IN SILICO STRUCTURAL ANALYSIS, PHYSICOCHEMICAL CHARACTERIZATION AND HOMOLOGY MODELING OF *Arabidopsis thaliana* Na^+/H^+ EXCHANGER 2 (AtNHX2) PROTEIN

B.S. THESIS

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

Submitted by: Prateem Das

ID: 12136014

Biotechnology Program

Department of Mathematics and Natural Sciences

BRAC University

Bangladesh

May 2017

“You’re only given one little spark of madness. You mustn’t loose it.”
– Robin Williams

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 2 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:

_____

Prateem Das

Certified:

_____

Dr. Aparna Islam
Supervisor
Associate Professor
Biotechnology Program
Department of Mathematics and Natural Sciences
BRAC University, Dhaka

Acknowledgement

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 obligated 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 grateful to Eusra Mohammad, Lecturer BRAC University for aiding and providing me with valuable insights regarding the technical aspects of this project.

I would like to express my sincere thanks to Abira Khan (Lecturer, University of Dhaka), for her help during the initial stages of this study.

Sincerely thanking Mohammad Rafid Feisal, for without his help this would be next to impossible.

I would like to thank the seniors, laboratory officers, laboratory assistants as well as teaching assistants at the department for their support and advice.

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

Abstract

Along with climate change, salinity has a detrimental effect on plants. It can cause susceptibility towards pathogens, halt proper growth and thus lead to food insecurity. Therefore, it is necessary to comprehend the ways in which plants adapt to abiotic stress. NHX- antiporter facilitates the exchange of Na^+ and H^+ across the membranes and these proteins help plants to tackle salinity by compartmentalizing Na^+ from cytoplasm to vacuoles via generated electrochemical gradient by H^+ pumps, hence considered as one of the significant protein that is capable to catalyze Na^+ in vacuoles. These classes of proteins are exhibited in many plants such as Arabidopsis Thaliana. As such, it is necessary to determine the structure of the protein encoded by the NHX gene in Arabidopsis thaliana. Determining the structure can reveal the protein-protein interaction network and identify regions such as, active sites, in the protein. Also, the structure between AtNHX1 and AtNHX2 can be compared to find the similarities and differences between them using in silico approach. Bio-computational analyses of the target protein were performed using an array of online bioinformatics tools and databases and the homology model was developed using 4 different software (I-TASSER, Phyre^2, Easy Modeller and Swiss Model) and the best model was selected upon evaluation. In addition, the secondary structural motifs were identified within the model. The results suggested that structure generated by Swiss Model, was the best amongst the four. It had the highest stereo-chemical quality scores and was considered to be the least unusual. The model consisted of α/β/γ 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^+. These factors were also present in AtNHX1, which further validates the structure. However, the position of the β-hairpin in AtNHX1 is different to AtNHX2. The β-hairpin on AtNHX1 was found in between amino acid 197 (Serine) till 199 (Aspartine). The second strand ranged from residue 209 (Phenylalanine) till residue 211 (Leucine). In AtNHX2, the beta hairpin spanned across amino acid residue 366 (Serine) and 367(Asparagine) and the other from amino acid residue 373 (Glutamine) to 374 (Serine). Moreover, the two proteins exhibited different classes of β-hairpin. However, further research is necessary to understand functional difference in between AtNHX1 and AtNHX2.

Table of content:

Chapter 1

1.1 Introduction

In various molecular pathways, new proteins are continuously discovered on a routine basis. However, most of the structures of the proteins are unknown. The structures of the proteins are necessary for understanding their molecular function, the mechanism in which they act, and their overall role in the biological pathway (Ramachandran & Dokholyan, 2012).

Membrane proteins are fundamental part in understanding the chemistry of the cell. These proteins transport vital steroid molecules and ions through the plasma membrane thus initiating various chained reactions and altering the chemistry of the cells (Elofsson & Heijne, 2007). Hence, the membrane protein acts as a media for the cells to interact with the external environment (Komatsu, Konishi, & Hashimoto, 2007). The methods that are currently available for attaining atomic-resolution structures of these biomolecules are through X-ray crystallography and NMR spectroscopy. However, it is very difficult to get high-resolution, three-dimensional structures of these proteins, as the quantity required for three-dimensional structure generation is very high in concentration and purity (Ramachandran & Dokholyan, 2012).

The plant cells are composed of several membrane systems performing numerous specialized functions. For instance, the plasma membrane functions as a communication interface with the outside environment for the exchange of substances and signal transduction (Komatsu, Konishi, & Hashimoto, 2007).

The structure of the protein determines its function. The regions of protein that span across the membrane, the orientation across the membrane, fold type of the amino acid sequences all contribute towards the function (Elofsson & Heijne, 2007). Moreover, fundamental structures, such as Alpha helix and Beta-sheets play crucial role in protein characterization. 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, Larsson, von Heijne , & Sonnhammer, 2001). However, β-barrel membrane proteins are difficult to identify by sequence gazing (Elofsson & Heijne, 2007).

1.2 Salinity, the problems and solutions

Salinity refers to a high concentration of soluble salts on soil. Salinity in soil, is one of the current major abiotic problems faced globally both by the producers of agriculture products and the scientific community. Ions and minerals are vital in growth of plants as it is required for growth, various enzyme activity and hormonal responses. However, over abundance of the ions can lead to plant growth depletion and can even cause death. One of the major reasons of increasing salinity is considered to be global warming. This escalation in temperature causing overall rises in the sea level and thus increasing salinity in soil (Feisal, 2015).

Salinity in soil is measured in a unit called EC_e. This stands for electrical conductivity of a particular soil sample. The higher the conductivity the more is the concentration of the salt present in the soil. EC_e is measured by making a saturated paste extract of the soil sample. Then the conductivity of the soil paste is measured. An EC_e value of $1dSm^{-1}$ values for 10mM of sodium chloride present in the solution and a value of a soil sample of 40mM or above is considered to be saline (Munns, 2005).

Soil sample can be classified into three different categories based on the concentration of the salt contamination in the soil. It can be sodic. Sodic soils have low concentration of soluble salts but have a high exchangeable sodium ions percentage. This phenomenon is denoted as ESP (excess exchangeable sodium). The soil sample is considered as ESP if the value of EC_e is equal or more than 15. As a result the clay part of the soil separates from the humus and the drainage of the soil becomes poor. The soil gets waterlogged when the condition is wet and gets hard when dry (Brown, 2005). In addition to being sodic, the soil can also be alkaline. Alkaline soils are a type of sodic soil with high pH of 8.5 to 10.

These overall affects the growth of the plants and determine the vegetation of the whole area.

Salinity is a major problem for growth of plants because it affects mechanisms within plants, controlling shoot and root growth. The mechanisms controlling this phase of the growth response are not specific to salinity however they are caused by factors associated with water stress. Hormone signals are responsible for leaf growth and most of these chemical signals come from roots. In dry or in saline conditions salinity effects on root elongation rate thus affecting the water retention on plant's leaves and this leading to poor leaf growth (Hu & Schmidhalter, 1998) (Jeschke, 1984) (Jeschke, Aslam, & Greenway, 1986).

A stress in salinity can cause adverse affect on plant shoot to root growth ratio causing primary and secondary stress leading to death (Yeo, 1998). Primary stress includes:

- Water stress or osmotic stress
- Sodium specific stress

Salinity causes soil water potential to be more negative than plant cell. Sodium ion has smaller radius than potassium ion therefore having high charge to mass ratio, disrupting and lowering hydrophobic interaction with proteins (Amtmann & Sanders, 1999) and causing adaptive structural changes, such as, closure of stomata to reduce water loss. These actions causes less photosynthesis in the system thus causing ROS accumulation (reactive oxidative species) and also further causing membrane disorganization, metabolic toxicity, attenuated nutrient acquisition (Yeo, 1998) (Duan, 2001).

Plants can tolerate and survive salinity mainly in two possible responses from the organism (Kumar & Mosa, 2015), which includes:

- Physiological response
- Bio-chemical response

1.3 Physiological response

- Salt Exclusion: Most of the plants use a method known as salt exclusion and this provides certain level of tolerance in plants. It is the ability to exclude salts occurs through filtration at the surface of the root. Root membranes prevent salt from entering while allowing the water to pass through. The red mangrove is an example of a salt-excluding species (Gilbert, Chang, & Rojas, 2002). The salt exclusion can be very high as much as by 98% in bread wheat plants and barley (Munns, 2005).

- Avoidance: Avoidance is the process of keeping the salt ions away from the parts of the plant where they are harmful (Munns, 2005; Brown, 2005).

- Salt excretion/extrusion: Plants that follows this process remove salt through glands or bladders or cuticle located on each leaf (Gilbert, Chang, & Rojas, 2002).

- Salt Dilution: this is achieved by dilution of ions in the tissue of the plant by maintaining succulence. Plants achieve this by increasing their storage volume by developing thick, fleshy, succulent structures. Succulence is mainly a result of vacuoles of mesophyll cells filling with water and increasing in size (Gilbert, Chang, & Rojas, 2002).

1.4 Bio-chemical response

- Compartmentalization of ions: Ions are compartmentalized in both cellular and organ level. In organ level, the roots retain and filter most of the ions in comparison to shoots, mainly leaves. On a cellular level, ions are compartmentalized in vacuoles thus protecting the enzymes.

- Tolerance: These methods include osmotic adjustments and hormone productions, such as, ABA stress hormones, which make plants resistant to the excess salts.

All of these methods depend on cellular ion homeostasis, which mainly depends on net intracellular toxic ion uptake and subsequent vacuolar compartmentalization without toxic ion accumulation on cystol (Kumar & Mosa, 2015).

Negative electric potentials influx on cells in roots (mainly caused by sodium ions) occurs through ion channels present in cell membranes. These ions mainly include calcium, magnesium and sulphur ions (Ca^{2+}, Mg^{2+}, SO^{4-}). These ions results in production of ROS (reactive oxygen species), such as, oxygen free radicals, hydrogen peroxide and hydroxyl ions (O^{2-}, H_2O_2, OH^-), which induces water stress, ion toxicity, nutritional disorders, reduction of cellular division and cellular expansion, alteration of metabolic process, membrane disorganization and geno-toxicity (Kumar & Mosa, 2015).

However, the level of effect varies within plant species as many of the of the plants developed various methods, as discussed above, which generally maintains a high potassium to sodium ratio, mostly on shoots, to avoid affects of salinity. These result in halophytic species that can grow in high salt concentration (Kumar & Mosa, 2015).

Transportation of sodium ion is done over various plasma membranes through various groups of protein transporters, such as, SOS, NSCCs that are generalized as the following categories (Waters, Gilliham, & Hrmova, 2013):

- Uniporters and Sympoters
- antiporters

1.4.1 Uniporters and Sympoters

These transporters include proteins include “plant high affinity transporters” (HKTs), “cat-ion chloride co transporters” (CCCs) and other similar proteins. The natures of the proteins are that it can transport ions in one direction, either inside or the outside of the cells. Some can only transport the sodium ions while others can transport sodium ions and potassium ions. (Corratge-Faillie, Jabnoune,

Zimmermann, Very, Fizames, & Sentenac, 2010) (Waters, Gilliham, & Hrmova, 2013).

Sympoters are group of transporters that transports osmolytes (compounds affecting osmosis) to tackle stress and induce osmoprotection. They mainly transport charged molecules, such as, glycine betaine, dimethyl-1, 1,4-carboxyl sulfonium propionate (DMPS) or neutral compounds, such as, sucrose, fructose, glycerol, raffinose. These compounds affect the osmotic potential of the plant cells (Delauney & Vermas, 1993) (Trossat, Rathinasabapathi, Weretylnik, Shen, & Huang, 1998) (Nanjo, Kobayashi, Yoshiba, Kakubari, Yamaguchi-Shinozaki, & Shinozaki, 1999).

Both of these protein channels can transport ions only in one direction (figure 1).

1.4.2 Antiporters

Antiporters are carriers which can transport ions both ways, from inside to outside of the cells and vice versa. These groups of proteins operate at a cellular level regulating and compartmentalization of ions inside the cells. These include channel proteins like Na^+/H^+ antiporters (NHX), cat-ion antiporters (CAX), auto inhibited Ca^{2+}-ATPase (ACA) and proton pumps, such as, V-H^+-P Phase and Vacoular H^+-ATPase (Blumwald & Poole, 1985) (Rodriguez-Rosaic, Galvez, & Huertas, 2009) (figure 1).

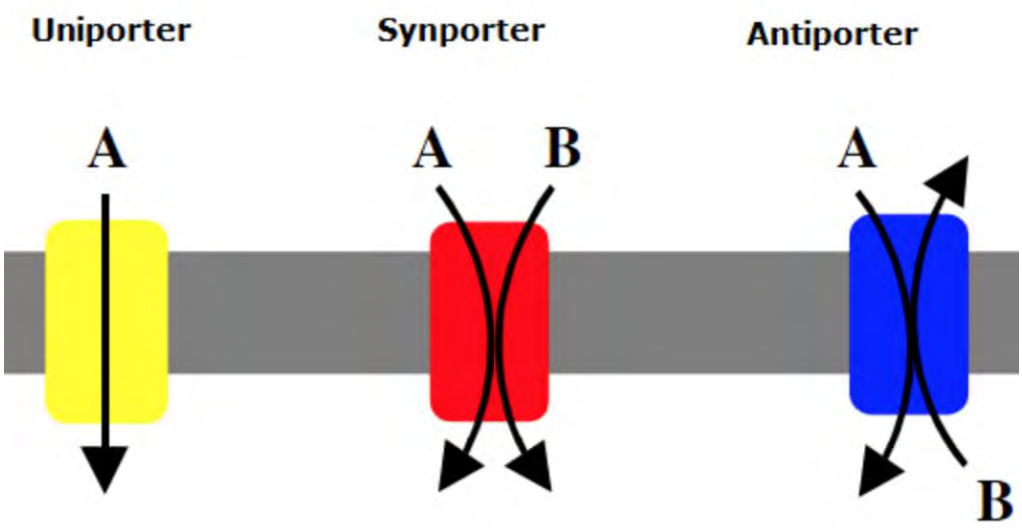


Figure 1: Different type of transporters present in a plant

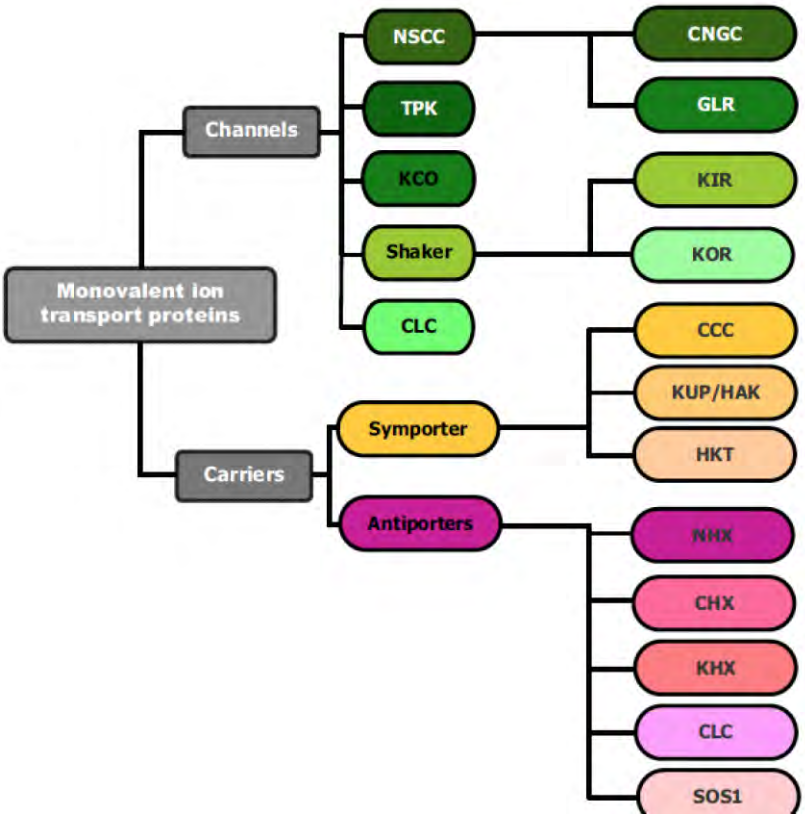


Figure 2: Total channels and carriers that regulate the osmotic potential of the plants (Mian, Senadheera, & Maathuis, 2011)

#	Proteins	Abbreviation
1	NSCC	Non-selective cation channel
2	TPK	Two-pore K+ channel
3	KCO	K+ outward rectifying channel
4	CLC	Voltage gated Cl- channel
5	CNGC	Cyclic nucleotide gated channel
6	GLR	Glutamate like receptor
7	KIR	Shaker type K+ inward rectifier
8	KOR	Shaker type K+ outward rectifier
9	CCC	Cat-ion chloride co-transporter
10	KUP/HAK	K+ uptake permease
11	HKT	High affinity K+ transporter
12	NHX	Na+/H+ exchanger
13	CHX	Cat-ion/H+ exchanger
14	KHX	K+/H+ exchanger
15	SOS	Salt overly sensitive transporter

Table 1: the total types of transporters present in the plant cells.

Figure 2 and table 1 demonstrates the various transporters present in a typical plant cell. On the other hand, figure 3 shows the distribution of each of these transporters throughout the plant.

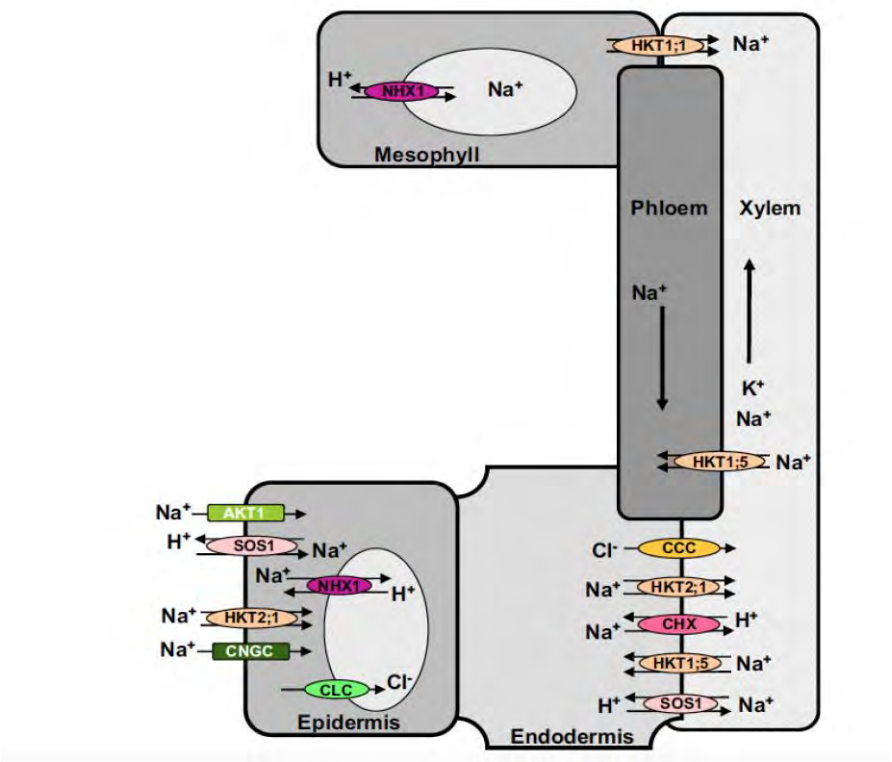


Figure 3: Distribution of the various transporter proteins in a plant (Mian, Senadheera, & Maathuis, 2011)

1.5 NHX transporters

The NHX transport protein is by far the most well studied and most observed in plant ion transport system, as it is one of the most heavily relied protein channel used by plants for osmoregulation. It is a multipurpose transport protein serving various functions. It plays a major role in recirculation and distribution of the sodium ions from shoots to roots (Sunarpi, Motoda, & Kubo, 2005). The NHX proteins are tonoplast bound and are driven by electrochemical gradient of the protons generated by other transport proteins (Rodriguez-Rosaic, Galvez, & Huertas, 2009). First NHX antiporter proteins activity was reported and directly measured in red beet storage tissue later as the study continued it was found predominantly on 60 different plant species (Blumwald & Poole, 1985). NHX proteins are involved in both transportation of sodium ions and potassium ions. Broadly they are divided into two groups, Class Class II and I. Class II resides mainly on endosomes and helps in potassium ions homeostasis and Class I NHX

proteins resides on tonoplast (Biasini, et al., 2014). Overall there are eight isoforms of NHX proteins found in the system of *Arabidopsis Thaliana* (Yokoi, Quintero, & Cubero, 2002). AtNHX1, AtNHX2, AtNHX3 and AtNHX4 are all intracellular mainly found in vacuole membranes, AtNHX5 and AtNHX6 are found on endosomes and the last two of the members AtNHX7/SOS1 and AtNHX8 are divergent members found in various parts of the plants (Bassil, Coku, & Blumwald, 2012) (Yokoi, Quintero, & Cubero, 2002). The C-terminal end of the NHX protein (*Arabidopsis*) regulates the cat-ion selectivity through the binding to calmodulin like protein forming AtCaMI5 in a calcium ions and pH dependent manner (Yamaguchi, Aharon, Sottosanto, & Blumwald) (figure 4).

1.6 NHX1 and NHX2 transporter proteins

Considering all of the transporters present in a plant cell, NHX1 protein plays a major role in plant salt tolerance. It is usually found in plant tonoplast and plasma membrane thus helps plants tackle salt tolerance on a cellular level. (Kumar & Mosa, 2015). It is considered as the major protein the plants use for regulation of cellular ion homeostasis (Feisal, 2015). On the other hand, NHX2 protein is usually found in the small vacuoles present in the plant cells and endosomes. Though NHX2 protein is involved in ion homeostasis, it is not as heavily involved as NHX1 protein. However, it is believed that NHX2 protein is involved in flowering as it is mainly involved in potassium ion regulation. (Wang, LI, Zhang, yang , Zheng, & Xie, 2007)

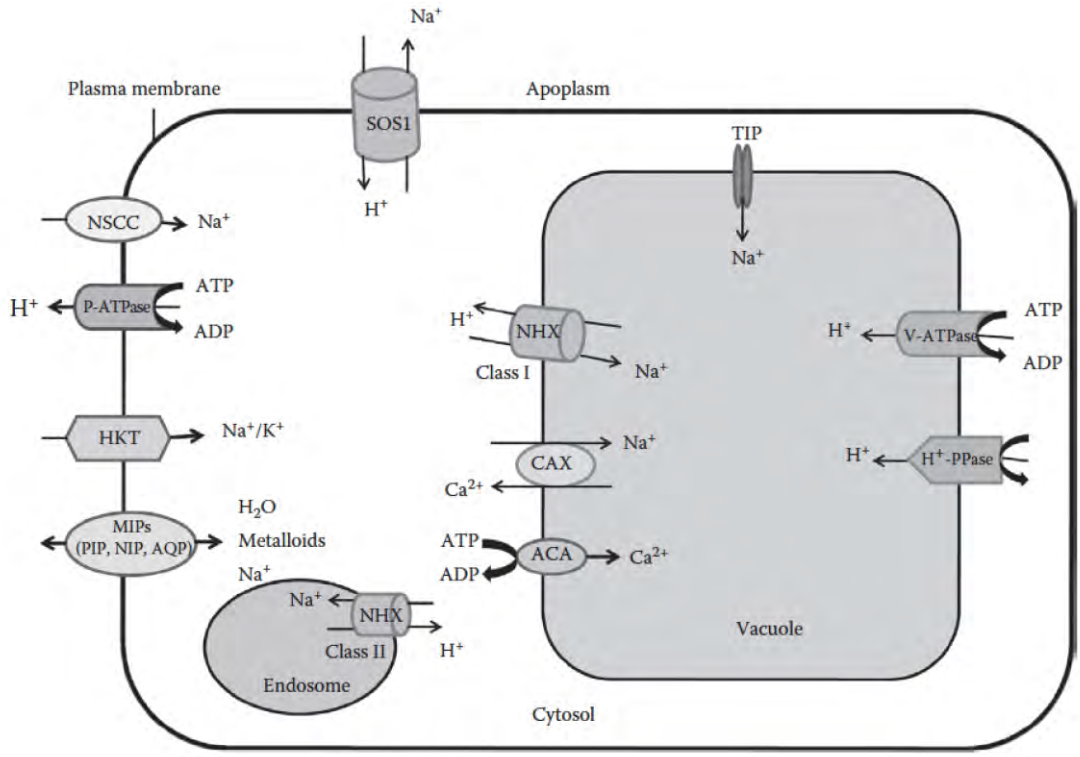


Figure 4: Schematic representation of primary and secondary transport in the plant cells. Electrogenic H+ transport (H^+-ATPase in the plasma membrane and vacuolar membrane, H^+-PPiase in the vacuolar membrane) generates gradients of pH and electrical potential difference across the cell and vacuolar membranes. Compartmentalization of sodium ions occurs, entering into the cells, into vacuoles by NHX and different proteins (Kumar & Mosa, 2015)**.**

1.7 Benefit of determining the structure (in silico) of the Na+/H+ exchanger 2 (NHX2) of *Arabidopsis thaliana*:

NHX protein is one of the most well documented and well-studied transporter proteins. However, only NHX 1 is well documented and studied and not done with latter proteins of the same family group. By predicting the three dimensional structure of the protein along with its physio-chemical analysis will help us to understand better why these proteins are clustered in the same family and on the same note why they are different.

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 different, novel proteins (Kalberg, 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 in better understanding of the function and property of the protein thus shedding new lights to already existing knowledge of the overall network of the proteins involved in stabilizing the osmotic potential of a plant cell.

1.8 Homology modeling and its significance

Currently the methods required for determination of any protein structure is through NMR spectroscopy and X-ray crystallography, which demands high amount of protein in purest form. However, obtaining that quantity of protein is next to impossible for now. Moreover, NMR has size restrictions, which is one of the reasons why there are many proteins out there that are still not yet modeled. Comparative modeling on the other hand gives an alternate reliable method in speculating the structure and predicts the function accordingly (Ramachandran & 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 & Lesk, 1986). Therefore various, reliable and significantly similar protein structures (templates) can be used for model generation and molecular model query whose structure is unknown.

Chapter 2

Materials, Method and Process

2.1 The Approach and the work plan

In this study, various online tools and databases were accessed and used to gain and analyze the desired gene and protein sequences, *in silico* method. These include several online and offline software accessed to predict the three-dimensional structure, the physiochemical characteristics and the structure validation of the target protein. The steps taken for achieving the results are all listed in figure 5 (a, b,c).

