configuration (-cp) and the standard GROMACS installation includes the solvent configuration (-cs). Here spc216 has been used as the solvent which is a generic equilibrated 3-point solvent model. The output has been named as PROTEIN_solvated.gro and the solvate has been told about the topology file (topol.top) so that it can be updated by the module. Now how many water molecules have been added would be kept tracked and would be written to the updated topology to reflect the changes made. The system now being solvated contains a charged protein that would be taken care of in the next step by adding ions.

iv Adding ions :

After learning the net charge the protein (based on its amino acid composition) ions are added to our system as life cannot prevail at a net charge. For adding ions within GROMACS, the tool genion is used which reads throughout the topology file and the water molecules are replaced with the ions that have been specified in the command. The input file is with an extension of .tpr; produced by the GROMACS grompp module. Both the coordinate and topology files, describing the molecules have been processed by the grompp module in order to generate an atomic-level input (.tpr) that possesses all the parameters of all the atoms in the system. For producing a .tpr file with grompp a .mdp file is required. Mdp file is the molecular dynamics parameter file containing specified parameters. Grompp would assemble those parameters and with the coordinates and topology information, a .tpr file would be produced. The .tpr file is assembled with the following command:

- gmx grompp -f p_ions.mdp -c PROTEIN_solvated.gro -p topol.top -o p_ions.tpr

Hence , an atomic-level description of our system in the binary file p_ions.tpr file is produced. This file is passed to genion

- gmx genion -s p_ions.tpr -o PROTEIN__neutral.gro -p topol.top -pname NA -nname CL -neutral

After being prompted, "SOL" is chosen for including ions. In the genion command, the structure/state file (-s) is provided as an input file in turn producing a .gro file as output (-o). The topology file has been updated (-p) that would reflect the elimination of the water molecules and inclusion of the ions. Here, genion has been asked to incorporate only the ions necessary, positive or negative, for neutralizing the net charge on the protein. This can be also done by specifying the number and type of the charges in the command. -pname and -nname denotes the positive and negative ion names respectively.

v) Energy minimization:

After the solvated, neutral system has been assembled in the previous stage, in order to ensure that there are no steric clashes and hindrances, the protein structure has been relaxed with the help of the energy minimization (EM) process. Quite large forces would be produced and

structure distortion would occur if the molecular dynamics is started immediately without going through the energy minimization step because of the added hydrogen and broken hydrogen bond in water if there are any. To make sure this doesn't happen it is important and expected to first conduct a brief energy minimization process. The .mdp file where the nonbonded interactions, required parameters and other settings are inscribed is the MINIMp.mdp file. Like the ions adding stage the structure, topology, and simulation parameters have been assembled by the grompp into a binary input file (.tpr) The energy minimization was run through the GROMACS MD engine:

- gmx grompp -f MINIMp.mdp -c PROTEIN_neutral.gro -p topol.top -o EMp.tpr

The topol.top file has again been updated while running genbox and genion. For invoking mdrun to carry out the EM:

- gmx mdrun -v -deffnm em

The -v flag makes the process verbose by displaying every step of progress on the screen. The file names of the input and the output are being defined by the -deffnm flag.

There are two important features from the result given at the end of the energy minimization(EM) step that decide where the EM step conducted has been successful or not. The first one is the potential energy given at the end of the EM process, E_{pot} being negative. For a small protein, like PSY1R kinase or SERK1 kinase in water, it would be on the order of 10^5-10^6. This order actually depends on the system size and number of water molecules.

F_{max} is the second most significant factor that denotes the maximum force. Its limit has been set in minimization.mdp - "emtol = 1000.0" - implying a target for F_{max} , not crossing 1000 kJ mol^{-1} nm^{-1}. An occurrence of a reasonable E_{pot} with F_{max} > emtol is certainly possible. In such cases, the system is predicted to not be stable or dependable enough for conducting the simulation process. In this situation, it would be necessary to check and evaluate the minimization parameters to find out the reasons and edit them as required. The em.edr file produced from the energy minimization run consisting of all the energy notes can be analyzed with the help of the GROMACS energy module:

- gmx energy -f em.edr -o PROTEIN_POTENTIAL.xvg

"Potential" has been selected for the input at the prompt. The E_{pot} average is shown and a file called "PROTEIN_POTENTIAL.xvg" is produced. Xmgrace is used as the plotting tool and the generated plot should show the nice and steady convergence of E_{pot}.

vi) Equilibration:

After energy minimization has ensured a reasonable structure regarding geometry and solvent orientation, real dynamics has begun. The solvent and ions around the protein must be equilibrated at the commencing of this step. If dynamics is attempted at this point, the system

may very likely fall apart. This is because the solvent has been mostly optimized within itself but not necessarily with the solute. Hence, it must be dropped at the proper temperature for simulating and forming the correct orientation about the solute which is the protein. After attaining the right temperature that is dependent on kinetic energies, pressure will have to be applied to the system till it achieves the right density.

The posre.itp file that has been generated a long time ago is meant to be used at this step. The posre.itp file applies a position restraining force on the heavy atoms excluding hydrogen atoms. Mobility is granted only after dealing with a considerable amount of energy penalty. The equilibration of the solvent around the protein without the added variable of structural changes within the protein can be done using the position restraints.

Equilibration has been conducted in two consecutive steps. The initial step is performed via an NVT ensemble that is constant Number of particles, Volume, and Temperature. The .mdp file has been formulated that contains all the required and specified parameters. While conducting NVT, the timeframe is dependent on the constituents of the system, but the system temperature should attain a plateau at the value specified. In case the temperature has not yet been optimized or stabilized, an extra 40-110 ps should be enough. Here 1ns NVT equilibration is conducted for all the protein and protein complexes.
The run was conducted by calling the grompp and mdrun as it was done in the energy minimization step :

- gmx grompp -f NVTp_file.mdp -c em.gro -r em.gro -p topol.top -o NVTp_file.tpr

- gmx mdrun -deffnm nvt

For analyzing the temperature progression graph is plotted :

- gmx energy -f nvt.edr -o temp.xvg

“Temperature” is selected from the prompt and the temp.xvg graph produced is later observed for analysis. After the temperature has been stabilized by NVT equilibration, the equilibration of pressure is conducted under an NPT ensemble also known as the "isothermal-isobaric" ensemble, wherein the Number of particles, Pressure, and Temperature are all constant and closely mimic the experimental conditions. The .mdp file used for 1ns NPT equilibration was formulated. Grompp and mdrun have been called, same as the NVT equilibration and the -t flag was included in the checkpoint file from the NVT equilibration that has all the important state variables for recontinuing the simulation. This file must be included for conserving the velocities produced during NVT. The coordinate file (-c) has been the concluding output from the NVT simulation that has been later used for NPT simulation. In this case, it is the nvt.gro file.

- gmx grompp -f NPTp_file.mdp -c nvt.gro -r nvt.gro -t nvt.cpt -p topol.top -o NPTp_file.tpr

- gmx mdrun -deffnm npt

For analyzing the temperature progression graph is plotted :

- gmx energy -f npt.edr -o pres.xvg

At the prompt the pressure is picked off the system and the graph is observed.

vii) Performing the production simulation:

After successful execution of the two steps of equilibration, since the system has been well-equilibrated at the specified temperature and pressure, the position restraints can be released now. The production MD has been run for the collection of data and the method is just like we have performed earlier. The checkpoint file possessing the preserved pressure coupling data has been forwarded to the grompp. All the parameters needed are contained in the .mdp file which is the MDp_file.mdp in this step. A 5ns MD simulation was run for the protein and 10 ns was run for the protein complexes. The mdp file was edited following that -

- gmx grompp -f MDp_file.mdp -c npt.gro -t npt.cpt -p topol.top -o MDp_file.tpr

Then mdrun was executed:

- gmx mdrun -deffnm md_file

After a certain amount of time, in this case, a few days each MD run was completed.

viii Analyzing the trajectory data:

After the protein and protein-complexes have been simulated, some analyses have been run on the system which has been presented in the "Results and Discussion" section of this paper. For a small protein-like PSY1R kinase and SERK1 kinase, a simulation of 5ns(50,0000 steps) was quite enough for the water to equilibrate surrounding the protein, however, during a large membrane system with the slow lipid motions several nanoseconds of relaxation would be required. The sole strategy to get this surely is by observing the potential energy and following that, extending the equilibration till it has been converged.

There is minimal difference between equilibration and production run. The position restraints and pressure coupling have been turned off and how often the output coordinates would be written to be analyzed has already been stated in the mdp file. In this case for every 5,000 steps coordinates have been recorded. How long the simulation would continue would depend on what

is being studied, and that should have been decided before the commencement of any simulation process.

In this project, a 5ns simulation was performed for all the selected protein structures and a 10ns simulation was performed for all the docking complexes. Next after the simulation output trajectory has been analyzed.

3.5 Protein-protein docking :

In this project both of our molecules - receptor and its partner were proteins, hence protein-protein docking was conducted. After molecular dynamics of the selected protein structures, the best protein models of PSY1R kinase and SERK1 kinase are subjected to energy minimization to eliminate steric clashes. Elimination of serious steric clashes or hindrances in some structures can be quite taxing as a number of promising refinement servers do not consider 3D structures with severe steric clashes. But fortunately, this problem can be tackled using server Chiron has been used for the energy minimization process of the structures that are to be docked. By accepting protein structures of different sizes, this server makes an effort to sort out and fix clashes with minimal backbone perturbation in comparison to the input structure (Ramachandran, S., Kota, 2011). Later this step, after looking through various docking tools, the online docking tool ClusPro was used.

The protein complexes produced were then again subjected to molecular dynamics simulation. The data that has been collected afterward have been analyzed and studied thoroughly in order to find some discerning insights.

Chapter 4: Result and Discussion

4.1 Sequence characterization of PSY1R receptor

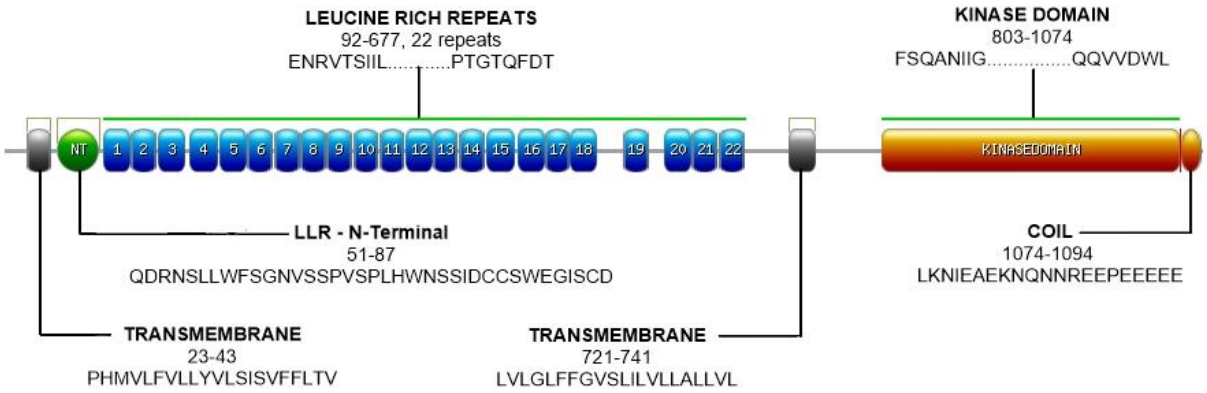

Figure 4.1: Sequence analysis of the PSY1R receptor.

Retrieving the sequence from Uniprot and by using PROSITE this diagram is created which identifies different portions of the whole PSY1R sequence. Here in this project, the Kinase domain, marked in orange from sequence 803 -1074 has been employed to continue the experimental process.

PSY1R kinase domain sequence retrieved from UniProt:

>sp|Q9C7S5|803-1074
FSQANIIGCGGFGLVYKATLDNGTKLAVKKLTGDYGMMEKEFKAEVEVLSRAKHENLV
ALQGYCVHDSARILIYSFMENGSLDYWLHENPEGPAQLDWPKRLNIMRGASSGLAYMH
QICEPHIVHRDIKSSNILLDGNFKAYVADFGLSRLILPYRTHVTTELVGTLGYIPPEYGQAW
VATLRGDVYSFGVVMLELLTGKRPMEVFRPKMSRELVAWVHTMKRDGKPEEVFDTLLR
ESGNEEAMLRVLDIACMCVNQNPMKRPNIQQVVDWL

4.2 Analysis of physicochemical properties:

Table 4.1 : Results of ProtPram tool:

Number of amino acids: 272		
Molecular weight: 30784.63		
Theoretical pI: 6.49		
Number of negatively charged residues (Asp + Glu): 32		
Number of positively charged residues (Arg + Lys): 30		
Instability index (II) : 39.58 This classifies the protein as stable.		
Aliphatic index: 91.36		
Grand average of hydropathicity (GRAVY): -0.156		
Amino acid composition:		
Ala (A)	17	6.2%
Arg (R)	15	5.5%
Asn (N)	12	4.4%
Asp (D)	13	4.8%
Cys (C)	5	1.8%
Gln (Q)	8	2.9%
Glu (E)	19	7.0%
Gly (G)	22	8.1%
His (H)	8	2.9%
Ile (I)	13	4.8%
Leu (L)	30	11.0%
Lys (K)	15	5.5%
Met (M)	12	4.4%
Phe (F)	9	3.3%
Pro (P)	12	4.4%
Ser (S)	13	4.8%
Thr (T)	11	4.0%
Trp (W)	5	1.8%
Tyr (Y)	11	4.0%
Val (V)	22	8.1%

Atomic composition:	
Carbon C	1379
Hydrogen H	2163
Nitrogen N	373
Oxygen O	392
Sulfur S	17
Formula:	$C_{1379}H_{2163}N_{373}O_{392}S_{17}$
Total number of atoms:	4324

Table 4.2 : Results of Pepstats

Charge = 2.0			
Property	Residues	Number	Mole%
Tiny	(A+C+G+S+T)	68	25.000
Small	(A+B+C+D+G+N+P +S+T+V)	127	46.691
Aliphatic	(A+I+L+V)	82	30.147
Aromatic	(F+H+W+Y)	33	12.132
Non-polar	(A+C+F+G+I+L+M +P+V+W+Y)	158	58.088
Polar	(D+E+H+K+N+Q+R +S+T+Z)	114	41.912
Charged	(B+D+E+H+K+R+Z)	70	25.735
Basic	(H+K+R)	38	13.971
Acidic	(B+D+E+Z)	32	11.765

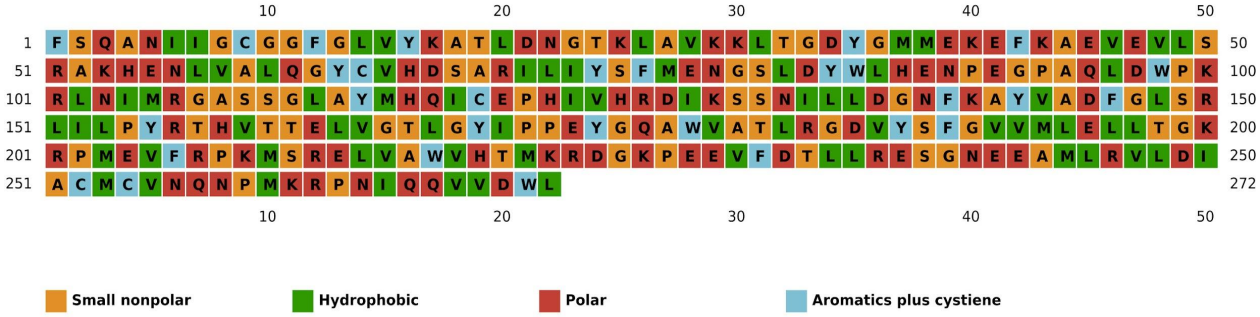

Figure 4.2: Residues type at a glance by PSIPRED

Analysis of various physical and chemical properties of PSY1R kinase was performed using ProtParam, Pepstats and also PSIPRED. It was revealed that the PSY1R kinase consists of 272 AA and its molecular weight is 30784.63 kDa.

The amino acid composition of the protein showed that the most prevailing amino acid is leucine which accounts for 11% of the primary structure whereas cysteine and tryptophan have the lowest percentage of amino acid and makes up 1.8% of its primary structure. With a branched hydrocarbon side chain, leucine is classified as nonpolar aliphatic amino acid and one of the essential ones of the 8. Since it's hydrophobic in nature it tends to be in the protein interior. Leucine being an aliphatic, hydrophobic favor substituting the amino acids having a similar nature and inclined to be inside the hydrophobic cores of the protein. Being hydrophobic, it has an inclination to be within the alpha helices rather than in beta-strands. Leucine side-chain being highly non-reactive, is rarely a part of the functional properties of the protein, even though it plays a part in recognition of substrates, especially binding and recognition of ligands that are hydrophobic as lipids. Besides, since there is evidence of a higher concentration of leucine around the perimeter of residues of active sites, it shows that hydrophobic residues like leucine aid in the process of forming a non-aqueous environment helping the substrate binding and catalysis between the polar residues. (Chou, P. Y., & Fasman, G. D. (1973)).

Meanwhile, Cysteine residues are adept at forming disulfide bonds playing an important role in structural stability and folding. The low concentration of cysteine residues indicates that the stability of the protein sources from different other interactions, such as hydrophobic interactions, ionic bonds and hydrogen bonds since the probability of forming disulfide bonds is relatively lower.

Another residue Tryptophan is also found to be in the least percentage 1.8 besides Cysteine. Even being one of the least abundant, it is one of the most expensive to be produced. (Hrazdira and Jensen, 1992). The level of tryptophan in plants is usually low compared to the other amino acids. While animals use tryptophan as an essential amino acid required for many pathways and protein synthesis plants use this as precursors for the synthesis of auxin, phytoalexins, glucosinolates, and both indole and anthranilate derived alkaloids. Because of these metabolites, it can be claimed that the biosynthesis of tryptophan has a direct effect on plant development, regulating plant defense responses and pollination (Kutchan, 1995). Tryptophan is 1.8% which is a very standard amount for tryptophan shows that the kinase domain may contribute to the biosynthesis pathways to a considerable extent.

The pH at which the net charge of the protein is zero is considered as the isoelectric point(pI) of that protein. The isoelectric point (pI) has been found to be 6.49 which says that the combined number of acidic residues (negatively charged) is almost as equal to that of the basic residues (positively charged). The aliphatic index of a protein can be defined by the relative volume

covered by the aliphatic side chains that include alanine, valine, isoleucine, and leucine Aliphatic index by being directly proportional to how stable the structure of the protein is, also acts as a positive indicator for higher thermostability of the protein. The aliphatic index of PSY1R kinase being 91.36 indicates that the protein is stable in different ranges of temperatures. Whereas the instability index being quite lower, as low as 39.58 shows that it is quite stable. The GRAVY value that stands for Grand Average of Hydropathy, is computed by dividing the sum total of hydropathy values of all the amino acids by the total number of residues in the sequence. The hydropathy value of an amino acid is an index for the hydrophobic or hydrophilic characteristics of its sidechain (Kyte and Doolittle, 1982). A negative value for GRAVY implies protein having polar nature, whereas a positive value implies the protein to be of more non-polar nature. Therefore, a GRAVY value of -0.156 tells us that our protein under inspection is quite polar i.e., hydrophilic in nature.

Again from the PSIPRED physicochemical prediction, we can see an overall visual representation of the properties of the residues in the kinase domain. The small nonpolar residues and hydrophobic residues comprise most part of it whereas quite a large amount of polar residues and very few aromatic plus cysteine residues are also observed.

In plants, the aromatics phenylalanine, tryptophan and tyrosine function as precursors in the production of a wide variety of compounds playing critical roles in the growth and maturing of plants, procreation, resistance and environmental responses meanwhile cysteine gives stability to the tertiary structure. The disulfide bridges formed by these residues connect the fragments within a polypeptide at times located at a greater distance from each other thus providing firmness to the tertiary structure. The role of cysteine in the tertiary structure of proteins is obvious. But a very little amount of cysteine and tryptophan residues in this protein sequence may result in very low amounts of disulfide bridges in both intramolecular and intermolecular interactions, hence compromising the overall rigidness of the structure to some extent.

4.3 Domain architecture analysis by InterPro:

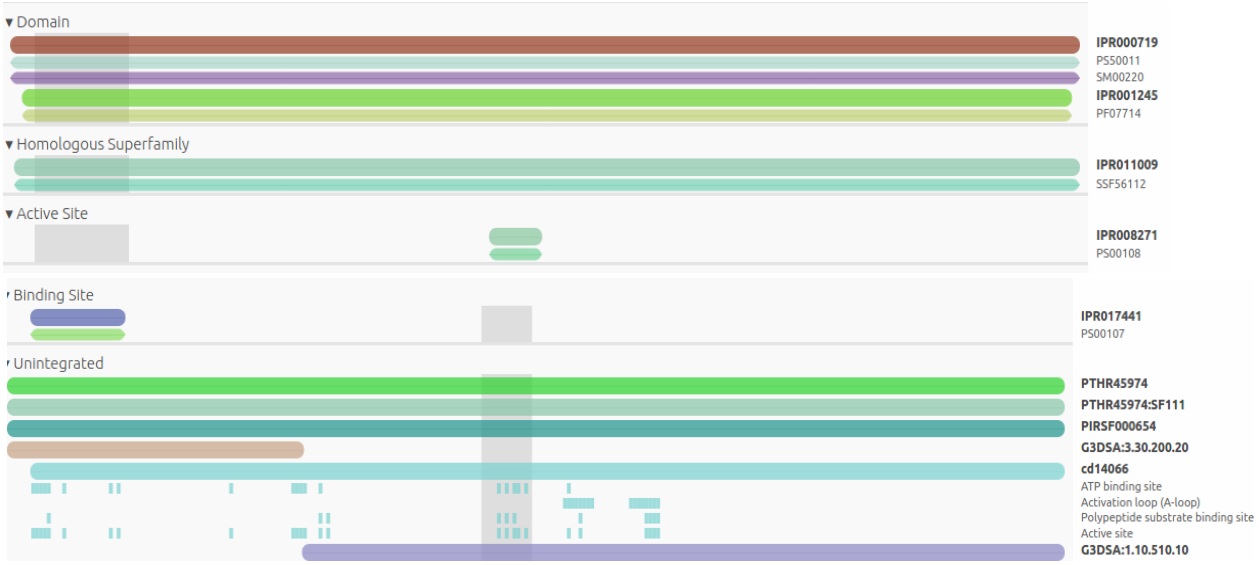

Figure 4.3: Domain architecture of the PSY1R kinase by InterPro.

Domain architecture analysis by InterPro showed that PSY1R kinase belongs to homologous Protein kinase-like domain superfamily (IPR011009) and domain relatedness is found with the Protein kinase domain (IPR000719). Protein kinases and phosphoprotein phosphatases mediate protein phosphorylation which is a reversible process and has a significant part in a lot of cellular activities. Gamma phosphate transfer from nucleotide triphosphates is catalyzed by the protein kinase resulting in a conformational change that affects the functional properties of the protein. Phosphoprotein phosphatases catalyze the reverse process. Meanwhile, the reverse process is catalyzed by phosphoprotein phosphatases.

Protein kinases can be classified into three broad classes that characterize with respect to substrate specificity -

- Serine and threonine protein kinases
- Tyrosine-protein kinases
- Protein kinases with dual specificity, for example, both Thr and Tyr can be phosphorylated by Mitogen-activated protein kinase kinase (MEK) on target proteins.

The active site of PSY1R kinase belongs to Serine/threonine-protein kinase(IPR008271) whereas the binding site belongs to Protein kinase, ATP binding site (IPR017441).

4.4 Secondary structure prediction by PSIPRED:

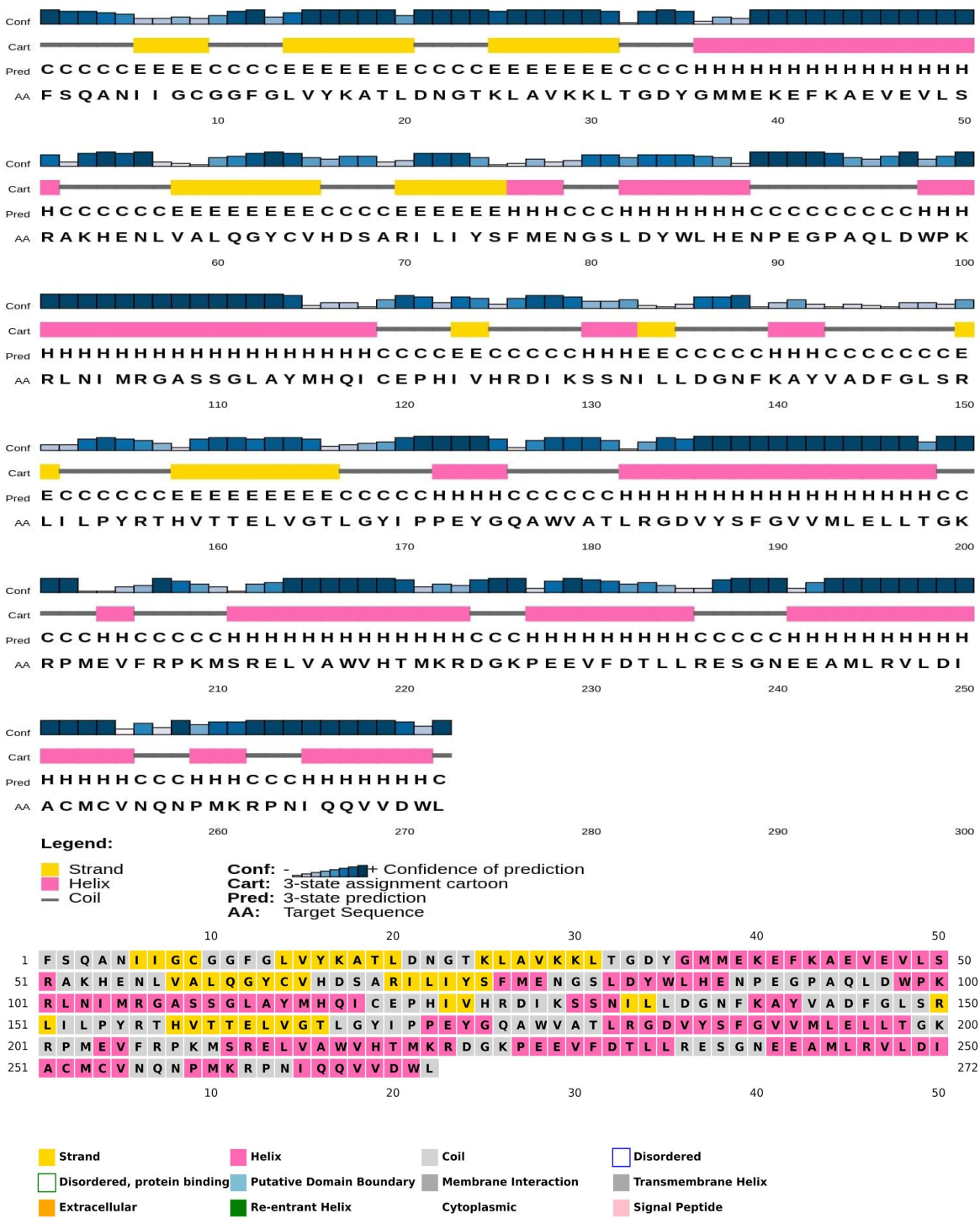

Figure 4.4: Secondary structure prediction by PSIPRED.

