

***IN SILICO* STRUCTURAL ANALYSIS,
PHYSICOCHEMICAL CHARACTERIZATION AND
HOMOLOGY MODELING OF *Arabidopsis Thaliana*
NA⁺/H⁺ EXCHANGER 1 (AtNHX1) PROTEIN**



Inspiring Excellence

B.S. THESIS

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

Submitted by

Mohammad Rafid Feisal

Student ID: 11236003

Biotechnology Program

Department of Mathematics and Natural Sciences

BRAC University

Bangladesh

September 2015

In memory of Alice A. Islam, my English teacher, mentor and friend. I miss you.

“Remember, remember, this is now, and now, and now. Live it, feel it, cling to it. I want to become acutely aware of all I’ve taken for granted.” – Sylvia Plath

Declaration

I hereby solemnly declare that the research work embodying the results reported in this thesis entitled “*In silico* structural analysis, physicochemical characterization and homology modeling of *Arabidopsis thaliana* Na⁺/H⁺ exchanger 1 protein” submitted by the undersigned has been carried out under the supervision of **Dr. Aparna Islam**, Associate Professor, Biotechnology Program, Department of Mathematics and Natural Sciences, BRAC University, Dhaka. It is further declared that the research work presented here is original and any part of this thesis has not been submitted to any other institution for any degree or diploma.

Candidate:

Mohammad Rafid Feisal

Certified:

Dr. Aparna Islam

Supervisor

Associate Professor

Biotechnology Program

Department of Mathematics and Natural Sciences

BRAC University, Dhaka

Acknowledgement

Foremost, I would like to express my utmost gratitude to the Almighty and His blessings, for providing me with the strength and perseverance needed to successfully complete this project.

I offer my sincere gratitude to **Professor A. A. Ziauddin Ahmad**, Chairperson, Department of Mathematics and Natural Sciences, BRAC University and **Professor Naiyyum Choudhury**, former Coordinator of the Biotechnology and Microbiology Program of the Department of Mathematics and Natural Sciences, BRAC University for their exemplary guidance and support during my tenure as a student under the Biotechnology Program at BRAC University.

I am indebted to my supervisor Associate Professor **Dr. Aparna Islam**, Department of Mathematics and Natural Sciences, BRAC University. She believed in me when I doubted myself and for that she has my earnest appreciation. Throughout the tenure of this research, Dr. Islam has constantly encouraged me to explore new ideas. Her scrutinous criticism and helpful advice have refined and enhanced my capability immensely. I could not have imagined having a better advisor and mentor for my undergraduate study.

I am much obliged to **Dr. ABM Md. Khademul Islam**, Associate Professor, Department of Genetic Engineering and Biotechnology, University of Dhaka for aiding and providing me with valuable insights regarding the technical aspects of this project.

I am grateful to all the faculty members of the Department of Mathematics and Natural Sciences for their unwavering support and guidance throughout the entire period of my bachelor's degree.

I would like to express my thanks to Mr. Samsad Razzaque for his help during the initial stages of this study. Furthermore, I would like to thank the seniors, laboratory officers, laboratory assistants as well as teaching assistants at the department for their support and advice. My profound gratitude goes to my friends Mr. Salman Khan Promon and Mr. Wasif Kamal for their unstinting support during the writing of this thesis.

Finally, I would like thank my parents for their sheer devotion to my education, future and happiness. They offered me their undying support, encouragement and listened to my frustrations with patience. I am eternally grateful for having them in my life.

September, 2015

Mohammad Rafid Feisal

Abstract

Climate changes have detrimental effects on the plants such as an increased susceptibility towards pathogens or ill health leading to food insecurity. This is eminently observed in developing countries such as Bangladesh. Hence, it is of crucial importance to comprehend the mechanisms the plants use to adapt to environmental stresses, such as, salinity. NHX-type antiporter facilitates the exchange of Na^+ for H^+ across the membranes. It sequesters Na^+ from the cytoplasm to vacuoles via the electrochemical H^+ gradient generated by two H^+ - pumps. In *Arabidopsis thaliana*, NHX1 encodes the vacuolar sodium or proton antiporter and it is identified as a significant salt tolerance determinant that is able to catalyze Na^+ accumulation in vacuoles. As such, it is necessary to determine the structure of the protein encoded by the NHX1 gene in *Arabidopsis thaliana*. This would allow us to establish the regions (e.g. active sites and secondary structural motifs) in the protein that affect the function and protein-protein interaction network in regard to the mechanisms involved in salinity tolerance. The objective of this study is to predict the three-dimensional structure of *Arabidopsis thaliana* sodium/hydrogen exchanger 1 protein via homology modeling and examine its physicochemical properties using *in silico* approaches. Biocomputational analyses of the target protein were performed using an array of online bioinformatics tools and databases and the homology model was developed using 3 different softwares (I-TASSER, Phyre2 and Easymodeller) and the best model was selected upon evaluation. In addition, the secondary structural motifs were identified within the model. The results suggested that the EasyModeller model, EM_Model 01, was the best amongst the three. It had the highest stereochemical quality scores and was considered to be the least unusual. The model consisted of $\alpha/\beta/\gamma$ topology where a single β -sheet constituted the β -hairpin as observed in the secondary structure schematic and topology diagrams. It was predicted that the presence of the β -hairpin allowed the protein to act as a membrane channel protein to facilitate the exchange of Na^+ and H^+ .

Table of Contents

Content	Page Number
Abstract	vi
Chapter 1: INTRODUCTION	1
Chapter 2: MATERIALS AND METHODS	10
Chapter 3: RESULTS AND DISCUSSION	36
Chapter 4: REFERENCES	78

CHAPTER 1:
INTRODUCTION

Chapter 1: Introduction

1.1 Membrane Proteins:

Geneticists, molecular biologists and cell biologists all are uncovering new proteins which are important in certain biological pathways and processes on a routine basis. However, due to limited knowledge regarding atomic structures of these proteins, the molecular functions or the mechanisms that are involved cannot be deciphered (Ramachandran and Dokholyan, 2012).

Membrane proteins play a crucial role within the cell system. Their activities range from transport of small molecules to the complex signaling pathways (Elofsson and Heijne, 2007). In plants, these proteins are of immense importance. This is because the plant cells are composed of several membrane systems performing several specialized functions. For instance, the plasma membrane functions as a communication interface with the outside environment for the exchange of substances (e.g. protons, anions and cations) and also as an information mediator (i.e. signal transduction) (Komatsu *et al.*, 2007).

Even though membrane proteins are of prime importance, it is still extremely difficult to attain high resolution three-dimensional (3D) structures of these proteins. But such knowledge is important to predict their topology (i.e. the transmembrane regions as well as their orientation across the membrane) and fold type, based on the amino acid sequence to understand the mode of function (Elofsson and Heijne, 2007). At present they stand for less than 1% of the structures present in the Protein Data Bank (Berman *et al.*, 2000). However, the number of experimentally known membrane protein structures is constantly increasing (White, 2004, Oberai *et al.*, 2006).

The two fundamental components that comprise the integral membrane proteins are: the α -helix and β -barrel. The helix-bundle proteins are present in all the cellular membranes and thus represent approximately 20-25% of all the open reading frames (ORFs) in the completely sequenced genomes (Krogh *et al.*, 2001). However, β -barrel membrane proteins are difficult to identify by sequence gazing. Therefore, their numbers remain undecided (Elofsson and Heijne, 2007).

1.2 The cation/ H⁺ exchangers (CPAs):

For this reason, among the membrane proteins, the cation/H⁺ exchangers are crucial. Ion and pH homeostasis are elemental regulators of cellular processes that establish and control plant growth. The H⁺-translocating enzymes are vital to the establishment and maintenance of cellular ion and pH balance. These generate the H⁺ electrochemical potential gradients and the cation/H⁺ exchangers. They use these gradients to couple the passive transport of H⁺ to the movement of cations against their electrochemical potential (Blumwald, 1987).

In plants, a number of monovalent cation/H⁺ transporters have been identified. These are classified into the large CPA family (Bassil *et al.*, 2012). The activity of these cation/H⁺ antiporters (CPAs) is extremely vital to the growth, cell turgor regulation, cellular osmotic adjustment and development of the plants. Additionally, coupled cation/H⁺ exchanges have an important function in regulation of ionic composition and pH of the internal milieu of endosomes and vacuoles which has an effect on the vesicular cargo composition, processing, vesicular movement as well as protein trafficking (Pardo *et al.*, 2006, Rodríguez-Rosales *et al.*, 2009).

The coupled exchange of K⁺ or Na⁺ for H⁺ is known to occur across membranes of all organisms, from prokaryotes to higher eukaryotes (Brett *et al.*, 2005, Pardo *et al.*, 2006, Rodríguez-Rosales *et al.*, 2009, Chanroj *et al.*, 2012, Orlowski and Grinstein, 2011). This K⁺(Na⁺)/H⁺ exchange is mediated by members of a family of transporters referred to as Na⁺/H⁺ antiporters (NHXs) in plants or Na⁺/H⁺ exchangers (NHEs) in animals.

The NHX functional groups appeared early in evolution and have conserved and fundamental cellular roles in plants (Bassil *et al.*, 2012). A number of recent publications have made significant contribution to our understanding of the roles of the NHX-type Na⁺/H⁺ antiporters in the regulation of vesicular trafficking, cell expansion, development and growth (Bassil *et al.*, 2011a).

In *Arabidopsis*, the NHX antiporters are comprised of six members; these intracellular members NHX1–NHX6 are again divided into two groups; a vacuolar group (NHX1–NHX4) and an endosomal group (NHX5 and NHX6)

based on localization and proposed cellular roles (Bassil *et al.*, 2012). Recent genetic evidence has confirmed that two of the most abundant vacuolar NHX antiporters in *Arabidopsis* primarily are K^+/H^+ exchangers which under normal growth conditions, are necessary for growth and development (Apse *et al.*, 2003, Rodríguez-Rosales *et al.*, 2008, Bassil *et al.*, 2011b, Barragán *et al.*, 2012). On the other hand, endosomal NHX antiporters have been shown to be decisive regulators of vesicle trafficking, particularly in the vacuole (Bassil *et al.*, 2012).

1.3 The vacuolar Na^+/H^+ antiporter:

Another important NHX-type antiporter is the vacuolar Na^+/H^+ antiporter. The Na^+/H^+ antiporters (exchangers) drive the exchange of Na^+ for H^+ across the membranes. Antiporters are found in animals, yeasts bacteria and plants; however antiporter activity within the vacuolar membranes has only been reported in yeast, algae and plants (Blumwald *et al.*, 2000). In plants the Na^+/H^+ antiporter within the vacuolar membranes sequesters Na^+ from the cytoplasm to vacuoles via the electrochemical H^+ gradient generated by two H^+ - pumps, namely, the vacuolar H^+ -inorganic pyrophosphatase and vacuolar H^+ -ATPase (Figure 1.1). The plant cells that undergo treatment with high salinity must uphold a higher K^+/Na^+ ratio in the cytoplasm and thus control the osmotic balance of the cell with the environment by gathering Na^+ in the vacuoles. Using the aforementioned process, the vacuolar Na^+/H^+ antiporter is believed to play vital role(s) (Fukuda *et al.*, 2004).

In *Arabidopsis thaliana*, NHX1 encodes the vacuolar sodium or proton antiporter. It is involved in salt tolerance, leaf development as well as ion homeostasis. It acts in low affinity electroneutral exchange of protons for cations like Na^+ or K^+ across the membranes. Furthermore, it can exchange Li^+ and Cs^+ with low affinity. This particular gene is engaged in the vacuolar ion compartmentalization that is needed for the cell volume regulation and cytoplasmic Na^+ detoxification (Sottosanto *et al.*, 2007, Mahdi, 2014).

1.4 Climate change and plant antiporters:

Plant stresses are the reasons behind food insecurity and therefore pose as a major threat to mankind (Jewell *et al.*, 2010). One of the biggest problems is

environmental stress and this is considered as a responsible phenomenon for reduction of crop yields (Hussain *et al.*, 2011).

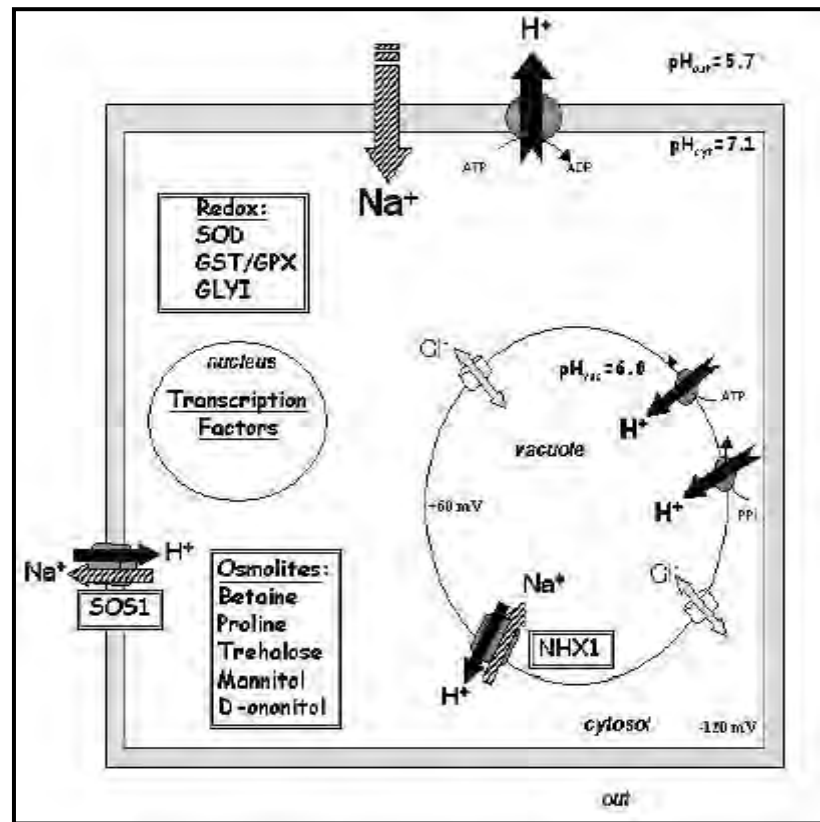


Figure 1.1: Schematic representation of primary and secondary transport in the plant cells. Electrogenic H^+ transport (H^+ -ATPase in the plasma membrane and vacuolar membrane, H^+ -PPiase in the vacuolar membrane) generates gradients of pH and electrical potential difference across the cell and vacuolar membranes. Na^+ ions enter the cell and can be translocated out of the cell or into the vacuole by the action of a plasma membrane Na^+/H^+ antiporter (SOS1) or a vacuolar Na^+/H^+ antiporter (NHX1), respectively (Source: Razzaque, 2011)

Climate changes have certain effects, such as, an increment in humidity may lead to a plant's increased susceptibility to pathogens or an increase in global temperatures may cause drought (Battisti and Naylor, 2009). The aforementioned factors endanger food security and therefore lead to social instability and poverty especially in the developing countries as well as throughout the world (Ronald, 2011). Therefore, it is very important to understand the mechanisms which plants use to adapt to environmental stresses and thus maintain food supplies on a global scale (Razzaque *et al.*, 2014). This will further allow us to understand how plants might be able to adapt to climate changes.

Plants respond to environmental stresses at both cellular and molecular level by changing the expression of many genes by means of different types of complex molecular signaling networks (Akpinar *et al.*, 2012). As such, knowledge of these pathways including identification of regulatory codes would enable us to develop stress tolerant plants through genetic manipulations (Razzaque *et al.*, 2014).

In their study, Razzaque and his colleagues (Razzaque *et al.*, 2014) using *in silico* methods focused on finding the connection between upregulated genes under different abiotic stress conditions using *Arabidopsis* as a model organism. They were able to identify common genes that are upregulated during various environmental stresses in *Arabidopsis thaliana* using freely available microarray datasets. They also proposed a protein-protein interaction network that may help comprehend the abiotic stress tolerance mechanism. Their study brought out 42 genes/transcription factors/enzymes that play vital roles during abiotic stress response. Thirty genes from those forty-two were highly correlated in all four datasets and only eight from those thirty genes were determined as highly responsive to the above abiotic stresses. One of the eight targeted genes/proteins is Na⁺/H⁺ exchanger (NHX1). According to their study each targeted protein brings more stress responsive molecules into a single string so that they can provide tolerance. For NHX1 protein it was observed that it connects with some cold responsive and drought response elements which elucidated its functional activity during targeted abiotic stress response. It has a strong physical binding affinity with other antiporters, like, NHX2, NHX3, CHX2, SOS1 etc. which makes it an important molecule in stress response mechanism (Mahdi, 2014).

1.5 Benefits of determining structure of Na⁺/H⁺ exchanger 1 (NHX1) of *Arabidopsis thaliana*:

In light of the aforementioned, it is necessary to determine the structure of NHX1 protein in *Arabidopsis thaliana*. This will allow us to verify the protein-protein interaction network. Predicting the three dimensional (3D) structure of the AtNHX1 protein would enable us to decipher the regions within the protein that play key roles in protein-protein interaction. For instance, this includes active sites or the secondary structural elements that affect the function of the protein in the network.

To date, no X-ray crystallographic structures for animal NHEs, or yeast or plant NHX antiporters are available (Bassil *et al.*, 2012). However, it is possible to attain structural models of the aforementioned proteins using homology modeling techniques.

1.6 Significance of structural analysis and physicochemical characterization using *in silico* approaches:

Usually analysis of a protein, which includes characteristics as well as determination of structure, can be done *in silico* which is offered through the use of bioinformatics tools and an array of various online databases.

During such studies, homologous proteins are identified and then compared in regard to their structural and functional properties to know the unfamiliar ones. Such data can then be used in laboratory experiments to establish properties and subsequently lead to discover novel proteins (Kallberg, 2002). The verifications can be used within bioinformatics so as to attain much more accurate and detailed results in terms of function and protein-protein interaction. Hence, *in silico* methods play a significant role and need to be used in collaboration with biology, biochemistry, medicine and so on.

1.7 Homology modeling and its significance:

Determination of the experimental structures of several proteins has technical challenges. The methods that are currently available for attaining atomic-resolution structures of biomolecules (X-ray crystallography and NMR spectroscopy) need pure preparations of proteins at concentrations which are higher than those at which the proteins exist in the physiological environment. Furthermore, the NMR has size restrictions. For these reasons atomic structures of many important proteins, concerning medical or biological aspects, are not present (Ramachandran and Dokholyan, 2012).

Comparative modeling or homology modeling is of great importance as it is a tool that bridges the gap between sequence and structure. Moreover this allows researchers to construct structural models of proteins that are complex to crystallize or for which structure determination via NMR spectroscopy is not

amenable (Ramachandran and Dokholyan, 2012). This whole process exploits the information of two proteins that have sequences related on an evolutionary scale and thus expected to have comparable structural features (Chothia and Lesk, 1986). Therefore, the known structure of the protein referred to as the “template” can be used to generate a molecular model of the query protein whose experimental structure is unknown (Figure 1.2).

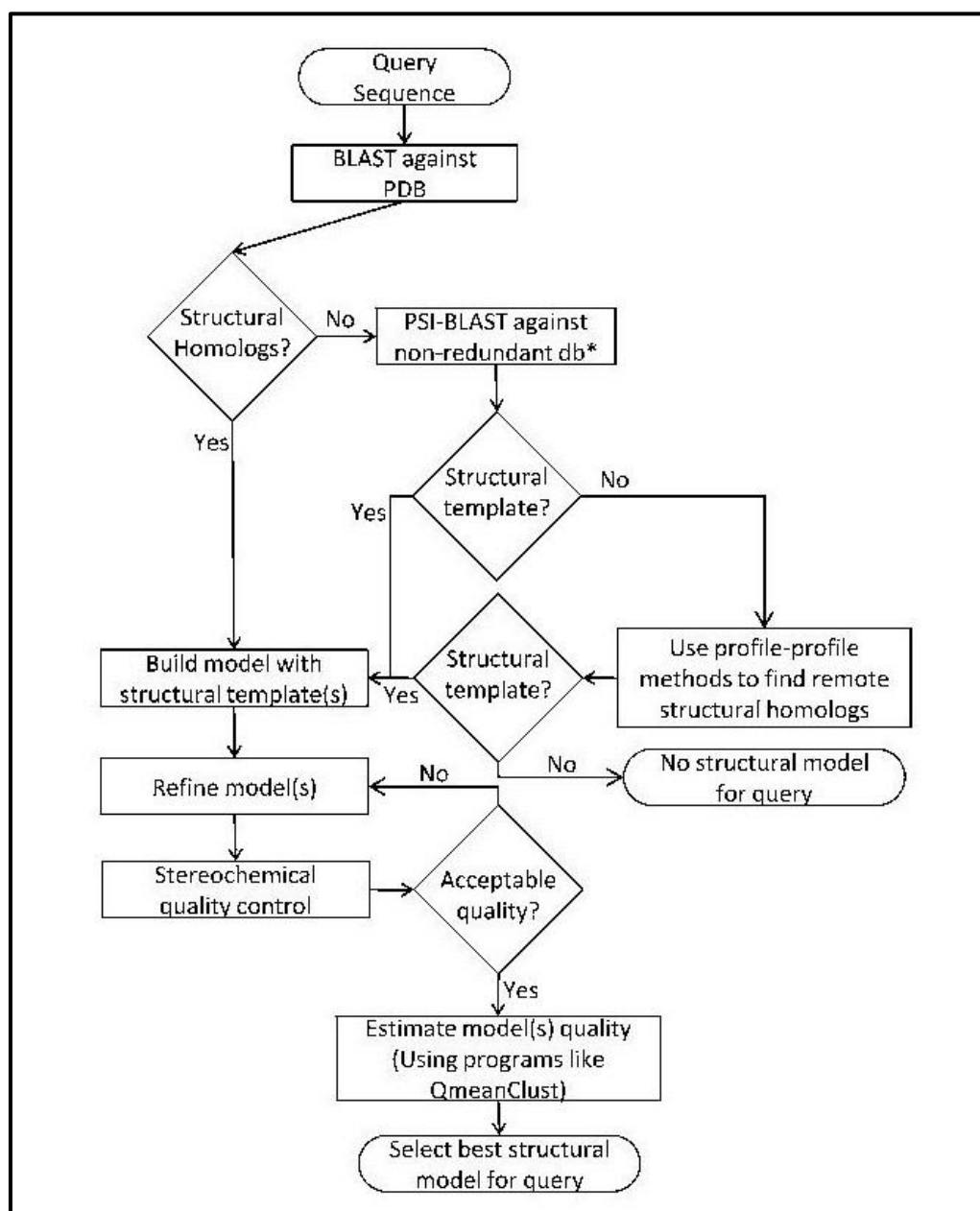


Figure 1.2: Flowchart of the steps followed in the construction of a comparative structural model (*database) (Source: Ramachandran and Dokholyan, 2012)

1.8 Current research objectives:

In context of the previous sections, the objectives of this particular research are:

1. To identify and select homologous sequences in relation to the query nucleotide and protein sequence
 - **Nucleotide query sequence:** *Arabidopsis thaliana* sodium/hydrogen exchanger 1 mRNA, complete *cds* (NCBI Reference Sequence: NM_122597.2)
 - **Protein query sequence:** sodium/hydrogen exchanger 1 [*Arabidopsis thaliana*] (NCBI Reference Sequence: NP_198067.1)
2. To compare both data sets and select the matching protein sequences with the nucleotide sequences
3. To generate phylogenetic trees of the selected sequences
4. To analyze physicochemical properties of the selected proteins using *in silico* methods
5. To predict transmembrane regions, secondary structure content and secondary structure of the query protein using *in silico* methods
6. To predict the three-dimensional (3D) structure of the query protein using homology modeling techniques and identify its structural motifs

CHAPTER 2:
MATERIALS AND METHODS

Chapter 2: Materials and Methods

2.1 Work plan:

In this study, different databases and online tools were used to attain and analyze the desired gene and protein sequences using *in silico* approaches. Several online and offline software were used to predict the three-dimensional structure of the target protein. The work plan(s) of the present study are illustrated which depicts the steps taken to attain the intended outcome (Figures 2.1-2.3).

2.2 Description and methods of different bioinformatics databases and tools/software used in this study:

2.2.1 Databases:

2.2.1.1 National Center for Biotechnology Information (NCBI):

Established in 1988 as a national resource for molecular biology information, the National Center for Biotechnology Information (NCBI) (Geer *et al.*, 2010) creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. All these databases are available online through the Entrez search engine (Figure 2.4 a).

URL link: <http://www.ncbi.nlm.nih.gov/>

The NCBI databases, namely, nucleotide and protein databases were used to retrieve the target nucleotide and protein sequences in relation to *Arabidopsis thaliana* in the current study. The AtNHX1 gene based on the nucleotide database was searched. The chosen sequence retrieved from the database was (Figure 2.4 b):

- Accession: NM_122597.2: *Arabidopsis thaliana* sodium/hydrogen exchanger 1 mRNA, complete cds

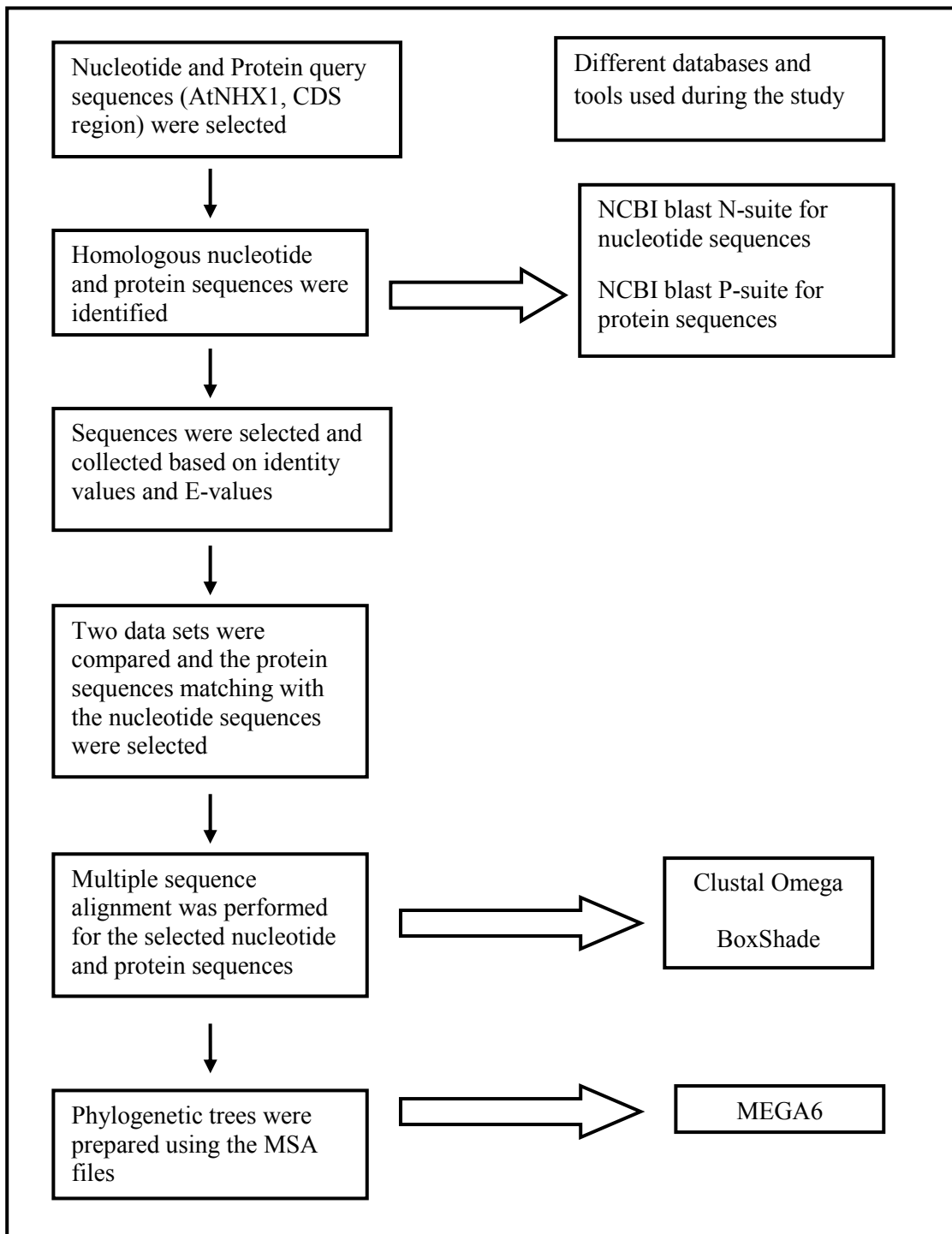


Figure 2.1: Experimental work plan for selection and identification of homologous sequences leading to phylogenetic tree generation