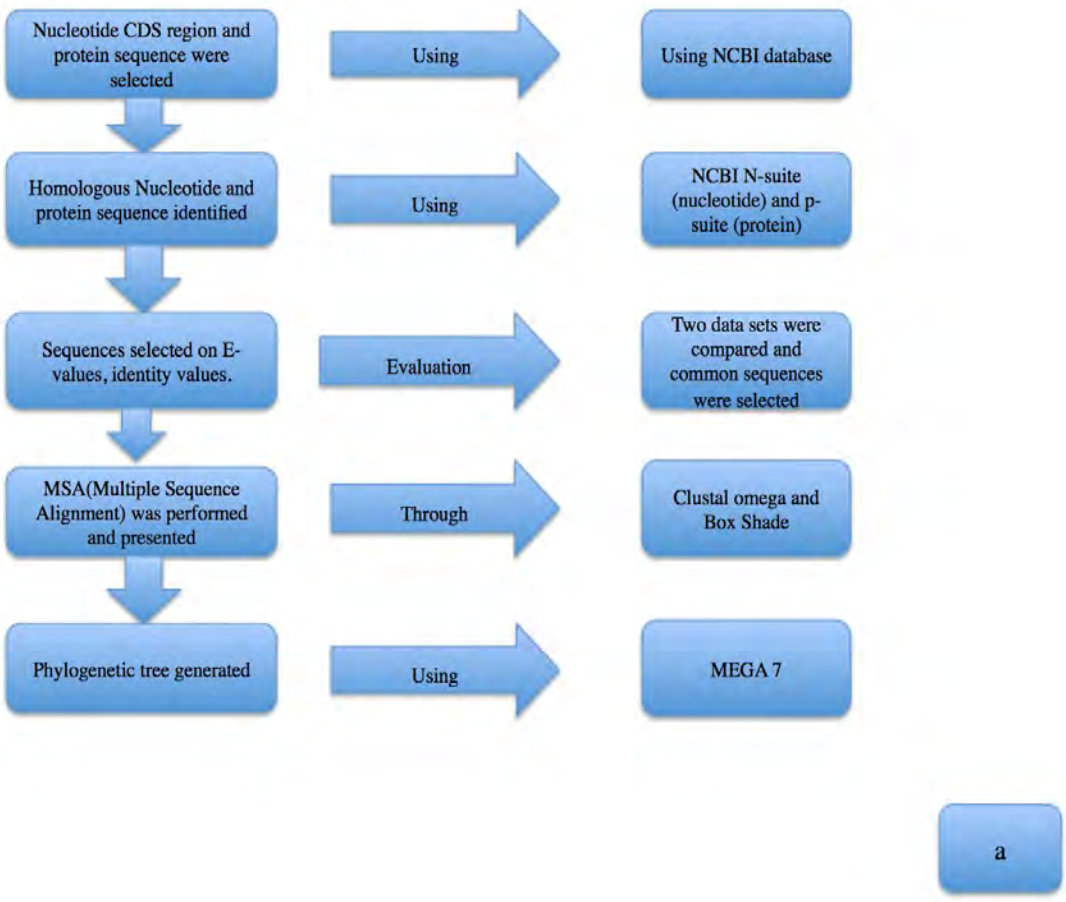


Figure 5 a: Data retrieval process and the characteristics of amino acid determination

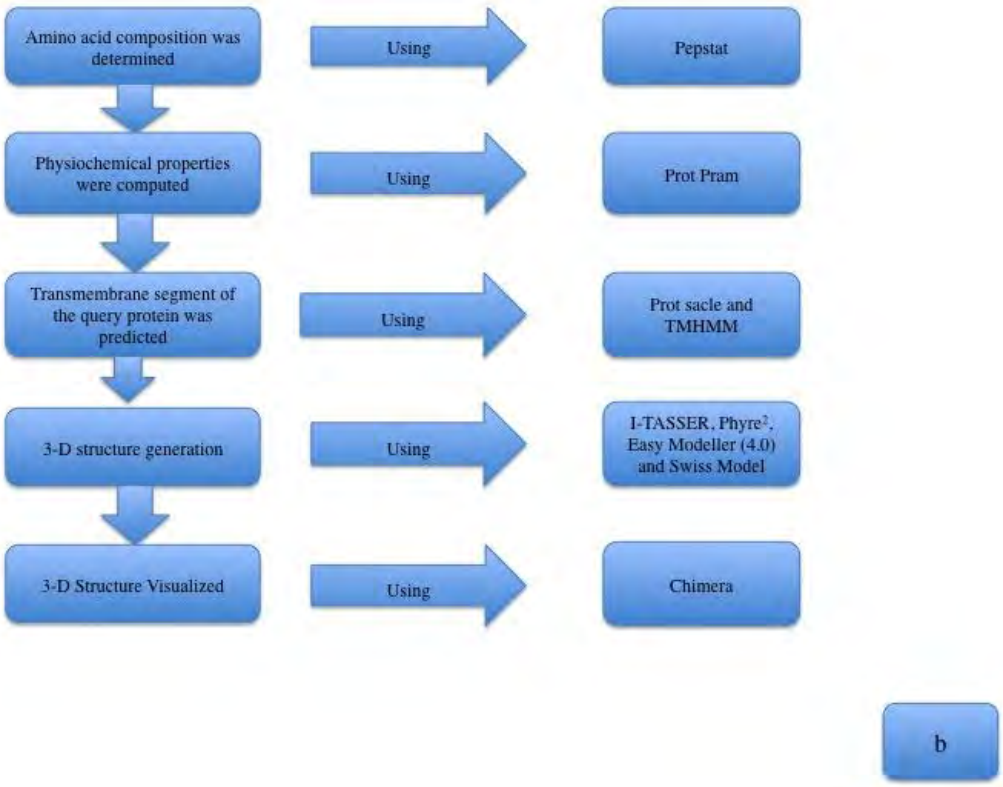


Figure 5 b: 2-Dimension and 3-dimensional structure generation

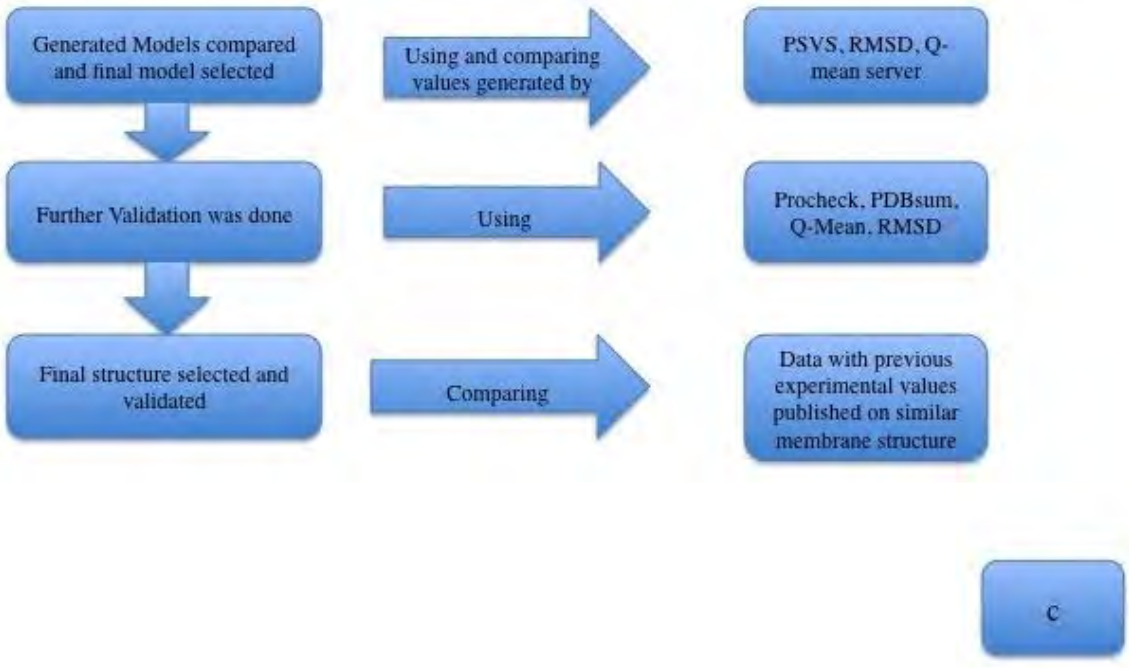


Figure 5 c: Model validation

Description of the different tools used and databases accessed during this process:

2.2. Databases:

2.2.1 NCBI (National Center for Biotechnology Information)

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

The National Center for Biotechnology (NCBI) generates databases accessible for public and focuses researches on computational biotechnology, develops software tools for genome data analysis and circulating biomedical information, helping in better understanding of the molecular behavior of different biochemical process affecting human health and disease (Geer, et al., 2010). The site was established in 1988 as a national resource for molecular biology information and uses Entrez as search engine for providing information (figure 6.1)

The database was accessed to (nucleotide and protein database) retrieve target nucleotide and protein sequences of the *Arabidopsis thaliana* relevance to the current study. The AtNHX2 gene and corresponding protein sequence was retrieved from the site. The accession ID: NM_111375 (*Arabidopsis thaliana* sodium/hydrogen exchanger 2 mRNA, complete *cds*)

The area of the coding region, CDS, which was within the entire sequence, was selected and from the graphics view (figure 6.2) as represented in green and then corresponding protein sequence in relation to the CDS region was retrieved. The retrieved Protein sequence accession ID was: NP_187154

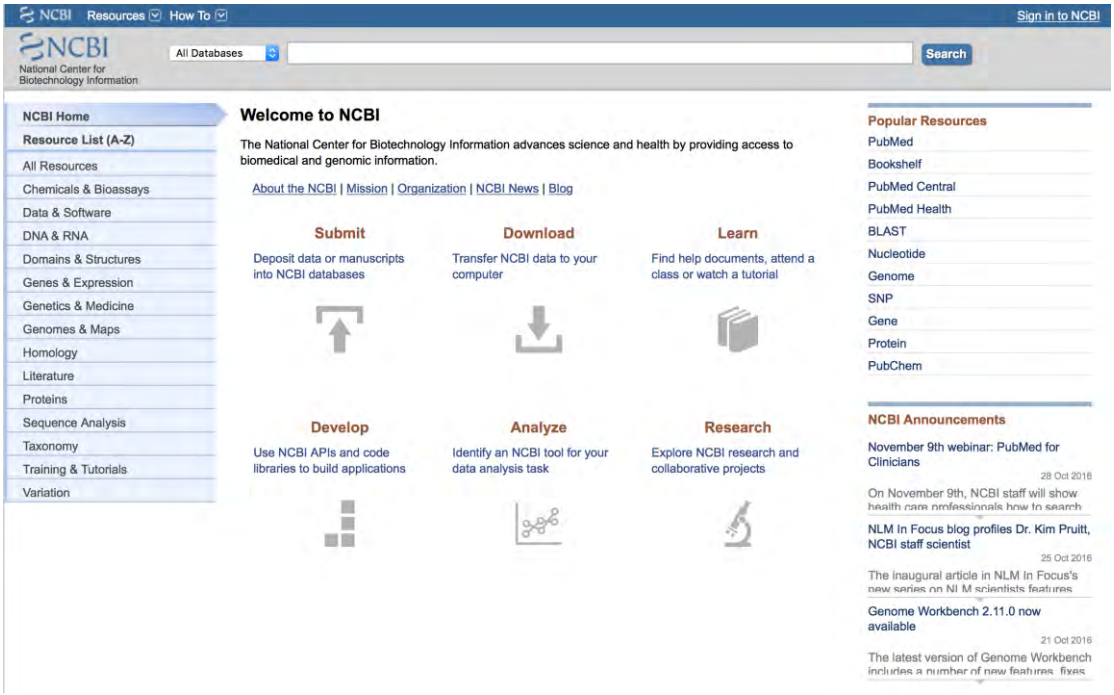


Figure 6.1: NCBI Homepage

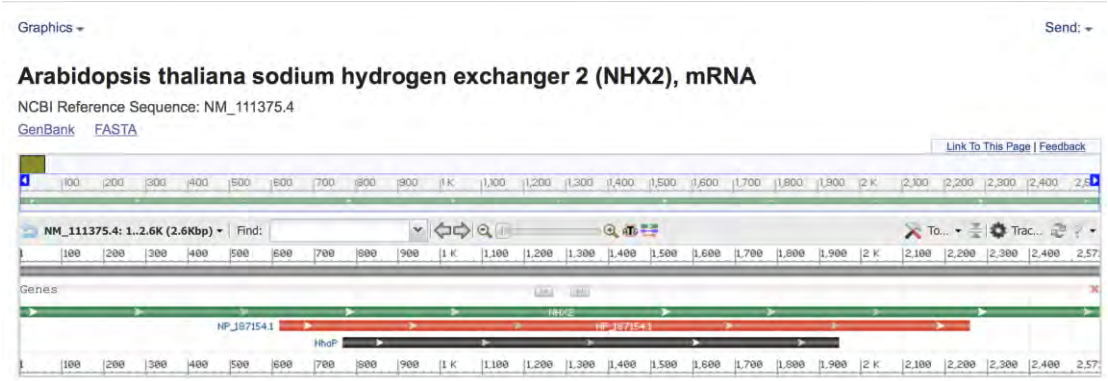


Figure 6.2: Graphical representation of the whole gene and the coding region of the sequence

2.2.2 Expert Protein Analysis System (ExPASy)

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

Operated by Swiss Institute of Bioinformatics, it is an extensible and integrative resource portal allowing access to various online database and software tools focusing in different

areas of life science. Along with access to several domains containing proteomics, genomics transcriptomics and population genetics it also has links to external institutions (Artimo, et al., 2012) (figure 7).

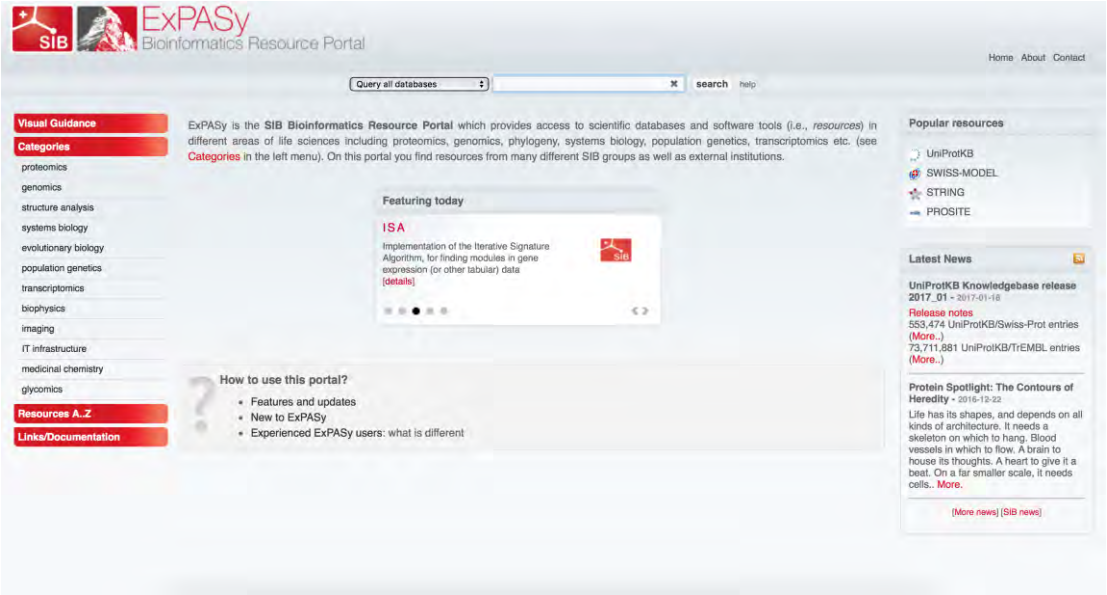


Figure 7: ExPASy Homepage

2.2.3 EMBL-EBI

The European Bioinformatics Institute (EBI) is a freely available resource site for life science experiment focusing on computational biology. It is apart of European Molecular Biology Laboratory (EMBL). The academic research institute is located on Welcome Trust Genome Campus in Hinxton near Cambridge in UK. (Figure 8)

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

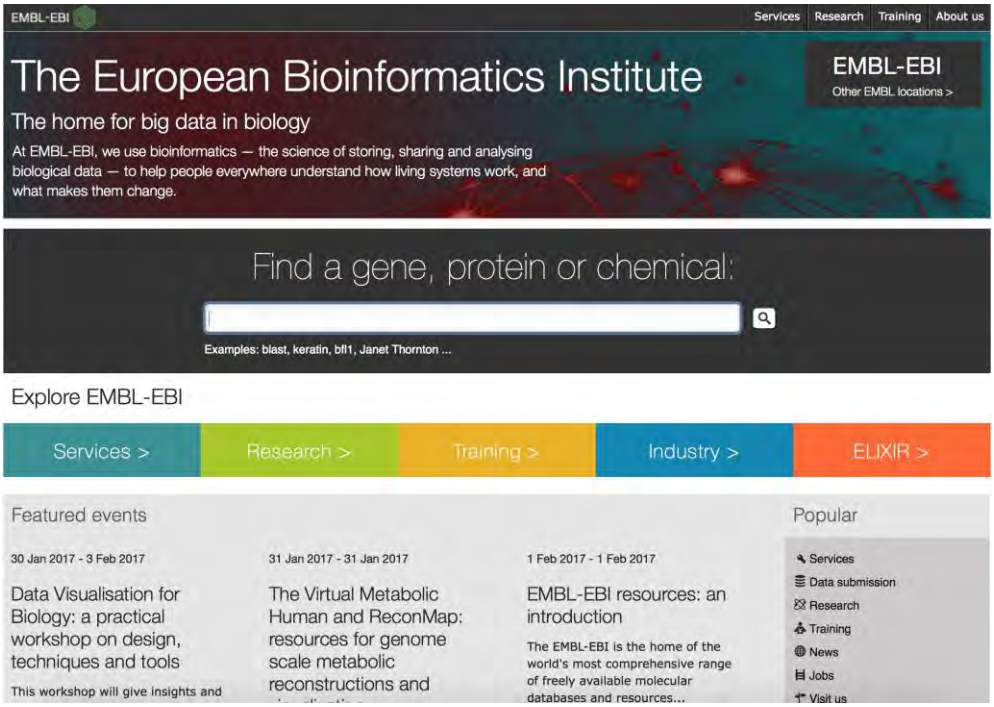


Figure 8: EMBL-EBI homepage

2.2.4 Protein Data Bank (PDB)

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

Protein Data Bank (PDB) is a crystallographic database of large biological molecules, such as, enzymes, transport proteins and nucleic acid. The data are obtained trough X-ray crystallography or NMR spectroscopy submitted by biologists from all over the world thus making it a vast three-dimensional (3D) library. Overseen by organization named world wide Protein Data Bank (wwPDB), it was accessed for template retrieval for generating homology modeling. (Figure 9)

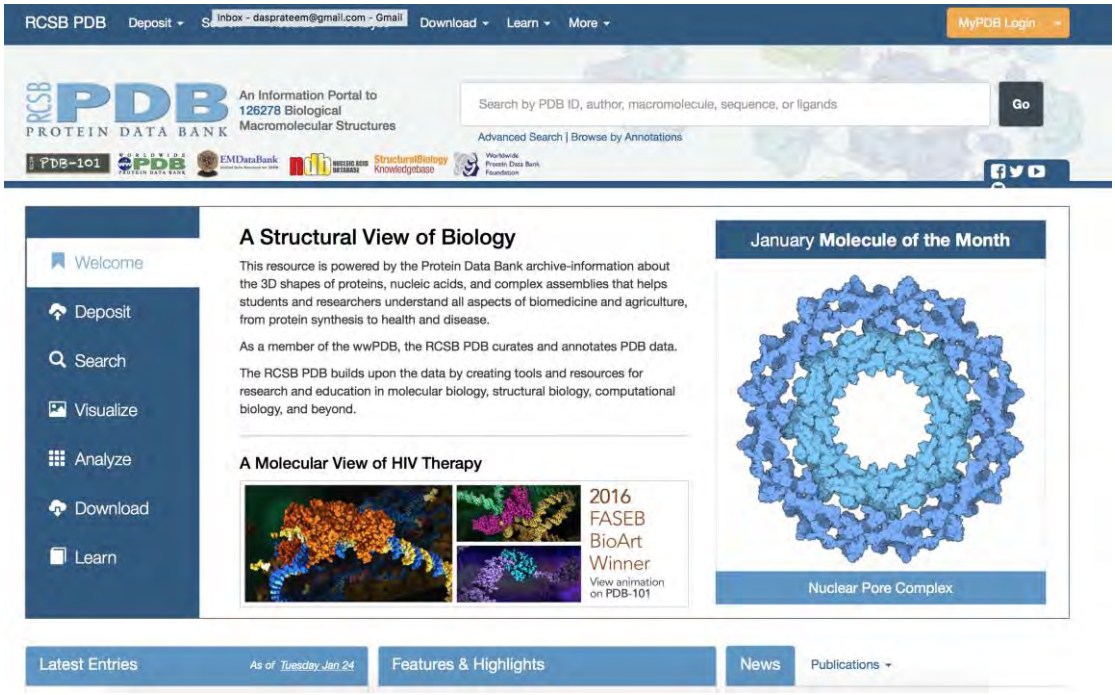


Figure 9: PDB homepage

2.2.5 String

URL link: http://string-db.org

Maintained by the Swiss Institute of Bioinformatics it is a database, which gives access on various experimental data on protein interactions and the state of the query protein. Various types of identifiers are recognized by the system and a full-text search on gene annotations is conducted in parallel to aid in the identification. Using the search results, STRING either recognizes automatically or asks the user to disambiguate, the organism of interest (Szklarczyk, et al., 2015) (figure 10).

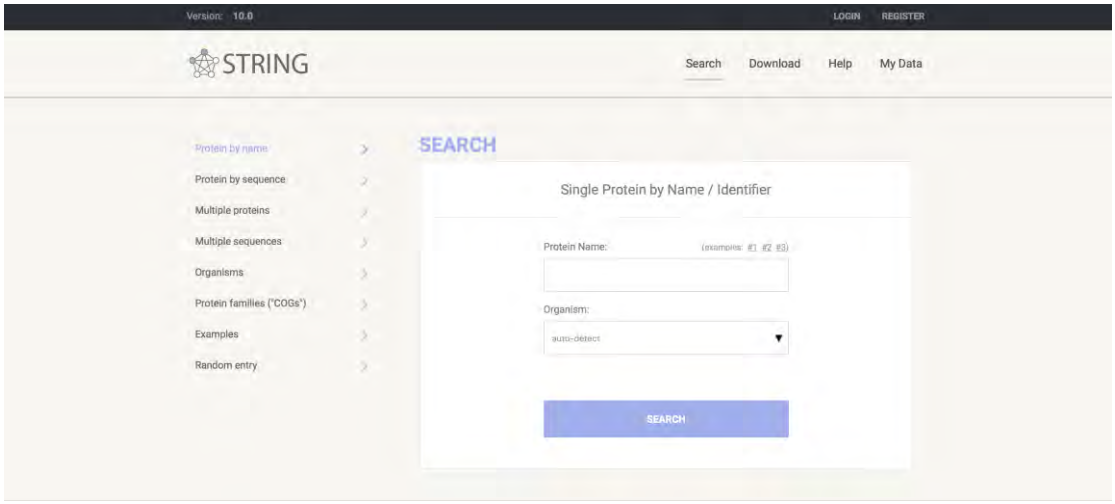


Figure 10: String Homepage

2.2 Tools

2.3.1 BLAST

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

The Basic Local Alignment Search tool (BLAST) recognizes similarities between sequences on local regions on a query gene. The program calculates statistical significance of matches between protein and nucleotide sequences with the sequence database (Coordinators, 2013) (Boratyn, et al., 2013) (Johnson, Zaretskaya, Raytselis, Merezhuk, McGinnis, & Madden, 2008). This tool can be used to deduce functional and evolutionary significance between two sequences and help in identifying members of gene families. In this study both N-suite and Blast P-suite were used (figure 11)

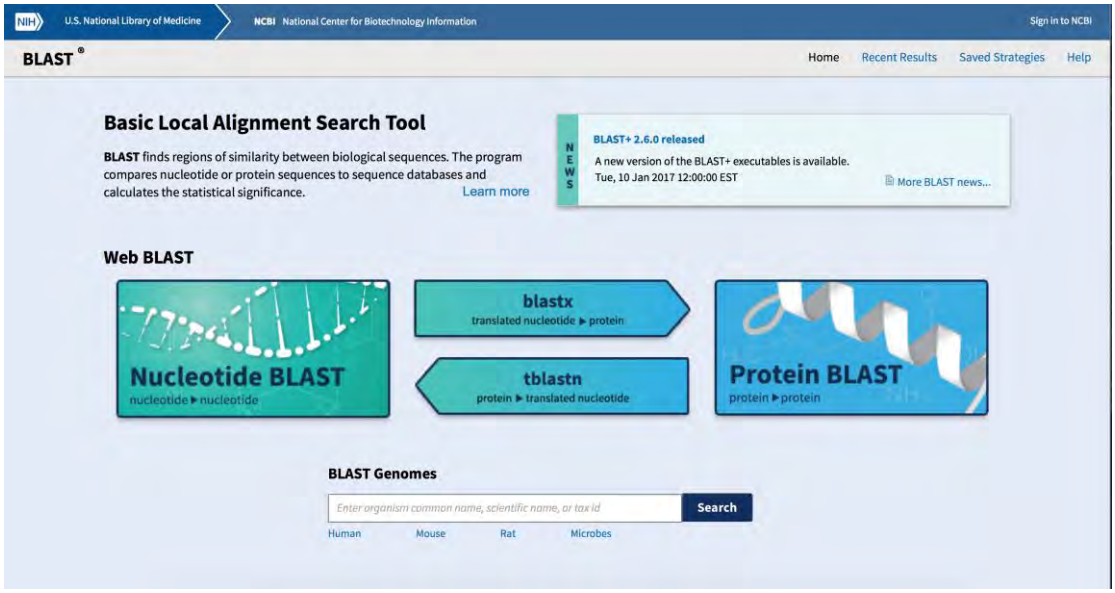


Figure 11: BLAST Homepage

2.3.2 Clustal Omega

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

Clustal Omega is a revised and refined version of the previously used Clustal series, namely Clustal W. Although it still uses HH-align and hidden Markov model, the system can now handle larger sequences and now can be ran from command line or online (Sievers & Higgins, 2014). (Figure 12)

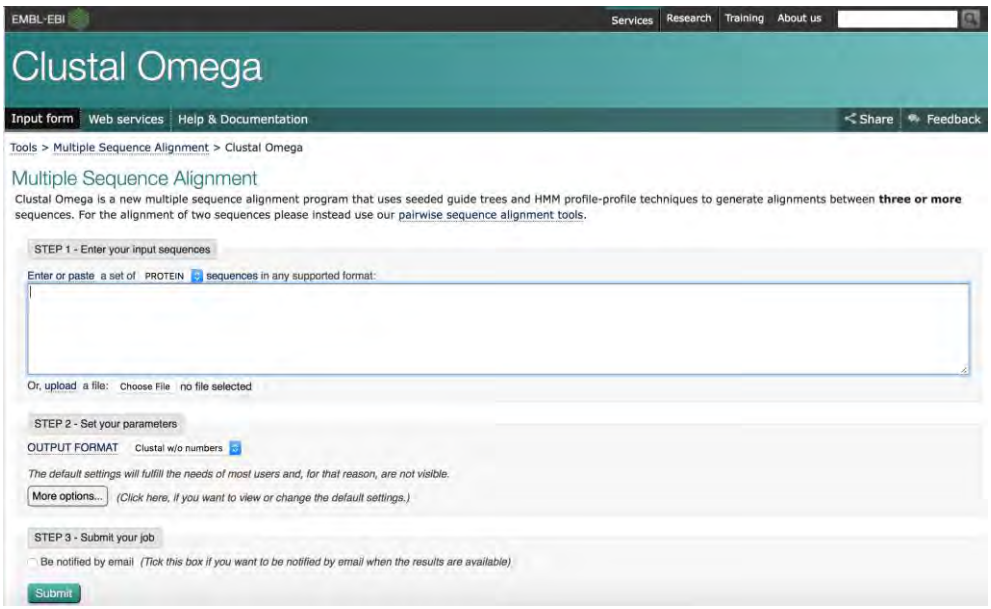


Figure 12: Clustal Omega Homepage

2.3.3 Molecular Evolutionary Genetics Analysis (MEGA)

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

Molecular Evolutionary Genetics Analysis (MEGA) is an integrated tool for sequence alignment, estimating divergence times, inferring phylogenetic tree, database mining, estimation of molecular evolution rate and testing evolutionary hypothesis. It is used for evolutionary history of reconstruction and theorizing extent and nature of forces, which selectively induce and shape the evolution of genes and species (Kumar, Stecher, & Tamura, 2015). This tool was used for generation of phylogenetic tree of selected protein sequences. Mega 7 was used for this current study (figure 13). Using the Phylogeny option in MEGA 7, 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. 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. For gaps and missing data, “complete deletion” was selected.

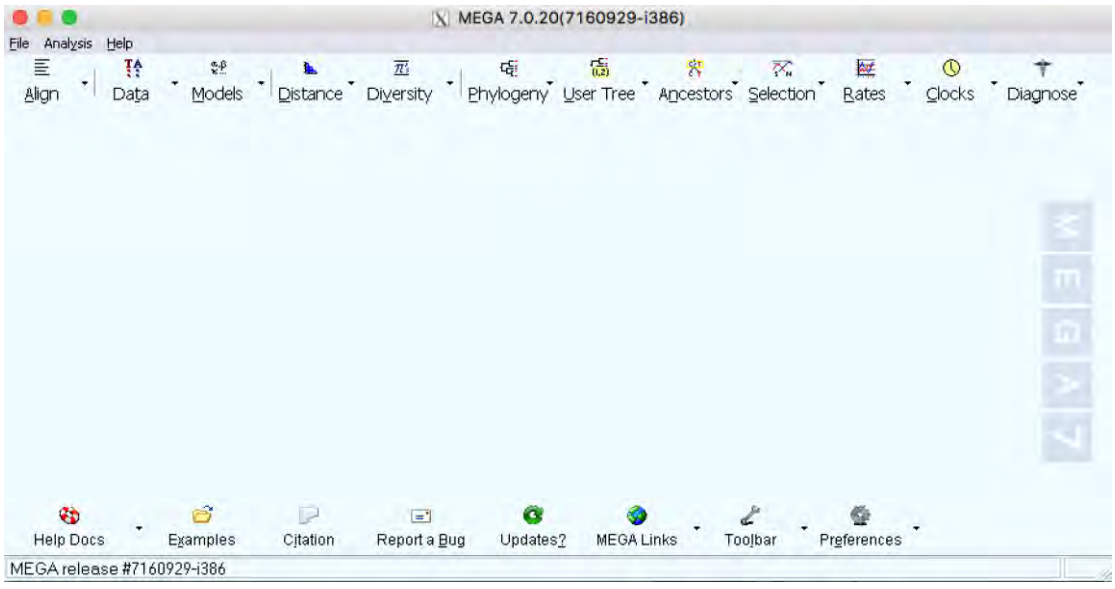


Figure 13: MEGA Homepage

2.3.4 BoxShade

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

Boxshade is a program, which it does not carry out the alignment but serves the purpose of viewing the multiple sequence alignment done by other tools, as such Clustal Omega. This program is known for pretty printing multiple alignment output. The output format selected for the current study was RTF new (figure 14).

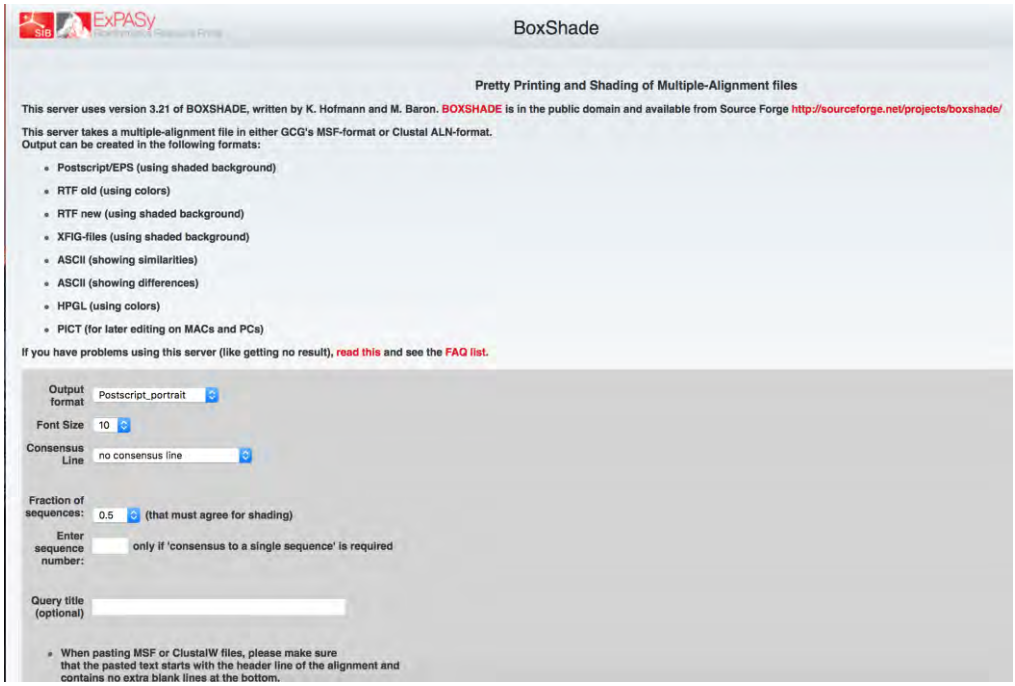


Figure 14: Boxshade Homepage

2.3.5 Pepstats

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

Pepstat is an online analysis tool for statistical protein analysis calculation. Part of EMBL-EBI website it provides the user with various physiochemical analysis of the query protein, such as, the isoelectric point, molecular weight extinction coefficient as well as amino acid compositions of the various query proteins by their FASTA sequence (figure 15) (Li, et al., 2015) (McWilliam, et al., 2013).