Secondary structure output that has been predicted using PSIPRED implies that the PSY1R kinase is composed of both beta sheets (strands, presented in yellow) and alpha-helix(presented with pink) with alpha-helix prevailing the most part of it. There are also few regions with coils. According to PSIPRED, there are 14 alpha-helix, 9 beta-strands and 22 coils. The confidence of prediction is very high around the helix region.

4.5 Prediction of conserved regions:

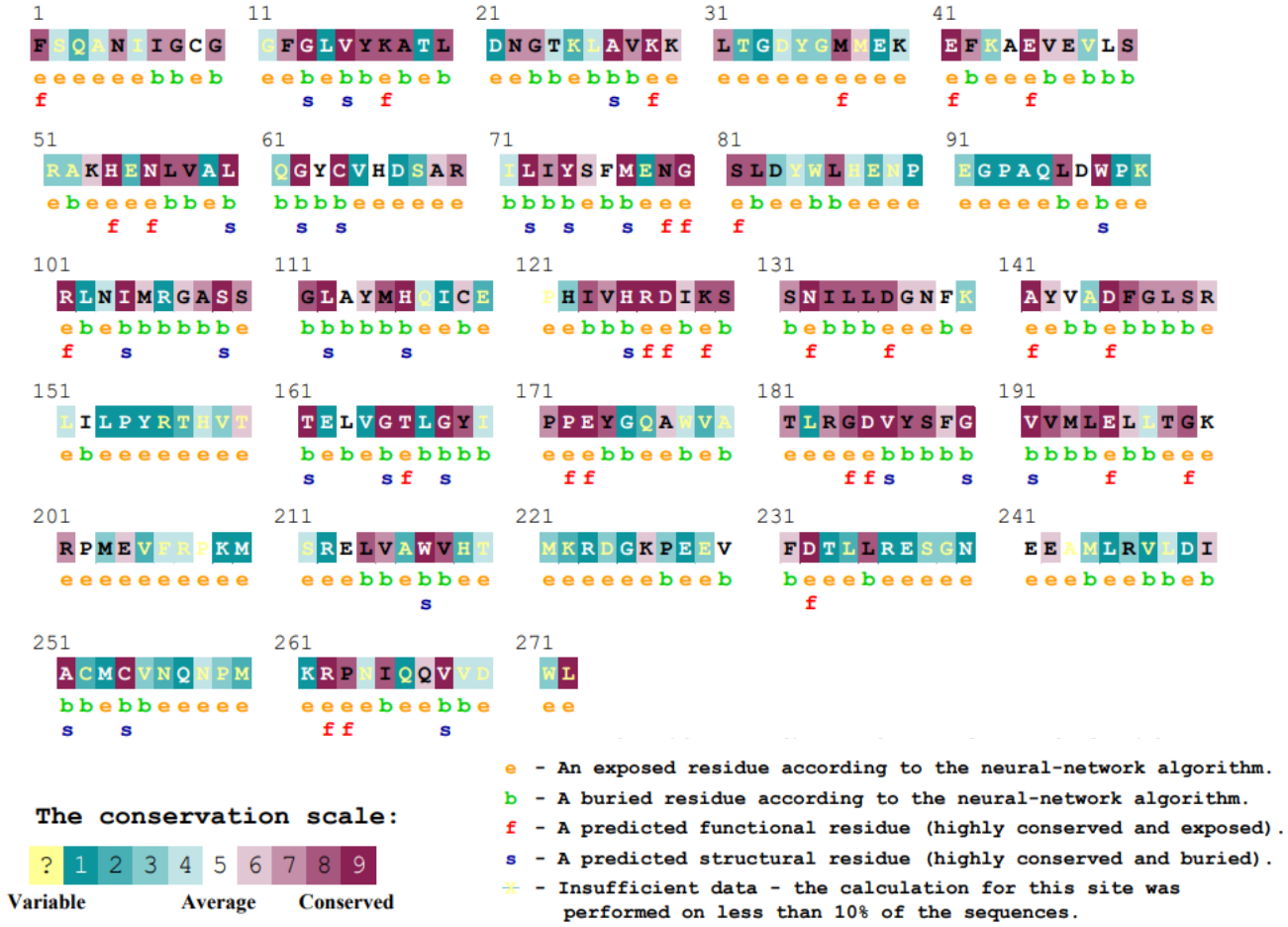

Figure 4.5: Conserved region predicted by ConSurf.

The ConSurf tool was utilized for the estimation of evolutionary conservation of residues in the PSY1R kinase domain. According to the ConSurf results, the majority of the residues that are placed high on the ConSurf conservation scale were exposed functional residues denoted by "f", some are also buried and structural residues denoted by "s". Most of the exposed residues of the protein are placed low on the conservation scale indicating they are variable residues. From the distribution shown in the chart, the exposed residues seemed to be more than the buried ones

along with 29 predicted functional residues and 25 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. In order to support the structural integrity of the proteins, hydrophobic cores are usually formed by the buried residues while the functional properties of the protein are associated with the exposed residues that are mainly hydrophilic.

4.6 Blast Results:

The target protein sequence was subjected to blast using P-Suite (protein-protein BLAST) of the BLAST software against the protein data bank (PDB) database. Upon completion of the 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, pink medium and black indicating the lowest score. The higher alignment scores the more significant hits. The summary table shows all the sequences in the database that showed significant matches to the query sequence.
The results were sorted 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 that is close to zero signifies that the hit has to be significant but 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 50 blast hits, the top 12 sequences were selected. The FASTA format of each of the top sequences was retrieved from the NCBI protein database.

Accession	Description
☑ 5LPV_A	Crystal structure of the BRI1 kinase domain (865-1160) in complex with AMPPNP and Mn from Arabidopsis thaliana [Arabidopsis thaliana]
☑ 5LPB_A	Crystal structure of the BRI1 kinase domain (865-1160) in complex with ADP from Arabidopsis thaliana [Arabidopsis thaliana]
☑ 5LPZ_A	Crystal structure of the BRI1 kinase domain (865-1196) in complex with ADP from Arabidopsis thaliana [Arabidopsis thaliana]
☑ 4OH4_A	Crystal structure of BRI1 in complex with BKI1 [Arabidopsis thaliana]
☑ 3TL8_A	The AvrPtoB-BAK1 complex reveals two structurally similar kinaseinteracting domains in a single type III effector [Arabidopsis thaliana]
☑ 3UIM_A	Structural basis for the impact of phosphorylation on plant receptor-like kinase BAK1 activation [Arabidopsis thaliana]
☑ 4Q5J_A	Crystal structure of SeMet derivative BRI1 in complex with BKI1 [Arabidopsis thaliana]
☑ 6M0U_A	Chain A, Symbiosis receptor kinase SymRK [Arachis hypogaea]
☑ 4L68_A	Structure of the psedudokinase domain of BIR2, an immune regulator of the RLK/Pelle family [Arabidopsis thaliana]
☑ 5XD6_A	CARK1 phosphorylates ABA receptors [Arabidopsis thaliana]
☑ 5TOS_A	Botrytis-induced kinase 1 (BIK1) from Arabidopsis thaliana [Arabidopsis thaliana]
☑ 6CTH_A	Crystal Structure of Pathogenesis-related Protein 1G (PR-1G) Kinase Domain from Cacao [Theobroma cacao]

Figure 4.6: Description and Accession number of the PDB bank of the top 12 sequences from BLAST result.

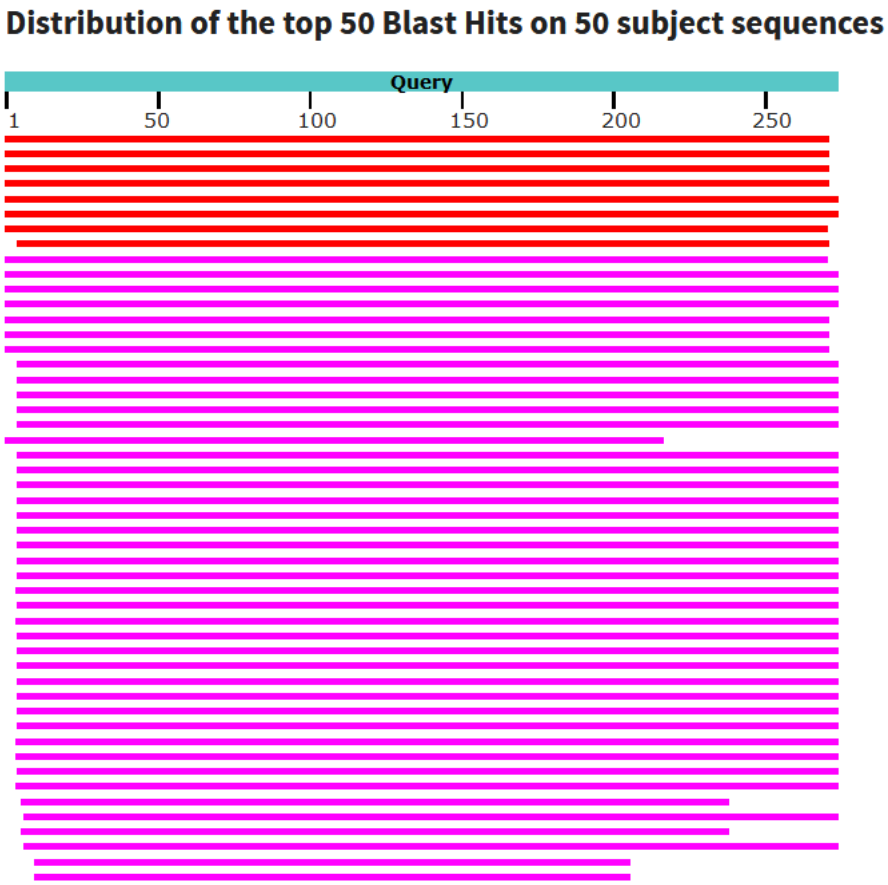

Figure 4.7: Graphical representation of the blast result of the top 50 sequences

Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Arabidopsis thaliana	224	224	98%	3e-72	43.96%	296	5LPV_A
Arabidopsis thaliana	224	224	98%	5e-72	43.96%	296	5LPB_A
Arabidopsis thaliana	224	224	98%	6e-72	43.96%	332	5LPZ_A
Arabidopsis thaliana	224	224	98%	1e-71	43.96%	333	4OH4_A
Arabidopsis thaliana	224	224	100%	2e-71	41.09%	349	3TL8_A
Arabidopsis thaliana	218	218	100%	3e-69	40.00%	326	3UIM_A
Arabidopsis thaliana	217	217	98%	6e-69	43.38%	341	4Q5J_A
Arachis hypogaea	210	210	97%	2e-66	41.95%	304	6M0U_A
Arabidopsis thaliana	184	184	98%	4e-56	38.46%	340	4L68_A
Arabidopsis thaliana	174	174	100%	1e-52	37.86%	305	5XD6_A
Arabidopsis thaliana	174	174	100%	1e-51	37.98%	395	5TOS_A
Theobroma cacao	169	169	100%	1e-50	36.86%	309	6CTH_A

Figure 4.8: Top 12 sequences from the blast result of the PSY1R kinase domain

4.7 Results of Multiple Sequence Alignment:

The multiple sequence alignment (MSA) method is a series of algorithmic solutions for the alignment of evolutionarily related sequences by taking into account all the evolutionary events such as mutations, insertions, deletions and rearrangements under certain conditions (Chatzou, M., Magis, C., 2016). Multiple sequence alignment of the selected protein sequences was performed using the MUSCLE server and the output was in clustalw file. In order to show the MSA results in a viewable format, the 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 PSY1R 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.

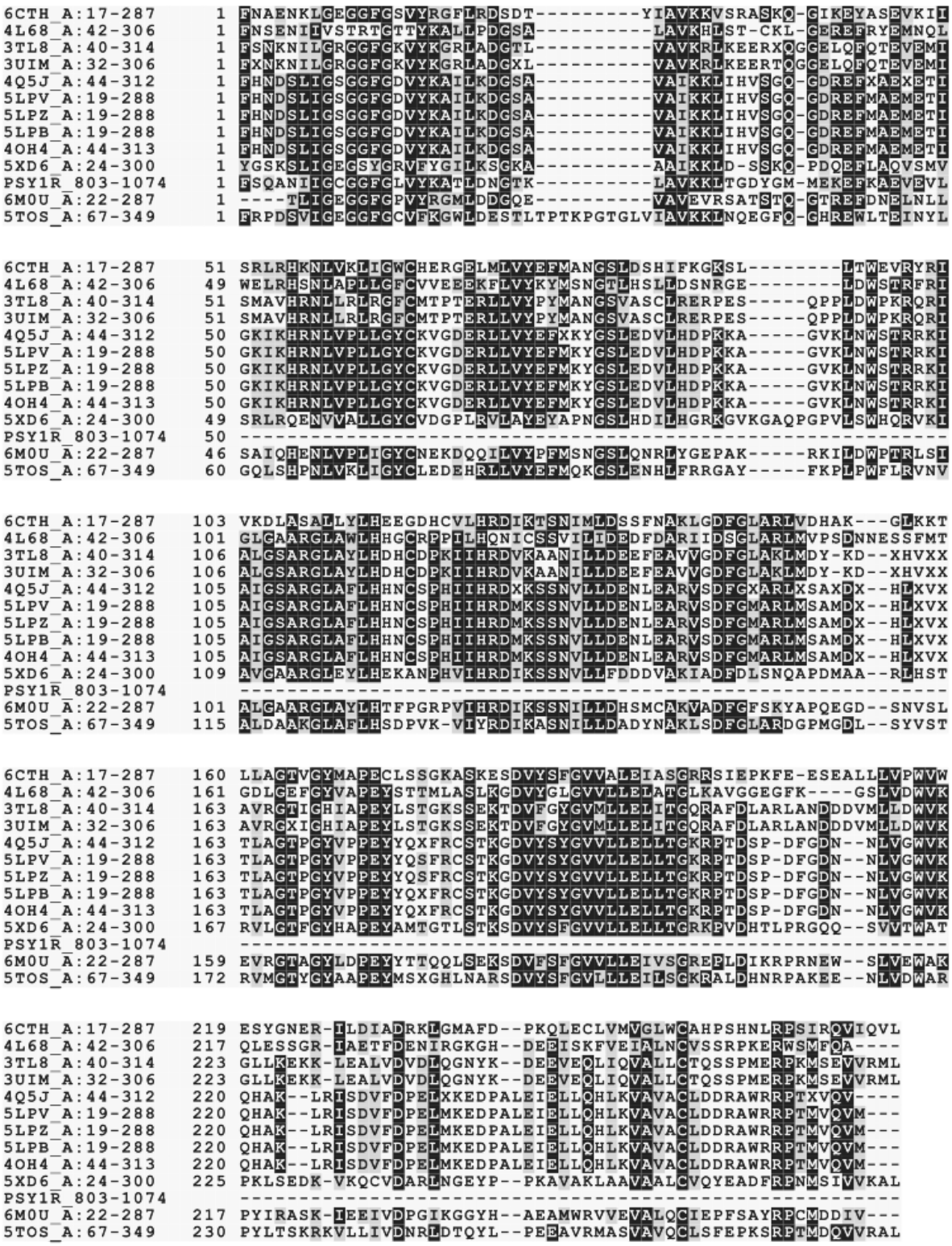

Figure 4.9: Multiple sequence alignment view using the BoxShade server in postscript portrait format.

4.8 Phylogenetic tree generation:

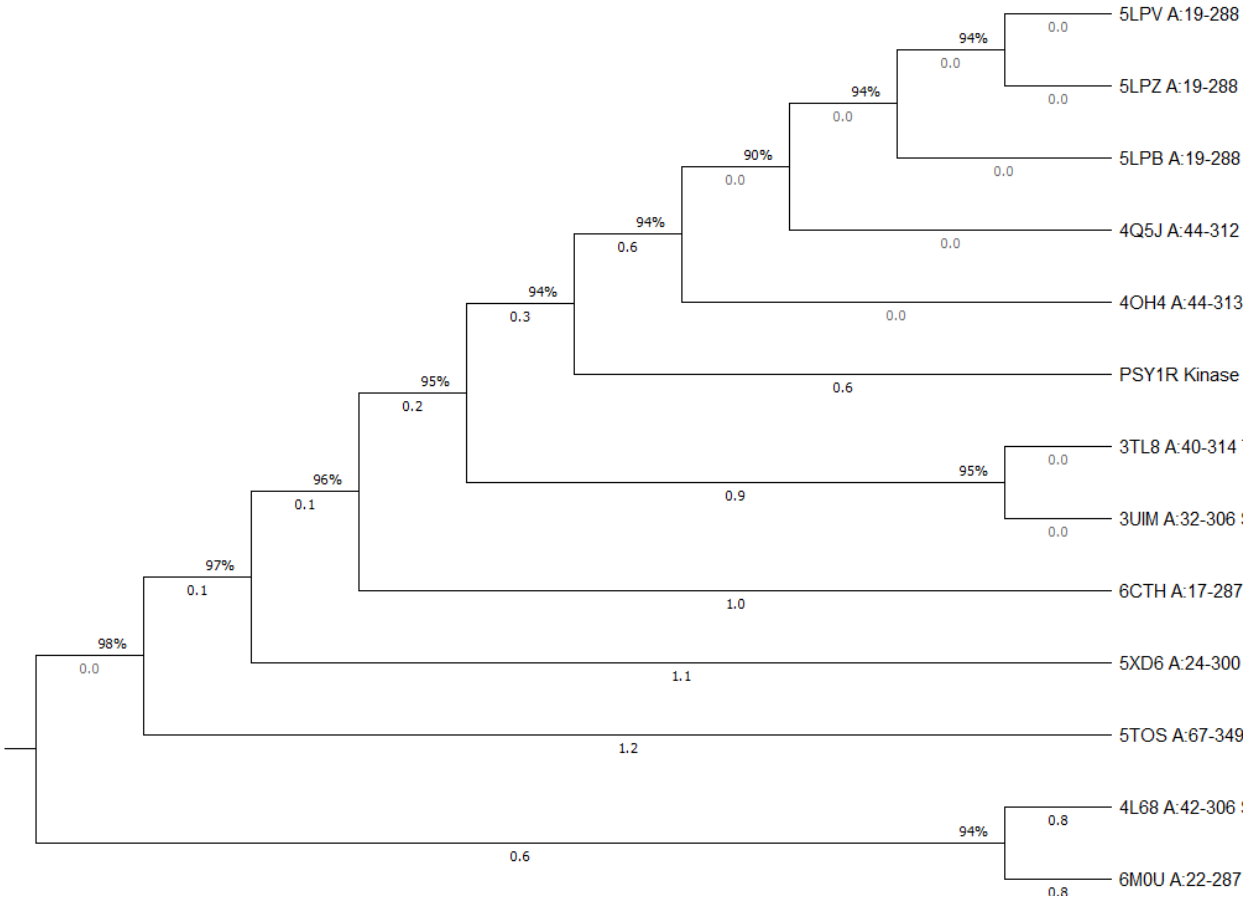

Figure 4.10: Phylogenetic tree, WAG model by MEGA-X

Accession	Description
☑ 5LPV_A	Crystal structure of the BRI1 kinase domain (865-1160) in complex with AMPPNP and Mn from Arabidopsis thaliana [Arabidopsis thaliana]
☑ 5LPB_A	Crystal structure of the BRI1 kinase domain (865-1160) in complex with ADP from Arabidopsis thaliana [Arabidopsis thaliana]
☑ 5LPZ_A	Crystal structure of the BRI1 kinase domain (865-1196) in complex with ADP from Arabidopsis thaliana [Arabidopsis thaliana]
☑ 4OH4_A	Crystal structure of BRI1 in complex with BKI1 [Arabidopsis thaliana]
☑ 3TL8_A	The AvrPtoB-BAK1 complex reveals two structurally similar kinaseinteracting domains in a single type III effector [Arabidopsis thaliana]
☑ 3UIM_A	Structural basis for the impact of phosphorylation on plant receptor-like kinase BAK1 activation [Arabidopsis thaliana]
☑ 4Q5J_A	Crystal structure of SeMet derivative BRI1 in complex with BKI1 [Arabidopsis thaliana]
☑ 6M0U_A	Chain A, Symbiosis receptor kinase SymRK [Arachis hypogaea]
☑ 4L68_A	Structure of the psedudokinase domain of BIR2, an immune regulator of the RLK/Pelle family [Arabidopsis thaliana]
☑ 5XD6_A	CARK1 phosphorylates ABA receptors [Arabidopsis thaliana]
☑ 5TOS_A	Botrytis-induced kinase 1 (BIK1) from Arabidopsis thaliana [Arabidopsis thaliana]
☑ 6CTH_A	Crystal Structure of Pathogenesis-related Protein 1G (PR-1G) Kinase Domain from Cacao [Theobroma cacao]

Figure 4.11: Description and Accession number of the PDB bank of the top 12 sequences from BLAST result, used for the construction of the phylogenetic tree.

A phylogenetic tree can be defined as a diagram in the shape of ascending branches that represent evolutionary relationships among organisms mainly based on hypotheses but not facts. Most of the modern systems of classification are done by evaluating the evolutionary relationships among organisms, i.e phylogeny. Classification systems on phylogeny sub-categorize species or other groups in such ways that they comprehend how they evolved from their common ancestors. A phylogenetic tree can be described as a display of information about the inferred evolutionary relationships between a set of sequences. In a phylogenetic tree, the species under inspection are shown at the end-tip of lines known as the tree branches. Each branch point which is also known as an internal node presents a division of lineage. Each node acts as the most recent common ancestor of all the groups descended from that branch point while each horizontal line illustrates a series of ancestors that leads up to the end species. Two descendants that split from the same node are called sister groups. Moving forward through the horizontal dimension from nodes to tips represents moving forward in time. In addition to this, the branch lengths imply the amount of evolutionary change that occurred with time. The greater the branch length, the greater the amount of genetic change. Lastly, the vertical dimension in a phylogenetic tree has no significance as branches can be swapped at any internal nodes.

From the results obtained PSY1R *Arabidopsis thaliana* has undergone the greatest amount of character change compared to the other groups as it has the second-largest branch length with 1.3 and 1.4 being the largest. The other sequences also went through an almost similar level of changes as the lowest branch distance is 1.3. Few groups are seen to have the same amount of changes, also splitting from the same node; it can be assumed that they are closely related to each other. 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.

From the tree it can be observed that the most closely related groups or sister groups to PSY1R([*Arabidopsis thaliana*]) are **5LPV** (Crystal structure of the BRI1 kinase domain (865-1160) in complex with AMPPNP and Mn [*Arabidopsis thaliana*]), **5LPZ** (Crystal structure of the BRI1 kinase domain (865-1196) in complex with ADP [*Arabidopsis thaliana*]), **5LPB** (Crystal structure of the BRI1 kinase domain (865-1160) in complex with ADP from Arabidopsis thaliana [*Arabidopsis thaliana*]), **4Q5J** (Crystal structure of SeMet derivative BRI1 in complex with BKI1 [*Arabidopsis thaliana*]) and **4OH4**(Crystal structure of BRI1 in complex with BKI1 [*Arabidopsis thaliana*]) as they all arise from the same node as PSY1R, have 0.0 distance among themselves and also from the same species *Arabidopsis thaliana*.

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 is the greater the statistical confidence of the node.

4.9 Results of Protein modelling:

For PSY1R kinase, single template modeling of query sequence was done using AIDA, FFAS03,(PS)2-v2, Raptor X, Spark X, CEthreader, and Swiss Model. For both FFAS03 and the 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, M4T, PRIMO, IntFOLD5, Phyre2, Robetta, and HHpred Modeller. Meanwhile, for SERK1 kinase modelling has been done using MODELLER in both basic and advanced mode. The basic modelling uses a single template whereas the Advanced modelling uses multiple templates. On account of the DOPE score, the produced models were judged.

i) PSY1R kinase domain modelling result :

Table 4.3: Total accounts of the tools and templates used for PSY1R kinase protein modelling:

<table>
<tr><th>Tool</th><th>Techniques</th><th>No. of models</th><th>Template number</th><th>Templates PDB ID</th></tr>
<tr><td>Robetta</td><td>Ab initio, Homology modelling</td><td>5 models</td><td>Automated</td><td>NA</td></tr>
<tr><td>Raptor X (New server)</td><td>Threading (distance-based)</td><td>5 models</td><td>Automated</td><td>NA</td></tr>
<tr><td>CEthreader</td><td>Threading</td><td>5 models</td><td>Single template based</td><td><table><tr><td>Model 1</td><td>6cthA</td></tr><tr><td>Model 2</td><td>3tl8A</td></tr><tr><td>Model 3</td><td>2nruB</td></tr><tr><td>Model 4</td><td>6bfnA</td></tr><tr><td>Model 5</td><td>2qkwB</td></tr></table></td></tr>
</table>

<table>
<tr><td>Sparks X</td><td>Threading</td><td>10 models</td><td>Single template based</td><td><table><tr><td>Model 1</td><td>3tl8_A</td></tr><tr><td>Model 2</td><td>4q5j_A</td></tr><tr><td>Model 3</td><td>3hgk_A</td></tr><tr><td>Model 4</td><td>5uiq_A</td></tr><tr><td>Model 5</td><td>6j5t_H</td></tr><tr><td>Model 6</td><td>4m66_A</td></tr><tr><td>Model 7</td><td>4f0f_A</td></tr><tr><td>Model 8</td><td>4mwi_A</td></tr><tr><td>Model 9</td><td>6j5t_I</td></tr><tr><td>Model 10</td><td>4c8b_A</td></tr></table></td></tr>
<tr><td>(PS2)-V2</td><td>Threading</td><td>1 model</td><td>Single template based</td><td>2qkw_B</td></tr>
<tr><td>FFASO3</td><td>Homology modelling</td><td>2 models</td><td>Single template based</td><td><table><tr><td>Model 1</td><td>5lpz_A</td></tr><tr><td>Model 2</td><td>5lpz_A</td></tr></table></td></tr>
<tr><td>SWISS model</td><td>Homology modelling</td><td>2 models</td><td>Single template based</td><td><table><tr><td>Model 1</td><td>4p5j_A</td></tr><tr><td>Model 2</td><td>5lpz_A</td></tr></table></td></tr>
<tr><td>LOMETS</td><td>Threading</td><td>5 models</td><td>Multiple templates based</td><td>6cth_A, 3tl8_A, 2qkw_B</td></tr>
<tr><td>M4T</td><td>Homology modelling</td><td>1 model</td><td>Multiple templates based</td><td>6ege_A, 6cth_A, 5lpb_A, 5lpz_A, 3tl8_D</td></tr>
<tr><td>I-TASSER</td><td>Threading</td><td>1 model</td><td>Multiple templates based</td><td>6cth_A, 3tl8_A, 6j5t</td></tr>
<tr><td>AIDA</td><td>Ab initio method</td><td>1 model</td><td>Multiple templates based</td><td>5lpb, 5lpv</td></tr>
</table>

PRIMO	Homology modelling	4 models	Multiple templates based	5lpw, 3tl8, 5lpv, 3uim, 5lpz, 4q5j, 4oh4
Phyre 2	Homology modelling	1 model	Multiple templates based	4xi2_A, 5ebz_F, 2fo0_A, 1opl_A, 2h8h_A, 1y57_A
HHPRED	Homology modelling	3 models	Multiple templates based	Model 1: 6J5T, 4M68, 3S95, 3KFA, 5UT3, 2V62 Model 2: 2QKW, 6CTH, 3UIM, 5XD6, 5LPB, 5TOS, 6VPM, 2H34, 4EQM, 4B6L Model 3: 2QKW, 6CTH, 3UIM, 5XD6, 5LPB, 5TOS, 4B6L,4OH4, 2W5A, 5UT3, 6J5T

i) SERK1 kinase domain modelling results :

For SERK1 kinase modelling, its sequence has been first retrieved from the Uniprot and then by Blastp, top three templates have been selected for modelling.

