"*In silico* Structural Analysis, Physicochemical Characterization and Homology Modeling of transmembrane protein 43 (TMEM 43)"

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

Submitted by: Mir Sohayeb Rabbi

Student ID: 13336003

Biotechnology Program

Department of Mathematics and Natural Sciences

BRAC University

September 2017

Dedicated to

My parents and loved ones

Declaration

I hereby solemnly declare that the research work embodying the analysis and results reported in this thesis entitled **"*In silico* structural analysis, physicochemical characterization and homology modeling of transmembrane protein 43 (TMEM 43)"** submitted by the undersigned has been carried out under the supervision of **Ms. Eusra Mohammad**, Lecturer, Biotechnology Program, Department of Mathematics and Natural Sciences, BRAC University, Dhaka. It is further declared that the research work presented here is original and no part of this thesis has been submitted to any other institution for any degree or diploma.

Candidate:

_____

Mir Sohayeb Rabbi

Certified:

_____

Ms. Eusra Mohammad
Supervisor

Lecturer
Biotechnology Program
Department of Mathematics and Natural Sciences
Brac University, Dhaka

Acknowledgements

First and foremost, I wish to express my utmost gratitude to the Almighty for giving me the strength, determination and understanding that helped me to successfully complete this project.

I express my heartfelt appreciation and special thanks to **Professor A. A. Ziauddin Ahmad**, Chairperson of the Department of Mathematics and Natural Sciences, BRAC University and **Dr. Naiyyum Choudhury,** former coordinator of the Biotechnology and Microbiology Program of the Department of Mathematics and Natural Sciences, BRAC University for giving me their continued support and exemplary guidance during my tenure as a student at BRAC University.

I have taken various initiatives for this project. However, it would not have been possible without the kind support and help of my supervisor, Lecturer **Ms. Eusra Mohammad**, Department of Mathematics and Natural Sciences, BRAC University. I am highly indebted to my supervisor for believing in me and giving me the opportunity to work under her supervision. I owe my profound gratitude to her for taking keen interest in my project work and for guiding me all along till the completion of the project. I could not have imagined completing the project without her constant supervision and advice as well as her valuable insight regarding the technical aspects of this project.

I also thank the Department of Mathematics and Natural Science, BRAC University for providing me with the many facilities and opportunities throughout the entire period of my bachelor's degree.

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

Sincerely,
Mir Sohayeb Rabbi

Abstract

Arrhythmogenic right ventricular cardiomyopathy (ARVC) is a rare disease of the heart muscle. ARVC is an inherited condition, which means that it is passed on through families. It is caused by a change or mutation in one or more genes. ARVC can result from mutations in at least eight genes and each mutation is related to different types of ARVC. Many of these genes are involved in the function of desmosomes, which are structures that attach heart muscle cells to one another. Mutations in the genes responsible for ARVC often impair the normal function of desmosomes. Without normal desmosomes, cells of the myocardium detach from one another and die. One of the mutations responsible for a specific type of this disease is called ARVC type 5. A heterozygous missense mutation in the transmembrane protein 43 (TMEM 43) gene, (Ser358Leu) has been genetically identified to cause autosomal dominant ARVC type 5. Although TMEM 43 is an inner nuclear membrane protein, its presence at the intercalated discs has been confirmed. It is an integral membrane protein that spatially and functionally organizes protein complexes of the inner nuclear membrane and, therefore, has the potential to cause pathological changes of the nuclear envelope. However, the exact mechanism by which this protein carries out its function is yet unknown due to limited research on this area. A 3D structure of this protein is yet to be developed which would have facilitated understanding the function of this protein. Thus it is necessary to determine the structure of the protein encoded by TMEM 43 gene which could help to unravel how it interacts with other proteins at the inner nuclear membrane. The aim of this study is to predict the three-dimensional structure of transmembrane protein 43 via homology modeling and to examine its physicochemical properties using *in silico* approaches. Bio-computational analyzes of the target protein were performed using an array of online bioinformatics tools and databases and the homology model was developed using two different software programs (I-TASSER and SwissModel), whereby the best model was selected upon evaluation. In addition, the secondary structural motifs were identified within the model. This project provides an insight about the possible functions of the protein in maintaining nuclear envelope, given that further research is performed to prove its acceptance and efficiency.

Contents

Abstract iii

Chapter 1: Introduction 1

1.1 Membrane Proteins 2

Chapter 2: Transmembrane Proteins 4

2.1 transmembrane protein 43 (TMEM 43) 5

2.2 Properties of transmembrane Protein 43 7

2.3 Role of transmembrane Protein 43 9

2.4 Aim of the Project 11

Chapter 3: Benefits of Determining TMEM 43 Structure 12

3.1 Arrhythmogenic Right Ventricular Cardiomyopathy Type 5 (ARVC 5) 13

3.2 Significance of determining the structure of TMEM 43 18

3.3 Significance of homology modeling using in silico approaches 19

Chapter 4: Materials and Methods 20

4.1 Work Plan 21

4.2 Software tools and method used in each step to analyze the protein 22

4.2.1 Protein Sequences 22

4.2.2 Homology 22

4.2.2.1 Blast 23

4.2.2.2 Clustal Omega 24

4.2.2.3 BoxShade 25

4.2.3 Phylogenetic 25

4.2.3.1 Phylogeny.fr 26

4.2.4 Amino Acid Composition 27

4.2.4.1 Pepstats 27

4.2.5 Protein Characteristic Analysis 28

4.2.5.1 ProtParam 28

4.2.6 Prediction of Transmembrane Segments 29
4.2.6.1 Prediction via ProtScale 29
4.2.6.2 Prediction via TMHMM 30
4.2.7 Prediction of Molecular Structure 31
4.2.7.1 SOPMA 31
4.2.7.2 I-TASSER 32
4.2.7.3 SwissModel 33
4.2.8 Model Validation 34
4.2.8.1 PROCHECK 34

Chapter 5: Results 36
5.1 Protein Analysis 37
5.1.1 Protein Sequences 37
5.1.2 Homology 37
5.1.2.1 Blast Result 37
5.1.2.2 MSA using Clustal Omega and BoxShade 42
5.1.3 Phylogenetic 44
5.1.4 Amino Acid Composition 45
5.1.5 Protein Characteristic Analysis 47
5.1.6 Prediction of Transmembrane Segments 49
5.1.6.1 Prediction via ProtScale 49
5.1.6.2 Prediction via TMHMM 51
5.1.7 Prediction of Molecular Structure 53
5.1.7.1 Results from SOPMA 53
5.1.7.2 Results from SwissModel 55
5.1.7.3 Results from I-TASSER 58
5.1.8 Model Validation 66
5.1.8.1 Selection of the best model between the 5 I-TASSER models 66
5.1.8.2 Selection of final the model 71

Chapter 6: Discussion 74
6.1 Discussion 75

Chapter 7: Conclusion 78

7.1 Conclusion 79

Bibliography 80

CHAPTER 1:

Introduction

1.1 Membrane Proteins

Phospholipid bilayers along with membrane proteins makeup the biological membranes vital for life (Müller, Wu, & Palczewski, 2008). Therefore, it is not really astounding that around one-third of all proteins synthesized in eukaryotes turn out to be a part of these structures (Müller et al., 2008). Proteins of the biological membranes are insoluble in water. Thus they are located in phospholipid bilayers such as cell membranes and membrane bound organelles (Müller et al., 2008). The role of these proteins are vital as they act as carriers of small molecules such as nutrients, electrolytes and important metabolic cofactors as well as taking part in many signaling pathways (Elofsson & von Heijne, 2007). Furthermore, Müller et al. (2008) described their functions are to enable cells to interact with each other and perceive changes in their surroundings, act as binding sites for ligands, participate in catalytic activity and most importantly provide structural stability enabling the cells to retain their shape and size.

Membrane proteins take up almost 50% of the volume of membranes with around 30% of the genes encoding them (Müller et al., 2008). But these approximations are not totally reliable (Ahram, Litou, Fang, & Al-Tawallbeh, 2006; Elofsson & von Heijne, 2007; Remm & Sonnhammer, 2000). Nevertheless, this uncertainty does not undermine the fact that membrane proteins are encoded by a vast number of genes (Müller et al., 2008). Many new proteins are being discovered but due to inadequate information about their atomic structures, their molecular functions cannot be fully understood (Ramachandran & Dokholyan, 2012). At the present time, membrane proteins account for less than 1% of the known structures available in the Protein Data Bank (Berman, 2000). Establishing the structure of these proteins is decisive in comprehending their functions (Müller et al., 2008).

Membrane proteins are basically two types: integral membrane proteins and peripheral membrane proteins (Figure 1.1). Integral membrane proteins, also known as transmembrane proteins, are permanently attached to the membrane whereas peripheral membrane proteins are temporarily anchored to the lipid bilayer. Until now, only two structural elements have been described that comprise the integral membrane proteins: β-barrels and α-helices. These structural components enhance stability of the protein by

hydrogen bonding and eliminate any water molecule inside the membrane proteins (Müller et al., 2008).

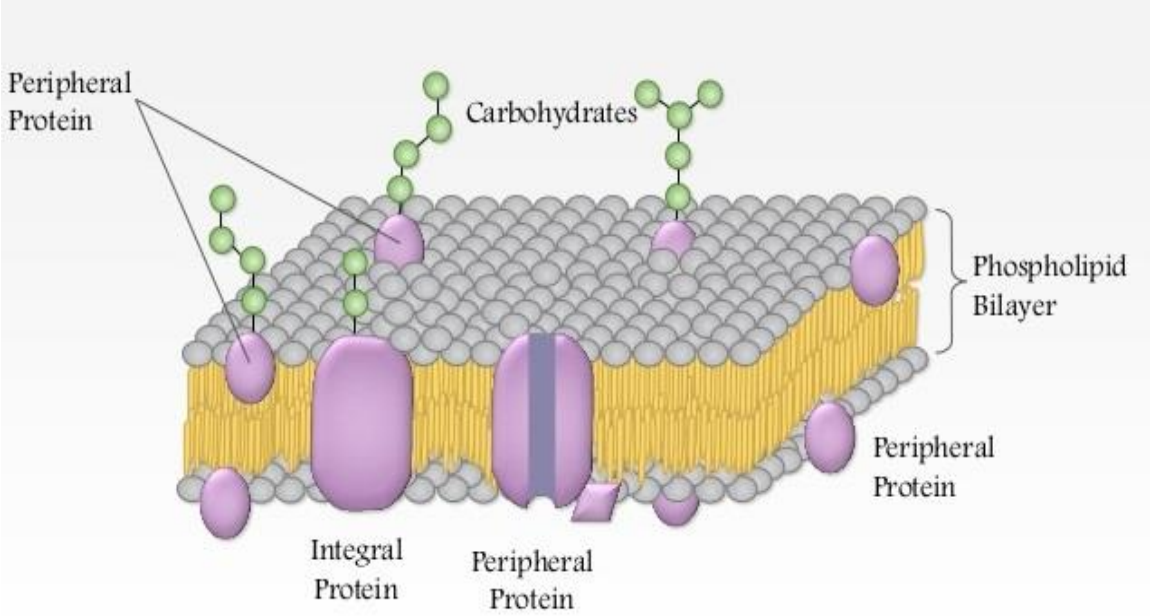


Figure 1.1: Integral and peripheral membrane protein

CHAPTER 2:

transmembrane protein 43 (TMEM 43)

2.1 transmembrane protein 43 (TMEM 43)

In eukaryotic cells, the nucleus is bordered by a double layered membrane called the nuclear envelope (NE) with the outer nuclear membrane (ONM) connected to the rough endoplasmic reticulum and an inner nuclear membrane (INM) facing the nucleoplasm (Dreger, Bengtsson, Schoneberg, Otto, & Hucho, 2001). The INM is the site where most of the transmembrane proteins are harbored and approximately 80 INM proteins have known structure (Bengtsson & Otto, 2007; Schirmer, Florens, Guan, Yates, & Gerace, 2003; Schirmer & Foisner, 2007). Each part of the nuclear envelope vary from each other in terms of their protein components. At the inner nuclear membrane, these transmembrane proteins are bonded to chromatin and lamina (Gruenbaum, Margalit, Goldman, Shumaker, & Wilson, 2005). Among them, a fairly novel integral membrane protein of the INM is the transmembrane protein 43 (TMEM 43).

TMEM 43, also known as protein LUMA, is a presumed membrane protein whose structure and function is yet to be determined (Siragam et al., 2014). According to Bengtsson and Otto (2007) LUMA is present in vertebrates, insects, plants, many unicellular eukaryotes and even in few bacteria, although how bacteria derived this protein is still relatively unclear. Thus it is widely distributed among species. Bengtsson and Otto (2007) has described TMEM 43 to be a unique transmembrane protein that has been greatly conserved over many evolutions and has stressed that no other studied integral membrane protein is that greatly conserved. They also stated that human LUMA is 93% similar to murine LUMA. Although this protein is mostly present at the inner nuclear membrane, it is actually a protein of the endoplasmic reticulum (ER), thus most of its sequence is present in ER lumen (Bengtsson & Otto, 2007). Dreger et al. (2001) visualized the inner nuclear membrane after performing immunofluorescence staining and found that LUMA localizes at the INM. Furthermore, Bengtsson and Otto (2007) ran their own tests with HeLa cells and NIH 3T3 cells and confirmed the presence of LUMA along with emerin and nuclear lamins (nuclear intermediate filament proteins) (Figure 2.1). The portion of LUMA found elsewhere other than the nuclear membrane is presumed to exist at the non-nuclear domains of the endoplasmic reticulum (Bengtsson & Otto, 2007). TMEM 43 protein is 400 amino acids long and is predicted to be a crucial component of

the nuclear envelope as mutations in the protein have proved to be fatal in some cases (Siragam et al., 2014).