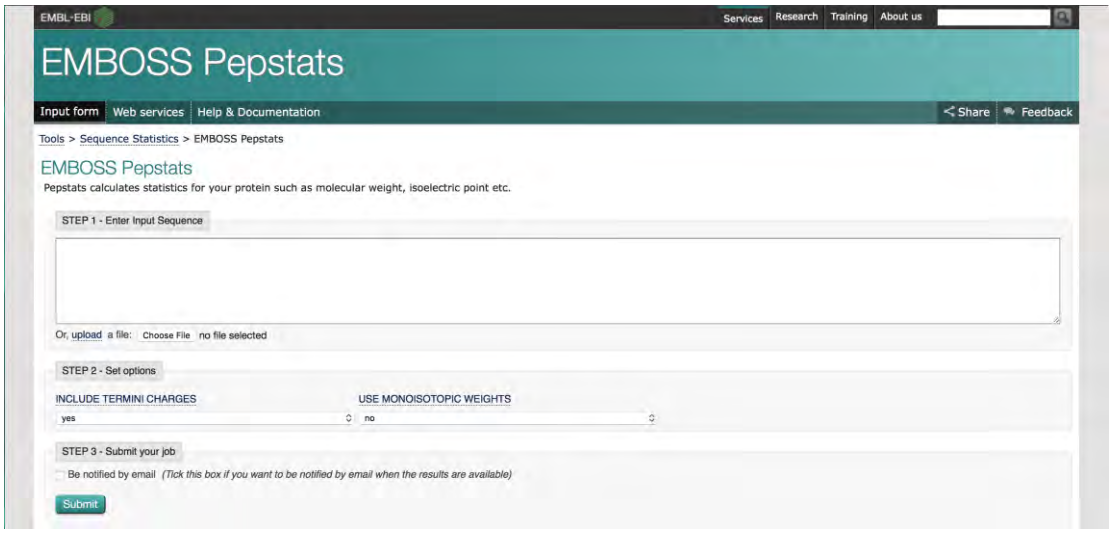


Figure 15: Pepstats Homepage

2.3.6 ProtParam

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

Part of ExPASy server the ProtParam is an online tool that computes various physiochemical properties that can be attained from previous mentioned sources. The computed parameters include deduction of molecular weight, theoretical pI, amino acid composition extinction coefficient, half-life, instability index and grand average of hydropathicity (GRAVY) (Gasteiger, et al., 2005). The FASTA sequence was uploaded and result was analyzed (figure 16).

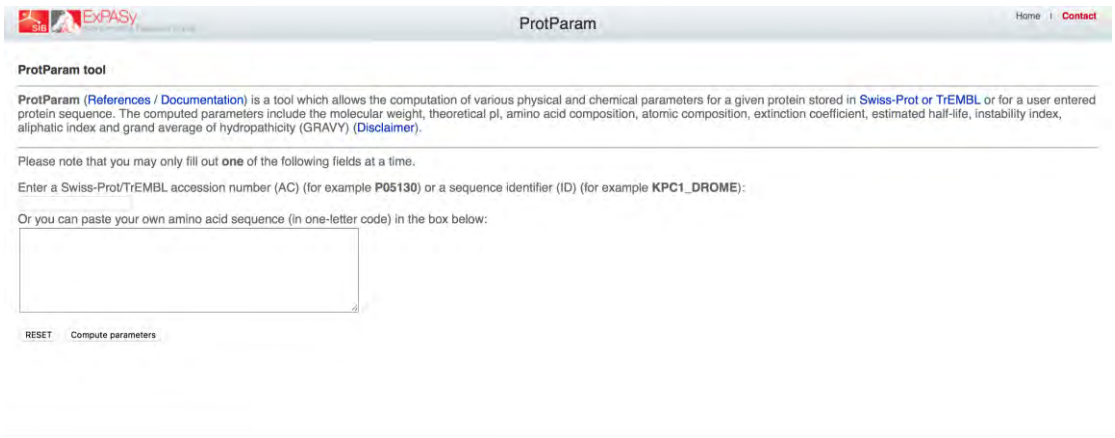


Figure 16: ProtParam Homepage

2.3.7 Prot Scale

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

Prot Scale is an online tool, which is available at ExPASy server. It allows user to compute and represent the profile produced by a sequence of amino acid on a scale that is of two dimensions in nature. 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 scales 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 (Gasteiger, et al., 2005).

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 are best suited for membrane spanning protein query. Analyzing the shape of the plot gives information about partial structure of the protein. For instance, if a stretch of about 20 amino acids shows positive for hydrophobicity, these amino acids may be part of alpha helix spanning across a lipid bilayer, which is composed of hydrophobic fatty acids. On the converse, amino acids with high hydrophilicity indicate that these residues are in contact with solvent, or water, and that they are therefore likely to reside on the outer surface of the protein.

In this study the FASTA sequence was uploaded in the input window with window size 19 and scale selected as Hphob. / Kyte & Doolittle for trans-membrane region detection (Claverie & Notredame, 2006) (figure 17)

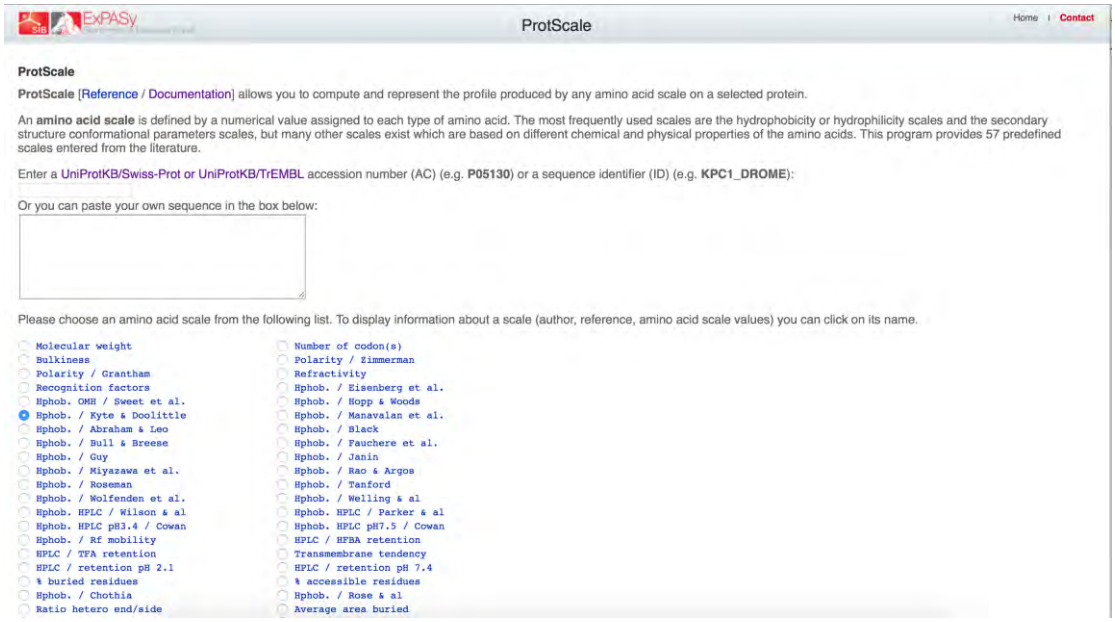


Figure 17: Prot Scale Homepage

2.3.8 TMHMM

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

The TMHMM server is an online tool that 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. (Krogh, Larsson, von Heijne , & Sonnhammer, 2001)

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 18)

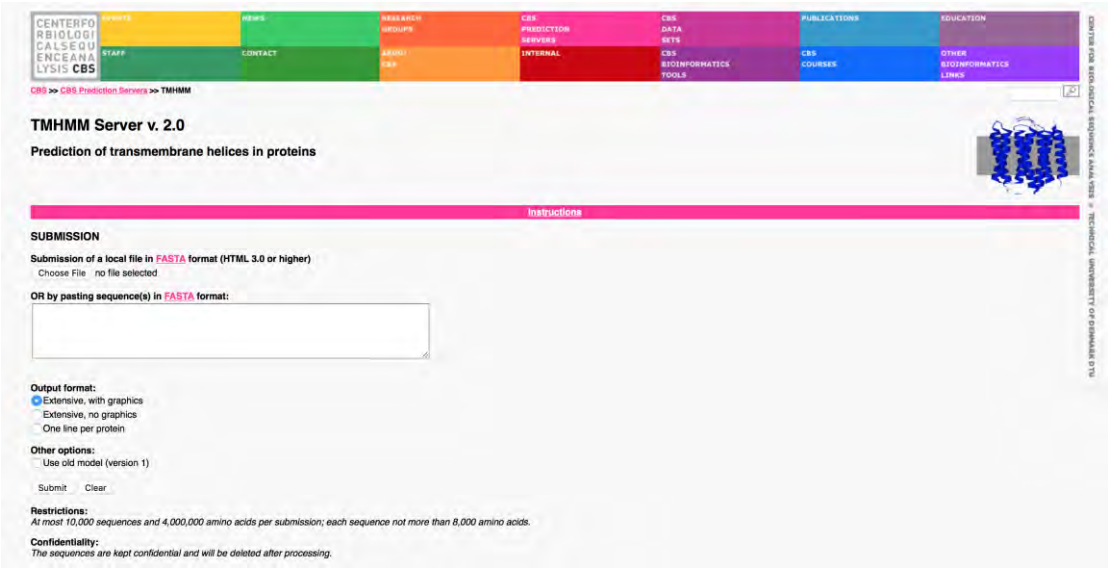


Figure 18: TMHMM Homepage

2.3.9 PSIPRED

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

PSIPRED 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 thus predicting a protein's secondary structure (beta sheets, alpha helices and coils) from the primary sequence and making it a easy and accurate secondary structure prediction method. (Buchan, Minneci, Nugent, Bryson , & Jones, 2013) (Jones, 1999)

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 19).

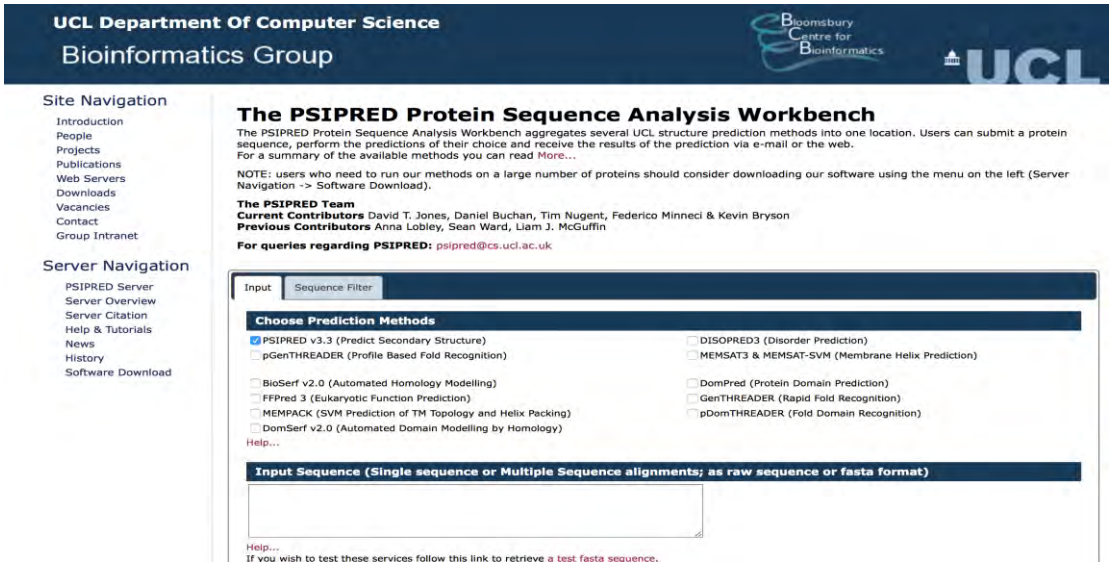


Figure 19: PSIPRED Homepage

2.3.10 Iterative Threading ASSEmbly Refinement (I-TASSER)

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

I-TASSER is an online server, which is a hierarchical method for protein structure and function prediction from amino acid sequences (figure 20). 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 (Zhang, 2008) (Yang, Yan, Roy, Xu, Poisson, & Zhang, 2015) (Roy, Kucukural, & Zhang, 2010).

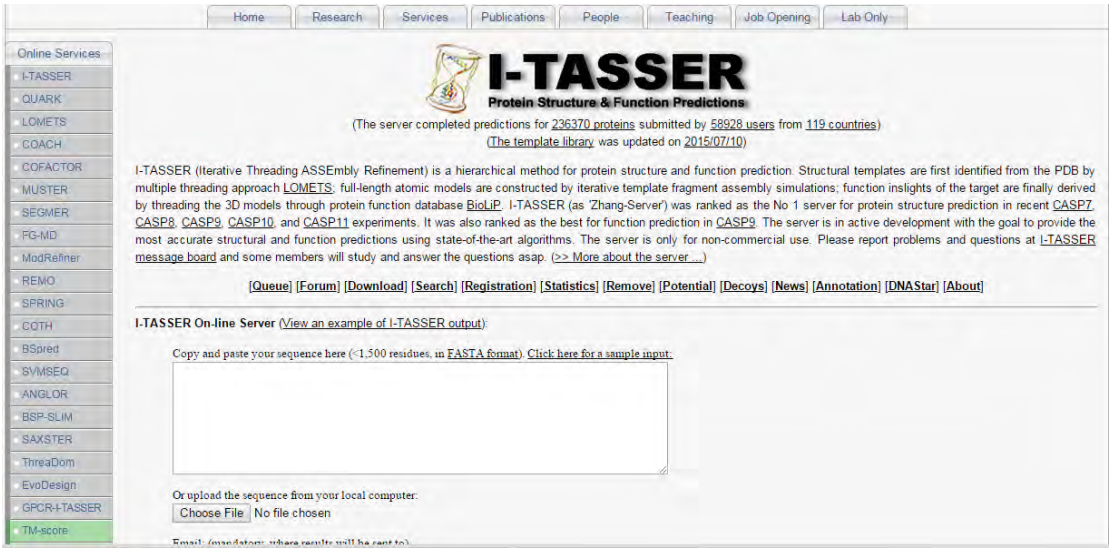


Figure 20: I-TASSER homepage

2.3.11 Protein Homology/analogy Recognition Engine V 2.0 (Phyre^2)

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

Phyre^2 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. Phyre^2 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 (Kelley, Mezulis, Yates, Wass, & Sternberg, 2015). In this study, the FASTA sequence of the query protein was entered and the intensive mode was selected to attain 3D models (figure 21)

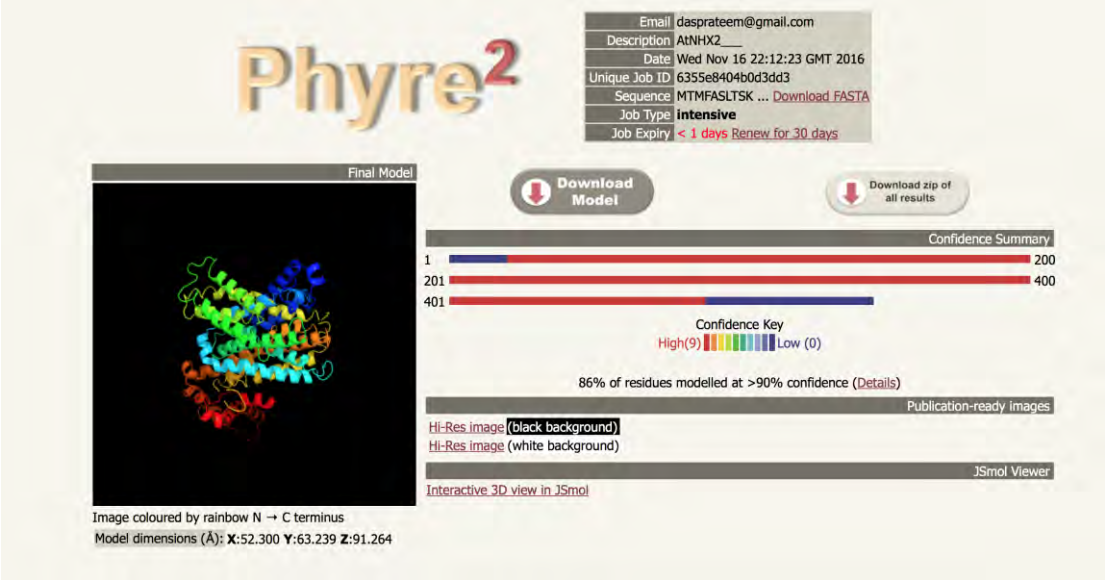


Figure 21: Phyre^2 Homepage

2.3.12 Swiss Model

URL link: https://swissmodel.expasy.org

SWISS-MODEL is a fully automated protein structure homology-modelling server, accessible via the ExPASy web server, or from the program DeepView (Swiss Pdb-Viewer). The purpose of this server is to make protein modeling accessible to all biochemists and molecular biologists worldwide (Biasini, et al., 2014) (Bordoli, Kiefer, Arnold, Benkert, Battey, & Schwede, 2009) (Arnold, Bordoli, Kopp, & Schwede, 2006)(Figure 22).

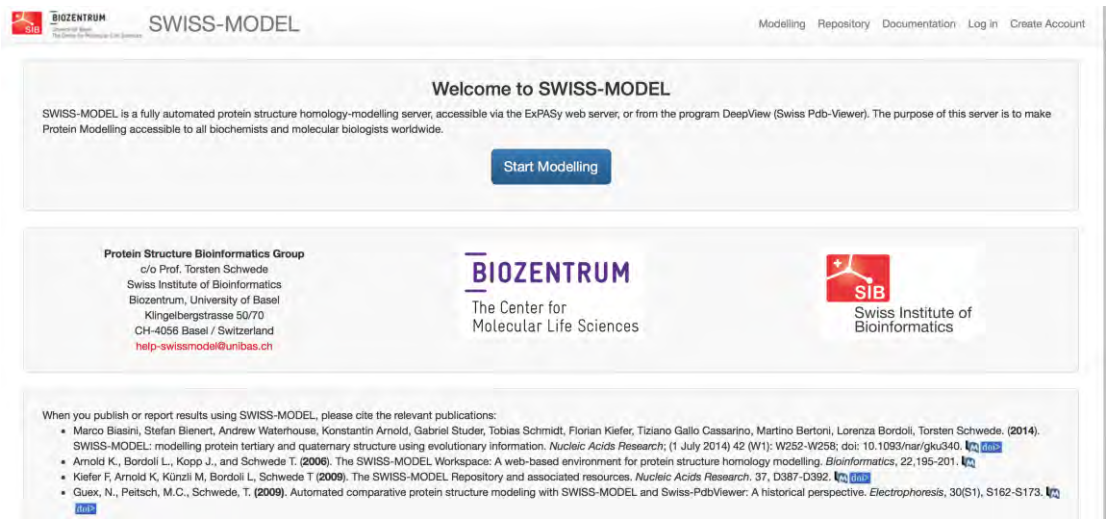


Figure 22: Swiss Model

2.3.13 Easy Modeller

MODELLER is used for homology or comparative modeling of protein three-dimensional structures. The user provides an alignment of a sequence to be modeled with known related structures and MODELLER automatically calculates a model containing all non-hydrogen atoms. MODELLER implements comparative protein structure modeling by satisfaction of spatial restraints, and can perform many additional tasks, including de novo modeling of loops in protein structures, optimization of various models of protein structure with respect to a flexibly defined objective function, multiple alignment of protein sequences and/or structures, clustering, searching of sequence databases, comparison of protein structures (Eswar, et al., 2006) (Fiser, Do, & Sali) (Martí-Renom , Stuart, Fiser, Sánchez, Melo, & Šali, 2000) (Sali, 1995). (Figure 23.1)

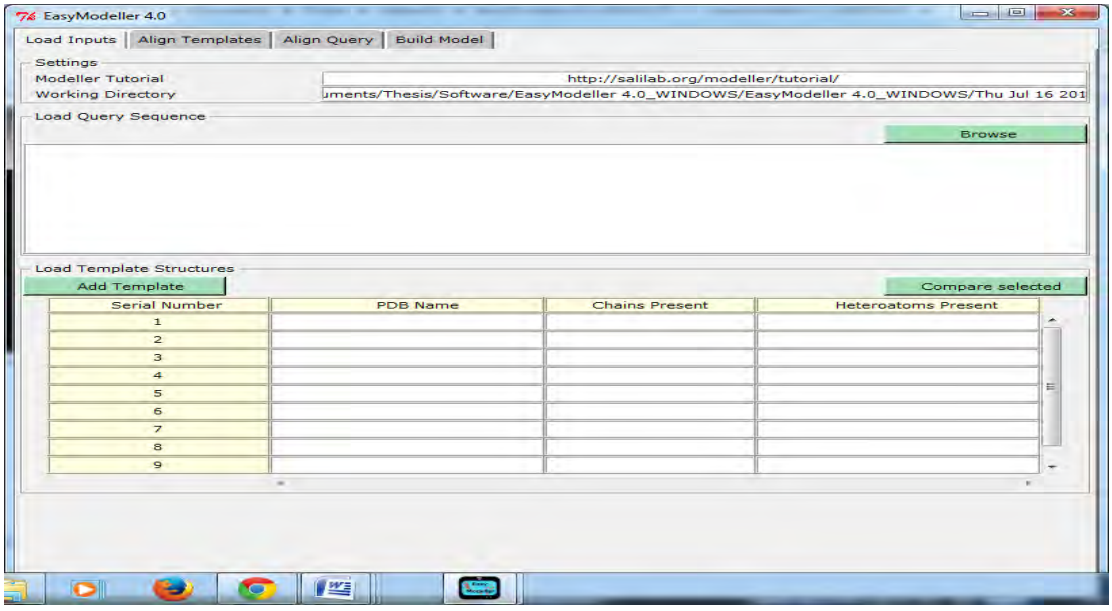


Figure 23.1: Easy Modeller

Since the whole process for this software is manual, NCBI was again used, expending the same accession ID of the previously used *Arabidopsis* Blast (p-Suite) software. However instead of using protein- protein Blast (Blastp) PSI-BLAST (Position-Specific Iterated BLAST) was used and Protein Data Bank was selected as databank.

Then these templates are uploaded into the easy modeller software and one of the templates is selected by validation. It is done via weighted pair-group average clustering method based on a distance matrix.

The higher the resolution factor (R-factor) and higher the Sequence Similarity the more preferable is the template for model generation. R-Factor determines the resolution of the image generated by the NMR- spectroscopy during the modeling of the template and the similarity score determines the similarity between the template and the query.

Then the templates were aligned using the Easy Modeler align software and consecutively alignment was also done with the query sequence.

The software uses software, such as, python 2.7 or the earlier versions to generate codes. It also cans intake code manually. For this experiment, python 2.7 was used to generate codes and run the simulation.

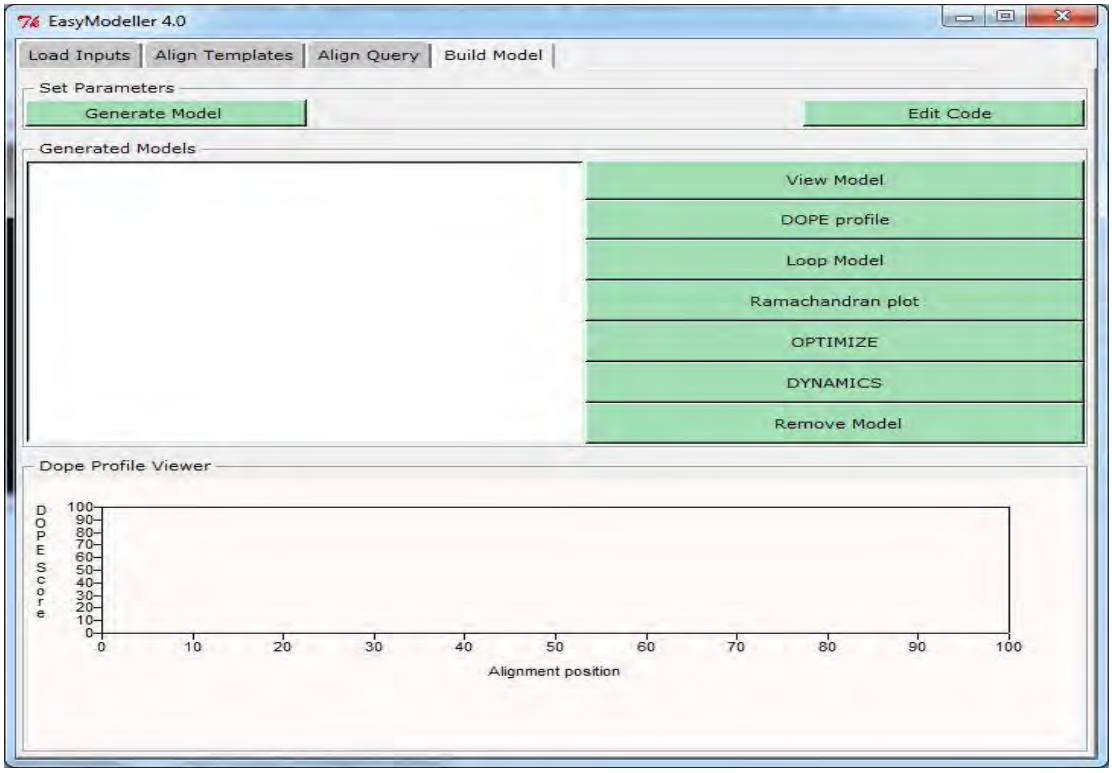


Figure 23.2: model generation GUI (Graphical user interface) of Easy Modeller

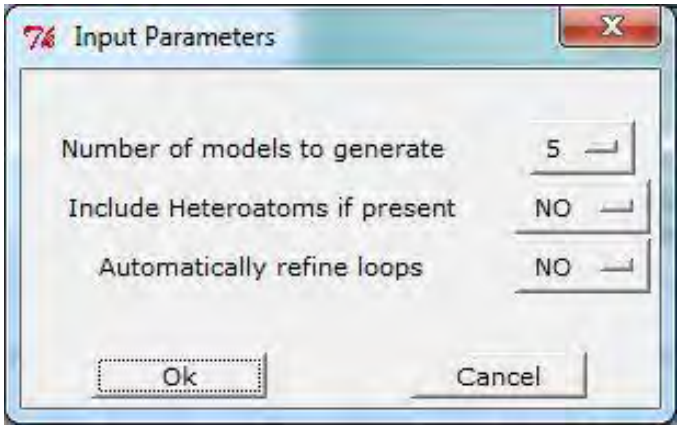


Figure 23.3: Input Parameter of Easy Modeller

2.3.14 Swiss Model

URL link: https://swissmodel.expasy.org

SWISS-MODEL is a fully automated protein structure homology-modelling server, accessible via the ExPASy web server, or from the program DeepView (Swiss Pdb-Viewer). The purpose of this server is to make Protein Modelling accessible to all biochemists and molecular biologists worldwide (figure 24).

Selection of templates are based on the scores the templates generate respect to the query sequence. The models are generated automatically afterwards.