The sequence of SERK1 kinase retrieved from UniProt:

>sp|Q94AG2|302-589

FSNKNILGRGGFGKVYKGRLADGTLVAVKRLKEERTPGGELQFQTEVEMISMAVHRNLL
RLRGFCMTPTERLLVYPYMANGSVASCLRERPPSQPPLDWPTRKRIALGSARGLSYLHD
HCDPKIIHRDVKAANILLDEEFEAVVGDFGLAKLMDYKDTHVTTAVRGTIGHIAPEYLST
GKSSEKTDVFGYGIMLLELITGQRAFDLARLANDDDVMLLDWVKGLLKEKKLEMLVDP
DLQTNYEERELEQVIQVALLCTQGSPMERPKMSEVVRMLEGDGLAEKWDEWQ

Table 4.4: Selected top three templates for SERK1 kinase, modelling of SERK1 kinase using those templates and DOPE score observation :

Single template modelling	Template 1: The AvrPtoB-BAK1 complex reveals two structurally similar kinase interacting domains in a single type III effector PDB ID: 3TL8	
	Model 1	**DOPE score: -32128.91992**
	Model 2	DOPE score: -31845.19141

	Model 3	DOPE score: -31938.53516
	Model 4	**DOPE score: -31983.61914**
	Model 5	**DOPE score: -32241.73438**
Single template modelling	Template 2 : Structural basis for the impact of phosphorylation on plant receptor-like kinase BAK1 activation PDB ID : 3UIM	
	Model 6	DOPE score: -29739.67969
	Model 7	DOPE score: -29837.80664
	Model 8	DOPE score: -29772.79102
	Model 9	DOPE score: -29947.79102
	Model 10	DOPE score: -29944.02930
Multiple template modelling	Template 1 : The AvrPtoB-BAK1 complex reveals two structurally similar kinase interacting domains in a single type III effector PDB ID : 3TL Template 2 : Structural basis for the impact of phosphorylation on plant receptor-like kinase BAK1 activation PDB ID : 3UIM Template 3: Crystal structure of the BRI1 kinase domain (865-1160) in complex with AMPPNP and Mn from Arabidopsis thaliana PDB ID :5LPV	
	Model 11	DOPE score: -31419.162109
	Model 12	DOPE score: -31971.312500
	Model 13	DOPE score: -31743.593750
	Model 14	**DOPE score: -32140.437500**
	Model 15	DOPE score: -31676.367188

The DOPE or Discrete Optimized Protein Energy model score is assigned for selecting the best structure from all of the models built by MODELLER. Since the lower the DOPE score is, the better the model quality, four models with the lowest DOPE score were selected for further evaluation in the next stage.

4.10 Results of protein model evaluation :

For testing the structural quality and integrity of the 3D structure of the modelled proteins, the structures are judged by using Verify3D, ERRAT and Ramachandran distribution plot generated by PROCHECK and Molprobity server. The statistics of non-bonded interactions between different atom types (using ERRAT), the compatibility of an atomic 3D protein model with its own primary amino acid sequence (using Verify 3D), the rotation angle of the residues and whether they are positioned in an allowed or disallowed region (observing Ramachandran distribution plot) were taken into accounts by inspecting the quantitative values. Ramachandran distribution plot displayed how stable each model of the protein is on the basis of rotations of the polypeptide that are allowed and disallowed as a result of steric hindrance. Ramachandran Z score, a statistical metric, was also taken into account that shows how 'normal' or okay a model is compared to a reference set of high-resolution structures (Hooft *et al*., 1997).

Table 4.5: Table containing values for evaluation on different parameters of PSY1R kinase 3D models :

	Models	ERRAT Goal: >50 %	VERIFY 3D Goal: >80%	Ramachandran score (Procheck)	Ramac han-dra n score (Molpr obity) Goal: >98%	Ramach -andran Outliers (Molpro bity) Goal: <0.05%	RMS Z-SCORE (Molprobit y) Goal: abs(Z score) < 2
1.	1.Raptor X Model -1	44.8	67.65% F	82.6% core 14.8% allowed 1.3% general 1.3% disallowed	85.93%	5.19%	-3.78 ± 0.40
2.	Model - 2	53.012	78.68%	83.5% core 15.3% allowed 1.3% general 0.0% disallowed	87.41%	4.81%	-4.07 ± 0.38
3.	Model - 3	47.4104	77.94%	81.4% core 16.1% allowed 2.1% general 0.4% disallowed	82.96%	7.41%	-3.58 ± 0.40
4.	Model -4	42.7419	82.35% P	80.9% core 17.8% allowed 0.8% general 0.4% disallowed	86.30%	5.56%	-3.78 ± 0.39
5.	Model -5	56.4516	81.99% P	84.3% core 13.6% allowed 1.7% general 0.4% disallowed	90.00%	4.07%	-3.87 ± 0.39

6.	6.LOMETS, Model - 1	37.4046	84.93% P	89.8% core 8.1% allowed 1.7% general 0.4% disallowed	92.22%	1.48%	1.13 ± 0.50
7.	Model - 2	50.3788	82.35% P	91.1% core 7.6% allowed 1.3% general 0.0% disallowed	94.44%	0.74%	1.06 ± 0.49
8.	Model -3	46.3878	82.35% P	87.7% core 10.6% allowed 1.7% general 0.0% disallowed	92.96%	1.11%	0.57 ± 0.48
9.	Model -4	38.8462	70.96%	87.3% core 11.9% allowed 0.0% general 0.8% disallowed	91.48%	1.85%	0.56 ± 0.50
10.	Model -5	33.9695	78.68%	87.7% core 9.3% allowed 2.1% general 0.8% disallowed	91.11%	2.59%	0.67 ± 0.51
11.	SparkX ,Model -1	79.8479	86.76% P	91.1% core 7.6% allowed 0.8% general 0.4% disallowed	95.56%	0.74%	0.37 ± 0.47
12.	Model -2	66.1597	69.12%	87.7% core 9.7% allowed 1.7% general 0.8% disallowed	91.48%	1.48%	-1.02 ± 0.45
13.	Model -3	61.7424	78.31%	78.4% core 15.7% allowed 5.1% general 0.8% disallowed	82.22%	6.30%	-3.08 ± 0.44
14.	Model -4	81.8182	72.79%	87.3% core 8.9% allowed 1.3% general 2.5% disallowed	92.22%	2.96%	-0.28 ± 0.48
15.	Model -5	64.2586	73.16%	83.9% core 13.6% allowed 2.1% general 0.4% disallowed	88.89%	3.70%	-1.41 ± 0.50
16.	sparkx Model - 6	51.6129	68.38%	86.0% core 11.4% allowed 2.1% general 0.4% disallowed	90.37%	2.59%	-0.90 ± 0.45
17.	Model - 7	60.2273	76.84%	87.7% core 11.0% allowed 1.3% general 0.0% disallowed	92.22%	1.48%	-0.27 ± 0.49

18.	Model - 8	44.697	66.18%	86.0% core 11.4% allowed 0.4% general 2.1% disallowed	91.48%	2.96%	-0.30 ± 0.47
19.	Model -9	36.0465	60.29%	79.7% core 18.6% allowed 1.3% general 0.4% disallowed	86.30%	2.96%	-2.35 ± 0.43
20.	Model -10	51.8939	67.65%	87.3% core 10.6% allowed 1.7% general 0.4% disallowed	91.85%	1.85%	-0.02 ± 0.48
21.	CEthreader, model -1	54.3307	84.19% P	87.3% core 8.9% allowed 3.0% general 0.8% disallowed	90.74%	3.33%	1.12 ± 0.49
22.	Model -2	42.5287	76.84%	86.9% core 11.0% allowed 0.4% general 1.7% disallowed	92.96%	2.22%	1.14 ± 0.52
23.	Model - 3	33.0769	84.56% P	87.3% core 9.3% allowed 2.1% general 1.3% disallowed	90.74%	2.96%	0.82 ± 0.51
24.	Model -4	49.0494	81.62% P	86.9% core 11.0% allowed 0.8% general 1.3% disallowed	93.70%	1.48%	0.64 ± 0.51
25.	Model -5	50	71.69%	89.4% core 8.9% allowed 1.3% general 0.4% disallowed	95.56%	1.48%	0.95 ± 0.49
26.	IntFold - model -1	87.1595	80.15% P	85.2% core 8.1% allowed 4.2% general 2.5% disallowed	87.04%	8.52%	-0.72 ± 0.51
27.	Model 2	76.1364	82.72% P	91.5% core 6.8% allowed 0.8% general 0.8% disallowed	95.56%	0.74%	-0.34 ± 0.44
28.	Model 3	76.1364	82.72% P	91.5% core 6.8% allowed 0.8% general 0.8% disallowed	95.56%	0.74%	-0.34 ± 0.44
29.	Model 4	76.1364	82.72% P	91.5% core 6.8% allowed 0.8% general 0.8% disallowed	95.56%	0.74%	-0.34 ± 0.44

30.	Model 5	81.8533	80.51% P	85.2% core 9.3% allowed 5.1% general 0.4% disallowed	89.63%	7.04%	-0.34 ± 0.50
31.	MT4	78.7879	71.32%	90.3% core 8.9% allowed 0.4% general 0.4% disallowed	92.59%	1.48%	NA
32.	I-TASSER	100	93.75% P	81.8% core 13.6% allowed 3.0% general 1.7% disallowed	87.78%	1.85%	-4.08 ± 0.40
33.	AIDA	7.34694	14.34%	36.1% core 30.9% allowed 17.6% general 15.5% disallowed	46.07%	31.09%	-7.35 ± 0.32
34.	PRIMO, Model-1	78.3133	75.19%	90.7% core 7.1% allowed 1.8% general 0.4% disallowed	95.38%	1.92%	0.11 ± 0.46
35.	Model -2	82.8685	83.97% P	92.9% core 5.3% allowed 1.3% general 0.4% disallowed	95.00%	1.15%	0.38 ± 0.48
36.	Model -3	76.8924	77.86%	90.7% core 8.0% allowed 0.4% general 0.9% disallowed	95.77%	0.38%	0.58 ± 0.48
37.	Model -4	79.3651	75.57%	90.3% core 8.0% allowed 1.3% general 0.4% disallowed	95.00%	1.15%	0.10 ± 0.47
38.	phyre2	59.0734	83.82% P	80.1% core 16.9% allowed 1.3% general 1.7% disallowed	88.15%	4.07%	-1.17 ± 0.48
39.	FFAS03 model 1	71.9697	75.37%	91.1% core 6.8% allowed 1.7% general 0.4% disallowed	95.56%	1.48%	0.01 ± 0.46
40.	Model 2	86.3813	92.65% P	91.8% core 7.0% allowed 0.8% general 0.4% disallowed	93.70%	0.74%	0.37 ± 0.46

41.	Psv -2	51.9084	88.24% P	83.1% core 13.1% allowed 2.5% general 1.3% disallowed	87.78%	4.07%	-1.37 ± 0.50
42.	Hhpred Model-1	55.3435	90.77% P	85.5% core 12.3% allowed 2.1% general 0.0% disallowed	92.19%	2.23%	0.11 ± 0.53
43.	Model -2	79.1667	87.87%	89.8% core 7.6% allowed 1.7% general 0.8% disallowed	4.07%	1.48%	0.29 ± 0.48
44.	Model -3	63.2576	78.68%	90.7% core 8.1% allowed 1.3% general 0.0% disallowed	94.81%	1.48%	0.36 ± 0.47
45.	Swiss model 1	89.916	78.44%	91.0% core 7.3% allowed 0.9% general 0.9% disallowed	95.13%	1.12%	-1.53 ± 0.46
46.	Model 2	88.6179	79.55%	88.4% core 10.7% allowed 0.4% general 0.4% disallowed	91.01%	2.25%	-0.76 ± 0.50
47.	Robetta, model 1	96.0938	94.49% P	91.1% core 7.6% allowed 0.0% general 1.3% disallowed	96.30%	0.00%	0.63 ± 0.49
48.	Model 2	96.4567	94.49% P	91.1% core 8.1% allowed 0.4% general 0.4% disallowed	96.67%	0.74%	1.04 ± 0.50
49.	Model 3	97.6744	95.59% P	90.7% core 7.2% allowed 1.3% general 0.8% disallowed	95.93%	1.48%	0.93 ± 0.49
50.	Model 4	94.0711	94.12% P	94.1% core 5.1% allowed 0.8% general 0.0% disallowed	97.78%	0.37%	0.89 ± 0.48
51.	Model 5	96.0784	96.69% P	90.7% core 8.5% allowed 0.4% general 0.4% disallowed	98.71%	1.11%	0.90 ± 0.50

The best 8 models of PSY1R kinase out of 51 models were selected using the evaluation table for performing molecular dynamics simulation at 5ns.

Table 4.6 : Evaluation of best eight models of PSY1R kinase:

	ERRAT Goal: >50 %	VERIFY 3D Goal: >80%	Ramachandran score (Procheck)	Ramachandran score (Molprobity) Goal: >98%	Ramachandran Outliers (Molprobity) Goal: <0.05%	RMS Z-SCORE (Molprobity) Goal: abs(Z score) < 2
IntFold Model 2	76.1364	82.72% P	91.5% core 6.8% allowed 0.8% general 0.8% disallowed	95.56%	0.74%	-0.34 ± 0.44
I-TASSER	100	93.75% P	81.8% core 13.6% allowed 3.0% general 1.7% disallowed	87.78%	1.85%	-4.08 ± 0.40
SparkX ,Model -1	79.8479	86.76% P	91.1% core 7.6% allowed 0.8% general 0.4% disallowed	95.56%	0.74%	0.37 ± 0.47
FFAS03 Model- 2	86.3813	92.65% P	91.8% core 7.0% allowed 0.8% general 0.4% disallowed	93.70%	0.74%	0.37 ± 0.46
Robetta, Model- 1	96.0938	94.49% P	91.1% core 7.6% allowed 0.0% general 1.3% disallowed	96.30%	0.00%	0.63 ± 0.49
Robetta, Model- 4	94.0711	94.12% P	94.1% core 5.1% allowed 0.8% general 0.0% disallowed	97.78%	0.37%	0.89 ± 0.48
Robetta, Model- 5	96.0784	96.69% P	90.7% core 8.5% allowed 0.4% general 0.4% disallowed	98.71%	1.11%	0.90 ± 0.50
PRIMO Model -2	82.8685	83.97% P	90.7% core 7.1% allowed 1.8% general 0.4% disallowed	95.00%	1.15%	0.38 ± 0.48

The best four models out of 15 models of SERK1 kinase with the lowest DOPE score were selected for further evaluation and for performing molecular dynamic simulation at 5 ns.

Table 4.7: Evaluation of best four models of SERK1 kinase:

Model number	ERRAT Goal: >50 %	VERIFY 3D Goal: >80%	Ramachandran score (Procheck)	Ramachandran score (Molprobity) Goal: >98%	Ramachandran Outliers (Molprobity) Goal: <0.05%	RMS Z-SCORE (Molprobity) Goal: abs(Z score) < 2
Model no. 1	86.7857	93.75%	93.2% core 4.8% allowed 1.6% general 0.4% disallowed	95.45%	0.70%	0.57 ± 0.47
Model no.4	79.6429	91.67%	93.2% core 4.4% allowed 2.0% general 0.4% disallowed	96.85%	1.05%	1.00 ± 0.46 (close to 1)
Model no. 5	84.9462	96.88%	93.2% core 4.4% allowed 2.0% general 0.4% disallowed	95.45%	0.70%	0.61 ± 0.46
Model no. 14	84.8375	91.32%	91.6% core 6.0% allowed 1.2% general 1.2% disallowed	95.10%	2.80%	0.94 ± 0.47

4.11 Result of Molecular Dynamics Simulation:

The GROMACS software suite has been used to run molecular dynamics simulations in a biomolecular system to analyze the conformations of the protein models of both PSY1R kinase and SERK1 kinase that were finally selected by first being filtered on the basis of the above-mentioned criteria. The selected PDB files were checked beforehand in case there are missing residues or atoms or any kind of irregularities or anomalies. The 8 selected models of PSY1R kinase and 4 selected models of SERK1 kinase were energy minimized and equilibrated with 1 ns NVT(temperature) and NPT(pressure) before the final MD production run at 5ns.

The most important property to look for is whether or not the protein is stable. The standard way to understand this can be the root mean square deviation (RMSD) of all atoms. The calculation of RMSD is the most worthwhile feature for structural comparison among different

conformations of the same molecule. This analysis represents a modification in the secondary structure of the protein while the simulation is run. Values of RMSD fluctuating at different nm represent a conformational modifying process from the very start to the end of the simulation. So the less fluctuation indicated by the RMSD graphs the better and stable the 3D structure. Hence this is the most important criteria to take into account for selecting the best 3D structures produced.

GROMACS features a program that prompts a fit group and a group to compute RMSD to perform the production and visualization of the RMSD graphs. Backbone has been chosen for both the groups when prompted. The result is written to rmsd.xvg, and with the installed program Grace a final graph was produced with the command "xmgrace rmsd.xvg."

Again, vibrations that rely on local structure flexibility around the equilibrium are not that random. Moreover, the root mean square fluctuation (RMSF) of every residue is quite forward to calculate over the period of the trajectory. Meanwhile, there are also two more important properties that are convenient to be analyzed, which are the protein size defined by the "radius of gyration" and the hydrogen bonds number.

Next, the radius of gyration (R_g) was observed which is more likely the measure of the protein compactness. A relatively steady value of R_g represents that the protein is stably folded. If a protein starts unfolding, its R_g will change over time. It is to be noted that some components tend to periodically deform a bit of structure during the process of simulation, but they also reform very quickly. The overall structure that is observed finally is found to be rather stable over a five ns period of simulation, as seen from the observations of the graphs of some of the structures.
The overall agreement is reasonable for a protein this small and a brief simulation and the observations are given below :

RMSD graphs produced by the best 8 models of PSY1R kinase are shown below :

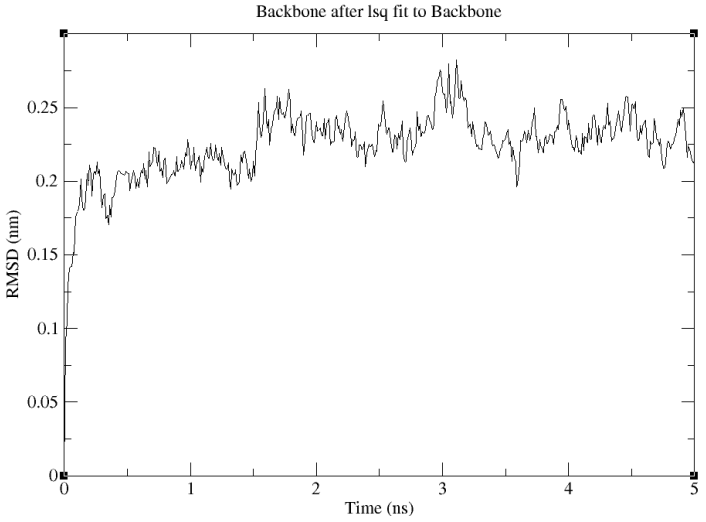

Figure 4.12: RMSD graph of the model Sparkx

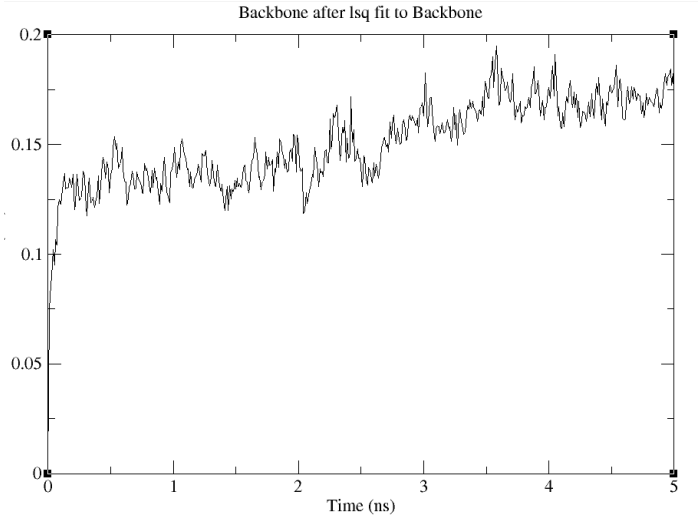

Figure 4.13: RMSD graph of the model Robetta (5th)

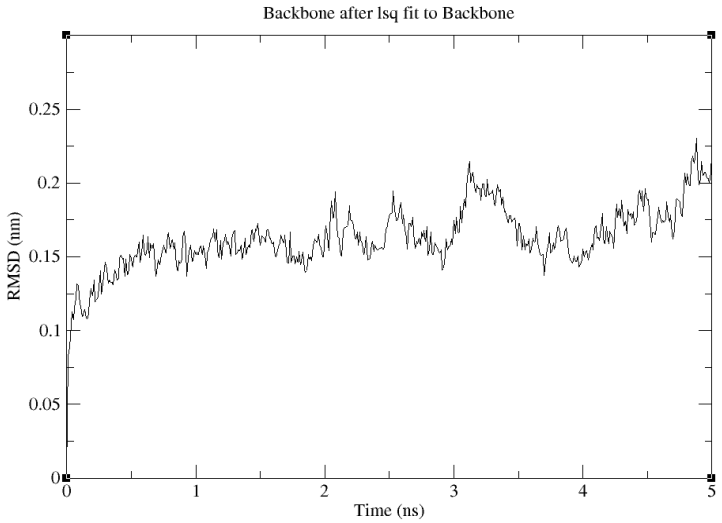

Figure 4.14: RMSD graph of the model Robetta(4th)

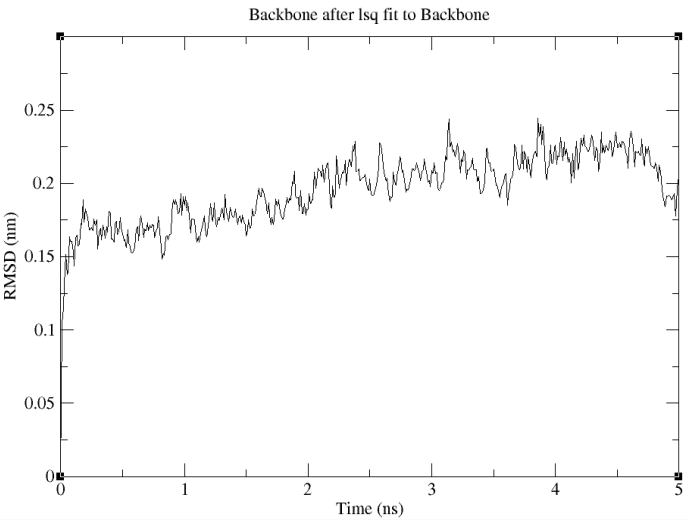

Figure 4.15: RMSD graph of the model FFAS03

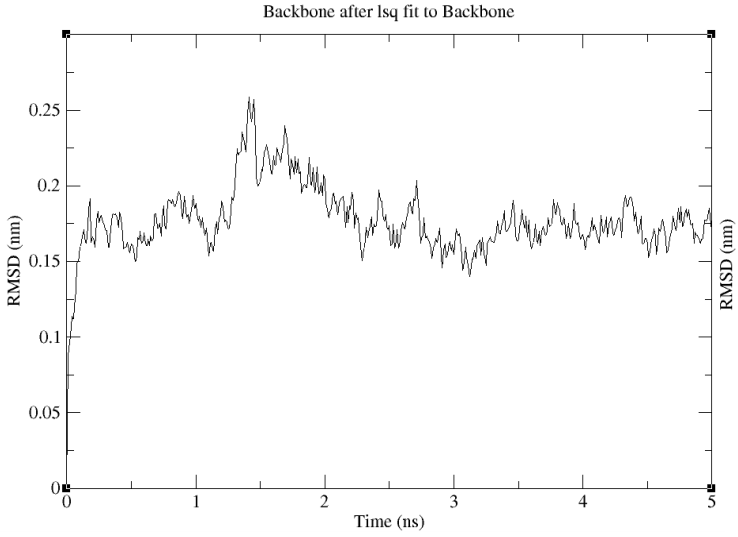

Figure 4.16: RMSD graph of the model IntFold

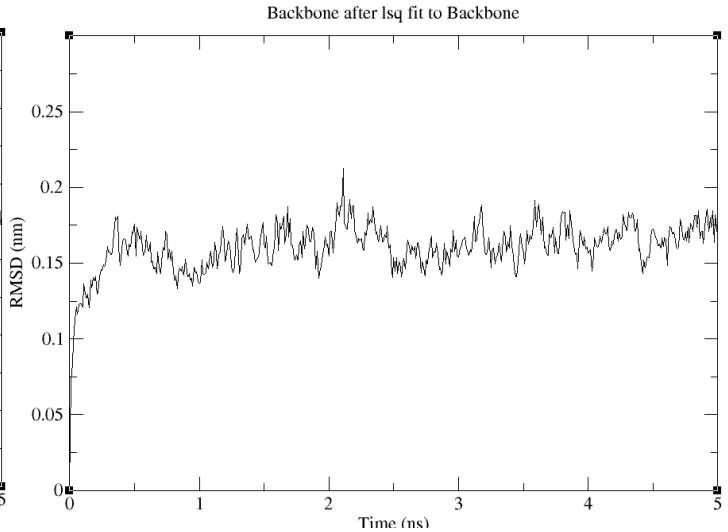

Figure 4.17: RMSD graph of the model Robetta(1st)

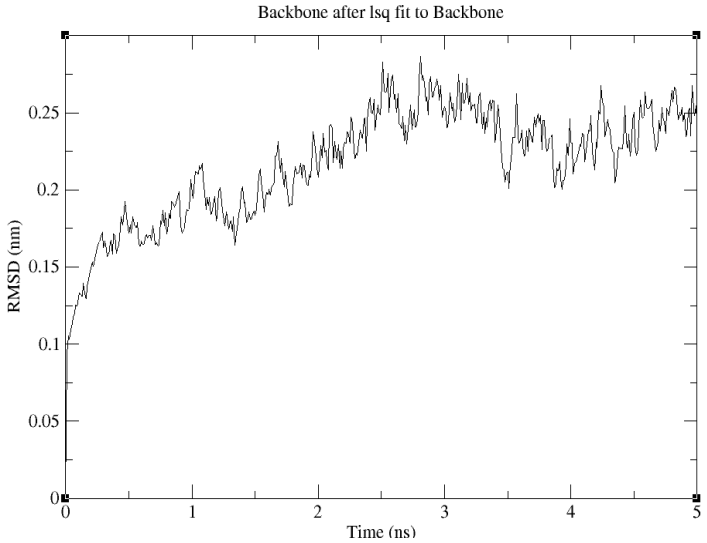

Figure 4.18: RMSD graph of the model I-TASSER

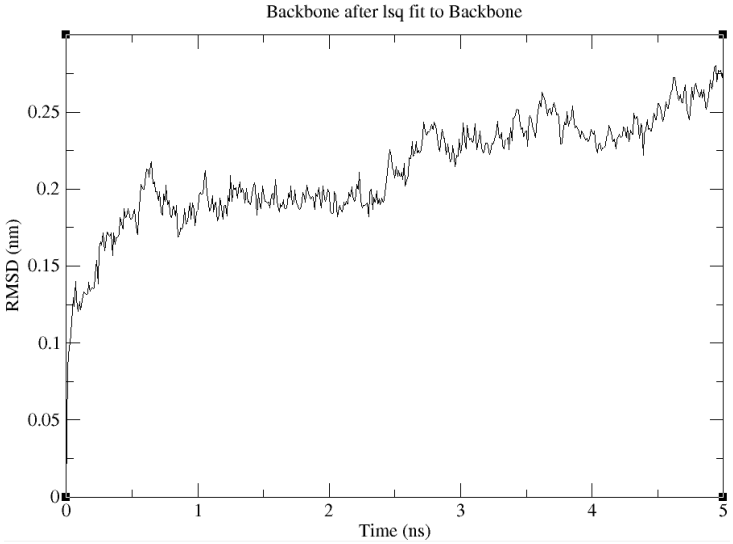

Figure 4.19: RMSD graph of the model PRIMO

The best RMSD graph was observed with the Robetta1 model with the lowest amount of fluctuation from the start to the end of the simulation, hence hinting at possessing the most structural stability.