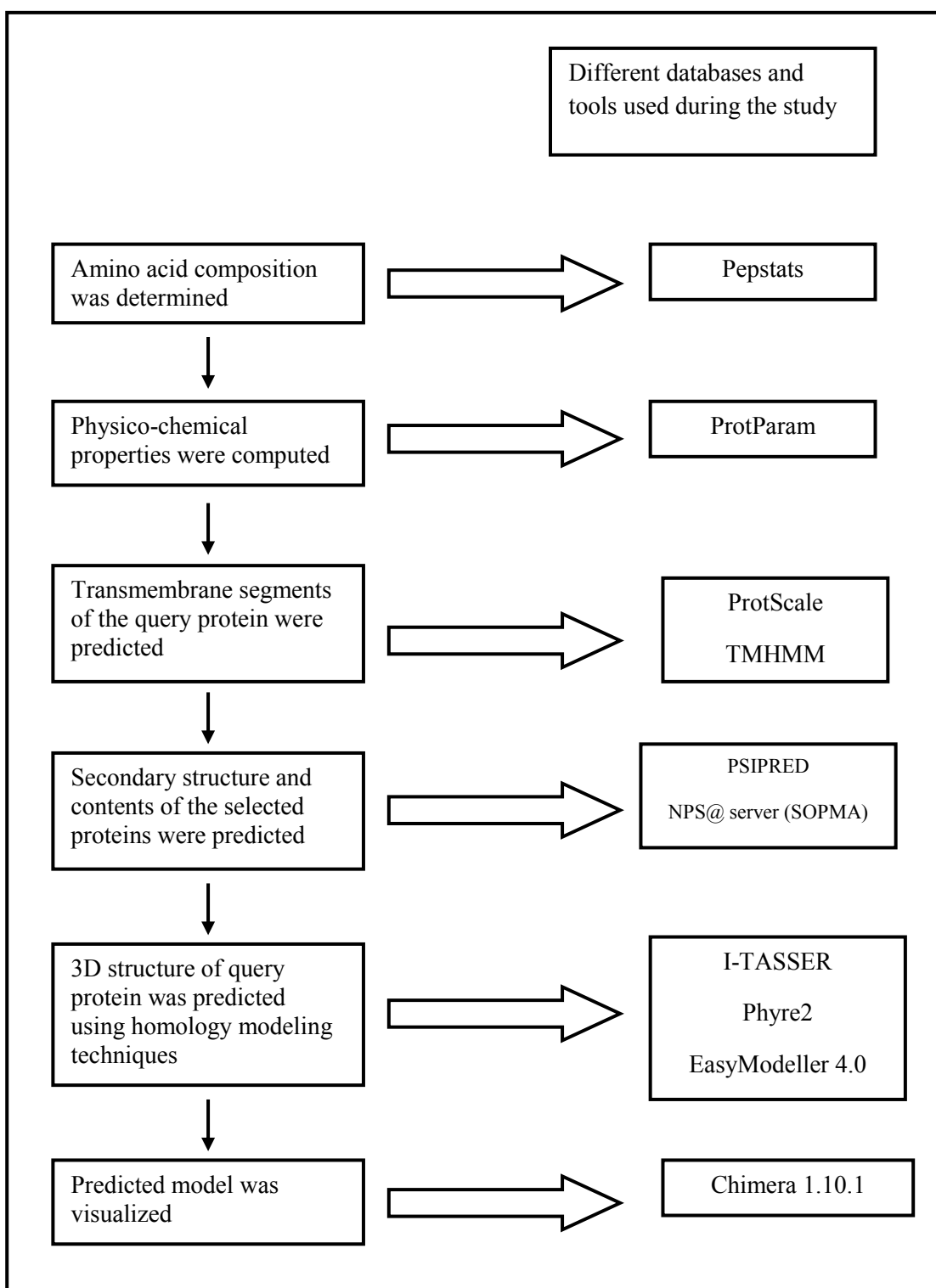


Figure 2.2: Experimental work plan for analysis of selected protein sequences using *in silico* approaches

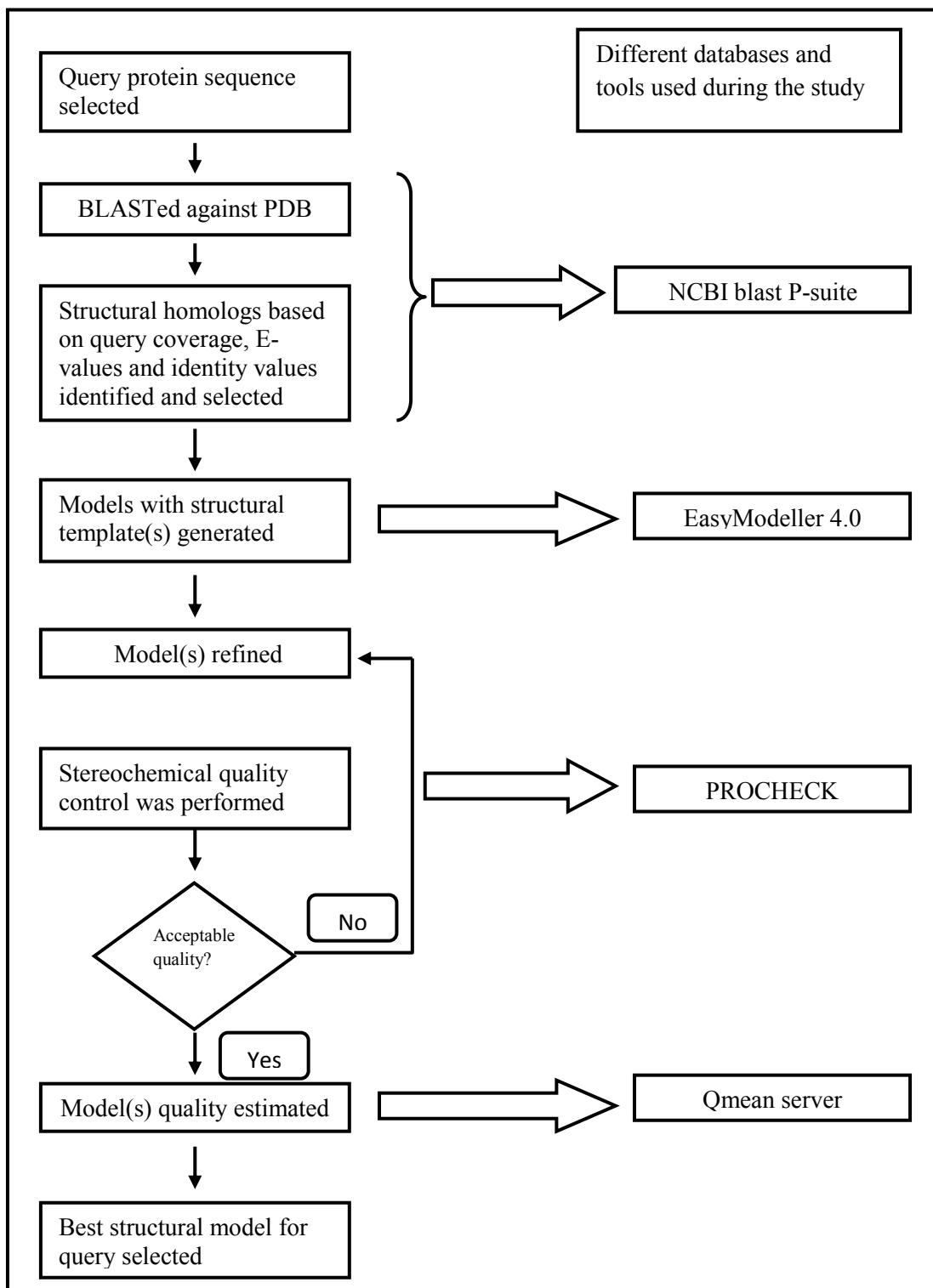


Figure 2.3: Experimental work plan depicting steps taken to attain homology model via EasyModeller 4.0

The area of interest within the entire sequence was the CDS (Coding Sequence) region (Location: 471 to 2087; Span: 1617; Product: 538). This was observed in the Graphics view of the query gene (Figure 2.4 c). The green streak represented the entire sequence present within the gene. The red streak represented the CDS region and the black streak represented the NhaP-type Na⁺/H⁺ or K⁺/H⁺ antiporter [Inorganic ion transport and metabolism] region. From the Graphics view, the corresponding protein sequence in relation to the CDS region of the gene was retrieved. The retrieved (query) protein sequence was:

- Accession: NP_198067.1: Sodium/hydrogen exchanger 1
[Arabidopsis thaliana]

2.2.1.2 Expert Protein Analysis System (ExPASy):

ExPASy (Artimo *et al.*, 2012) is a bioinformatics resource portal operated by the Swiss Institute of Bioinformatics (SIB) and in particular the SIB Web Team. It is an extensible and integrative portal which allows access to scientific resources, databases and software tools in different areas of life sciences. Scientists can access a wide range of resources in several different domains, such as, proteomics, genomics, phylogeny/evolution, systems biology, population genetics and transcriptomics. On this portal one would find resources from many different SIB groups as well as external institutions (Figure 2.5).

URL link: <http://www.expasy.org/>

2.2.1.3 EMBL-EBI:

The European Bioinformatics Institute (EBI) is an academic research institute located on the Wellcome Trust Genome Campus in Hinxton near Cambridge (UK). It is part of the European Molecular Biology Laboratory (EMBL). It provides freely available resources for life science experiments which lead to basic research in computational biology (Figure 2.6).

URL link: <http://www.ebi.ac.uk/>

NCBI Resources How To Sign in to NCBI

NCBI National Center for Biotechnology Information

All Databases Search

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News](#)

Submit
Deposit data or manuscripts into NCBI databases

Download
Transfer NCBI data to your computer

Learn
Find help documents, attend a class or watch a tutorial

Develop
Use NCBI APIs and code libraries to build applications

Analyze
Identify an NCBI tool for your data analysis task

Research
Explore NCBI research and collaborative projects

Popular Resources

- PubMed
- Bookshelf
- PubMed Central
- PubMed Health
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

NCBI Announcements

July 15th webinar: "E-Direct: Bringing the E-Utilities to the UNIX Command Line" 01 Jul 2015

In two weeks, NCBI staff will introduce E-Direct a simple, easy-to-use interface for the E-Utilities.

Tree Viewer version 1.5 improves performance

a

Arabidopsis thaliana sodium/hydrogen exchanger 1 mRNA, complete cds

NCBI Reference Sequence: NM_122597.2

[FASTA](#) [Graphics](#)

Go to:

LOCUS NM_122597 2334 bp mRNA linear PLN 22-JAN-2014

DEFINITION Arabidopsis thaliana sodium/hydrogen exchanger 1 mRNA, complete cds.

ACCESSION NM_122597

VERSION NM_122597.2 GI:30690553

KEYWORDS RefSeq.

SOURCE Arabidopsis thaliana (thale cress)

ORGANISM [Arabidopsis thaliana](#)
Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; Gunneridae; Pentapetalae; rosids; malvids; Brassicales; Brassicaceae; Camelineae; Arabidopsis.

REFERENCE 1 (bases 1 to 2334)

AUTHORS Swarbreck,D., Lamesch,P., Wilks,C. and Huala,E.

CONSTRM Arabidopsis TAIR10 Release

TITLE Direct Submission

JOURNAL Submitted (18-FEB-2011) The Arabidopsis Information Resource, Department of Plant Biology, Carnegie Institution, 260 Panama Street, Stanford, CA, USA

COMMENT REVIEWED [REFSEQ](#): This record has been curated by TAIR. This record is derived from an annotated genomic sequence (NC_003076).
On May 13, 2003 this sequence version replaced gi:18421084.

FEATURES

source Location/Qualifiers

1..2334

/organism="Arabidopsis thaliana"

/mol_type="mRNA"

/db_xref="taxon:3702"

/chromosome="5"

/ecotype="Columbia"

b

100 200 300 400 500 600 700 800 900 1K 1,100 1,200 1,300 1,400 1,500 1,600 1,700 1,800 1,900 2K 2,100 2,334

NM_122597.2: 1..2.3K (2.3Kbp) Find: Tools Configure

1 100 200 300 400 500 600 700 800 900 1K 1,100 1,200 1,300 1,400 1,500 1,600 1,700 1,800 1,900 2K 2,100 2,334

Genes

NP_198067.1

HsdP

c

Figure 2.4: (a) NCBI Homepage, (b) GenBank profile for NM_122597.2 and (c) The Graphics view of NM_122597.2

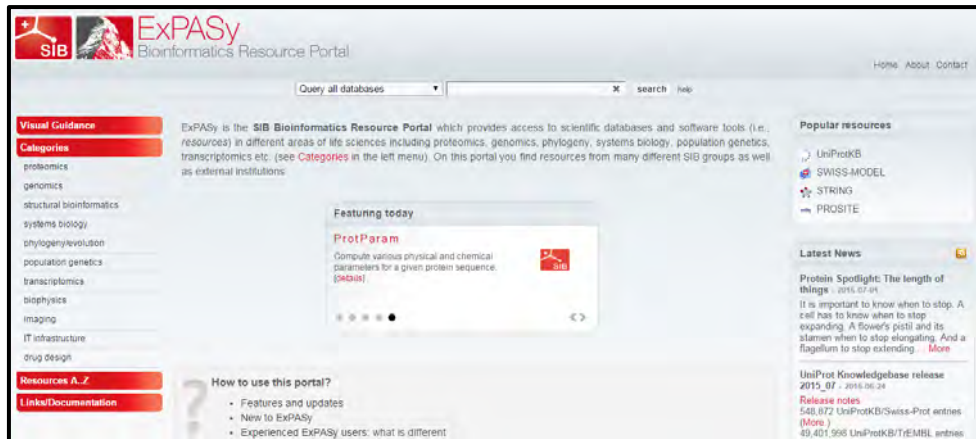


Figure 2.5: ExPASy homepage

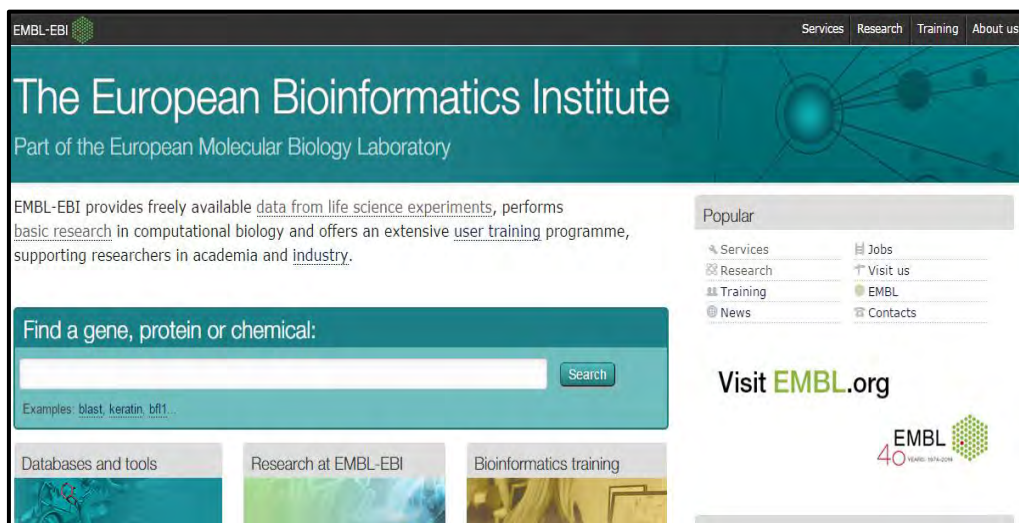


Figure 2.6: EMBL-EBI homepage

2.2.1.4 Protein Data Bank (PDB):

The Protein Data Bank (PDB) (Berman *et al.*, 2000) is a crystallographic database for the three-dimensional (3D) structural data of large biological molecules, such as, proteins and nucleic acids. The data, typically obtained by X-ray crystallography or NMR spectroscopy and submitted by biologists and biochemists from around the world, are freely accessible on the Internet via websites of its member organizations (PDBe, PDBj and RCSB). The PDB is overseen by an organization called the Worldwide Protein Data Bank, wwPDB. The RCSB PDB was used to retrieve structural templates required to generate the query protein homology model (Figure 2.7).

URL link: <http://www.rcsb.org/pdb/home/home.do>

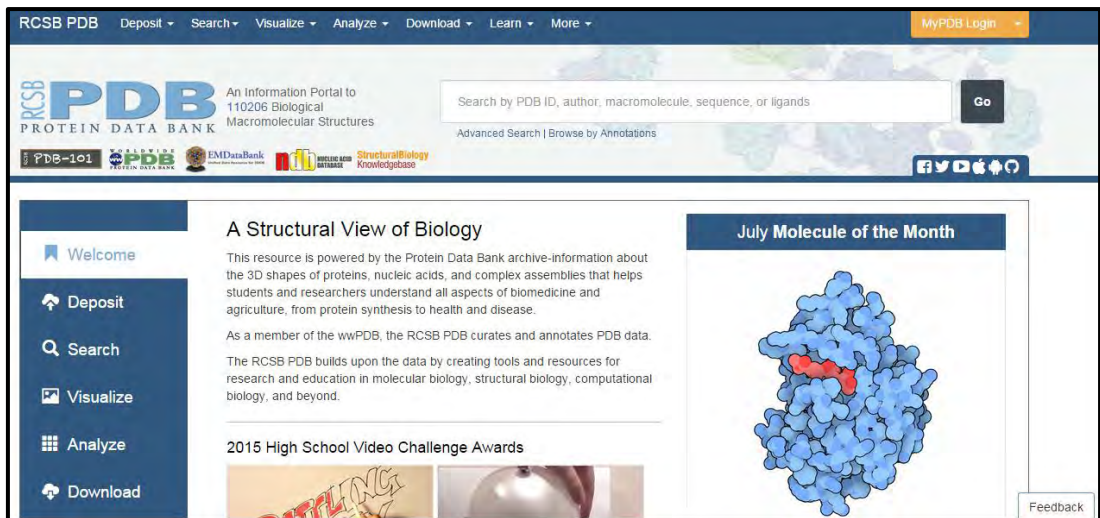


Figure 2.7: PDB homepage

2.2.2 Tools and software:

2.2.2.1 BLAST:

The Basic Local Alignment Search Tool (BLAST) (Coordinators, 2013, Boratyn *et al.*, 2013, Johnson *et al.*, 2008) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families. In this study both blast N-suite and blast P-suite were used (Figure 2.8).

URL link: <http://blast.ncbi.nlm.nih.gov/Blast.cgi>

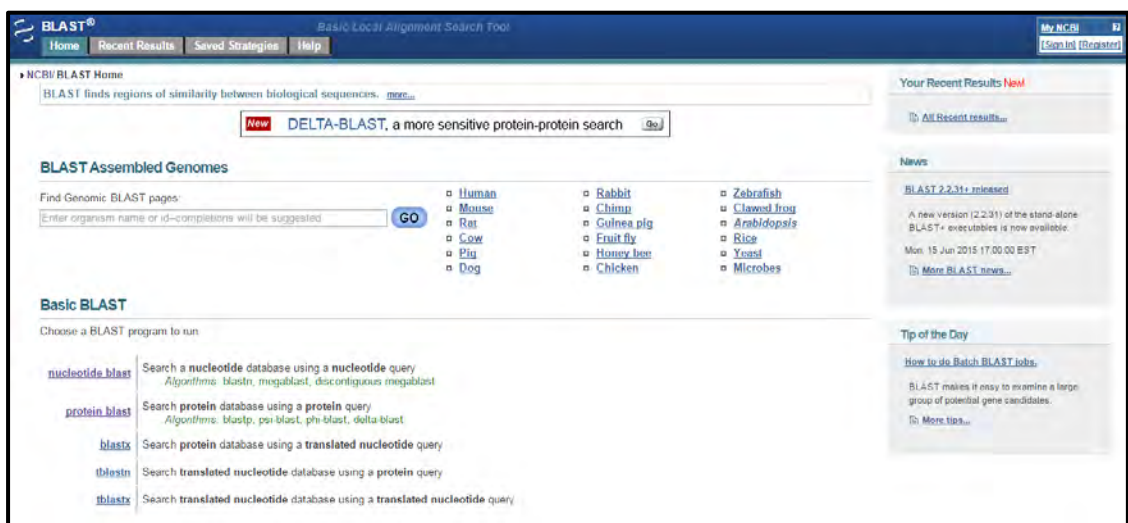


Figure 2.8: BLAST Homepage

2.2.2.2 Clustal Omega:

Clustal Omega is a completely rewritten and revised version of the widely used Clustal series of programs for multiple sequence alignment. It can deal with very large numbers of DNA/RNA or protein sequences. The accuracy of the program has been considerably enhanced over earlier Clustal programs, through the use of the HAlign method for aligning profile hidden Markov models. The program currently is used from the command line or can be run on line (Sievers and Higgins, 2014) (Figure 2.9).

URL link: <http://www.ebi.ac.uk/Tools/msa/clustalo/>

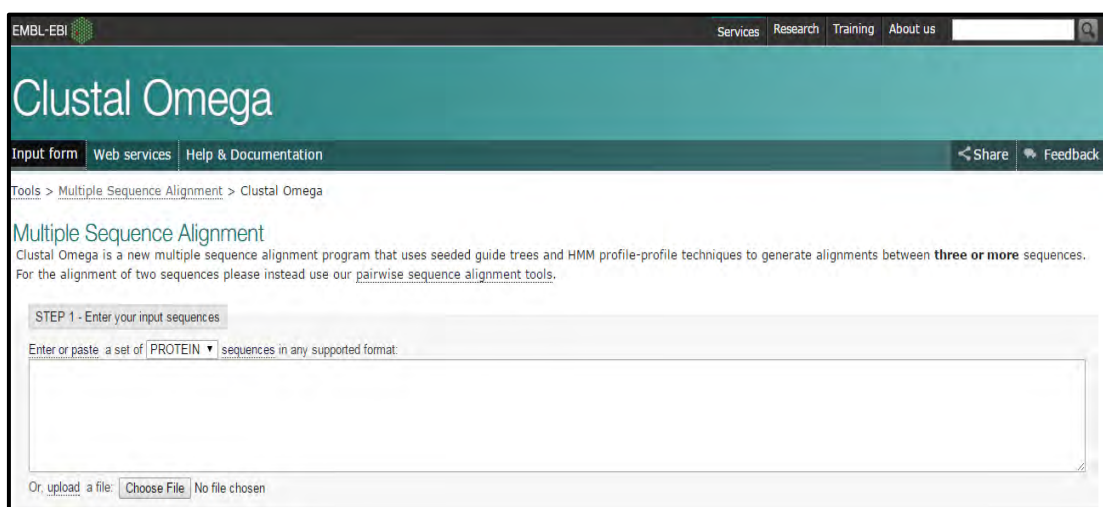


Figure 2.9: Clustal Omega homepage

2.2.2.3 BoxShade:

BoxShade is a program for pretty-printing multiple alignment output. The program itself does not carry out alignment of the selected nucleotide or protein sequences, as such, a multiple sequence alignment (MSA) programs like Clustal Omega or Clustal W2 needs to be used. Following so, the outputs of the programs are used as inputs for BoxShade to attain publishable images of the MSA results. The output format selected for the current study was RTF new (Figure 2.10).

URL link: http://www.ch.embnet.org/software/BOX_form.html

2.2.2.4 Molecular Evolutionary Genetics Analysis (MEGA):

Molecular Evolutionary Genetics Analysis (MEGA) is an integrated tool for conducting sequence alignments, estimating divergence times, inferring phylogenetic

trees, online database mining, molecular evolution rate estimation, inferring ancestral sequences and testing evolutionary hypotheses. It is used by biologists for reconstruction of evolutionary histories of species and hypothesizing/theorizing the extent and nature of the selective forces that shape the evolution of genes as well as species. The software is available online and can be downloaded (Figure 2.11).

URL link: <http://www.megasoftware.net/>

In this current study, MEGA 6 (Tamura *et al.*, 2013) was used to generate phylogenetic trees for both nucleotide and protein sequences.

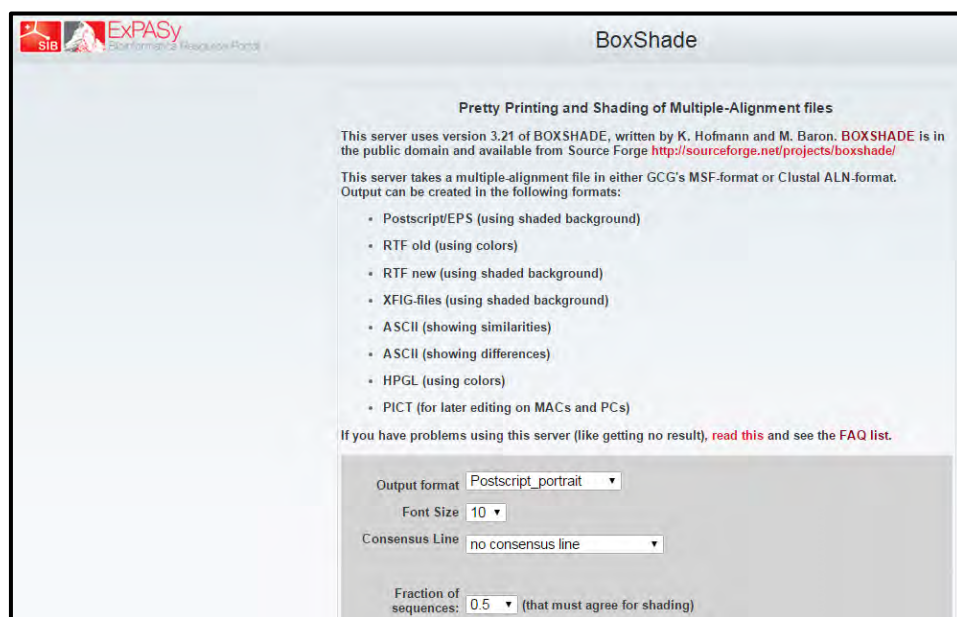


Figure 2.10: BoxShade Homepage

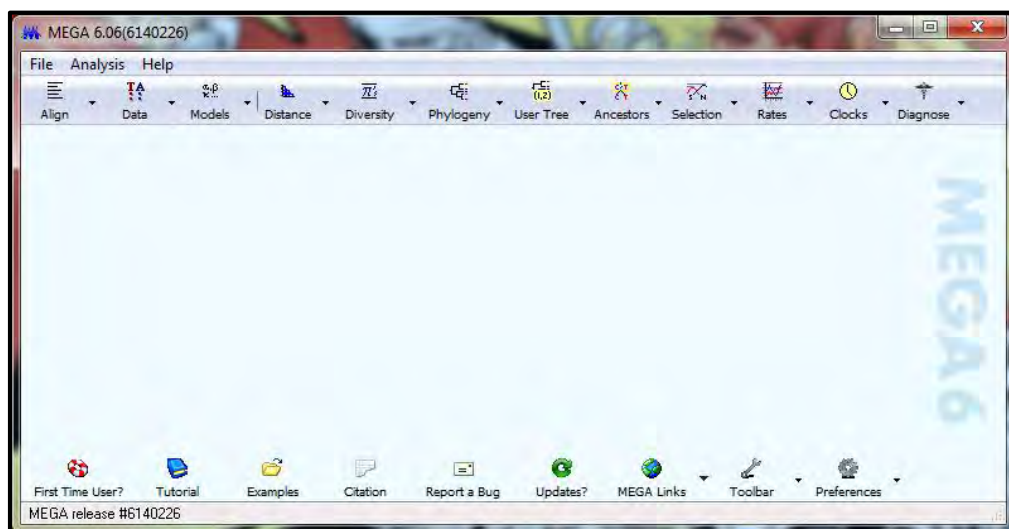


Figure 2.11: MEGA6 opening window

2.2.2.4.1 Parameters used for phylogenetic tree generation in MEGA 6:

2.2.2.4.1.1: Nucleotide sequence data:

For the generation of nucleotide sequence based phylogenetic trees, in the input data section Nucleotide sequences were selected and it was confirmed as the protein-coding nucleotide sequence data. For the selection of genetic code, the Standard option was selected (Figure 2.12).

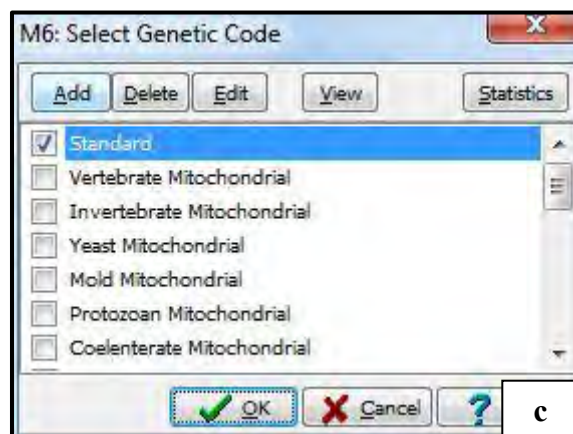
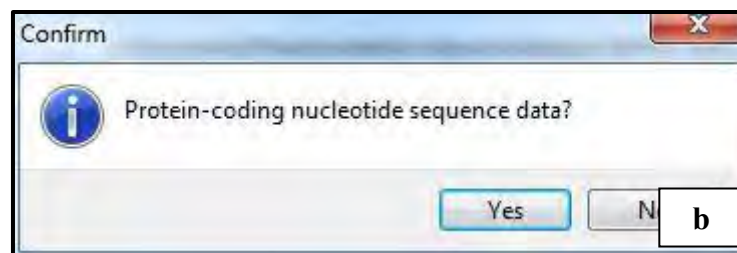
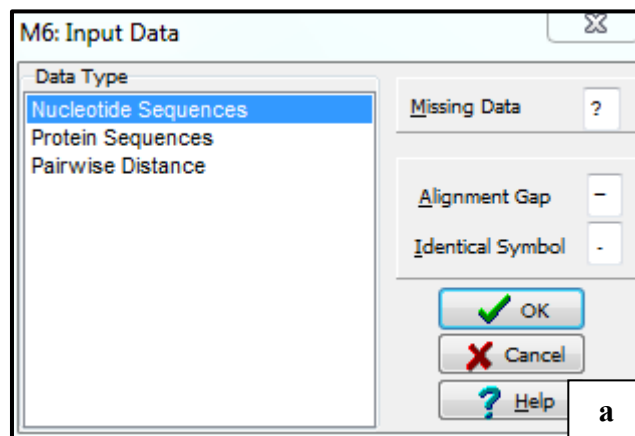


Figure 2.12: Steps taken for setting parameters for nucleotide sequence data. (a) Input data, (b) Confirmation and (c) Selection of genetic code

Using the Phylogeny option in MEGA 6, a Neighbor-Joining (NJ) phylogenetic tree was constructed. In the Analysis Preferences section of the software, in the statistical method, NJ method was selected. For the test of phylogeny the Bootstrap method was selected to determine the robustness with replicates set at 500. This was done as the NJ method does not have any clade support measure. The Nucleotide Substitutions type was selected. The model used was Maximum Composite Likelihood. Transitions and Transversions were selected as the substitutions to be included. For Rates and Patterns, the Rates were set to Gamma distributed (G) and the gamma parameter was set to 2. The Pattern among lineages was set to Homogenous and for gaps and missing data, Complete deletion was selected (Figure 2.13).