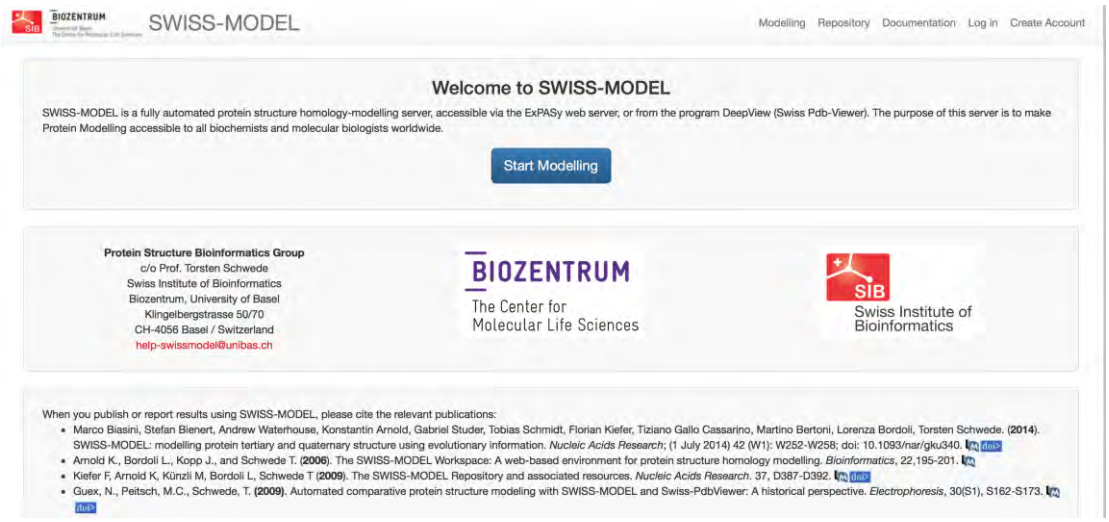


Figure 24: Swiss Model Homepage

2.3.15 PROCHECK:

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

PROCHECK is a downloadable software available at the EMBL-EBI website which checks the stereo-chemical 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, which assesses and evaluates the protein PDB coordinate models (Laskowski, Rullmann, MacArthur, Kaptein, & Thornton , 1996) (Laskowski, MacArthur, Moss, & Thornton)(figure 25)

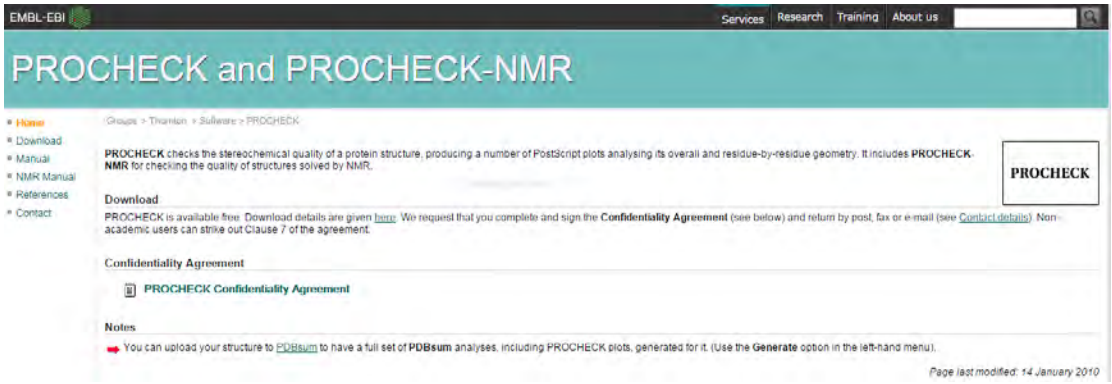


Figure 25: PROCHECK Homepage

2.3.15 (a) PDB sum

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

In this current study, the PROCHECK web server available at the PDBsum Generate section of the PDBsum server 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 (de Beer , Berka, Thornton, & Laskowski, 2013) (figure 26 a).

Figure 26 a: PDB sum Generate Homepage

2.3.15 (b) PSVS

Protein structure validation software is a server for the software of PROCHECK. PSVS (Protein Structure Validation Server) systematically evaluates the quality of protein structures, and reports a large set of quality scores and constraint analyses. PSVS defines a set of standard quality scores and constraint analyses that are reported for each structure. In addition to experimental constraints, this set encompasses a number of parameters evaluating different aspects of structure quality, including fold and packing, local residue separation, deviations of bond length and bond angles, backbone and side-chain torsion angle conformation, and atomic overlaps. Quality scores are reported based on calibration with a set of high-resolution X-ray crystal structures. A Z-score for the query structure, calculated for the score from each tool, provides an evaluation of the quality of the structure compared to this set of X-ray crystal structures. This server returns a concise validation report that includes a standard set of graphs and tables (de Beer , Berka, Thornton, & Laskowski, 2013)(figure 26 b)

Figure 26 b: PSVS Homepage

2.3.16 Q-Mean Server

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

Q-Mean (Qualitative Model Energy ANalysis) is a server that focuses on structure validation. It does by scoring the Q mean value and Z score.

Q mean score determines the composite scoring function, which is able to derive both global (i.e. for the entire structure) and local (i.e. per residue) error estimates on the basis of one single model. QMEAN score has been extended to an absolute quality estimate that is denoted as Z-score. It is a measure for the absolute quality for the model. The Q mean score of the model is compared with high-resolution reference structure solved by X-ray crystallography. The Q mean has a score between 1-0 and higher the value the higher the stability of the structure. Z-score also has a high value and a negative value or a score less than 0 suggests poor structure or the protein is a trans-membrane protein (Benkert, Biasini, & Schwede, Toward the estimation of the absolute quality of individual protein structure models , 2011) (Benkert, Künzli, & Schwede, QMEAN server for protein model quality estimation, 2009) (Benkert, Tosatto, & Schomburg, QMEAN: A comprehensive scoring function for model quality assessment, 2008) (Figure 27).

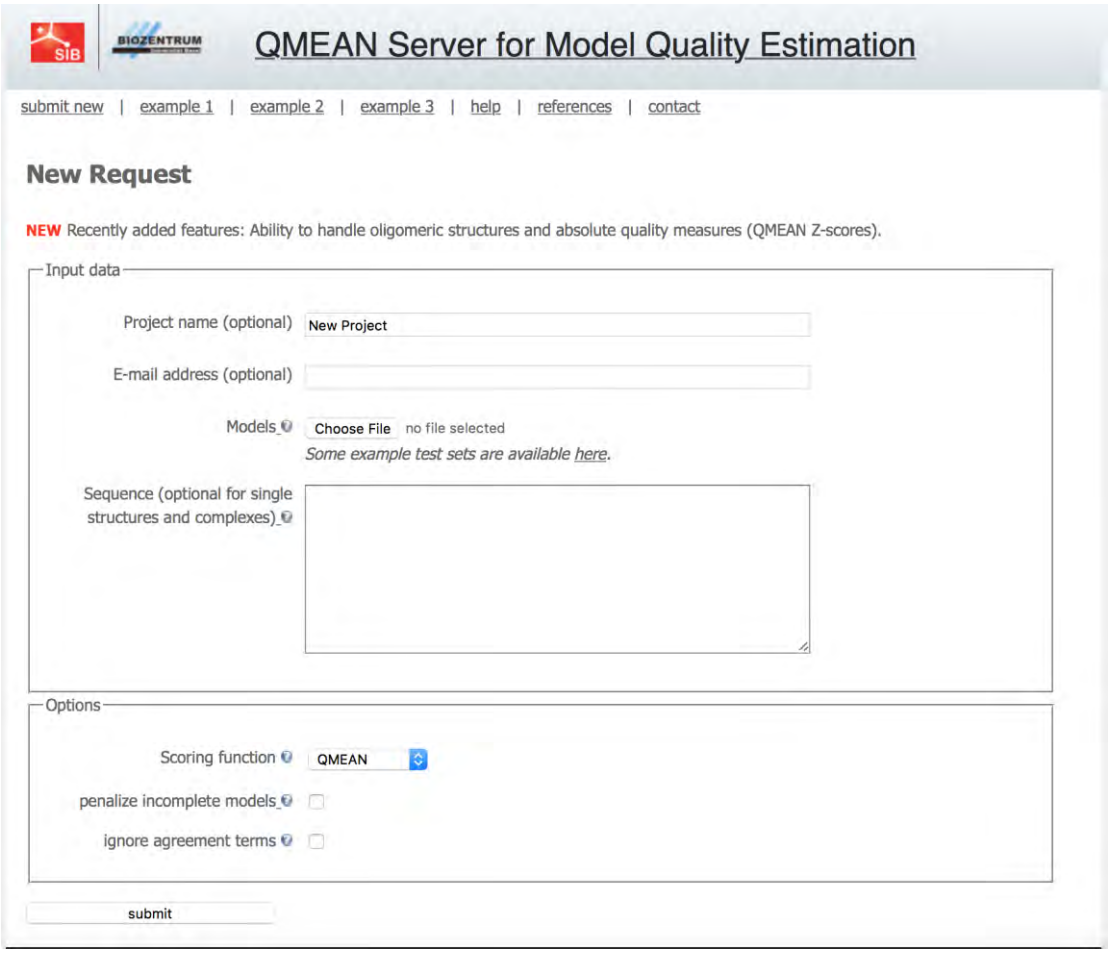


Figure 27: Q-mean Homepage

2.3.17 UCSF Chimera

URL link: http://www.cgl.ucsf.edu/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 (Pettersen, et al., 2004). This tool can produce high quality animations and images. It can be downloaded from the UCSF Chimera website. In this current study Chimera Version 1.10.1 was used (Figure 28)

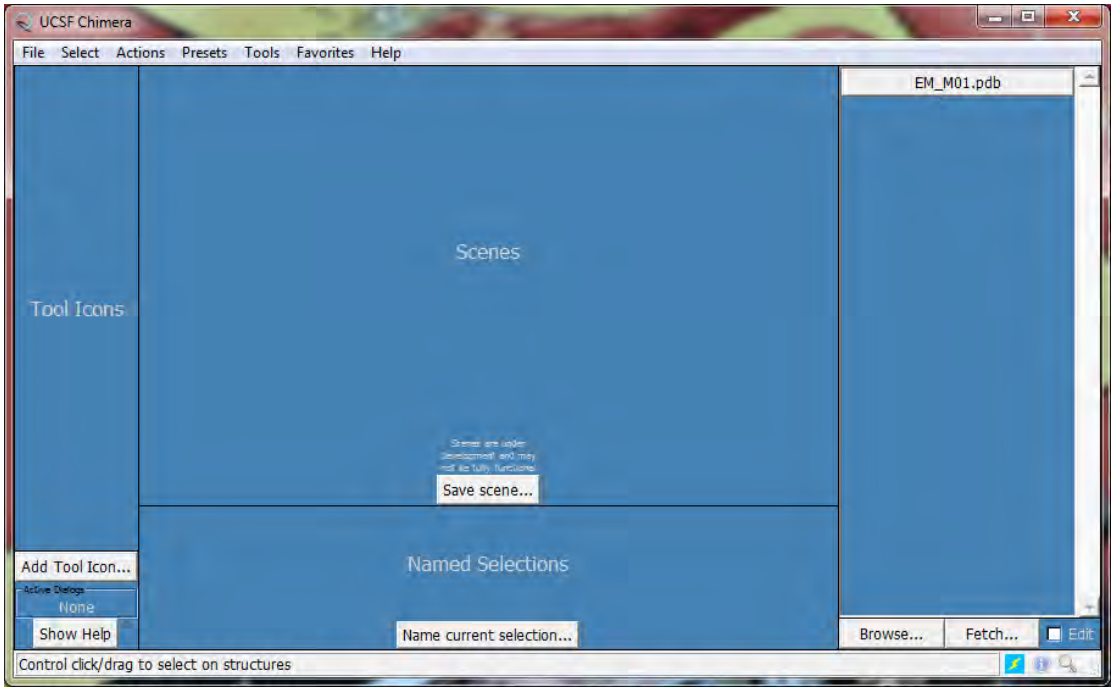


Figure 28: Chimera GUI

Chapter 3

3.1 RESULTS

3.1.1 Target FASTA Sequence Retrieved from NCBI databases:

The target sequence was achieved by surfing NCBI database. It was analyzed with the initial target source of *Arabidopsis* and the coding region was obtained. The target FASTA nucleotide sequence of sodium/hydrogen exchanger 2 was attained.

3.1.1.1 Retrieved FASTA sequence of the target sequence:

>NM_111375.4 Arabidopsis thaliana sodium hydrogen exchanger 2 (NHX2), mRNA

GAAAAATTGTTAGAAAAAAAGAAAATCTATTGTCCTTGTTATTCAGGAGA
ATACTTGTCTACGACGAGGTCCACAGTAACTCCGAAGACATCTAGAGTCT
AGAGTGATCTAGTGAATCTATAAAGCCGCCAAATCGTTAAATCTTTCCTTT
TAAACATTTTTGATTTCTTTTACGTTTGCGTTACGAATCATCTTCTGCCTCT
CTCTCAACGCAACTCAATCCACCCTACACGTGTTCCTCTCTTTTCCCGTCTA
AATTTCTCAGTCTCGCCGTGAATTTTGTTCTTCTCTTGTTTTCCGGGCGAAT
CAAAATCCGTCTCATTTCCAATCGCGTCTCCGTTGGTATTTTACATGAACG
TTTAGCGGAATCTCGGGTTTAGCTTCGGAGCTTTGCATTTGATCTTCGTCG
TCTGTGTTGTCCTCTCTCGGATTTTGTTTTTTCGCCGCTGCAAATCAGAGAA
ATCTATTCCTGGAAGTGATAAGATCGCTTGGAAGCGTTTCAGGGAGTTGTT
GGAATGGAGGGTCCGGAAAGATGAGAGATTTTGAATACGCAGTTATTCTT
TGATTTCTTAATAGATTTTACACGTCATTGGTGAAGAAAAAGGGAAAAAG
AAAGATGACAATGTTCGCCTCTTTAACCTCTAAAATGCTATCGGTGTCAAC
TTCTGATCACGCATCTGTCGTTTCACTTAATCTCTTTGTTGCCCTTCTATGT
GCTTGTATCGTCATTGGCCATCTTTTGGAGGAGAATCGATGGATGAACGA
ATCCATCACTGCTTTATTGATTGGGCTTGGCACTGGTGTCGTCATATTGTT
GATTAGTAGAGGGAAAAACTCACATCTCTTGGTCTTTAGTGAAGATCTCTT
CTTTATATATCTTTTGCCACCCATAATATTCAATGCAGGGTTTCAAGTAAA
AAAGAAGCAGTTTTTCCGAAATTTTGTAACTATTATGGCTTTTGGCGCCAT
TGGGACCGTAGTTTCTTGCACCATAATATCTCTAGGTGCAATTCAGTTCTT
TAAGAAATTAGACATTGGGACCTTTGACTTGGGCGATTTTCTTGCAATCGG
CGCCATATTTGCTGCAACCGACTCTGTATGCACACTACAGGTTCTCAATCA

AGATGAGACACCTTTGCTTTACAGTCTTGTATTTGGAGAGGGCGTTGTGAA
TGATGCCACATCTGTTGTGCTCTTCAATGCTATTCAGAGTTTTGACCTCAC
CCACCTTAACCATGAAGCAGCTTTTCAATTTCTTGGGAACTTTTTTTATCTG
TTTCTCTTGAGCACCGGACTTGGTGTCGCAACTGGTCTGATAAGTGCTTAT
GTCATCAAGAAACTGTATTTTGGAAGGCACTCGACTGATCGAGAAGTTGC
CCTCATGATGCTTATGGCTTATCTTTCATATATGCTTGCTGAGCTATTCGCC
TTGAGTGGTATCCTAACTGTATTTTTCTGTGGGATTGTGATGTCCCATTAC
ACTTGGCACAATGTCACCGAGAGCTCAAGAATTACTACCAAGCATGCCTT
TGCTACTTTGTCGTTTCTCGCTGAGACTTTTATTTTCCTCTACGTTGGAATG
GATGCATTGGACATAGAGAAATGGAGATTCGTGAGTGACAGCCCGGGGA
CATCAGTTGCAGTGAGCTCAATTCTAATGGGTCTAGTCATGCTTGGAAGA
GCAGCTTTTGTCTTTCCTCTTTCCTTCTTATCAAACTTAGCCAAGAAGCATC
AGAGCGAGAAAATCAGCATCAAGCAGCAAGTTGTGATCTGGTGGGCTGGT
CTAATGAGAGGTGCTGTATCTATGGCTCTTGCCTACAATAAGTTTACAAGA
TCAGGGCACACAGAATTGCGCGGGAATGCAATCATGATTACCAGTACAAT
AACCGTCTGTCTTTTTAGCACCATGGTGTTTGGTATGCTAACCAAACCACT
GATTAGATACCTAATGCCACACCAAAAAGCGACCACCAGTACCACGAGTA
TGTTATCGGACGATAGCACTCCGAAATCAATCCACATTCCGCTCCTCGATG
GTGAACAGCTAGATTCATTTGAGTTACCTGGGAGCCACCAGGACGTGCCA
CGACCAAACAGCCTTCGAGGTTTCCTCATGCGCCCCACACGGACTGTCCA
CTATTACTGGAGACAGTTTGATGATGCCTTCATGCGTCCTGTGTTTGGTGG
TCGCGGATTCGTTCCCTTTGTCCCTGGTTCTCCGACTGAGAGAAGCAGCCA
TGATCTTAGTAAACCTTGAGGAGAAAGATATATAGAAACTTAACCAAAAA
ACTTCTTCTTGCTCTTCCCTCTTATGGTGACTAGTATTGGTGATGTAAATGT
ATTTTTCGTTCTTCAAATTTACATATTCTTCTGTAAATTTGTTATTATTCGA
TGATGAAGAAGCTTCTTACGTTTTTGAGAGACGTGTGGGGGTTCATGTGTG
TTACTATTAGTCTTTGTTTTGTTTAACAAAAAAACTACCAAGTTTTTGACT
ATGCATCATGGGGTTTGTTGTTGTTGGTCAGAACTTAGAACTTGACTAACT
CTGCTTCGGTTAGGATCTGATTGGTA

The N-Suite of BLAST algorithm was used for the target sequence analysis and mega-blast option was selected for high accuracy. Total summary was obtained, graphical in nature that consisted of a visual representation of the sequences in the data bank that is similar to the query sequence. The red bars represented high similarity between the sequences along with the regions where similarities occur. Only the coding regions of the sequence were selected and the rest of the sequence was eliminated for better result. This included nucleotide sequences from 617 to 2257 that were provided to N-Suite where CDS region reclined.

A total of 17 sequences were retrieved from the result that the BLAST produced. The result was highly specific because of high accuracy set before query and selections of sequences based on E-values and identity scores. 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 numbers of times the database match have occurred randomly. It provides with a criterion that is more objective than that of the percentage-of-similarity (identity values). Therefore the lower the E-value the higher is the authenticity thus sequences with zero values are selected and any sequence similarity by chance is avoided (Claverie & Notredame, 2006). A graphical summary of the BLAST N-Suite result is listed below (figure 29 a) followed by graphical representation of the query gene (figure 29 b).

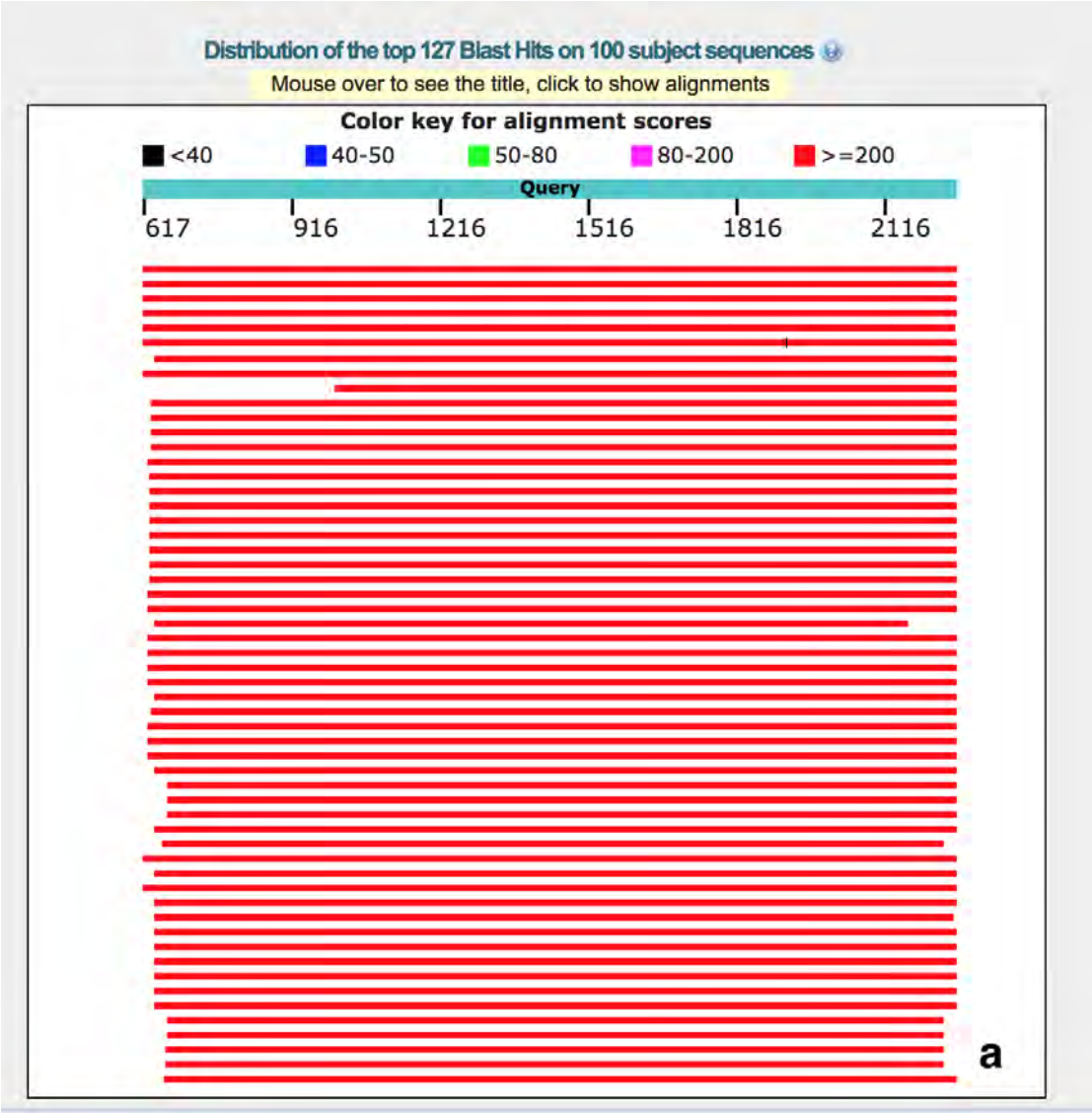


Figure 29 a: A graphical summary of the BLAST results using blast N-suite for the target nucleotide sequence

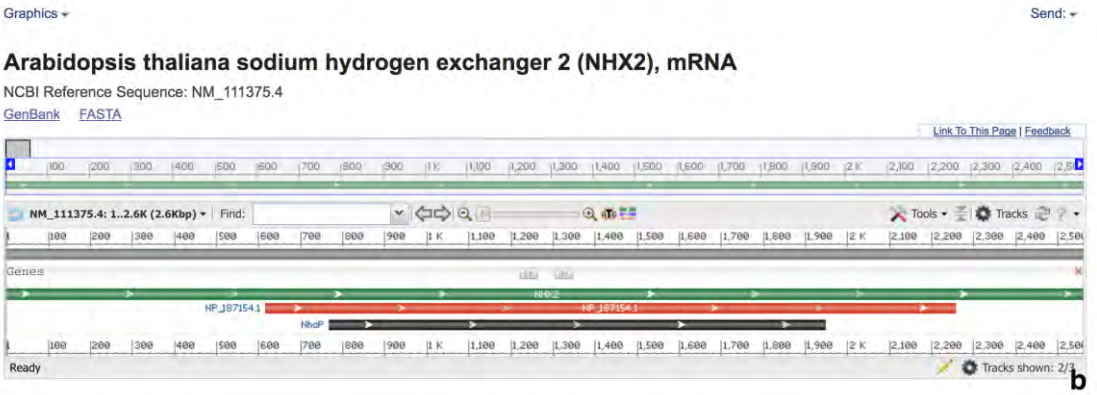


Figure 29 b: Graphical representation of the whole gene and the coding region of the query sequence. Green bar representing the whole gene, the red bar representing part of gene representing promoter and the CDS and finally the black bar representing the gene that codes for the protein itself.

3.1.1.2 FASTA sequence for the target protein:

The target protein FASTA sequence of sodium/hydrogen exchanger 2 (Arabidopsis thaliana) attained:

>NP_187154.1 sodium hydrogen exchanger 2 [Arabidopsis thaliana]

MTMFASLTSKMLSVSTSDHASVVSLNLFVALLCACIVIGHLLEENRWMNESIT
ALLIGLGTGVVILLISRGKNSHLLVFSEDLFFIYLLPPIIFNAGFQVKKKQFFRNF
VTIMAFGAIGTVVSCTIISLGAIQFFKKLDIGTFDLGDFLAIGAIFAATDSVCTL
QVLNQDETPLLYSLVFGEGVVNDATSVVLFNAIQSFDLTHLNHEAAFQFLGN
FFYLFLLSTGLGVATGLISAYVIKKLYFGRHSTDREVALMMLMAYLSYMLAE
LFALSGILTVFFCGIVMSHYTWHNVTESSRITTKHAFATLSFLAETFIFLYVGM
DALDIEKWRFVSDSPGTSVAVSSILMGLVMLGRAAFVFPLSFLSNLAKKHQSE
KISIKQQVVIWWAGLMRGAVSMALAYNKFTRSGHTELRGNAIMITSTITVCLF
STMVFGMLTKPLIRYLMPHQKATTSTTSMLSDDSTPKSIHIPLLDGEQLDSFEL
PGSHQDVPRPNSLRGFLMRPTRTVHYYWRQFDDAFMRPVFGGRGFVPFVPG
SPTERSSHDLSKP

The list of sequences that produced significant alignment was retrieved from numerous sequences generated via P-Suite of the BLAST software. 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 (figure 30). The red bars indicated highly similar sequences amongst different species of plants. The protein sequences that were selected underwent same screening method as the selection of the nucleotide sequence, which required the sequences to have a greater than or equal value of 80% of identity score and zero E-Value. This was done to conduct a valid comparison between the two sets of data. 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.

Table 2: List of nucleotide sequences selected and categorized according to their score and their corresponding accession ID for protein sequence.

	Organism	Blast n (NM_111375)	Protein accession ID of the result nucleotide sequences
1	*Arabidopsis thaliana*	NM_001337555	NP_001326201
2	*Arabidopsis thaliana*	NM_001337552	NP_001319475
3	*Arabidopsis thaliana*	AF490586	AAM08403
4	*Arabidopsis thaliana*	NM_001337554	NP_001326211
5	*Brassica napus*	AY189676	AAO38856
6	*Eutrema halophilum*	FJ713100	ACN76859
7	*Eutrema halophilum*	DQ490966	ABF48496
8	*Brassica napus*	GU192449	ACZ92142
9	*Olimarabidopsis pumila*	KC200248	AGG11160
10	*Arabidopsis thaliana*	NM_122597	NP_198067
11	*Arabidopsis thaliana*	EF596738	ABQ58865
12	*Arabidopsis thaliana*	AK226586	BAE98703
13	*Arabidopsis thaliana*	AF510074	AAM34759
14	*Arabidopsis thaliana*	AF106324	AAD16946
15	*Olimarabidopsis pumila*	JF357965	AEA51351
16	*Arabidopsis thaliana*	AY685183	AAT95387
17	*Arabidopsis thaliana*	AF056190	AAF21755
18	*Arabidopsis thaliana*	NM_122597	NP_198067

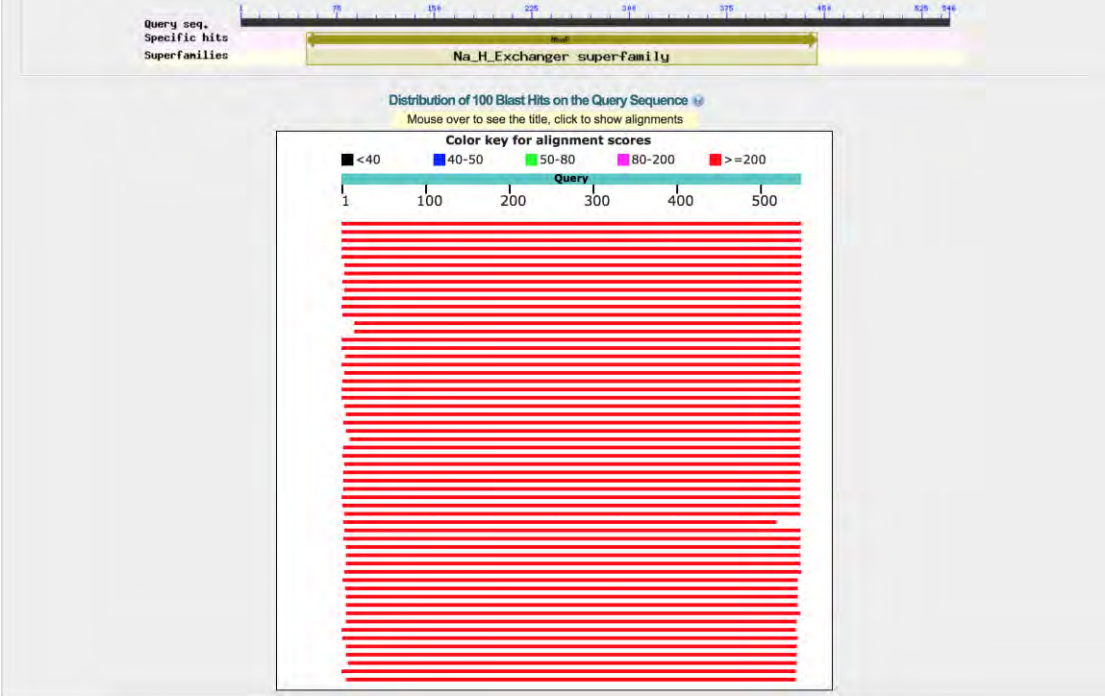


Figure 30: A graphical summary of the BLAST results using blast P-suite for the target protein sequence

It is to be taken to an account that the names of the final selected protein and nucleotide sequence names were altered matching to their corresponding sequence names and sequences that were putative, hypothetical or partial were avoided for further data validation.

The selected nucleotide sequences and protein sequences were tabulated and compared and common hits (for both the protein and the nucleotide sequences) were further tabulated in a separate table (table 3).