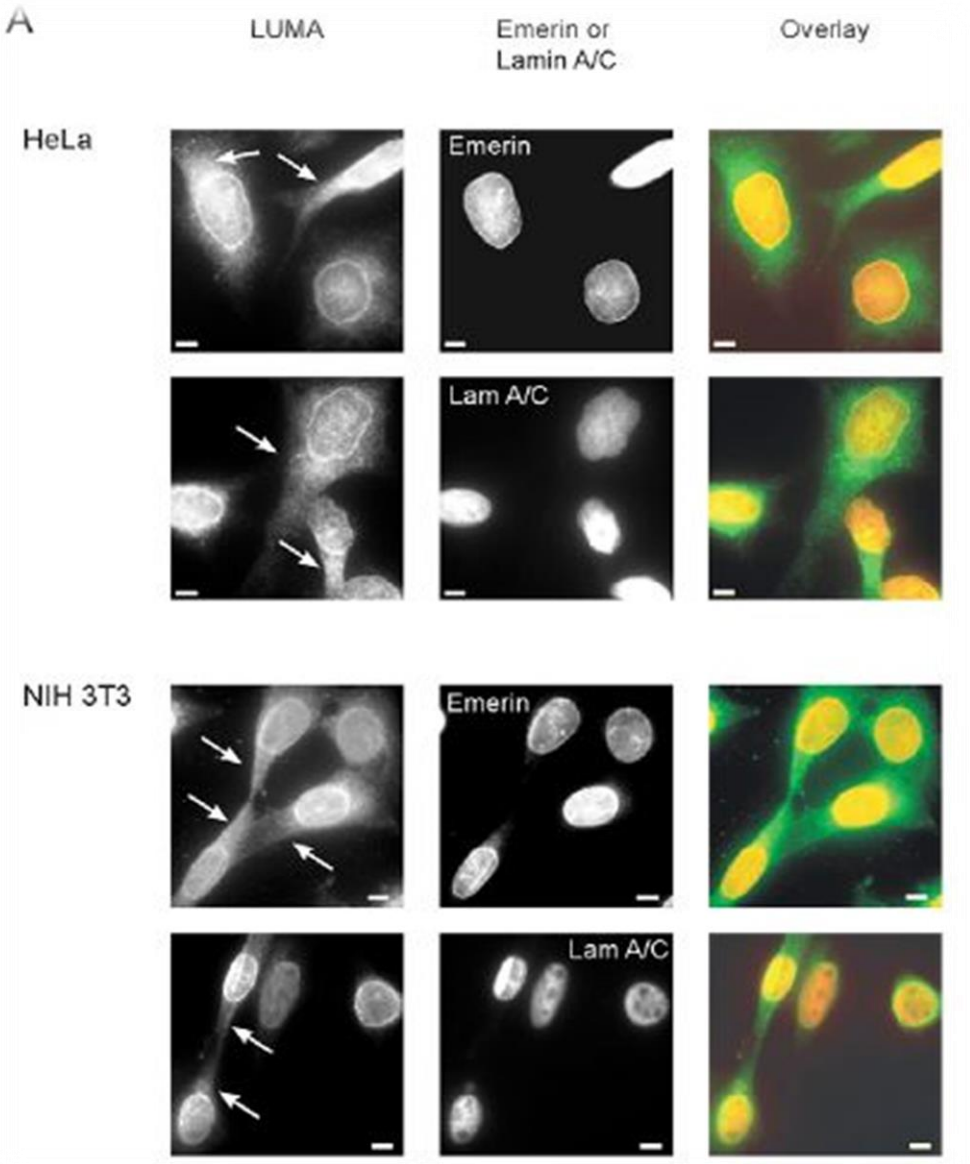


Figure 2.1: Existence of LUMA along with emerin or lamin A/C in HeLa and NIH 3T3 cells, which were subjected to immunofluorescence staining. They are localized at the nuclear envelope. The arrows show the non-nuclear portion of LUMA at the endoplasmic reticulum. In the overlay images, LUMA is seen in green color and emerin or lamin A/C in red. (Source: Bengtsson & Otto, 2007)

2.2 Properties of transmembrane Protein 43

Despite being a novel protein, Bengtsson & Otto (2007) deduced some of the properties of the protein LUMA. They performed numerous in vitro analytical procedures to confirm that LUMA is actually a typical inner nuclear membrane protein. According to Östlund, Sullivan, Stewart, & Worman (2006) some inner nuclear membrane proteins are misplaced at the non-nuclear regions of the endoplasmic reticulum if lamins are absent. Bengtsson & Otto (2007) performed their own tests to see if that was also true for LUMA by using mouse embryonic fibroblasts (MEFs) missing A-type lamins in the presence of antibody specific to LUMA. The results were positive as they did not detect presence of LUMA at the nuclear envelope without A-type lamins. Thus it was confirmed that the retention of LUMA at the inner nuclear rim is dependent on A-type lamins (Figure 2.2).

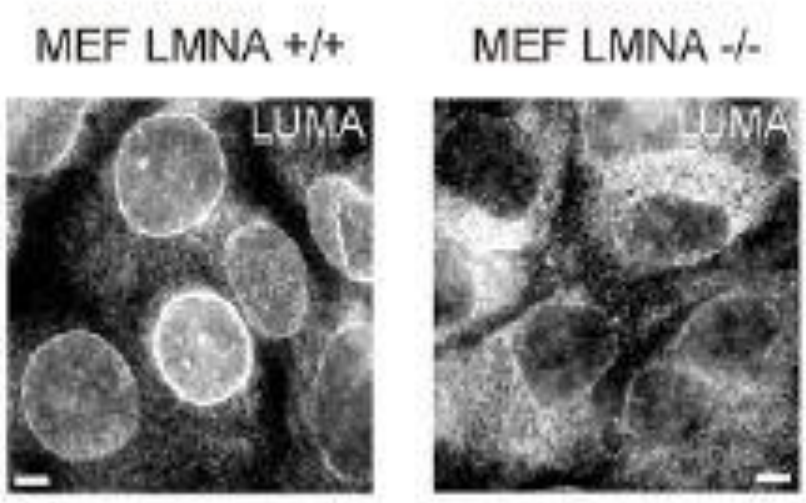


Figure 2.2: Images of normal cells containing A-type lamins (MEF LMNA+/+) and cells lacking A-type lamins (MEF LMNA–/–) after they were stained with anti-LUMA antibody. It shows A-type lamins is required to keep LUMA at the nuclear envelope. (Source: Bengtsson & Otto, 2007)

To study these transmembrane proteins it is important that they are extracted easily without affecting their structure and most of them can be extracted by using detergents. However, inner nuclear membrane proteins and lamins cannot be extracted in detergents at low ionic strength and requires a concentration of at 1 M or 2 M $NaCl$ (Foisner & Gerace, 1993). Bengtsson & Otto (2007) compared extraction of LUMA by treating isolated nuclear envelopes with Triton X-100 at a concentration of 0.1M and 1M$NaCl$. It revealed that the fraction of LUMA extracted from isolated nuclei was very low in 0.1M $NaCl$ compared to the amount extracted in 1M $NaCl$ which supported that LUMA is no different than other inner nuclear membrane proteins. From this result they concluded that there is a strong interaction between LUMA and nuclear lamins. Furthermore, they also added that the protein has the ability to self-oligormerize through its predicted transmembrane domains.

It is still quite unknown how the protein is transported to the nuclear envelope. Usually proteins of the endoplasmic reticulum get transported to the inner nuclear membrane due to the interactions between these proteins and components of the nucleus (Ellenberg et al., 1997; Lusk, Blobel, & King, 2007; Ohba, Schirmer, Nishimoto, & Gerace, 2004). According to Bengtsson & Otto (2007) LUMA is anchored to the nuclear envelope as a result of more than a few interactions with 'anchor proteins' and due to their property of self-oligomerization.

To understand the interaction of LUMA with emerin, Bengtsson & Otto (2007) performed immunoprecipitation of LUMA with emerin, lamin A/C and a few other proteins of the membrane from homogenous C2C12 cells. The ideology behind the experiment was that if they interact with each other they will immunoprecipitate together forming a pellet. After completion of their experiment they found LUMA co-immunoprecipitated with both lamin A/C and lamin B2 with emerin present in both cases (Figure 2.3). Thus it was confirmed that there is an interaction of LUMA with both emerin and lamins (both A and B type) and that the interaction is specific.

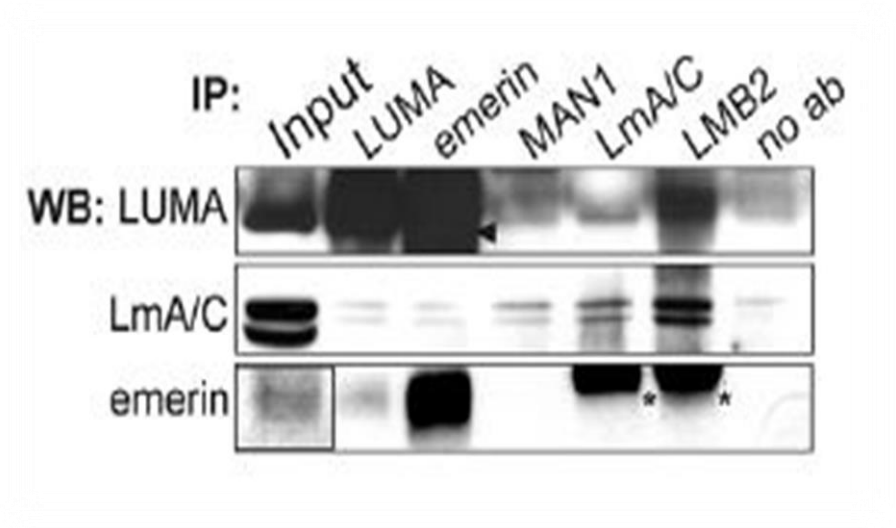


Figure 2.3: Image of the result of immunoprecipitation of the different proteins from C2C12 cells in the presence of antibodies specific to the proteins. LUMA immunoprecipitated with emerin as well as with lamin A/C and lamin B2. (Source: Bengtsson & Otto, 2007)

2.3 Role of transmembrane Protein 43

Like TMEM 43, another conserved LEM-domain containing protein, emerin, is also an integral membrane protein that largely localizes at the inner nuclear membrane of the nuclear envelope. It is an important component of the nuclear lamina; required for the proper assembling of the nuclear envelope (Berk, Tifft, & Wilson, 2013). Emerin along with few other LEM-domain proteins interact with lamins and if any one of these proteins is not available, the other proteins fail to co-assemble causing failure in proper chromosome segregation in mitosis and nuclear assembly following mitosis (Berk et al., 2013). TMEM 43 is apparently needed to retain emerin at the inner nuclear membrane and can affect emerin distribution (Bengtsson & Otto, 2007). To examine the role of LUMA on retaining emerin, Bengtsson & Otto (2007) conducted experiments on HeLa cells by reducing the expression of LUMA with the help of GFP-tagged specific micro RNAs (miRNAs). The miRNA named Hmi48 was used to prevent the translation of the protein LUMA. They observed that the presence of emerin was considerably reduced meaning they were redistributed to some other places resulting in an altered nuclear

structure (Figure 2.4). Thus LUMA is essential to retain emerin at the inner nuclear membrane.

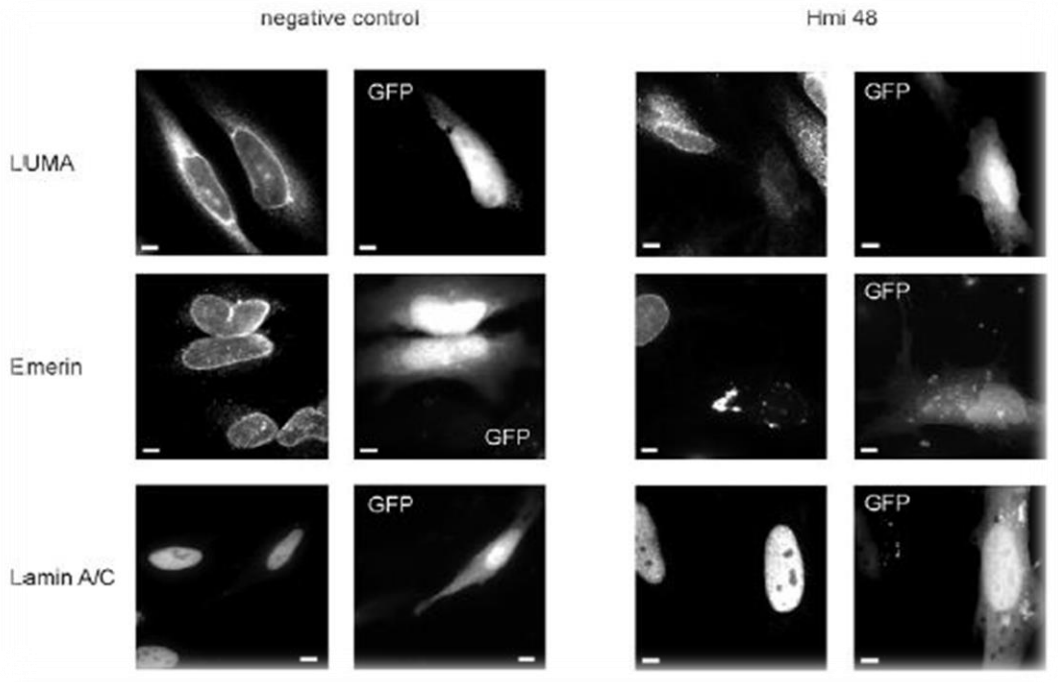


Figure 2.4: Images of the result observed after blocking expression of LUMA. Transfected HeLa cells with either a negative control for miRNA or the Hmi 48 containing vector designed to stop translation of LUMA mRNA were observed after 72 hours following transfection. For visualization, immunofluorescence staining was done and the GFP tagged transfected cells were easily detected. It was observed, cells containing the Hmi 48 failed to retain emerin. (Source: Bengtsson & Otto, 2007)

2.4 Aim of the Project

The objectives of this particular research are:

1. Identification and selection of homologous sequences in relation to the query protein sequence.
2. Comparison of the homologous sequences for conserved regions in other species.
3. Construction of a phylogenetic tree using the selected homologous sequences.
4. Analysis of physicochemical properties of the query protein using *in silico* methods.
5. Prediction of transmembrane regions, secondary structure of the protein in study using *in silico* methods.
6. Prediction of the three-dimensional (3D) structure of the query protein using homology modeling techniques.
7. Model validation of the 3D structures obtained and selection of the best model for the protein.