Radius of gyration graphs produced by the best 8 models of PSY1R kinase are shown below :

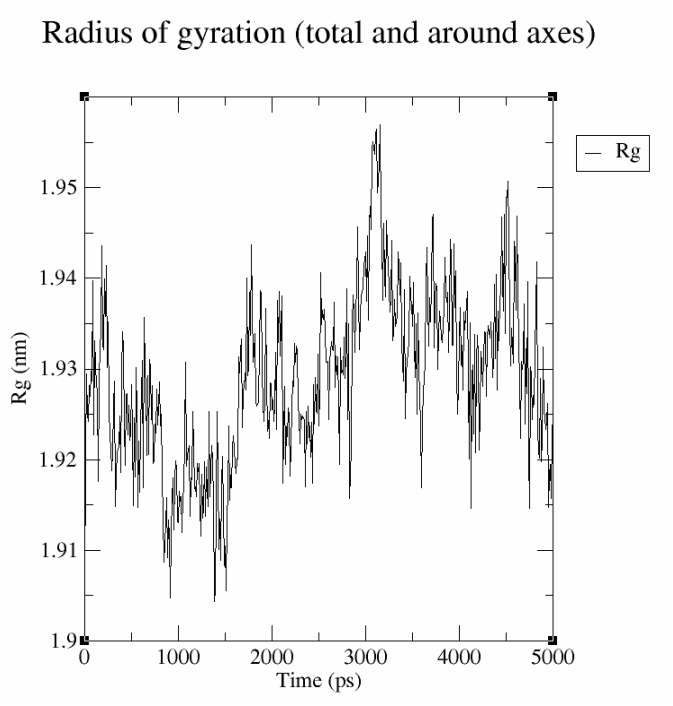

Figure 4.20: Rg graph of the model of SPARKS-X

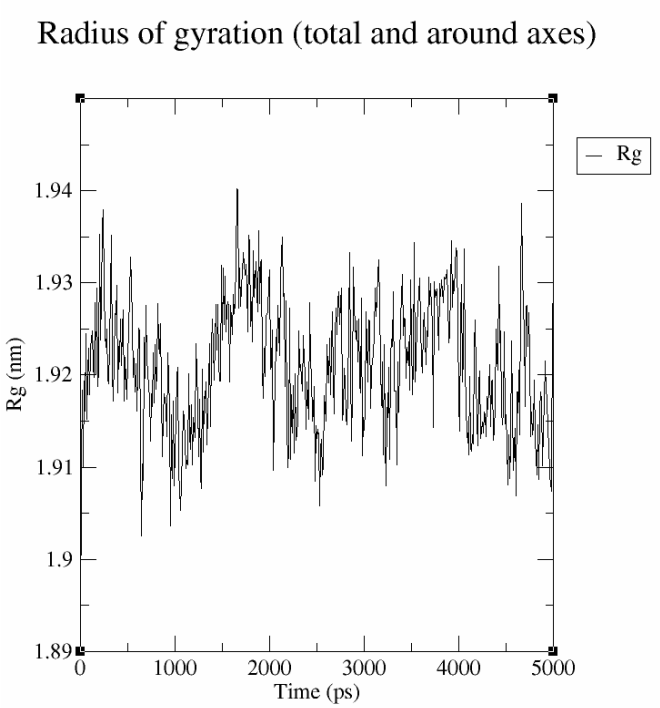

Figure 4.21: Rg graph of the model Robetta (5th)

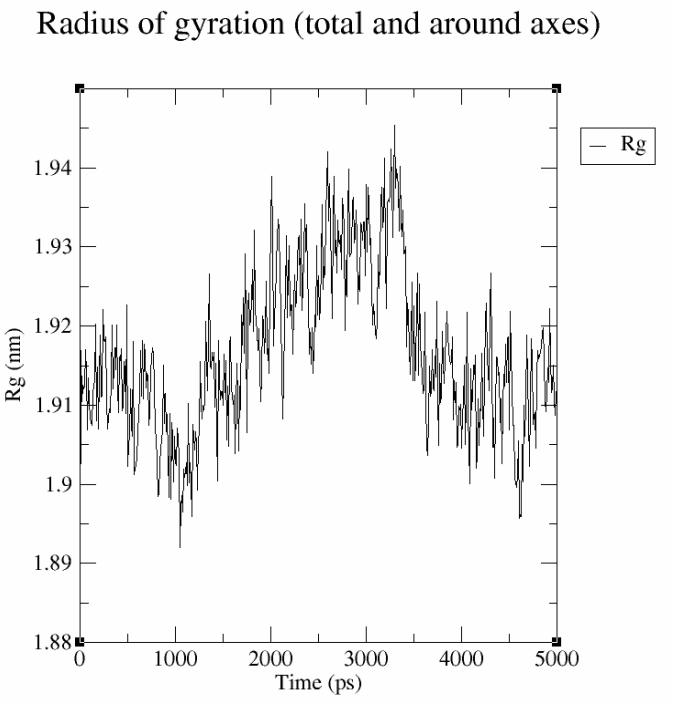

Figure 4.22: Rg graph of the model Robetta(4th)

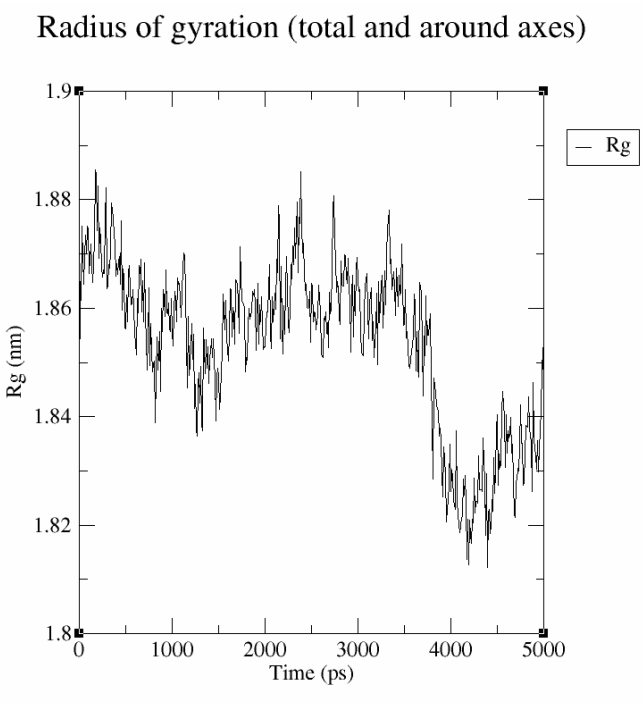

Figure 4.23: Rg graph of the model FFAS03

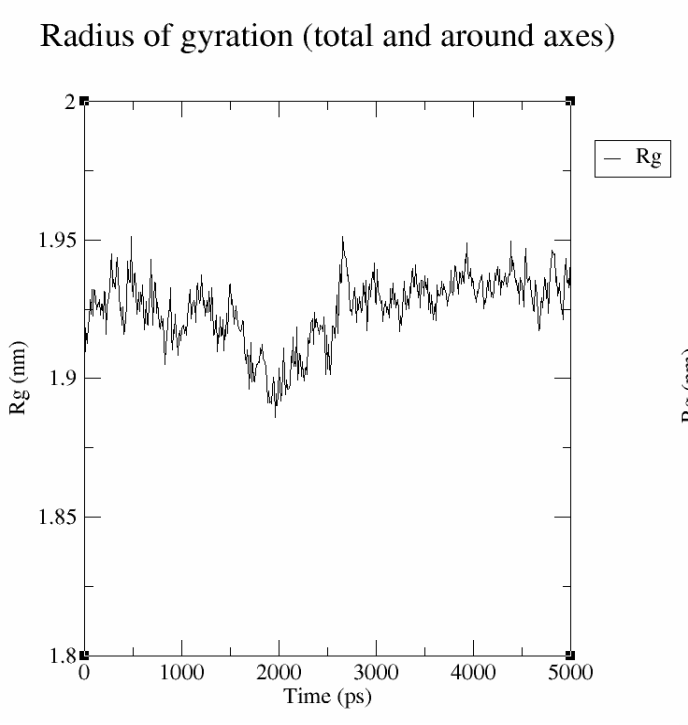

Figure 4.24: Rg graph of the model IntFold

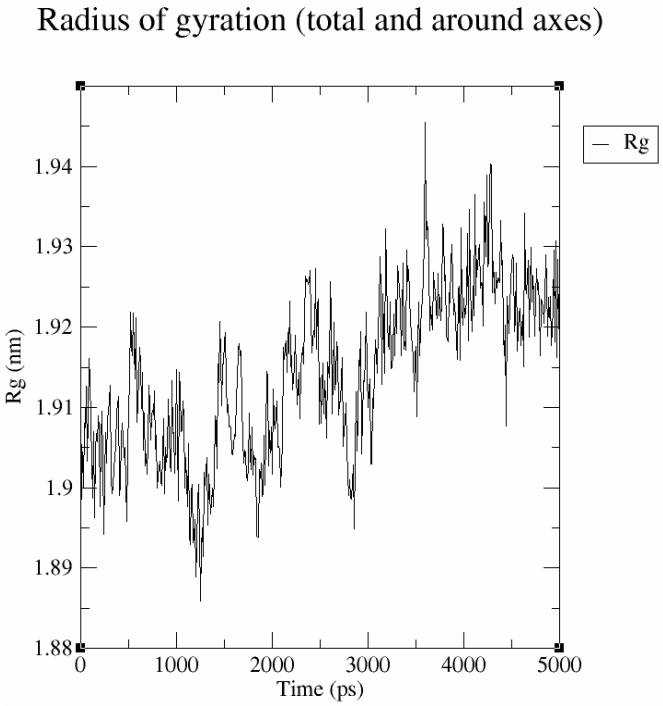

Figure 4.25: Rg graph of the model Robetta(1st)

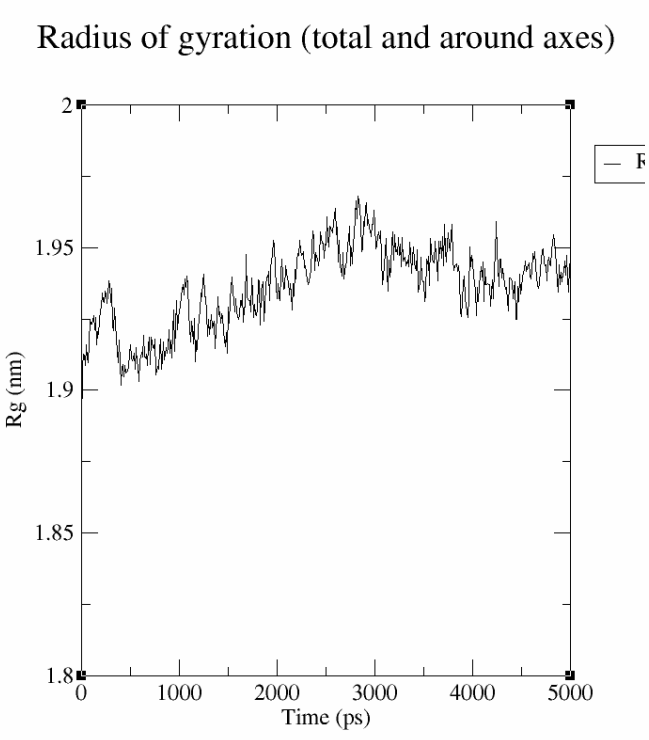

Figure 4.26: Rg graph of the model I-TASSER

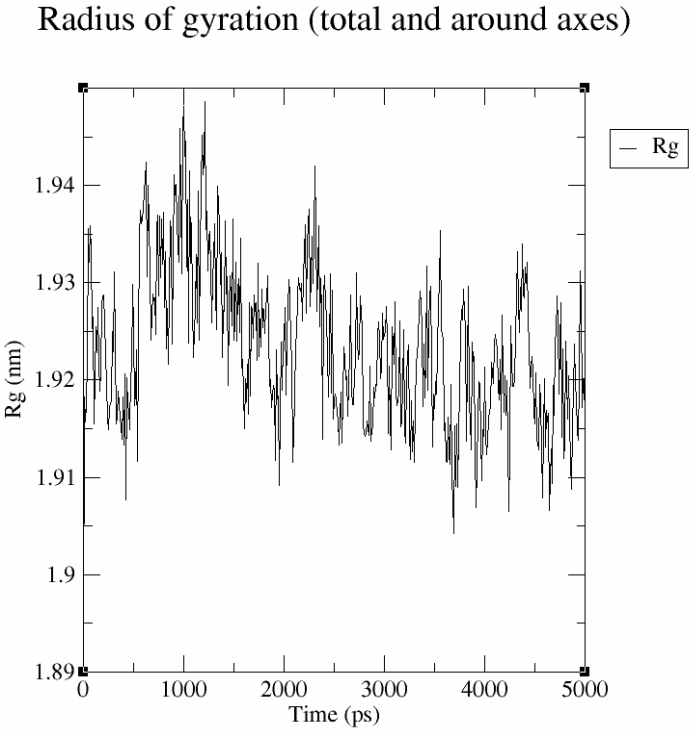

Figure 4.27: Rg graph of the model PRIMO

The deflection of most of the graphs of the radius of gyration can be seen around 189nm to 1.95 nm. By taking all the factors into consideration Robetta model 1 is taken for the further processes.

Views of Robetta model no 1 of PSY1R kinase on different angles :

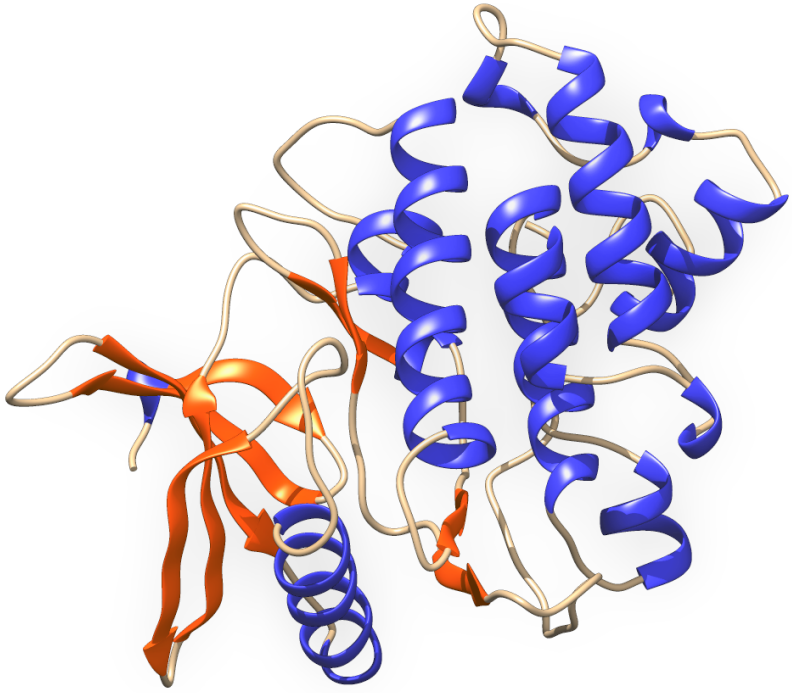

Figure 4.28: Model 1 of PSY1R kinase produced by Robetta (view 1)

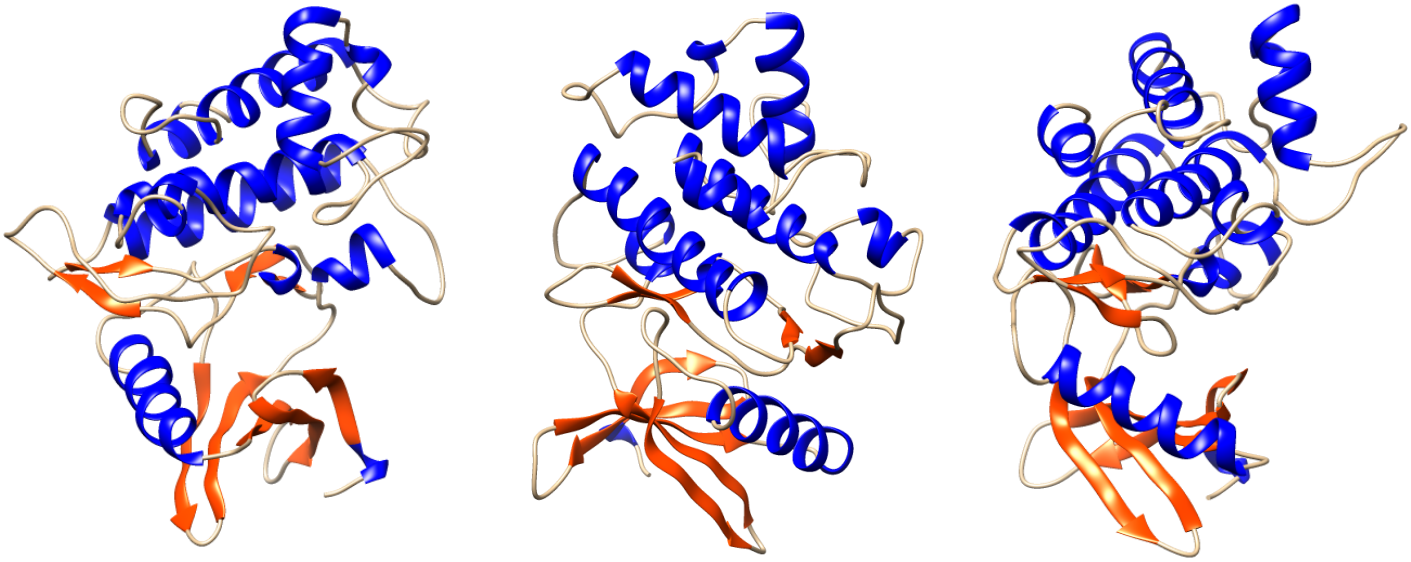

Figure 4.29: Model 1 of PSY1R kinase produced by Robetta (view 2, 3 and 4 respectively)

RMSD graphs produced by the best 4 models of SERK1 kinase are shown below :

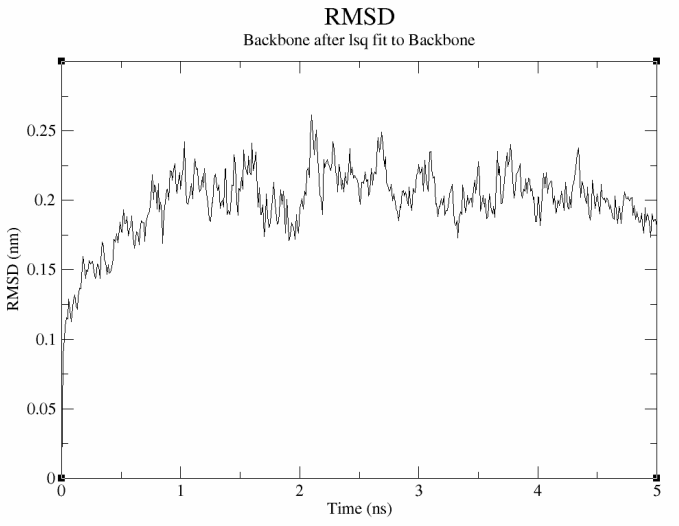

Figure 4.30: RMSD graph of Model no. 1

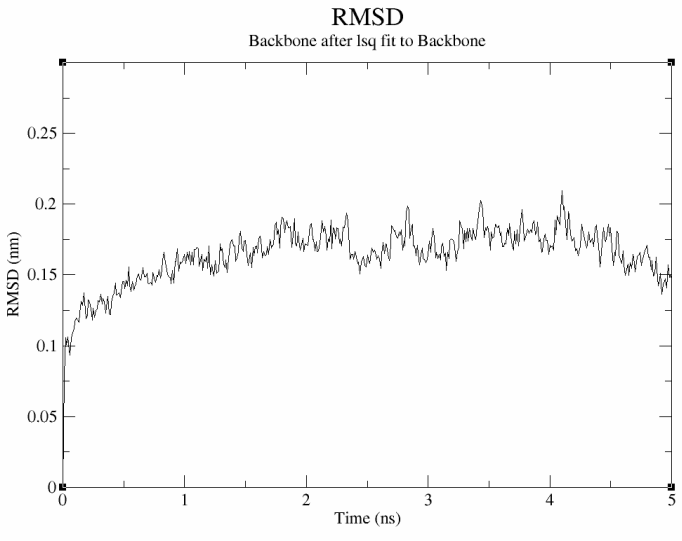

Figure 4.31: RMSD graph of Model no.4

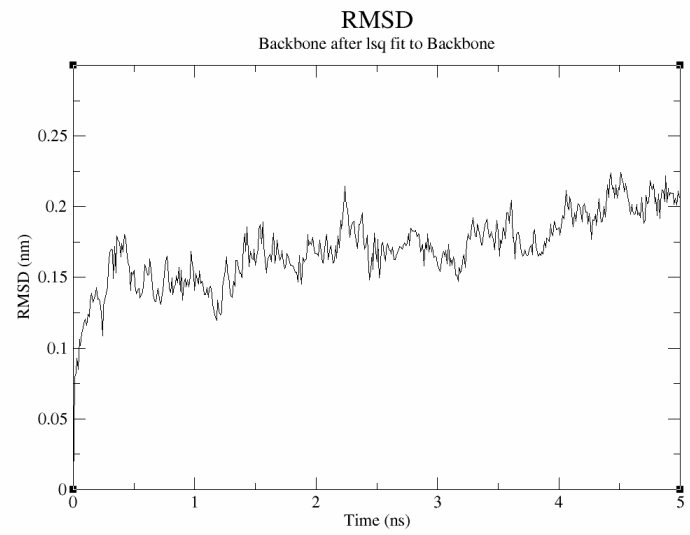

Figure 4.32: RMSD graph of Model no .5

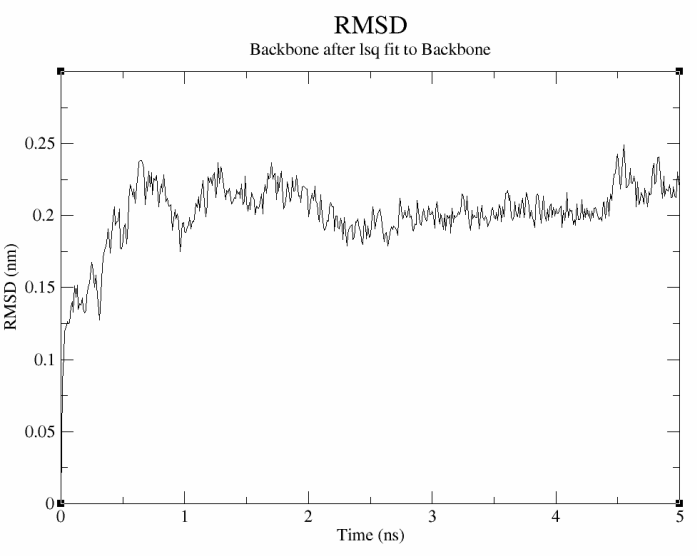

Figure 4.33: RMSD graph of Model no. 14

From the observation, it can be said that model number 4 of SERK1 kinase gave the best graph in RMSD.