Table 3: List of selected protein sequences categorized according to the score

#	Organism	Accession ID
1	*Brassica napus*	CDY27804
2	*Arabidopsis thaliana*	NP_198067
3	*Arabidopsis thaliana*	ABQ58865
4	*Arabidopsis thaliana*	AAM34759
5	*Eutrema halophilum*	ABF48496
6	*Nitraria sibirica*	BAR73031
7	*Pyrus betulifolia*	AGE13941
8	*Brassica napus*	CDY50586
9	*Brassica napus*	CDY19707
10	*Citrus reticulata*	AAT95387
11	*Brassica napus*	AAO38856
12	*Ricinus communis*	NP_001310702
13	*Cochlearia hollandica*	AFF57538
14	*Brassica napus*	CDY0002
15	*Brassica napus*	ACZ92142
16	*Olimarabidopsis pumila*	AEA51351
17	*Olimarabidopsis pumila*	AGG11160
18	*Morus notabilis*	AIL23819
19	*Nitraria tangutorum*	AID55215
20	*Arabidopsis thaliana*	AAT36679
21	*Arabidopsis Thaliana*	NP_198067

Common hits were then isolated and then further tabulated with their name changed in corresponding to the organism's name (table 4).

#	Organism	Blast n	Blast p	Matching Altered Name
Q	Arabidopsis thaliana (query sequence)	NM_111375	NP_187154	***A.thaliana Q***
1	Brassica napus	AY189676	AAO38856	***B.napus 1***
2	Eutrema halophilum	DQ490966	ABF48496	***E. halophilum 1***
3	Brassica napus	GU192449	ACZ92142	***B.napus 2***
4	Arabidopsis thaliana	NM_122597	NP_198067	***A.thaliana 1***
5	Olimarabidopsis pumila	KC200248	AGG11160	***O.pumila 1***
6	Arabidopsis thaliana	EF596738	ABQ58865	***A.thaliana 2***
7	Arabidopsis thaliana	AF510074	AAM34759	***A.thaliana 3***
8	Olimarabidopsis pumila	JF357965	AEA51351	***O.pumila 2***
9	Arabidoopsis thaliana	AY685183	AAT95387	***A.thaliana 4***
10	Arabidoopsis thaliana	NM122597	NP_198067	***A.thaliana 5***

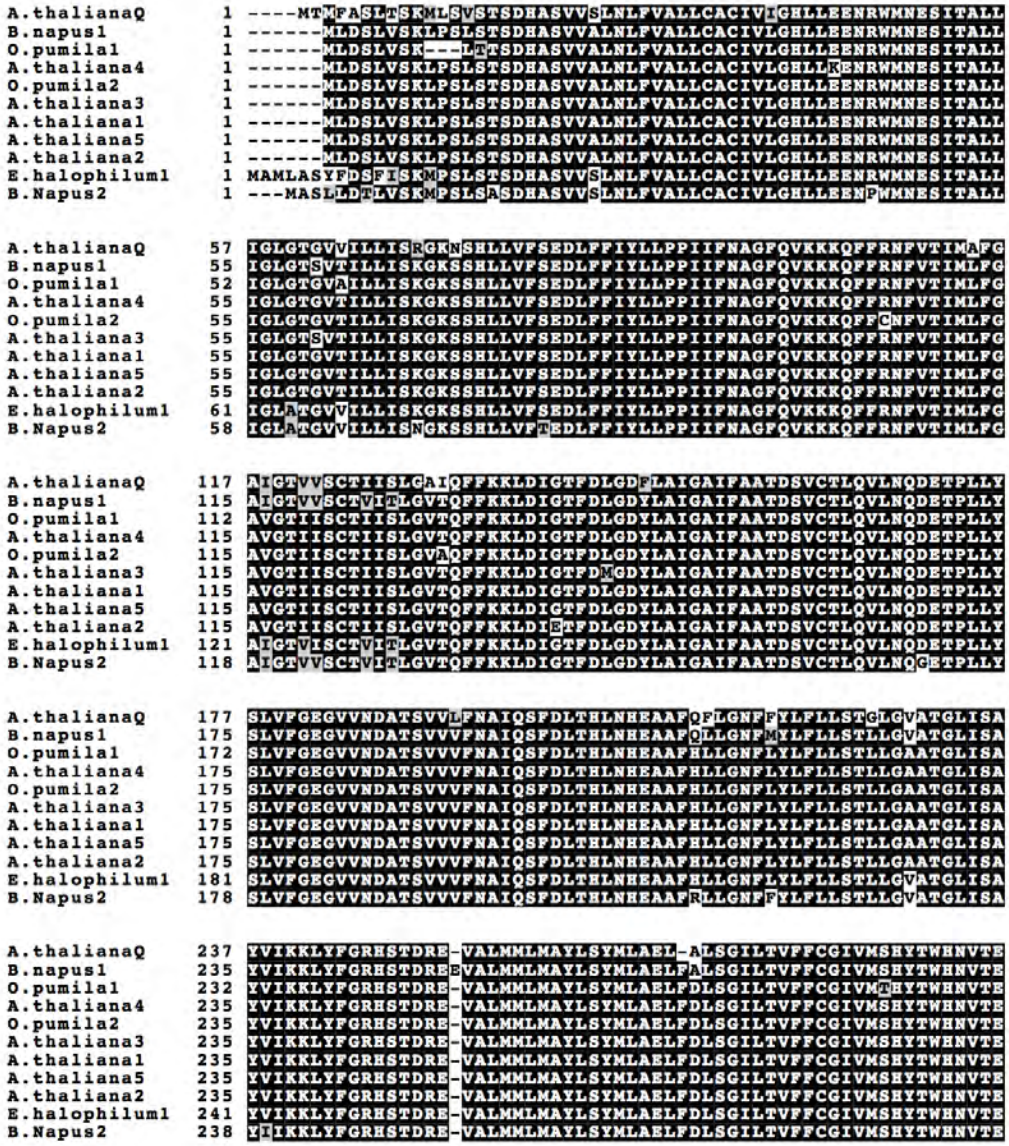
Table 4: List of the organisms along with the accession ID, common hits (common for both N-Suite and P-Suite) and names altered for convenience.

3.1.2 Multiple Sequence Alignment (MSA) results using Clustal Omega

The selected ten sequences were aligned in such a manner that rewrites the selected sequences and where all of the same features plan in the same columns. Clustal Omega was used, as alignment software and the purpose of this were to put amino acids in same column to predict the stability of the query amino acid sequence. Amino acids usually subjected to two evolutionary subsets, which are surface loops and core regions. The surface loops are more prone to change since their usual spots are on the surface of the protein and are far more exposed to the outer environment thus making them prone to rapid evolution. These regions are self-contained, and are mostly free of the evolutionary constraints imposed by the conserved core of the domains. The loops are the softer portion present on the surface of the proteins that connect to more rigid portions. However, the core regions are far more conserved, rigid and serve as the backbone for the protein and provide the protein with an identity. The core regions are less prone to rapid evolutions unlike surface loops. (Blouin, Butt , & Roger, 2004) Therefore, through multiple sequence alignment this phenomenon can be identified. Gap rich regions give an indication of surface loops as they are poorly defined and the core regions, the counterparts, are gap free regions (Claverie & Notredame, 2006).

3.1.3 Visualization and interpretation of the final alignment

The Multiple Sequence Alignment was carried out by Clustal Omega, which was then downloaded and imported to BoxShade sequence alignment editor and the identical or similar amino acids were shaded (Saitou & Nei, 1987) using the afore-mentioned tool and output file was downloaded. It was observed that the sequences of the species were highly conserved and there were very little surface loops present in the amino acid sequences and most were composed of conserved, gap free regions (Blouin, Butt , & Roger, 2004) (Bassil, Coku, & Blumwald, 2012). (Ramachandran & Dokholyan, 2012)


A.thalianaQ 1 ----MTMFASLTSKMLSVSTSDHASVVSLNLFVALLCACIVIGHLLEENRWMNESITALL
B.napus1 1 ------MLDSLVSKLPSLSTSDHASVVALNLFVALLCACIVLGHLLEENRWMNESITALL
O.pumila1 1 ------MLDSLVSK---LTTSDHASVVALNLFVALLCACIVLGHLLEENRWMNESITALL
A.thaliana4 1 ------MLDSLVSKLPSLSTSDHASVVALNLFVALLCACIVLGHLLKENRWMNESITALL
O.pumila2 1 ------MLDSLVSKLPSLSTSDHASVVALNLFVALLCACIVLGHLLEENRWMNESITALL
A.thaliana3 1 ------MLDSLVSKLPSLSTSDHASVVALNLFVALLCACIVLGHLLEENRWMNESITALL
A.thaliana1 1 ------MLDSLVSKLPSLSTSDHASVVALNLFVALLCACIVLGHLLEENRWMNESITALL
A.thaliana5 1 ------MLDSLVSKLPSLSTSDHASVVALNLFVALLCACIVLGHLLEENRWMNESITALL
A.thaliana2 1 ------MLDSLVSKLPSLSTSDHASVVALNLFVALLCACIVLGHLLEENRWMNESITALL
E.halophilum1 1 MAMLASYFDSFISKMPSLSTSDHASVVSLNLFVALLCACIVLGHLLEENRWMNESITALL
B.Napus2 1 ---MASLLDTLVSKMPSLSASDHASVVSLNLFVALLCACIVLGHLLEENPWMNESITALL
A.thalianaQ 57 IGLGTGVVILLISRGKNSHLLVFSEDLFFIYLLPPIIFNAGFQVKKKQFFRNFVTIMAFG
B.napus1 55 IGLGTSVTILLISKGKSSHLLVFSEDLFFIYLLPPIIFNAGFQVKKKQFFRNFVTIMLFG
O.pumila1 52 IGLGTGVAILLISKGKSSHLLVFSEDLFFIYLLPPIIFNAGFQVKKKQFFRNFVTIMLFG
A.thaliana4 55 IGLGTGVTILLISKGKSSHLLVFSEDLFFIYLLPPIIFNAGFQVKKKQFFRNFVTIMLFG
O.pumila2 55 IGLGTGVTILLISKGKSSHLLVFSEDLFFIYLLPPIIFNAGFQVKKKQFFCNFVTIMLFG
A.thaliana3 55 IGLGTSVTILLISKGKSSHLLVFSEDLFFIYLLPPIIFNAGFQVKKKQFFRNFVTIMLFG
A.thaliana1 55 IGLGTGVTILLISKGKSSHLLVFSEDLFFIYLLPPIIFNAGFQVKKKQFFRNFVTIMLFG
A.thaliana5 55 IGLGTGVTILLISKGKSSHLLVFSEDLFFIYLLPPIIFNAGFQVKKKQFFRNFVTIMLFG
A.thaliana2 55 IGLGTGVTILLISKGKSSHLLVFSEDLFFIYLLPPIIFNAGFQVKKKQFFRNFVTIMLFG
E.halophilum1 61 IGLATGVVILLISKGKSSHLLVFSEDLFFIYLLPPIIFNAGFQVKKKQFFRNFVTIMLFG
B.Napus2 58 IGLATGVVILLISNGKSSHLLVFTEDLFFIYLLPPIIFNAGFQVKKKQFFRNFVTIMLFG
A.thalianaQ 117 AIGTVVSCTIISLGAIQFFKKLDIGTFDLGDFLAIGAIFAATDSVCTLQVLNQDETPLLY
B.napus1 115 AIGTVVSCTVITLGVTQFFKKLDIGTFDLGDYLAIGAIFAATDSVCTLQVLNQDETPLLY
O.pumila1 112 AVGTIISCTIISLGVTQFFKKLDIGTFDLGDYLAIGAIFAATDSVCTLQVLNQDETPLLY
A.thaliana4 115 AVGTIISCTIISLGVTQFFKKLDIGTFDLGDYLAIGAIFAATDSVCTLQVLNQDETPLLY
O.pumila2 115 AVGTIISCTIISLGVAQFFKKLDIGTFDLGDYLAIGAIFAATDSVCTLQVLNQDETPLLY
A.thaliana3 115 AVGTIISCTIISLGVTQFFKKLDIGTFDMGDYLAIGAIFAATDSVCTLQVLNQDETPLLY
A.thaliana1 115 AVGTIISCTIISLGVTQFFKKLDIGTFDLGDYLAIGAIFAATDSVCTLQVLNQDETPLLY
A.thaliana5 115 AVGTIISCTIISLGVTQFFKKLDIGTFDLGDYLAIGAIFAATDSVCTLQVLNQDETPLLY
A.thaliana2 115 AVGTIISCTIISLGVTQFFKKLDIETFDLGDYLAIGAIFAATDSVCTLQVLNQDETPLLY
E.halophilum1 121 AIGTVLSCTVITLGVTQFFKKLDIGTFDLGDYLAIGAIFAATDSVCTLQVLNQDETPLLY
B.Napus2 118 AIGTVVSCTVITLGVTQFFKKLDIGTFDLGDYLAIGAIFAATDSVCTLQVLNQGETPLLY
A.thalianaQ 177 SLVFGEGVVNDATSVVLFNAIQSFDLTHLNHEAAFQFLGNFFYLFLLSTGLGVATGLISA
B.napus1 175 SLVFGEGVVNDATSVVVFNAIQSFDLTHLNHEAAFQLLGNFMYLFLLSTLLGVATGLISA
O.pumila1 172 SLVFGEGVVNDATSVVVFNAIQSFDLTHLNHEAAFHLLGNFLYLFLLSTLLGAATGLISA
A.thaliana4 175 SLVFGEGVVNDATSVVVFNAIQSFDLTHLNHEAAFHLLGNFLYLFLLSTLLGAATGLISA
O.pumila2 175 SLVFGEGVVNDATSVVVFNAIQSFDLTHLNHEAAFHLLGNFLYLFLLSTLLGAATGLISA
A.thaliana3 175 SLVFGEGVVNDATSVVVFNAIQSFDLTHLNHEAAFHLLGNFLYLFLLSTLLGAATGLISA
A.thaliana1 175 SLVFGEGVVNDATSVVVFNAIQSFDLTHLNHEAAFHLLGNFLYLFLLSTLLGAATGLISA
A.thaliana5 175 SLVFGEGVVNDATSVVVFNAIQSFDLTHLNHEAAFHLLGNFLYLFLLSTLLGAATGLISA
A.thaliana2 175 SLVFGEGVVNDATSVVVFNAIQSFDLTHLNHEAAFHLLGNFLYLFLLSTLLGAATGLISA
E.halophilum1 181 SLVFGEGVVNDATSVVVFNAIQSFDLTHLNHEAAFHLLGNFLYLFLLSTLLGVATGLISA
B.Napus2 178 SLVFGEGVVNDATSVVVFNAIQSFDLTHLNHEAAFRLLGNFFYLFLLSTLLGVATGLISA
A.thalianaQ 237 YVIKKLYFGRHSTDRE-VALMMLMAYLSYMLAEL-ALSGILTVFFCGIVMSHYTWHNVTE
B.napus1 235 YVIKKLYFGRHSTDREEVALMMLMAYLSYMLAELFALSGILTVFFCGIVMSHYTWHNVTE
O.pumila1 232 YVIKKLYFGRHSTDRE-VALMMLMAYLSYMLAELFDLSGILTVFFCGIVMTHYTWHNVTE
A.thaliana4 235 YVIKKLYFGRHSTDRE-VALMMLMAYLSYMLAELFDLSGILTVFFCGIVMSHYTWHNVTE
O.pumila2 235 YVIKKLYFGRHSTDRE-VALMMLMAYLSYMLAELFDLSGILTVFFCGIVMSHYTWHNVTE
A.thaliana3 235 YVIKKLYFGRHSTDRE-VALMMLMAYLSYMLAELFDLSGILTVFFCGIVMSHYTWHNVTE
A.thaliana1 235 YVIKKLYFGRHSTDRE-VALMMLMAYLSYMLAELFDLSGILTVFFCGIVMSHYTWHNVTE
A.thaliana5 235 YVIKKLYFGRHSTDRE-VALMMLMAYLSYMLAELFDLSGILTVFFCGIVMSHYTWHNVTE
A.thaliana2 235 YVIKKLYFGRHSTDRE-VALMMLMAYLSYMLAELFDLSGILTVFFCGIVMSHYTWHNVTE
E.halophilum1 241 YVIKKLYFGRHSTDRE-VALMMLMAYLSYMLAELFDLSGILTVFFCGIVMSHYTWHNVTE
B.Napus2 238 YIIKKLYFGRHSTDRE-VALMMLMAYLSYMLAELFDLSGILTVFFCGIVMSHYTWHNVTE

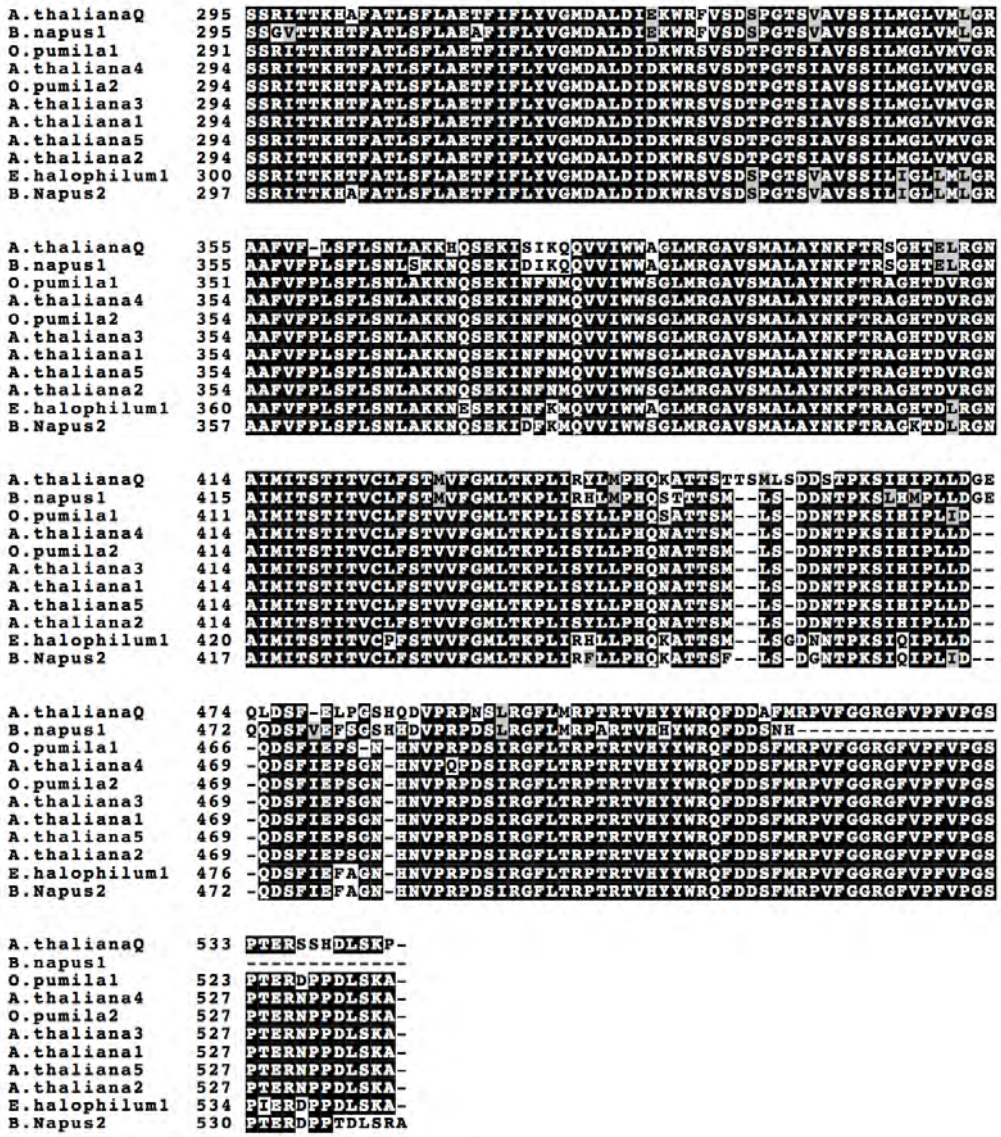


Figure 31: Multiple sequence alignment of the 10 selected protein sequences where A.thalianiaQ is the query sequence. It was generated using Clustal Omega and BoxShade was used to convert it into a publishable format. Black shaded regions indicate similar residues, followed by white or gap rich regions, which are gap rich regions.

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 demonstrated by the fact that the NHX functional groups appeared early on in evolution and have conserved and essential cellular roles in plants (Bassil, Coku, & Blumwald, 2012).

3.1.4 Phylogenetic tree generation by MEGA

Bootstrap consensus tree was constructed using MEGA 7 software by importing the MSA file on MEGA 7.

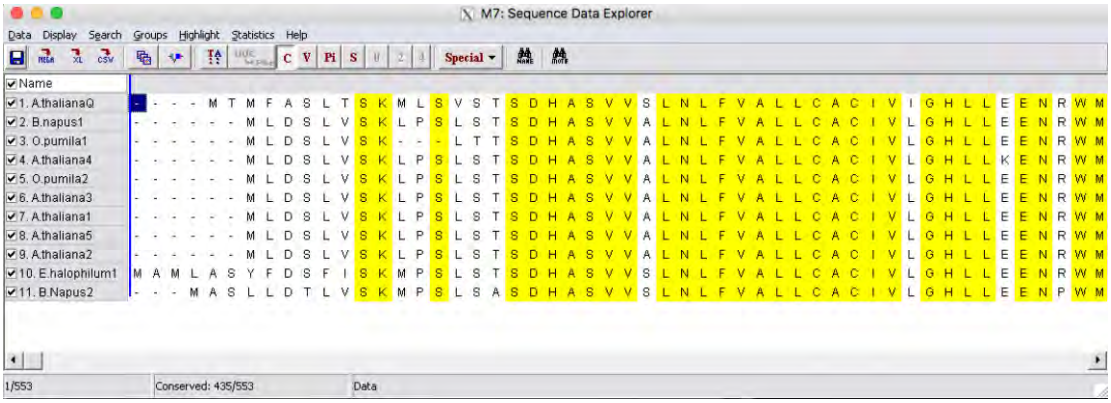


Figure 32: Sequence Data Explorer GUI of MEGA 7 showing conserved and variable sites of the protein sequences of 10 selected and query sequences.

The software revealed 435 conserved regions and 118 variable sites over a total span of 553 regions (figure 32).

The neighbor joining method was executed to generate a bootstrap consensus tree with node statistics (figure 33) (Saitou & Nei, 1987) (Miseta & Custora, 2000). It revealed that there were significant amount of selected species related with the query sequences supported by the fact that node statistics score were greater than 85, thus the clades formed during divergence were strongly supported. However, this cannot be true in case of all of the species that were compared with the query list. Neighbor-Joining method is dependent on the evolutionary distance, therefore it can be deduced that the query sequence (*A. thalianaQ*) is closely related to *B.napus 1* and *E.halophilum 1* as they had the least degree of divergence from the query sequence itself.

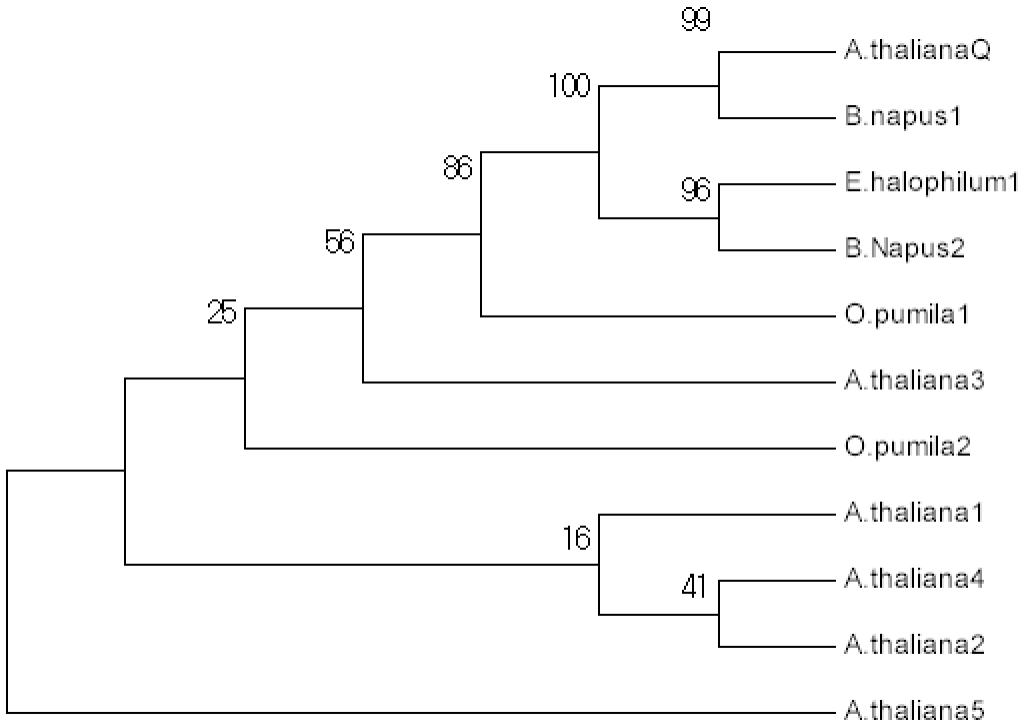


Figure 33: A bootstrap consensus tree with node statistics for the selected eleven protein sequences along with the query sequence

3.1.5 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. It was seen from the result that most abundant amino acid accounted for was leucine which was around ~13% of the total protein makeup with least being cysteine and tryptophan, both being around ~1%. Cysteine is responsible for di-sulfide bonds present in proteins, which is a necessary for structure formation between two separate polypeptide chains (Miseta & Custora, Relationship Between the Occurrence of Cysteine in Proteins and the Complexity of Organisms, 2000) and also plays a major role in folding and stability (Betz, 1993) (Miseta & Csutora, Molecular Biology and Evolution , 2000). Since there are low amount of cysteine therefore the stability of the protein is provided by other procedures. The amino acid composition makes it unlikely to be expressed as a protein in the inclusion body of yeast. Results tabulated in table 5.

3.1.6 Analysis of physicochemical properties

Computation of various physical and chemical parameters of the selected protein sequences was performed using the ProtParam tool and tabulated. The computed Isoelectric Point (pI) of the proteins was ~7.00 (6.92) 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. It also indicates the overall nature of the protein since it is involved in ion transport thus affecting The Extinction Coefficients (EC) of the proteins were all same for each of the eleven organisms, with a slight variation seen in *B.napus 1,* which was below the average. 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 a range of temperatures. Results tabulated in table 6.

	A.thaliana Q	***B.napus 1***	***E.halophilum 1***	***B.napus 2***	***A.thaliana 1***	***O.pumila 1***	***A.thaliana 2***	***A.thaliana 3***	***O.pumila 2***	***A.thaliana 4***	***A.thaliana 5***
Molecular weight	60278.51	57148.70	60436.70	59931.13	59513.42	59131.00	59585.49	59561.48	59430.35	59484.43	59513.42
Residues	546	515	545	542	538	534	538	538	538	538	538
Average Residue Weight	110.806	110.968	110.893	110.574	110.620	110.732	110.754	110.709	110.465	110.566	110.620
Charge	10.0	3.0	7.5	7.0	5.5	4.5	4.5	5.5	4.5	6.5	5.5
Isoelectric Point	8.1605	6.7929	7.6731	7.8963	7.2311	7.0628	7.0628	7.2311	7.0574	7.4339	7.2311
A280 Molar Extinction Coefficients	52370	50880	53860	52370	53860	53860	53860	53860	53860	53860	53860
cystine bridges	52745	51255	54235	52745	54235	54235	54235	54235	54235	54235	54235

	A.thaliana Q	***B.napus 1***	***E.halophilum 1***	***B.napus 2***	***A.thaliana 1***	***O.pumila 1***	***A.thaliana 2***	***A.thaliana 3***	***O.pumila 2***	***A.thaliana 4***	***A.thaliana 5***
A280 Molar Extinction Coefficients (1mg/ml)	0.869	0.890	0.891	0.874	0.905	0.911	0.904	0.904	0.906	0.905	0.905
cystine bridges	0.875	0.897	0.897	0.880	0.911	0.917	0.910	0.911	0.913	0.912	0.911
Improbability of expression in inclusion bodies	0.719	0.586	0.701	0.711	0.706	0.662	0.688	0.706	0.694	0.717	0.706
Property of each residue (Mole %)											
Tiny	31.618	30.680	30.275	30.812	30.669	30.712	30.483	30.669	30.855	30.669	30.669
Small	49.081	49.126	49.908	50.554	50.929	50.749	50.743	50.929	51.115	50.929	50.929
Aliphatic	33.640	34.369	34.495	34.871	34.201	34.457	34.201	34.015	34.387	34.201	34.201
Aromatic	14.706	14.563	14.679	14.207	14.312	14.419	14.312	14.312	14.312	14.312	14.312

Non-polar	60.478	59.417	61.101	61.439	60.409	60.487	60.223	60.223	60.781	60.409	60.409
Polar	39.522	40.583	38.899	38.561	39.591	39.513	39.777	39.777	39.219	39.591	39.591
Charged	17.279	17.864	17.431	16.790	16.914	17.228	17.100	16.914	16.729	16.729	16.914
Basic	10.294	10.097	10.092	9.594	9.665	9.738	9.665	9.665	9.480	9.665	9.665
Acidic	6.985	7.767	7.339	7.196	7.249	7.491	7.435	7.249	7.249	7.063	7.249
Highest amino acid residue	L (69)	L (72)	L (71)	L (72)	L (71)	L (69)	L (71)	L (70)	L (71)	L (71)	L (71)

Table 5: Amino acid composition attained by PEPSTATS analysis tool depicting molecular weight, residues, average residue weight, charge, isoelectric point, Molar Extinction Coefficients, improbability of expression in inclusion bodies, amino acids grouped according to their physical properties.