CHAPTER 3:

Benefits of Determining

TMEM 43 Structure

3.1 Arrhythmogenic Right Ventricular Cardiomyopathy Type 5 (ARVC-5)

Arrhythmogenic right ventricular cardiomyopathy (ARVC) is a disease of the myocardium which leads to deposition of adipose and fibrous tissue by replacing the cardiomyocytes mainly in the right ventricle which can cause ventricular arrhythmias and sudden cardiac arrests (Rajkumar, Sembrat, McDonough, Seidman, & Ahmad, 2012). An unusual form of arrhythmogenic right ventricular cardiomyopathy, ARVC type 5 is triggered by a missense mutation in the TMEM 43 genes located on chromosome 3p23 (Merner et al., 2008). It is a heterozygous mutation which they predicted to be deleterious (c.1073 C > T) initiating a p.Ser358Leu substitution in the final protein molecule. Other forms of ARVC such as (ARVC 7,8,9,10,12) are different from ARVC-5 because unlike ARVC-5 they are triggered due to mutations in genes synthesizing desmosomal protein (for example Desmoplakin, Junction Plakoglobin, Plakophilin) (Siragam et al., 2014).

Although the TMEM 43 is an endoplasmic protein localized mainly at the nuclear envelope, Franke et al. (2014) has confirmed that this protein is also found at the cardiac intercalated (IC) discs which explains why its mutation is associated with the above mentioned disease. According to Siragam et al. (2014), the effects of the mutation at the IC disc are still not clear due to limited research regarding this protein, but they conducted their own experiments and developed a hypothesis on how the mutation in TMEM 43 triggers ARVC-5. They created constructs of both normal and mutant human TMEM 43 gene marked by GFP which were then used to transfect HL-1 atrial cell line. This was done to evaluate the effect of the mutant protein at the IC discs and detect changes in their localization in affected cells.

It was observed that the transfected HL-1 cells containing the mutant form of the protein (TMEM43-S358l) aggregated into phagosome like structures (Figure 3.1 B and D) and had reduced expression at the nuclear envelope and endoplasmic reticulum compared to the cells transfected with the normal form of the protein (TMEM43-WT), (Figure 3.1 A and C) (Siragam et al., 2014). The formation of structures like phagosomes was not observed in the cells containing the wild type protein. It was described that, in cells containing TMEM43-S358L protein, phagosome like structures were formed due to the

property of self-oligomerization of the proteins which later attracted some endogenous TMEM 43 to the complex structure.

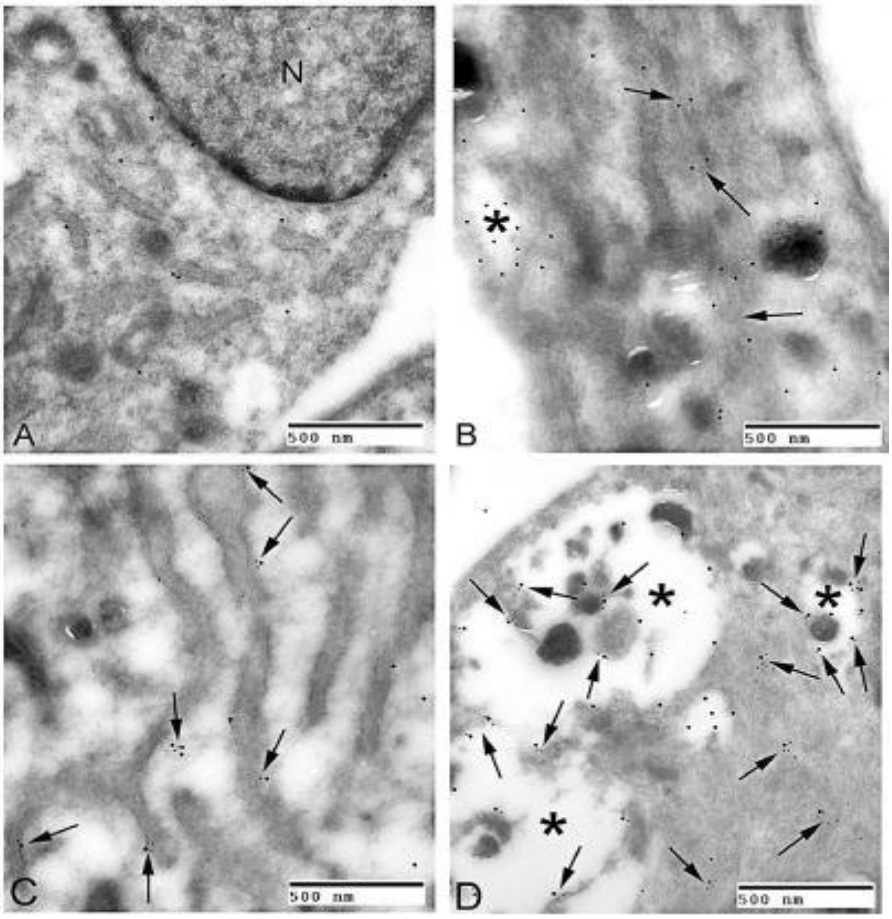


Figure 3.1: Images of HL-1 cells under electron microscope containing TMEM43 labeled with immunogold particles. (A & B) GFP labeled by using 15nm gold particles. (A) TMEM43-WT cells labeled with immunogold (large dot) was found in the nuclear envelope, endoplasmic reticulum (arrows). (B) TMEM43-S358L cells marked with immunogold particles (large dot) were concentrated into microfilaments in the cytoplasm (arrows) and unknown endosomal structures which were mostly vacuoles (asterisk). TMEM 43 was absent on nuclear envelope or endoplasmic reticulum. (C and D) GFP labeled with 15 nm gold particles and TMEM43 labeled with 5 nm gold particles. (C) Both TMEM 43 (small dots) and GFP (large dots) co-localized on the endoplasmic reticulum (arrows) in cells containing the wild type protein. (D) TMEM 43 (small dots) and GFP (large dots) were localized in large vacuolated structures (asterisk) and as cytoplasmic clusters (arrows) in cells containing mutant protein. (Source: Siragam et al., 2014).

Siragam et al. (2014) were also successful in explaining the effect at the IC discs of mutant TMEM 43. They found that in cells containing mutant TMEM 43, IC disc protein ZO-1 was markedly reduced compared to cells with wild type TMEM 43. The protein ZO-1 is important in maintaining proper exchange of signals at the junctions between the cardiac cells as they are responsible for retaining α-catenin and Junction Plakoglobin (JUP) at the cell-cell junction (Toyofuku et al., 2001). Thus, reduced ZO-1 levels also lead to loss of α-catenin JUP at the junctions. All these proteins are associated with electrical conductions at the cell-cell junctions of the IC discs. Furthermore, in TMEM43-S358L cells the proportion of phosphorylated to non-phosphorylated Cx43 (P2:P0 = 0.62) was also reduced in comparison to TMEM43-WT cells (P2: P0 = 0.97). The decrease in the level of phosphorylated Cx43 was also due to reduction of ZO-1 amounts. Phosphorylated state of Cx43 is more prevalent at the gap junction while non-phosphorylated Cx43 resides mainly inside cells. Presence of phosphorylated Cx43 is vital for proper functioning of the adherens junctions at the IC discs. The localization of TMEM43 at the intercalated disc along with ZO-1 and JUP is shown in figure 3.2.

These changes led to reduced permeability of the gap junctions at the intercalated discs with decreased intercellular communication at the gap junctions in TMEM43-S358L cells. To support this, Siragam et al. (2014) recorded conduction velocities and electrograms from both wild type and mutant protein containing cells (Figure 3.3 a & b). They observed that while the cells containing the normal protein have a rhythmic and regular beating pattern, the cells containing the mutant protein had a much delayed and asymmetrical rhythm. To quantify this, the conduction velocity decreased by 40% in TMEM43-S358L cells. Thus, the presence of mutant TMEM 43 contributes to decreased conduction at gap junctions between cells. Furthermore, the protein is thought to take part in an adipogenic pathway and contains receptor for PPARγ responsible for adipogenesis and any mutation in the protein can hamper the pathway which is predicted to be the reason of the deposition of fibrous tissue replacing the myocardium (Merner et al., 2008). Due to inadequate knowledge about the protein and insufficient researches regarding its function, how it causes adipogenesis is yet to be properly understood (Merner et al., 2008).

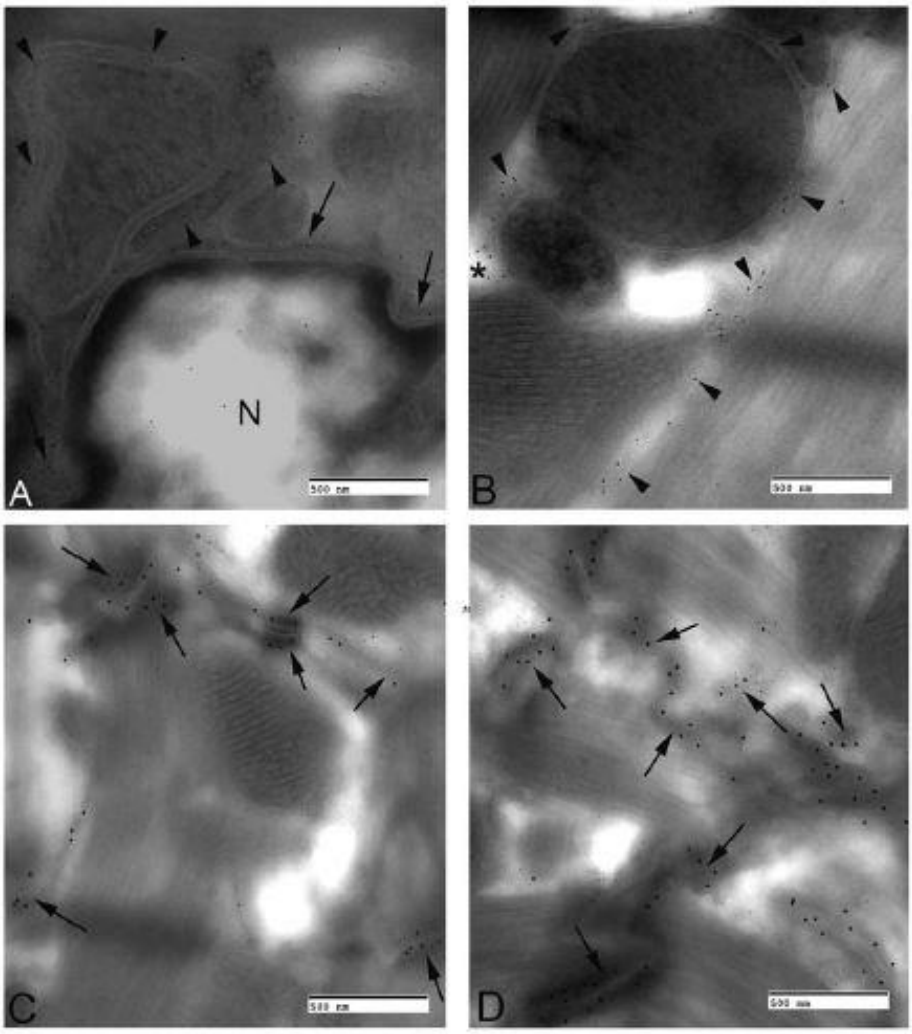


Figure 3.2: Localization of TMEM43 along with ZO-1 and JUP at the IC disc of normal murine cardiac tissue. (A) Image of a perinuclear region of the myocardium showing TMEM 43 labeled with immunogold (arrows), inside the nucleus (N) and at the sarcotubular network (arrowheads). (B) Image of the sarcoplasm of cardiac muscle cells. Sarcoplasmic reticulum represented by (arrowheads) on vesicles inside the sacrotubular system (asterisk). (C) Presence of TMEM43 along with ZO-1. The arrows specify gap junctions containing both TMEM43 and ZO-1. (D) Shows TMEM43 co-localizes with JUP in the gap junctions (arrows) at the intercalated disc. (Source: Siragam et al., 2014).

The life expectancy of the affected indivuals were reduced notably. Sudden cardiac death occurred in several cases (Merner et al., 2008). It was estimated that the average life expectancy of affected males is 41 years compared to 83 years in healthy males. Affected females, however, have a longer life expectancy of 71 years compared to 83 years in unaffected females (Merner et al., 2008). Thus affected males are at greater risk of dying due to heart failure than affected females.