Radius of gyration graphs produced by the best 4 models of SERK1 kinase are shown below :

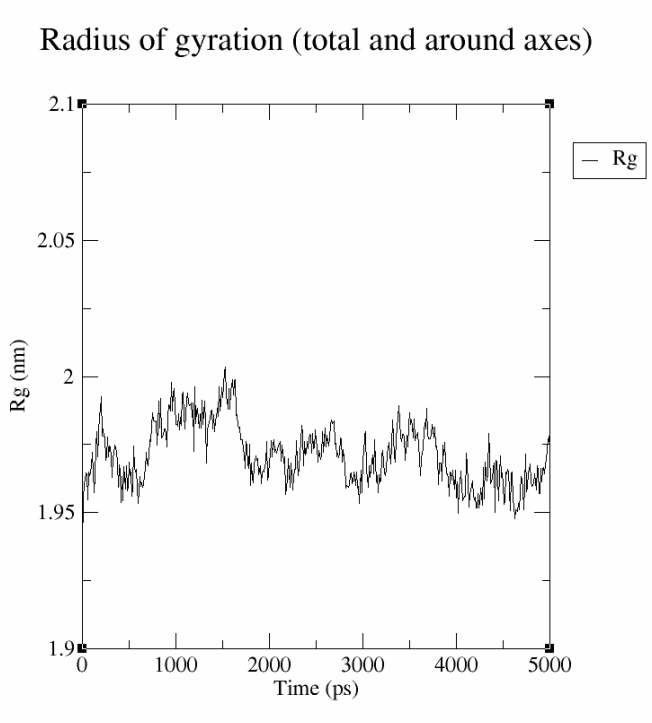

Figure 4.34: Rg graph of Model no.1

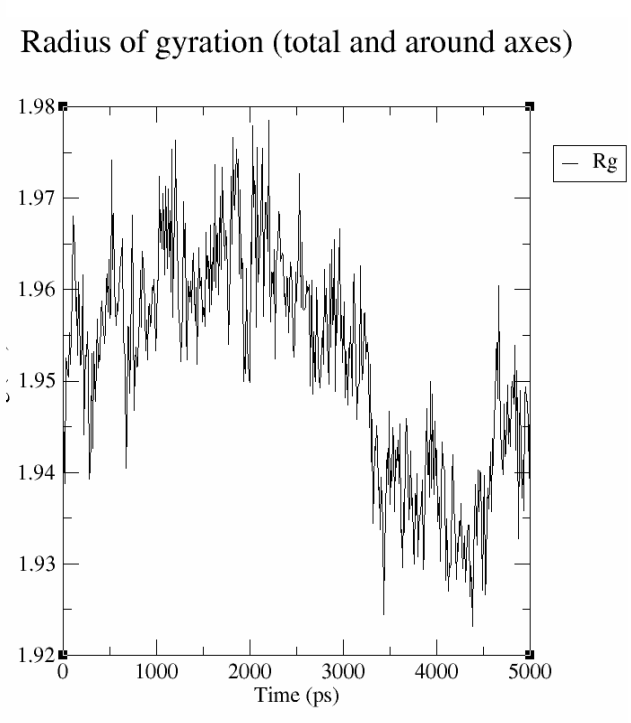

Figure 4.35: Rg graph of Model no.4

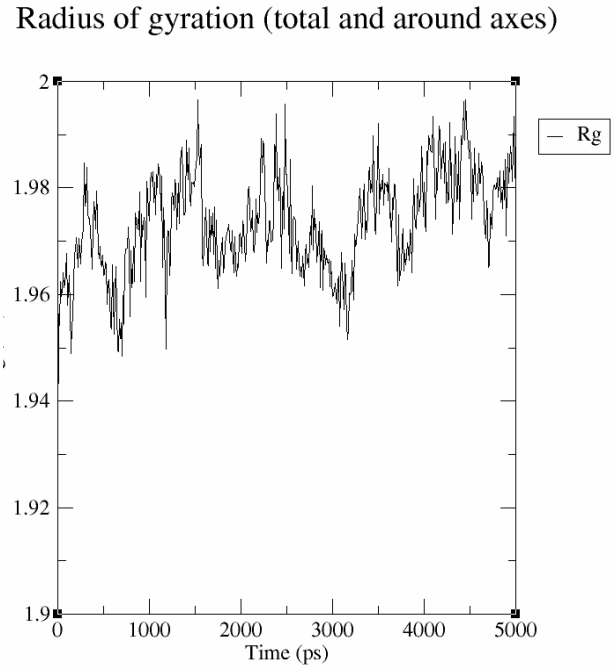

Figure 4.36: Rg graph of Model no.5

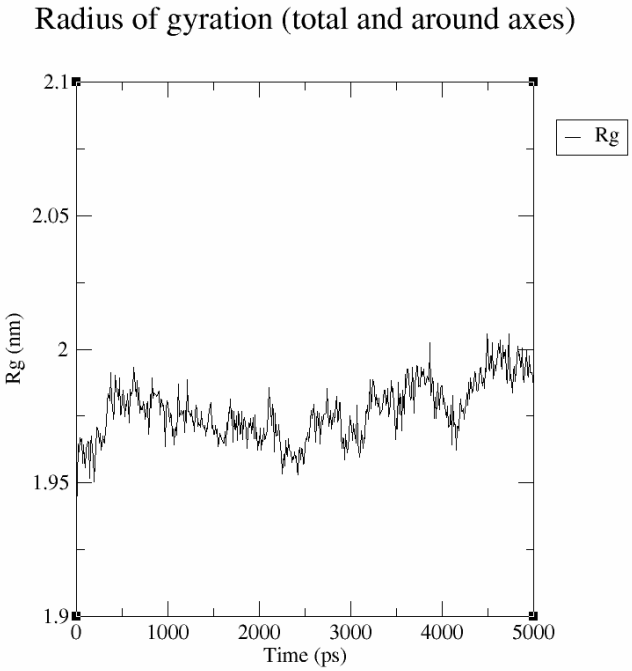

Figure 4.37: Rg graph of Model no.14

The deflection of most of the graphs of radius of gyration can be seen around 1.95nm to1.99 nm. Model no.4 was selected for the next steps as this model has shown the lowest amount of fluctuation, hence indicating toward more stability.

SERK1 kinase structure of the selected Model 4 on various angles :

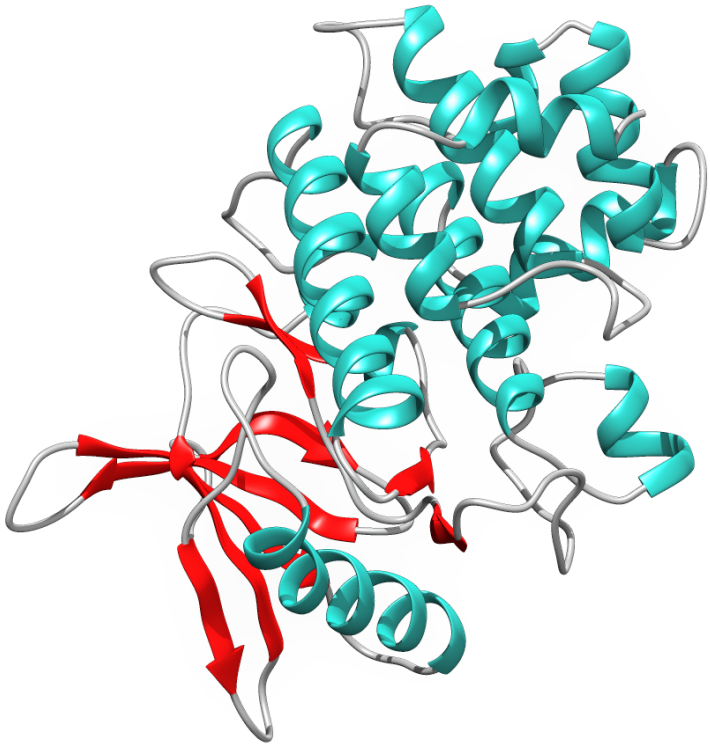

Figure 4.38 : Model 4 of SERK1 kinase produced by MODELLER (view 1)

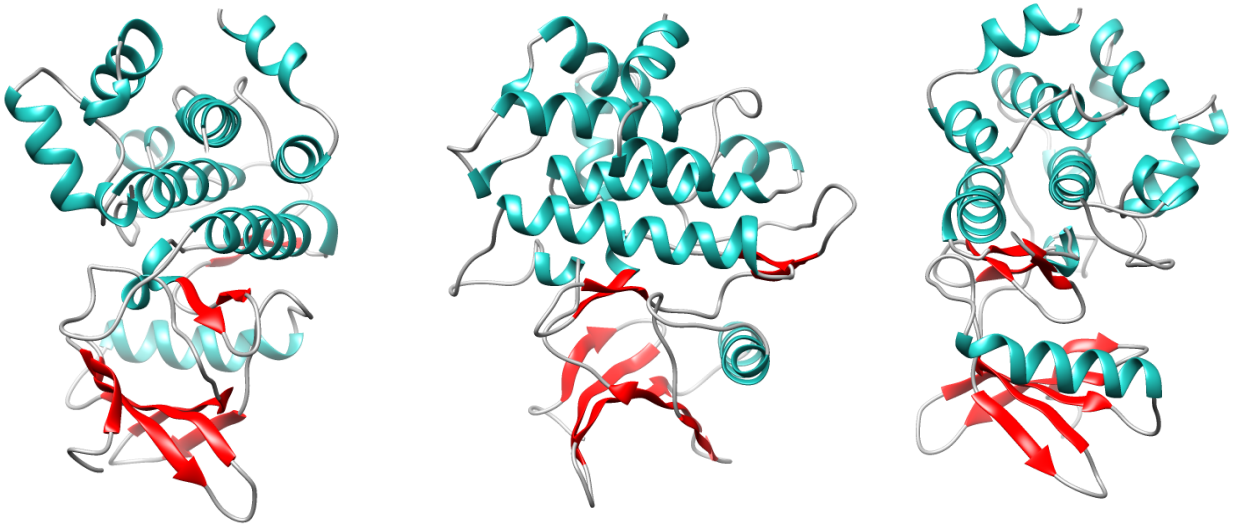

Figure 4.39: Model 4 of SERK1 kinase produced by MODELLER(view 2, 3 and 4 respectively)

4.12 Results of Protein-Protein Docking :

Docking was performed between PSY1R kinase and SERK1 kinase using the protein-protein docking tool ClusPro. Before docking, the selected structure of PSY1R kinase and SERK1 kinase were energy minimized to lessen steric clashes with the help of Chiron. Ten docking complexes on the "balanced" state are selected for inspection that have beeb produced by the ClusPro.

Views of the ten complexes at a glance are given below:

In the complexes , PSY1R kinase has been shown in red color whereas SERK1 kinase has been presented in the color green .

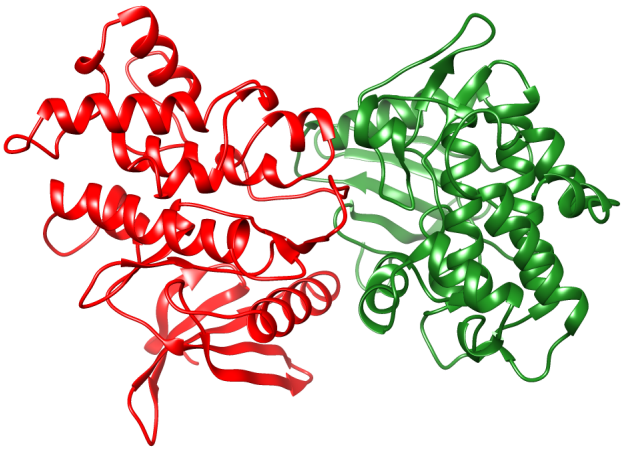

Figure 4.40: protein complex 1

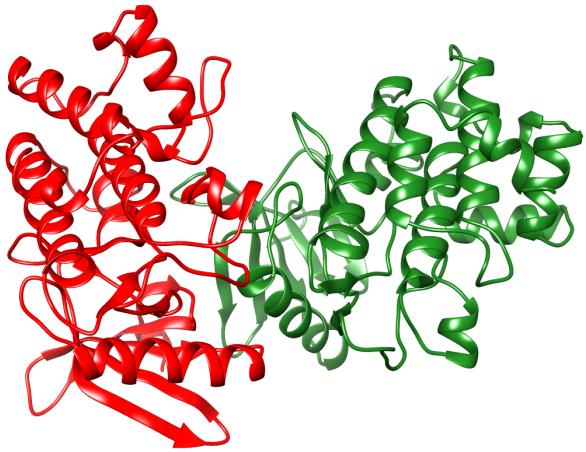

Figure 4.41: protein complex 2

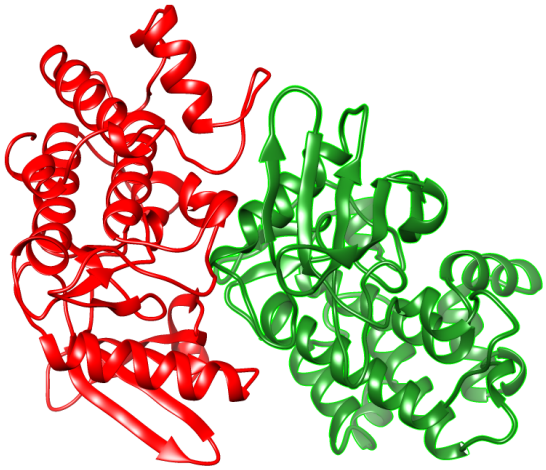

Figure 4.42: protein complex 3

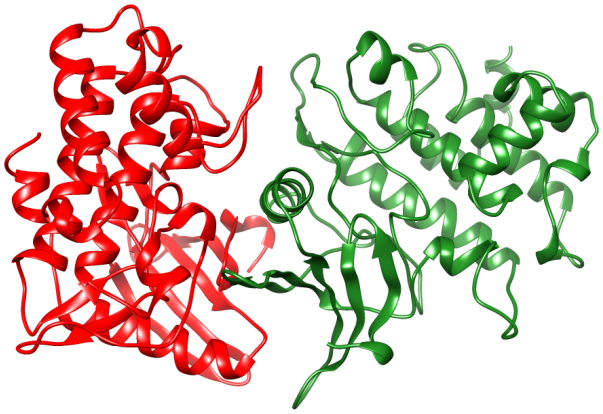

Figure 4.43: protein complex 4

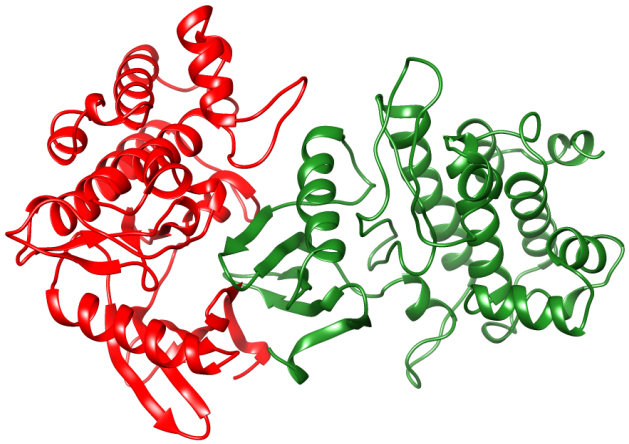

Figure 4.44: protein complex 5

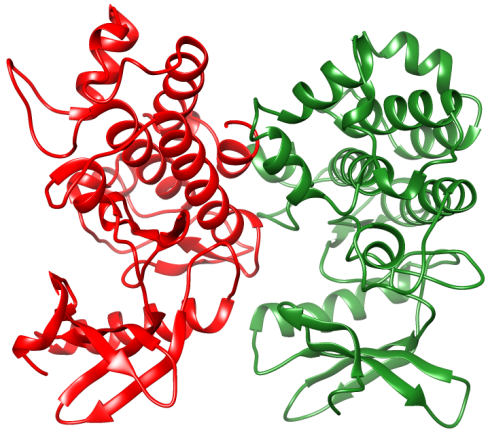

Figure 4.45: protein complex 6

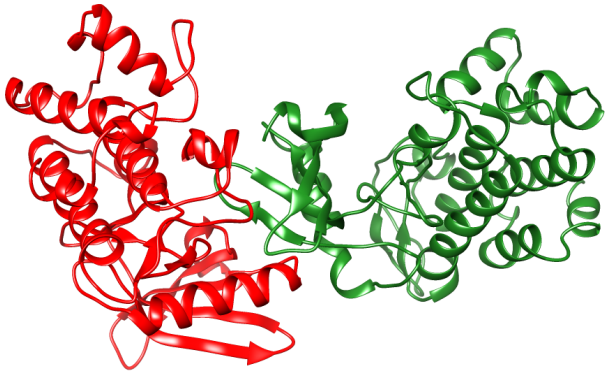

Figure 4.46: protein complex 7

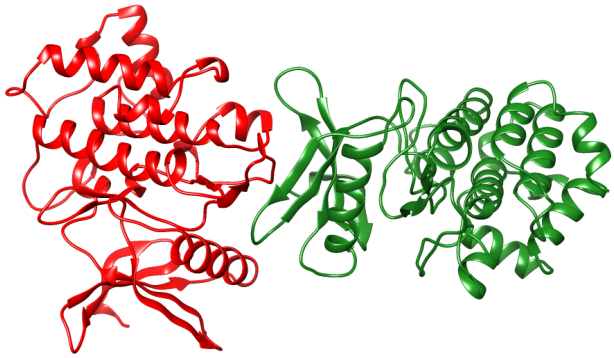

Figure 4.47: protein complex 8

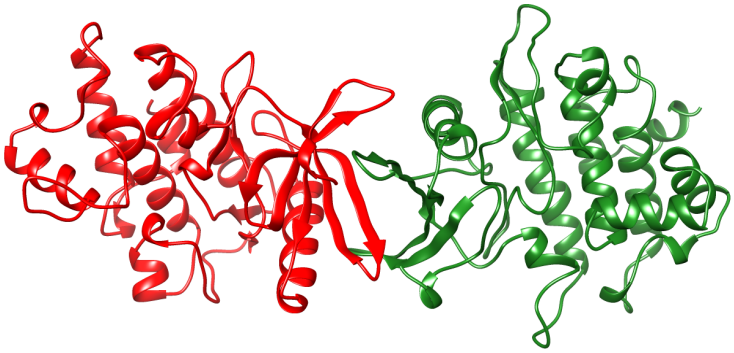

Figure 4.48: protein complex 9

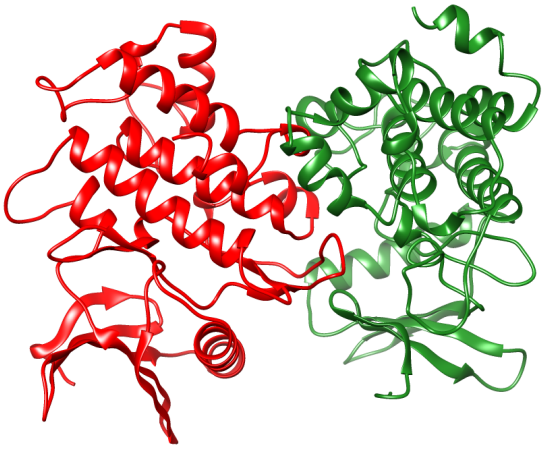

Figure 4.49: protein complex 10

4.13 Results of molecular dynamics simulation of the protein-protein complexes :

The ten complexes produced by docking have been subjected to a 10 ns molecular dynamics simulation after being energy minimized and equilibrated at 1ns NVT (temperature) and NPT (pressure).

RMSD graphs produced by the best 10 protein-protein complexes models are shown below :

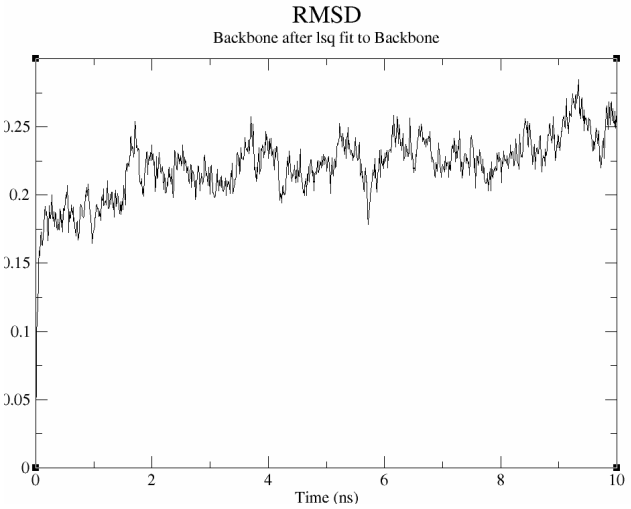

Figure 4.50: RMSD graph of protein complex 1

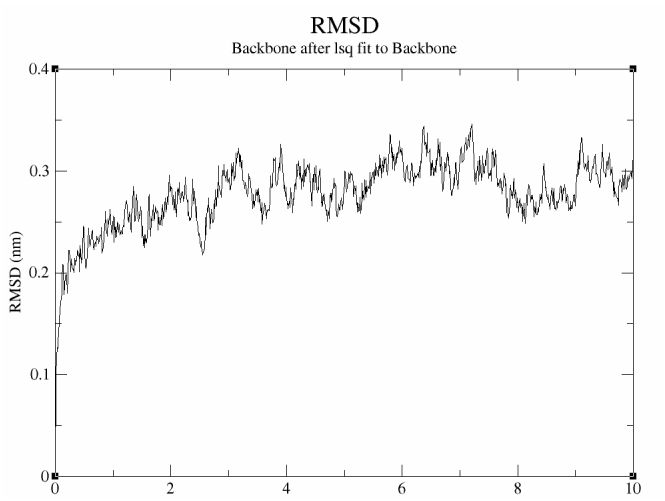

Figure 4.51: RMSD graph of protein complex 2

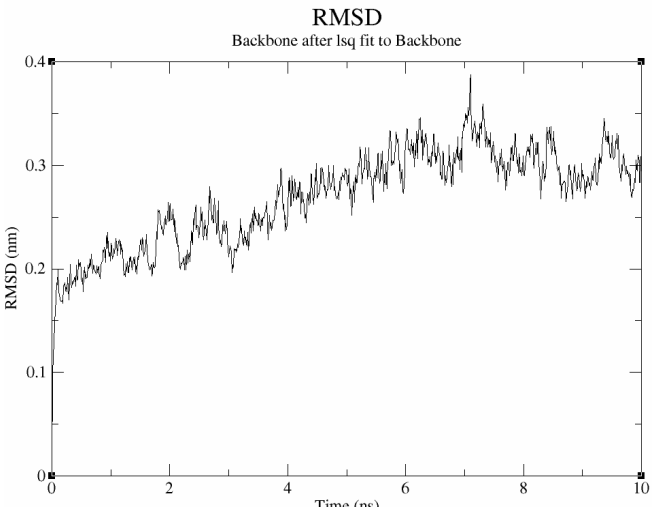

Figure 4.52: RMSD graph of protein complex 3

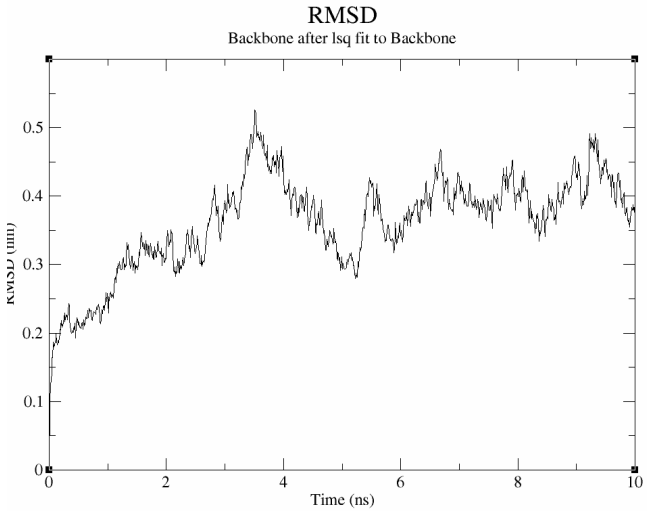

Figure 4.53: RMSD graph of protein complex 4

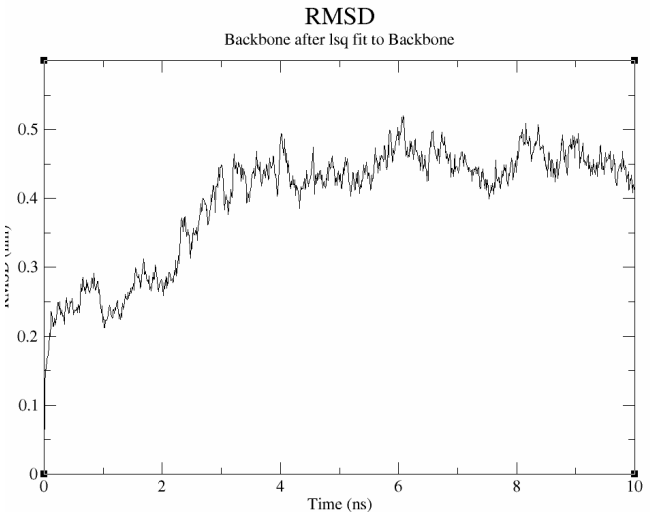

Figure 4.54: RMSD graph of protein complex 5

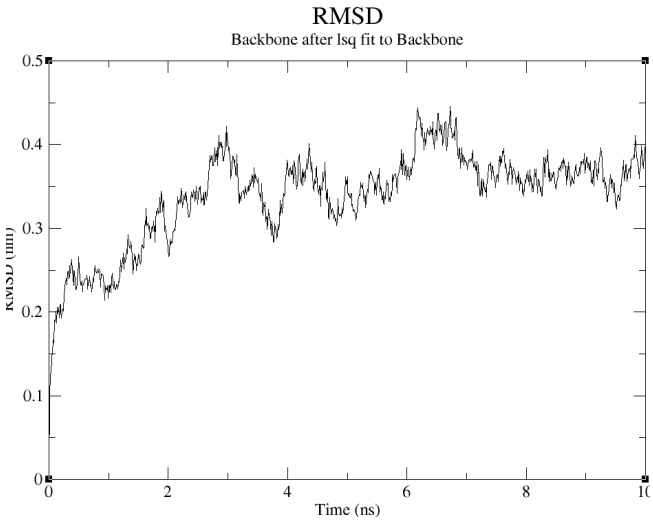

Figure 4.55: RMSD graph of protein complex 6

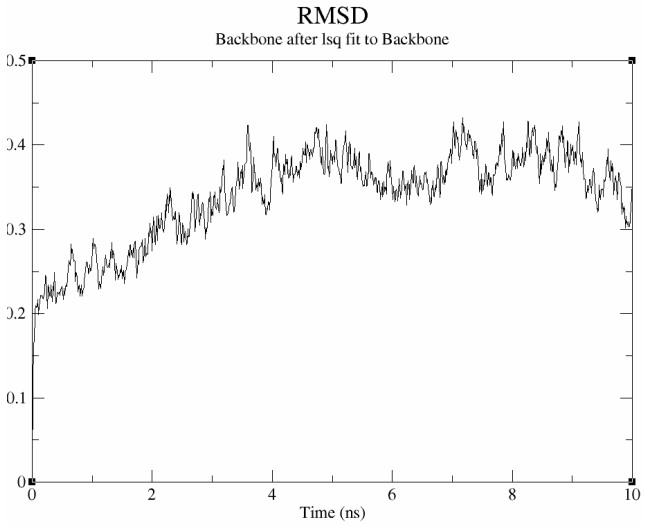

Figure 4.56: RMSD graph of protein complex 7

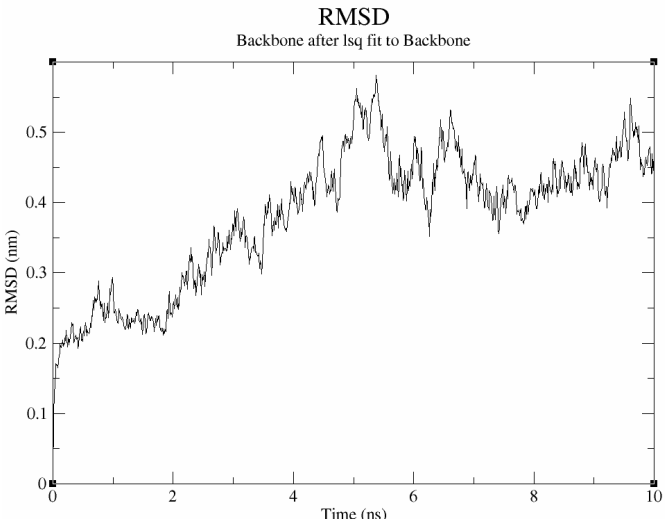

Figure 4.57: RMSD graph of protein complex 8

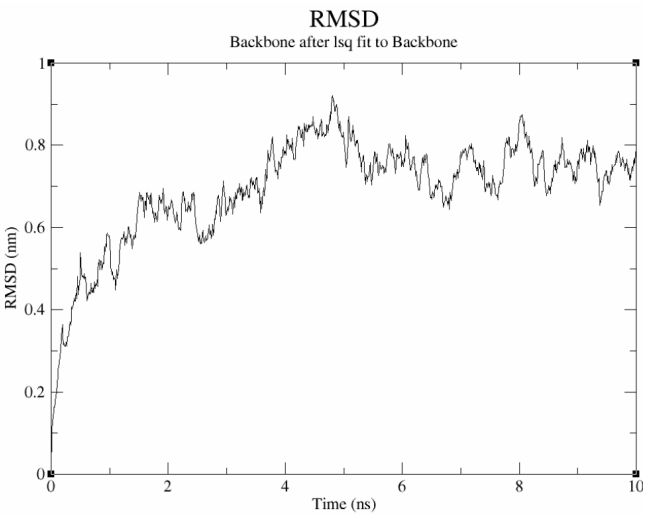

Figure 4.58: RMSD graph of protein complex 9

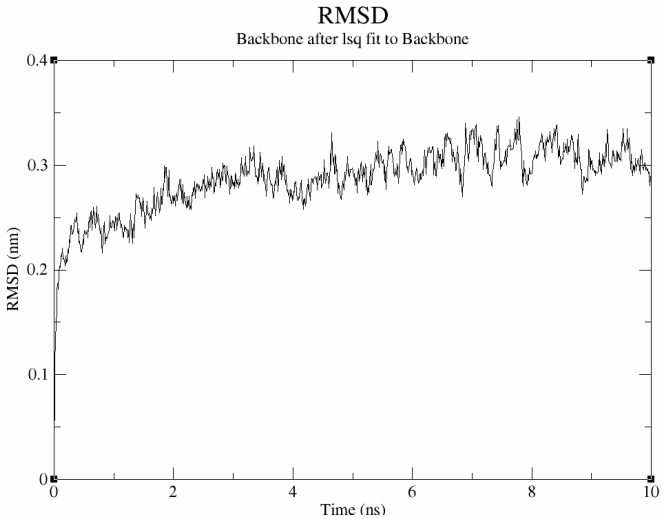

Figure 4.59: RMSD graph of protein complex 10

The RMSD graph of protein complexes 1, 2 and 10 seems comparative stable with less amount of fluctuation. Later the Radius of gyration analysis is observed.