Figure 2.13: Parameters used for the construction of phylogenetic trees for nucleotide sequence data

2.2.2.4.1.2 Protein sequence data:

For the generation of protein sequence based phylogenetic trees, in the input data section Protein sequences were selected (Figure 2.14).

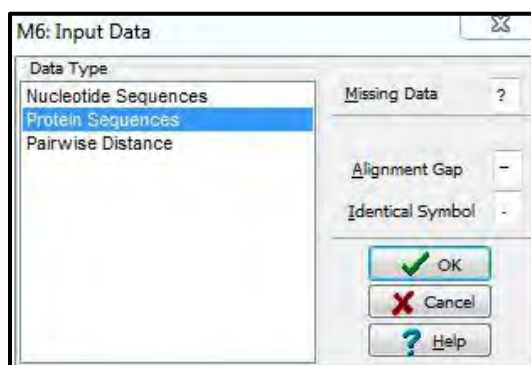


Figure 2.14: Input data type selection for protein data

Using the Phylogeny option in MEGA 6, a Neighbor-Joining (NJ) phylogenetic tree was constructed. In the Analysis Preference section, in the statistical method NJ method was selected. The Bootstrap method was selected for the test of phylogeny with replicates set at 500. Amino acid Substitutions type was selected. The model used was Poisson model. For Rates and Patterns, the Rates were set to Gamma distributed (G) and the gamma parameter was set to 2. The Pattern among lineages was set to Homogenous and for gaps and missing data, Complete deletion was selected (Figure 2.15).

2.2.2.5 Pepstats:

Pepstats analysis tool (Li *et al.*, 2015, McWilliam *et al.*, 2013) is an online tool that is able to calculate the statistics of protein properties. It is available at the EMBL-EBI website. It provides the user with a range of properties such as molecular weight, isoelectric points, extinction coefficients as well as the amino acid composition in terms of percentages. In this current study, the FASTA sequences of the selected proteins along with the query protein were submitted and the results attained were analyzed to determine the most abundant and least common amino acids in the proteins (Figure 2.16).

URL link: http://www.ebi.ac.uk/Tools/seqstats/emboss_pepstats/

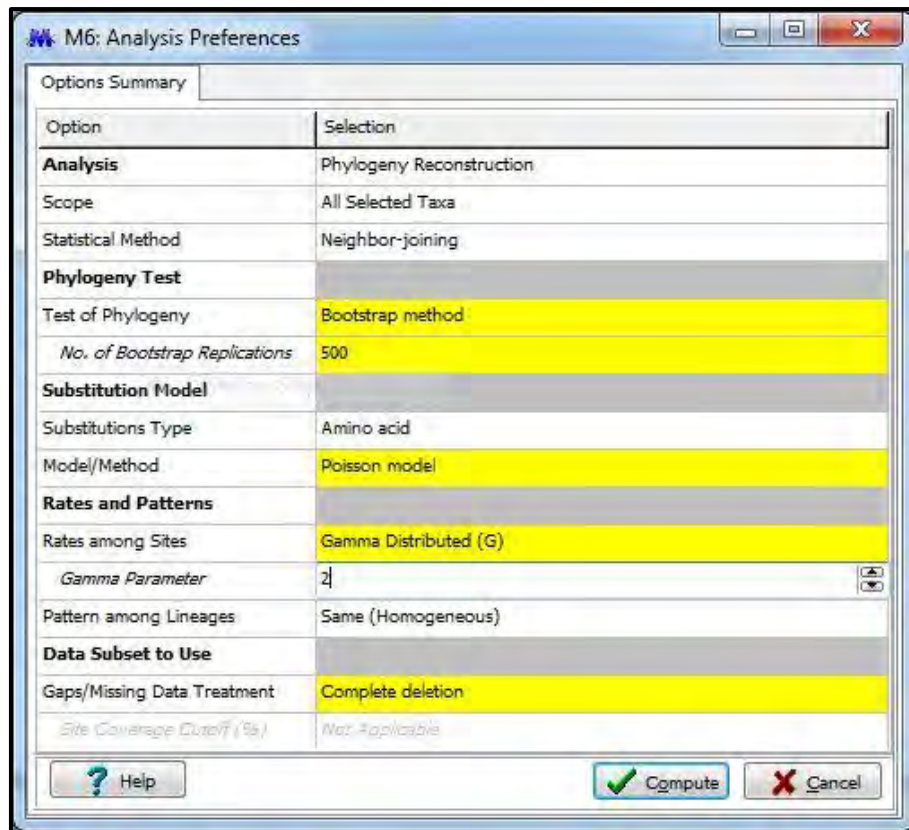


Figure 2.15: Parameters used for the construction of phylogenetic trees for protein sequence data

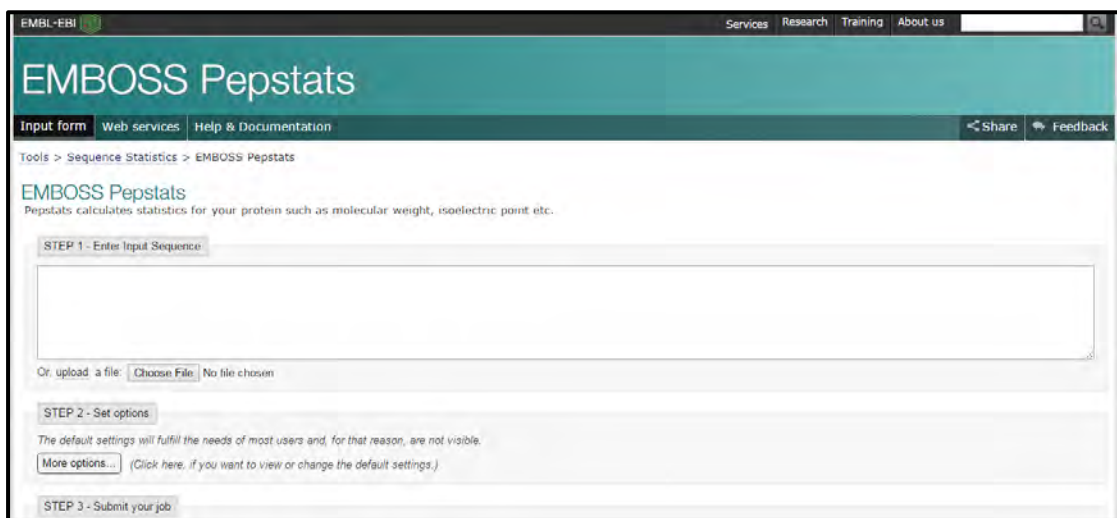


Figure 2.16: Pepstats homepage at EMBL-EBI website

2.2.2.6 ProtParam:

ProtParam (Gasteiger *et al.*, 2005) is an online tool which is available at the ExPASy server. It computes various physicochemical properties that can be deduced from a protein sequence attained from the Swiss-Prot or TrEMBL or it can be user entered

protein sequence. The computed parameters include the molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index and grand average of hydropathicity (GRAVY). In this study, the FASTA sequences of the selected proteins along with the query were uploaded and the results attained were analyzed (Figure 2.17).

URL link: <http://web.expasy.org/protparam/>

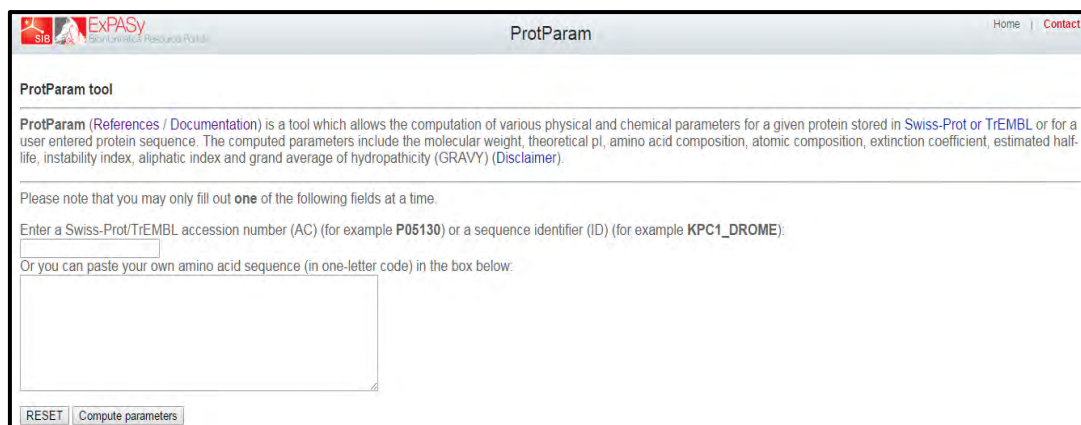


Figure 2.17: ProtParam Homepage

2.2.2.7 ProtScale:

ProtScale (Gasteiger *et al.*, 2005) is an online tool which is available at the ExPASy server. It allows the user to compute and represent (in the form of two dimensional plot) the profile produced by any amino acid scale of a selected protein. An amino acid scale is defined by a numerical value assigned to each type of amino acid. The most often used scales are the hydrophobicity scales. Most of these scale were derived from experimental studies on partitioning of peptides in apolar and polar solvents, with the goal of predicting membrane-spanning segments that are highly hydrophobic, and secondary structure conformational parameter scales. It can be used with 50 different pre-defined scales. The scale values for the 20 amino acids, as well as a literature reference, are provided on ExPASy for each of these scales. To generate data for a plot, the protein sequence is scanned with a sliding window of a given size. At each position, the mean scale value of the amino acids within the window is calculated, and that value is plotted for the midpoint of the window. The window size is the number of amino acids analyzed at a time needed to determine the points of hydrophobicity or hydrophilic regions. Window sizes of 19 or 21 will make

hydrophobic, membrane-spanning domains stand out rather clearly (e.g. typically >1.6 on the Kyte-Doolittle scale) (Figure 2.18).

URL link: <http://web.expasy.org/protscale/>

In this study, the raw FASTA sequence (excluding the header) of the target protein was pasted on to the input window. Next the parameters were set. The amino acid scale selected was Hphob. / Kyte & Doolittle (Kyte and Doolittle, 1982) with the window size set to 19 as it is the best window value for detection of transmembrane regions. The rest of the parameters were set at default.

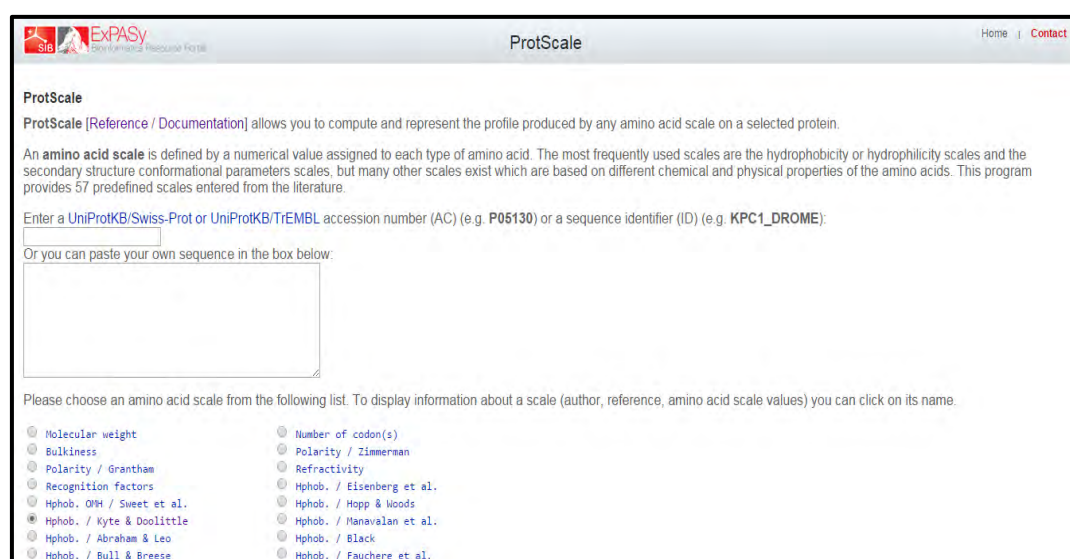


Figure 2.18: ProtScale homepage

2.2.2.8 TMHMM:

The TMHMM (Krogh *et al.*, 2001) server is an online tool which is available at the Center for Biological Sequence Analysis (CBS) prediction servers. It is capable of predicting the membrane protein topologies of protein sequences. It can clearly predict transmembrane helices and capable of discriminating between soluble and membrane proteins with both sensitivity and specificity. In this study, TMHMM server version 2.0 was used. The FASTA sequence of the query protein was pasted onto the input section of the server and rest of the parameters were set at default (Figure 2.19).

URL link: <http://www.cbs.dtu.dk/services/TMHMM/>

Figure 2.19: TMHMM server homepage

2.2.2.9 Self Optimized Prediction Method (SOPMA):

Self Optimized Prediction Method (SOPMA) tool (Geourjon and Deléage, 1995) is an online tool which is available at the Network Protein Sequence Analysis (NPS@) server (Combet *et al.*, 2000). It allows the user to predict the secondary structure of proteins. It correctly predicts 69.5% of amino acids for a three-state description of the secondary structure (alpha-helix, beta-sheet and coil) in a whole database containing 126 chains of non-homologous (less than 25% identity) proteins. The NPS@ server is an interactive Web server dedicated to protein sequence analysis and available for the biologist community (Figure 2.20).

URL link for NPS@ server: <http://npsa-devel.ibcp.fr/>.

URL link for SOMPA tool: https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_sopma.html

In this current study, to attain quantitative values for the amount of alpha-helices, beta sheets and coils present within the amino acid stretch of the protein the SOPMA tool was used. The FASTA sequence of the query protein was pasted onto the input window and the results attained were recorded.

SOPMA SECONDARY STRUCTURE PREDICTION METHOD

[\[Abstract\]](#) [\[NPS@ help\]](#) [\[Original server\]](#)

Sequence name (optional) :

Paste a protein sequence below : [help](#)

Output width :

Parameters

Number of conformational states :

Similarity threshold :

Window width :

Figure 2.20: SOPMA homepage

2.2.2.10 CYS_REC:

The CYS_REC tool (<http://www.softberry.com/berry.phtml>) is an online program that is used to identify SS-bonding states of cysteines and location of disulphide bridges of proteins. It is available at the SoftBerry website under the protein structure analysis section. In this study, the query protein FASTA sequence was entered and then computed. The results attained were then recorded and analyzed (Figure 2.21).

URL link:

http://linux1.softberry.com/berry.phtml?topic=cys_rec&group=programs&subgroup=propt

2.2.2.11 PSIPRED:

PSIPRED (Buchan *et al.*, 2013, Jones, 1999) is an easy and accurate secondary structure prediction method. It incorporates two feed-forward neural networks which perform an analysis on output obtained from PSI-BLAST (Position Specific Iterated - BLAST). The PSIPRED Protein Sequence Analysis Workbench aggregates several UCL structure prediction methods into one location. It can predict a protein's secondary structure (beta sheets, alpha helices and coils) from the primary sequence. The users can submit a protein sequence, perform the predictions of their choice and

receive the results of the prediction via e-mail or the web. In this current study the FASTA sequence of the query protein was submitted to attain a graphical representation of the secondary structure of the target protein. The results attained were then analyzed (Figure 2.22).

URL link: <http://bioinf.cs.ucl.ac.uk/psipred/>

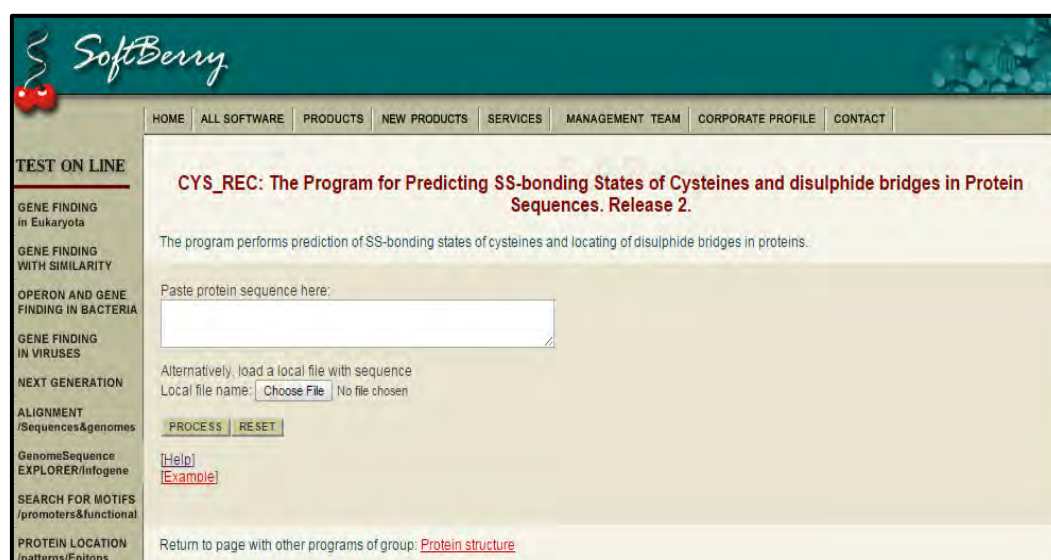


Figure 2.21: CYC_REC homepage at the SoftBerry website

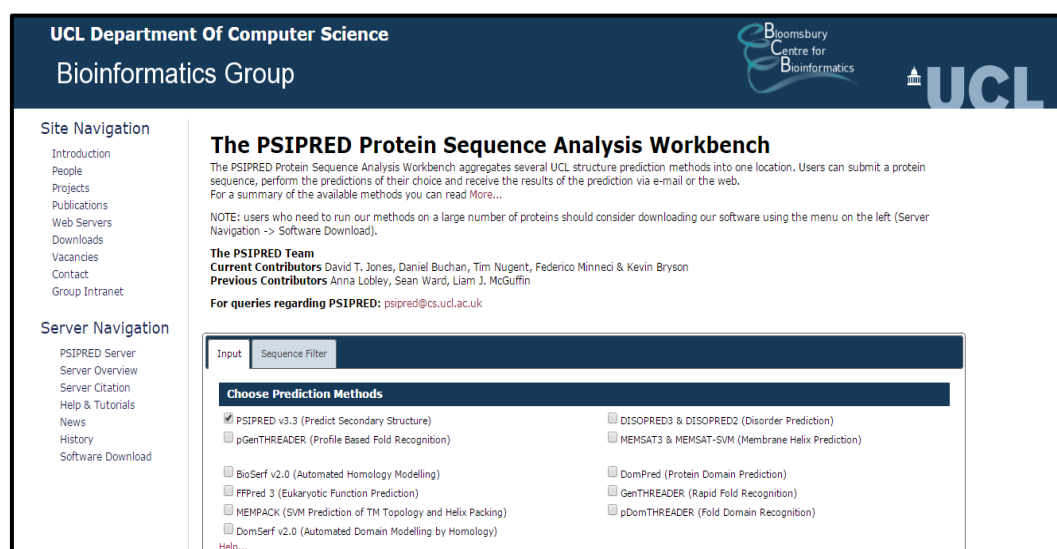


Figure 2.22: PSIPRED homepage

2.2.2.12 Iterative Threading ASSEMBLY Refinement (I-TASSER):

I-TASSER (Roy *et al.*, 2010, Yang *et al.*, 2015, Zhang, 2008) is an online server, which is a hierarchical method for protein structure and function prediction from

amino acid sequences. It detects structural templates from PDB by a process called fold recognition/threading. Full-length atomic models are generated, using the templates, by iterative template fragment assembly simulations. The server's main goal is to produce the most accurate structure and function predictions using state-of-the-art algorithms (Figure 2.23).

URL link: <http://zhanglab.ccmb.med.umich.edu/I-TASSER/>

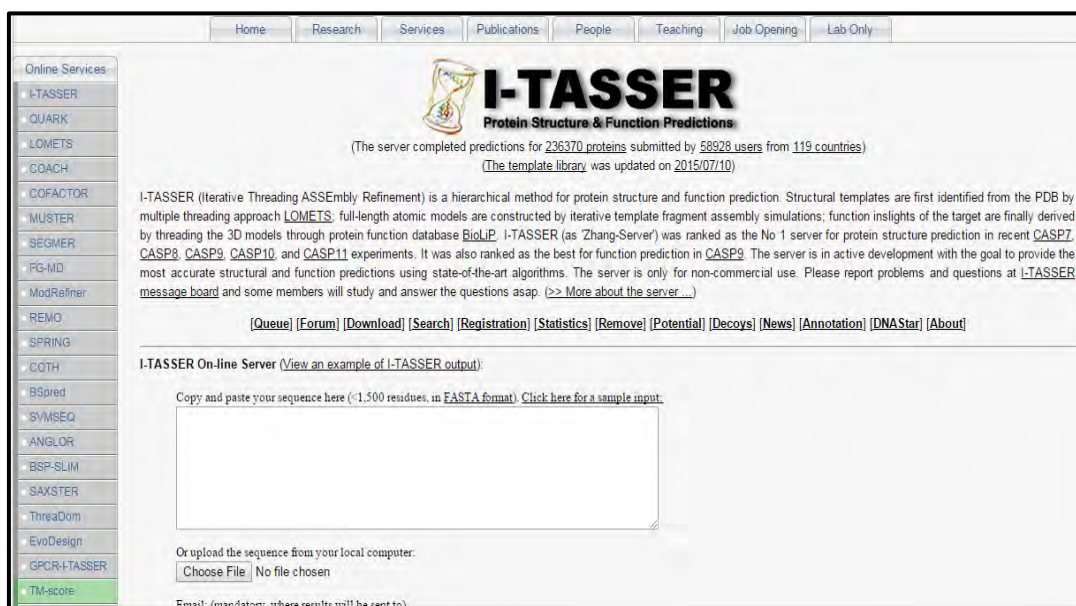


Figure 2.23: I-TASSER homepage

2.2.2.13 Protein Homology/analogy Recognition Engine V 2.0 (Phyre2):

Phyre2 (Kelley *et al.*, 2015) is a suite of tools available on the web which is used to predict and analyze protein structure, function and mutations. It provides biologists with an easy and insightful interface to the state-of-the-art protein bioinformatics tools. Phyre2 replaces Phyre which is the original version of the server. It uses advanced remote homology detection methods to build 3D models, predict ligand binding sites as well as several other features for the user's protein sequence. In this study, the FASTA sequence of the query protein was entered and the intensive mode was selected to attain 3D models (Figure 2.24).

URL link: <http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index>



Figure 2.24: Phyre2 homepage

2.2.2.14 EasyModeller 4.0:

EasyModeller 4.0 (Kuntal et al., 2010) is a graphic user interface (GUI) for homology modeling using MODELLER (Eswar *et al.*, 2006, Fiser *et al.*, 2000, Martí-Renom *et al.*, 2000, \leq ali, 1995, Sali, 1995) in the backend and is available for both Windows and Linux platforms. To generate models using EasyModeller software, the user should have MODELLER and Python preinstalled (Figure 2.25 a).

An experimental work plan of the steps that were taken to utilize EasyModeller 4.0 is shown in Figure 2.3.

URL Link: <http://modellergui.blogspot.com/>

2.2.2.14.1 Steps taken using EasyModeller:

The FASTA sequence of the query protein was uploaded onto the EasyModeller interface. Next the structural templates attained via blast P-suite were downloaded from the RCSB PDB and uploaded to the software. The single template to be used for homology modeling was selected via comparison between the templates. It was selected based on sequence identity to the query protein and crystallographic

resolutions. Next the selected template was aligned with the query protein and afterwards five models were generated where the input parameters were set at default. The best models were then selected based on molpdf values, DOPE scores and GA341 values (Figure 2.25 b-c).

2.2.2.15 PROCHECK:

PROCHECK (Laskowski *et al.*, 1993, Laskowski *et al.*, 1996) is a downloadable software available at the EMBL-EBI website which checks the stereochemical quality of a protein structure. It produces a number of PostScript plots analyzing the protein's overall and residue-by-residue geometry. The PROCHECK tool provides the user with Ramachandran plots (Lovell *et al.*, 2003, Ramachandran *et al.*, 1963) which assesses and evaluates the protein PDB coordinate models (Figure 2.26).

URL link: <http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/>

In this current study, the PROCHECK web server available at the PDBsum Generate section of the PDBsum server (de Beer *et al.*, 2013) was used to assess and evaluate the homology models of the query protein attained from various homology modeling online tools and software as discussed in the previous sections (Figure 2.27).

URL link: <https://www.ebi.ac.uk/thornton-srv/databases/pdbsum/Generate.html>

2.2.2.16 UCSF Chimera

Molecular graphics and analyses were performed with the UCSF Chimera package (Pettersen *et al.*, 2004). Chimera is developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIGMS P41-GM103311). It is a highly extensible program for interactive visualization and analysis of molecular structures and related data. High quality animations and images can be produced by this tool. It can be downloaded from the UCSF Chimera website. In this current study Chimera Version 1.10.1 was used (Figure 2.28).