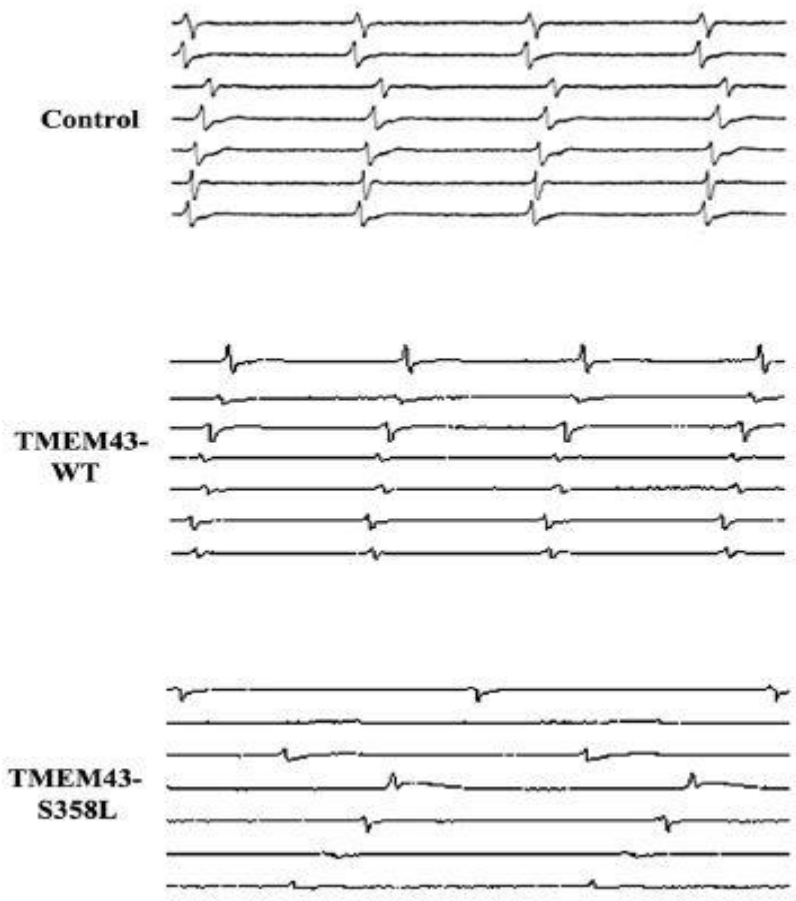


Figure 3.3a: Electrograms were documented using MEA electrodes. Both control and normal had a more regular and rhythmic beating while the mutant cells had slower and irregular beating. (Source: Siragam et al., 2014).

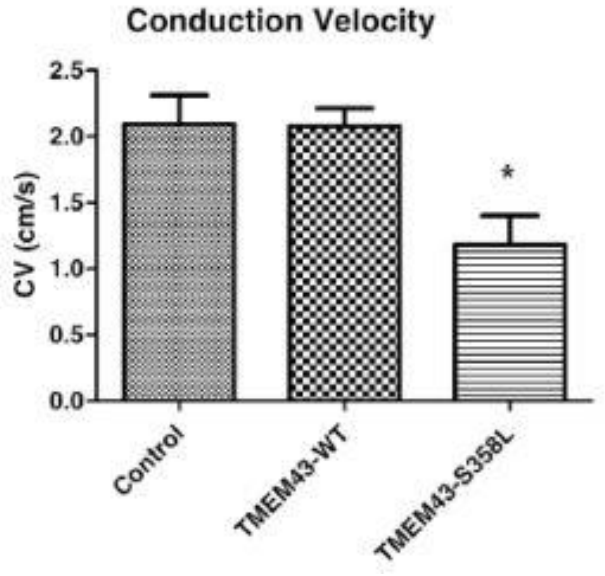


Figure 3.3b: Conduction velocity chart showing reduced conduction velocity in TMEM43-S358L cells compared to TMEM43-WT and control cells. (Source: Siragam et al., 2014).

3.2 Significance of determining the structure of TMEM 43

Taking into consideration the aforementioned, the importance of determining the structure of transmembrane protein 43 is quite clear. As it has been mentioned that it is still not clear how the protein works, determination of its three dimensional structure will allow us to decipher the regions that are responsible for its interaction with other inner nuclear membrane proteins and how it organizes other protein complexes to maintain the nuclear envelope. Furthermore, developing a molecular structure of the mutated protein and comparing it to the native protein structure would reveal the differences in the two structures. This would help understand the changes in the function of the mutated protein and how it leads to the dominant negative effect in ARVC-5. This study is only limited to determining the structure of the protein LUMA which is a crucial step in elucidating the role the protein in the nuclear envelope and the pathological mechanisms triggering ARVC-5.

3.3 Significance of homology modeling using *in silico* approaches

The determination of the molecular structure of this protein is quite challenging. The methods that are frequently used to determine protein structures are X-ray crystallography and NMR spectroscopy which are fairly accurate but are very costly (Schmidt & Lamzin, 2002). In addition, the purification process of some proteins also acts as limitations to these processes since it requires purified proteins at a very high concentration. Due to the challenge of maintaining the natural state of the protein following crystallization, computational approaches such as homology modeling are becoming increasingly popular in predicting protein structure (Aloy & Russell, 2006). Homology creates a molecular structure of a protein which is dependent on comparing homologous protein sequences where the unknown structure is deduced based on comparison with similar, known structure of other proteins (Šali & Blundell, 1993). It is also a way to determine the structure of proteins that are not suitable for prediction using either X-ray crystallography or NMR spectroscopy (Ramachandran & Dokholyan, 2012). Computationally created protein structures are analogous to the structures determined by NMR spectroscopy and they are important tools in reducing the gap between protein sequences and their structures.

Chapter 4:

Materials and Methods

4.1 Work Plan

The study was performed to analyze and predict a three-dimensional structure of the protein Luma using *in silico* techniques. The work plan was divided into a number of stages which started with obtaining the desired protein gene and ended with molecular structure determination. In each stage different software programs and databases were used to aid the analysis of the protein. The steps performed to obtain the desired result are demonstrated in a flowchart (Figure 4.1).

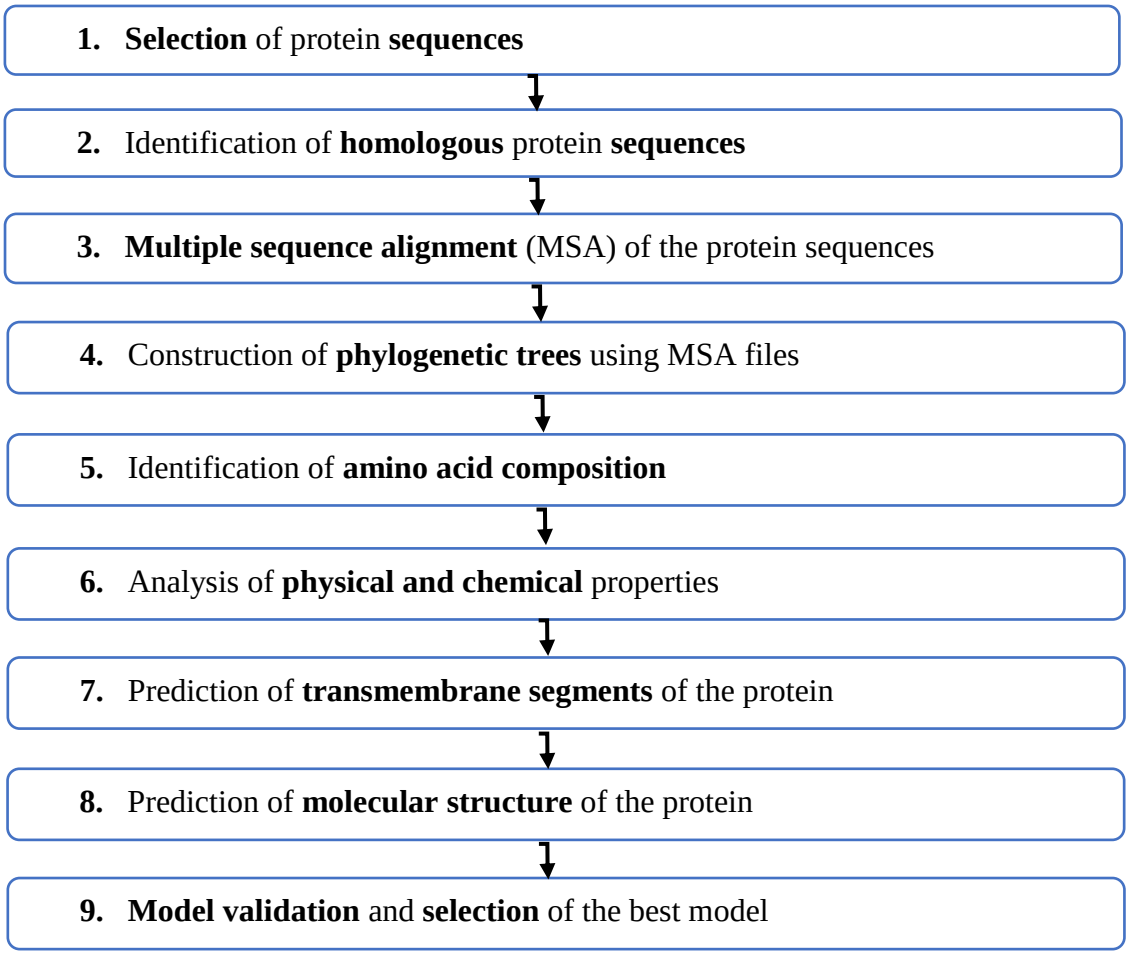


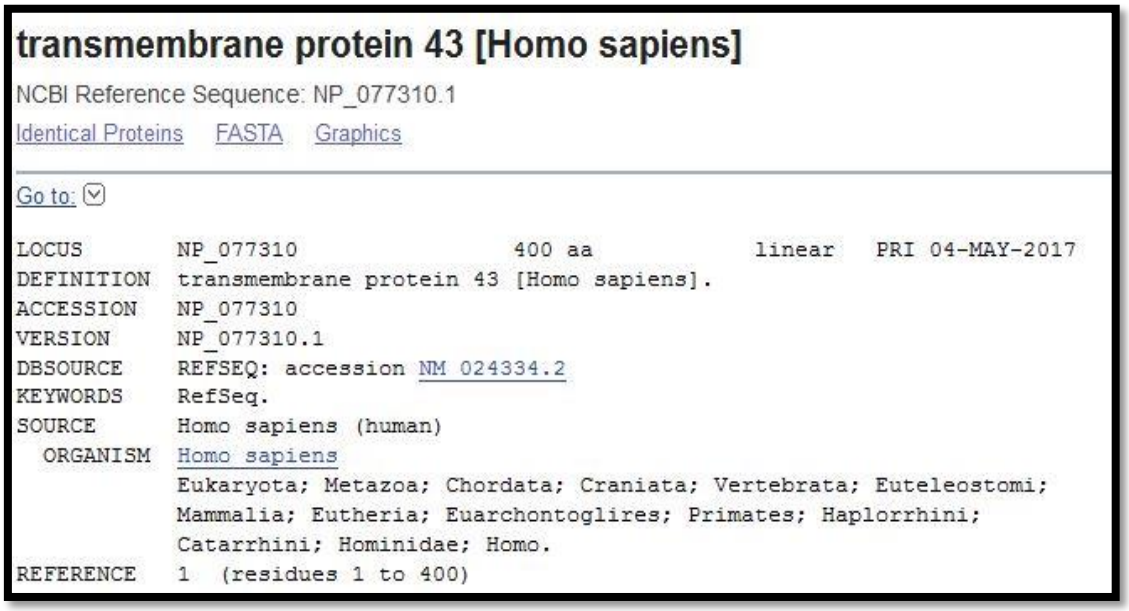
Figure 4.1: Experiment work plan for analysis of the desired protein leading to its structure and function determination

4.2 Software tools and method used in each step to analyze the protein

4.2.1 Protein Sequences

The target protein sequence was obtained from the National Center for Biotechnology Information (NCBI) in FASTA format. The retrieved protein sequence is shown below (Figure 4.2):

- Accession : NP_077310: transmembrane protein 43 [Homo sapiens]



transmembrane protein 43 [Homo sapiens]

NCBI Reference Sequence: NP_077310.1

Identical Proteins FASTA Graphics

Go to:

```
LOCUS       NP_077310               400 aa            linear   PRI 04-MAY-2017
DEFINITION  transmembrane protein 43 [Homo sapiens].
ACCESSION   NP_077310
VERSION     NP_077310.1
DBSOURCE    REFSEQ: accession NM_024334.2
KEYWORDS    RefSeq.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (residues 1 to 400)
```

Figure 4.2: GenPept profile for NP_077310

4.2.2 Homology

Homology is the condition of shared ancestry, that is, similarity in sequence. To identify if a specific sequence is similar to another sequence, the new sequence is compared with sequences already determined and stored in a database. This helps us to identify the protein family they belong to and the function they perform. The software tools used to find out the homology of different species with reference to protein LUMA were BLAST, Clustal Omega and BoxShade.