The radius of gyration graphs produced by the best 10 protein-protein complex models are shown below :

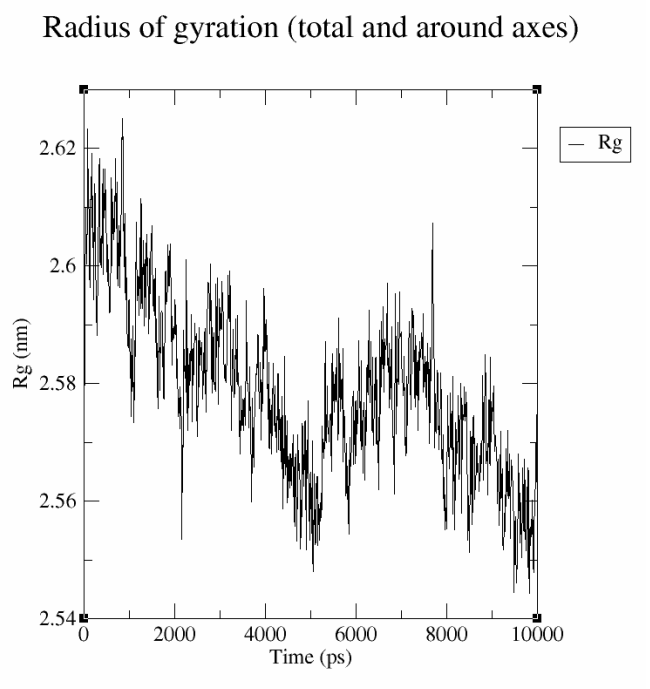

Figure 4.60 :Rg graph of Protein complex 1

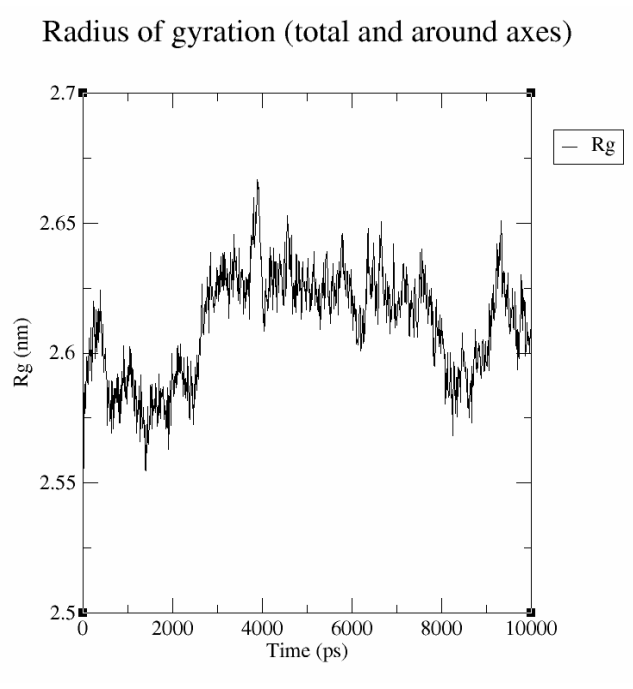

Figure 4.61: Rg graph of protein complex

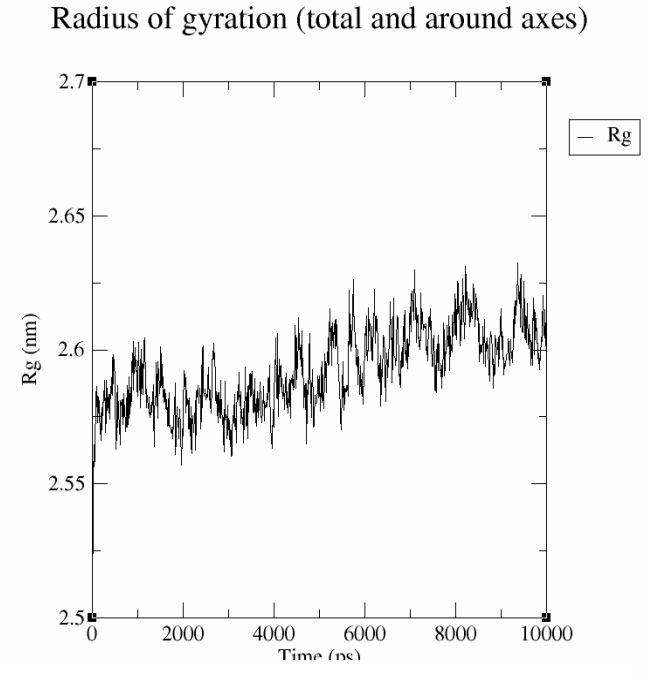

Figure 4.62: Rg graph of protein complex 3

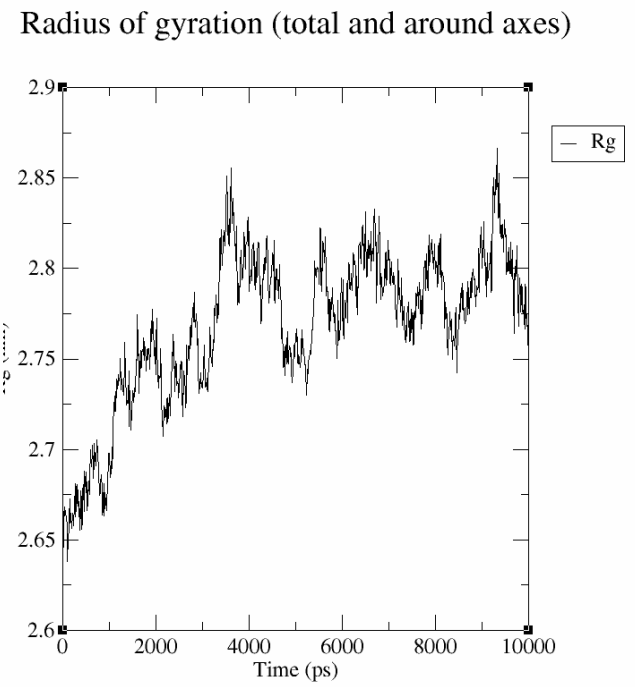

Figure 4.63: Rg graph of protein complex 4

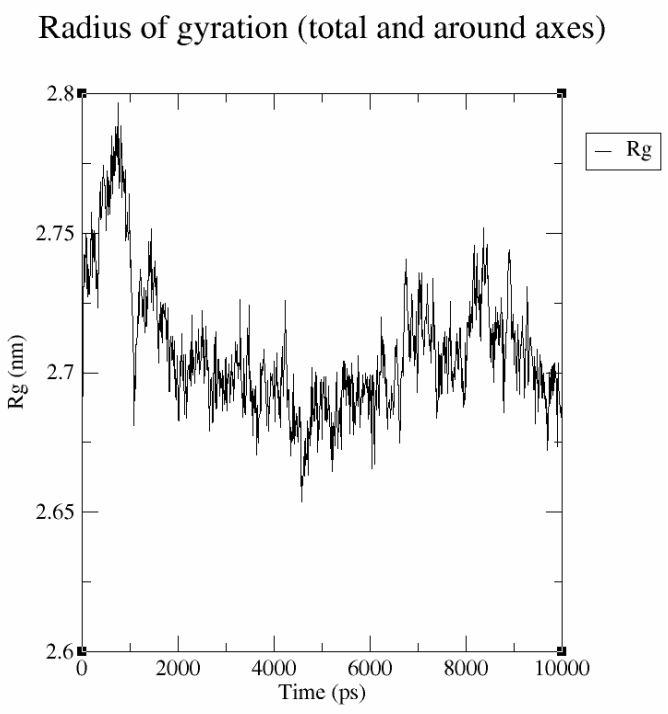

Figure 4.64 : Rg graph of protein complex 5

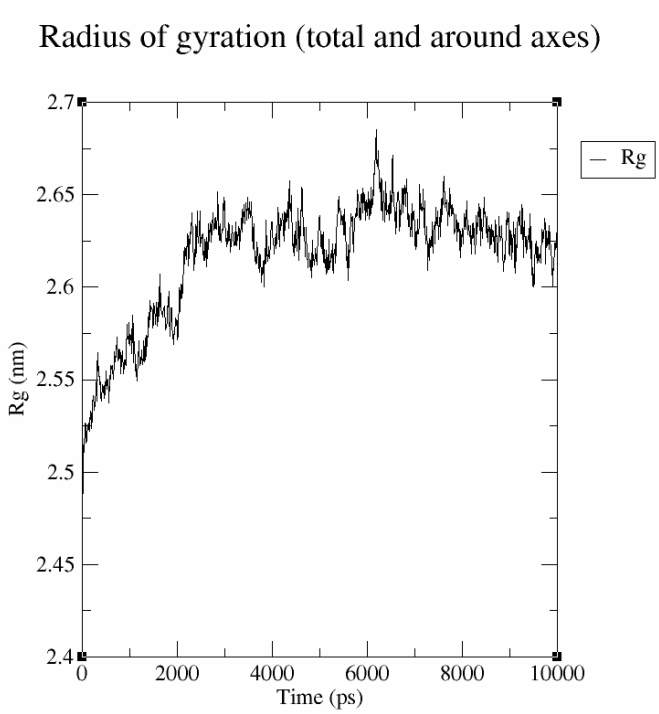

Figure 4.65: Rg graph of protein complex 6

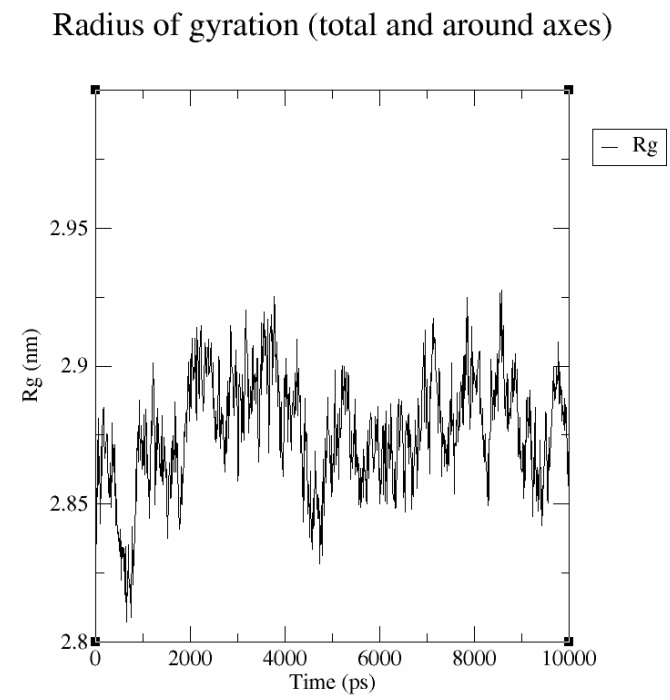

Figure 4.66: Rg graph of protein complex 7

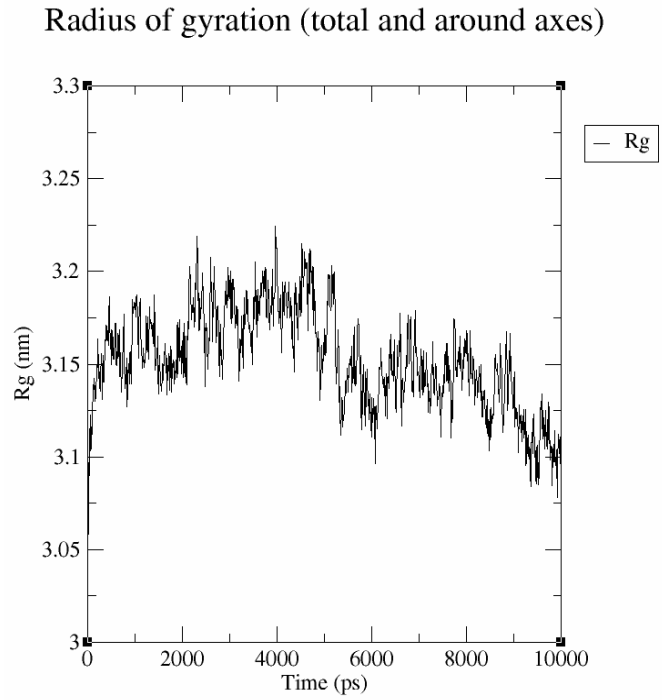

Figure 4.67 : Rg graph of protein complex 8

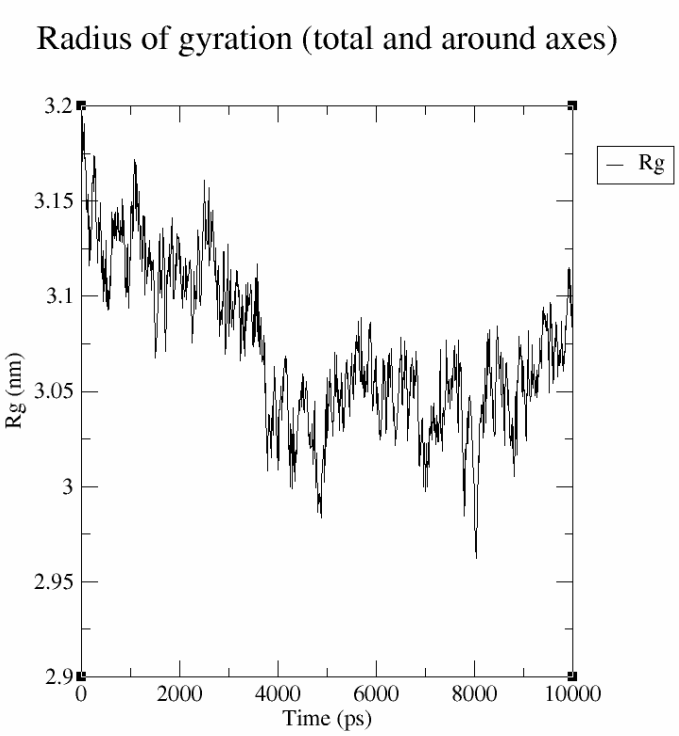

Figure 4.68: Rg graph of protein complex 9

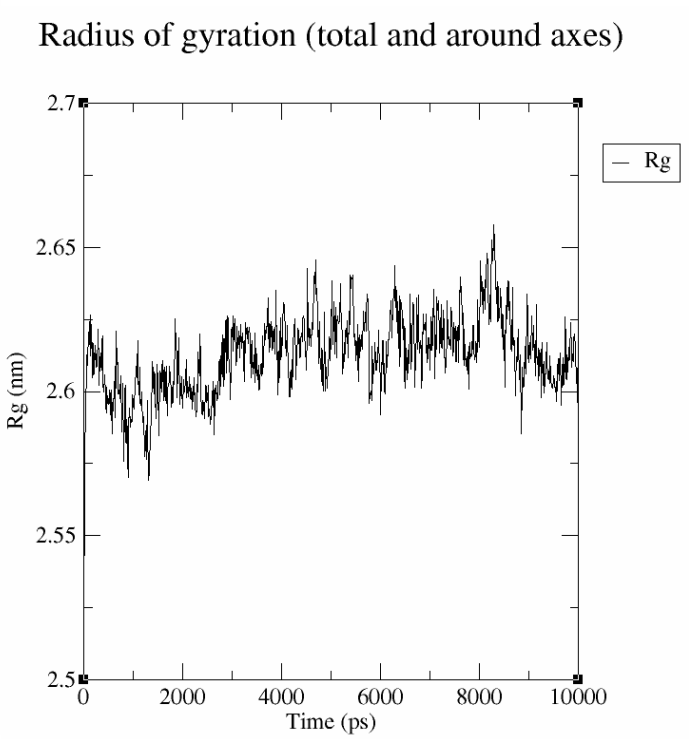

Figure 4.69: Rg graph of protein complex 10

From the observation, it can be said that protein complex 10 gave the best graph for the radius of gyration.

The best three protein-protein complexes were further evaluated using ERRAT, Verify 3D and Ramachandran distribution plot.

Table 4.8: Results of evaluation of best three protein-protein complexes selected after molecular dynamics simulation:

	ERRAT Goal: >50 %	VERIFY 3D Goal: >80%	Ramachandran score (Procheck)	Ramachandran score (Molprobity) Goal: >98%	Ramachandran Outliers (Molprobity) Goal: <0.05%	RMS Z-SCORE (Molprobity) Goal: abs(Z score) < 2
Protein complex 1	90.71%	89.7579	83.9% core 14.6% allowed 0.6% general 0.8% disallowed	89.93%	0.90%	-3.14 ± 0.31
Protein complex 2	98.21%	93.3333	83.1% core 15.1% allowed 1.0% general 0.8% disallowed	92.66%	0.35%	-3.73 ± 0.31
Protein complex 10	92.14%	90.8752	82.7% core 15.1% allowed 1.0% general 1.2% disallowed	92.66%	0.35%	-3.77 ± 0.32

Taking all the evaluating factors into account and by giving more importance to the quality of the RMSD graph, protein complex 10 has been selected for further analysis and inspection of the formation of the bond before and after MD simulation.

Views of Protein complex 10 of PSY1R kinase and SERK1 kinase on four different angles, after simulation :

PSY1R kinase has been represented in the color blue and SERK1 kinase has been represented in the color cyan.

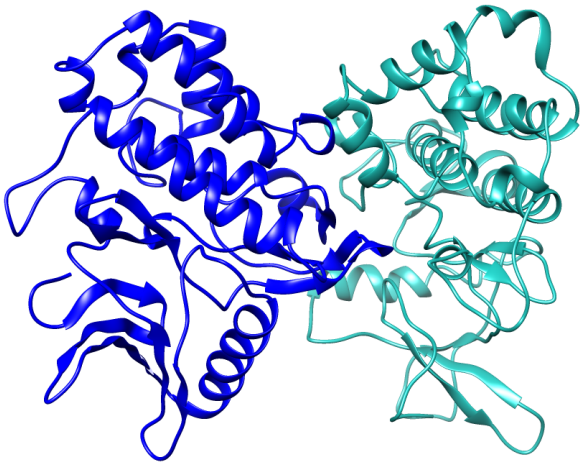

Figure 4.70: Protein complex 10 (view 1)

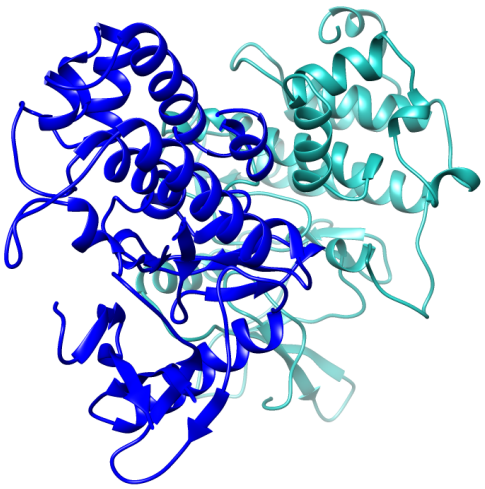

Figure 4.71 : Protein complex 10 (view 2)

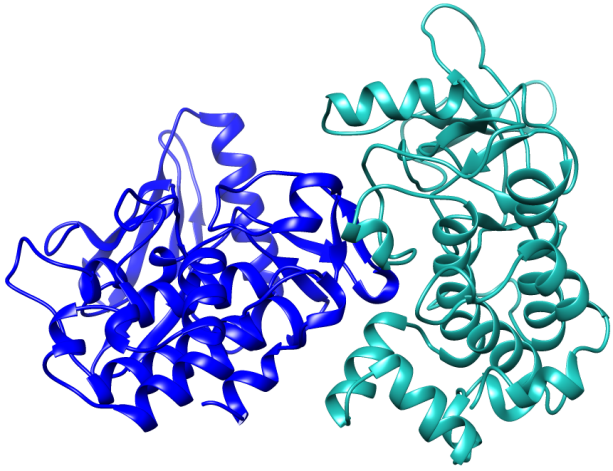

Figure 4.72: Protein complex 10 (view 3)

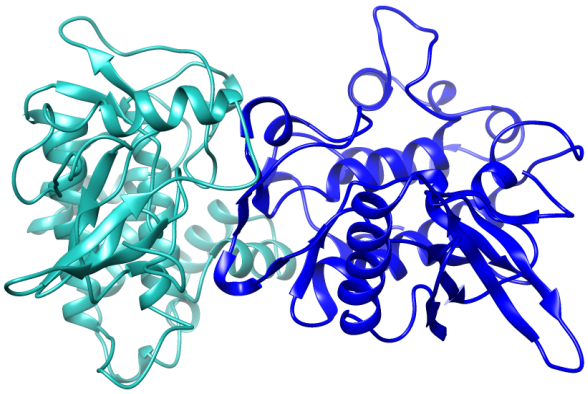

Figure 4.73: Protein complex 10 (view 4)

4.14 Results of Protein-Protein interactions :

Interactions between protein and protein (PPIs) are accountable for a variety of biological purposes within the functioning cells. These interactions are the key to unfolding the mechanism of many diseases in plants and animals. Besides that, within the protein-protein interactions, there are "hot spots", that are very well specified interface regions that are accountable for the majority of the binding energy. PPIs modulators are designed to target those “hot spots”. The structure-based designing of these modulators also requires a three-dimensional structure of the protein complex to be observed. This is just an instance of one of the most useful purposes of the protein protein docking method. Analysis of protein-protein interactions offers a wide area of purposes ranging from understanding mechanisms of different pathways to designing various tools in order to manipulate them.
Detail accounts of protein-protein interactions on different criteria are given in the following chart:

Table 4.9: Protein-protein interactions chart:

Hydrophobic Interactions within 5 Angstroms					
Before simulation					
Position	**Residue**	**Chain**	**Position**	**Residue**	**Chain**
172	PRO	A	164	VAL	B
172	PRO	A	176	LEU	B
177	ALA	A	149	LEU	B
177	ALA	A	164	VAL	B
179	VAL	A	168	ILE	B
182	LEU	A	204	PHE	B
259	PRO	A	168	ILE	B
259	PRO	A	171	ILE	B
259	PRO	A	176	LEU	B
260	MET	A	168	ILE	B
260	MET	A	207	ALA	B

260	MET	A	217	LEU	B
After simulation					
174	TYR	A	168	ILE	B
177	ALA	A	168	ILE	B
179	VAL	A	168	ILE	B
182	LEU	A	204	PHE	B
259	PRO	A	168	ILE	B
259	PRO	A	176	LEU	B
259	PRO	A	204	PHE	B
260	MET	A	203	ALA	B
260	MET	A	204	PHE	B
260	MET	A	207	ALA	B
260	MET	A	217	LEU	B

Protein-Protein disulfide Bridges
Before simulation
NO PROTEIN-PROTEIN disulfide BRIDGES FOUND
After simulation
NO PROTEIN-PROTEIN disulfide BRIDGES FOUND

Protein-Protein Main Chain-Main Chain Hydrogen Bonds
Before simulation
NO PROTEIN-PROTEIN MAIN CHAIN-MAIN CHAIN HYDROGEN BONDS FOUND
After simulation
NO PROTEIN-PROTEIN MAIN CHAIN-MAIN CHAIN HYDROGEN BONDS FOUND

Protein-Protein Main Chain-Side Chain Hydrogen Bonds

Dd-a = Distance Between Donor and Acceptor
Dh-a = Distance Between Hydrogen and Acceptor
A(d-H-N) = Angle Between Donor-H-N
A(a-O=C) = Angle Between Acceptor-O=C
MO = Multiple Occupancy
Note that angles that are undefined are written as 999.99

Before simulation

DONOR				ACCEPTOR				PARAMETERS				
POS	**CHAIN**	**RES**	**ATOM**	**POS**	**CHAIN**	**RES**	**ATOM**	**MO**	**Dd-a**	**Dh-a**	**A(d-H-N)**	**A(a-O=C)**
176	A	GLN	OE1	164	B	VAL	O	1	3.15	3.45	64.79	122.55
176	A	GLN	OE1	164	B	VAL	O	2	3.15	3.15	80.17	122.55
183	A	ARG	NH2	207	B	ALA	O	1	2.65	1.68	150.40	131.54
183	A	ARG	NH2	207	B	ALA	O	2	2.65	3.19	49.75	131.54
258	A	ASN	ND2	171	B	ILE	O	1	3.26	2.71	112.00	170.71
258	A	ASN	ND2	171	B	ILE	O	2	3.26	3.17	85.94	170.71
261	A	LYS	NZ	216	B	MET	O	-	2.69	9.99	999.99	150.19
267	A	GLN	OE1	211	B	ASN	O	1	3.36	2.91	105.09	81.37
267	A	GLN	OE1	211	B	ASN	O	2	3.36	4.21	30.59	81.37
267	A	GLN	NE2	211	B	ASN	O	1	2.71	2.11	112.95	116.33
267	A	GLN	NE2	211	B	ASN	O	2	2.71	3.27	48.82	116.33
12	B	PHE	N	160	A	THR	OG1	-	3.13	2.75	104.94	999.99
163	B	ALA	N	176	A	GLN	OE1	-	2.89	2.07	145.81	170.93
164	B	VAL	N	176	A	GLN	OE1	-	2.96	2.03	168.05	124.53
165	B	ARG	NE	176	A	GLN	O	-	3.06	2.13	162.19	146.31
165	B	ARG	NH2	177	A	ALA	O	1	3.39	2.99	103.35	89.51

165	B	ARG	NH2	177	A	ALA	O	2	3.39	3.46	77.80	89.51
169	B	GLY	N	260	A	MET	SD	-	3.22	2.68	116.94	999.99
202	B	ARG	NE	155	A	TYR	O	-	3.13	2.19	163.09	130.70
211	B	ASN	OD1	262	A	ARG	O	1	3.08	2.59	107.44	139.46
211	B	ASN	OD1	262	A	ARG	O	2	3.08	3.44	61.96	139.46
After simulation												
DONOR				**ACCEPTOR**				**PARAMETERS**				
POS	**CHAIN**	**RES**	**ATOM**	**POS**	**CHAIN**	**RES**	**ATOM**	**MO**	**Dd-a**	**Dh-a**	**A(d-H-N)**	**A(a-O=C)**
183	A	ARG	NH2	207	B	ALA	O	1	3.28	3.76	55.42	101.72
183	A	ARG	NH2	207	B	ALA	O	2	3.28	2.62	124.11	101.72
257	A	GLN	NE2	176	B	LEU	O	1	3.13	2.89	93.83	124.08
257	A	GLN	NE2	176	B	LEU	O	2	3.13	2.54	112.43	124.08
32	B	LYS	NZ	160	A	THR	O	-	3.31	9.99	999.99	138.73
202	B	ARG	NE	155	A	TYR	O	-	3.17	3.10	85.10	156.29
202	B	ARG	NH2	155	A	TYR	O	1	3.26	2.85	103.85	125.04
202	B	ARG	NH2	155	A	TYR	O	2	3.26	3.96	41.08	125.04