Organisms	Sequence Length	MW	pI	EC (Cys residues not reduced)	EC (Cys residues reduced)	Ii	Comment	Ai	GRAVY	-R	+R
A.thalinia Q	546	60278.51	8.14	52745	52370	36.48	Stable	103.40	0.465	38	40
B.napus1	515	57148.70	6.30	51255	50880	35.71	Stable	106.74	0.463	40	34
E.halophilum1	545	60436.70	7.25	54235	53860	35.62	Stable	106.61	0.477	40	40
B.Napus2	542	59931.13	7.67	52745	52370	33.52	Stable	107.73	0.508	39	40
A.thaliana1	538	59513.42	6.73	54235	53860	32.71	Stable	106.71	0.458	39	37
O.pumila1	534	59131.00	6.56	54235	53860	32.24	Stable	106.97	0.471	40	37
A.thaliana2	538	59585.49	6.56	54235	53860	33.68	Stable	106.71	0.452	40	37
A.thaliana3	538	59561.48	6.73	54235	53860	32.86	Stable	105.99	0.453	39	37
O.pumila2	538	59430.35	6.56	54235	53860	32.78	Stable	106.90	0.475	39	36
A.thaliana4	538	59484.43	6.94	54235	53860	32.44	Stable	106.71	0.459	38	37
A.thaliana5	538	59513.42	6.73	54235	53860	32.71	Stable	106.71	0.458	39	37

Table 6: Parameters for the protein encoded by AtNHX2 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)

3.1.7 Prediction via ProtScale tool

The Protscale is used for measuring the trans-membrane regions present within a protein. ProtScale provides with a hydrophobicity profile presented in two-dimensional plot form, which depicts potential trans-membrane regions from a query sequence. The sequence was uploaded and the software predicted 12 trans-membrane regions over a span of 544 amino acids (figure 34). This was obtained by taking account of all of the peaks above 1.6, which is the recommended threshold level of Kyte & Doolittle scale, which showed 12 peaks overall (Kyte & Doolittle, 1982).

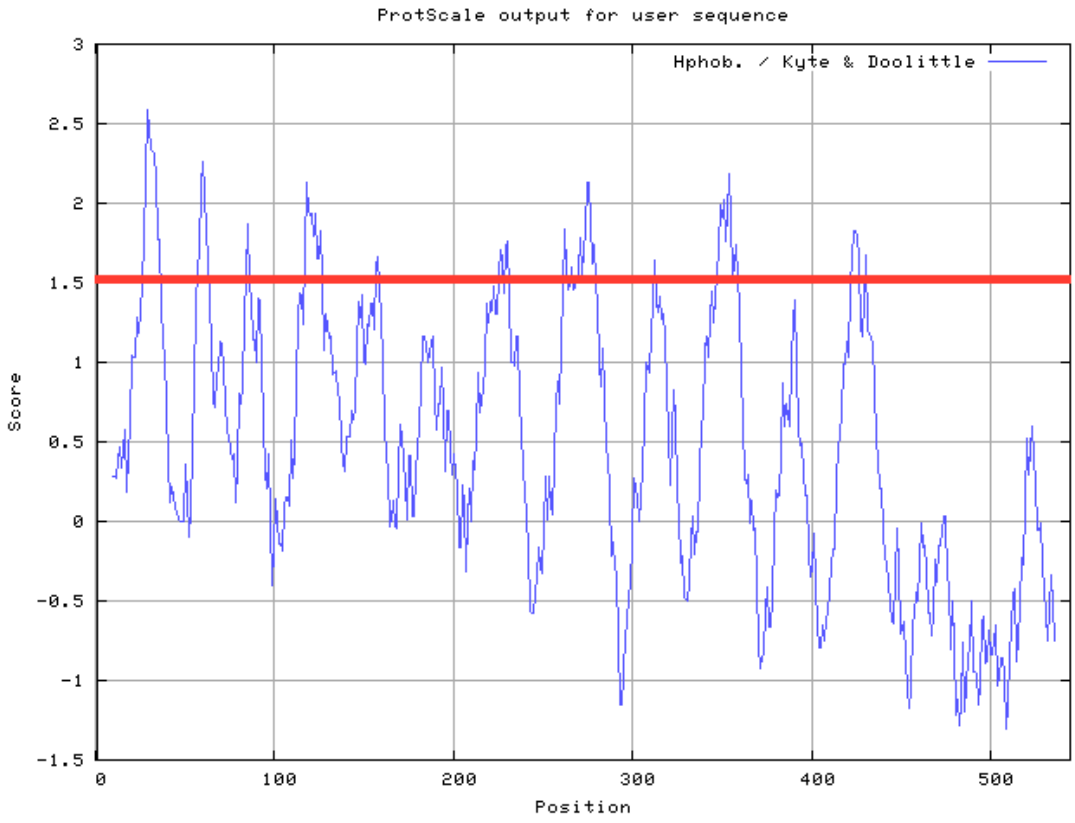


Figure 34: ProtScale output of AtNHX2 (*A.thaliana Q*)

3.1.8 Trans-membrane prediction via TMHMM server

Similar to ProtScale, TMHMM server was also used for trans-membrane prediction and revealed two-dimensional plot with depiction of potential trans-membrane segments. However, the server predicts the intrinsic and the extrinsic in nature. The blue line represent intrinsic trans-membrane segment and the pink lines indicated

extrinsic trans-membrane helix. Twelve peaks were also observed here. This indicated twelve trans-membrane segments, thus supporting the previous conclusion (figure 35). In an experiment for analysis of hydropathy plot it was seen that the trans-membrane antiporter protein, primarily AtNHX (Yamaguchi, Apse, Shi, & Blumwald, 2003) are functional as a channel protein if arranged in tetramers (Swarbreck, Colaço, & Davies, 2013) and furthermore, it was predicted that the trans-membrane 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, Yamagishi, & Tomaselli, 1998). It was also found that transporter/ antiporter proteins are comprised of numerous amounts of helices to form long trans-membrane alpha helices packed together for functional feature (Elofsson & Heijne, 2007). Moreover, the fact that more than half of the total amino acid residues reside in trans-membrane region (~240 of 544) establishes and solidifies the fact that the query protein (AtNHX2) of *A.thaliana Q,* spans across the membrane.

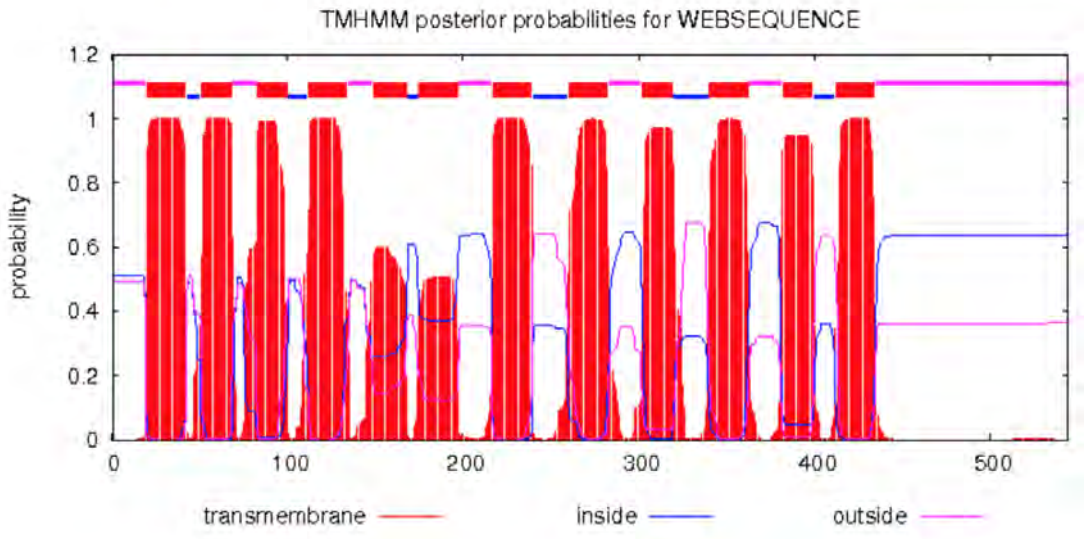


Figure 35: TMHMM output for AtNHX2 (*A.thalianaQ*)

Predicted	Trans-membrane Helices	12
#	Number of the amino acids in Trans-membrane Helices	238.90
#	Spanning outside	7
#	Spanning inside	6

Table 7: Number of trans-membrane helix present in the query sequence of the AtNHX2 along with their properties.

Trans-membrane Helix deposition	Position of helix on amino acid chains
1	20-42
2	51-68
3	83-100
4	112-134
5	149-168
6	175-197
7	217-239
8	260-282
9	302-319
10	340-362
11	382-399
12	412-434

Table 8: Positions of the predicted trans-membrane Helices predicted by TMHMM

3.1.9 Graphical representation of secondary structures in *Arabidopsis thaliana* sodium/hydrogen exchanger 2 protein

PISPRED was used for determination of secondary structure. Overall there were 15 helix, 24 coils and 10 beta strands present in the query sequence of AtNHX2 (figure 36.2).

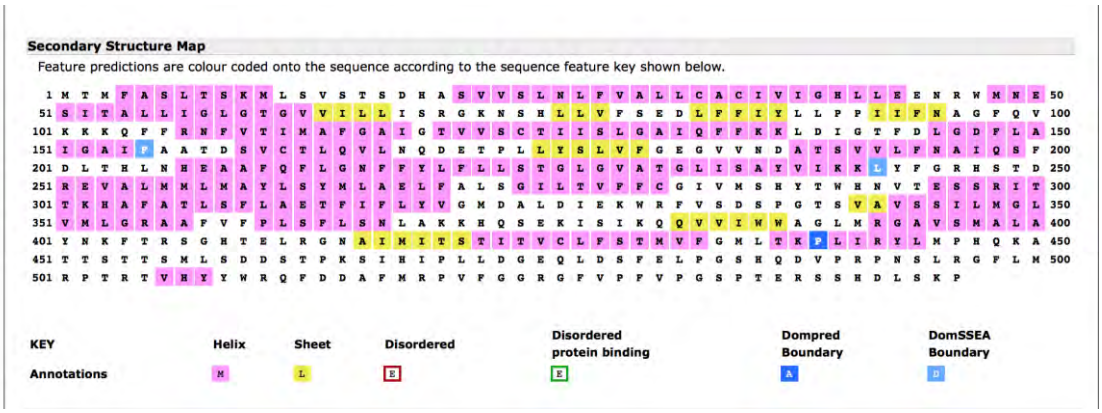
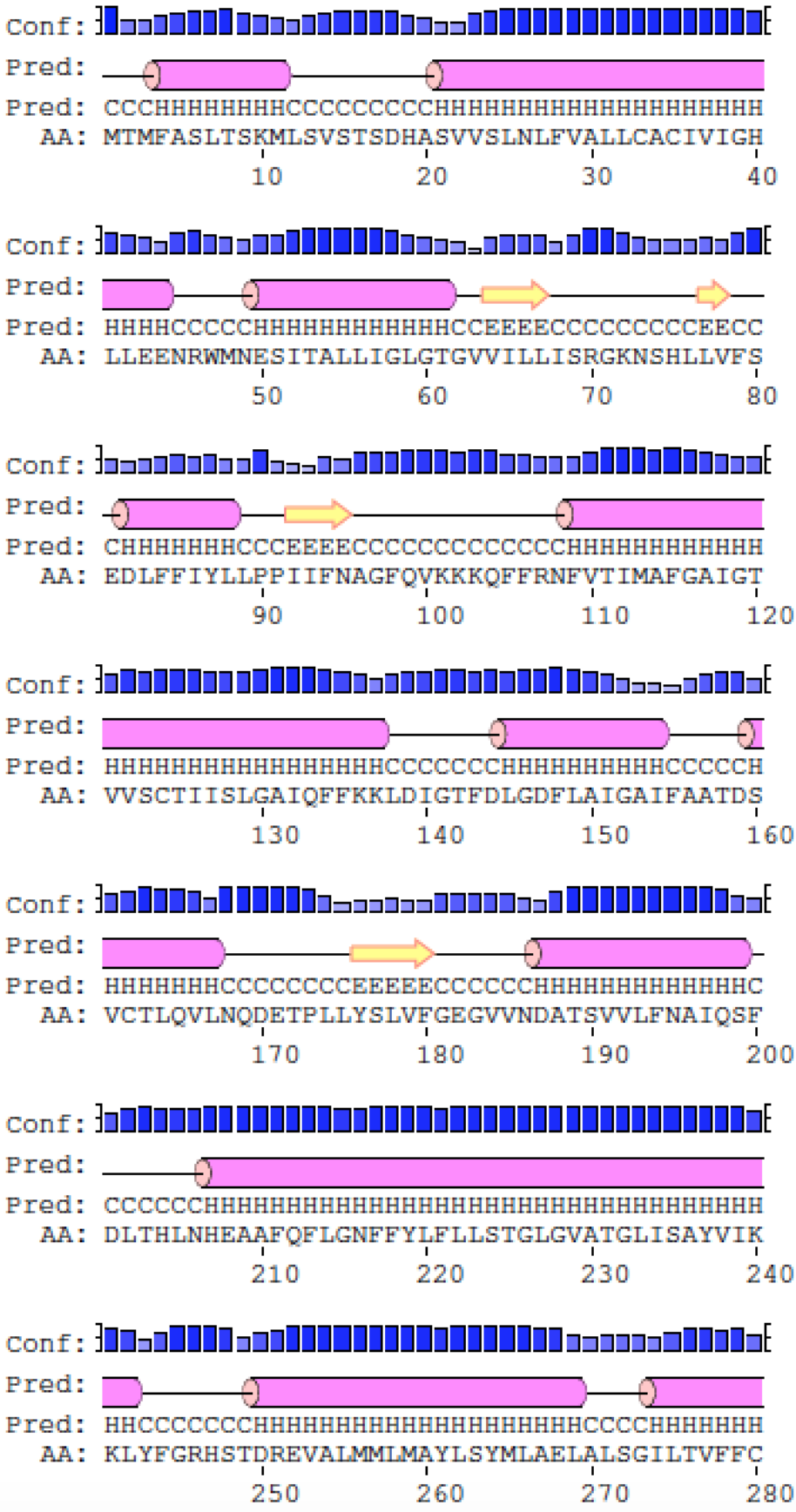
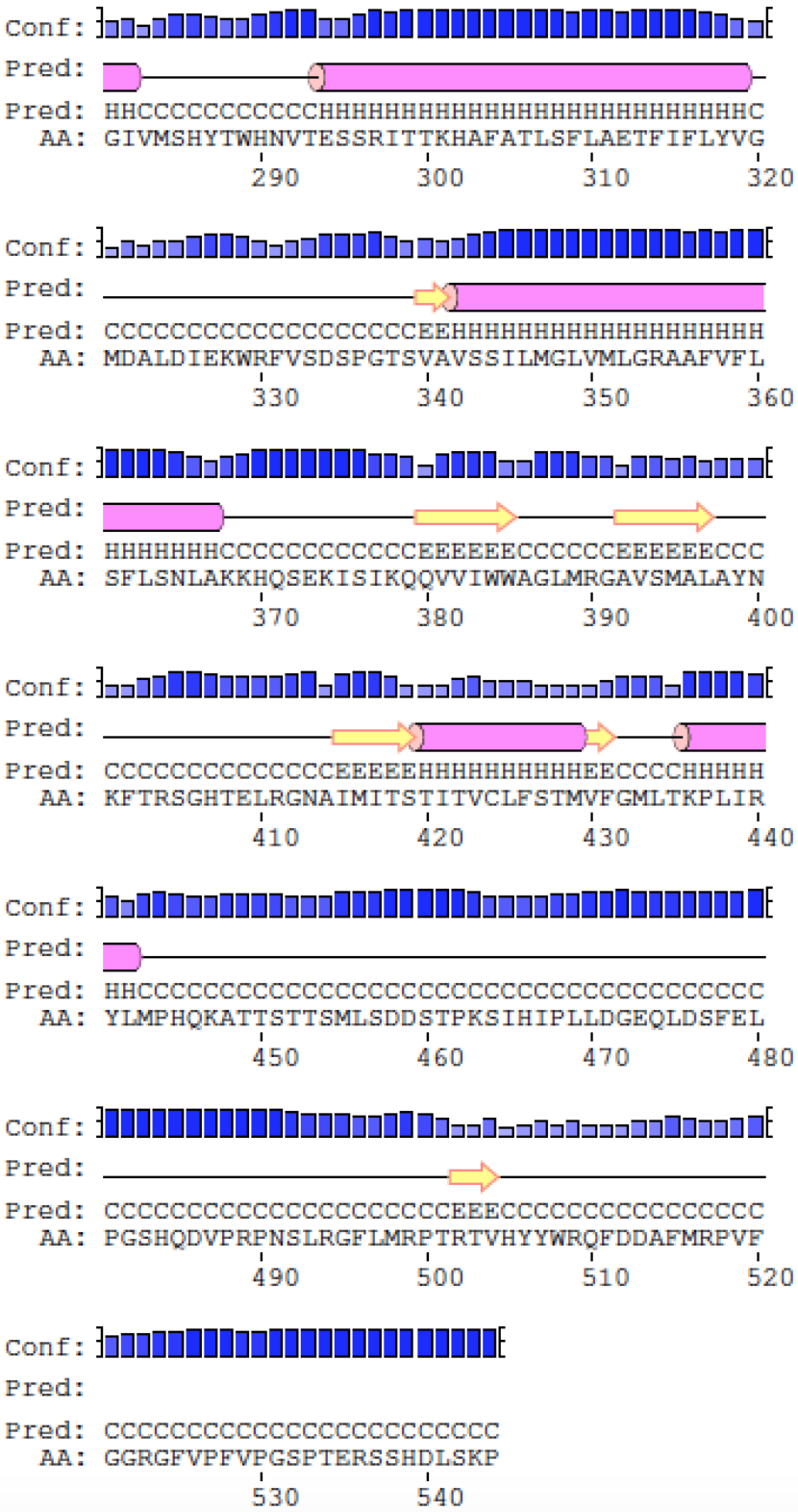


Figure 36.1: Secondary Structure map of AtNHX2 obtained via PSIPRED. The Purple box indicates the amino acid that forms helical structure and the yellow box indicates the amino acids forming beta sheets.


Conf:
Pred:
Pred: CCCHHHHHHHHCCCCCCCCCHHHHHHHHHHHHHHHHHHH
AA: MTMFASLTSKMLSVSTSDHASVVSLNLFVALLCACIVIGH
10 20 30 40
Conf:
Pred:
Pred: HHHHCCCCCHHHHHHHHHHHHCCEEEECCCCCCCCCEECC
AA: LLEENRWMNESITALLIGLGTGVVILLISRGKNSHLLVFS
50 60 70 80
Conf:
Pred:
Pred: CHHHHHHHCCCEEEECCCCCCCCCCCCCHHHHHHHHHHH
AA: EDLFFIYLLPPIIFNAGFQVKKKQFFRNFVTIMAFGAIGT
90 100 110 120
Conf:
Pred:
Pred: HHHHHHHHHHHHHHHHHCCCCCCHHHHHHHHHHCCCCCH
AA: VVSCTIISLGAIQFFKKLDIGTFDLGDFLAIGAIFAATDS
130 140 150 160
Conf:
Pred:
Pred: HHHHHHHCCCCCCCCEEEEECCCCCCHHHHHHHHHHHHC
AA: VCTLQVLNQDETPLLYSLVFGEGVVNDATSVVLFNAIQSF
170 180 190 200
Conf:
Pred:
Pred: CCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
AA: DLTHLNHEAAFQFLGNFFYLFLLSTGLGVATGLISAYVIK
210 220 230 240
Conf:
Pred:
Pred: HHCCCCCCCHHHHHHHHHHHHHHHHHHHHCCCCHHHHHH
AA: KLYFGRHSTDREVALMMLMAYLSYMLAELALSGILTVFFC
250 260 270 280


Conf:
Pred:
Pred: HHCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHC
AA: GIVMSHYTWHNVTESSRITTKHAFATLSFLAETFIFLYVG
290 300 310 320
Conf:
Pred:
Pred: CCCCCCCCCCCCCCCCCCCEEHHHHHHHHHHHHHHHHHH
AA: MDALDIEKWRFVSDSPGTSVAVSSILMGLVMLGRAAFVFL
330 340 350 360
Conf:
Pred:
Pred: HHHHHHHCCCCCCCCCCCCEEEEEECCCCCCEEEEEECCC
AA: SFLSNLAKKHQSEKISIKQQVVIWWAGLMRGAVSMALAYN
370 380 390 400
Conf:
Pred:
Pred: CCCCCCCCCCCCCCEEEEEHHHHHHHHHHEECCCCHHHHH
AA: KFTRSGHTELRGNAIMITSTITVCLFSTMVFGMLTKPLIR
410 420 430 440
Conf:
Pred:
Pred: HHCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
AA: YLMPHQKATTSTTSMLSDDSTPKSIHIPLLDGEQLDSFEL
450 460 470 480
Conf:
Pred:
Pred: CCCCCCCCCCCCCCCCCCCCEEECCCCCCCCCCCCCCCC
AA: PGSHQDVPRPNSLRGFLMRPTRTVHYYWRQFDDAFMRPVF
490 500 510 520
Conf:
Pred:
Pred: CCCCCCCCCCCCCCCCCCCCCCC
AA: GGRGFVPFVPGSPTERSSHDLSKP
530 540

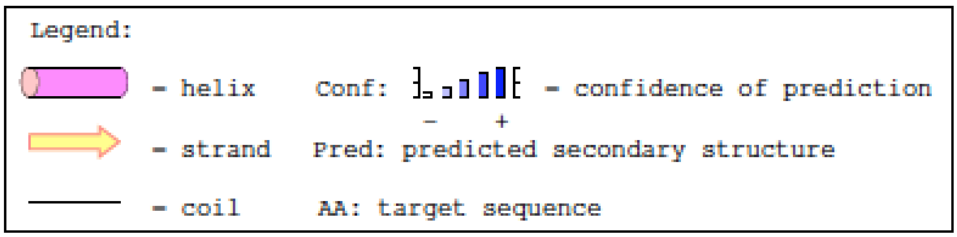


Figure 36.2: Graphical representation of the predicted secondary structures present within the target protein AtNHX2

As per the legend provided, the pink cylinders represented the alpha helices and the yellow arrows represented beta strands. The black threads like structures represent coils. As discussed earlier, the trans-membrane proteins are dominated by alpha helices and secondly by beta strands. Beta strands can round together to form beta barrels, which can also act like channels.

Type	Number
Helix	15
Beta sheets	10
Coils	24

Table 9: Total number of secondary structures present in the query sequences predicted by PSIPRED

3.1.10 Homology Modeling

The 3D models of the target protein were constructed using four protein structure homology model-building programs: I-TASSER, PHYRE, SWISS Model and EasyModeller. Molecular graphics and analyses were performed with the UCSF Chimera package.

3.1.10.1 Homology Modeling using I-TASSER

In this method the target sequences are threaded using a representative PDB structure library (Zhang, 2008). Profile-Profile Alignment (PPA), Hidden Markov Model, PSI-

BLAST profiles, Needleman-Wunsch and Smith-Waterman alignment algorithms (Suganya, Sudevan, Kalva, & Saleena, 2014), are used to search for the possible folds in the query protein. The software automatically selects the best-ranked templates, based on scores, to generate 5 models and then further evaluates in between those 5 models that were generated. The Z-score of the template 4cz9A was high among the query templates (3.91) that were matched with the query sequence. It an also be deduced that Iden1 (0.22) value was greater than that of the Iden2 (0.20) value, on a very low margin which made the template ideal for modeling structure. This indicated conserved structural motifs in the query sequence in comparison to the template.

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 means a good alignment and vice versa.

Note: The final models are prioritized based on their C-score, SCOP and RMSD. The confidence score or the C-score is calculated based on the significance of threading template alignments and convergence parameters of the structure assembly simulations. It is scored between -5 to 2 and higher the value more is the confidence in the structure.

Template modeling score or the TM score determines the structural similarity between 2 protein models including the query protein sequence and the template. It is scoring range of 0 to1, with 1 being the perfect match. However, score below 0.17 indicates or corresponds to randomly selected unrelated protein and higher than 0.5 is an indication of the same fold in the SCOP/CATH.

#	PDB Hit	Iden1	Iden2	Cov	Norm. Z-Score
1	4cz8A	0.21	0.19	0.72	2.13
2	1qgrA	0.08	0.20	0.99	2.23
3	4cz9A	0.22	0.20	0.72	3.91
4	5ezmA	0.10	0.19	0.89	1.01
5	4cz8A	0.21	0.19	0.71	2.43
6	1b3uA	0.09	0.17	0.97	1.30
7	4czbA	0.17	0.20	0.71	6.55
8	2bkuB	0.09	0.19	1.00	2.16
9	4cz8A	0.21	0.19	0.71	2.50
10	4cz8A	0.23	0.19	0.71	4.52

Table 10: Top ranking templates used by I-TASSER to model the target protein

SCOP (Structural classification of proteins) determines the classification and function of the proteins by structure and CATH (Class, Architecture, Topology, Homology) also is a scoring method that determines the structure of the protein.
RMSD (Root Mean Square Deviation) calculates the average distance between the atoms (backbone atoms) of superimposed proteins between the template and the query protein.

#Model Number	C-score
1	-0.95
2	-1.36
3	-2.36
4	-3.17
5	-3.17

Table 11: C-Scores of the generated models

The best model had a C-value of -0.95 and RMSD value of 9.7

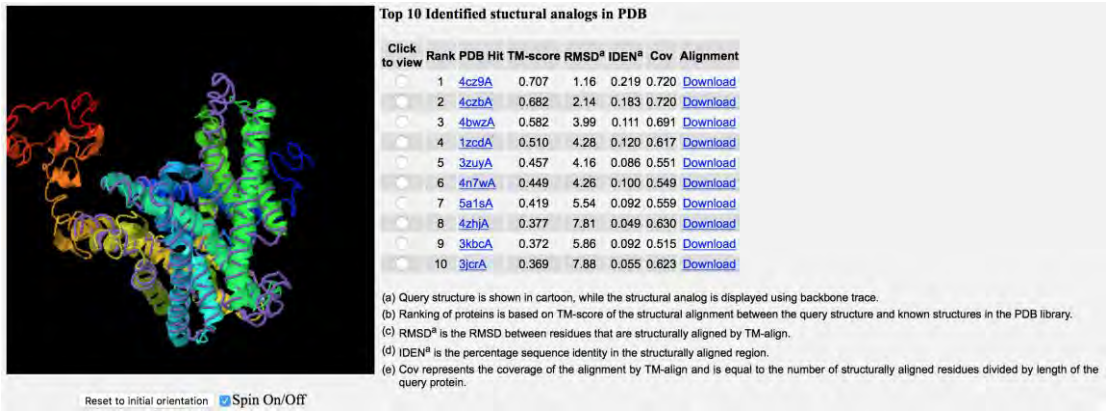
The I-TASSER server provided a list of the top ten identified structural analogs (table 12). Structurally conserved residues and motifs were observed via visual inspection. As such, the preferred structural analogs based on the parameters provided were the first ranking PDB ID entries.

Rank	PDB Hit	TM-Score	RMSD[a]	IDEN[a]	Cov
1	4cz9A	0.707	1.16	0.219	0.720
2	4czbA	0.682	2.14	0.183	0.720
3	4bwzA	0.582	3.99	0.111	0.691
4	1zcdA	0.510	4.28	0.120	0.617
5	3zuyA	0.457	4.16	0.086	0.551
6	4n7wA	0.449	4.26	0.100	0.549
7	5a1sA	0.419	5.54	0.092	0.559
8	4zhjA	0.377	7.81	0.049	0.630
9	3kbcA	0.372	5.86	0.092	0.515
10	3jcrA	0.369	7.88	0.055	0.623

Table 12: Top ten identified structural analogs in PDB provided by I-TASSER

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, 4cz9A was the preferred choice. The TM-score was highest (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 37).



Top 10 Identified stuctural analogs in PDB

Click to view	Rank	PDB Hit	TM-score	RMSD[a]	IDEN[a]	Cov	Alignment
	1	4cz9A	0.707	1.16	0.219	0.720	Download
	2	4czbA	0.682	2.14	0.183	0.720	Download
	3	4bwzA	0.582	3.99	0.111	0.691	Download
	4	1zcdA	0.510	4.28	0.120	0.617	Download
	5	3zuyA	0.457	4.16	0.086	0.551	Download
	6	4n7wA	0.449	4.26	0.100	0.549	Download
	7	5a1sA	0.419	5.54	0.092	0.559	Download
	8	4zhjA	0.377	7.81	0.049	0.630	Download
	9	3kbcA	0.372	5.86	0.092	0.515	Download
	10	3jcrA	0.369	7.88	0.055	0.623	Download

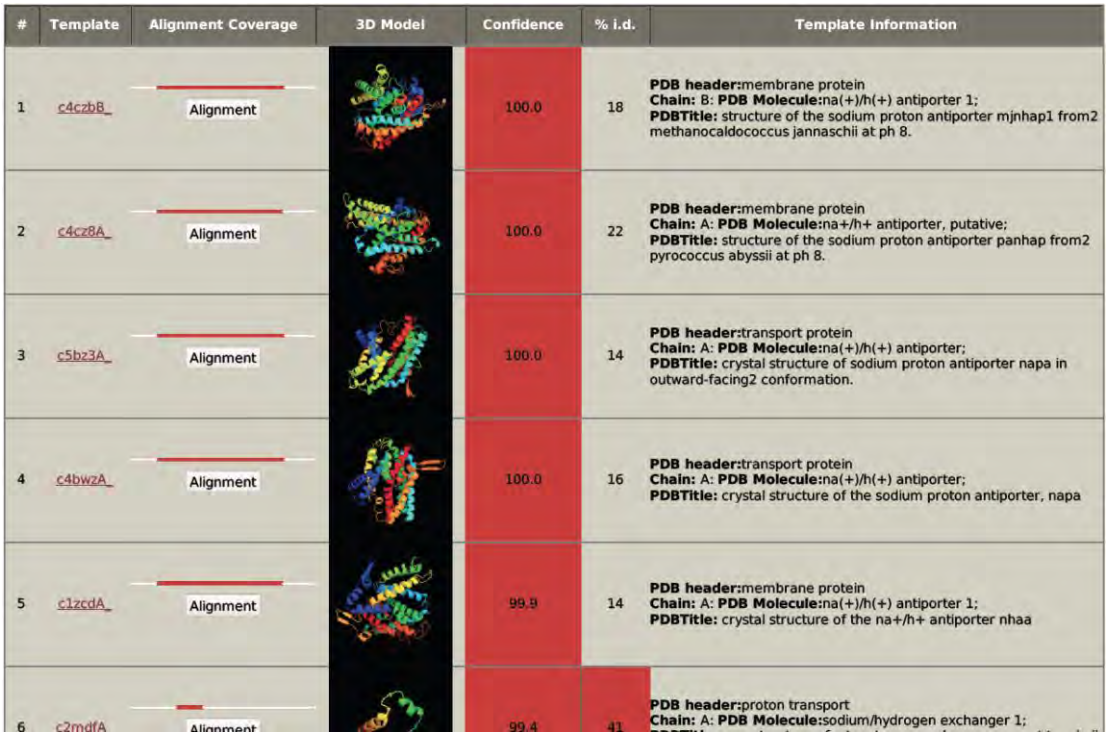
(a) Query structure is shown in cartoon, while the structural analog is displayed using backbone trace.
(b) Ranking of proteins is based on TM-score of the structural alignment between the query structure and known structures in the PDB library.
(c) RMSD[a] is the RMSD between residues that are structurally aligned by TM-align.
(d) IDEN[a] is the percentage sequence identity in the structurally aligned region.
(e) 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.