4.2.2.1 BLAST

The Basic Local Alignment Search Tool is specifically designed to search nucleotide and protein databases. It takes input the query sequence (DNA or protein) and searches either the DNA or protein databases for levels of similarity ranging from very low similarity to perfect matches. The database that is searched is a collection of numerous nucleotide and protein sequences which are accessible from websites such National Center for Biotechnology Information (NCBI). For each alignment reported, an Expect (E) Value is presented which show the statistical significance of the match. The findings are displayed in descending order in terms of significance, tables, graphics and alignments. In this study the blast P-suite (for protein blast) was used from the NCBI website (Figure 4.3).
URL Link: https://blast.ncbi.nlm.nih.gov/Blast.cgi

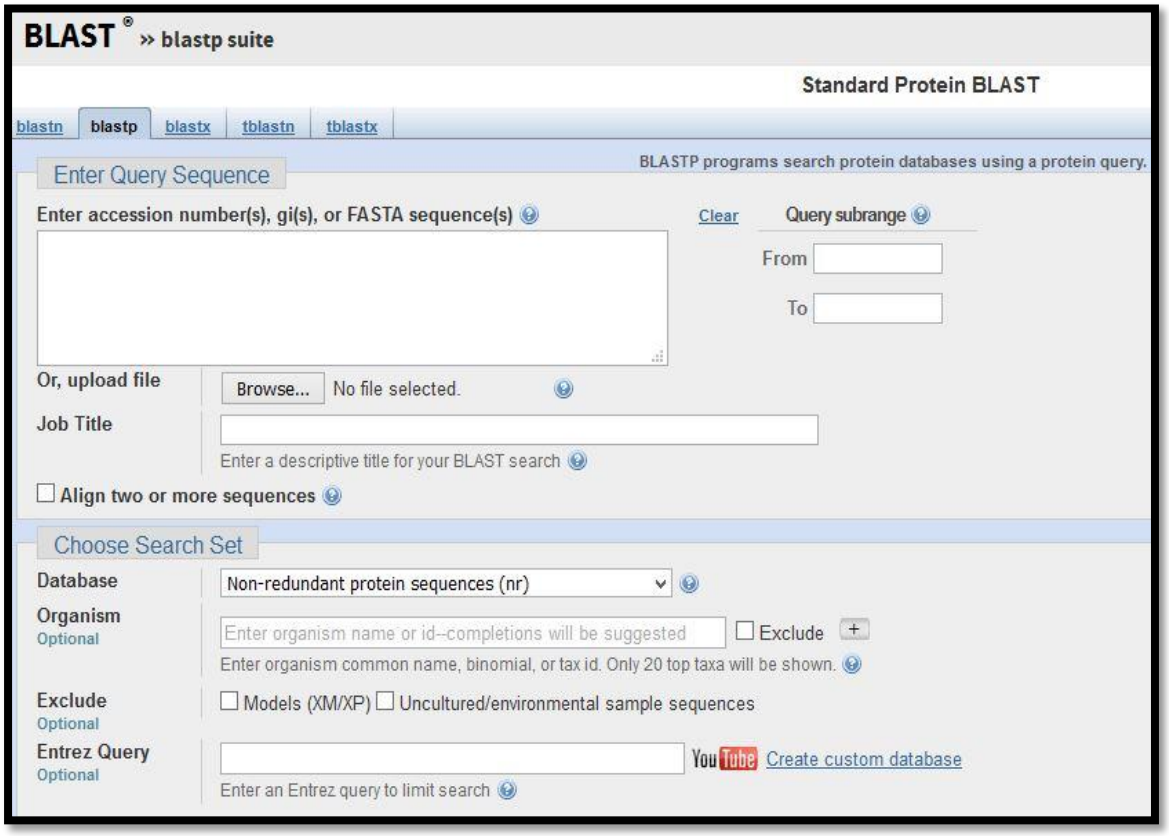


Figure 4.3: Blast P-suite Homepage

4.2.2.2 Clustal Omega

Clustal Omega is the current standard version of the Clustal series for multiple sequence alignment which can effectively align protein sequences rapidly and correctly. Clustal Omega uses a modified iterative progressive alignment method and can align over 10,000 sequences quickly and accurately. Clustal Omega is very useful for finding evidence of conserved function in DNA and protein sequences. The results from multiple sequence alignment can be used to construct a phylogenetic tree and predict protein structure. The program was run online where the protein sequences were input in the FASTA format (Figure 4.4).

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

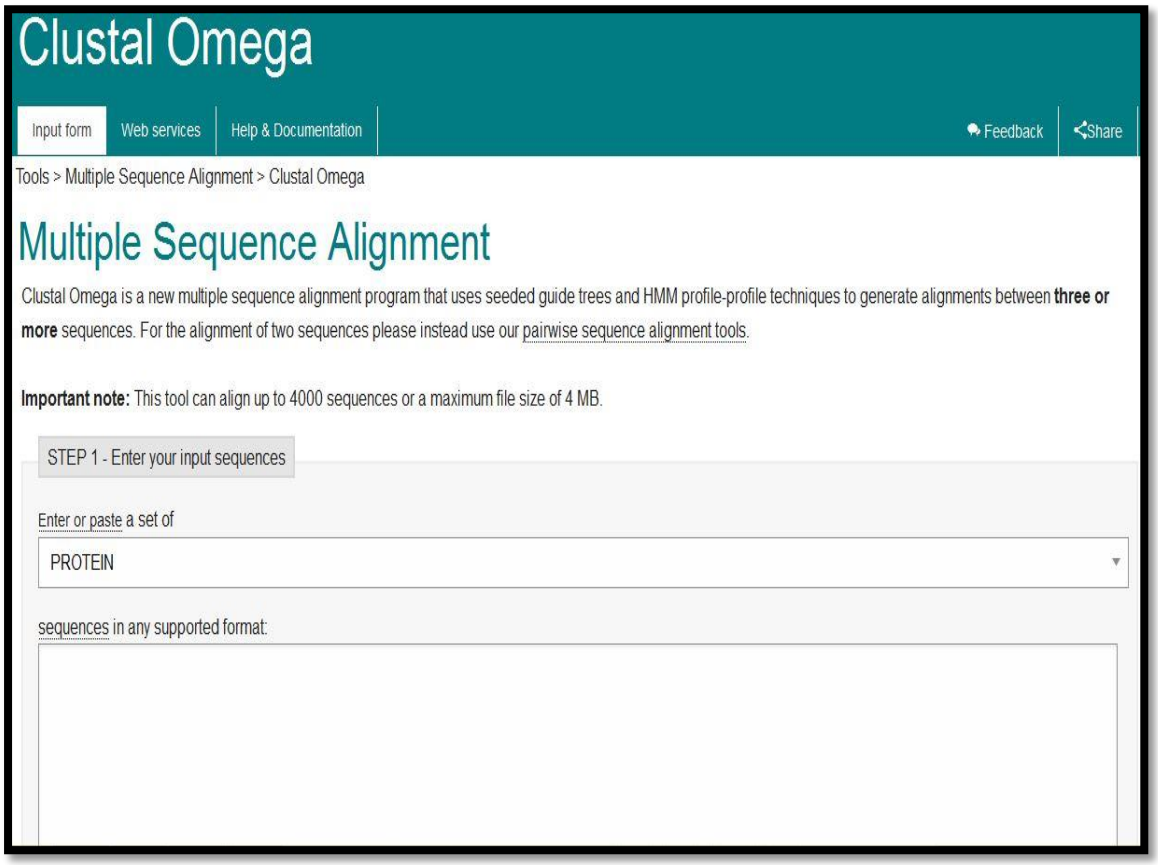


Figure 4.4: Clustal Omega Homepage

4.2.2.3 BoxShade

BoxShade is a program for pretty-printing multiple alignment output. The program is not responsible for carrying out alignments but for creating good-looking printouts from multiple aligned protein or DNA sequences. The outputs of Clustal Omega were used as the input for BoxShade. The online tool facilitated the complete visualization of the alignments generated by Clustal Omega. The output format was chosen as the RTF new. The BoxShade server was used from the bioinformatics resource portal ExPASy (Expert Protein Analysis System) (Figure 4.5).

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

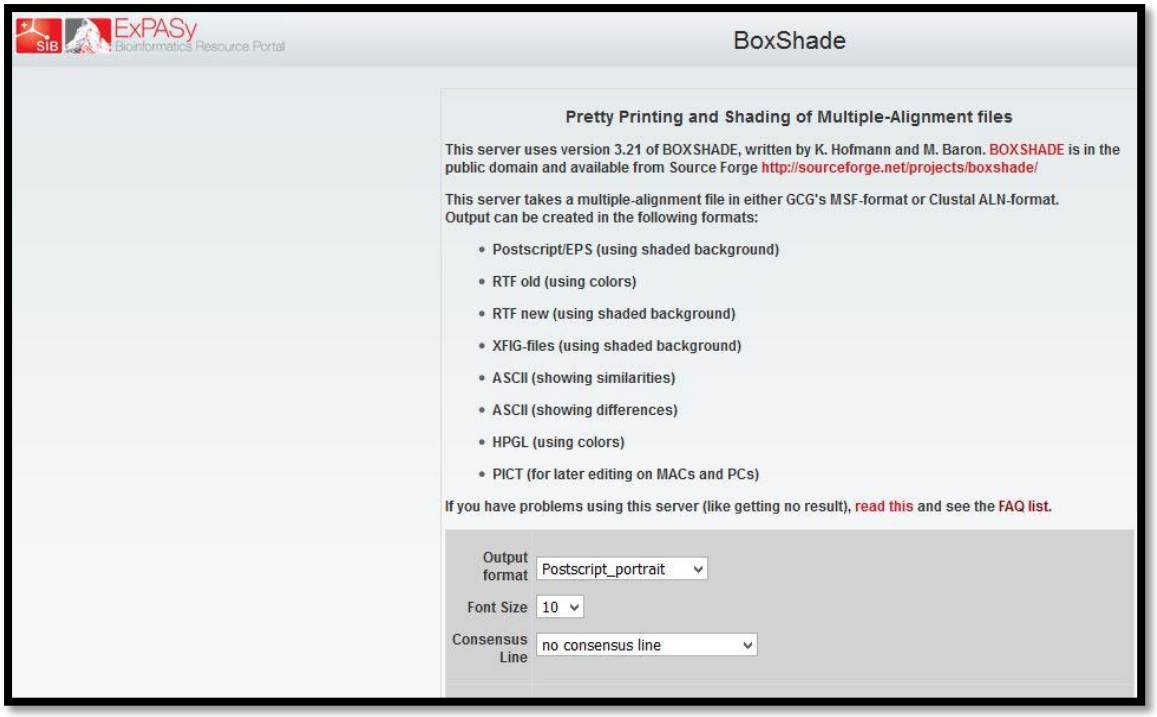


Figure 4.5: BoxShade Homepage

4.2.3 Phylogenetic

Phylogenetic, also known as evolutionary biology is the study of phylogeny. It focuses on the relationships among biological entities – organisms, species or genes in terms of evolutionary similarities and differences. Thus evolutionary relationships can be used to

classify organisms (taxonomy). In this study, the phylogenetic tree was constructed using Phylogeny.fr.

4.2.3.1 Phylogeny.fr

Phylogeny.fr is an online high performance platform that is intended to provide analysis of phylogenetic relationship between sequences by creating a phylogenetic tree using different bioinformatics programs. The software has three main modes by which a phylogenetic tree can be constructed. In this project, the 'One Click mode' was used to recreate a robust phylogenetic tree from the selected protein sequences. The protein sequences were copied and pasted in FASTA format and the process was started by clicking on the submit button. The parameters were set as default (Figure 4.6).
URL Link: http://www.phylogeny.fr/simple_phylogeny.cgi

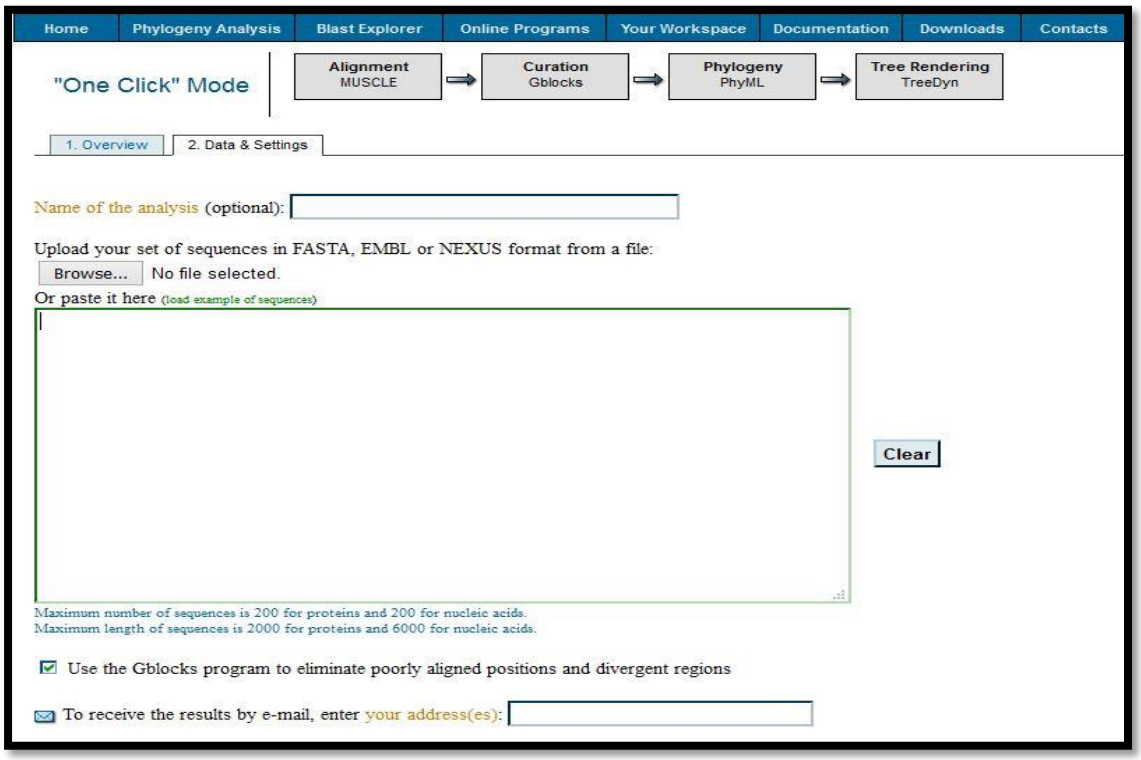


Figure 4.6: Phylogeny.fr Homepage

4.2.4 Amino Acid Composition

The amino acid composition of the protein of interest was determined by Pepstats.