Protein-Protein Side Chain-Side Chain Hydrogen Bonds

Dd-a = Distance Between Donor and Acceptor
Dh-a = Distance Between Hydrogen and Acceptor
A(d-H-N) = Angle Between Donor-H-N
A(a-O=C) = Angle Between Acceptor-O=C
MO = Multiple Occupancy
Note that angles that are undefined are written as 999.99

Before simulation

DONOR				**ACCEPTOR**				**PARAMETERS**				
POS	**CHAIN**	**RES**	**ATOM**	**POS**	**CHAIN**	**RES**	**ATOM**	**MO**	**Dd-a**	**Dh-a**	**A(d-H-N)**	**A(a-O=C)**

156	A	ARG	NH1	89	B	GLU	OE1	1	2.71	3.31	47.36	999.99
156	A	ARG	NH1	89	B	GLU	OE1	2	2.71	1.81	145.83	999.99
156	A	ARG	NH1	89	B	GLU	OE2	1	2.81	3.70	27.27	999.99
156	A	ARG	NH1	89	B	GLU	OE2	2	2.81	2.00	134.96	999.99
156	A	ARG	NH2	89	B	GLU	OE2	1	2.75	3.64	28.06	999.99
156	A	ARG	NH2	89	B	GLU	OE2	2	2.75	1.94	135.77	999.99
183	A	ARG	NH2	211	B	ASN	OD1	1	2.72	3.33	46.40	999.99
183	A	ARG	NH2	211	B	ASN	OD1	2	2.72	1.71	170.83	999.99
257	A	GLN	NE2	177	B	SER	OG	1	2.89	3.58	42.07	999.99
257	A	GLN	NE2	177	B	SER	OG	2	2.89	1.88	166.65	999.99
261	A	LYS	NZ	219	B	ASP	OD2	-	2.57	9.99	999.99	999.99
264	A	ASN	ND2	212	B	ASP	OD1	1	2.89	3.54	44.44	999.99
264	A	ASN	ND2	212	B	ASP	OD1	2	2.89	2.09	132.51	999.99
264	A	ASN	ND2	212	B	ASP	OD2	1	3.00	3.67	43.59	999.99
264	A	ASN	ND2	212	B	ASP	OD2	2	3.00	2.01	161.08	999.99
267	A	GLN	OE1	211	B	ASN	ND2	1	2.75	1.68	167.60	999.99
267	A	GLN	OE1	211	B	ASN	ND2	2	2.75	3.20	55.19	999.99
267	A	GLN	NE2	214	B	ASP	OD1	1	2.91	3.24	62.70	999.99
267	A	GLN	NE2	214	B	ASP	OD1	2	2.91	2.00	146.74	999.99
271	A	TRP	NE1	214	B	ASP	OD1	-	3.04	2.37	131.12	999.99
271	A	TRP	NE1	214	B	ASP	OD2	-	2.95	2.20	139.44	999.99
177	B	SER	OG	257	A	GLN	NE2	-	2.89	9.99	999.99	999.99
208	B	ARG	NH1	117	A	GLN	OE1	1	2.78	1.81	153.48	999.99
208	B	ARG	NH1	117	A	GLN	OE1	2	2.78	3.19	57.81	999.99
208	B	ARG	NH1	122	A	HIS	NE2	1	2.89	3.72	32.48	999.99

208	B	ARG	NH1	122	A	HIS	NE2	2	2.89	1.95	153.73	999.99
208	B	ARG	NH2	122	A	HIS	NE2	1	3.21	4.14	23.43	999.99
208	B	ARG	NH2	122	A	HIS	NE2	2	3.21	2.40	136.06	999.99
211	B	ASN	ND2	267	A	GLN	OE1	1	2.75	3.35	47.07	999.99
211	B	ASN	ND2	267	A	GLN	OE1	2	2.75	2.08	120.80	999.99
212	B	ASP	OD1	264	A	ASN	ND2	1	2.89	1.97	141.32	999.99
212	B	ASP	OD1	264	A	ASN	ND2	2	2.89	3.71	34.86	999.99
212	B	ASP	OD2	264	A	ASN	ND2	1	3.00	2.14	135.32	999.99
212	B	ASP	OD2	264	A	ASN	ND2	2	3.00	3.88	30.69	999.99
214	B	ASP	OD1	267	A	GLN	NE2	1	2.91	3.08	70.68	999.99
214	B	ASP	OD1	267	A	GLN	NE2	2	2.91	2.04	135.11	999.99
After simulation												
DONOR				**ACCEPTOR**				**PARAMETERS**				
POS	**CHAIN**	**RES**	**ATOM**	**POS**	**CHAIN**	**RES**	**ATOM**	**MO**	**Dd-a**	**Dh-a**	**A(d-H-N)**	**A(a-O=C)**
156	A	ARG	NE	89	B	GLU	OE1	-	3.50	3.05	106.78	999.99
183	A	ARG	NH1	211	B	ASN	OD1	1	2.77	3.66	26.88	999.99
183	A	ARG	NH1	211	B	ASN	OD1	2	2.77	1.88	147.83	999.99
183	A	ARG	NH2	211	B	ASN	OD1	1	2.87	3.80	24.37	999.99
183	A	ARG	NH2	211	B	ASN	OD1	2	2.87	2.03	141.24	999.99
88	B	ARG	NH1	155	A	TYR	OH	1	2.99	3.53	51.64	999.99
88	B	ARG	NH1	155	A	TYR	OH	2	2.99	2.83	88.90	999.99
88	B	ARG	NH2	155	A	TYR	OH	1	3.13	3.76	46.58	999.99
88	B	ARG	NH2	155	A	TYR	OH	2	3.13	2.97	90.63	999.99
208	B	ARG	NH1	117	A	GLN	OE1	1	3.45	4.10	46.33	999.99
208	B	ARG	NH1	117	A	GLN	OE1	2	3.45	3.24	93.17	999.99

208	B	ARG	NH2	117	A	GLN	OE1	1	3.43	4.07	46.58	999.99
208	B	ARG	NH2	117	A	GLN	OE1	2	3.43	3.25	92.02	999.99

Protein-Protein Ionic Interactions within 6 Angstroms					
Before simulation					
Position	**Residue**	**Chain**	**Position**	**Residue**	**Chain**
122	HIS	A	205	ASP	B
156	ARG	A	89	GLU	B
162	GLU	A	32	LYS	B
261	LYS	A	219	ASP	B
After simulation					
Position	**Residue**	**Chain**	**Position**	**Residue**	**Chain**
156	ARG	A	89	GLU	B
158	HIS	A	128	ASP	B
Protein-Protein Aromatic- Aromatic Interactions					
Before simulation					
NO PROTEIN-PROTEIN AROMATIC-AROMATIC INTERACTIONS FOUND					
After simulation					
NO PROTEIN-PROTEIN AROMATIC-AROMATIC INTERACTIONS FOUND					

Protein-Protein Aromatic-Sulphur Interactions within 5.3 Angstroms
Before simulation
NO PROTEIN-PROTEIN AROMATIC-SULPHUR INTERACTIONS FOUND
After simulation

Position	Residue	Chain	Position	Residue	Chain	D(Centroid-Sulphur)	Angle
204	PHE	B	260	MET	A	4.27	148.56
Protein-Protein Cation-Pi Interactions within 6 Angstroms							
Before simulation							
Position	**Residue**	**Chain**	**Position**	**Residue**	**Chain**	**D(cation-Pi)**	**Angle**
204	PHE	B	183	ARG	A	5.95	31.38
After simulation							
Position	**Residue**	**Chain**	**Position**	**Residue**	**Chain**	**D(cation-Pi)**	**Angle**
155	TYR	A	202	ARG	B	5.96	68.20
155	TYR	A	88	ARG	B	4.57	47.83
178	TRP	A	32	LYS	B	5.96	127.36

Table 4.10: Summary of the whole protein-protein interactions :

Interaction types:	Before Simulation	After Simulation
Hydrophobic Interactions within 5 Angstroms	12	11
Protein-Protein disulfide Bridges	0	0
Protein-Protein Main Chain-Main Chain Hydrogen Bonds	0	0
Protein-Protein Main Chain-Side Chain Hydrogen Bonds	21	8
Protein-Protein Side Chain-Side Chain Hydrogen Bonds	36	13
Protein-Protein Ionic Interactions	4	2
Protein-Protein Aromatic- Aromatic Interactions	0	0
Protein-Protein Aromatic-Sulphur Interactions	0	1
Protein-Protein Cation-Pi Interactions	1	3

4.15 Inspections and observations of some of the protein-protein interactions after MD simulation:

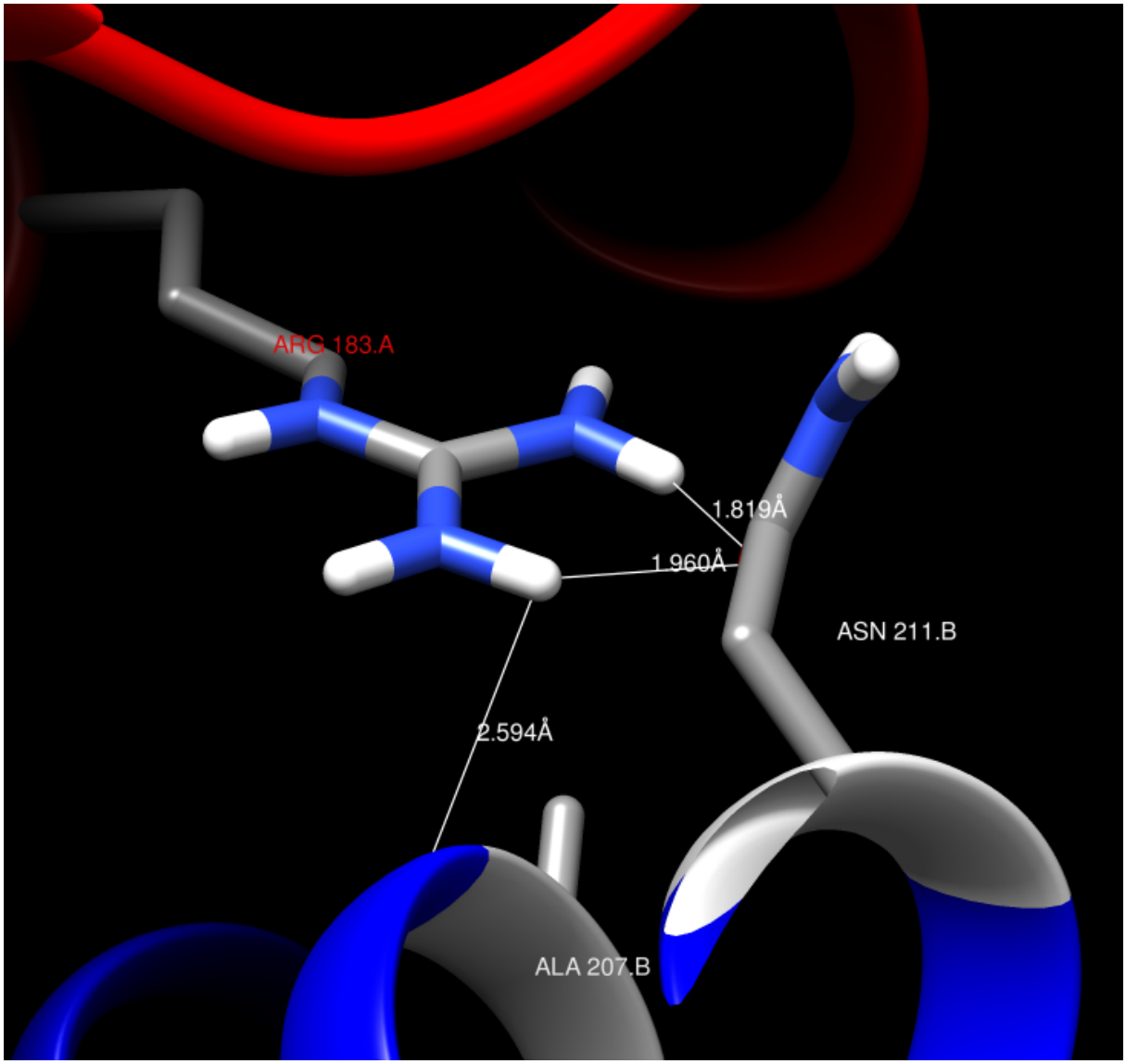

Figure 4.74: H Bonds between ARG 183.A and ASN 211.B and between ARG 183.A and ALA 207.B (Zoom view)
H bonds between ARG 183.A 2HH1 - ASN 211.B OD1= distance 1.819A
H bonds between ARG 183.A 2HH2 - ASN 211.B OD1= distance 1.960A
These two are protein protein side chain-side chain H bonds
H bonds between ARG 183.A 2HH1 - ALA 207.B O = distance 2.594A
This one is the protein protein main chain - side chain H bond.

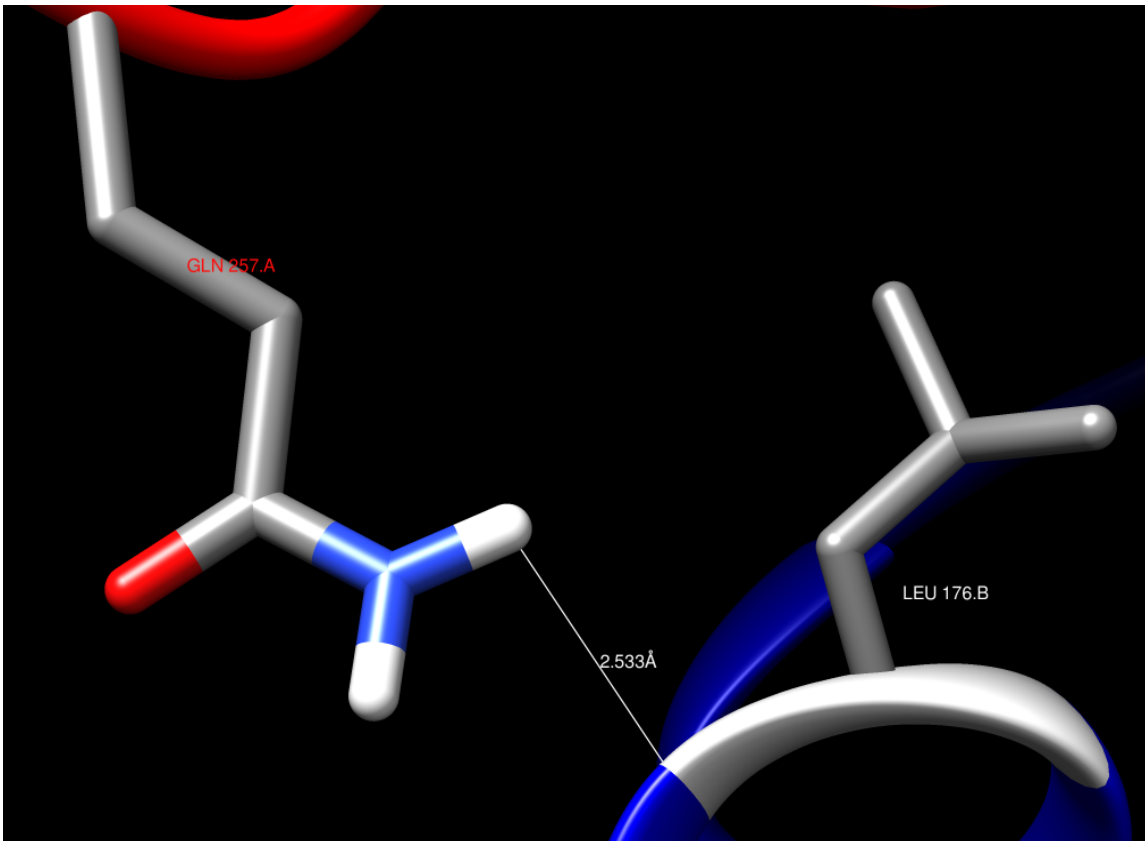

Figure 4.75: H bond between GLN 257.A and LEU 176.B
H bond between GLN 257.A 1HE2 and LEU 176.B O = distance 2.533A

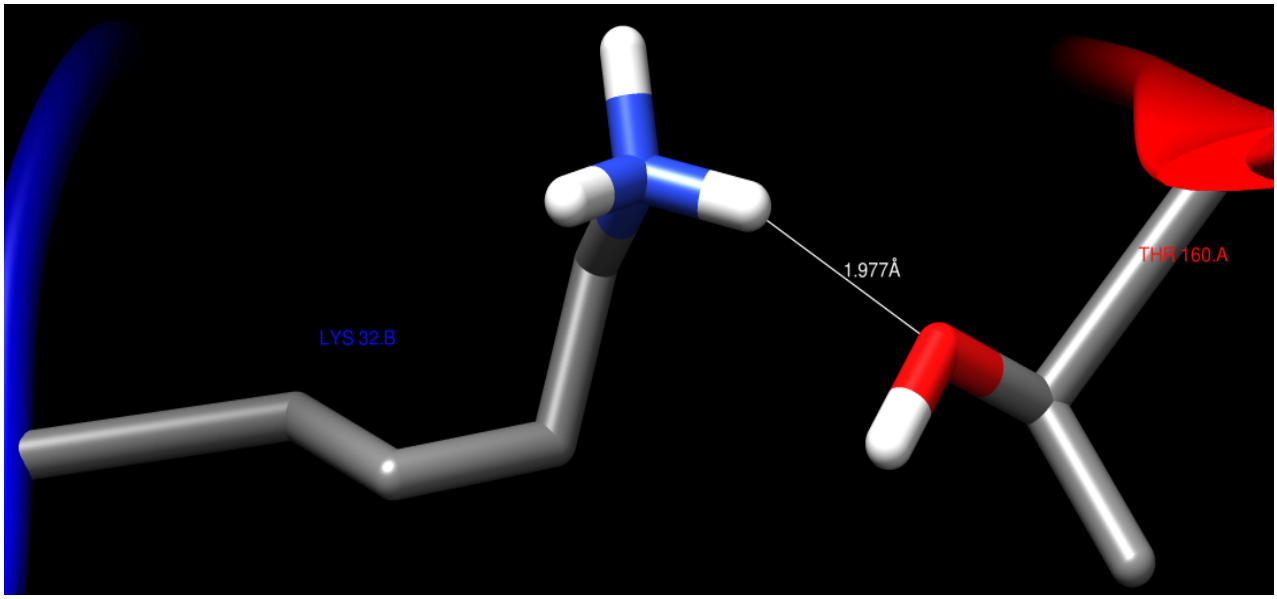

Figure 4.76: H bond between LYS 32.B and THR 160.A
H bond between LYS 32.B HZ3 and THR 160.A OG1 = distance 1.977A

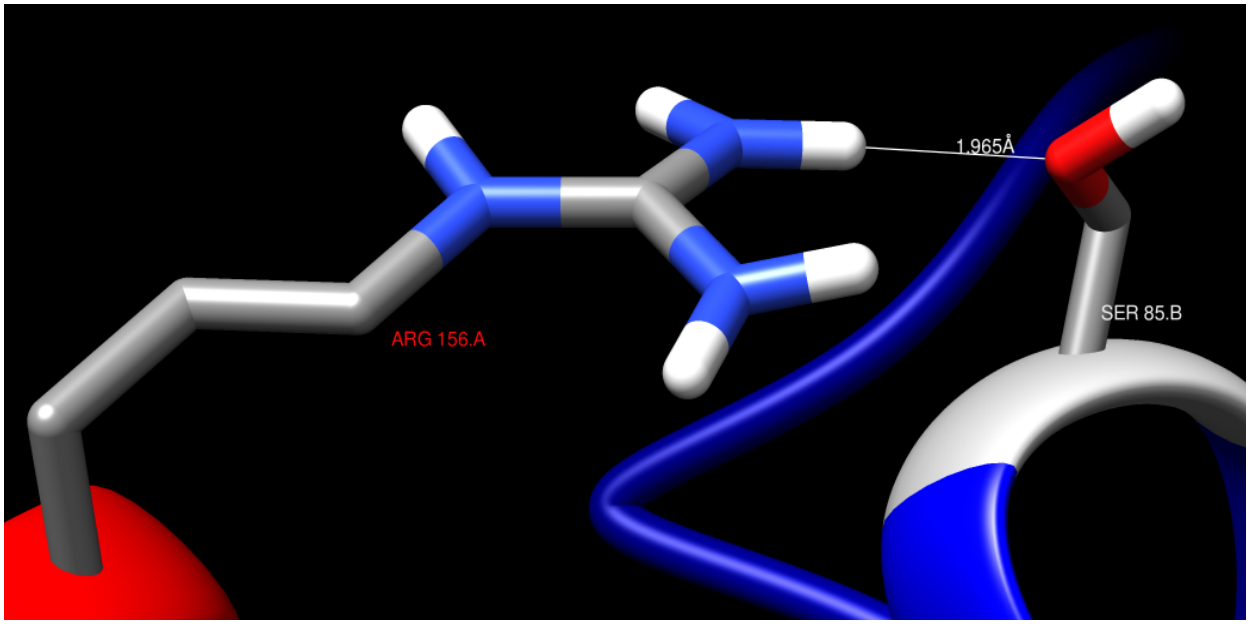

Figure 4.77: H bond between ARG 156.A and SER 85.B
H bond between ARG 156.A 2HH2 - SER 85.B OG1 = distance 1.965A

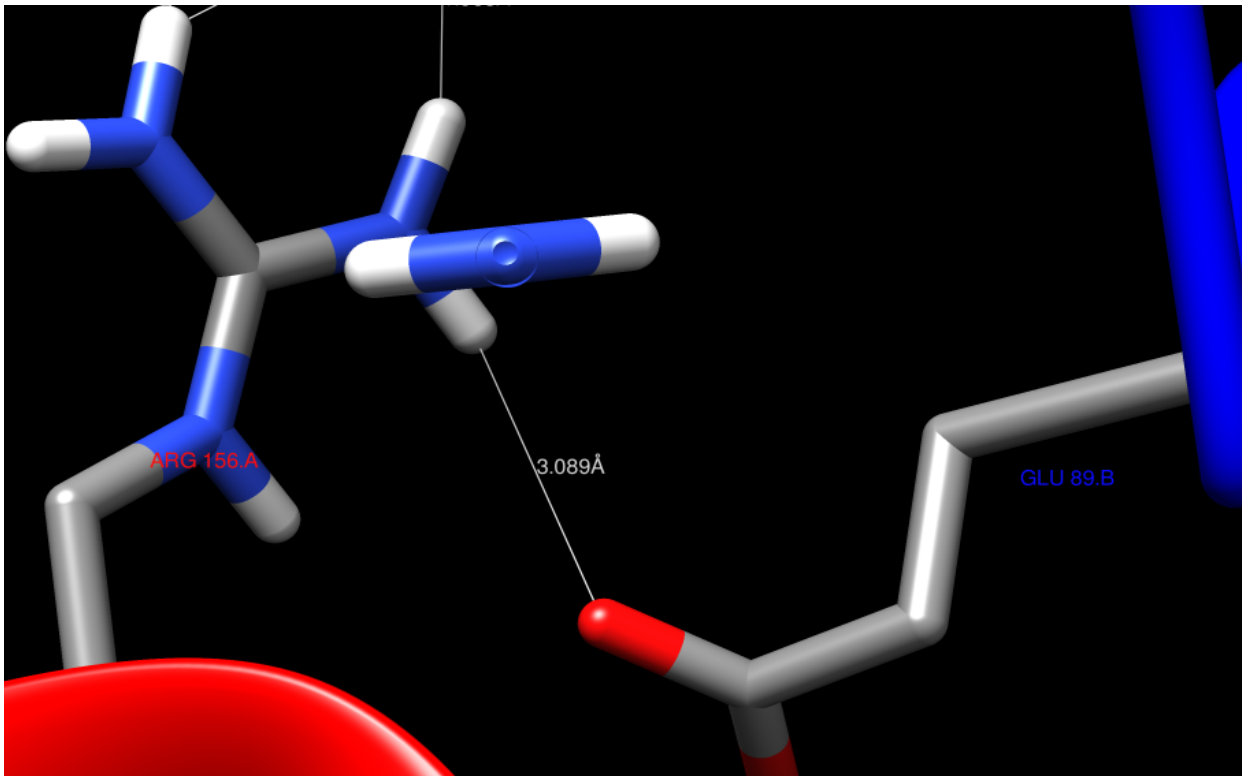

Figure 4.78: H bond between ARG 156.A and GLU 89.B

H bond ARG 156.A 1HH2 - GLU 89.B OE1 = distance 3.089A
Both of these bonds protein protein side-chain -side chain hydrogen bonds.