URL link: <http://www.cgl.ucsf.edu/chimera/>

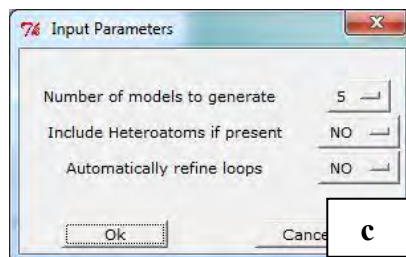
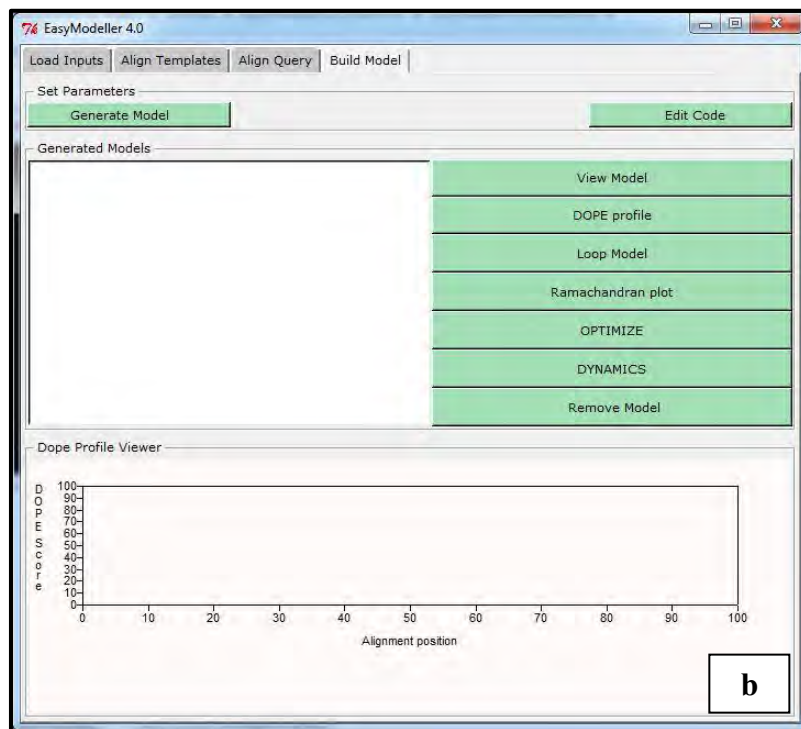
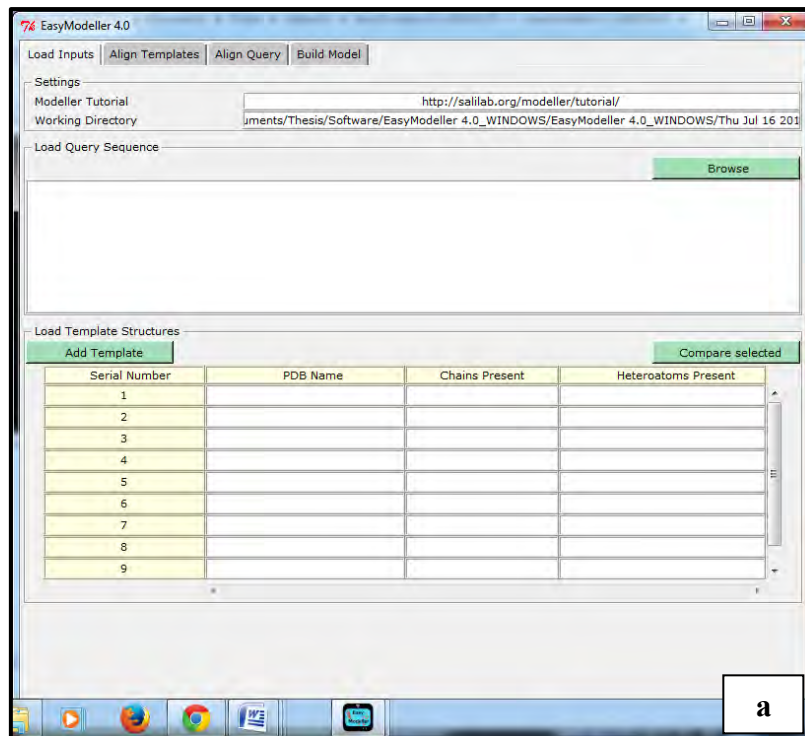


Figure 2.25: (a) EasyModeller 4.0 GUI, (b) Model generation GUI on EasyModeller 4.0 and (c) Input parameters

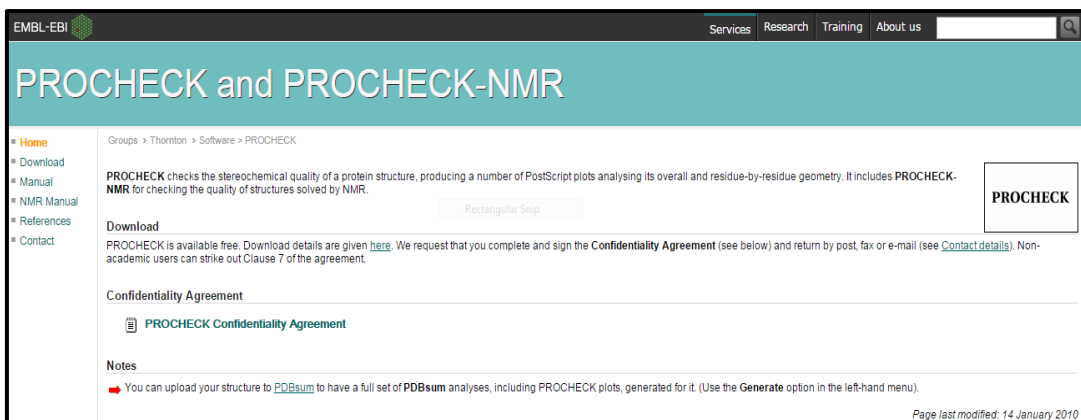


Figure 2.26: PROCHECK homepage

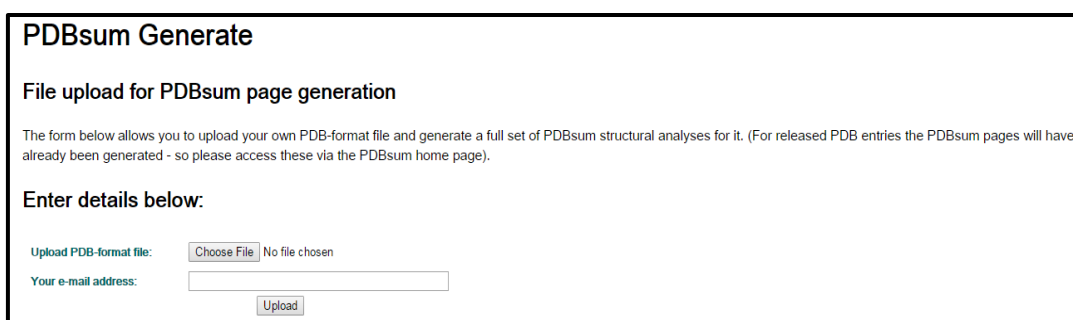


Figure 2.27: PDBsum Generate homepage

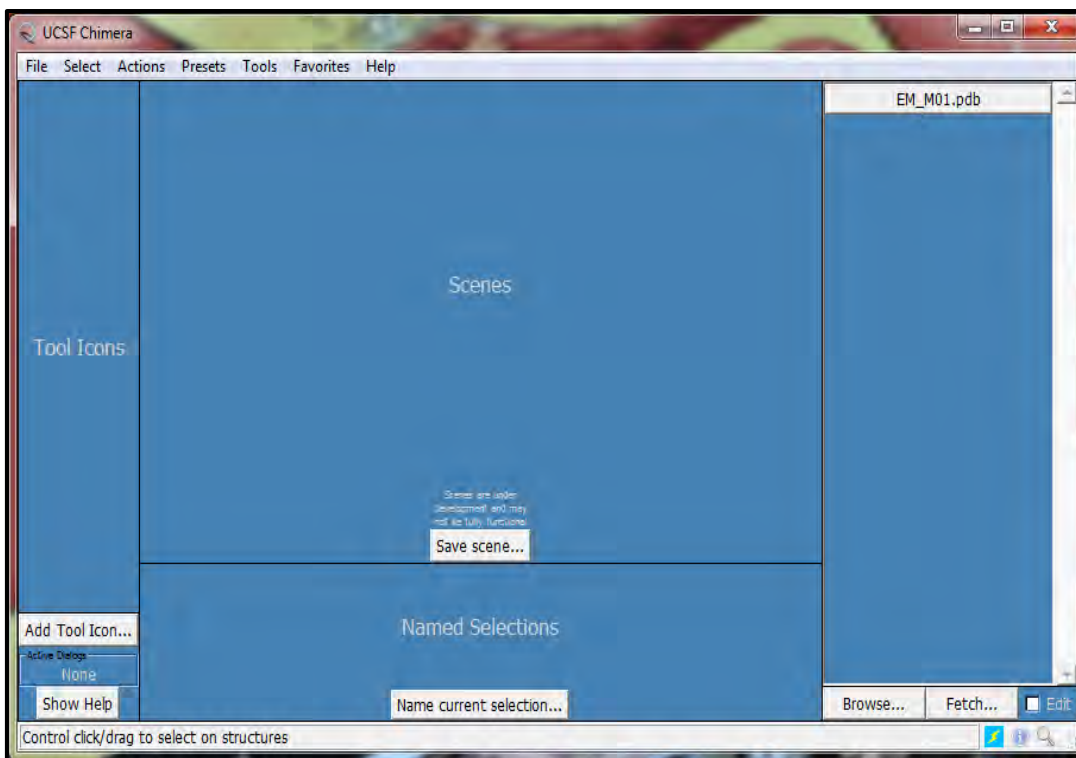
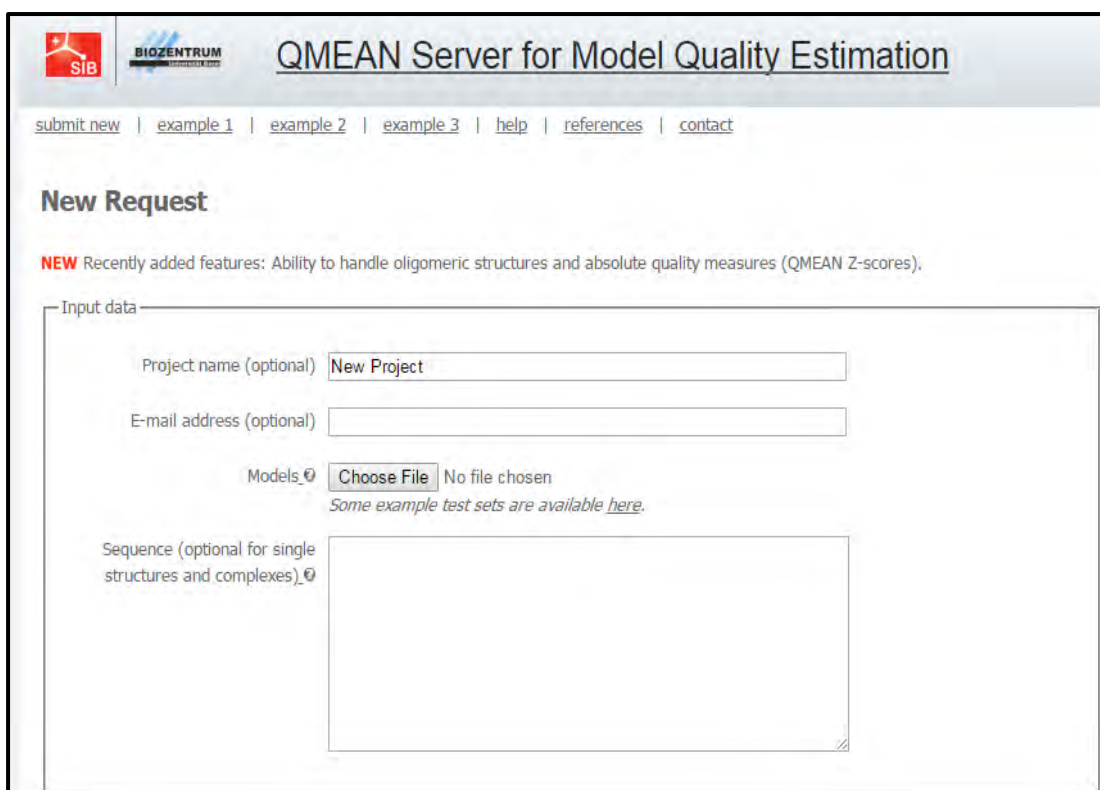


Figure 2.28: Chimera GUI

2.2.2.17 QMEAN server

The QMEAN server (Benkert *et al.*, 2009) provides access to scoring functions for the quality estimation of protein structure models. This allows the models to be ranked and also identify the potentially unreliable regions within the proteins. The QMEAN (Benkert *et al.*, 2008) is a composite scoring function which is capable of deriving both global (i.e. for the entire structure) and local (i.e. per residue) error estimates on the basis of a single protein model. The QMEAN Z-score (Benkert *et al.*, 2011) provides an estimate of the absolute quality of a model by relating it to reference structures already determined by X-ray crystallography. It is an approximation of the “degree of nativeness” of the structural features observed in a model by describing the likelihood that a model is of comparable quality to high-resolution experimental structures. It is an online server which is part of the ExpASy resource portal. In this current study, the models were uploaded and the results attained were analyzed (Figure 2.29).

URL link: <http://swissmodel.expasy.org/qmean/cgi/index.cgi>



The screenshot displays the QMEAN Server homepage. At the top, there are logos for SIB and BIOZENTRUM, followed by the title "QMEAN Server for Model Quality Estimation". Below the title is a navigation menu with links: [submit new](#), [example 1](#), [example 2](#), [example 3](#), [help](#), [references](#), and [contact](#). The main heading is "New Request". A red "NEW" banner indicates recently added features: "Ability to handle oligomeric structures and absolute quality measures (QMEAN Z-scores)". The "Input data" section contains several form fields: "Project name (optional)" with the value "New Project", "E-mail address (optional)", "Models" with a "Choose File" button and the text "No file chosen" and a link to "Some example test sets are available here.", and "Sequence (optional for single structures and complexes)" with a large text area.

Figure 2.29: QMEAN Server homepage

CHAPTER 3:
RESULTS AND DISCUSSION

Chapter 3: Results and Discussion

3.1 Target FASTA Sequences retrieved from the NCBI databases:

The available antiporter gene was analyzed with the initial target source of *Arabidopsis*. The coding sequence of the antiporter gene and protein (AtNHX1) were retrieved using the NCBI databases. The FASTA sequences attained are as follows:

3.1.1 Nucleotide FASTA sequence for the target gene:

The target nucleotide FASTA sequence of the *Arabidopsis thaliana* sodium/hydrogen exchanger 1 mRNA, complete *cds* attained:

>gi|30690553:471-2087 Arabidopsis thaliana sodium/hydrogen exchanger 1 mRNA, complete cds

```
ATGTTGGATTCTCTAGTGTCGAAACTGCCTTCGTTATCGACATCTGATCAC
GCTTCTGTGGTTGCGTTGAATCTCTTTGTTGCACTTCTTTGTGCTTGTATTG
TTCTTGGTCATCTTTTGGAAAGAGAATAGATGGATGAACGAATCCATCACC
GCCTTGTTGATTGGGCTAGGCACTGGTGTACCATTTTGTTGATTAGTAAA
GGAAAAAGCTCGCATCTTCTCGTCTTTAGTGAAGATCTTTTCTTCATATAT
CTTTTGCCACCCATTATATTC AATGCAGGGTTTCAAGTAAAAAAGAAGCA
GTTTTTCCGCAATTCGTGACTATTATGCTTTTTGGTGCTGTTGGGACTATT
ATTTCTGCACAATCATATCTCTAGGTGTAACACAGTTCTTTAAGAAGTTG
GACATTGGAACCTTTGACTTGGGTGATTATCTTGCTATTGGTGCCATATTT
GCTGCAACAGATTCAGTATGTACACTGCAGGTTCTGAATCAAGACGAGAC
ACCTTTGCTTTACAGTCTTGTATTCGGAGAGGGTGTGTGAATGATGCAAC
GTCAGTTGTGGTCTTCAACGCGATTCAGAGCTTTGATCTCACTCACCTAAA
CCACGAAGCTGCTTTTCATCTTCTTGGA AACTTCTTGTATTTGTTTCTCCTA
AGTACCTTGCTTGGTGCTGCAACCGGTCTGATAAGTGCGTATGTTATCAAG
AAGCTATACTTTGGAAGGCACTCAACTGACCGAGAGGTTGCCCTTATGAT
GCTTATGGCGTATCTTTCTTATATGCTTGCTGAGCTTTTCGACTTGAGCGGT
ATCCTCACTGTGTTTTTCTGTGGTATTGTGATGTCCCATTACACATGGCAC
AATGTAACGGAGAGCTCAAGAATAACAACAAAGCATAACCTTTGCAACTTT
GTCATTTCTTGCGGAGACATTTATTTTCTTGTATGTTGGAATGGATGCCTTG
GACATTGACAAGTGGAGATCCGTGAGTGACACACCGGGAACATCGATCGC
```

AGTGAGCTCAATCCTAATGGGTCTGGTCATGGTTGGAAGAGCAGCGTTTCG
TCTTCCGTTATCGTTTCTATCTAACTTAGCCAAGAAGAATCAAAGCGAGA
AAATCAACTTTAACATGCAGGTTGTGATTTGGTGGTCTGGTCTCATGAGAG
GTGCTGTATCTATGGCTCTTGCATACAACAAGTTTACAAGGGCCGGGCAC
ACAGATGTACGCGGGAATGCAATCATGATCACGAGTACGATAACTGTCTG
TCTTTTTAGCACAGTGGTGTGGTATGCTGACCAAACCACTCATAAGCTA
CCTATTACCGCACCAGAACGCCACCACGAGCATGTTATCTGATGACAACA
CCCCAAAATCCATACATATCCCTTTGTTGGACCAAGACTCGTTCATTGAGC
CTTCAGGGAACCACAATGTGCCTCGGCCTGACAGTATACGTGGCTTCTTGA
CACGGCCCACTCGAACCGTGCATTACTACTGGAGACAATTTGATGACTCCT
TCATGCGACCCGTCTTTGGAGGTCGTGGCTTTGTACCCTTTGTTCCAGGTT
CTCCAAGTACGAGAGAAACCCTCCTGATCTTAGTAAGGCTTGA

3.1.2 FASTA sequence for the target protein:

The target protein FASTA sequence of sodium/hydrogen exchanger 1 [Arabidopsis thaliana] attained:

>gi|15240448|ref|NP_198067.1| sodium/hydrogen exchanger 1 [Arabidopsis thaliana]

MLDSLVS KLPSLSTSDHASVVALNLFVALLCACIVLGHLLLEENRWMNESITAL
LIGLGTGVTILLISKGKSSHLLVFESEDLFFIYLLPPIIFNAGFQVKKKQFFRNFT
IMLFGAVGTIISCTIISLGVTQFFKKLDIGTFDLGDYLAIGAIFAATDSVCTLQVL
NQDETPLLYSLVFGEGVVNDATSVVVFNAIQSFDLTHLNHEAAFHLLGNFLY
LFLLSTLLGAATGLISAYVIKKLYFGRHSTDREVALMMLMAYLSYMLAELFD
LSGILTVFFCGIVMSHYTWHNVTESSRITTKHTFATLSFLAETFIFLYVGM DAL
DIDKWRVS DTPGTSIAVSSILMGLVMVGRAAFVFPLSFLSNLAKKNQSEKIN
FNMQVVIWWSGLMRGAVSMALAYNKFTRAGHTDVRGNAIMITSTITVCLFS
TVVFGMLTKPLISYLLPHQNATTSMLSDDNTPKSIHIPLLDQDSFIEPSGNHNV
PRPDSIRGFLTRPTRTVHYYWRQFDDSFMRPVFGGRGFVPFVPGSPTERNPPD
LSKA

3.2 Nucleotide Analysis:

3.2.1 BLAST results of the nucleotide sequence:

The target nucleotide sequence was blasted using the NCBI blast N-suite. The megablast option for highly similar sequences was selected. A graphical summary of the BLAST results was attained.

A visual representation of the BLAST results for the target nucleotide sequence was attained where the top red streak/bar indicated the query sequence (Figure 3.1). Each bar represented a portion of another sequence that is similar to the query sequence along with the region of the sequence where the similarity occurs. The red bars indicated highly similar sequences which were valid as the parameters to search only for the highly similar sequences were set.

A list of sequences that produced significant alignments were retrieved (Figure 3.2). From the list, 21 sequences along with the query sequence based on their identity values and E-values were selected. The range set, to be used for sequence selection, for identity values was 80% to 100%. This means that the selected sequences had a genomic configuration that was 80% to 100% identical to that of the query sequence. The E-values are the expected values. It can be defined as the number of times the database match may have occurred randomly. It provides us with a criterion which is more objective than that of the percentage-of-similarity (identity values) (Claverie and Notredame, 2006). As such, a good match would be considered one that is highly unlikely to occur just by chance. Therefore, the sequences that had low E-values (especially $<10^{-4}$) were selected. Since lower the E-value, higher the authenticity of the selected sequence.

To utilize the sequences in a suitable and easy manner the names of the sequence identifiers were altered so that they could be easily recognized. The sequences that were selected were categorized according to organism, accession ID and the corresponding altered name and then tabulated (Table 3.1).

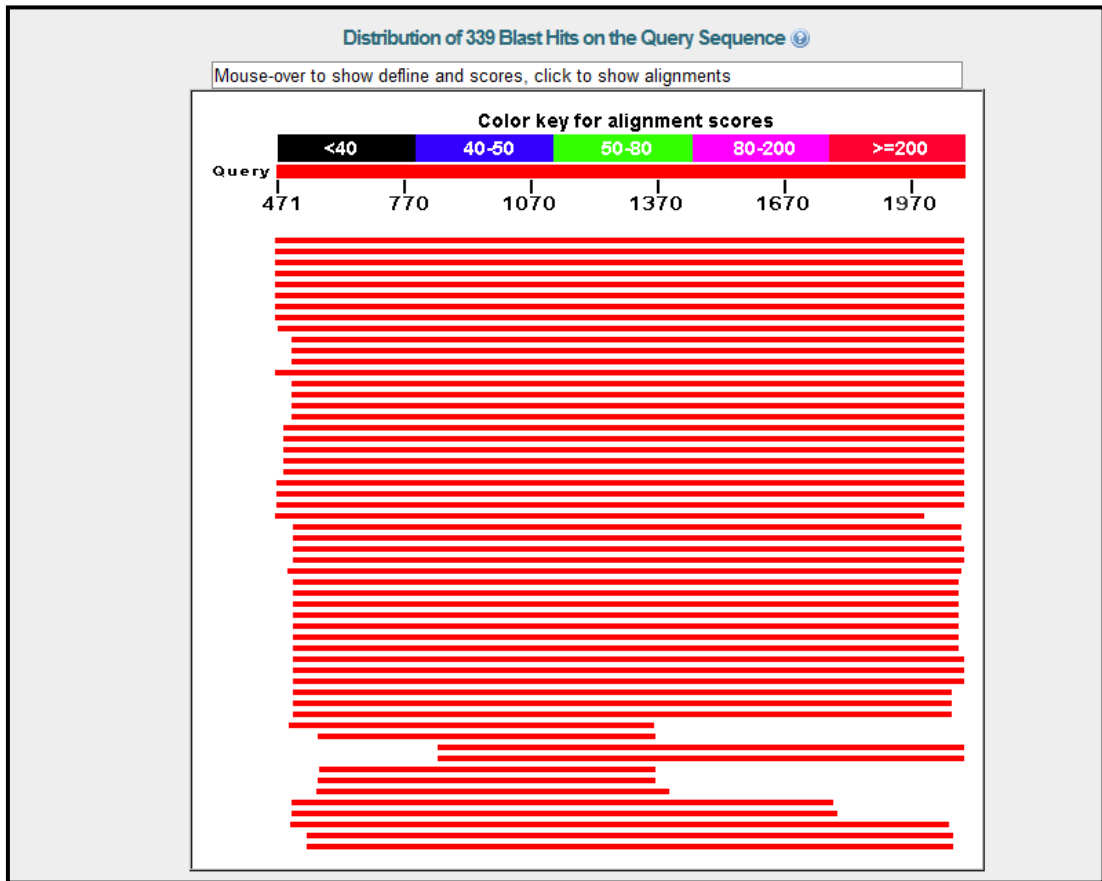


Figure 3.1: A graphical summary of the BLAST results using blast N-suite for the target nucleotide sequence

Sequences producing significant alignments:

Select: All None Selected: 0

Alignments Download GenBank Graphics Distance tree of results

Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/> Arabidopsis thaliana mRNA for Na+/H+ exchanger, complete cds, clone: RAF1.07-14-P04	2987	2987	100%	0.0	100%	AK226586.1
<input type="checkbox"/> Arabidopsis thaliana sodium/hydrogen exchanger 1 mRNA, complete cds	2987	2987	100%	0.0	100%	NM_122597.2
<input type="checkbox"/> Arabidopsis thaliana sodium proton exchanger Nhx1 mRNA, partial cds	2981	2981	99%	0.0	100%	AF106324.1
<input type="checkbox"/> Olimarabidop [Go to alignment for Arabidopsis thaliana sodium proton exchanger Nhx1 mRNA, partial cds]	2976	2976	100%	0.0	99%	JF357965.1
<input type="checkbox"/> Arabidopsis thaliana Na+/H+ antiporter mRNA, complete cds	2976	2976	100%	0.0	99%	EF596738.1
<input type="checkbox"/> Arabidopsis thaliana Na+/H+ exchanger (NHX1) mRNA, complete cds	2976	2976	100%	0.0	99%	AF056190.1
<input type="checkbox"/> Arabidopsis thaliana sodium proton exchanger (NHX1) mRNA, complete cds	2964	2964	100%	0.0	99%	AY685183.1
<input type="checkbox"/> Arabidopsis thaliana Na+/H+ antiporter (NHX1) mRNA, complete cds	2964	2964	100%	0.0	99%	AF510074.1
<input type="checkbox"/> Arabidopsis lyrata subsp. lyrata hypothetical protein, mRNA	2638	2638	99%	0.0	96%	XM_002874357.1
<input type="checkbox"/> Capsella rubella hypothetical protein (CARUB_v10000654mg) mRNA, complete cds	2460	2460	97%	0.0	95%	XM_006287383.1
<input type="checkbox"/> PREDICTED: Camelina sativa sodium/hydrogen exchanger 1-like (LOC104771239), transcript variant X2, mRNA	2422	2422	97%	0.0	94%	XM_010495738.1
<input type="checkbox"/> PREDICTED: Camelina sativa sodium/hydrogen exchanger 1-like (LOC104771239), transcript variant X1, mRNA	2422	2422	97%	0.0	94%	XM_010495737.1
<input type="checkbox"/> Olimarabidopsis pumila NHX-like protein (NHX1) mRNA, complete cds	2410	2410	100%	0.0	94%	KC200248.1

Figure 3.2: A segment of nucleotide sequences that produced significant alignments

Table 3.1: List of nucleotide sequences selected and categorized according to organism, accession ID and corresponding altered name where A.thalQ is the query sequence

Organisms	Accession ID	Altered name
<i>Arabidopsis thaliana</i>	NM_001084641.1	A.thal1
<i>Arabidopsis thaliana</i>	NM_122597.2	A.thalQ
<i>Arabidopsis thaliana</i>	HE802897.1	A.thalEco
<i>Arabidopsis thaliana</i>	EF596738.1	A.thal2
<i>Arabidopsis thaliana</i>	AF510074.1	A.thal3
<i>Arabidopsis thaliana</i>	AK226586.1	A.thal4
<i>Eutrema halophilum</i>	FJ713100.1	E.halo1
<i>Eutrema halophilum</i>	DQ995339.1	E.halo2
<i>Eutrema halophilum</i>	DQ490966.1	E.halo3
<i>Eutrema salsugineum</i>	XM_006394928.1	E.sal1
<i>Eutrema salsugineum</i>	XM_006394927.1	E.sal2
<i>Capsella rubella</i>	XM_006299101.1	Cap.rubella1
<i>Capsella rubella</i>	XM_006287383.1	Cap.rubella2
<i>Cenchrus americanus</i>	HQ283439.1	P.glaucum1
<i>Cenchrus americanus</i>	DQ228817.1	P.glaucum2
<i>Olimarabidopsis pumila</i>	JF357965.1	O.pumila
<i>Thlaspi arvense</i>	JQ435892.1	T.arvense
<i>Brassica juncea</i>	HQ848294.1	B.juncea
<i>Brassica napus</i>	GU192449.1	B.napus
<i>Cochlearia anglica</i>	JQ435894.1	C.anglica
<i>Populus trichocarpa</i>	AC210556.1	Pop.trichocarpa

3.2.2 Multiple Sequence Alignment (MSA) results using Clustal Omega:

Sequence alignment of the selected 21 sequences was performed using Clustal Omega. The multiple sequence alignment was performed so as to rewrite the selected sequences in manner so that the similar features end up in the same columns. The objective behind the multiple sequence alignment is to put nucleotides in the same column since they are similar based on a particular criterion such as: structural similarity, evolutionary similarity, functional similarity and sequence similarity (Claverie and Notredame, 2006). However, in this study, the concern was sequence similarity. This means that the nucleotides that were in the same columns yielded alignments with maximum similarity. Since, the sequences are closely related, on the basis of the selected sequences attained from the BLAST output which have high identity values, it can be deduced that their functional, evolutionary and structural similarities are equivalent to that of the sequence similarity. This indicated that the sequences similar to that of the coding sequence of the AtNHX1 gene would probably encode for proteins that have structures and functions similar to that of the antiporter protein.

3.2.3 Phylogenetic tree generation by MEGA 6:

A bootstrap consensus phylogenetic tree with node statistics for the aligned 21 sequences was attained via the MEGA 6 software (Figure 3.3). The MSA file was downloaded from the Clustal Omega tool and analyzed using the MEGA 6 software. Based on the MEGA 6 analysis output, for the 21 sequences, there were 2272 conserved sites and 3371 variable sites over a span of 94004 sites (Figures 3.4 - 3.5).

The Neighbor-Joining method (Saitou and Nei, 1987) was implemented to generate the phylogenetic tree. It is based on evolutionary distance data. As mentioned earlier, bootstrapping was performed as the neighbor-joining method does not have any clade support measure. The node statistics represent the bootstrap percentage values for each node. The clades were poorly supported based on the node statistics (<80). Most of the sequences were highly diverged from the query sequence i.e. A.thalQ. This suggested that most of them were distantly genetically related. The closest sequences to A.thalQ were C.anglica and B.juncea followed by Pop.trichocarpa as they had the least distance from the query.