4.2.4.1 Pepstats

Pepstats is one of the EMBOSS programs that is used to analyze protein sequences. It is an online analysis tool that is used to calculate the statistics of protein properties. The program reads one or more protein sequences and writes an output file with various statistics on the protein properties such as amino acid composition, molecular weight, number of residues, charge, isoelectric point, and etcetera. Pepstats was accessed from the EMBL-EBI website where the sequence of the protein in study was submitted in FASTA format (Figure 4.7).

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

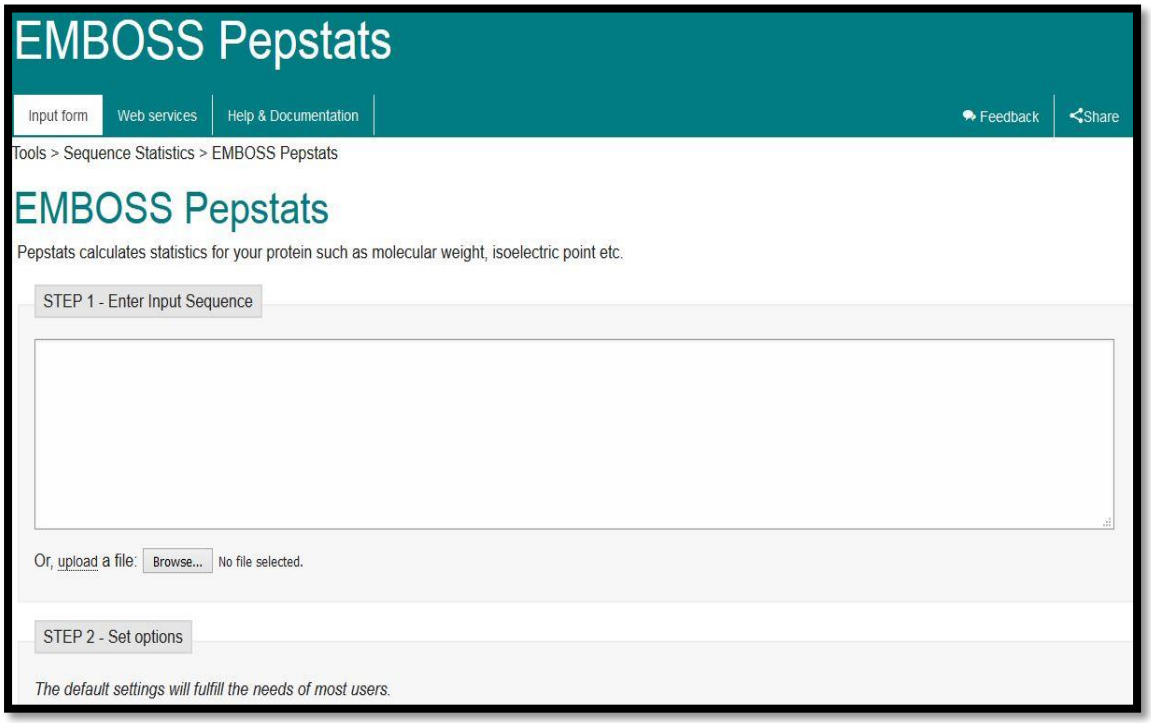


Figure 4.7: Pepstats Homepage in the EMBL-EBI website

4.2.5 Protein Characteristic Analysis

The characteristic of the protein in the study was analyzed with ProtParam.

4.2.5.1 ProtParam

This is an online tool capable of computing the various physical and chemical properties that can be inferred from a given protein sequence. The properties that it can compute of a given protein includes molecular weight, theoretical pI, amino acid composition, estimated half-life, instability index, aliphatic index and grand average of hydropathicity (GRAVY). The FASTA sequence of the query protein was used as input and the tool was accessed from the bioinformatics resource portal ExPASy (Expert Protein Analysis System) (Figure 4.8).
URL link: http://web.expasy.org/protparam/

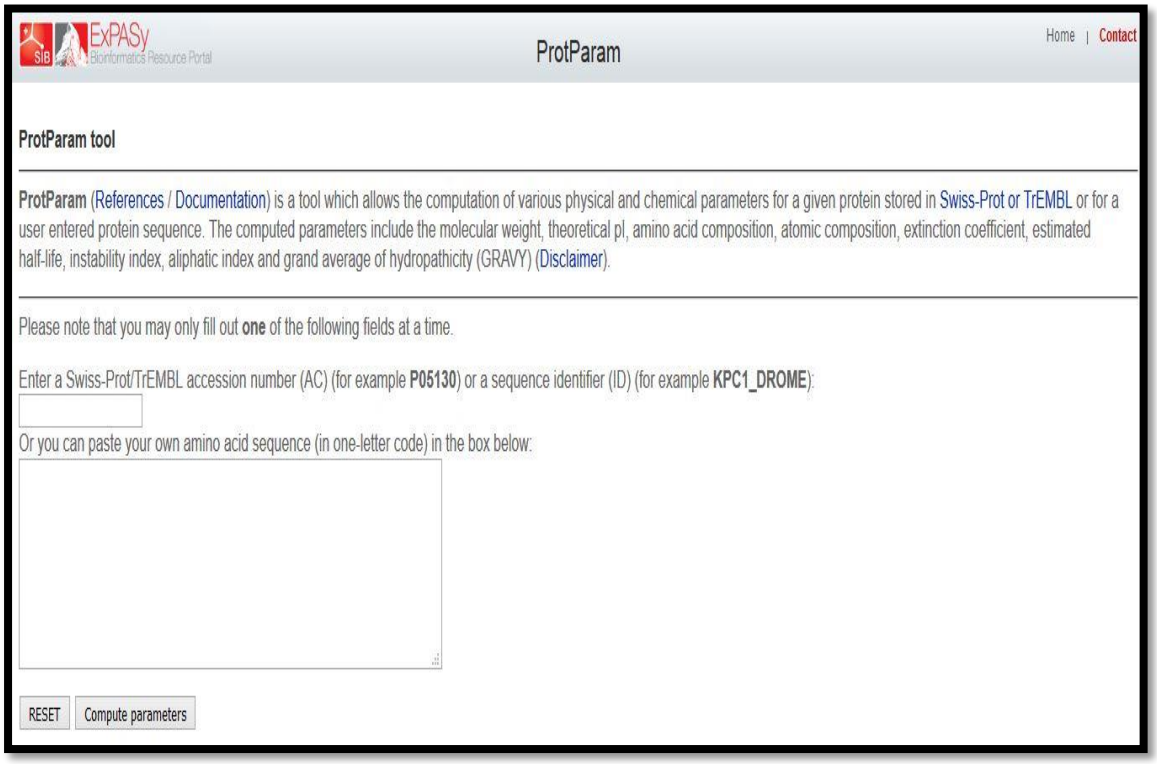


Figure 4.8: ProtParam Homepage in the ExPASy server

4.2.6 Prediction of Transmembrane Segments

4.2.6.1 Prediction via ProtScale

Predicting the fold of transmembrane proteins is potentially easier than water soluble proteins due to severely restricted way in which a protein can be embedded in the membrane. Using hydropathicity plots, transmembrane helices can be predicted by examining the hydrophobic and hydrophilic regions of that protein. Transmembrane helices are buried in the non-polar phase of the lipid membrane whilst other parts (loops) exist in more polar solution. Hydropathy plots are used to visualize hydrophobicity over the length of a peptide sequence (Kyte and Doolittle). The hydropathy scale is established on the hydrophobic and hydrophilic properties of the 20 amino acids used. To make a hydropathicity plot, a protein sequence and window size are chosen, where window size refers to the number of amino acids examined at a time to determine a point of hydrophobic character. The moving "window" determines the summed hydropathy at each point in the sequence (Y coordinate). These sums are then plotted against their respective positions (X coordinate). Such plots are useful in determining membrane spanning regions of membrane bound proteins.

ProtScale (Gasteiger et al., 2005) allows to compute and represent (in the form of a two-dimensional plot) the profile produced by any amino acid scale on a selected protein. An amino acid scale is defined by a numerical value assigned to each type of amino acid, of which hydrophobicity scales are most frequently used with the goal of predicting membrane-spanning segments that are highly hydrophobic, and secondary structure conformational parameter scales. To generate data for a plot, the protein sequence is scanned with a sliding window of a given size. At each position, the mean scale value of the amino acids within the window is calculated, and that value is plotted for the midpoint of the window.

In this study, the protein sequence was input in the FASTA format and the amino acid scale selected was Hphob /Kyte & Doolittle (Kyte and Doolittle, 1982) with the window size of 19 as detection of hydrophobic, membrane-spanning domains is best suited at this

window. The ProtScale tool is available at the bioinformatics resource portal ExPASy (Expert Protein Analysis System) (Figure 4.9).

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

Figure 4.9: ProtScale Homepage in the ExPASy server

4.2.6.2 Prediction via TMHMM

TMHMM is an online tool used to predict membrane protein topology based on a hidden Markov model and developed by Anders Krogh and Erik Sonnhamme. It predicts transmembrane helices and can discriminate between soluble and membrane proteins with high degree of accuracy. It can correctly predict 97-98 % of the transmembrane helices. The FASTA sequence of the query protein was used as input with all the parameters set to default. TMHMM Server version 2.0 was used (Figure 5.0).

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

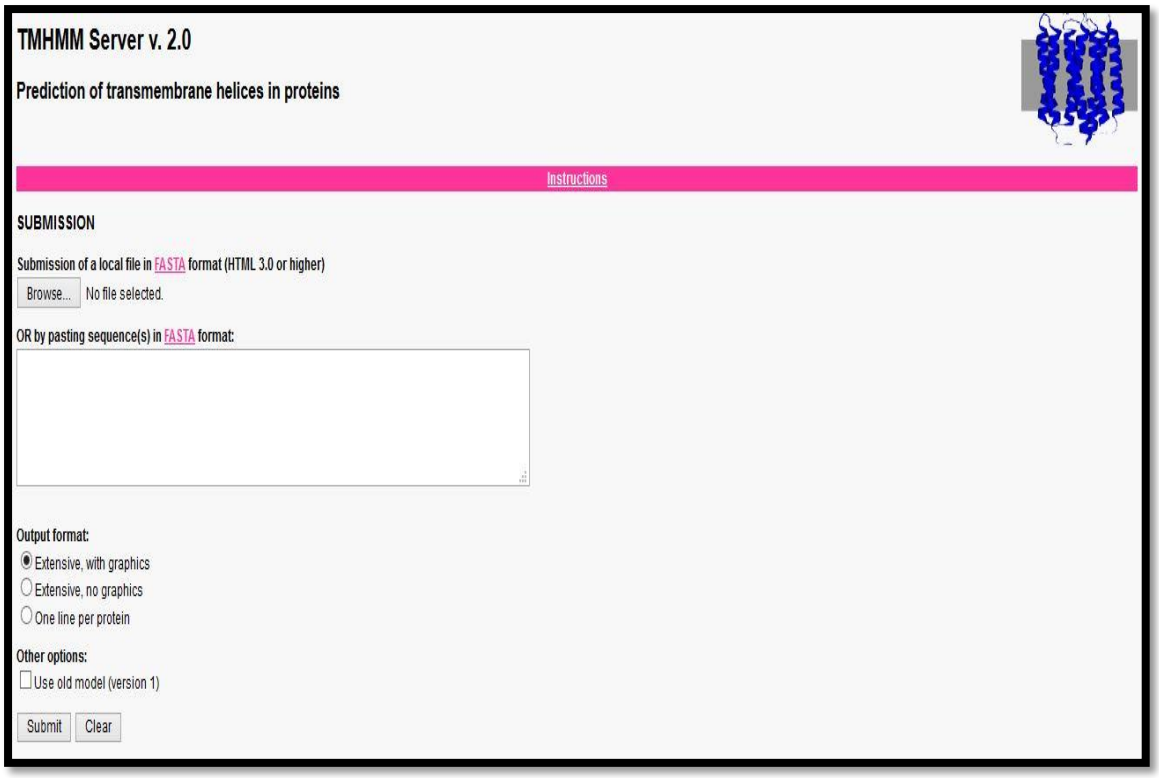


Figure 5.0: TMHMM Server version 2.0 Homepage

4.2.7 Prediction of Molecular Structure

A prediction of the three-dimensional structure of a protein can be done using the amino acid sequence (primary structure) of the respective protein. In other words, it uses the primary structure of the protein to predict its secondary structure along with its folding into the three dimensional tertiary structure. Before determining the 3D structure, the secondary structure of the protein was developed using SOPMA. The molecular structure of the transmembrane protein 43 was then conducted using I-TASSER and SwissModel.

4.2.7.1 SOPMA

Self-Optimized Prediction Method (SOPMA) is an online tool used to determine the secondary structure of proteins. It has a higher success rate when it comes to predicting the secondary structure of the proteins and can accurately predict 69.5% of amino acids for a three-state description of the secondary structure (α-helix, β-sheet and coil) in a whole database containing 126 chains of non-homologous (less than 25% identity)

proteins (Geourjon & Deléage, 1995). To obtain the secondary structure, the FASTA sequence of the protein was used as input with all the other parameters selected to default. The tool was accessed from the Network Protein Sequence Analysis (NPS@) server (Figure 5.1).