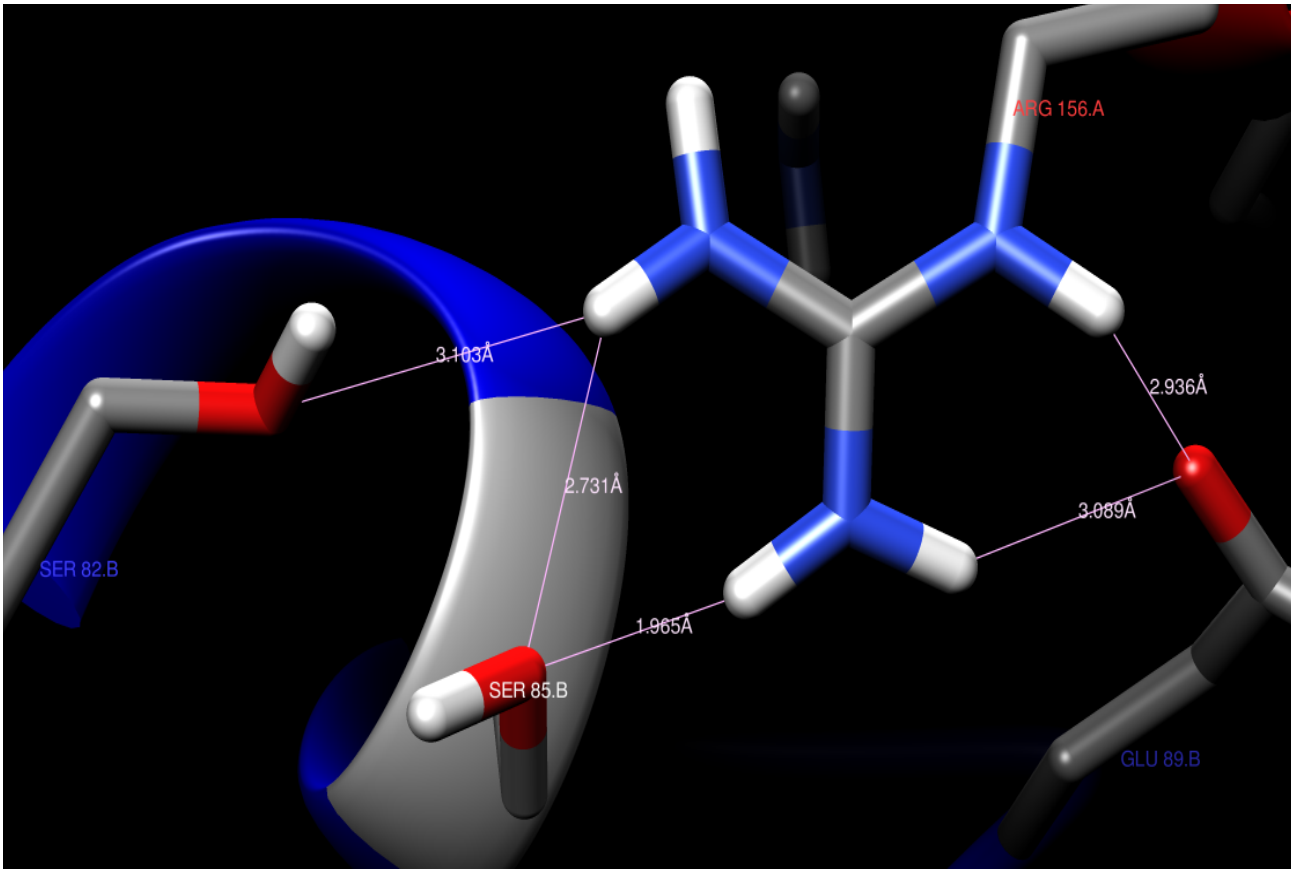

Figure 4.79: H bond between ARG 156.A with GLU 89.B , SER 85.B and SER 82.B

H bond between ARG 156.A HE - GLU 89.B OE1 = distance 2.936A
H bond between ARG 156.A 1HH2 - GLU 89.B OE1 = distance 3.089A
H bond between ARG 156.A 2HH2 - SER 85.B OG = distance 1.965A
H bond between ARG 156.A 2HH1 - SER 85.B OG = distance 2.731A
H bond between ARG 156.A 2HH1- SER 82.B OG = distance 3.103A

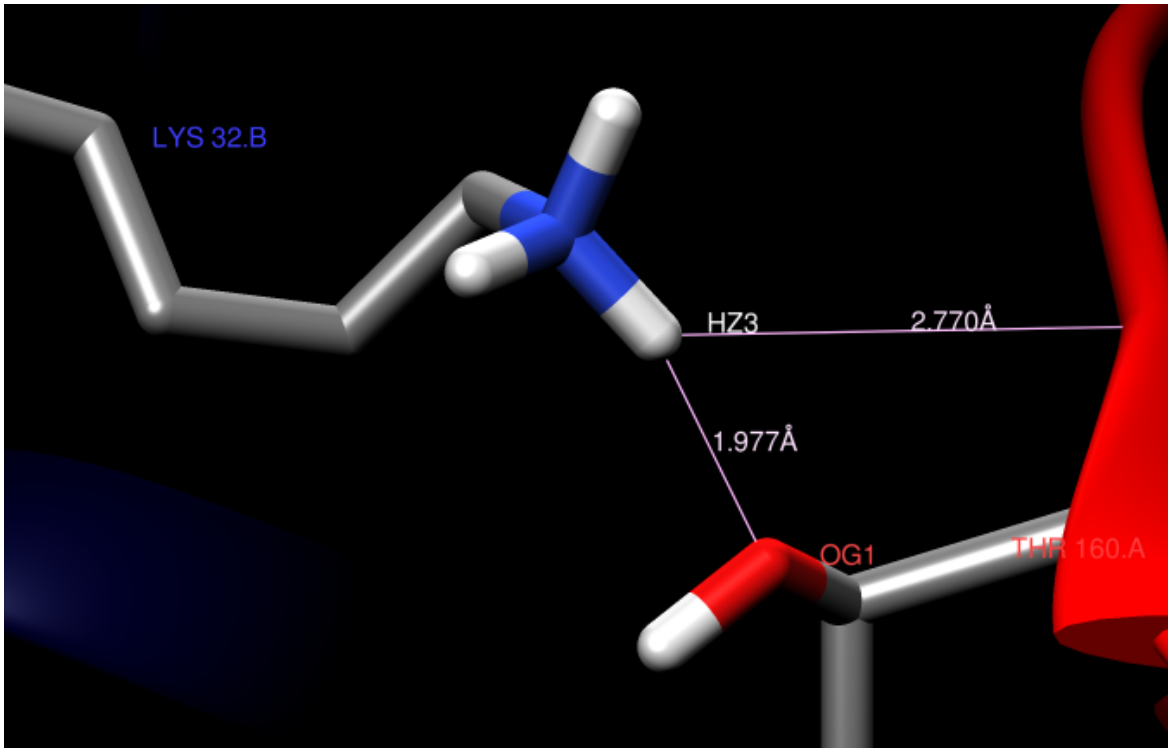

Figure 4.80: H bonds between THR 160.A and LYS 32.B

H bond between LYS 32.B HZ3 and THR 160.A OG1 = distance 1.977A
This is protein protein side chain-side chain interaction
H bond between LYS 32.B HZ3 and THR 160.A O = distance 2.770A
This is protein protein side chain- main chain interaction

Discussion:

After analyzing the protein-protein interactions, a number of hydrogen bonds are found to be evidently present. Hydrogen bonding is a polar bond that is formed by interactions between two polar components, one of them being hydrogen and the other being a highly electronegative element. Salt bridges are relatively lower in number at the protein-protein interaction interface. Among the residues having a distance of not more than 4 Å which are oppositely charged, salt bridges are formed. In the observed figure of the protein-protein complex, no salt bridges were found.

Disulfide bonds have a low affinity for interface regions and are hardly found at the protein-protein interaction interface, especially in the case of intracellular protein and homo complexes. Meanwhile, for oligomerization of permanent complexes from their monomers hydrophobic interactions mostly dominate the interactions. Also, by comparing taxonomically it has been revealed that the protein-protein interfaces are polar to a great extent which are more likely to exploit electrostatic interactions or Van der Waal force of attractive or repulsive interaction. Hence, the aromatics use π electron clouds to interact with the cations of their partners' side chains, apart from the salt bridges.

A central role is played by hydrogen bonds and salt bridges while protein binding. The protein surfaces that basically comprise the binding interfaces are more hydrophilic in nature being covered mostly with hydrophilic residues whereas the protein interior is more hydrophobic. Besides, from the studies of Warshel and Russell in 1984 it can be said that electrostatic complementarity to the charge distribution of the binding substrates is often provided by the protein active sites (Warshel and Russell, 1984; Cherfilset al.,1991; Novotny and Sharp, 1992; Creighton, 1993). Binding interfaces have been found to form more salt bridges and hydrogen bonds in comparison to the protein interiors. Apart from their density differences in these two places, hydrogen bonds and salt bridges on the binding surface and within the interior are also found to contribute relatively differently in energetics i.e, the study of energy or thermodynamics.

In some early studies, it has been shown that (Xuet al., 1997), electrostatic interactions participate more in binding than in folding. Hence the interfacial H bonds and salt bridges playing the dominant role in the electrostatic interactions between proteins carries out a significant role in binding more than the intra monomer hydrogen bonds and infolding salt bridges. Again, protein-protein binding is often mediated with the help of bound water molecules unlike the intramolecular bond (Creighton, 1993; Bhatet al., 1994). Ordered water molecules bridge the network of the hydrogen bond between the proteins and consequently help the complexes to be more stabilized (Bhatet al., 1994; Helms and Wade, 1995). Thus, this difference represents that even though the types of interactions in both of the processes are similar their relative role is relatively different based on being interfacial and within the protein.

In spite of hydrophobic interactions having a critical role in binding, they cannot dominate the way they dominate in folding (Tsaiet al., 1997). Besides, a large variability has been observed among the interfaces that are able to be rationalized quite straightforwardly. This is because the chains have first folded to a configuration having the most stability with its core being buried with a large proportion of hydrophobic residues. The compactly packed hydrophobic residues inside push the hydrophilic residues to the surface hence limiting the probable hydrophobic effect on the interface.

Being responsible for a greater proportion of polar charged surface residues, hydrogen bonds have much importance in the scope of stabilizing the interacting molecules. But again a significant role is played by the geometry considering constraints created by bond lengths and angles in order to ensure their overall quality. For the determination of binding specificity, hydrogen bonds and salt bridges have much significance (Fersht, 1984; Honig and Yang, 1995).

In the pattern recognition process specified protein-protein interactions play an important role. Complementarity in geometry and stability in energetics are the two prime reasons for binding to occur. Hydrophobic effect, hydrogen bonds and salt bridges have much to contribute to the processes of energetics. A hydrogen bond or a salt bridge is very much capable of contributing free energy that is favorable to the binding. But a hydrogen bond that has not been fulfilled or a charge that has been isolated can considerably destabilize the binding due to the desolvation effect when it is in the protein interface (Bartlett and Marlowe, 1984; Gaoet al., 1996; Xuet al., 1997). Due to these reasons the hydrogen bonds and salt bridges are highly selective in interacting between the proteins that eventually confer binding specificity. While the hydrogen bonds and salt bridges have dominance in binding between the protein domains, the hydrophobic effects dominate the stabilization of the protein complex structure.

For most of the part, arginine is notably one of the most prevailing residues found in the protein-protein interfaces. Its significance complements the adaptability of its side chain during molecular interactions. The high participation of arginine in the interfaces can be traced back to its higher capability to donate hydrogen bonds. Additionally, it has been found that the participation of arginine in cation interactions is frequent on protein complex interfaces that eventually contribute to the binding energy to form the complex. Therefore cation interactions are also included in the criteria of characterizing interactions on protein interfaces here. The fact that cation–pi interactions have much significance in protein-protein interfaces is beneficial for protein docking studies to a great extent. The presence of Cation–interactions can be used as a measure to separate the docking results that are chemically relevant from its false positives. The number of cation-pi interactions has been increased about three times after the simulation which can be considered as a positive change.(Crowley, Peter and Adel, 2005)

Disulfide bond formation invokes a reaction among two cysteine residues' the sulfhydryl (SH) side chains. Sulphide anion of one sulfhydryl group of a cysteine residue acting as a nucleophile

attacks the side chain of another cysteine for creating a disulfide bond. In this process, electrons are released for transfer. Proper disulfide bonds decrease entropic choices further that in turn ease the folding progression to the native state by restraining the misfolded conformations. Thus, eventually contributing to the protein's stability. The same principle is reflected in the case of protein docked complexes too. The rise in the native structures' stability which resulted due to the formation of a certain disulfide bond is found to be directly proportional to the number of residues in between the cysteines involved. This means the higher the number of residues in the loops of disulfide the more support and firmness is imparted to the native structure (Gautam and Peter, 2013). However, there are no disulfide interactions or bridges observed in our protein-protein complex under introspection before and after the simulation.

Hydrophobic residues on the interfaces are critical for stabilizing the protein-protein docked complexes (Keskin, O., Gursoy, A., 2008). The process of protein-protein complex formation in an aqueous medium is found to be entropy-driven (Yan C, Wu F, 2008). The reason behind this is the observation that the burial of hydrophobic surface patches generates a high gain in entropy and hence provides a driving force to form stable protein complexes. The interfaces being more conserved are more significant for maintaining the protein-protein interactions through evolution (Yan C, Wu F, 2008).

Meanwhile, intense hydrophobic interactions can be formed among the large hydrophobic side chains of the aromatic residues. Moreover, further contributions can be made by the positioning of two aromatic rings parallelly that form a compact pack with finer fit in geometry. (Keskin, O., Gursoy, A., 2008) In our observation in this study, quite a number of hydrophobic interactions are found in both before and after simulation. From this, it can be claimed that the structure before and after the simulation has been quite stabilized.

Conclusions and Perspectives:

PSKα and PSY1 signaling that is mediated by the respective receptors PSKR1 and PSY1R employ in shifting the hormone homeostasis in favor of the JA pathway whereas negatively regulate SA accumulation and signaling (Mosher, S., & Kemmerling, B., 2013). The fact that PSKα induced growth promotion is encouraged by auxin adds another degree of complexity (Mosher, S., & Kemmerling, B., 2013). The involvement of PSKα and PSY1 signaling in growth development and defense have the possibility to include crosstalk between different phytohormones including auxin. Moreover, auxin has been found to induce a suppressive effect on SA responses and vice versa (Mosher, S., & Kemmerling, B., 2013). Decrease amount of crown gall sizes instigated by *Agrobacterium* infection in the mutants of *pskr1* has been reported by Loivamaki, et al. This points and invokes the fact that auxin plays a significant part in *Agrobacterium/Arabidopsis* interaction and alongside manipulates the senescence negatively. Since every aspect of PSKα/PSY1-signaling can be comprehended by different actions of auxin, there is a possibility for auxin might be the central regulator that balance out the defense responses in order to encourage the growth of the plant and prevent senescence (Mosher, S., & Kemmerling, B., 2013).

For the past few years, several researches have shown and claimed that sulfated peptides are quite significant in signaling employed by plants to encourage growth and development processes along with controlling the stress responses that include PAMP responses too. As plants are continuously undergoing stresses and different environmental and growth problems it is very important for plants to respond to these stress signals at the right time with efficacy. But unfortunately, enforcement of stress responses at times comes at the expense of reduction or compromising the growth in plants. Inappropriate regulation and extended incitement of these responses can lead to reduced or halted growth in plants, even resulting in the death of the cells.

In this research study, PSY1R, one of the most important receptors which used to be known for its significance for plant growth and regulation has been taken under inspection for its newly found effects on the plant defense system. The kinase domain, which is taken as the catalytic domain and is responsible for major function, has been examined for developing a better understanding on how this receptor works. For having a clear view on the whole story regarding PSY1R receptor signaling and its consequences in a complex network, more research and studies on hormone pathways in PSKα/PSY1 signaling are needed.

REFERENCES:

[1] Mosher, S., Seybold, H., Rodriguez, P., Stahl, M., Davies, K. A & Kemmerling, B. (2013). The tyrosine‐sulfated peptide receptors PSKR1 and PSY1R modify the immunity of Arabidopsis to biotrophic and necrotrophic pathogens in an antagonistic manner. *The Plant Journal*, *73*(3), 469-482.

[2] Mosher, S., & Kemmerling, B. (2013). PSKR1 and PSY1R-mediated regulation of plant defense responses. *Plant Signaling & Behavior*, *8*(5), e24119.

[3] Oehlenschlæger, C. B., Gersby, L., Ahsan, N., Pedersen, J. T., Kristensen, A., Solakova, T. V. & Fuglsang, A. T. (2017). Activation of the LRR receptor-like kinase PSY1R requires transphosphorylation of residues in the activation loop. *Frontiers in plant science*, *8*, 2005.

[4] Mahmood, K., Kannangara, R., Jørgensen, K., & Fuglsang, A. T. (2014). Analysis of peptide PSY1 responding transcripts in the two Arabidopsis plant lines: wild type and psy1r receptor mutant. *BMC genomics*, *15*(1), 1-12.

[5] Mosher, S. (2013). *The tyrosine-sulfated peptide receptors PSKR1 and PSY1R modulate Arabidopsis immune responses* (Doctoral dissertation, Universitätsbibliothek Tübingen).

[6] Tost, A. S., Kristensen, A., Olsen, L. I., Axelsen, K. B., & Fuglsang, A. T. (2021). The PSY Peptide Family—Expression, Modification and Physiological Implications. *Genes*, *12*(2), 218.

[7] Gou, X., & Li, J. (2020). Paired receptor and coreceptor kinases perceive extracellular signals to control plant development. *Plant physiology*, *182*(4), 1667-1681.

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

[9] Jones, J. & Dangl, J. (2006). The plant immune system. Nature, 444(7117), 323-329.

[10] Bigeard, J., Colcombet, J. and Hirt, H. (2015). Signaling Mechanisms in Pattern-Triggered Immunity (PTI). Molecular Plant, 8(4), 521-539.

[11] Ranf, S. (2017). Sensing of molecular patterns through cell surface immune receptors. Current Opinion in Plant Biology, 38, 68-77.

[12] 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.

[13] Saijo, Y., Loo, E. and Yasuda, S. (2018). Pattern recognition receptors and signaling in plant microbe interactions. The Plant Journal, 93(4), 592-613.

[14] Choi, H. & Klessig, D. (2016). DAMPs, MAMPs, and NAMPs in plant innate immunity. BMC Plant Biology, 16(1).

[15] 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.

[16] 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.

[17] National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – [cited 2021 Apr 06]. Available from: NCBI website

[18] Tufchi, M., & Kumar, N. Recent Advances in Protein Bioinformatics. *Crop Improvement*, 53-61.

[19] Sussman, J. L., Lin, D., Jiang, J., Manning, N. O., Prilusky, J., Ritter, O., & Abola, E. E. (1998). Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta Crystallographica Section D: Biological Crystallography*, *54*(6), 1078-1084.

[20] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool J Mol Biol 215: 403–410.

[21] Dayhoff, M. O. (1972). A model of evolutionary change in proteins. *Atlas of protein sequence and structure*, *5*, 89-99.

[22] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, Philip E. Bourne, The Protein Data Bank, *Nucleic Acids Research*, Volume 28, Issue 1, 1 January 2000, Pages 235–242.

[23] UniProt Consortium. (2015). UniProt: a hub for protein information. *Nucleic acids research*, *43*(D1), D204-D212.

[24] Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics*, *5*(1), 1-19.

[25] 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.

[26] Garg, V. K., Avashthi, H., Tiwari, A., Jain, P. A., Ramkete, P. W., Kayastha, A. M., & Singh, V. K. (2016). MFPPI–multi FASTA ProtParam interface. *Bioinformation*, *12*(2), 74.

[27] Li, W., Cowley, A., Uludag, M., Gur, T., McWilliam, H., Squizzato, S & Lopez, R. (2015). The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic acids research*, *43*(W1), W580-W584.

[28] Glaser, F., Pupko, T., Paz, I., Bell, R. E., 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.

[29] McGuffin, L. J., Bryson, K., & Jones, D. T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics*, *16*(4), 404-405.

[30] Wu, S., & Zhang, Y. (2007). LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic acids research*, *35*(10), 3375-3382.

[31] Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., & Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction. *Nature methods*, *12*(1), 7-8.

[32] Zheng, W., Wuyun, Q., Li, Y., Mortuza, S. M., Zhang, C., Pearce, R. & Zhang, Y. (2019). Detecting distant-homology protein structures by aligning deep neural-network based contact maps. *PLoS computational biology*, *15*(10), e1007411.

[33] Xu, D., Jaroszewski, L., Li, Z., & Godzik, A. (2014). AIDA: ab initio domain assembly server. *Nucleic acids research*, *42*(W1), W308-W313.

[34] Eswar, N., Eramian, D., Webb, B., Shen, M. Y., & Sali, A. (2008). Protein structure modeling with MODELLER. In *Structural proteomics* (pp. 145-159). Humana Press.

[35] Söding, J., Biegert, A., & Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic acids research*, *33*(suppl_2), W244-W248.

[36] Hildebrand, A., Remmert, M., Biegert, A., & Söding, J. (2009). Fast and accurate automatic structure prediction with HHpred. *Proteins: Structure, Function, and Bioinformatics*, *77*(S9), 128-132.

[37] Fernandez-Fuentes, N., Madrid-Aliste, C. J., Rai, B. K., Fajardo, J. E., & Fiser, A. (2007). M4T: a comparative protein structure modeling server. *Nucleic acids research*, *35*(suppl_2), W363-W368.

[38] McGuffin, L. J., Atkins, J. D., Salehe, B. R., Shuid, A. N., & Roche, D. B. (2015). IntFOLD: an integrated server for modelling protein structures and functions from amino acid sequences. *Nucleic acids research*, *43*(W1), W169-W173.

[39] Kim, D. E., Chivian, D., & Baker, D. (2004). Protein structure prediction and analysis using the Robetta server. *Nucleic acids research*, *32*(suppl_2), W526-W531.

[40] Jaroszewski, L., Rychlewski, L., Li, Z., Li, W., & Godzik, A. (2005). FFAS03: a server for profile–profile sequence alignments. *Nucleic acids research*, *33*(suppl_2), W284-W288.

[41] Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., & Sternberg, M. J. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nature protocols*, *10*(6), 845-858.

[42] Hatherley, R., Brown, D. K., Glenister, M., & Tastan Bishop, Ö. (2016). PRIMO: an interactive homology modeling pipeline. *PLoS One*, *11*(11), e0166698.

[43] Waterhog Eisenberg, D., Lüthy, R., & Bowie, J. U. (1997).VERIFY3D: assessment of protein models with three-dimensional profiles. *Methods in enzymology*, *277*, 396-404.

[44] A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R. & Schwede, T. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic acids research*, *46*(W1), W296-W303.

[45] Chen, CC., Hwang, JK. & Yang, JM. $(PS)^2$-v2: template-based protein structure prediction server. *BMC Bioinformatics* **10,** 366 (2009).

[46] 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.

[47] Lengths, M., & Angles, M. (2018). Limitations of structure evaluation tools errat. *Quick Guideline Comput Drug Des*, *16*, 75.

[48] Laskowski, R. A., MacArthur, M. W., Moss, D. S., & Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of applied crystallography*, *26*(2), 283-291.

[49] Laskowski, R. A., MacArthur, M. W., & Thornton, J. M. (2006). PROCHECK: validation of protein-structure coordinates.

[50] Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J. & Richardson, D. C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography*, *66*(1), 12-21

[51] Williams, C. J., Headd, J. J., Moriarty, N. W., Prisant, M. G., Videau, L. L., Deis, L. N. & Richardson, D. C. (2018). MolProbity: More and better reference data for improved all‐atom structure validation. *Protein Science*, *27*(1), 293-315.

[52] Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., & Berendsen, H. J. (2005). GROMACS: fast, flexible, and free. *Journal of computational chemistry*, *26*(16), 1701-1718.

[53] Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., & Lindahl, E. (2015). GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, *1*, 19-25.

[54] Berendsen, H. J., van der Spoel, D., & van Drunen, R. (1995). GROMACS: a message-passing parallel molecular dynamics implementation. *Computer physics communications*, *91*(1-3), 43-56.

[55] Lindahl, E., Hess, B., & Van Der Spoel, D. (2001). GROMACS 3.0: a package for molecular simulation and trajectory analysis. *Molecular modeling annual*, *7*(8), 306-317.

[56] Cowan, R., & Grosdidier, G. (2000, February). Visualization tools for monitoring and evaluation of distributed computing systems. In *Proc. of the International Conference on Computing in High Energy and Nuclear Physics, Padova, Italy*.

[57] Robbins, A., Hannah, E., & Lamb, L. (2008). *Learning the vi and Vim Editors*. " O'Reilly Media, Inc.".

[58] Kozakov, D., Hall, D. R., Xia, B., Porter, K. A., Padhorny, D., Yueh, C., ... & Vajda, S. (2017). The ClusPro web server for protein–protein docking. *Nature protocols*, *12*(2), 255.

[59] Comeau, S. R., Gatchell, D. W., Vajda, S., & Camacho, C. J. (2004). ClusPro: a fully automated algorithm for protein–protein docking. *Nucleic acids research*, *32*(suppl_2), W96-W99.

[60] Comeau, S. R., Gatchell, D. W., Vajda, S., & Camacho, C. J. (2004). ClusPro: an automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics*, *20*(1), 45-50.

[61] Tina, K. G., Bhadra, R., & Srinivasan, N. (2007). PIC: protein interactions calculator. *Nucleic acids research*, *35*(suppl_2), W473-W476.

[62] Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD: visual molecular dynamics. *Journal of molecular graphics*, *14*(1), 33-38.

[63] DeLano, W. L. (2002). Pymol: An open-source molecular graphics tool. *CCP4 Newsletter on protein crystallography*, *40*(1), 82-92.

[64] Seeliger, D., & de Groot, B. L. (2010). Ligand docking and binding site analysis with PyMOL and Autodock/Vina. *Journal of computer-aided molecular design*, *24*(5), 417-422.

[65] Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry*, *25*(13), 1605-1612.

[66] Hansson, T., Oostenbrink, C., & van Gunsteren, W. (2002). Molecular dynamics simulations. *Current opinion in structural biology*, *12*(2), 190-196.

[67] Karplus, M., & McCammon, J. A. (2002). Molecular dynamics simulations of biomolecules. *Nature structural biology*, *9*(9), 646-652.

[68] Durrant, J. D., & McCammon, J. A. (2011). Molecular dynamics simulations and drug discovery. *BMC biology*, *9*(1), 1-9.

[69] Michaud‐Agrawal, N., Denning, E. J., Woolf, T. B., & Beckstein, O. (2011). MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *Journal of computational chemistry*, *32*(10), 2319-2327.

[70] Kaushik, A. C., Kumar, A., Bharadwaj, S., Chaudhary, R., & Sahi, S. (2018). *Bioinformatics techniques for drug discovery: applications for complex diseases*. Springer International Publishing.

[71] Rao, V. S., Srinivas, K., Sujini, G. N., & Kumar, G. N. (2014). Protein-protein interaction detection: methods and analysis. *International journal of proteomics*, *2014*.

[72] Wodak, S. J., & Janin, J. (1978). Computer analysis of protein-protein interaction. *Journal of molecular biology*, *124*(2), 323-342.

[73] Crowley, P. B., & Golovin, A. (2005). Cation-pi interactions in protein-protein interfaces. *Proteins*, *59*(2), 231–239.

[74] Shen MY, Sali A. Statistical potential for assessment and prediction of protein structures. *Protein Sci*. 2006;15(11):2507-2524.

[75] Yan C, Wu F, Jernigan RL, Dobbs D, Honavar V. Characterization of protein-protein interfaces. *Protein J*. 2008;27(1):59-70.

[76] Keskin, O., Gursoy, A., Ma, B., & Nussinov, R. (2008). Principles of protein-protein interactions: what are the preferred ways for proteins to interact?. *Chemical reviews*, *108*(4), 1225–1244.

[77] Sobolev, Oleg & Afonine, Pavel & Moriarty, Nigel & Hekkelman, Maarten & Joosten, Robbie & Perrakis, Anastassis & Adams, Paul. (2020). A global Ramachandran score identifies protein structures with unlikely stereochemistry. 10.1101/2020.03.26.010587.

[78] Zhang, L., & Skolnick, J. (1998). What should the Z-score of native protein structures be?. *Protein science : a publication of the Protein Society*, *7*(5), 1201–1207.

[79] Ramachandran, S., Kota, P., Ding, F., & Dokholyan, N. V. (2011). Automated minimization of steric clashes in protein structures. *Proteins*, *79*(1), 261–270.

[80] Chou, P. Y., & Fasman, G. D. (1973). Structural and functional role of leucine residues in proteins. *Journal of molecular biology*, *74*(3), 263–281.

[81] Chatzou, M., Magis, C., Chang, J. M., Kemena, C., Bussotti, G., Erb, I., & Notredame, C. (2016). Multiple sequence alignment modeling: methods and applications. *Briefings in bioinformatics*, *17*(6), 1009-1023.

[82] Comeau SR, Gatchell DW, Vajda S, Camacho CJ. ClusPro: a fully automated algorithm for protein-protein docking. *Nucleic Acids Res*. 2004;32(Web Server issue):W96-W99.

[83] Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., … Yeats, C. (2009). InterPro: the integrative protein signature database. *Nucleic acids research*, *37*(Database issue), D211–D215.