Reset to initial orientation Spin On/Off

Figure 37: Structure superposition of query protein (shown in cartoon) and template protein (shown in backbone)

3.1.10.2 Homology modeling using Phyre^2

The Phyre^2 tool modeled the query protein using multiple templates with the highest sequence coverage and confidence. The intensive mode was selected to achieve the desired output. In the Phyre^2 model of the query sequence, 469 residues (86%) modeled at greater than 90% accuracy. Template with highest confidence score was used to model the query protein.



#	Template	Alignment Coverage	3D Model	Confidence	% i.d.	Template Information
1	c4czbB_	Alignment		100.0	18	**PDB header:**membrane protein **Chain:** B: **PDB Molecule:**na(+)/h(+) antiporter 1; **PDBTitle:** structure of the sodium proton antiporter mjnhap1 from2 methanocaldococcus jannaschii at ph 8.
2	c4cz8A_	Alignment		100.0	22	**PDB header:**membrane protein **Chain:** A: **PDB Molecule:**na+/h+ antiporter, putative; **PDBTitle:** structure of the sodium proton antiporter panhap from2 pyrococcus abyssii at ph 8.
3	c5bz3A_	Alignment		100.0	14	**PDB header:**transport protein **Chain:** A: **PDB Molecule:**na(+)/h(+) antiporter; **PDBTitle:** crystal structure of sodium proton antiporter napa in outward-facing2 conformation.
4	c4bwzA_	Alignment		100.0	16	**PDB header:**transport protein **Chain:** A: **PDB Molecule:**na(+)/h(+) antiporter; **PDBTitle:** crystal structure of the sodium proton antiporter, napa
5	c1zcdA_	Alignment		99.9	14	**PDB header:**membrane protein **Chain:** A: **PDB Molecule:**na(+)/h(+) antiporter 1; **PDBTitle:** crystal structure of the na+/h+ antiporter nhaa
6	c2mdfA	Alignment		99.4	41	**PDB header:**proton transport **Chain:** A: **PDB Molecule:**sodium/hydrogen exchanger 1;

Figure 38: Templates used by Phyre^2 to model the protein with confidence score coloring

3.1.10.3 Homology modeling using Swiss Model

The query protein sequence was uploaded in the Swiss Model server. The software automatically scans through the PDB database and provides with an option of selection of template according to the corresponding scores. Out of all templates provided, three were selected based on highest coverage scores, sequence identity and similarity scores (figure 39).

Template	Seq Identity	Oligo-state	Found by	Method	Resolution	Seq Similarity	Coverage	Description
4cza.1.A	22.57	homo-dimer	HHblits	X-ray	3.20Å	0.30	0.70	NA+/H+ ANTIPORTER, PUTATIVE
4cza.1.B	22.57	homo-dimer	HHblits	X-ray	3.20Å	0.30	0.70	NA+/H+ ANTIPORTER, PUTATIVE
4cz9.1.A	22.31	homo-tetramer	HHblits	X-ray	3.50Å	0.30	0.70	NA+/H+ ANTIPORTER, PUTATIVE
4cz9.1.B	22.31	homo-tetramer	HHblits	X-ray	3.50Å	0.30	0.70	NA+/H+ ANTIPORTER, PUTATIVE
4cz8.1.A	22.31	homo-dimer	HHblits	X-ray	3.15Å	0.30	0.70	NA+/H+ ANTIPORTER, PUTATIVE
4cz8.1.B	22.31	homo-dimer	HHblits	X-ray	3.15Å	0.30	0.70	NA+/H+ ANTIPORTER, PUTATIVE
4czb.2.A	18.39	hetero-oligomer	HHblits	X-ray	3.50Å	0.28	0.71	NA(+)/H(+) ANTIPORTER 1
4d0a.1.A	18.49	homo-dimer	HHblits	2DX	6.00Å	0.28	0.70	NA(+)/H(+) ANTIPORTER 1
4czb.1.B	18.49	homo-dimer	HHblits	X-ray	3.50Å	0.28	0.70	NA(+)/H(+) ANTIPORTER 1
4czb.2.B	18.49	hetero-oligomer	HHblits	X-ray	3.50Å	0.28	0.70	NA(+)/H(+) ANTIPORTER 1
4czb.1.A	18.49	homo-dimer	HHblits	X-ray	3.50Å	0.28	0.70	NA(+)/H(+) ANTIPORTER 1
5bz2.1.A	14.79	homo-dimer	HHblits	X-ray	3.70Å	0.28	0.67	Na(+)/H(+) antiporter
5bz3.1.A	14.92	homo-dimer	HHblits	X-ray	2.30Å	0.28	0.66	Na(+)/H(+) antiporter
4bwz.1.A	14.17	homo-dimer	HHblits	X-ray	2.98Å	0.28	0.66	NA(+)/H(+) ANTIPORTER
4atv.1.A	15.02	homo-dimer	HHblits	X-ray	3.50Å	0.27	0.57	NA(+)/H(+) ANTIPORTER NHAA
4au5.1.A	14.70	homo-dimer	HHblits	X-ray	3.70Å	0.27	0.57	NA(+)/H(+) ANTIPORTER NHAA
1zcd.1.A	15.81	monomer	HHblits	X-ray	3.45Å	0.28	0.50	Na(+)/H(+) antiporter 1
3fi1.1.A	15.81	homo-dimer	HHblits	2DX	7.00Å	0.28	0.50	Na(+)/H(+) antiporter nhaA
5a1s.1.A	10.67	homo-dimer	HHblits	X-ray	2.50Å	0.25	0.27	CITRATE-SODIUM SYMPORTER
5a1s.1.B	10.67	homo-dimer	HHblits	X-ray	2.50Å	0.25	0.27	CITRATE-SODIUM SYMPORTER
5a1s.2.A	10.67	homo-dimer	HHblits	X-ray	2.50Å	0.25	0.27	CITRATE-SODIUM SYMPORTER
5a1s.2.B	10.67	homo-dimer	HHblits	X-ray	2.50Å	0.25	0.27	CITRATE-SODIUM SYMPORTER
4bwz.1.A	15.15	homo-dimer	HHblits	X-ray	2.98Å	0.30	0.12	NA(+)/H(+) ANTIPORTER
4cza.1.A	11.94	homo-dimer	HHblits	X-ray	3.20Å	0.28	0.12	NA+/H+ ANTIPORTER, PUTATIVE
4cza.1.B	11.94	homo-dimer	HHblits	X-ray	3.20Å	0.28	0.12	NA+/H+ ANTIPORTER, PUTATIVE
2mdf.1.A	45.83	monomer	HHblits	NMR	NA	0.40	0.09	Sodium/hydrogen exchanger 1
2e30.1.B	21.62	hetero-oligomer	HHblits	NMR	NA	0.29	0.07	Sodium/hydrogen exchanger 1
2bec.1.B	21.62	hetero-oligomer	HHblits	X-ray	2.70Å	0.29	0.07	Sodium/hydrogen exchanger 1
2k3c.1.A	41.94	monomer	HHblits	NMR	NA	0.40	0.06	TMIX peptide
1y4e.1.A	44.00	monomer	HHblits	NMR	NA	0.43	0.05	Sodium/hydrogen exchanger 1
2m7x.1.A	17.86	monomer	HHblits	NMR	NA	0.30	0.05	Na(+)/H(+) antiporter
2l0e.1.A	36.00	monomer	HHblits	NMR	NA	0.36	0.05	Sodium/hydrogen exchanger 1
2htg.1.A	59.09	monomer	HHblits	NMR	NA	0.45	0.04	NHE1 isoform of Na+/H+ exchanger 1

Figure 39: Swiss Model generated templates

Three models were generated and based on the selected templates, produced three models with almost identical score with 4czaA having slightly higher sequence identity score and template score.

# Models	Template	Template score (Identity Score)	Q-mean	Coverage	Sequence Similarity	Sequence identity
1	4czaA	22.57	-5.33	0.70	0.30	22.57
2	4cz8	22.31	-5.46	0.70	0.30	22.31
3	4cz9	22.31	-5.88	0.70	0.30	22.31

Table 13: Score of the models generated by individual templates on Swiss Model

3.1.10.4.1 Homology modeling using EasyModeller

EasyModeller is a manual-modeling tool where one is not restricted with any parameter. 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 40 a). The black 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 (less than 40). Out of 4 hits three were selected as templates based on high query cover, low E-value and higher identity scores (figure 40 b).

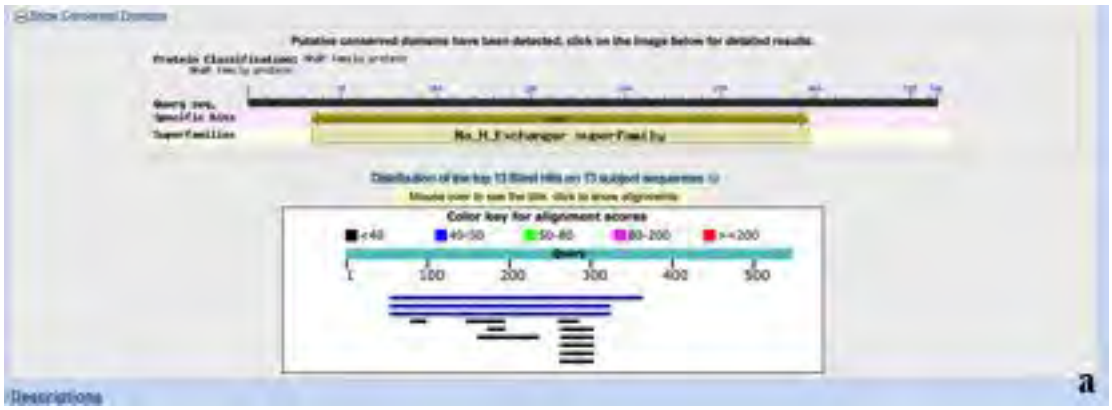


Figure 40 a: Graphical summary of BLAST results against PDB for template selection

Figure 40 b: Top 4 homologous sequences that produced significant alignments

Templates	Identity Scores	E-value	Query Cover
4CZ9_A	24%	2e-04	56%
4CZA_A	25%	3e-04	49%
4CZ8_A	25%	3e-04	49%
2MDF_A	45%	66e-04	8%

Table 14: List of selected templates attained using blast P-suite against Protein Data Bank (PDB)

A list of homologous sequences that produced significant alignments for the target protein sequence was attained. 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 14). These three templates were selected as they had the lowest E-values, moderate query coverage and identity values. These templates referred to the structure of sodium proton antiporter PaNhaP from Pyrococcus abyssii (Wöhlert, Kühlbrandt, & Yildiz, 2014). 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 41).

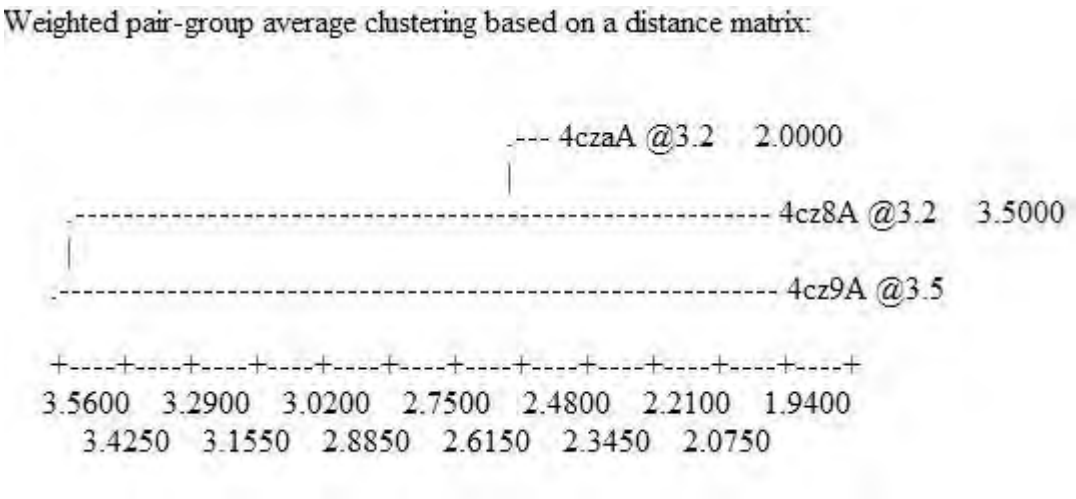


Figure 41: Comparison between selected templates using weighted pair-group average clustering based on distance matrix

#	Templates	Resolution factor	Similarity
1	4czaA	3.2	2.0000
2	4cz8A	3.2	3.5000
3	4cz9A	3.5	No value obtained

Table 15: Comparison between selected templates using weighted pair-group average clustering based on distance matrix

Note: The higher the resolution factor (R-factor) and higher the Sequence Similarity the more preferable is the template for model generation. R-Factor determines the resolution of the image generated by the NMR- spectroscopy during the modeling of the template and the similarity score determines the similarity between the template and the query.

The substrate ion in the dimeric and electroneutral sodium/proton antiporter was resolved PaNhaP for *Pyrococcus abyssi* at 3.2 Å. Furthermore, it was determined that the structure of the above-mentioned protein exhibits two different conformations at pH 8 and pH 4 (Wöhlert, Kühlbrandt, & Yildiz, 2014). 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 2 protein (table 15).

3.1.10.4.2 Template and query sequence alignment-using EasyModeller

The EasyModeller software was used to align the template with the query protein sequence (figure 42). The red blocks represented conserved regions and the amino acid residues were colored according to similarity. Additionally, the alignment accomplished through EasyModeller, also illustrated 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 an assignment 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 42: Template and query sequence alignment by EasyModeller

3.1.10.4.3 Models generated by EasyModeller

Five models were generated by EasyModeller (table 16). The models were selected based on low molpdf values, low DOPE (Discrete Optimized Protein Energy) and high GA341 scores.

Note:

Molpdf score: This determines the sum of restraint violation. The lower the value means less restrain violations thus proving a good model.

DOPE (Discreet optimized protein energy): It is an energy score that uses standard Modeller energy functions. This determines the least amount of energy required for 3-D structure formation.

GA341: This uses percentage sequence identity between model and template as parameter. It is scored 0 to 1 and higher value means higher similarity between the template and the query model.

According to the parameters mentioned above, models 2 and 4 were considered to be the outstanding in nature.

Model Number	Molpdf	GA341	DOPE score
1	3592.92700	0.17125	-58362.26172
2	3782.89941	0.29658	-59908.55078
3	3832.02393	0.07589	-59286.87109
4	3289.87402	0.11672	-58943.25781
5	4262.76270	0.14031	-57468.32422

Table 16: Summary of the five models generated by EasyModeller (EM)

3.1.11 Model Comparison

Models provided by I_TASSER, Phyre^2, EasyModeller and Swiss Model were all evaluated using various tools. These include PROCHECK for calculating Ramachandran Plot calculations and QMean server for determining model reliability.

G-Factor value and Root Mean Square Deviation (RMSD) values were attained to establish any unusual properties resided within the model and RMSD superimposes the predicted model with the template used calculating and predicting deviations from the experimented structure with the predicted structure.

3.1.11.1 Model Validation: EasyModeller

The two selected models generated by EasyModeller were evaluated using above-mentioned tools. The PROCHECK used PSVS server for evaluating the models and plot Ramachandran Plot. The results supported model 4 over model 2 (table 17) with higher "favored regions" in model 4 (84.7%) than that of model 2 (81.2%). 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. Both of the predicted structure was well above the margins however, model 4 having higher G-Factor value and higher favored regions, was chosen over model 2.

#	Model 2	Model 4
Pro-check G-factor	-0.081	0.046
Ramachandran Plot Summary from Pro-check	---	---
Most favored regions	81.2%	84.7%
Additionally allowed regions	13.5%	11.2%
Disallowed Regions	1.2%	2.0%

Table 17: Selection of models based on scores generated by validation software.

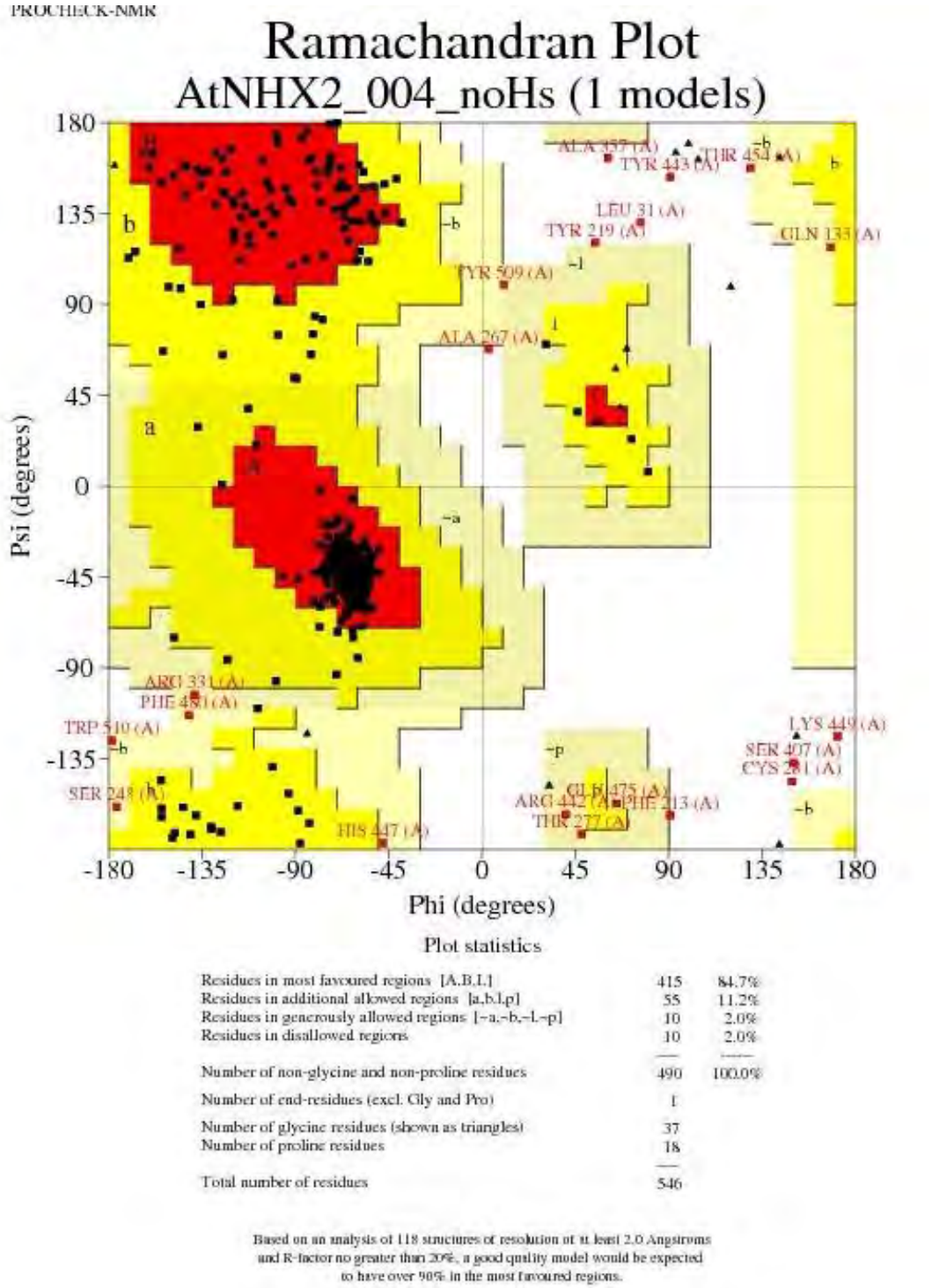


Figure 43: Ramachandran Plot of the model generated by EasyModeller

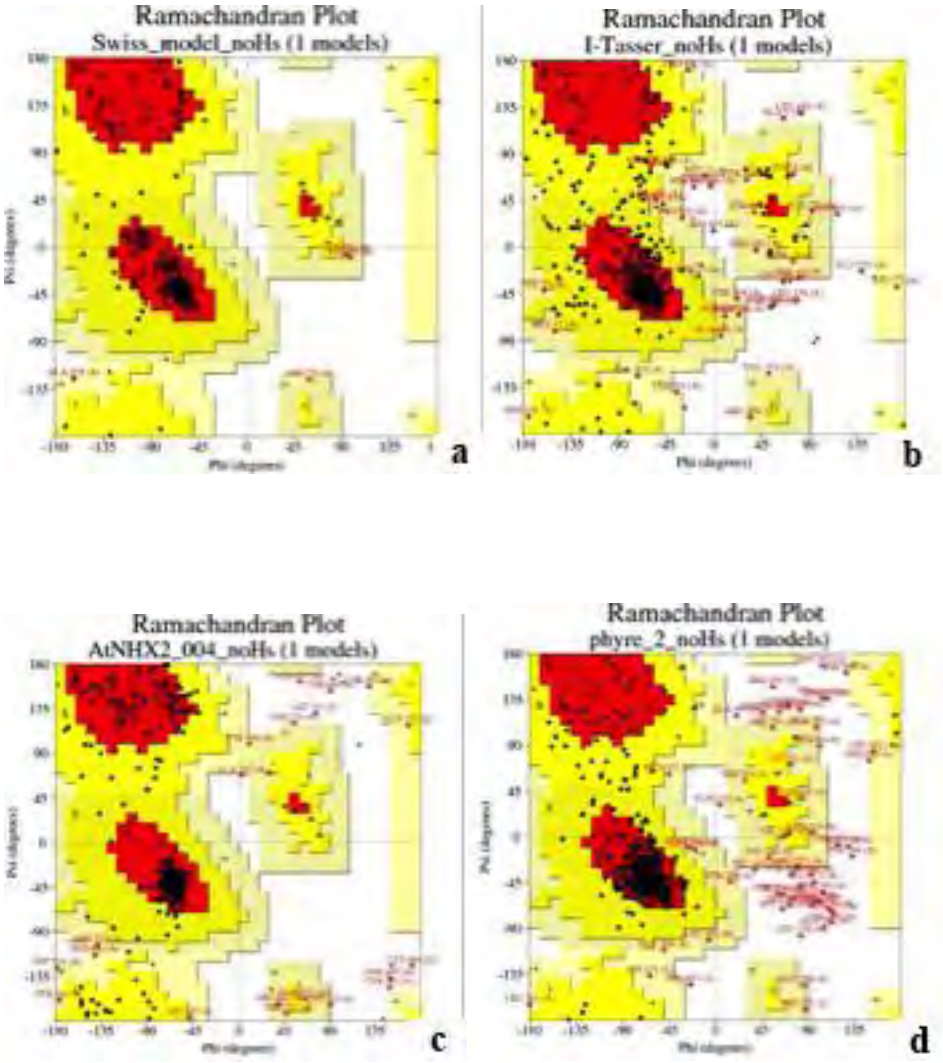
3.1.11.2 Selection of final model between models generated by I_TASSER, Phyre^2, EasyModeller, Swiss Model

Final comparison was conducted between 4 models generated by Easymodeller, Phyre 2, Swiss Model and I_TASSER. For the final model validation Ramachandran Plot calculations, G-Factor value, RMSD values were compared. The Q-Mean and the Q-Mean Z scores were also taken into account.

The preferential of the Ramachandran Plot was Swiss Model, which had the highest favored regions with a value of 91.5%, which is by far the highest of the four models generated. Moreover, in comparison to the other structures, the model generated by Swiss Model had the least additionally allowed regions (7.4%), generously allowed regions (1.1%) and disallowed regions (0.00%). 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 models did not have that high score. Furthermore, the G-Factor score of the I_TASSER and Phyre^2 had negative scores (-0.72 and -0.865 respectively), which meant that there were high amount of unusualness within the structures generated by those models. However, the models generated by the Swiss Model and EasyModeller had positive scores with Swiss Model outweighing EasyModeller with a score of 0.186 to 0.046. The RMSD scores were also decent for Swiss Model with a score of 0.712Å and the Q-mean score was around 0.428, which was highest out of all yielded results. This meant that out of all structures predicted the model predicted by Swiss Model was most stable. The Z- score was also the highest (-3.97) in comparison to other models generated. It should be noted that the trans-membrane proteins.

#	Phyre[2]	I-TASSER	Swiss Model	Easy Modeler
Ramachandran plot	---	---	---	---
Most Favored Regions	74.9%	70.6%	91.5%	84.7%
Additionally allowed regions	14.3%	22.7%	7.4%	11.2%
Generously allowed regions	4.1%	4.1%	1.1%	2.0%
Disallowed regions	6.7%	2.7%	0.0%	2.0%
G-Factor	-0.856	-0.72	0.186	0.046
RMSD	2.003	0.00	0.712	2.346
Q-Mean	0.317	0.416	0.428	0.237
Z-Score	-5.13	-3.99	-3.97	-6.06

Table 18: Selection of final model between models generated by I-TASSER, Phyre[2] and EasyModeller and Swiss Model.



3. EasyModeller 4. Phyre[2]

Figure 44: Ramachandran plots by PROCHECK: (a) Swiss Model (b) I-TASSER (c) EasyMdeller (d) Phyre[2]

3.1.12 Model Validation

The PDBsum Generate's ProMotif provided a summary of the secondary structure components present within the final selected model Schematic and topology diagrams showing the secondary structural elements within the model were attained from PDBsum tool. The significant structures that were mentioned by the ProMotif in the query structures were beta hairpin that was of class 3:5 in nature. This meant that within each monomer one hand of the beta hairpin arm interacts with the other part of

the beta hairpin to help form a tertiary structure, however not interacting with the rest of the protein (Sibanda , Blundell, & Thorton, 1989). The class 3:5 beta hairpin structure is formed with the help of two strands, one from amino acid residue 366 (Serine) and 367(Asparagine) and the other from amino acid residue 373 (Glutamine) to 374 (Serine). All of the amino acids, afore-mentioned, are polar in nature. These amino acids help in hydrogen bond formation thus providing more stability in soluble state (Alba, Rico, & Jimenez, 1999). The β-hairpin is the essential structural subunit of a trans-membrane β-barrel (Wimbley, 2003) thus forming a channel for possible ion transportation (Sansom & Kerr, 1995).

It also found beta turns, which are composed of 4 residues of consecutive amino acids and occurs if the center is not helical in nature out of total 9 classes of beta turns, 4 were exhibited (Class: 1,2,4,8) and total 23 beta turns were reported. Within these 23 reported beta turns reported, 14 exhibited hydrogen bond formation, which can be considered very low. However, this was compensated by other secondary structures, namely helices and helix-helix interacts, which easily fulfilled hydrogen bond formation ration. Finally, there was another significant secondary structure found in the amino acid structure, which was gamma turns. The gamma turns can make a chain or a sheet turn 40 degrees and occurs when 3 amino acid residues, formed by hydrogen bonds within the first two, causes the third amino acid to turn at an angle of 40 degrees.