URL link:

https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_sopma.html

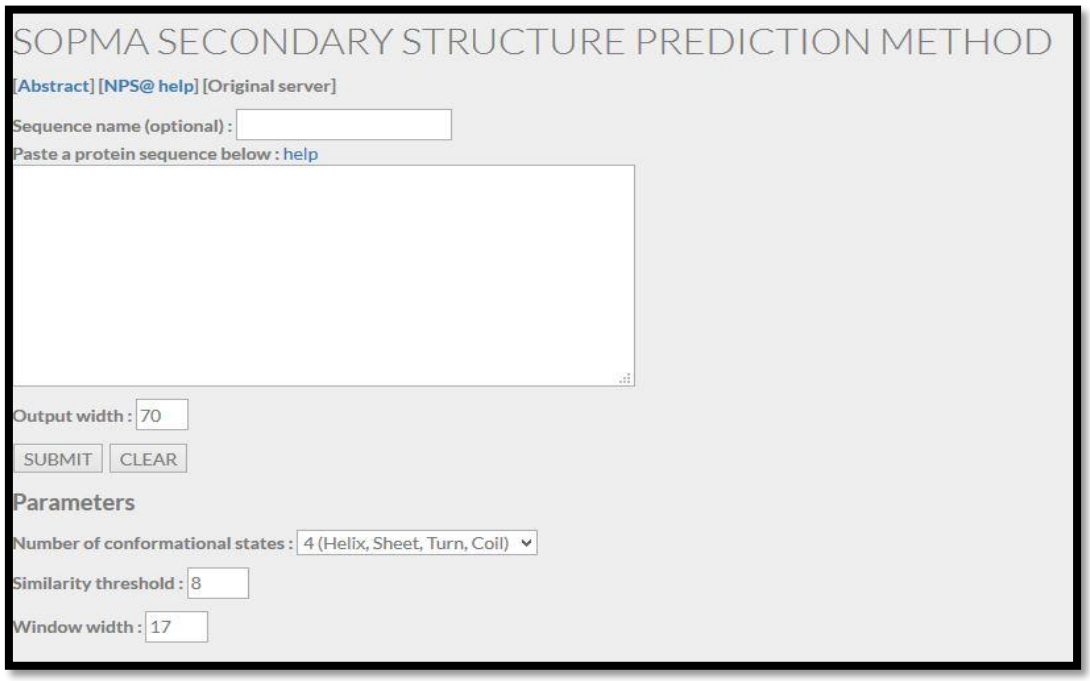


Figure 5.1: SOPMA Homepage

4.2.7.2 I-TASSER

I-TASSER (Iterative Threading ASSEmbly Refinement) is a combined platform used for protein structure and function prediction from amino acid sequences. At first, it generates full-length atomic models by many threading alignments and iterative template fragment assembly simulations and then infers the function of the protein by structurally matching the 3D models with other known proteins (Roy, Kucukural, & Zhang, 2010). This protocol provides new insights and guidelines for designing of on-line server systems for the state-of-the-art protein structure and function predictions. The server is developed for the most accurate structural and function predictions using state-of-the-art algorithms (Figure 5.2).

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

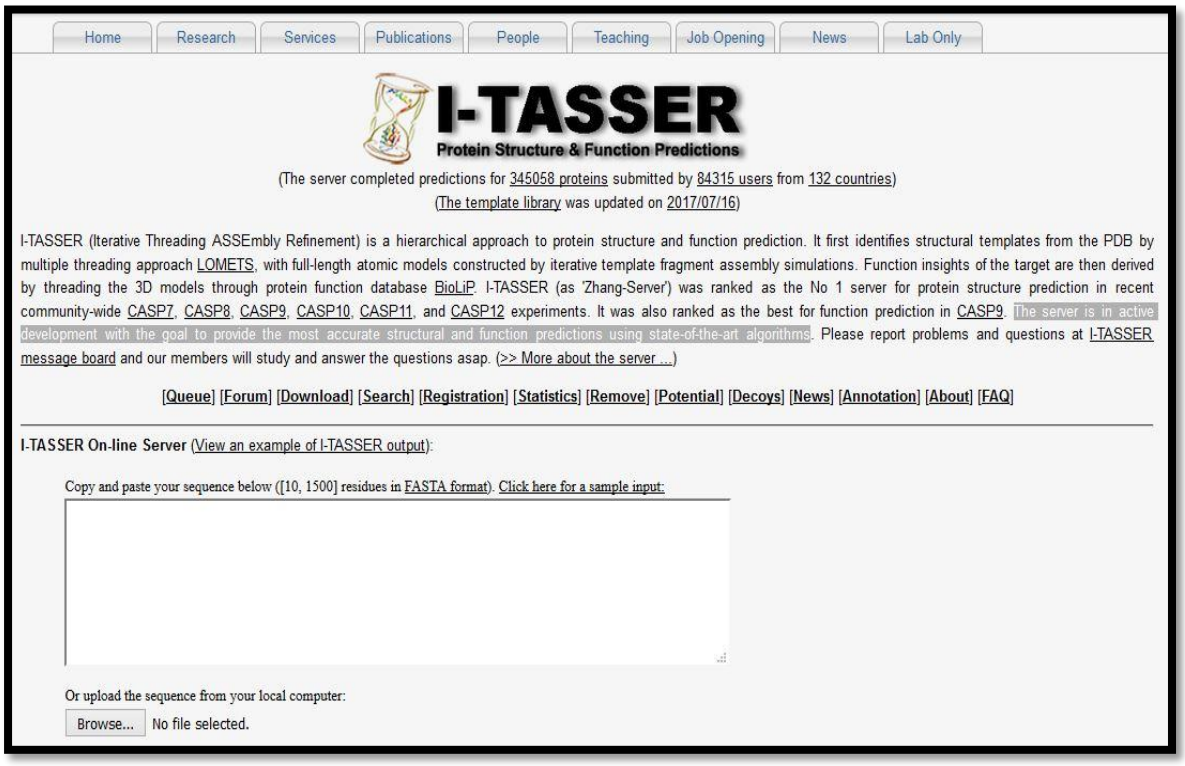


Figure 5.2: I-TASSER Homepage

4.2.7.3 SWISS-MODEL

SWISS-MODEL is a fully automated online server used for protein structure and homology modeling. It first identifies a structural template, aligns the target sequence with the template structure and develops a model which is then evaluated. It uses specialized software, updated protein sequence, structure database and each step can be repeated interactively until a satisfying modelling result is achieved (Arnold, Bordoli, Kopp, & Schwede, 2006; Biasini et al., 2014; Guex, Peitsch, & Schwede, 2009; Kiefer, Arnold, Künzli, Bordoli, & Schwede, 2009). The FASTA sequence was used as input from which a number of models were constructed. The tool was accessed from the bioinformatics resource portal ExPASy (Expert Protein Analysis System) (Figure 5.3).
URL link: https://swissmodel.expasy.org/

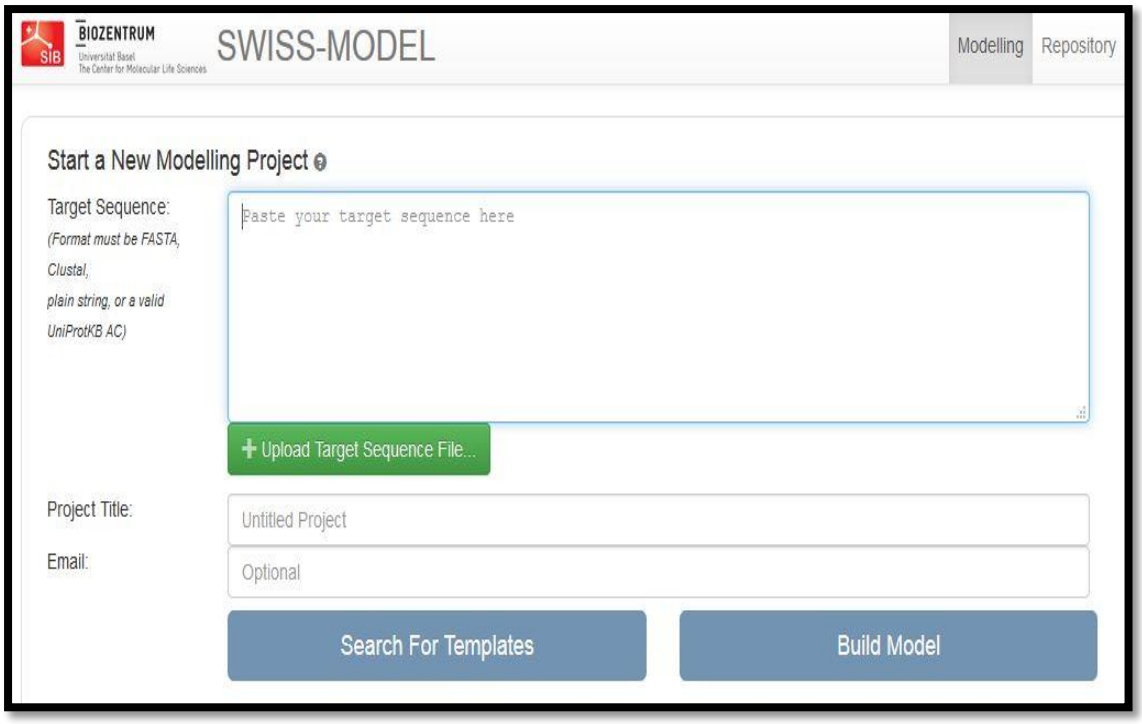


Figure 5.3: SWISS-MODEL Homepage

4.2.8 Model Validation

To evaluate the models provided by I-TASSER and SwissModel, PROCHECK was used for calculating Ramachandran plot calculations to validate the structures.

4.2.8.1 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 (Figure 5.4).
URL Link: http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/refs.html

Here, the PROCHECK web server available at the 'PDBsum Generate' section of the PDBsum server was used to evaluate the homology models of the query protein attained

from various homology modeling online tools and software as discussed in the previous sections (Figure 5.5)

URL Link: http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/Generate.html

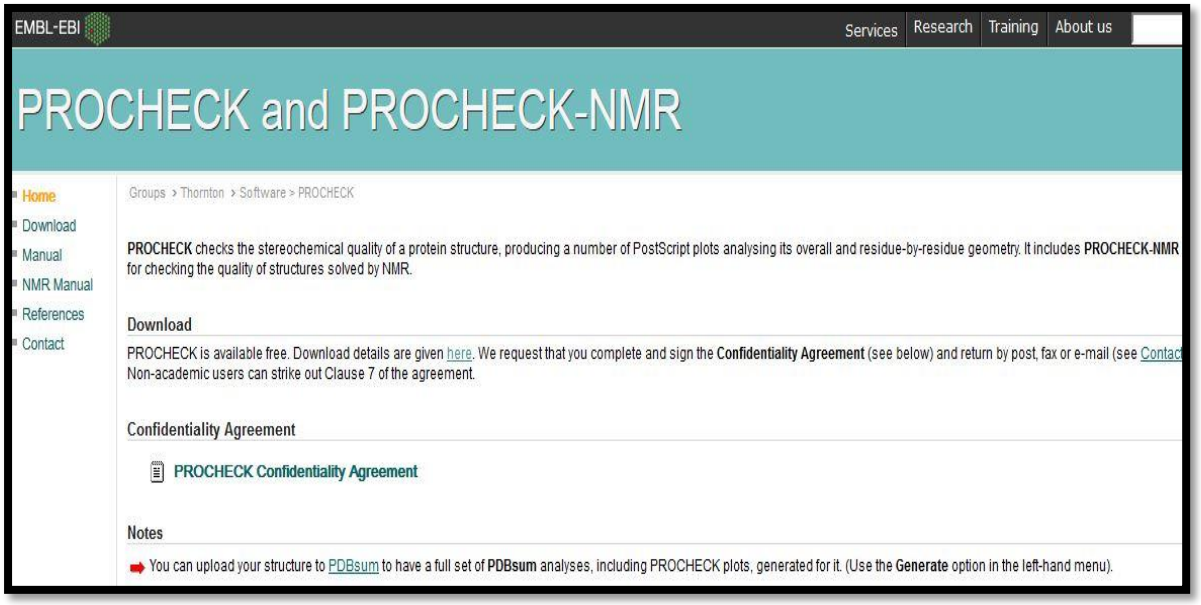


Figure 5.4: PROCHECK Homepage

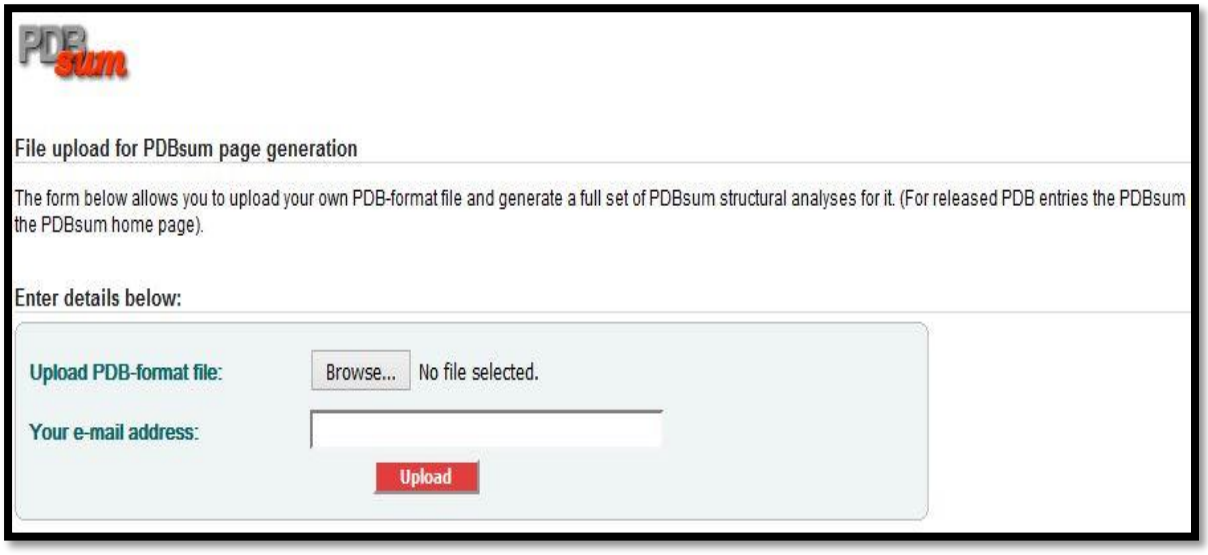


Figure 5.5: PDBsum Generate Homepage

Chapter 5:

Results

5.1 Protein Analysis

5.1.1 Protein Sequence

The target protein sequence of the transmembrane protein 43 (TMEM 43), [Homo sapiens] was obtained in FASTA format from NCBI under the protein database. The accession number is **NP_077310.1**.