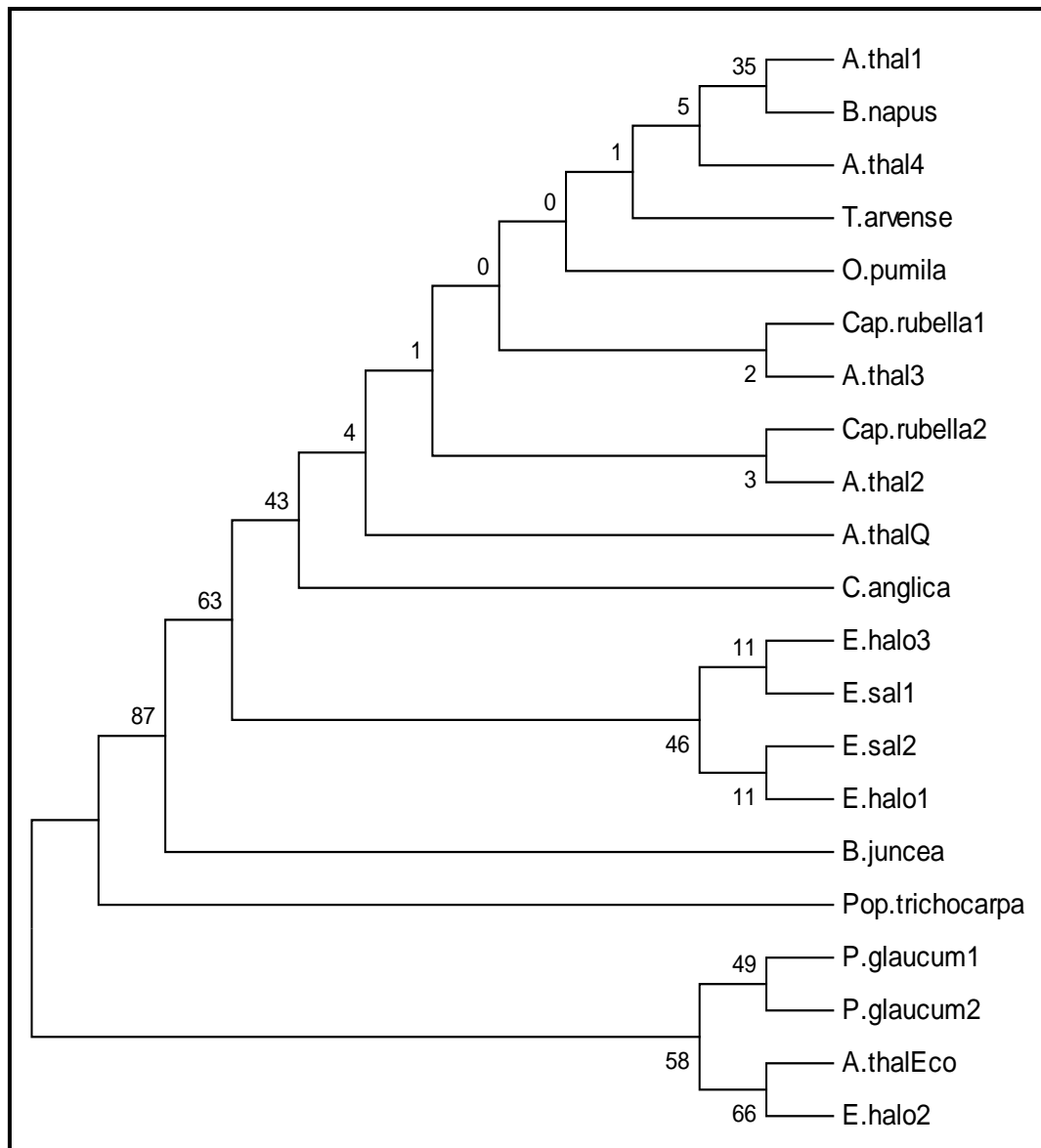


Figure 3.3: Bootstrap consensus tree with node statistics for the selected twenty one nucleotide sequences along with the query sequence

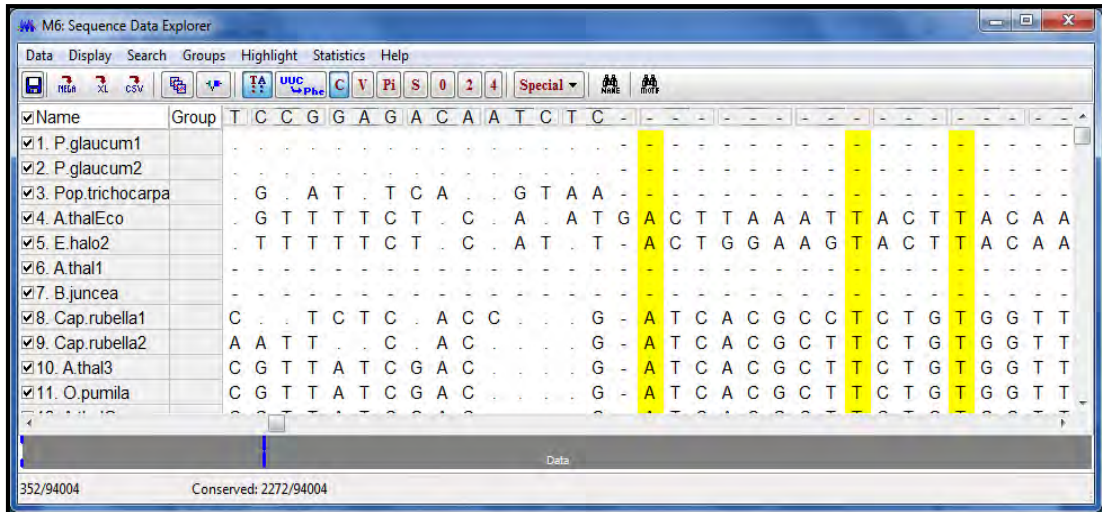


Figure 3.4: Number of conserved sites in the selected 21 nucleotide sequences

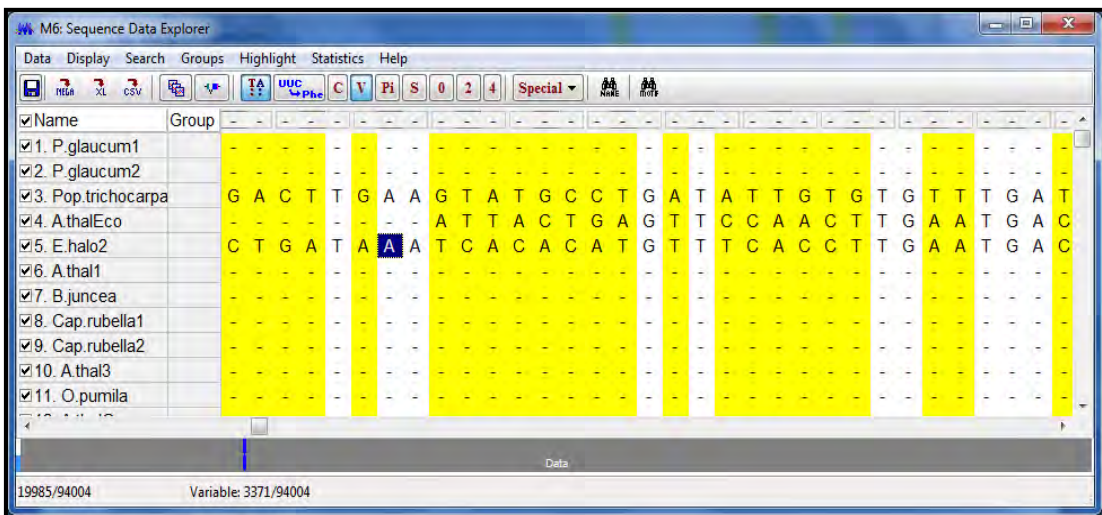


Figure 3.5: Number of variable sites in the selected 21 nucleotide sequences

3.3 Protein analysis

3.3.1 BLAST results of the protein sequence:

The target protein sequence was blasted using the NCBI blast P-suite. A graphical summary of the BLAST results was attained. A visual representation of the BLAST results for the target nucleotide sequence was attained where the top red streak/bar indicated the query protein sequence. The red bars indicated highly similar sequences amongst different species of plants (Figure 3.6).

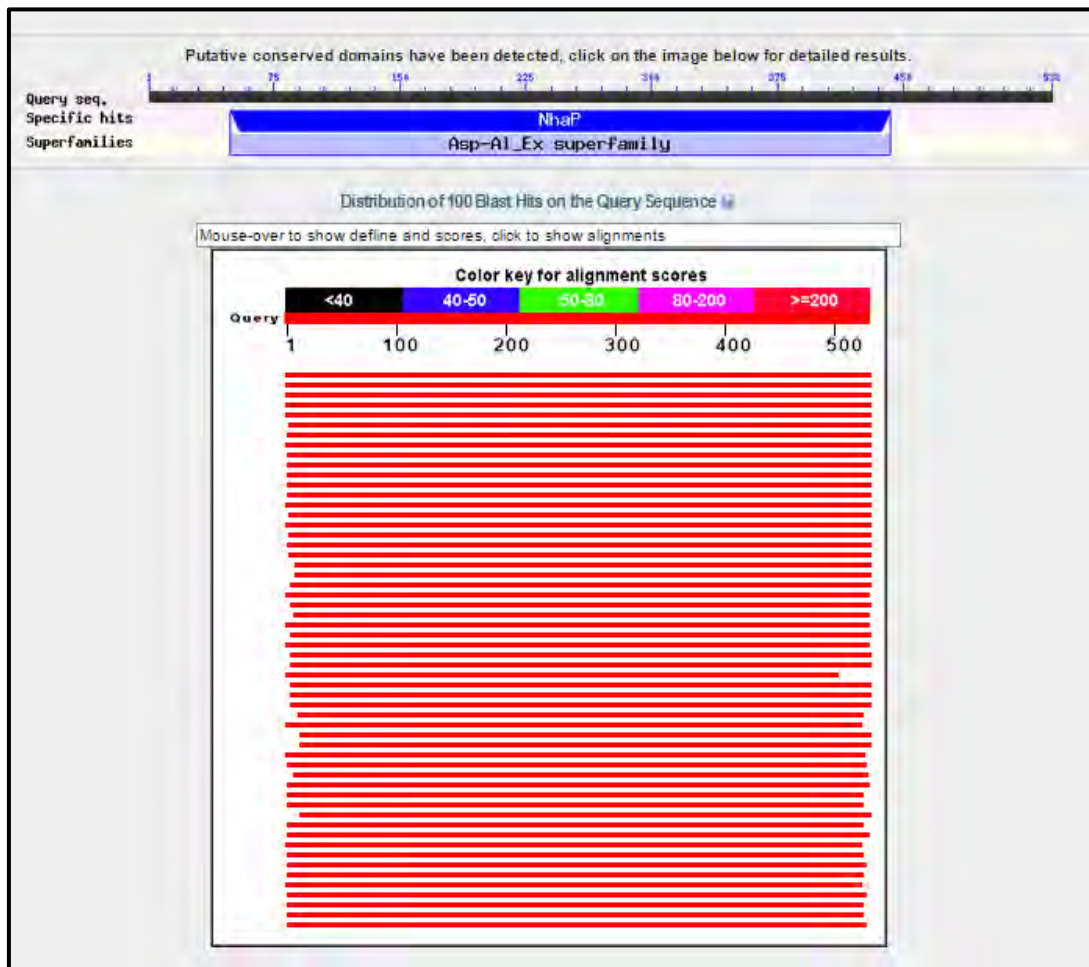


Figure 3.6: A graphical summary of the BLAST results using blast P-suite for the target protein sequence

A list of sequences that produced significant alignments were retrieved (Figure 3.7). A total of 100 sequences were present within the list. The sequences in the list were cross-checked with the selected sequences from the nucleotide analysis data (Table 3.1). This was done so that the subsequent steps could be undertaken using the same sequences. Therefore, a valid comparison could be conducted between the two sets of data. The sequences were selected based on identity values (ranging from 80% to 100%) and low E-values ($< 10^{-4}$). In total seven sequences (including the query sequence) were selected that matched with the nucleotide analysis data. The selected protein sequences were then tabulated and categorized according to organism, protein accession ID, corresponding nucleotide accession ID and the matching altered name (Table 3.2).

Sequences producing significant alignments:

Select: [All](#) [None](#) Selected: 0

Alignments [Download](#) [GenPept](#) [Graphics](#) [Distance tree of results](#) [Multiple alignment](#)

Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/> sodium/hydrogen exchanger 1 [Arabidopsis thaliana]	1097	1097	100%	0.0	100%	NP_198067.1
<input type="checkbox"/> Na+/H+ antiporter [Arabidopsis thaliana]	1094	1094	100%	0.0	99%	ABQ58865.1
<input type="checkbox"/> Na+/H+ antiporter [Arabidopsis thaliana]	1094	1094	100%	0.0	99%	AAM34759.1
<input type="checkbox"/> sodium proton exchanger [Arabidopsis thaliana]	1093	1093	100%	0.0	99%	AAT95387.1
<input type="checkbox"/> Na+/H+ antiporter-like protein [Olimarabidopsis pumila]	1092	1092	100%	0.0	99%	AEA51351.1

Figure 3.7: A segment of the list of protein sequences that produced significant alignments

Table 3.2: List of selected protein sequences categorized according to organism, protein accession ID, corresponding nucleotide accession ID and the matching altered name

SL	Organisms	Protein Accession ID	Corresponding Nucleotide Accession ID	Matching Altered name
1	<i>Arabidopsis thaliana</i>	NP_198067.1	NM_122597.2	A.thalQ
2	<i>Arabidopsis thaliana</i>	AAM34759.1	AF510074.1	A.thal3
3	<i>Olimarabidopsis pumila</i>	AEA51351.1	JF357965.1	O.pumila
4	<i>Capsella rubella</i>	XP_006287445.1	XM_006287383.1	Cap.rubella2
5	<i>Eutrema salsugineum</i>	XP_006394990.1	XM_006394928.1	E.sal1
6	<i>Brassica napus</i>	ACZ92142.1	GU192449.1	B.napus
7	<i>Eutrema halophilum</i>	ABF48496.1	DQ490966.1	E.halo3

3.3.2 Multiple Sequence Alignment (MSA) results using Clustal Omega:

Multiple sequence alignment (MSA) of the seven selected sequences was performed using Clustal Omega. The Clustal Omega alignment file was imported into the BoxShade sequence alignment editor. The identical or similar amino acids were shaded using the aforesaid tool and an output file was downloaded (Figure 3.8).

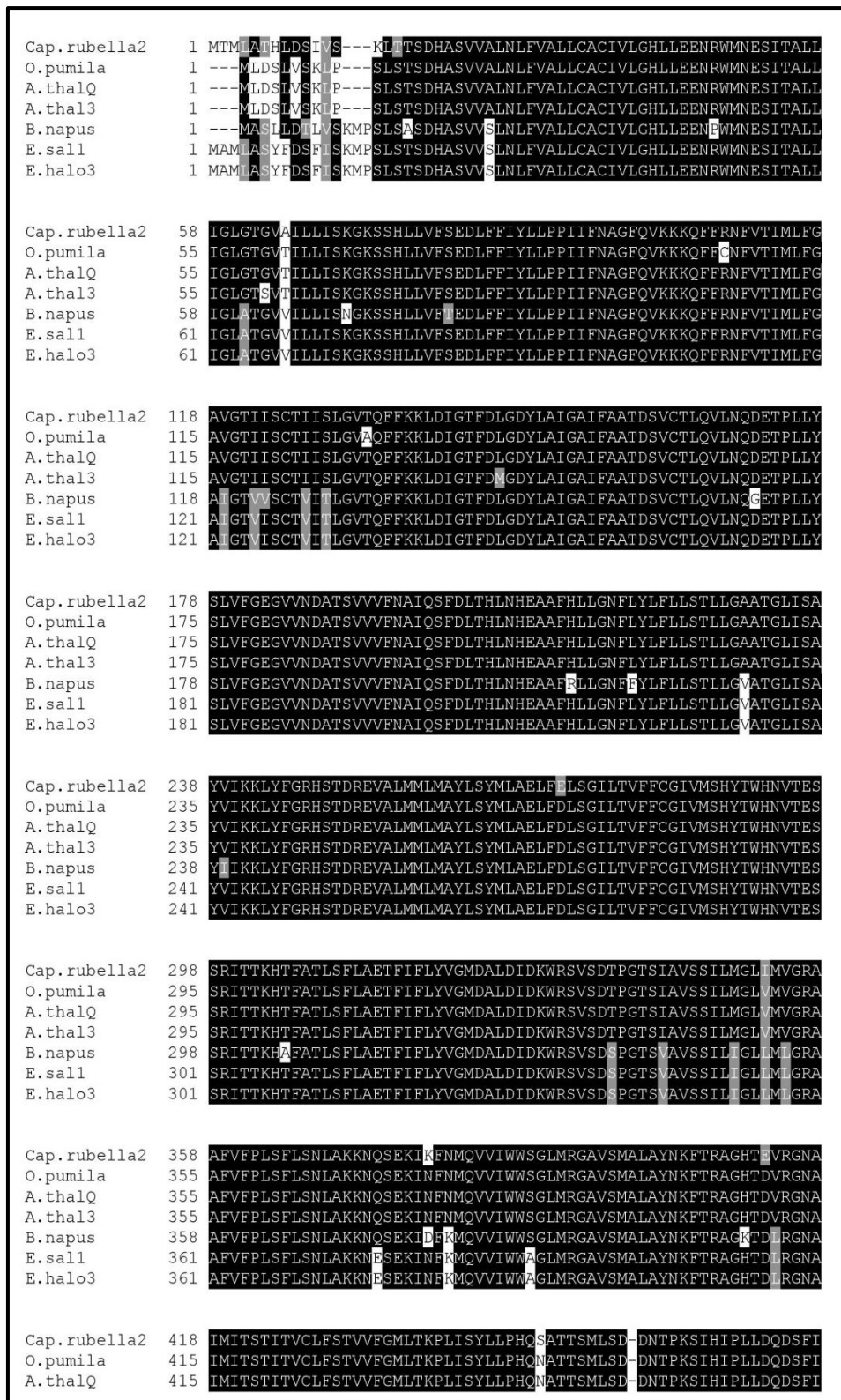


Figure 3.8: Multiple sequence alignment of the seven selected protein sequences where A.thalQ is the query sequence. It was generated using Clustal Omega and BoxShade was used to convert it into a publishable format. Black shaded regions indicate similar residues.

A.thal3	415	IMITSTITVCLFSTVVFGLTKPLISYLLPHQ N ATTSMLS D -DNTPKSIHIPLLDQDSFI
B.napus	418	IMITSTITVCLFSTVVFGLTKPLIR R LLPHQKATTS F LSD-GNTPKSI Q IPL I LDQDSFI
E.sall1	421	IMITSTITVCLFSTVVFGLTKPLIR R LLPHQKATTSMLS D DNNTPKSI Q IPLLDQDSFI
E.halo3	421	IMITSTITVCL F STVVFGLTKPLIR R LLPHQKATTSMLS D GNNTPKSI Q IPLLDQDSFI
Cap.rubella2	477	EP S GNHNVPRPDSIRGFLTRP T RTVHYYWRQ F DDSFMRPVFGGRGFV P FPVGSPT E RD P P
O.pumila	474	EP S GNHNVPRPDSIRGFLTRP T RTVHYYWRQ F DDSFMRPVFGGRGFV P FPVGSPT E R N IP P
A.thalQ	474	EP S GNHNVPRPDSIRGFLTRP T RTVHYYWRQ F DDSFMRPVFGGRGFV P FPVGSPT E R N IP P
A.thal3	474	EP S GNHNVPRPDSIRGFLTRP T RTVHYYWRQ F DDSFMRPVFGGRGFV P FPVGSPT E R N IP P
B.napus	477	EF A GNHNVPRPDSIRGFLTRP T RTVHYYWRQ F DDSFMRPVFGGRGFV P FPVGSPT E RD P P
E.sall1	481	EF A GNHNVPRPDSIRGFLTRP T RTVHYYWRQ F DDSFMRPVFGGRGFV P FPVGSPT E RD P P
E.halo3	481	EF A GNHNVPRPDSIRGFLTRP T RTVHYYWRQ F DDSFMRPVFGGRGFV P FPVGSPT E RD P P
Cap.rubella2	537	DLSKA-
O.pumila	534	DLSKA-
A.thalQ	534	DLSKA-
A.thal3	534	DLSKA-
B.napus	537	TDLSRA
E.sall1	541	DLSKA-
E.halo3	541	DLSKA-

Figure 3.8 (continued): Multiple sequence alignment of the seven selected protein sequences where A.thalQ is the query sequence. It was generated using Clustal Omega and BoxShade was used to convert it into a publishable format. Black shaded regions indicate similar residues.

A protein consists of surface loops and core regions. The rapidly evolving subsets of a protein are referred to as the surface loops and are often apparent in the multiple sequence alignments (MSAs) and are poorly defined as gap-containing or gap-rich regions (Blouin *et al.*, 2004). Therefore, their counter parts are the core regions which in terms of MSA terminology are called the gap-free regions. The loops are the softer portion present on the surface of the proteins that connect to more rigid portions. The core regions, on the other hand, act as support walls for the proteins (Claverie and Notredame, 2006). The sequence variability in the proteins is linked to the structural variability of the surface loops, thus making them prone to rapid evolutions. These regions are self contained, and are mostly free of the evolutionary constraints imposed by the conserved core of the domains (Blouin *et al.*, 2004). The core regions are less prone to rapid evolutions unlike surface loops. As observed in Figure 3.8, the MSA results displayed protein blocks that mostly contained gap-free regions and only the first block showed very few gap-rich regions. Thus, it was concluded that the selected seven proteins mainly contained core regions and very few surface loops. As such, it supported that the protein is highly conserved amongst the plants, as the core regions do not undergo rapid evolutions and remain unchanged. This is further exemplified by

the fact that the NHX functional groups appeared early on in evolution and have conserved and essential cellular roles in plants (Bassil *et al.*, 2012).

3.3.3 Phylogenetic tree generation by MEGA 6:

A bootstrap consensus tree was generated for the selected seven protein sequences using the MEGA 6 software (Figure 3.9). The MSA file was imported and analyzed using MEGA 6. It revealed 479 conserved sites and 66 variable sites over a span of 546 sites (Figures 3.10 - 3.11).

The Neighbor-Joining method (Saitou and Nei, 1987) was implemented to generate the bootstrap consensus tree with node statistics. As the node statistics were greater than 85, it indicated that the clades formed during divergence were strongly supported. As previously mentioned, the Neighbor-Joining method is dependent on the evolutionary distance. As such, *A.thalQ* is closely related to *Cap.rubella2* and *B.napus* as they had the least degree of divergence from the query sequence itself.

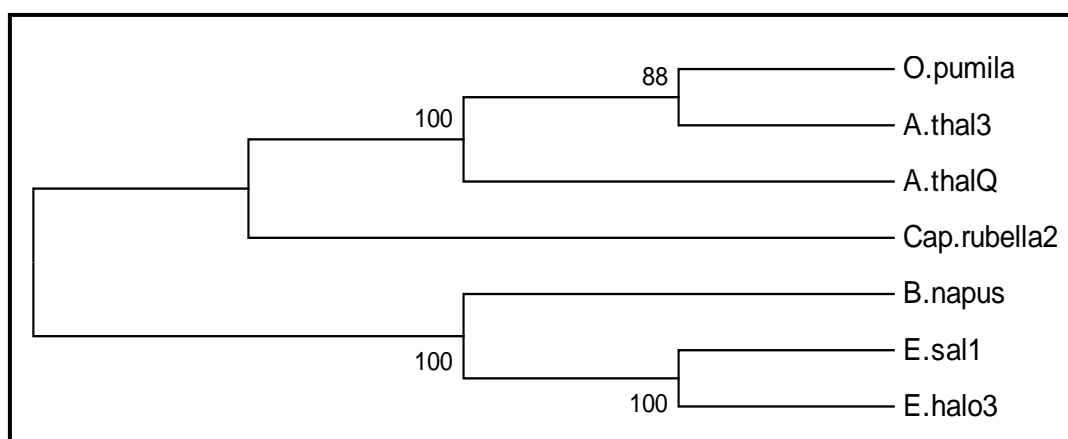


Figure 3.9: A bootstrap consensus tree with node statistics for the selected seven protein sequences along with the query sequence

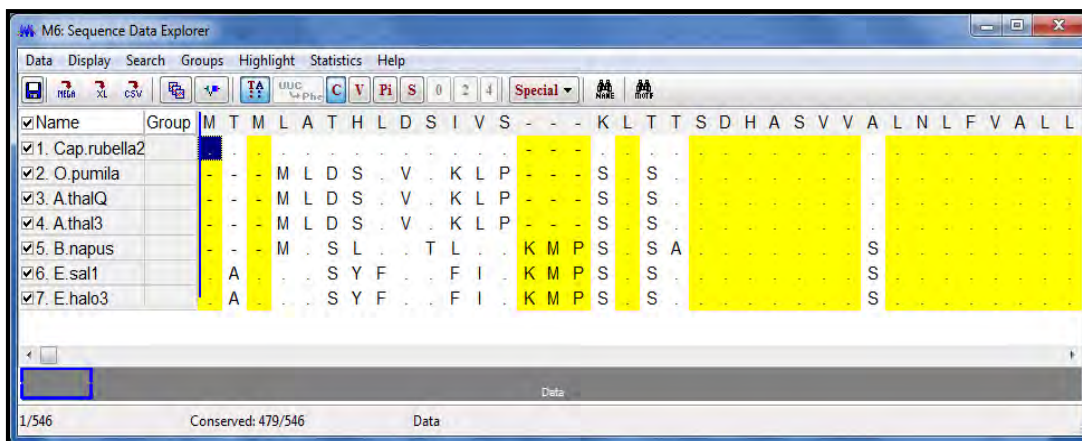


Figure 3.10: Number of conserved sites in seven selected protein sequences

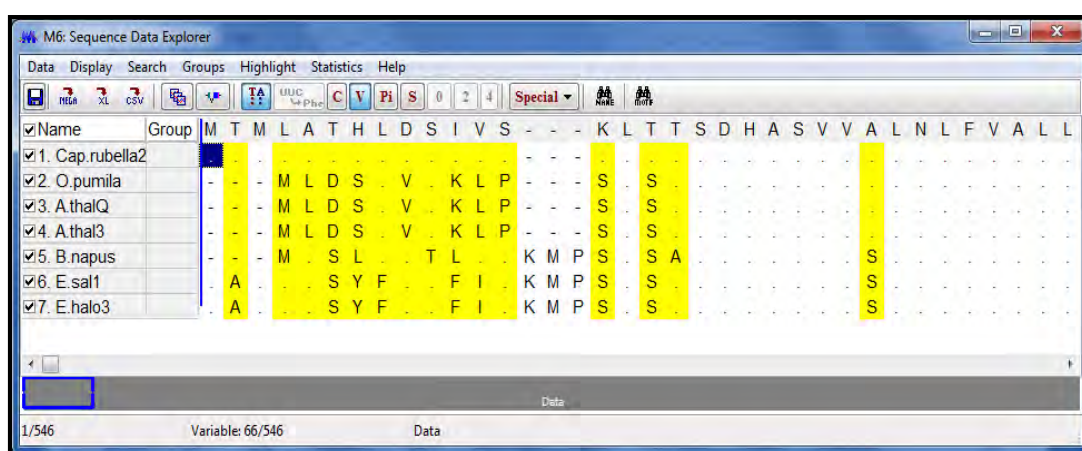


Figure 3.11: Number of variable sites in the seven selected protein sequences

3.3.4 Amino acid composition computation:

The amino acid composition of each of the selected seven sequences including the query sequence was computed using the PEPSTATS analysis tool (Table 3.3). It revealed that the most abundant amino acid was leucine which accounted for ~13% of the protein's primary structure. The least common amino acids were cysteine and tryptophan which accounted for ~1% of the protein's primary structure.

The cysteine residues are very important. This is because two cysteine residues may take part in formation of disulfide bonds between various parts of the same protein or between two separate polypeptide chains (Miseta and Csutora, 2000). The disulfide bonds play a major role in folding, stability and overall maintenance of the topology of the proteins (Betz, 1993, Darby and Creighton, 1995). The low amounts of cysteine residues indicated that the chances of disulfide bond formation were low. As such, it

was predicted that the proteins attain stability from other factors other than the formation of disulfide bonds.

Table 3.3: Amino acid composition based on most abundant residues and least common residues attained by PEPSTATS analysis tool. Here, Leu is leucine, Cys is cysteine and Trp is tryptophan along with their molecular percentage (%) values.

Organisms	Most abundant amino acid	Mole %	Least common amino acid	Mole %	Mole % of Cys residues
A.thalQ	Leu	13.197	Cys, Trp	1.115, 1.115	1.115
A.thal3	Leu	13.001	Cys, Trp	1.115, 1.115	1.115
O.pumila	Leu	13.197	Cys, Trp	1.301, 1.115	1.301
Cap.rubella2	Leu	12.939	Cys, Trp	1.109, 1.109	1.109
E.sal1	Leu	13.211	Cys, Trp	1.101, 1.101	1.101
B.napus	Leu	13.284	Cys, Trp	1.107, 1.107	1.107
E.halo3	Leu	13.028	Cys, Trp	1.101, 1.101	1.101

3.3.5 Analysis of physicochemical properties:

Computation of various physical and chemical parameters of the selected protein sequences was performed using the ProtParam tool and tabulated (Table 3.4). The computed Isoelectric Point (pI) of the proteins was ~6.95 on average; this indicated that the proteins are likely to precipitate in either acidic or basic buffers and can be maintained within a neutral buffer, such as, PBS (Phosphate-buffered saline) buffer. The Extinction Coefficients (EC) of the proteins were all same for each of the seven organisms, with a slight variation seen in B.napus. The Instability Indices (Ii) for the proteins were below 40, which indicated that they would remain stable within a solution. All the proteins had positive Grand Average Hydropathy (GRAVY) scores, which meant that they are hydrophobic in nature. The Aliphatic Index (Ai) evaluates the relative volume of the protein occupied by the aliphatic side chains. Based on the results attained, it indicated that Ai values were quite high, which indicated that the proteins would remain stable over an array of temperatures.

Table 3.4: Parameters for the protein encoded by AtNHX1 gene using the ProtParam program: molecular weight (MW) (g/mol); isoelectric point (pI); extinction coefficient (EC) ($M^{-1} cm^{-1}$); instability index (Ii); aliphatic index (Ai); grand average hydropathy (GRAVY); number of negative residues (-R); number of positive residues (+R)

Organisms	Sequence Length	MW	pI	EC (Cys residues not reduced)	EC (Cys residues reduced)	Ii	Comment	Ai	GRAVY	-R	+R
A.thalQ	538	59513.4	6.73	54235	53860	32.71	Stable	106.71	0.458	39	37
A.thal3	538	59561.4	6.73	54235	53860	32.86	Stable	105.99	0.453	39	37
O.pumila	538	59430.3	6.56	54235	53860	32.78	Stable	106.90	0.475	39	36
Cap.rubella2	541	59884.9	6.76	54235	53860	32.70	Stable	106.67	0.469	40	38
E.sal1	545	60498.7	6.94	54235	53860	34.82	Stable	106.61	0.471	41	40
B.napus	542	59931.1	7.67	52745	52370	33.52	Stable	107.73	0.508	39	40
E.halo3	545	60436.7	7.25	54235	53860	35.62	Stable	106.61	0.477	40	40

3.3.6 Prediction of transmembrane segments within the query protein sequence:

3.3.6.1 Prediction via ProtScale tool:

The ProtScale tool was used to predict the transmembrane segments present within the target protein AtNHX1. This was represented in the form of a two dimensional plot (Figure 3.12). The image seen in Figure 3.12 is the hydrophobicity profile returned by ProtScale using the Kyte & Doolittle Scale (Kyte and Doolittle, 1982). The peaks indicated the potential transmembrane regions present within the protein over a span of 538 amino acids. The recommended threshold level for the aforementioned scale is 1.6. There were twelve peaks thus indicating presence of 12 transmembrane regions within the target protein. The strongest signal was observed for the first peak which had the highest score.