Motifs	Numbers
Beta Sheet	1
Beta Hairpin	1
Beta Bulge	1
Beta Strands	2
Helices	25
Helix-Helix interacs	50
Beta turns	23
Gama turns	1

Table 19: Secondary structure summary of Swiss Model provided by ProMotif

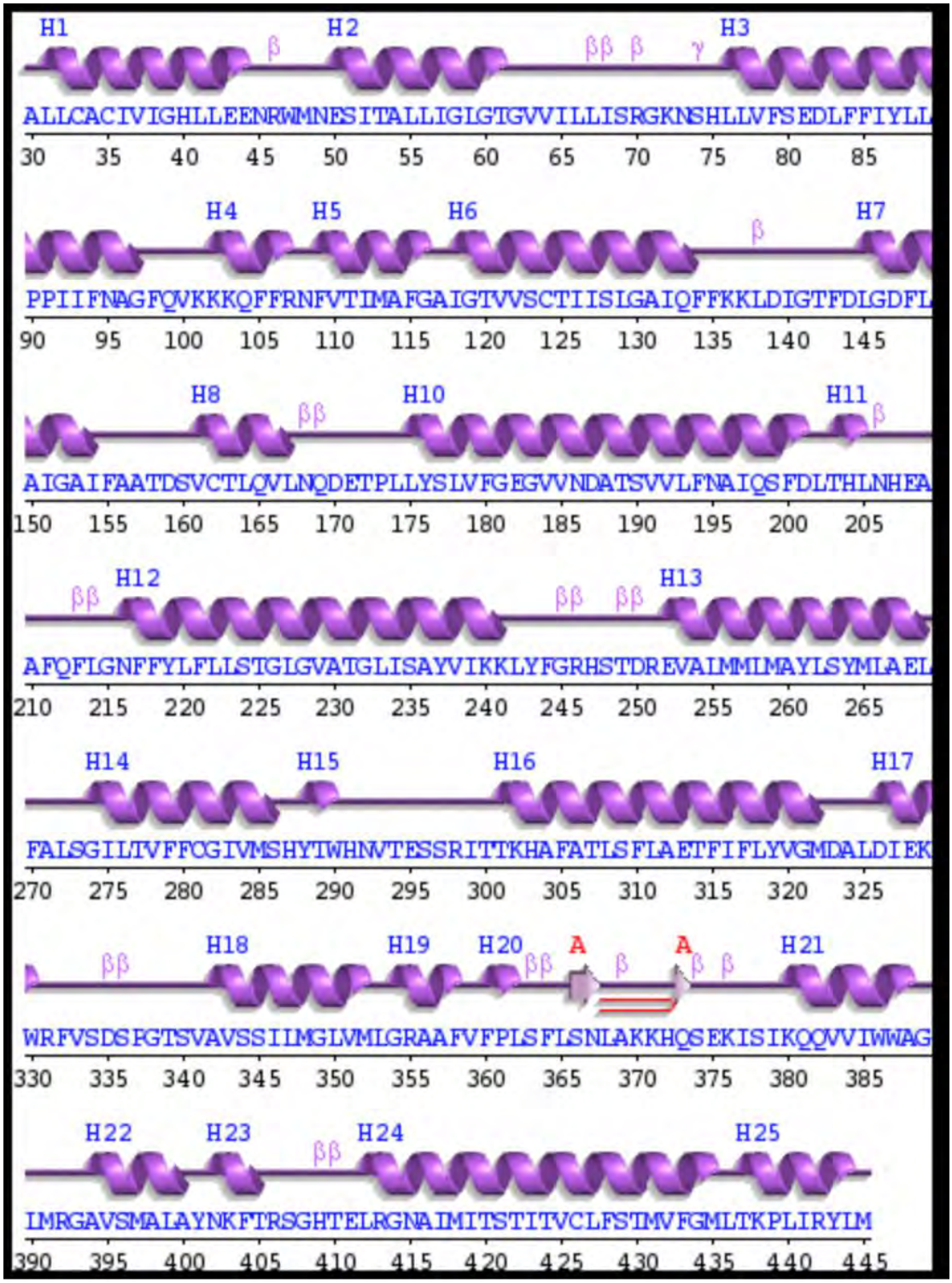


Figure 45: Schematic diagram showing the secondary structural elements in the final model attained from the PDBsum tool. The α-helices are labeled with the letter "H" and β-strands are lettered in the uppercase. β, γ and hairpin turns are also labeled.

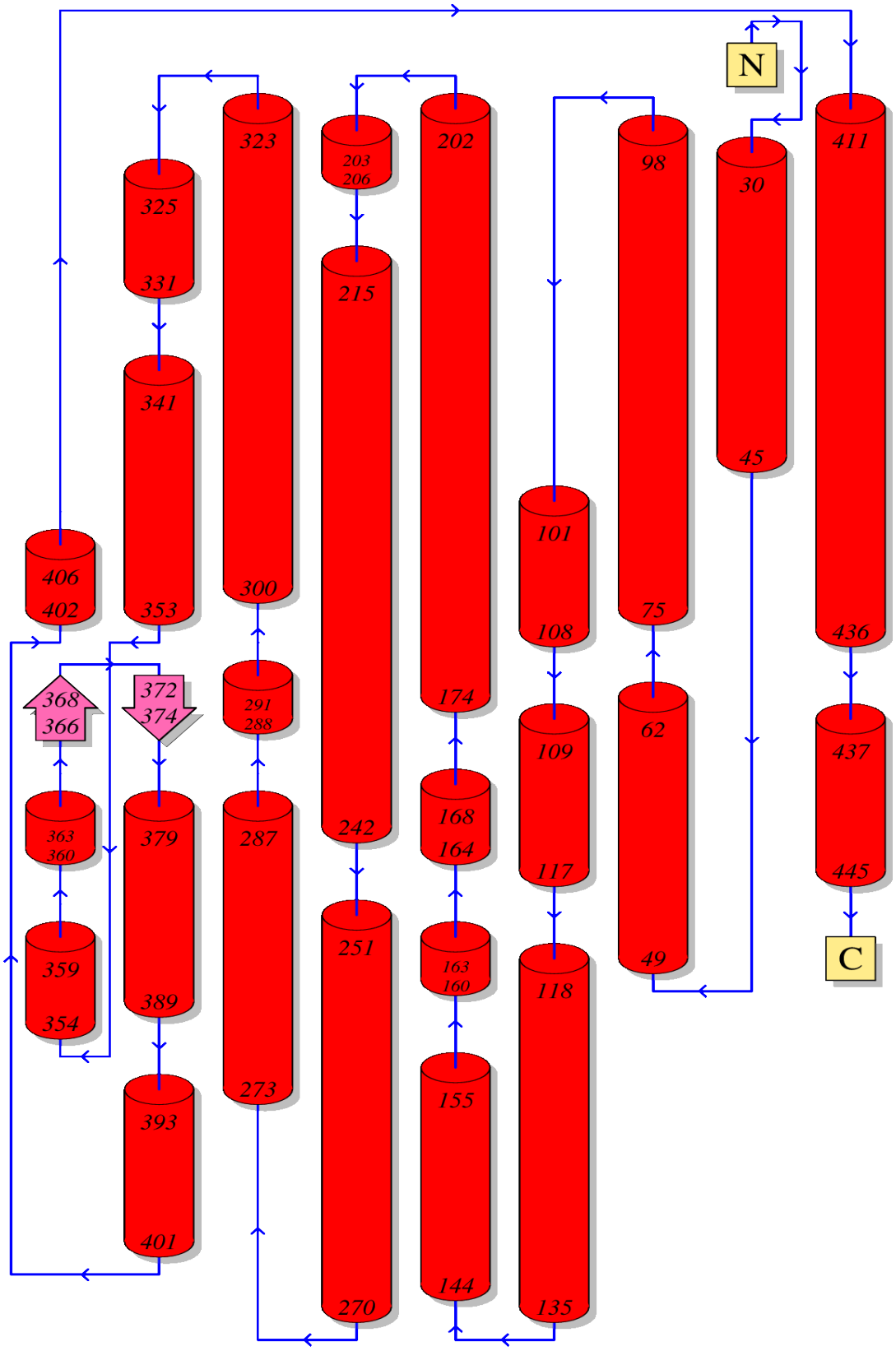


Figure 46: Topology diagram showing the secondary structural elements in the final model attained from the PDBsum tool. Helices are represented as cylinders and β-strands as arrows.

From the results predicted by PSIPRED there were far more alpha helices present within a structure. TMHMM server takes in account of the number of the helix and other structure present on the surface of the membrane. This data gets further validation from the TMHMM server where it states that there are overall 12 trans-membrane helixes spanning across the membrane protein establishing the fact that the helix membrane are responsible for the interaction of the sodium ions. However, if compared with the topology diagram, afore-mentioned by PDBsum (figure 46), it was seen the long alpha- helices, predicted by the two software, were further broken down and shorten by motifs, such as beta turns, which gave the protein its three-dimensional structure.

Since it is a channel protein it is most likeable to have tunnels and pores spanning through the structure and it is an anti-porter, it is likeable to have two tunnels. The result matched with the prediction and the final structure had the following tunnels, which can make it a transport protein and an anti-porter at the same instance.

Tunnel	1	2
Radius (Å)	1.11	1.21
Free R (Å)	1.15	1.26
Length (Å)	34.2	34.5
Residue Type	#	#
Positive (H, K, R)	1	1
Negative (D, E)	1	1
Neutral (S, T, N, Q)	2	3
Aliphatic (A, V, L, I, M)	6	9
Aromatic (F, Y, W)	4	2
Pro & Gly (P, G)	0	0
Cys (C)	0	0

Table 20: Physiochemical characteristics of the tunnels formed by the query protein as predicted by Mole 2 program in PDBSum.

It is to be noted that the tunnels less than 15 Å long are not considered because in an experiment it was found that around 65% of the total enzymes have pores longer than 15 Å so for an anti-porter tunnels should be feasibly larger (Pravda, et al., 2014). This fact followed by large abundance of aliphatic compounds required for stability inside of the protein is also matched thus giving further validation to the structure.

3.2 Discussion

3.2.1 Structural analysis

Tunnels are calculated using a method, which uses Voronoi diagram to map out and predict tunnel formation within the tertiary structure. Voronoi diagram is a partitioning of a plane into regions based on distance to points in a specific subset of the plane and then connecting them in such a way to map out a 3-Dimentional structure. The tunnels are computed using Mole 2.0. Two sets are calculated with and without ligands. First, the Voronoi diagram of the atomic centres of the protein is calculated. Then, any cavities are identified and their start and end points linked (Sehnal, et al., 2013). Various tunnel properties are computed and tabulated according to the amino acids lining the tunnel (table 20).

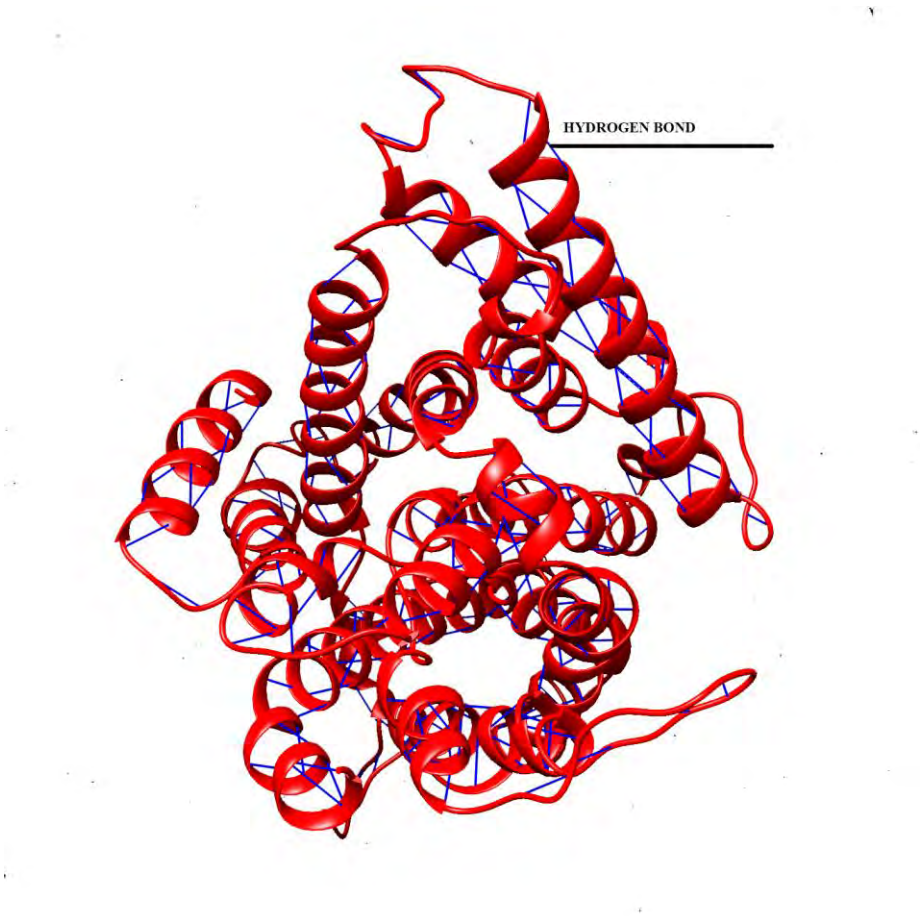


Figure 47: Final predicted structure of the AtNHX2 as viewed using Chimera

The structure, upon further research, has been found that it solely depends on the hydrogen bonds, especially hydrogen bonds present on the helical structure to interact with the hydrogen and the sodium molecule. Moreover, it has also been found that the helixes formed by the structure changes direction (parallel and anti-parallel) thus forming almost a crisscross structure and the overall tunnel. Furthermore, the channel validates for ion transport. This is supported by the fact that beta-turns reverses direction of a protein chain, in this case, helices (Petersen, Lundegaard, & Petersen, 2010). Hydrogen bonds, present in alpha helices have a central role in folding, stabilization and function of helical membrane. This brings regular structures, such as, alpha helix and beta structures together and makes them interact with each other. This phenomenon causes increase in stability and provides the structure with its function (Zheng & Kurgan, 2008). There ware 23 beta turns found in the result provided by PDBsum tool, which were almost equal to the alpha helix found in the structure (25) and almost half of helix-helix interacts (50). These values indicate the facts, mentioned above to be true with the structure (Senes, Ubarretxena-Belandia, & Engeman, 2001) (Admian & Liang, 2002) (Curran & Engelman, 2003).

In addition, it has been found that the hydrogen bonds break inside the helix structures in order for ion interaction with the protein thus increasing in the overall energy state of the protein structure conformation. However, each protein is designed to form the 3-Dimentional structures in their lowest energy state thus after ion transport it reverts back to its original state (Hanqiao, Zheng, & Yawen, 2004). The transport proteins are usually bell like in shape, which also resembles with the generated structure (Michalek, et al., 2012).

3.2.2 Structural difference between AtNHX1 and AtNHX2

From the experiment (Feisal, 2015), it was found that AtNHX1 had around 12 trans-membrane peaks at Protscale and TMHMM server, which was similar to the AtNHX2. Both of the proteins had peaks around similar areas of amino acids. In addition, theoretical pI value was also similar. AtNHX1 had a value of 6.95 and AtNHX2 had around 6.92 thus the proteins were similar in composition. However,

there was a significant visual difference in topology diagram in between AtNHX1 and AtNHX2. The N and C terminal of AtNHX2 was far closer than that of AtNHX1. Furthermore, there was a noticeable difference in between the position and the class of the beta hairpin that was present in AtNHX1 and in AtNHX2. The beta hairpin on AtNHX1 was found in between amino acid 197 (Serine) till 199 (Aspartine). The second strand ranged from residue 209 (Phenylalanine) till residue 211 (Leucine). In AtNHX2 the beta hairpin was from amino acid residue 366 (Serine) and 367(Asparagine) and the other from amino acid residue 373 (Glutamine) to 374 (Serine). Moreover, the class of hairpin formed was also different with AtNHX1 hairpin having 9:9 class and AtNHX2 having a class of 3:5. These might help in forming slightly different beta barrel formation thus leading to different functionality (Sibanda , Blundell, & Thorton, 1989). This phenomenon along with difference in topology diagram (PDBsum) might shed light to the difference in between the structure and function of AtNHX1 and AtNHX2. However, further research is necessary to shed light regarding functional difference in between AtNHX1 and AtNHX2.

Reference

- **Bibliography**

- Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., de Castro , E., et al. (2012). ExPASy: SIB bioinformatics resource portal . *Nucleic Acids Research* .
- Bassil, E., Coku, A., & Blumwald, E. (2012). Cellular ion homeostasis: emerging roles of intracellular NHX Na+/H+ antiporters in plant growth and development . *Journal of Experimental Botany* .
- Benkert, P., Biasini, M., & Schwede, T. (2011). Toward the estimation of the absolute quality of individual protein structure models . *Bioinformatics* .
- Benkert, P., Künzli, M., & Schwede, T. (2009). QMEAN server for protein model quality estimation. *Nucleic Acids Research* .
- Benkert, P., Tosatto, S. C., & Schomburg, D. (2008). QMEAN: A comprehensive scoring function for model quality assessment. *Proteins: Structure, Function, and Bioinformatics* .
- Boratyn, G. M., Camacho , C., Cooper, P. S., Coulouris , G., Fong, A., Ma, N., et al. (2013). BLAST: a more efficient report with usability improvements . *Nucleic Acids Research* .
- Buchan, D. W., Minneci, F., Nugent, T. C., Bryson , K., & Jones, D. T. (2013). Scalable web services for the PSIPRED Protein Analysis Workbench . *Nucleic Acids Research* .
- Coordinators, N. R. (2013). Database resources of the National Center for Biotechnology Information . *Nucleic Acids Research* .
- de Beer , T. A., Berka, K., Thornton, J. M., & Laskowski, R. A. (2013). PDBsum additions . *Nucleic Acids Research* .
- Eswar, N., Webb, B., Marti-Renom , M. A., Madhusudhan , M. S., Eramian, D., Shen, M. Y., et al. (2006). Comparative Protein Structure Modeling Using Modeller . *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* .
- Fiser, A., Do, R. K., & Sali, A. (n.d.). 'Modeling of loops in protein structures. *Protein Science : A Publication of the Protein Society* .

- Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S. e., Wilkins, M., Appel, R., et al. (2005). Protein Identification and Analysis Tools on the ExPASy Server . *The Proteomics Protocols Handbook* .
- Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., & Madden, T. L. (2008). NCBI BLAST: a better web interface', Nucleic Acids Research . *Nucleic Acids Research* .
- Krogh, A., Larsson, B., von Heijne , G., & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden markov model: application to complete genomes1 . *Journal of Molecular Biology* .
- Laskowski, R., MacArthur, M. W., Moss, D. S., & Thornton, J. M. (n.d.). PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography* .
- Laskowski, R., Rullmann, J. A., MacArthur, M., Kaptein, R., & Thornton , J. (1996). AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR. *Journal of Biomolecular NMR* .
- Martí-Renom , M. A., Stuart, A. C., Fiser, A., Sánchez, R., Melo, F., & Šali, A. (2000). Comparative Protein Structure Modeling of Genes and Genomes. *Annual Review of Biophysics and Biomolecular Structure* .
- McWilliam, H., Li, W., Uludag , M., Squizzato, S., Park, Y. M., Buso , N., et al. (2013). Analysis Tool Web Services from the EMBL-EBI . *Nucleic Acids Research* .
- Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., et al. (2004). UCSF Chimera—A visualization system for exploratory research and analysis . *Journal of Computational Chemistry* .
- Sali, A. (1995). Comparative protein modeling by satisfaction of spatial restraints. *Molecular Medicine Today* .
- Sievers , F., & Higgins, D. (2014). Clustal Omega, Accurate Alignment of Very Large Numbers of Sequences. *Multiple Sequence Alignment Methods Methods in Molecular Biology* .
- Wöhlert, D., Kühlbrandt, W., & Yildiz, Ö. (2014). Structure and substrate ion binding in the sodium/proton antiporter PaNhaP . *eLife* .
- Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., & Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction . *Nat Meth* .

- Admian, L., & Liang, J. (2002). Interhelical hydrogen bonds and spatial motifs in membrane proteins: Polar clamps and serine zippers . *Proteins: Structure, Function, and Bioinformatics* .
- Alba, E. D., Rico, M., & Jimenez, A. M. (1999). The turn sequence directs Beta-strand alignment in designed beta-hairpins. *Protein Science* .
- Amtmann, A., & Sanders, D. (1999). Mechanisrns of Na' uptake by plant cells. *Adv. Bot. Res. 29.75-1 12* .
- Arnold, K., Bordoli, L., Kopp, J., & Schwede, T. (2006). The SWISS-MODEL Workspace: A web-based environment for protein structure homology modelling. *Bioinformatics* .
- Bassil, E., Coku, A., & Blumwald, E. (2012). Cellular ion homeostasis: Emerging roles of intracellular NHX Na+/H+ antiporters in plant growth and development. *Journal of Experimental Botany* .
- Betz, S. F. (1993). Disulfide bonds and the stability of globular proteins . *Protein Science : A Publication of the Protein Society* .
- Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., et al. (2014). SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Research* .
- Blouin, C., Butt , D., & Roger, A. J. (2004). A study of loop regions in a ubiquitous GTP binding 81 domain. *Rapid evolution in conformational space:* , 608-616.
- Blumwald, E., & Poole, R. (1985). Na+/H+-antiport in isolated tonoplast vesicles from storage tissue of Beta vulgaris. *Plant Physiology* .
- Bordoli, L., Kiefer, F., Arnold, K., Benkert, P., Battey, J., & Schwede, T. (2009). Protein structure homology modelling using SWISS-MODEL Workspace. *Nature Protocols* .
- Brown, G. E. (2005). Jr Salinity Laboratory. *USSL* .
- Chothia , C., & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins . *The EMBO Journal* .
- Claverie, J. M., & Notredame, C. (2006). *Bioinformatics For Dummies. 2nd edn.* New Jersey: Wiley Publishing, Inc., p. 456.

- Corratge-Faillie, Jabnoune, M., Zimmermann, S., Very, A. A., Fizames, C., & Sentenac, H. (2010). Potassium and sodium transport in non-animal cells: the Trk/Ktr/HKT transporter family. *Cellular and Molecular Life Science* .
- Curran, A. R., & Engelman, D. M. (2003). Sequence motifs, polar interactions and conformational changes in helical membrane proteins . *Current Opinion in Structural Biology* .
- Delauney, A. J., & Vermas, D. P. (1993). Proline biosynthesis and osrnoregulation in plant. *Plant J* .
- Duan, S. (2001). Claning and characterization of an Arabidopsis hafiana vacuolar Na'M' antiport-AtNHX3.
- Elofsson, A., & Heijne, G. (2007). Membrane Protein Structure: Prediction versus Reality . *Annual Review of Biochemistry* .
- Feisal, M. R. (2015). IN SILICO STRUCTURAL ANALYSIS, PHYSICOCHEMICAL CHARACTERIZATION AND HOMOLOGY MODELING OF Arabidopsis thaliana NA+/H+ EXCHANGER 1 (AtNHX1) PROTEIN .
- Geer, L. Y., Marchler- Bauer, A., Geer , R. C., Han, L., He, J., He, S., et al. (2010). The NCBI BioSystems database. *Nucleic Acids Research, 38* .
- Gilbert, G. S., Chang, M. M., & Rojas, E. (2002). Fungal diversity and plant disease in mangrove forests: salt excretion as a possible defense mechanism. *Community Ecology* .
- Hanqiao, F., Zheng, Z., & Yawen, B. (2004, December). A protein folding pathway with multiple folding intermediates at atomic resolution. *Proceedings of the National Academy of Sciences of the United States of America* .
- Hu, Y., & Schmidhalter, U. (1998). Spatial distributions and net deposition rates of mineral elements in the elongating wheat (Triticum aestivum L.) leaf under saline soil conditions. *Planta* , 212–219.
- Jeschke, W. D. (1984). K+–Na+ exchange at cellular membranes, intracellular compartmentation of cations, and salt tolerance. In: Staples RC. *Salinity Tolerance in Plants: Strategies for Crop Improvement* , 37-66.
- Jeschke, W. D., Aslam, Z., & Greenway, H. (1986). Effects of NaCl on ion relations and carbohydrate status of roots and on osmotic regulation of

roots and shoots of Atriplex amnicola. . *Plant, Cell & Environment* , 559–569.

- Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices1 . *Journal of Molecular Biology* .
- Kalberg, Y. (2002). Bioinformatic methods in protein characterization. *PhD, Bioinformatics* .
- Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., & Sternberg, M. J. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols* .
- Komatsu, S., Konishi, H., & Hashimoto, M. (2007). The proteomics of plant cell membranes . *Journal of Experimental Botany* .
- Kumar, K., & Mosa, K. A. (2015). Ion Transporters: A Decisive Component of Salt Stress Tolerance in Plants.
- Kumar, S., Stecher, G., & Tamura, K. (2015). MEGA7: Molecular Evolutionary Genetics Analysis version 7.0. *Molecular Biology and Evolution* .
- Kyte , J., & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein . *Journal of Molecular Biology* .
- Li, W., Cowley, A., Ulndag, M., Gur, T., McWilliam, H., Squizzato , S., et al. (2015). The EMBL-EBI bioinformatics web and programmatic tools framework . *Nucleic Acids Research* .
- Marban, E., Yamagishi, T., & Tomaselli, G. F. (1998). Structure and function of voltage-gated sodium channels . *The Journal of Physiology* .
- Mian, A. A., Senadheera, P., & Maathuis, F. J. (2011). Improving Crop Salt Tolerance: Anion and Cation Transporters as Genetic Engineering Targets. *Plant Stress* .
- Michalek, M., DSönnichsen, F., Wechselberger, W., Dingley, A. J., Hung, C.-W., Kopp, A., et al. (2012). Structure and function of a unique pore-forming protein from a pathogenic acanthamoeba. *Nature Chemical Biology* .
- Miseta, A., & Csutora, P. (2000). Molecular Biology and Evolution . *Relationship Between the Occurrence of Cysteine in Proteins and the Complexity of Organisms* .

- Miseta, A., & Custora, P. (2000). Relationship Between the Occurrence of Cysteine in Proteins and the Complexity of Organisms. *Molecular Biology and Evolution* .
- Munns, R. (2005). Genes and salt tolerance: bringing them together. *New Phytologist* .
- Nanjo, T., Kobayashi, M., Yoshiba, Y., Kakubari, Y., Yamaguchi-Shinozaki, K., & Shinozaki, K. (1999). Antisense supression of proline degradation improves tolerance to freezing and saiinity in Arabidopsis thaliana. *FEBS Lett* .
- Petersen, B., Lundegaard, C., & Petersen, T. N. (2010). 'NetTurnP – Neural Network Prediction of Beta-turns by Use of Evolutionary Information and Predicted Protein Sequence Features . *PLoS ONE* .
- Pravda, L., Berka, K., Vařeková, R. S., Sehnal, D., Banáš, P., Laskowsk, R. A., et al. (2014). Anatomy of enzyme channels. *BMC Bioinformatics* .
- Ramachandran , G., & Dokholyan, N. (2012). Homology Modeling: Generating Structural Models to Understand Protein Function and Mechanism . *Computational Modeling of Biological Systems Biological and Medical Physics, Biomedical Engineering* .
- Rodriguez-Rosaic, M. P., Galvez, F. J., & Huertas, R. (2009). Plant NHX cation/proton antiporter. *Plant Signaling and Behavior* .
- Roy, A., Kucukural, A., & Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction . *Nature protocols* .
- Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees . *Molecular Biology and Evolution* , 406-425 .
- Sansom, M. S., & Kerr, I. D. (1995). Transbilayer pores formed by beta-barrels: molecular modeling of pore structures and properties . *Biophysical Journal* .
- Sehnal, D., Vařeková, R. S., Berka, K., Pravda, L., Navrátilová, V., Ionescu, C. M., et al. (2013). Mole 2.0; Advanced approach for analysis of biomacromolecular channels. *J. Cheminform* .
- Senes, A., Ubarretxena-Belandia, I., & Engeman, D. M. (2001). The Cα—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* .

- Sibanda , B. L., Blundell, T. L., & Thorton, J. M. (1989). Conformation of beta-hairpins in protein structures. A systematic classification with applications to modelling by homology, electron density fitting and protein engineering. *J.Mol. Biol* .
- Suganya, P. R., Sudevan, K., Kalva, S., & Saleena, L. M. (2014). HOMOLOGY MODELING FOR HUMAN ADAM12 USING PRIME, I-TASSER AND EASYMODELLER. *International Journal of Pharmacy & Pharmaceutical Sciences* .
- Sunarpi, T., Motoda, J., & Kubo, M. (2005). Enhanced salt tolerance mediated by AtHKT1 transporterinduced Na+ unloading from xylem vessels to xylem parenchyma cells. *The Plant Journal* .
- Swarbreck, S. M., Colaço, R., & Davies, J. M. (2013). Plant Calcium-Permeable Channels. *Plant Physiology* .
- Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2015, January). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acid Res.*
- Trossat, C., Rathinasabapathi, B., Weretylnik, E. A., Shen, T. L., & Huang, Z. H. (1998). Salinity promotes accumulation of 3-dimethylsuifoniopropionate and its precursor S-methyimethionine in chloroplast. *Plant Physiol* .
- Wang, W. Q., LI, Y., Zhang, Y. Y., yang , C. P., Zheng, N. Y., & Xie, Q. (2007). Comparative expression analysis of three genes from the Arabidopsis vacuolar Na+ /H+ antiporter (AtNHX) family in relation to abiotic stresses. *Chinese Science Bulletin* .
- Waters, S., Gilliham, M., & Hrmova, M. (2013). Plant high-affinity potassium (HKT) transporters involved in salinity tolerance: Structural insights to probe differences in ion selectivity. *International Journal of Molecular Sciences* .
- Wimbley, W. C. (2003). The versatile β-barrel membrane protein . *The versatile β-barrel membrane protein* .

- Yamaguchi, T., Aharon, G. S., Sottosanto, J. B., & Blumwald, E. (n.d.). Vacuolar Na+/H+ antiporter cation selectivity is regulated by calmodulin from within the vacuole in a Ca2+ and pH dependent manner. *Proceedings of the National Academy of Sciences of the United States of America* .
- Yamaguchi, T., Apse, M. P., Shi, H., & 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* .
- Yeo, A. R. (1998). MoleculAr biology of salt tolerance in the context of whole-plant physiology.
- Yokoi, S., Quintero, F. J., & Cubero, B. (2002). Differential expression and function of Arabidopsis thaliana NHX Na+/H+ antiporters in the salt stress response. *The Plant Journal* .
- Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction . *BMC Bioinformatics* .
- Zheng, C., & Kurgan, L. (2008). Prediction of beta-turns at over 80% accuracy based on an ensemble of predicted secondary structures and multiple alignments. *BMC Bioinformatics* .