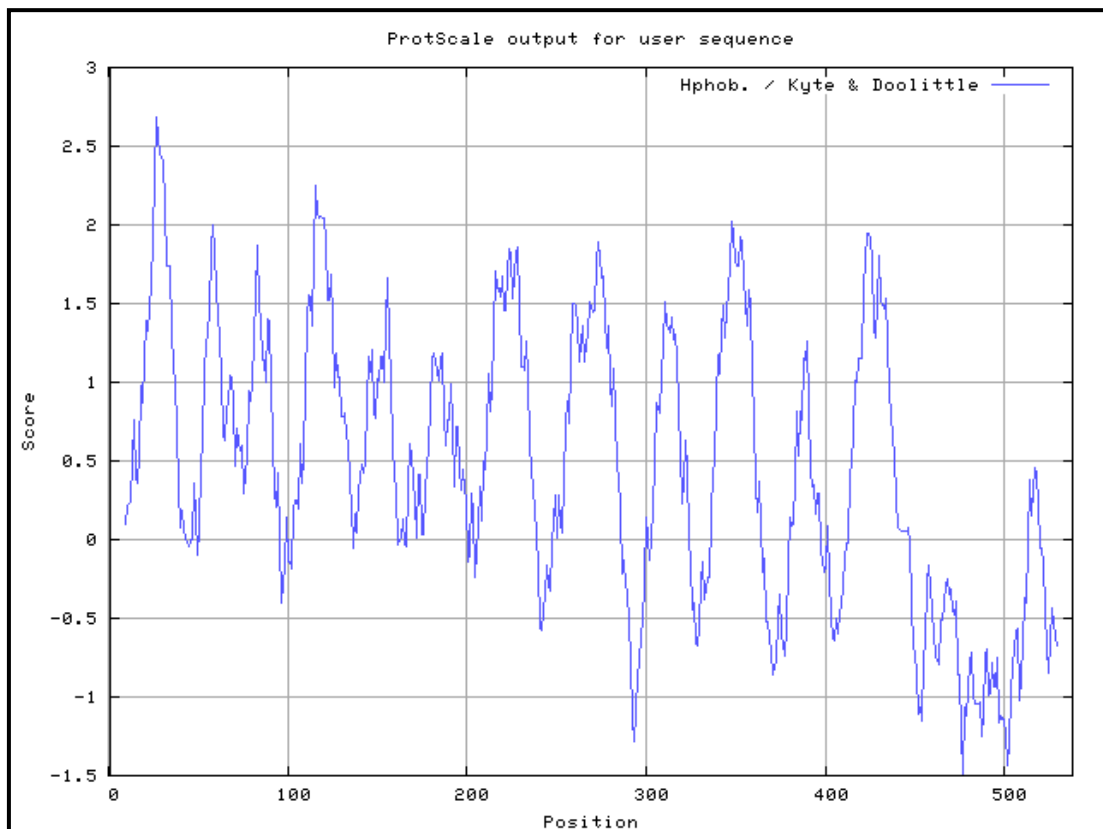


Figure 3.12: ProtScale output of AtNHX1 (A.thalQ)

3.3.6.2 Prediction via TMHMM server:

Similarly transmembrane region prediction was performed using TMHMM server. It revealed a two dimensional plot which depicted the potential transmembrane segments within the target protein (Figure 3.13).

The TMHMM server provided a much more detailed prediction of the transmembrane segments which included whether the transmembrane segments are intrinsic or extrinsic in nature. The blue lines indicated intrinsic transmembrane segments and the pink lines indicated the extrinsic ones. In comparison to the ProtScale hydrophobicity profile, similar patterns were observed. Twelve peaks were also observed here (Figures 3.12 and 3.13). This indicated twelve transmembrane segments, thus supporting the previous conclusion.

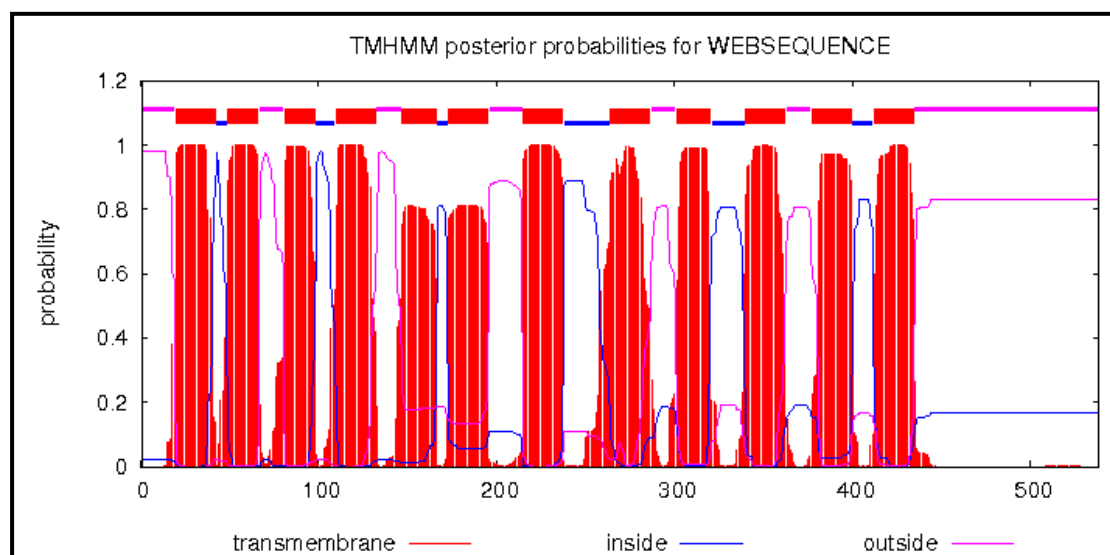


Figure 3.13: TMHMM output for AtNHX1 (A.thalQ)

Yamaguchi and his colleagues (Yamaguchi *et al.*, 2003), performed hydropathy plot analyses which indicated that AtNHX1 contained 10-12 transmembrane domains (Bassil *et al.*, 2012). Therefore, this further validated the previous results in regard to transmembrane segment prediction for AtNHX1 protein. Based on the results attained from both ProtScale and TMHMM, it was concluded that the target protein is a transmembrane-spanning protein that would function as channels when assembled into tetramers (Swarbreck *et al.*, 2013). Furthermore, it was predicted that the transmembrane segments form domains which can fold together so as to shape a central pore whose structural constituents determine the selectivity and conductance

properties of the channel (Marban *et al.*, 1998). This is a valid conclusion as the target protein, AtNHX1, is in fact an antiporter protein. In addition, it was predicted that a sufficiently large amount of alpha-helices were present amongst the transmembrane regions of the target protein. This is because, the plots suggested that the AtNHX1 protein constitutes of helix-bundle proteins that are built from long transmembrane α -helices that pack together into more or less complicated bundles (Elofsson and Heijne, 2007).

3.3.7 Secondary structure element prediction and prediction of disulfide bonds/bridges:

Protein is made of a sequence of amino acids that is folded in a 3D structure. The protein secondary structures are small groups of protein structures that exhibit particular prominent and regular characteristics that function as the intermediate building blocks of the overall 3D structure. It can be classified into three types, namely, α -helix, β - sheet and coil (Abe and Mamitsuka, 1997, Wang and Jardetzky, 2002).

To attain quantitative values for the amount of alpha-helices, beta sheets and coils present within the amino acid stretch of the protein the self optimized method for protein secondary structure prediction by consensus prediction from multiple alignments (SOPMA) tool available at the NPS@ server was used. The presence of disulphide bonds/bridges was analyzed using the CYS_REC tool which predicts the most probable bonding patterns between available cysteine residues. The results were tabulated (Table 3.5).

The selected seven sequences shared similar α -helical and extended strands/ β -sheet content. The analysis revealed that the α -helices were dominant amongst the secondary structures followed by the coils, extended strands/ β -sheets and β -turns. The data revealed that 9.11% of the target protein's (AtNHX1 of A.thalQ) secondary structure was composed of β -turns. No disulphide bridges were present for any of the proteins including the target protein when the CYS-REC tool was implemented.

Table 3.5: Predicted secondary structure content and disulphide bridges using NPS@ SOPMA and CYS-REC tools.

Organism	α-helix (%)	E-strands/ β-sheets (%)	Coil (%)	β-turn (%)	Disulfide bridge prediction (CYS-REC) (%)
A.thalQ	33.64	26.21	31.04	9.11	None
A.thal3	33.83	25.65	31.60	8.92	None
O.pumila	33.83	26.02	30.48	9.67	None
Cap.rubella2	35.12	26.43	29.39	9.06	None
E.sal1	39.08	24.59	28.26	8.07	None
B.napus	36.16	26.20	29.15	8.49	None
E.halo3	38.35	24.59	28..81	8.26	None

The α -helices and β -sheets are considered to be regular secondary structure elements. However, the residues that correspond to the turns structures do not form the regular secondary structure elements. The most common types of turns structure that exist in protein are β -turns structure (Elbashir *et al.*, 2013). β -turns can reverse the direction of a protein chain. Therefore, they are considered as the orienting structure (Petersen *et al.*, 2010). They have major impacts on protein folding as well. This is due to their ability to bring together and allow interactions between regular secondary structure elements. They play significant roles in stability and molecular recognition. The β -turns also play key roles in biological activities of peptides, such as, the bioactive structures that allow interaction with several other molecules like, enzymes, receptors and so on (Zheng and Kurgan, 2008).

In light of the above, it was predicted that the proteins attained stability due to the presence of β -turns which compensated for the lack of disulfide bond/bridge formation. Furthermore, the proteins all have high percentages of α -helices. It is of common knowledge that the hydrogen bonding is the most prominent feature of a α -helix. Hydrogen bonds have a central role in the folding, stabilization, and function of helical membrane proteins in general (Senes *et al.*, 2001, Adamian and Liang, 2002, Curran and Engelman, 2003). Therefore, a secondary factor that plays a role in the proteins' stability is the presence of extensive hydrogen bonds.

3.3.8 Graphical representation of secondary structures in *Arabidopsis thaliana* sodium/hydrogen exchanger 1 protein:

A secondary structure map and a graphical representation of the predicted secondary structures of the target protein were attained using PSIPRED (Figures 3.14-3.15).

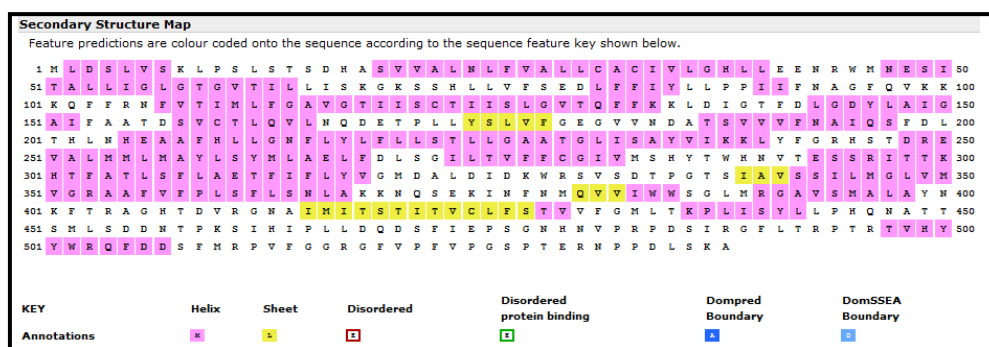


Figure 3.14: Secondary structure map of AtNHX1 attained from PSIPRED

As per the legend provided in Figure 3.15, the pink cylinders represented the α -helices and the yellow arrows represented beta strands. The black threads like structures were the coils. Most of the transmembrane proteins consist solely of α -helices that are present in the cytoplasmic membrane. A few of the membrane proteins constitutes of β -strands. These β -strands form the β -barrel topology which is a cylindrical structure composed of antiparallel β -sheets. These β -barrels are normally found in the outer transmembrane proteins (Xiong, 2006).

The extended strands or β -sheets linked to the α -helices may act as beta-barrels, and thus constructed the external transmembrane regions of the protein. Furthermore, the confidence of prediction observed throughout the predicted secondary structure was quite high. Thus it assured the plausibility of the prediction.

3.3.9 Homology Modeling:

The 3D models of the target protein were constructed using three protein structure homology model building programs I-TASSER, PHYRE and EasyModeller. Molecular graphics and analyses were performed with the UCSF Chimera package. Chimera is developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco.

3.3.9.1 Homology modeling using I-TASSER:

In this method the target sequences are threaded using a representative PDB structure library (Zhang, 2008). This is done to search for the possible folds by Profile-Profile Alignment (PPA), Hidden Markov Model, PSI-BLAST profiles, Needleman-Wunsch and Smith-Waterman alignment algorithms (Suganya *et al.*, 2014). A list of the top ranking templates used by I-TASSER to generate the models was attained (Table 3.6). I-TASSER used the PDB ID: 4cz8A as the template for modeling the *Arabidopsis thaliana* sodium/hydrogen exchanger 1 protein. The Z-score was greater than 1, which indicated a confident alignment between query and template. It also indicated that the template was likely to have the same fold as the query protein. It was observed that, Iden1 was slightly greater than Iden2 which indicated conserved structural motifs in the query sequence in comparison to the template.

Table 3.6: Top ranking templates used by I-TASSER to model the target protein

Rank	PDB Hit	Iden1	Iden2	Cov	Norm. Z-score
1	4cz8A	0.21	0.20	0.72	2.13
2	1qgrA	0.10	0.16	0.99	2.19
3	4cz9A	0.22	0.20	0.73	4.00
4	3c2gA	0.10	0.21	0.89	1.13
5	4cz8A	0.19	0.20	0.73	2.47
6	3wajA	0.09	0.20	0.84	1.31
7	4bwzA	0.16	0.20	0.71	9.08
8	4heaL	0.15	0.21	0.79	1.46
9	4bwzA	0.15	0.20	0.71	5.80
10	4a01A	0.11	0.21	0.95	1.06

Note: **Iden1** is the percentage sequence identity of the templates in the threading aligned region with the query sequence, **Iden2** is the percentage sequence identity of the whole template chains with query sequence, **Cov** represents the coverage of the threading alignment and is equal to the number of aligned residues divided by the length of query protein and **Norm. Z-score** is the normalized Z-score of the threading alignments. Alignment with a Normalized Z-score >1 mean a good alignment and vice versa.

The I-TASSER server predicted 5 models from which the model with the best C-score of -0.66 was selected (Figure 3.16 a). The range of the C-score i.e. the confidence score is -5 to 2. As such the selected model had a score closest to 2 in comparison to the other 4. The estimated TM-score was 0.71 which is greater than 0.5. Therefore, it was inferred that the model was of correct topology. The number of decoys used was 600 and the cluster density was 0.1136. As such, out of the 600 decoys used to generate the models, Model 01 appeared 11.36% of times, which indicated a good quality model.

A list of the top ten identified structural analogs was provided by the I-TASSER server (Table 3.7). Structurally conserved residues and motifs were observed via visual inspection. As such, the preferred structural analogs based on the parameters provided were the first two ranking PDB ID entries.

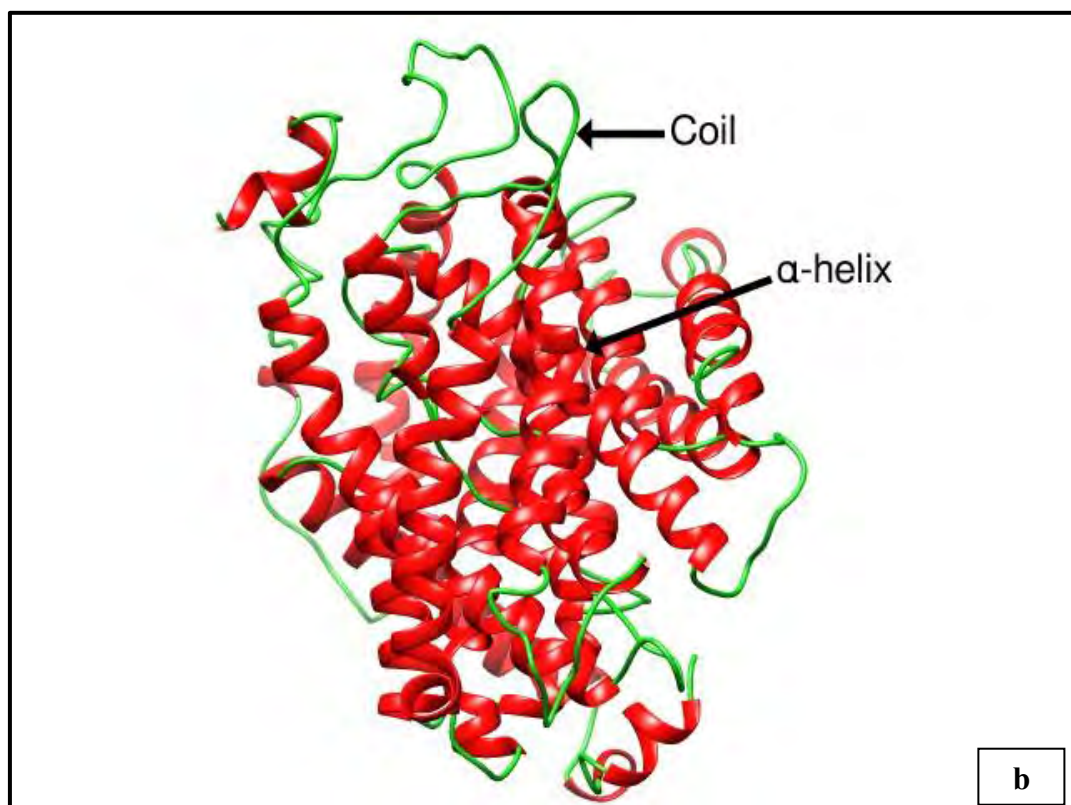
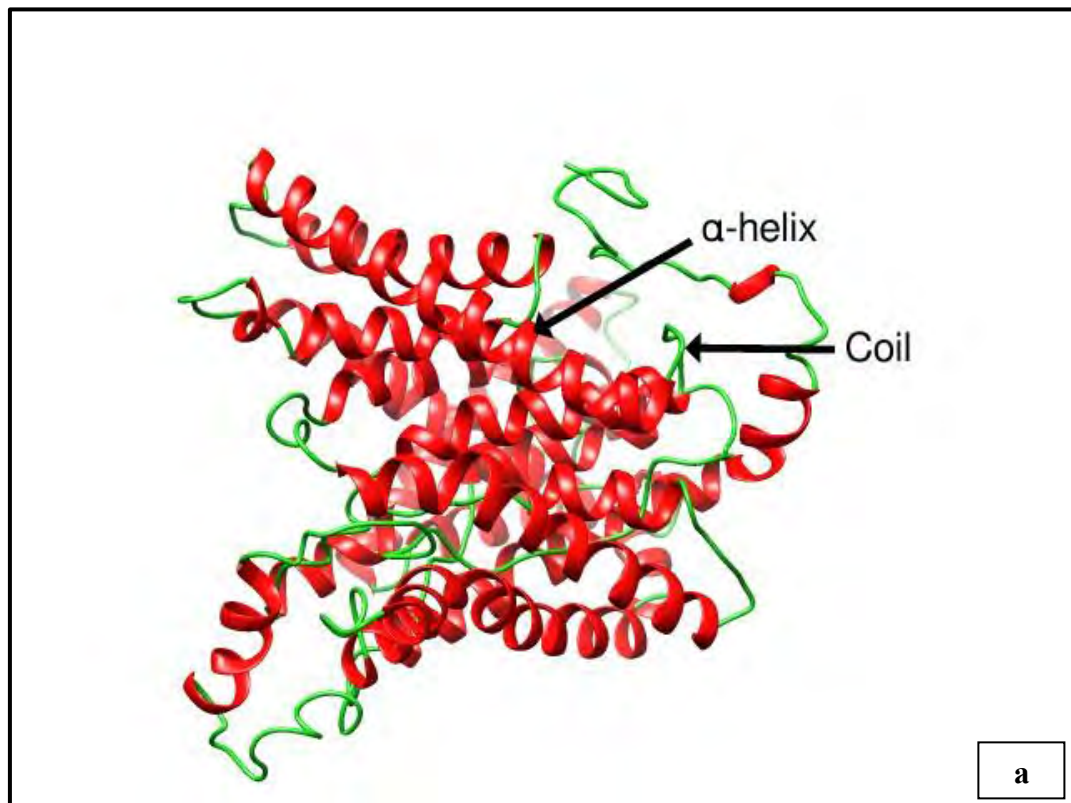


Figure 3.16: Ribbon diagrams of modeled *Arabidopsis thaliana* Na⁺/H⁺ exchanger 1 protein (a) I-TASSER model and (b) Phyre2 model. α-helices and coils are colored red and green, respectively

Table 3.7: Top ten identified structural analogs in PDB provided by I-TASSER

Rank	PDB hit	TM-score	RMSD ^a	IDEN ^a	Cov
1	4cz8A	0.718	1.09	0.219	0.729
2	4czbA	0.691	2.11	0.181	0.731
3	4bwzA	0.589	3.99	0.095	0.701
4	1zcdA	0.514	4.42	0.103	0.628
5	3zuyA	0.460	4.17	0.069	0.558
6	4n7wA	0.454	4.36	0.113	0.561
7	3kbcC	0.381	5.94	0.108	0.535
8	2wt5A	0.362	7.18	0.056	0.565
9	3cqmA	0.358	7.55	0.061	0.580
10	3f93C	0.353	7.78	0.047	0.584

Note: **RMSD^a** is the RMSD between residues that are structurally aligned by TM-align. **IDEN^a** is the percentage sequence identity in the structurally aligned region. **Cov** represents the coverage of the alignment by TM-align and is equal to the number of structurally aligned residues divided by length of the query protein.

Amongst the ten structural analogs, 4cz8A was the preferred choice. The TM-score was higher than 0.5 which indicated that this analog and the generated model had a similar topology. Therefore, it can be used to determine the structural class of the query protein (Figure 3.17).

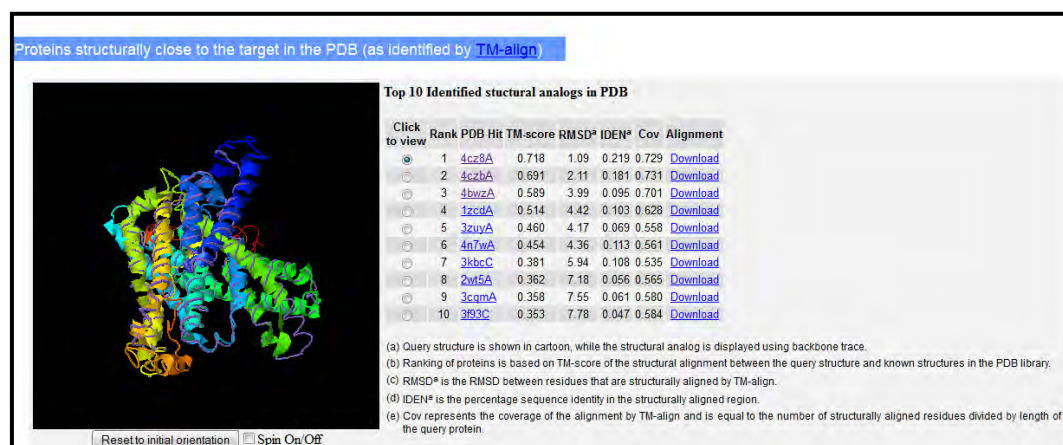


Figure 3.17: Structure superposition of query protein (shown in cartoon) and template protein (shown in backbone)

3.3.9.2 Homology modeling using Phyre2:

The Phyre2 tool modeled the query protein using multiple templates with the highest sequence coverage and confidence (Figure 3.16 b). The intensive mode was selected to achieve the desired output. In the Phyre2 model of the query sequence, 80% of the residues (~430 residues) were modeled with over 90% confidence. Four templates were used to model the query protein. The target sequence was covered by each template, color-coded by the confidence of the match to that template overall (Figure 3.18). Furthermore, 75 residues were modeled by *ab initio* (subject to unreliable modeling). Out of the 4 templates one had a borderline confidence level of 70%.

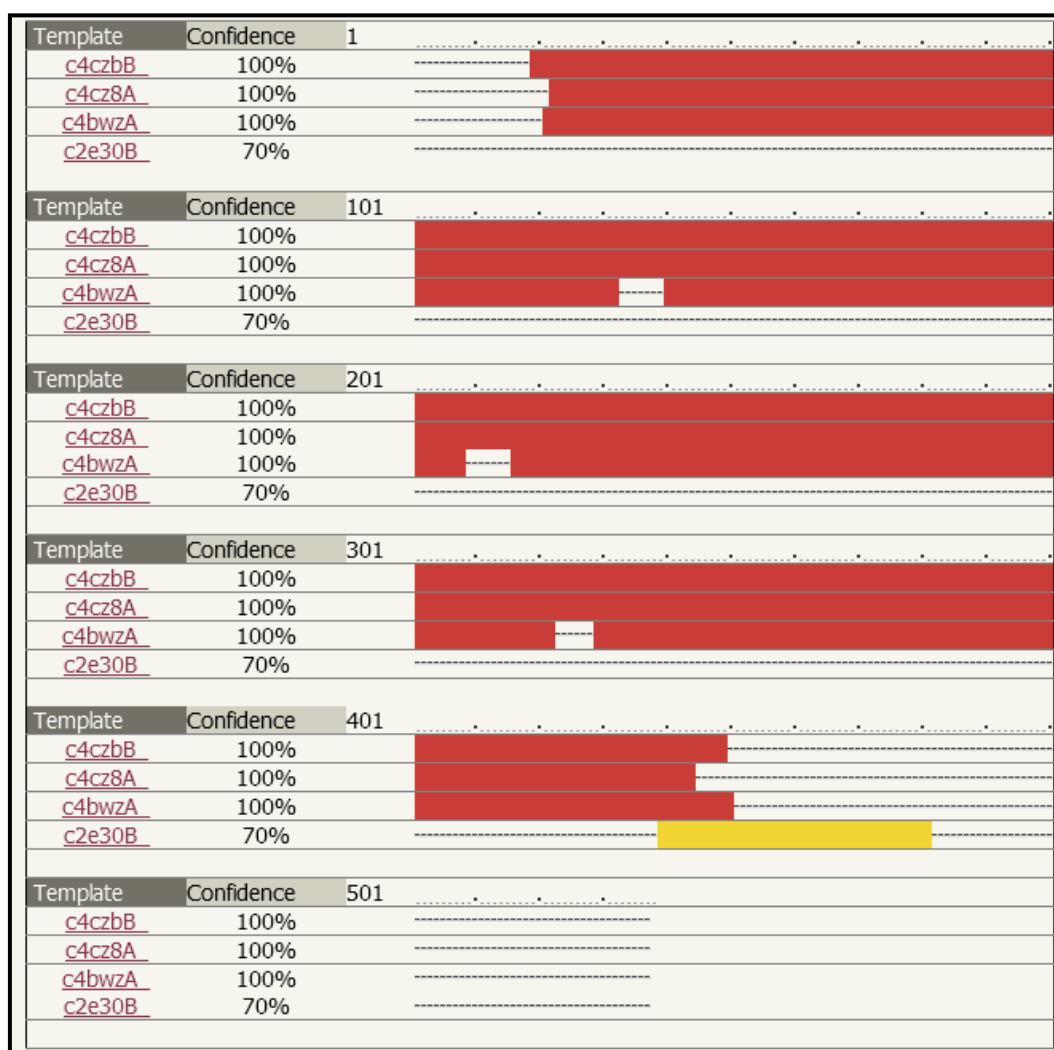


Figure 3.18: Templates used by Phyre2 to model the protein with confidence score coloring

3.3.9.3 Homology modeling using EasyModeller:

3.3.9.3.1 Template selection:

The target protein sequence was blasted using the blast P-suite against the Protein Data Bank (PDB) to attain templates required for homology modeling via EasyModeller. A graphical summary of the BLAST results was attained (Figure 3.19). The red streak represented the query sequence. Based on the BLAST results it was observed that there were very few structures that could match completely with the query sequence. The highest query coverage was between the ranges of 40-50 (%).

A list of homologous sequences that produced significant alignments for the target protein sequence was attained (Figure 3.20). From the top 13 results, sequences/structures were selected based on identity values and E-values. Based on the results attained and comparing the E-values and identity values, the top three structures were selected to act as templates for homology modeling of the target protein (Table 3.8). These three templates were selected as they had the lowest E-values, moderately good query coverage and identity values. These templates referred to the structure of sodium proton antiporter PaNhaP from *Pyrococcus abyssii*. Using the RCSB Protein Data Bank and using the PDB IDs, the PDB text files were downloaded to be used as templates.

Using the EasyModeller software, a comparison was made between the templates. This comparison was made using weighted pair-group average clustering based on distance matrix (Figure 3.21). Based on the three templates provided in the figure, 4CZAA and 4CZ8A were similar. However, 4CZ9A did not have any sequence similarity with the query sequence and it had a higher crystallographic resolution (3.5 Å against 3.2Å) thus it proved to be an unsuitable candidate for homology modeling. Between 4CZAA and 4CZ8A, the latter had a higher sequence identity to the query sequence than the former. Even though they had matching crystallographic resolution, 4CZ8A was chosen to be the appropriate template for homology modeling of the AtNHX1 antiporter protein.

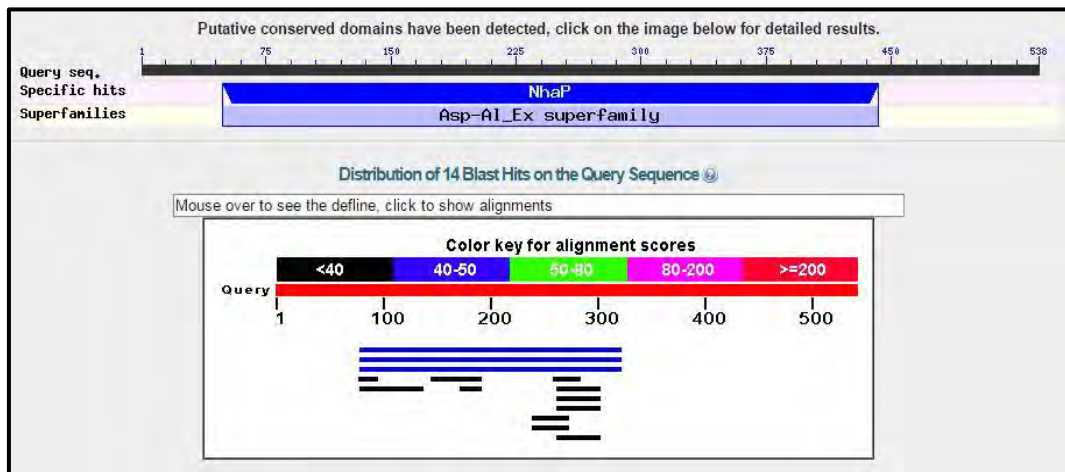


Figure 3.19: Graphical summary of BLAST results against PDB for template selection

Sequences producing significant alignments:

Select: All None Selected: 0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/> Chain A, Structure Of The Sodium Proton Antiporter Panhap From Pyrococcus Abyssii At Ph 8	43.9	43.9	44%	3e-04	25%	4CZ8_A
<input type="checkbox"/> Chain A, Structure Of The Sodium Proton Antiporter Panhap From Pyrococcus Abyssii With Bound Thallium Ion	43.9	43.9	44%	3e-04	25%	4CZA_A
<input type="checkbox"/> Chain A, Structure Of The Sodium Proton Antiporter Panhap From Pyrococcus Abyssii At Ph 4	43.9	43.9	44%	3e-04	25%	4CZ9_A
<input type="checkbox"/> Chain A, Nmr Structure Of A Two-transmembrane Segment Tm Vi-vii Of Nhe1	37.7	37.7	8%	0.003	43%	2MDF_A
<input type="checkbox"/> Chain A, Structural And Functional Characterization Of Tm Ix Of The Nhe1 Isoform Of The Na ⁺ /H ⁺ EXCHANGER	31.6	31.6	4%	0.34	50%	2K3C_A
<input type="checkbox"/> Chain A, Nmr Structure Of Transmembrane Segment Iv Of The Nhe1 Isoform Of The Na ⁺ /H ⁺ EXCHANGER	28.9	28.9	3%	2.6	58%	1Y4E_A
<input type="checkbox"/> Chain A, Structural And Functional Characterization Of Tm Vii Of The Nhe1 Isoform Of The Na ⁺ /H ⁺ EXCHANGER	28.1	28.1	3%	4.8	57%	2HTG_A
<input type="checkbox"/> Chain A, Tem1 Beta Lactamase Mutant S70a	30.0	30.0	7%	5.2	30%	1ZGG_A
<input type="checkbox"/> Chain A, A Triple Mutant In The Omega-loop Of Tem-1 Beta-lactamase Changes The Substrate Profile Via A Large Conformational Change And An Altered General	30.0	30.0	7%	5.2	30%	4RX3_A
<input type="checkbox"/> Chain A, Crystal Structures Of Chimeric Beta-lactamase Ctem-19m Showing Different Conformations	29.6	29.6	7%	6.5	30%	4QY5_A
<input type="checkbox"/> Chain A, Nmr Structure Of Chaperone Chz1 Complexed With Histone H2a-z-h2b Dimer	29.3	29.3	6%	7.3	39%	2JSS_A
<input type="checkbox"/> Chain A, Crystal Structure Of Yeast Swr1-z Domain In Complex With H2a-z-h2b Dimer	29.3	29.3	6%	8.1	39%	4M6B_A
<input type="checkbox"/> Chain A, Crystal Structure Of M68I/m69T Double Mutant Tem-1	29.3	29.3	7%	8.5	30%	4MEZ_A
<input type="checkbox"/> Chain B, Crystal Structure Of Pta Gtp-Specific Succinyl-Coa Synthetase In Complex With Gtp	29.3	29.3	10%	9.5	28%	2EP4_B

Figure 3.20: Top 13 homologous sequences that produced significant alignments

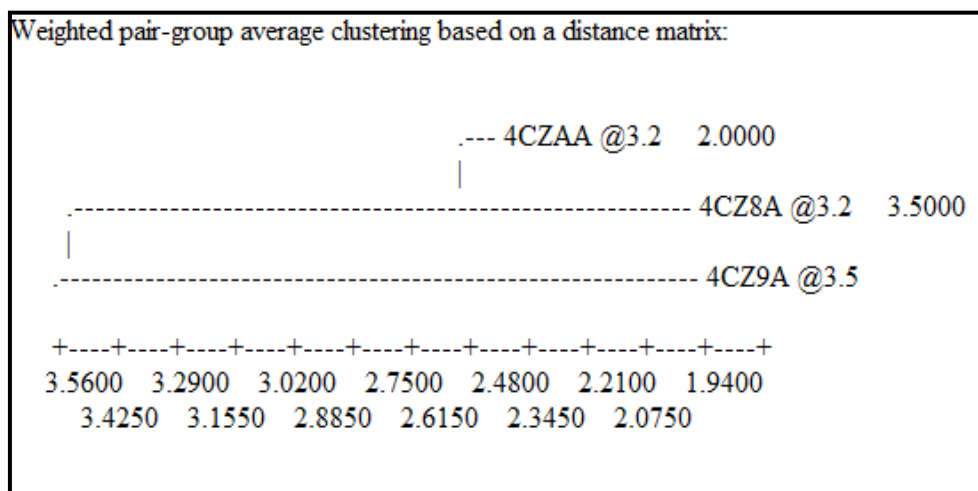


Figure 3.21: Comparison between selected templates using weighted pair-group average clustering based on distance matrix

Table 3.8: List of selected templates attained using blast P-suite against Protein Data Bank (PDB)

Description	Max score	Total score	Query coverage	E-value	Identity	Accession
Chain A, Structure Of The Sodium Proton Antiporter Panhap From <i>Pyrococcus Abyssii</i> At Ph 8 [<i>Pyrococcus abyssi</i> GE5]	43.9	43.9	44%	3e-04	25%	4CZ8_A
Chain A, Structure Of The Sodium Proton Antiporter Panhap From <i>Pyrococcus Abyssii</i> With Bound Thallium Ion [<i>Pyrococcus abyssi</i> GE5]	43.9	43.9	44%	3e-04	25%	4CZA_A
Chain A, Structure Of The Sodium Proton Antiporter Panhap From <i>Pyrococcus Abyssii</i> At Ph 4 [<i>Pyrococcus abyssi</i> GE5]	43.9	43.9	44%	3e-04	25%	4CZ9_A

Wöhlert and his colleagues (Wöhlert *et al.*, 2014), resolved the substrate ion in the dimeric and electroneutral sodium/proton antiporter PaNhaP for *Pyrococcus abyssi* at 3.2 Å. Furthermore, they determined the structure of the aforesaid protein in two different conformations at pH 8 and pH 4. The entry in the Protein Data Bank for the structure at pH 8 is 4CZ8. The template selected to generate model using EasyModeller was 4CZ8_A which indicated that the chain A of the template was used to model the *Arabidopsis thaliana* Na⁺/H⁺ exchanger 1 protein.

3.3.9.3.2 Template and query sequence alignment using EasyModeller:

The EasyModeller software was used to align the template with the query protein sequence (Figure 3.23). The red blocks depicted conserved regions and the amino acid residues were colored according to similarity. Furthermore, the alignment attained via EasyModeller, also depicted the predicted secondary structure based on the query and templates used. The predicted secondary structures in the alignment, namely, alpha

helices and beta sheets were observed along with their occurrence probability indicated as a function of color. A deeper shade of red indicated a higher confidence and a deeper shade of green represented a lower confidence level. Based on the tool's prediction it was observed that there were no beta strands and consisted mainly of alpha helices which had high confidence levels as indicated by shades of reddish-orange.

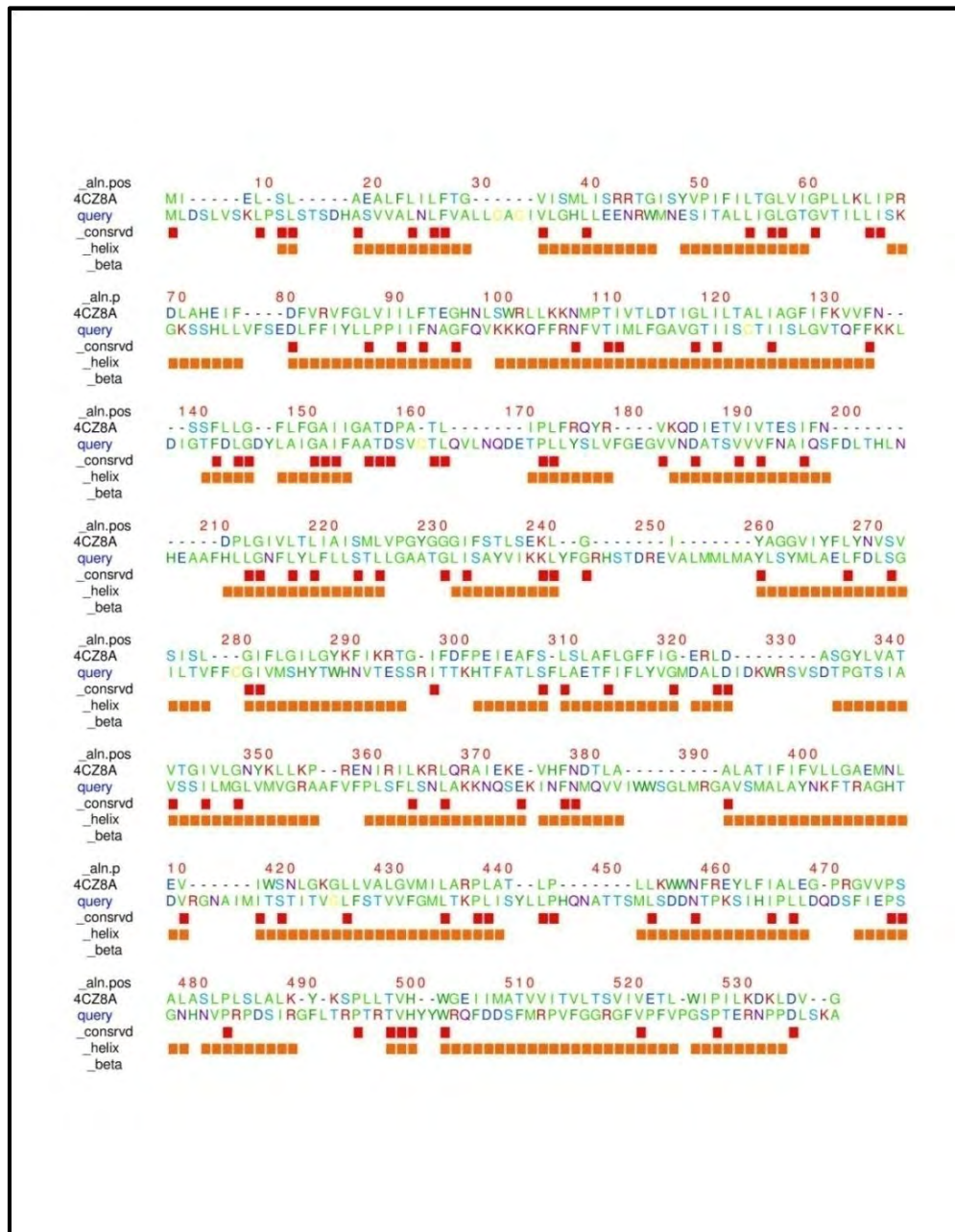


Figure 3.23: Template and query sequence alignment by EasyModeller

3.3.9.3.3 Models generated by EasyModeller:

Five models were generated by EasyModeller and summary of the models were provided (Table 3.9). The models were selected based on low molpdf values, low DOPE (Discrete Optimized Protein Energy) and high GA341 scores. Taking the data provided in the table into account, the best three models were the first, second and fifth models. These three models had the lowest DOPE score and the highest GA341 values (Figure 3.24).

Table 3.9: Summary of the five models generated by EasyModeller (EM)

Model Generated	molpdf	DOPE score	GA341
EM_Model 01	4291.60400	-61590.34375	0.63531
EM_Model 02	3983.05884	-61260.57422	0.47629
EM_Model 03	5198.48682	-58306.60156	0.40473
EM_Model 04	4416.02979	-59859.49219	0.11849
EM_Model 05	4183.39648	-61248.77344	0.46483

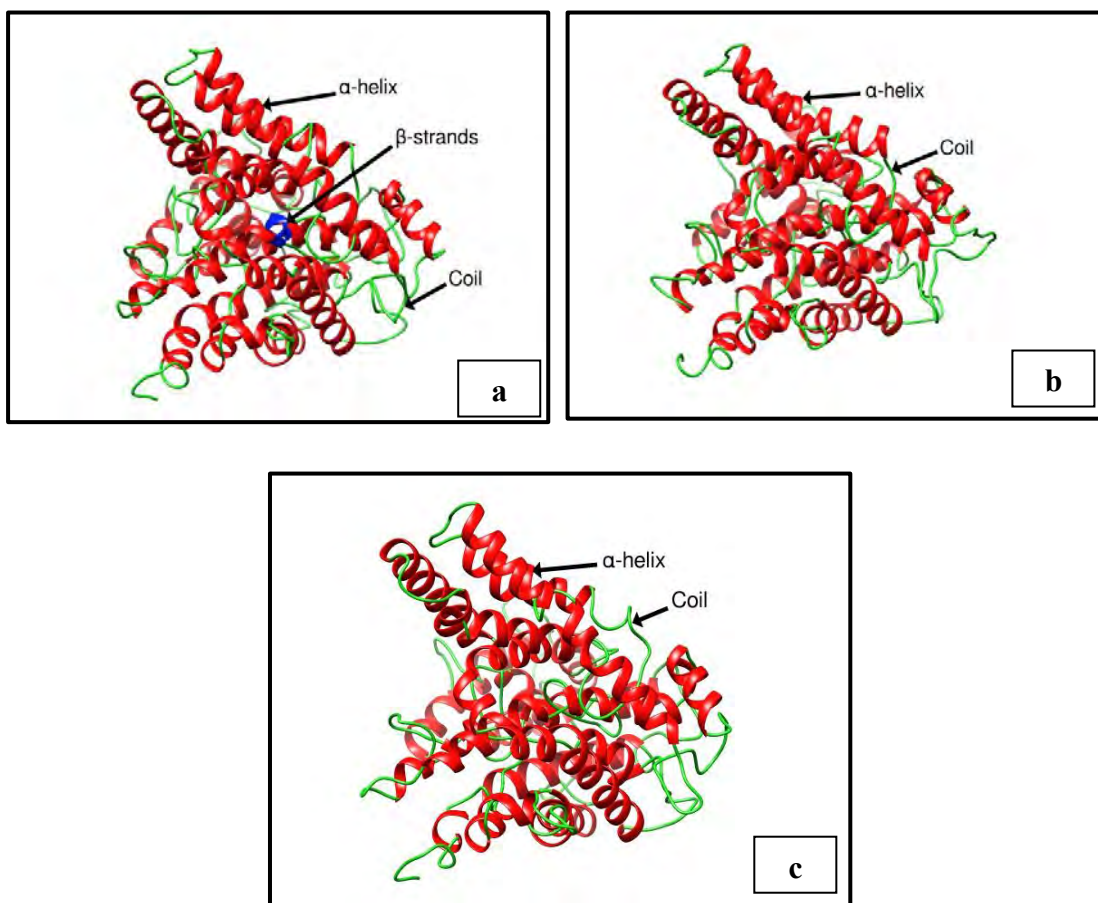


Figure 3.24: Selected models generated by EasyModeller. (a) EM_Model 01, (b) EM_Model 02 and (c) EM_Model 05

3.3.10 Model validation:

To evaluate the models provided by I-TASSER, Phyre2 and EasyModeller several tools were used. These included PROCHECK for calculating Ramachandran plot (Morris *et al.*, 1992) calculations and the QMean server for determining model reliability. The G-Factor values were used to establish the unusual properties of the models. Furthermore, the **Root Mean Square Deviation (RMSD)** values were attained by superimposing the models onto the selected template i.e. 4CZ8A using Chimera.

3.3.10.1 Selection of best model between the three EasyModeller models:

The three selected models generated by EasyModeller were evaluated using the aforementioned tools and a comparison was conducted between the three (Table 3.10). PROCHECK Ramachandran Plot supported EM_Model 01 as it had the highest number of residues in the most favored regions (82.1%). The G-Factor overall average was the highest (-0.38). The G-Factor provides a measure of how unusual a property is and as such a value below -0.5 is considered unusual and a value below -1.0 is considered highly unusual. EM_Model 01 was the least unusual amongst the three.

For QMEAN scores, the estimated model reliability is between 0-1 with higher values indicating more reliable candidates. The QMEAN score for EM_Model 01 was the lowest (0.204). This suggested that EM_Model 02 and EM_Model 05 were more reliable. The QMEAN Z-score provides an approximation of the absolute quality of a model by relating it to reference structures solved by X-ray crystallography. The QMEAN Z-scores were very low. This was justified as low Z-scores are applicable for membrane proteins.

The RMSD between the model and the template was calculated by superimposing the structure of the template on the predicted structure of *Arabidopsis thaliana* Na⁺/H⁺ exchanger 1 antiporter protein in order to assess the reliability of the model using Chimera. As such a low RMSD score would be preferable (0 would indicate it is exactly similar to the template). EM_Model 01 had the lowest RMSD score (0.595 Å). Overall the validation scores favored EM_Model 01 which suggested it was the best between the three EasyModeller models. Considering all three validation data, the EM_Model 01 was selected for further analysis.

Table 3.10: Comparative values of PROCHECK, G-Factor, QMean scores and RMSD between the template and all three EasyModeller modeled proteins

Validation		EM_Model 01	EM_Model 02	EM_Model 05
PROCHECK Ramachandran Plot	Most favored regions	82.1%	81.9%	81.1 %
	Additional allowed regions	12.5%	12.7%	13.5%
	Generously allowed regions	3.5%	4.0%	3.5%
	Disallowed regions	1.9%	1.5%	1.9%
G-Factor Overall Average		-0.38	-0.33	-0.35
Total QMean score		0.204	0.232	0.231
QMean Z- score		-6.54	-6.22	-6.22
RMSD		0.595 Å	0.606 Å	0.685 Å

3.3.10.2 Selection of final model between models generated by I-TASSER, Phyre2 and EasyModeller

A final comparison was conducted between EM_Model 01, I-TASSER model and Phyre2 model (Table 3.11). Ramachandran plots for the models EM_Model01, I-TASSER and Phyre 2 were provided by PROCHECK (Figure 3.25).

The selected best model from EasyModeller (EM_Model 01) indicated 82.1% of the residues in the most favored regions, 12.5% in the additional allowed regions, 3.5% in the generously allowed regions and 1.9% in the disallowed regions (Figure 3.25 a). These results revealed that the bulk of the amino acids are in the phi-psi distribution that is consistent with a right-handed alpha helix and beta strands. It also indicated that the model was reliable and of good quality. The other two models did not have such scores in comparison to EM_Model 01 (Figure 3.25 b-c). It had G-Factor score of -0.38 which is greater than -0.5 which indicated that the model was not unusual. However, I-TASSER model had a G-Factor score of -0.91 and the Phyre2 model had a score of -1.94. This suggested that both the models were highly unusual.

QMEAN score for EM_Model 01 was 0.204, I-TASSER model was 0.359 and Phyre2 model was 0.291. The scores suggested that the I-TASSER model was the most reliable amongst the three. RMSD between the template and the selected model EM_Model 01 was 0.595 Å, whereas I-TASSER model was 0.504 Å and the Phyre2 model was 1.805 Å.

All these results suggested that the EasyModeller model EM_Model 01 was the best amongst the three which had the highest stereochemical quality scores and was considered to be the least unusual. Therefore, it was considered to be comparatively robust and can be used in subsequent stages of analysis.

Table 3.11: Comparative values of PROCHECK, G-Factor, QMean scores and RMSD between the template and models generated by EasyModeller, I-TASSER and Phyre2

Validation		EasyModeller (EM_Model 01)	I-TASSER	PHYRE
PROCHECK Ramachandran Plot	Most favored regions	82.1%	68.8%	75.8%
	Additional allowed regions	12.5%	20.6%	14.0%
	Generously allowed regions	3.5%	7.5%	5.8%
	Disallowed regions	1.9%	3.1%	4.4%
G-Factor Overall Average		-0.38	-0.91	-1.94
Total QMean score		0.204	0.359	0.291
QMean Z-score		-6.54	-4.73	-5.53
RMSD		0.595 Å	0.504 Å	1.805 Å

3.3.11 Structural motifs present within the final model EM_Model 01:

The PDBsum Generate's ProMotif (Hutchinson and Thornton, 1996) provided a summary of the secondary structure components present within the final selected model (EM_Model 01) (Tables 3.12-3.13). Schematic and topology diagrams showing the secondary structural elements within the model were attained from PDBsum tool (Figures 3.26 - 3.27).

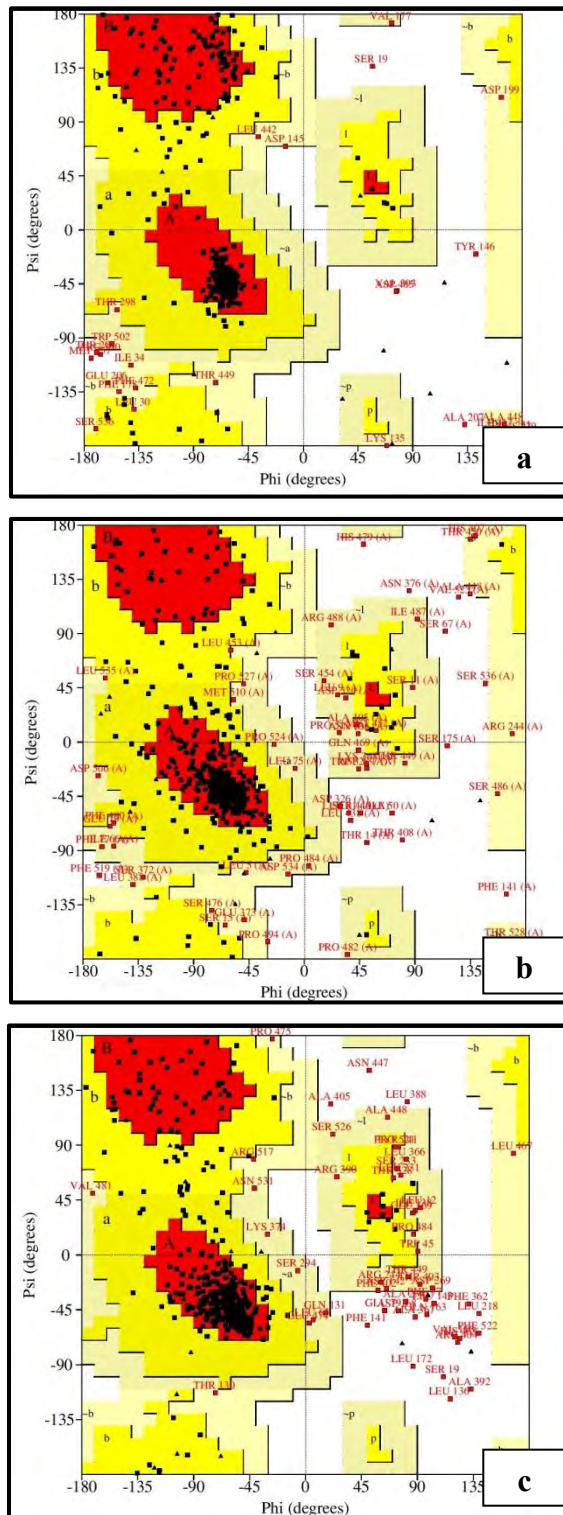


Figure 3.25: Ramachandran plots by PROCHECK. (a) EM_Model 01, (b) I-TASSER and (c) Phyre 2

Table 3.12: Secondary structure summary of EM_Model 01 provided by ProMotif

Motifs	Residues
Strand	6 (1.1%)
Alpha helix	293 (54.5%)
Other	239 (44.4%)
Total	538 (100%)

Table 3.13: Different motifs present within EM_Model 01 provided by ProMotif

Motifs	Quantity
Sheets	1
Beta Hairpin	1
Strands	2
Helices	28
Helix-helix interacs	63
Beta turns	56
Gamma turns	10

The selected protein model consisted of a single β -sheet (Table 3.14). This β -sheet was made of two strands which were antiparallel and hydrogen bonded. The first strand ranged from residue 197 (Serine) till residue 199 (Aspartine). The second strand ranged from residue 209 (Phenylalanine) till residue 211 (Leucine). Each strand was three residues in length. This constituted the β -hairpin as observed in the schematic and topology diagrams for secondary structural elements (Figures 3.26-3.27).

Table 3.14: Summary of beta hairpin provided by ProMotif

Strand 1			Strand 2			Hairpin class
Start	End	Length	Start	End	Length	
Ser197	Asp199	3	Phe209	Leu211	3	9:9

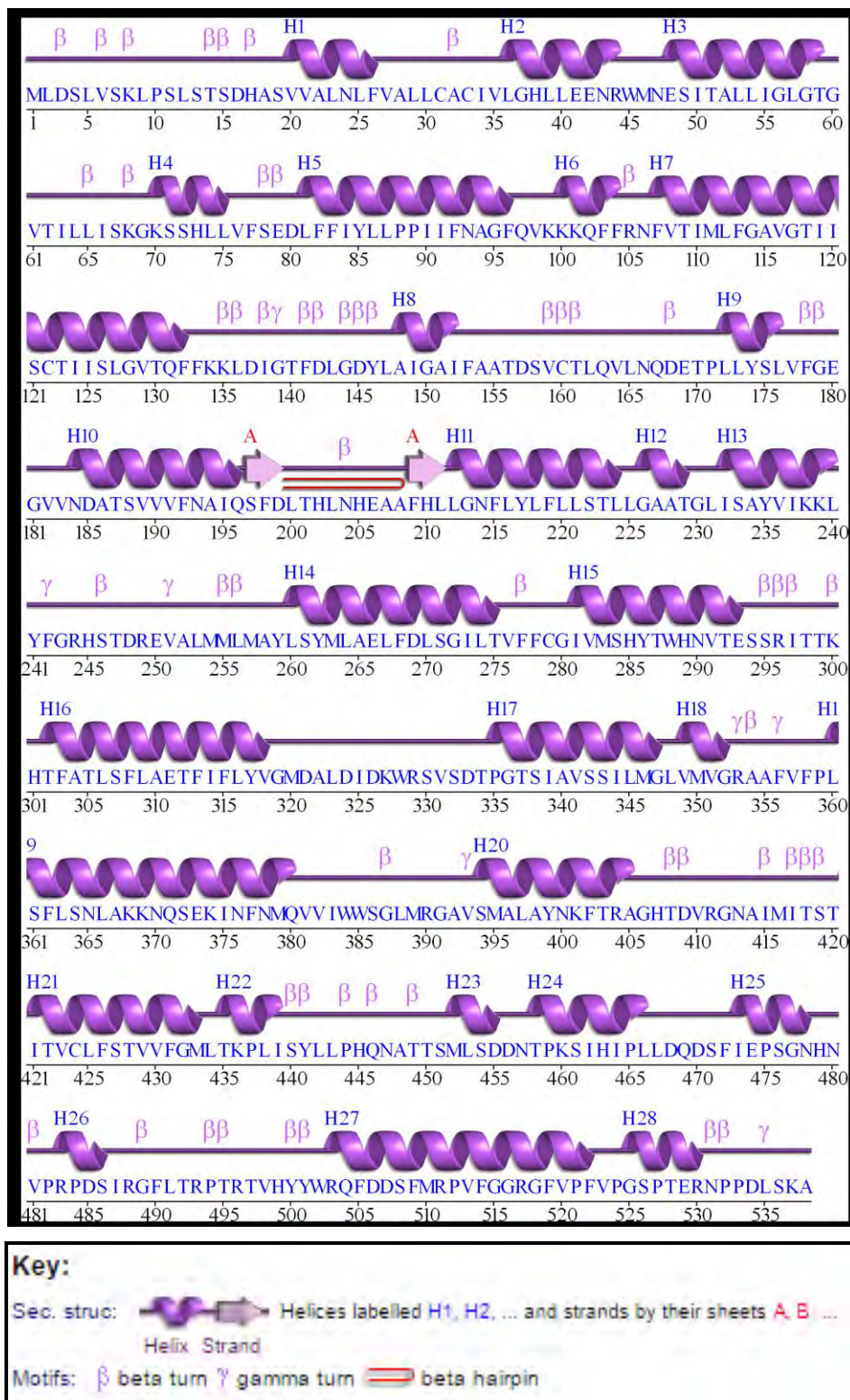


Figure 3.26: Schematic diagram showing the secondary structural elements in the final model attained from the PDBsum tool. The α -helices are labeled with the letter “H” and β -strands are lettered in the uppercase. β , γ and hairpin turns are also labeled.

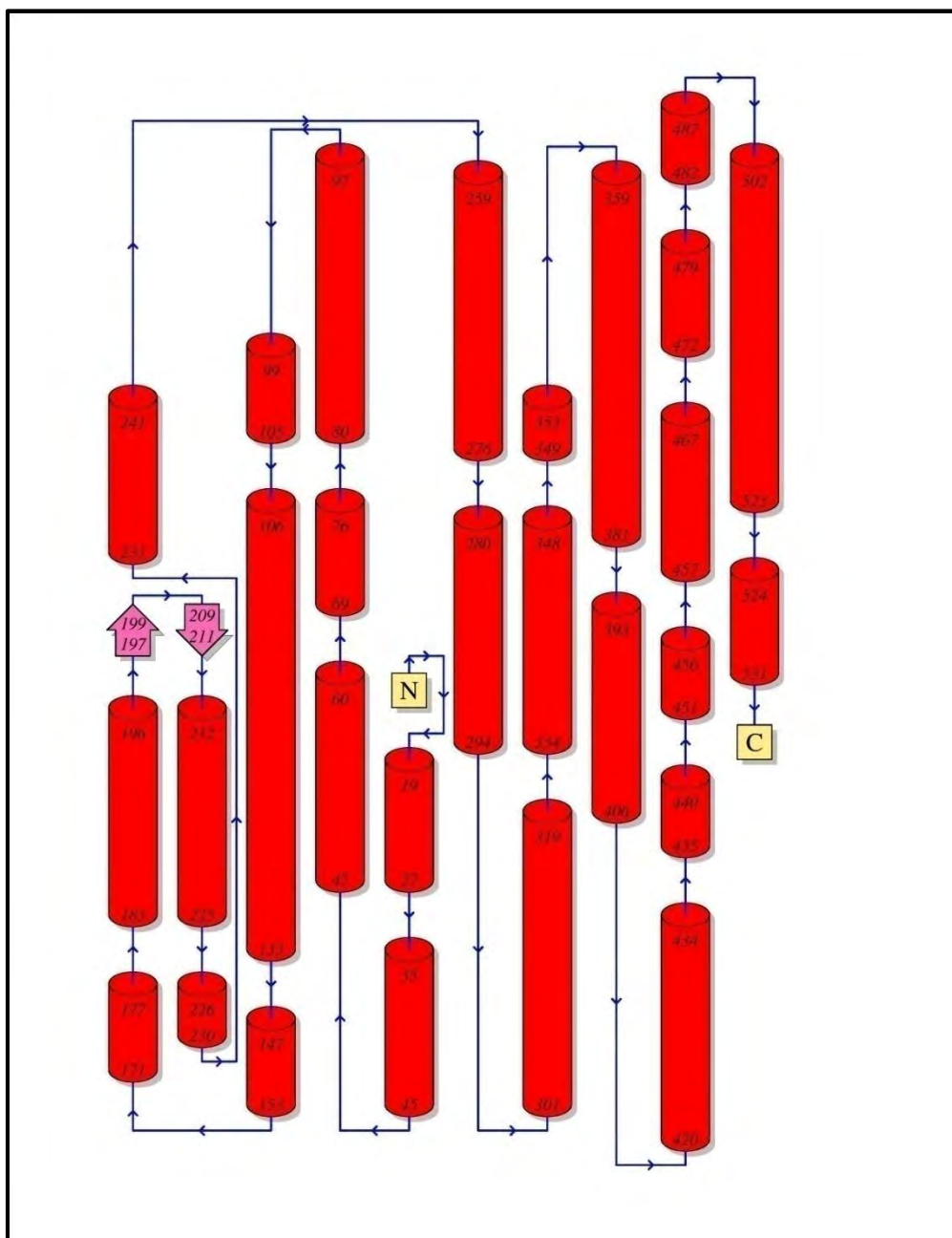


Figure 3.27: Topology diagram showing the secondary structural elements in the final model attained from the PDBsum tool. Helices are represented as cylinders and β -strands as arrows.

The classification of the β -hairpin i.e. 9:9 suggested that the end strands residues involved formed two hydrogen bonds (Sibanda *et al.*, 1989) (Figure 3.28). The β -hairpin is the essential structural subunit of a transmembrane β -barrel (Wimley, 2003). The transmembrane β -barrel may form transbilayer pores which act as possible models for a number of membrane channels (Sansom and Kerr, 1995).

Therefore, this suggested that the presence of the β -hairpin may allow the protein to act as membrane channel proteins while undergoing conformational changes due to changes in pH (Alberts *et al.*, 2002, Shaikh *et al.*, 2010). As such, the presence of channels would allow it to facilitate the exchange of Na^+ and H^+ . The immense number of β -turns indicated the presence of active sites and ligand binding surfaces within the protein model (Hutchinson and Thornton, 1996).

Prediction of secondary structures using NPS@ SOPMA and CYS-REC tools clearly showed that there were no disulfide bridges in the query sequence/structure. Under this circumstance it was predicted that the β -strands might have roles in giving this protein stability through protein folding and orienting structures for interaction. Hydrogen bonds in such cases play a key role. Figure 3.28 shows the H-bonds between the β -strands indicating the structural stability of the model.

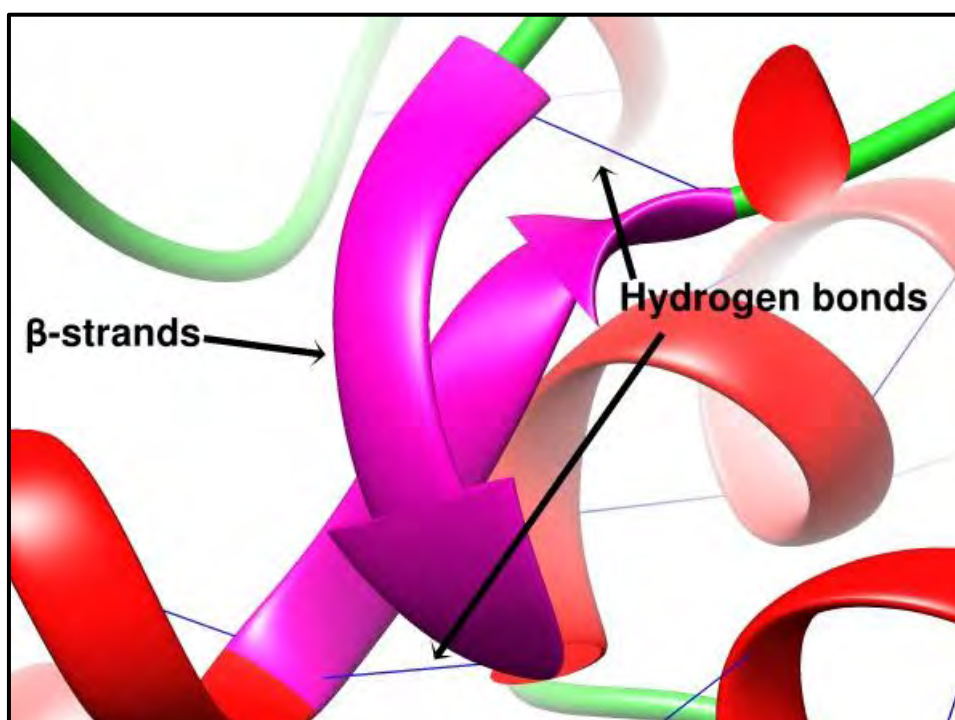


Figure 3.28: Hydrogen bonds between antiparallel β -strands forming β -hairpin within EM_Model 01

In future, further studies will be done to see how this structure interacts with its immediate molecules in signal transduction pathway. That information will reveal the interaction between various abiotic stress tolerances.

CHAPTER 4:
REFERENCES

Chapter 4: References

- Abe, N. and Mamitsuka, H. (1997) 'Predicting Protein Secondary Structure Using Stochastic Tree Grammars', *Machine Learning*, 29(2-3), pp. 275-301.
- Adamian, L. and Liang, J. (2002) 'Interhelical hydrogen bonds and spatial motifs in membrane proteins: Polar clamps and serine zippers', *Proteins: Structure, Function, and Bioinformatics*, 47(2), pp. 209-218.
- Akpinar, B. A., Avsar, B., Lucas, S. J. and Budak, H. (2012) 'Plant abiotic stress signaling', *Plant Signaling & Behavior*, 7(11), pp. 1450-1455.
- Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. (2002) 'Carrier Proteins and Active Membrane Transport', *Molecular Biology of the Cell*. 4th ed. New York: Garland Science.
- Apse, M. P., Sottosanto, J. B. and Blumwald, E. (2003) 'Vacuolar cation/H⁺ exchange, ion homeostasis, and leaf development are altered in a T-DNA insertional mutant of AtNHX1, the Arabidopsis vacuolar Na⁺/H⁺ antiporter', *The Plant Journal*, 36(2), pp. 229-239.
- Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., de Castro, E., Duvaud, S., Flegel, V., Fortier, A., Gasteiger, E., Grosdidier, A., Hernandez, C., Ioannidis, V., Kuznetsov, D., Liechti, R., Moretti, S., Mostaguir, K., Redaschi, N., Rossier, G., Xenarios, I. and Stockinger, H. (2012) 'ExPASy: SIB bioinformatics resource portal', *Nucleic Acids Research*, 40(Web Server issue), pp. W597-W603.
- Barragán, V., Leidi, E. O., Andrés, Z., Rubio, L., De Luca, A., Fernández, J. A., Cubero, B. and Pardo, J. M. (2012) 'Ion Exchangers NHX1 and NHX2 Mediate Active Potassium Uptake into Vacuoles to Regulate Cell Turgor and Stomatal Function in Arabidopsis', *The Plant Cell*, 24(3), pp. 1127-1142.
- Bassil, E., Coku, A. and Blumwald, E. (2012) 'Cellular ion homeostasis: emerging roles of intracellular NHX Na⁺/H⁺ antiporters in plant growth and development', *Journal of Experimental Botany*.

- Bassil, E., Ohto, M.-a., Esumi, T., Tajima, H., Zhu, Z., Cagnac, O., Belmonte, M., Peleg, Z., Yamaguchi, T. and Blumwald, E. (2011a) 'The Arabidopsis Intracellular Na(+)/H(+) Antiporters NHX5 and NHX6 Are Endosome Associated and Necessary for Plant Growth and Development', *The Plant Cell*, 23(1), pp. 224-239.
- Bassil, E., Tajima, H., Liang, Y.-C., Ohto, M.-a., Ushijima, K., Nakano, R., Esumi, T., Coku, A., Belmonte, M. and Blumwald, E. (2011b) 'The Arabidopsis Na(+)/H(+) Antiporters NHX1 and NHX2 Control Vacuolar pH and K(+) Homeostasis to Regulate Growth, Flower Development, and Reproduction', *The Plant Cell*, 23(9), pp. 3482-3497.
- Battisti, D. S. and Naylor, R. L. (2009) 'Historical Warnings of Future Food Insecurity with Unprecedented Seasonal Heat', *Science*, 323(5911), pp. 240-244.
- Benkert, P., Biasini, M. and Schwede, T. (2011) 'Toward the estimation of the absolute quality of individual protein structure models', *Bioinformatics*, 27(3), pp. 343-350.
- Benkert, P., Künzli, M. and Schwede, T. (2009) 'QMEAN server for protein model quality estimation', *Nucleic Acids Research*, 37(Web Server issue), pp. W510-W514.
- Benkert, P., Tosatto, S. C. E. and Schomburg, D. (2008) 'QMEAN: A comprehensive scoring function for model quality assessment', *Proteins: Structure, Function, and Bioinformatics*, 71(1), pp. 261-277.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. (2000) 'The Protein Data Bank', *Nucleic Acids Research*, 28(1), pp. 235-242.
- Betz, S. F. (1993) 'Disulfide bonds and the stability of globular proteins', *Protein Science : A Publication of the Protein Society*, 2(10), pp. 1551-1558.
- Blouin, C., Butt, D. and Roger, A. J. (2004) 'Rapid evolution in conformational space: A study of loop regions in a ubiquitous GTP binding

domain', *Protein Science : A Publication of the Protein Society*, 13(3), pp. 608-616.

- Blumwald, E. (1987) 'Tonoplast vesicles as a tool in the study of ion transport at the plant vacuole', *Physiologia Plantarum*, 69(4), pp. 731-734.
- Blumwald, E., Aharon, G. S. and Apse, M. P. (2000) 'Sodium transport in plant cells', *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1465(1-2), pp. 140-151.
- Boratyn, G. M., Camacho, C., Cooper, P. S., Coulouris, G., Fong, A., Ma, N., Madden, T. L., Matten, W. T., McGinnis, S. D., Merezhuk, Y., Raytselis, Y., Sayers, E. W., Tao, T., Ye, J. and Zaretskaya, I. (2013) 'BLAST: a more efficient report with usability improvements', *Nucleic Acids Research*, 41(W1), pp. W29-W33.
- Brett, C. L., Donowitz, M. and Rao, R. (2005) *Evolutionary origins of eukaryotic sodium/proton exchangers*.
- Buchan, D. W. A., Minneci, F., Nugent, T. C. O., Bryson, K. and Jones, D. T. (2013) 'Scalable web services for the PSIPRED Protein Analysis Workbench', *Nucleic Acids Research*, 41(Web Server issue), pp. W349-W357.
- Chanroj, S., Wang, G., Venema, K., Zhang, M. W., Delwiche, C. F. and Sze, H. (2012) 'Conserved and Diversified Gene Families of Monovalent Cation/H⁺ Antiporters from Algae to Flowering Plants', *Frontiers in Plant Science*, 3.
- Chothia, C. and Lesk, A. M. (1986) 'The relation between the divergence of sequence and structure in proteins', *The EMBO Journal*, 5(4), pp. 823-826.
- Claverie, J.-M. and Notredame, C. (2006) *Bioinformatics For Dummies*. 2nd edn.: Wiley Publishing, Inc., p. 456.
- Combet, C., Blanchet, C., Geourjon, C. and Deléage, G. (2000) 'NPS@: Network Protein Sequence Analysis', *Trends in Biochemical Sciences*, 25(3), pp. 147-150.

- Coordinators, N. R. (2013) 'Database resources of the National Center for Biotechnology Information', *Nucleic Acids Research*, 41(D1), pp. D8-D20.
- Curran, A. R. and Engelman, D. M. (2003) 'Sequence motifs, polar interactions and conformational changes in helical membrane proteins', *Current Opinion in Structural Biology*, 13(4), pp. 412-417.
- Darby, N. and Creighton, T. (1995) 'Disulfide Bonds in Protein Folding and Stability', in Shirley, B. (ed.) *Protein Stability and Folding Methods in Molecular Biology*TM: Humana Press, pp. 219-252.
- de Beer, T. A. P., Berka, K., Thornton, J. M. and Laskowski, R. A. (2013) 'PDBsum additions', *Nucleic Acids Research*.
- Elbashir, M., Wang, J., Wu, F.-X. and Wang, L. (2013) 'Predicting beta-turns in proteins using support vector machines with fractional polynomials', *Proteome Science*, 11(S1), pp. 1-10.
- Elofsson, A. and Heijne, G. v. (2007) 'Membrane Protein Structure: Prediction versus Reality', *Annual Review of Biochemistry*, 76(1), pp. 125-140.
- Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M.-y., Pieper, U. and Sali, A. (2006) 'Comparative Protein Structure Modeling Using Modeller', *Current protocols in bioinformatics / editorial board, Andreas D. Baxevanis ... [et al.]*, 0 5, pp. Unit-5.6.
- Fiser, A., Do, R. K. and Sali, A. (2000) 'Modeling of loops in protein structures', *Protein Science : A Publication of the Protein Society*, 9(9), pp. 1753-1773.
- Fukuda, A., Chiba, K., Maeda, M., Nakamura, A., Maeshima, M. and Tanaka, Y. (2004) 'Effect of salt and osmotic stresses on the expression of genes for the vacuolar H⁺-pyrophosphatase, H⁺-ATPase subunit A, and Na⁺/H⁺ antiporter from barley*', *Journal of Experimental Botany*, 55(397), pp. 585-594.

- Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S. e., Wilkins, M., Appel, R. and Bairoch, A. (2005) 'Protein Identification and Analysis Tools on the ExPASy Server', in Walker, J. (ed.) *The Proteomics Protocols Handbook*: Humana Press, pp. 571-607.
- Geer, L. Y., Marchler-Bauer, A., Geer, R. C., Han, L., He, J., He, S., Liu, C., Shi, W. and Bryant, S. H. (2010) 'The NCBI BioSystems database', *Nucleic Acids Research*, 38(suppl 1), pp. D492-D496.
- Geourjon, C. and Deléage, G. (1995) 'SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments', *Computer applications in the biosciences : CABIOS*, 11(6), pp. 681-684.
- Hussain, S. S., Raza, H., Afzal, I. and Kayani, M. A. (2011) 'Transgenic plants for abiotic stress tolerance: current status', *Archives of Agronomy and Soil Science*, 58(7), pp. 693-721.
- Hutchinson, E. G. and Thornton, J. M. (1996) 'PROMOTIF--a program to identify and analyze structural motifs in proteins', *Protein Science : A Publication of the Protein Society*, 5(2), pp. 212-220.
- Jewell, M., Campbell, B. and Godwin, I. (2010) 'Transgenic Plants for Abiotic Stress Resistance', in Kole, C., Michler, C., Abbott, A. & Hall, T. (eds.) *Transgenic Crop Plants*: Springer Berlin Heidelberg, pp. 67-132.
- Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S. and Madden, T. L. (2008) 'NCBI BLAST: a better web interface', *Nucleic Acids Research*, 36(suppl 2), pp. W5-W9.
- Jones, D. T. (1999) 'Protein secondary structure prediction based on position-specific scoring matrices1', *Journal of Molecular Biology*, 292(2), pp. 195-202.
- Kallberg, Y. (2002) *Bioinformatic methods in protein characterization*. PhD, Bioinformatics, Karolinska Institutet, Samuelssonsalen,

Scheelelaboratoriet, Tomtebodavägen 6, Karolinska Institutet [Online]
Available at: <http://hdl.handle.net/10616/39896> (Accessed.

- Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. and Sternberg, M. J. E. (2015) 'The Phyre2 web portal for protein modeling, prediction and analysis', *Nat. Protocols*, 10(6), pp. 845-858.
- Komatsu, S., Konishi, H. and Hashimoto, M. (2007) 'The proteomics of plant cell membranes', *Journal of Experimental Botany*, 58(1), pp. 103-112.
- Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E. L. L. (2001) 'Predicting transmembrane protein topology with a hidden markov model: application to complete genomes¹', *Journal of Molecular Biology*, 305(3), pp. 567-580.
- Kuntal, B. K., Aparoy, P. and Reddanna, P. (2010) 'EasyModeller: A graphical interface to MODELLER', *BMC Research Notes*, 3, pp. 226-226.
- Kyte, J. and Doolittle, R. F. (1982) 'A simple method for displaying the hydropathic character of a protein', *Journal of Molecular Biology*, 157(1), pp. 105-132.
- Laskowski, R., Rullmann, J. A., MacArthur, M., Kaptein, R. and Thornton, J. (1996) 'AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR', *Journal of Biomolecular NMR*, 8(4), pp. 477-486.
- Laskowski, R. A., MacArthur, M. W., Moss, D. S. and Thornton, J. M. (1993) 'PROCHECK: a program to check the stereochemical quality of protein structures', *Journal of Applied Crystallography*, 26(2), pp. 283-291.
- Li, W., Cowley, A., Uludag, M., Gur, T., McWilliam, H., Squizzato, S., Park, Y. M., Buso, N. and Lopez, R. (2015) 'The EMBL-EBI bioinformatics web and programmatic tools framework', *Nucleic Acids Research*, 43(W1), pp. W580-W584.

- Lovell, S. C., Davis, I. W., Arendall, W. B., de Bakker, P. I. W., Word, J. M., Prisant, M. G., Richardson, J. S. and Richardson, D. C. (2003) 'Structure validation by C α geometry: ϕ, ψ and C β deviation', *Proteins: Structure, Function, and Bioinformatics*, 50(3), pp. 437-450.
- Mahdi, R. (2014) *Identification of synchronized role of transcription factors, genes and enzymes in Arabidopsis thaliana under four Abiotic stress responsive pathways*. Master of Science in Biotechnology, BRAC University [Online] Available at: <http://hdl.handle.net/10361/3166> (Accessed.
- Marban, E., Yamagishi, T. and Tomaselli, G. F. (1998) 'Structure and function of voltage-gated sodium channels', *The Journal of Physiology*, 508(Pt 3), pp. 647-657.
- Martí-Renom, M. A., Stuart, A. C., Fiser, A., Sánchez, R., Melo, F. and Šali, A. (2000) 'Comparative Protein Structure Modeling of Genes and Genomes', *Annual Review of Biophysics and Biomolecular Structure*, 29(1), pp. 291-325.
- McWilliam, H., Li, W., Uludag, M., Squizzato, S., Park, Y. M., Buso, N., Cowley, A. P. and Lopez, R. (2013) 'Analysis Tool Web Services from the EMBL-EBI', *Nucleic Acids Research*, 41(W1), pp. W597-W600.
- Miseta, A. and Csutora, P. (2000) 'Relationship Between the Occurrence of Cysteine in Proteins and the Complexity of Organisms', *Molecular Biology and Evolution*, 17(8), pp. 1232-1239.
- Morris, A. L., MacArthur, M. W., Hutchinson, E. G. and Thornton, J. M. (1992) 'Stereochemical quality of protein structure coordinates', *Proteins: Structure, Function, and Bioinformatics*, 12(4), pp. 345-364.
- Oberai, A., Ihm, Y., Kim, S. and Bowie, J. U. (2006) 'A limited universe of membrane protein families and folds', *Protein Science : A Publication of the Protein Society*, 15(7), pp. 1723-1734.
- Orłowski, J. and Grinstein, S. (2011) 'Na⁺/H⁺ Exchangers', *Comprehensive Physiology*: John Wiley & Sons, Inc.

- Pardo, J. M., Cubero, B., Leidi, E. O. and Quintero, F. J. (2006) 'Alkali cation exchangers: roles in cellular homeostasis and stress tolerance', *Journal of Experimental Botany*, 57(5), pp. 1181-1199.
- Petersen, B., Lundegaard, C. and Petersen, T. N. (2010) 'NetTurnP – Neural Network Prediction of Beta-turns by Use of Evolutionary Information and Predicted Protein Sequence Features', *PLoS ONE*, 5(11), pp. e15079.
- Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C. and Ferrin, T. E. (2004) 'UCSF Chimera—A visualization system for exploratory research and analysis', *Journal of Computational Chemistry*, 25(13), pp. 1605-1612.
- Ramachandran, G. N., Ramakrishnan, C. and Sasisekharan, V. (1963) 'Stereochemistry of polypeptide chain configurations', *Journal of Molecular Biology*, 7(1), pp. 95-99.
- Ramachandran, S. and Dokholyan, N. (2012) 'Homology Modeling: Generating Structural Models to Understand Protein Function and Mechanism', in Dokholyan, N.V. (ed.) *Computational Modeling of Biological Systems Biological and Medical Physics, Biomedical Engineering*: Springer US, pp. 97-116.
- Razzaque, S. (2011) *CLONING OF AN ANTIporter GENE (NUX1) FROM PCRAMPLICONS INTO A RECOMBINATION COMPETENT VECTORCONTAINING A CONSTITUTIVE PROMOTER (CaMV35S)*. Masters of Science in Biotechnology, BRAC University [Online] Available at: <http://hdl.handle.net/10361/1665> (Accessed).
- Razzaque, S., Mahdi, R. and Islam, A. (2014) 'Identification of Synchronized Role of Transcription Factors, Genes, and Enzymes in Arabidopsis thaliana under Four Abiotic Stress Responsive Pathways', *Computational Biology Journal*, 2014, pp. 13.

- Rodríguez-Rosales, M. P., Gálvez, F. J., Huertas, R., Aranda, M. N., Baghour, M., Cagnac, O. and Venema, K. (2009) 'Plant NHX cation/proton antiporters', *Plant Signaling & Behavior*, 4(4), pp. 265-276.
- Rodríguez-Rosales, M. P., Jiang, X., Gálvez, F. J., Aranda, M. N., Cubero, B. and Venema, K. (2008) 'Overexpression of the tomato K⁺/H⁺ antiporter LeNHX2 confers salt tolerance by improving potassium compartmentalization', *New Phytologist*, 179(2), pp. 366-377.
- Ronald, P. (2011) 'Plant Genetics, Sustainable Agriculture and Global Food Security', *Genetics*, 188(1), pp. 11-20.
- Roy, A., Kucukural, A. and Zhang, Y. (2010) 'I-TASSER: a unified platform for automated protein structure and function prediction', *Nature protocols*, 5(4), pp. 725-738.
- Saitou, N. and Nei, M. (1987) 'The neighbor-joining method: a new method for reconstructing phylogenetic trees', *Molecular Biology and Evolution*, 4(4), pp. 406-425.
- Sali, A. (1995) 'Comparative protein modeling by satisfaction of spatial restraints', *Molecular Medicine Today*, 1(6), pp. 270-277.
- Sansom, M. S. and Kerr, I. D. (1995) 'Transbilayer pores formed by beta-barrels: molecular modeling of pore structures and properties', *Biophysical Journal*, 69(4), pp. 1334-1343.
- Senes, A., Ubarretxena-Belandia, I. and Engelman, D. M. (2001) 'The Ca—H···O hydrogen bond: A determinant of stability and specificity in transmembrane helix interactions', *Proceedings of the National Academy of Sciences of the United States of America*, 98(16), pp. 9056-9061.
- Shaikh, S., Wen, P.-C., Enkavi, G., Huang, Z. and Tajkhorshid, E. (2010) 'Capturing Functional Motions of Membrane Channels and Transporters with Molecular Dynamics Simulation', *Journal of computational and theoretical nanoscience*, 7(12), pp. 2481-2500.

- Sibanda, B. L., Blundell, T. L. and Thornton, J. M. (1989) 'Conformation of β -hairpins in protein structures', *Journal of Molecular Biology*, 206(4), pp. 759-777.
- Sievers, F. and Higgins, D. (2014) 'Clustal Omega, Accurate Alignment of Very Large Numbers of Sequences', in Russell, D.J. (ed.) *Multiple Sequence Alignment Methods Methods in Molecular Biology*: Humana Press, pp. 105-116.
- Sottosanto, J., Saranga, Y. and Blumwald, E. (2007) 'Impact of AtNHX1, a vacuolar Na⁺/H⁺ antiporter, upon gene expression during short- and long-term salt stress in *Arabidopsis thaliana*', *BMC Plant Biology*, 7(1), pp. 18.
- Suganya, P. R., Sudevan, K., Kalva, S. and Saleena, L. M. (2014) 'HOMOLOGY MODELING FOR HUMAN ADAM12 USING PRIME, I-TASSER AND EASYMODELLER', *International Journal of Pharmacy & Pharmaceutical Sciences*.
- Swarbreck, S. M., Colaço, R. and Davies, J. M. (2013) 'Plant Calcium-Permeable Channels', *Plant Physiology*, 163(2), pp. 514-522.
- Tamura, K., Stecher, G., Peterson, D., Filipski, A. and Kumar, S. (2013) 'MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0', *Molecular Biology and Evolution*, 30(12), pp. 2725-2729.
- Wang, Y. and Jardetzky, O. (2002) 'Probability-based protein secondary structure identification using combined NMR chemical-shift data', *Protein Science : A Publication of the Protein Society*, 11(4), pp. 852-861.
- White, S. H. (2004) 'The progress of membrane protein structure determination', *Protein Science : A Publication of the Protein Society*, 13(7), pp. 1948-1949.
- Wimley, W. C. (2003) 'The versatile β -barrel membrane protein', *Current opinion in structural biology*, 13(4), pp. 404-411.

- Wöhlert, D., Kühlbrandt, W. and Yildiz, Ö. (2014) 'Structure and substrate ion binding in the sodium/proton antiporter PaNhaP', *eLife*, 3, pp. e03579.
- Xiong, J. (2006) *Essential Bioinformatics*. United States of America: Cambridge University Press, p. 352.
- Yamaguchi, T., Apse, M. P., Shi, H. and Blumwald, E. (2003) 'Topological analysis of a plant vacuolar Na(+)/H(+) antiporter reveals a luminal C terminus that regulates antiporter cation selectivity', *Proceedings of the National Academy of Sciences of the United States of America*, 100(21), pp. 12510-12515.
- Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J. and Zhang, Y. (2015) 'The I-TASSER Suite: protein structure and function prediction', *Nat Meth*, 12(1), pp. 7-8.
- Zhang, Y. (2008) 'I-TASSER server for protein 3D structure prediction', *BMC Bioinformatics*, 9, pp. 40-40.
- Zheng, C. and Kurgan, L. (2008) 'Prediction of beta-turns at over 80% accuracy based on an ensemble of predicted secondary structures and multiple alignments', *BMC Bioinformatics*, 9, pp. 430-430.
- ≤ali, A. (1995) 'Comparative protein modeling by satisfaction of spatial restraints', *Molecular Medicine Today*, 1(6), pp. 270-277.