>NP_077310.1 transmembrane protein 43 [Homo sapiens] RecName: Full=transmembrane protein 43; AltName: Full=Protein LUMA

MAANYSSTSTRREHVKVKTSSQPGFLERLSETSGGMFVGLMAFLLSFYLIFTNEG
RALKTATSLAEGLSLVVSPDSIHSVAPENEGRLVHIIGALRTSKLLSDPNYGVHLP
AVKLRRHVEMYQWVETEESREYTEDGQVKKETRYSYNTEWRSEIINSKNFDREI
GHKNPSAMAVESFMATAPFVQIGRFFLSSGLIDKVDNFKSLSLSKLEDPHVDIIRR
GDFFYHSENPKYPEVGDLRVSFSYAGLSGDDPDLGPAHVVTVIARQRGDQLVPFS
TKSGDTLLLLHHGDFSAEEVFHRELRSNSMKTWGLRAAGWMAMFMGLNLMTRI
LYTLVDWFPVFRDLVNIGLKAFAFCVATSLTLLTVAAGWLFYRPLWALLIAGLA
LVPILVARTRVPAKKLE

5.1.2 Homology

5.1.2.1 Blast Results

The target protein sequence was subjected to blast using the blastp (protein-protein BLAST) algorithm in the standard protein blast suite. The database against which the search was performed was the non-redundant protein sequence database. All algorithm parameters were set to default where maximum target sequence was 100 and the scoring matrix used was BLOSUM 62. Once the search was completed, a blast report was presented into three sections: a graphical summary (Figure 5.6a), a list of sequences producing significant alignments (Figure 5.6b), and the corresponding alignments (Figure 5.6c).

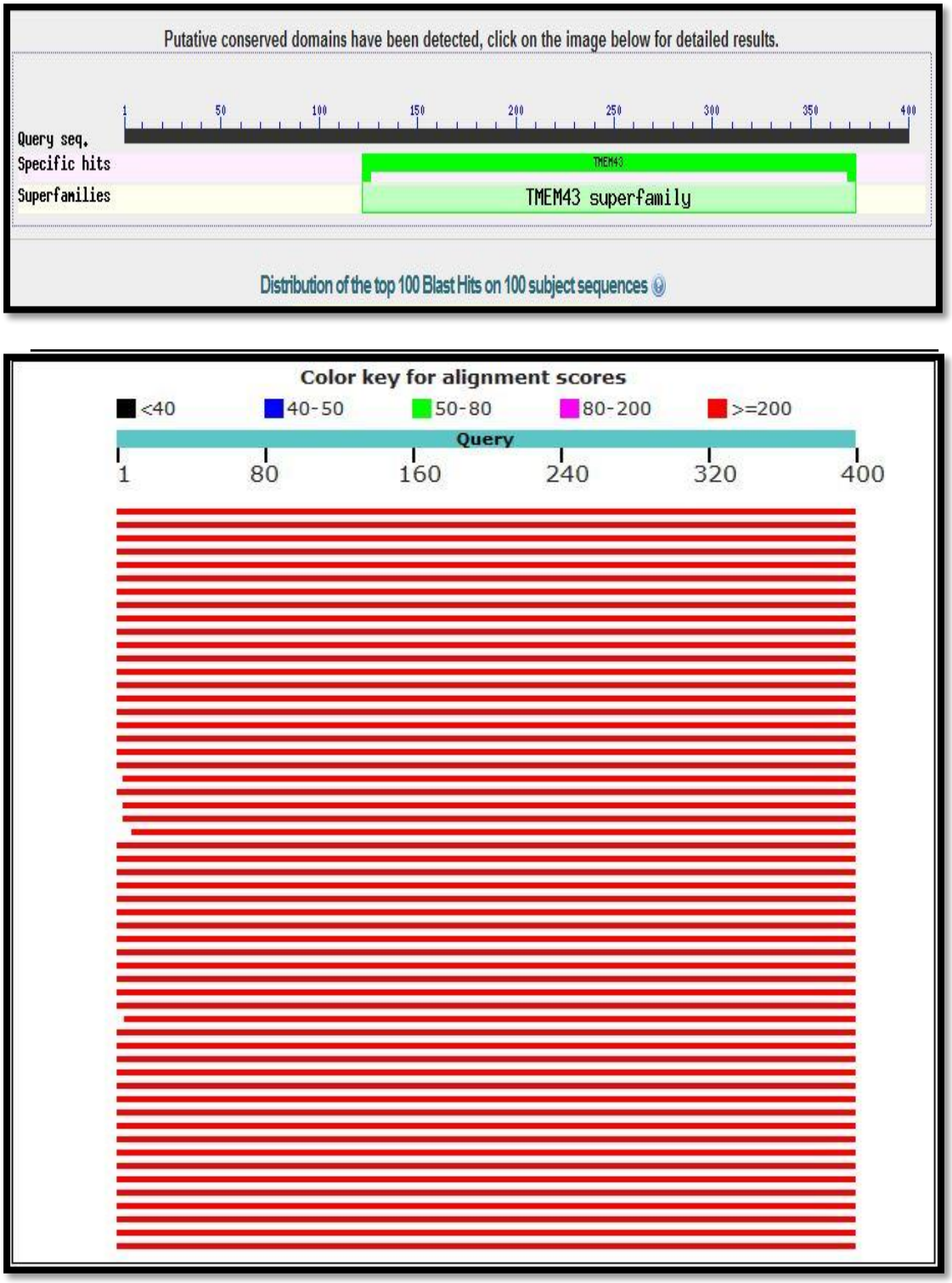


Figure 5.6a: A graphical overview of all the protein sequences' blastp hits for our query sequence

The graphical summary shows the alignments (as colored boxes) of protein sequences that matched our query sequence (the top red bar under the color key). The color keys represent the score (S) of the alignment, with red indicating the highest score and black indicating the lowest score. Thus higher alignment score means more significant hit. Of the aligned sequences, the most similar are placed closest to the query.

In this case, a total of 100 sequences were present in summary of which there were ninety four high scoring database matches that aligned to most of the query sequence. The remaining six bars (21-25 and 39) represented slightly lower-scoring matches that aligned to the query, indicating fewer matches to the query.

Sequences producing significant alignments:

Select: All None Selected:0

Alignments | Download | GenPept | Graphics | Distance tree of results | Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
transmembrane protein 43 [Homo sapiens]	821	821	100%	0.0	100%	NP_077310.1
TMEM43 [synthetic construct]	819	819	100%	0.0	99%	AKI70166.1
unnamed protein product [Homo sapiens]	818	818	100%	0.0	99%	BAC11350.1
PREDICTED: transmembrane protein 43 [Pan paniscus]	818	818	100%	0.0	99%	XP_003826279.1
hypothetical protein [Homo sapiens]	817	817	100%	0.0	99%	CAB66850.1
PREDICTED: transmembrane protein 43 [Pan troglodytes]	816	816	100%	0.0	99%	XP_001157301.1
PREDICTED: transmembrane protein 43 [Nomascus leucogenys]	816	816	100%	0.0	99%	XP_003265059.1
PREDICTED: transmembrane protein 43 [Gorilla gorilla gorilla]	814	814	100%	0.0	99%	XP_004033711.2
Transmembrane protein 43 [Homo sapiens]	813	813	100%	0.0	99%	AAH11719.1
transmembrane protein 43 [Pongo abelii]	810	810	100%	0.0	99%	NP_001125873.1
PREDICTED: transmembrane protein 43 [Cercocebus atys]	802	802	100%	0.0	98%	XP_011915636.1
PREDICTED: transmembrane protein 43 [Rhinopithecus roxellana]	802	802	100%	0.0	97%	XP_010382778.1

Figure 5.6b: A segment of the list of proteins that produced significant alignments

The summary table (Figure 5.6b) shows all the sequences in the database that showed significant match to the query sequence. The results were sorted in descending order in terms of the Expect value (E-value). The E-value is the number of alignments expected by chance with the same score. A number close to zero means that the hit has to be significant and not due to chance. Normally, $E < .05$ is required to be considered significant. The Max score is the highest alignment score between the query sequence and the database sequence while the total score is the summation of alignment scores of all segments of the same sequence that matched the query sequence. The query coverage is

the percentage of the query length that is included in the aligned segments. If BLAST could align all 400 nucleotides of a query against a hit, then that would be 100% coverage. Lastly, the identity is the percentage identity between the query and the hit in an amino acid to amino acid alignment. In this case, out of hundred sequences, ten sequences (including the query sequence) were selected based on E-values and identity values (ranging from 80% to 100%). The selected protein sequences are tabulated below according to species, protein, protein accession ID and identity values (Table 1.1).

SL	Species	Protein Accession ID	Protein Name	Identity Values
1	*Homo sapiens*	NP_077310.1	transmembrane protein 43	100%
2	*Pan paniscus* (pygmy chimpanzee)	XP_003826279.1	PREDICTED: transmembrane protein 43	99%
3	*Pongo abelii* (Sumatran orangutan)	NP_001125873.1	transmembrane protein 43	99%
4	*Cercocebus atys* (sooty mangabey)	XP_011915636.1	PREDICTED: transmembrane protein 43	98%
5	*Papio Anubis* (olive baboon)	XP_003906904.1	PREDICTED: transmembrane protein 43 isoform X1	97%
6	*Carlito syrichta* (Philippine tarsier)	XP_008068792.1	PREDICTED: transmembrane protein 43	95%
7	*Ursus maritimus* (polar bear)	XP_008696051.1	PREDICTED: transmembrane protein 43	94%
8	*Panthera pardus* (leopard)	XP_019289130.1	PREDICTED: transmembrane protein 43 isoform X2	93%

9	*Equus caballus* (horse)	XP_001489843.2	PREDICTED: transmembrane protein 43	92%
10	*Bos indicus* (even-toed unqulates)	XP_019839707.1	PREDICTED: transmembrane protein 43	91%

Table 1.1: List of selected homologous sequences for multiple sequence alignment

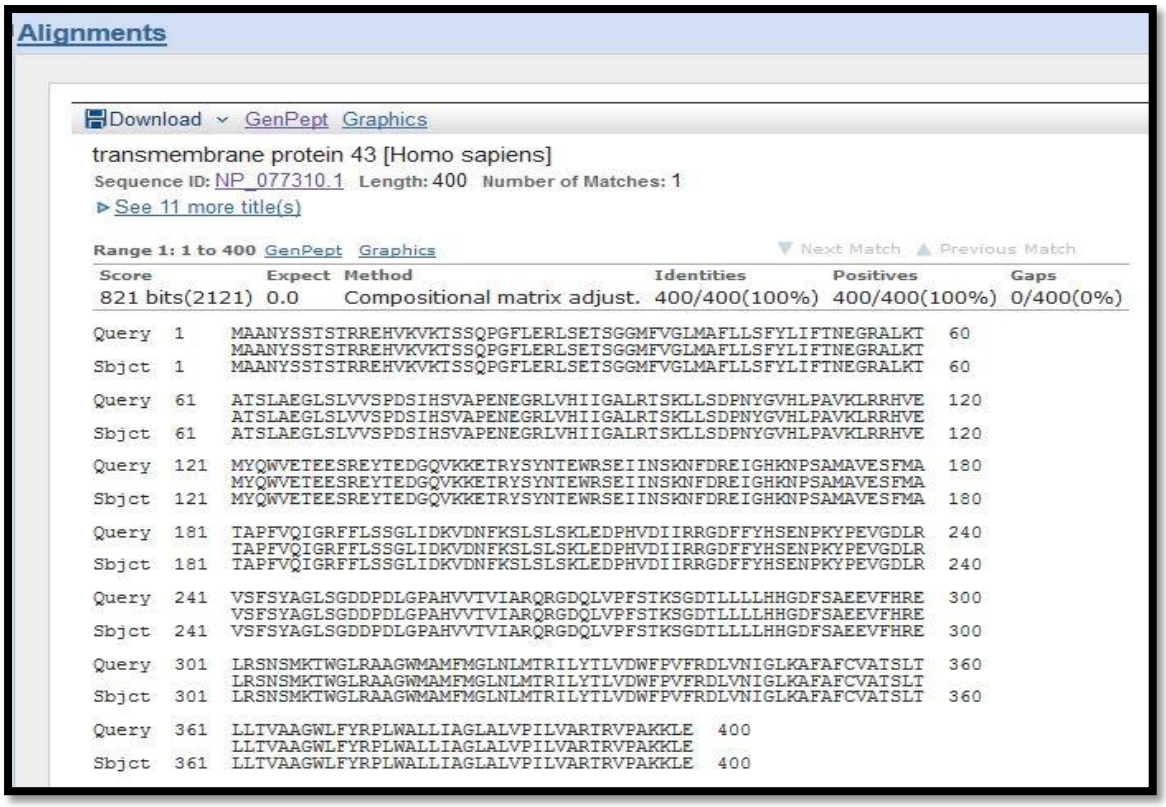
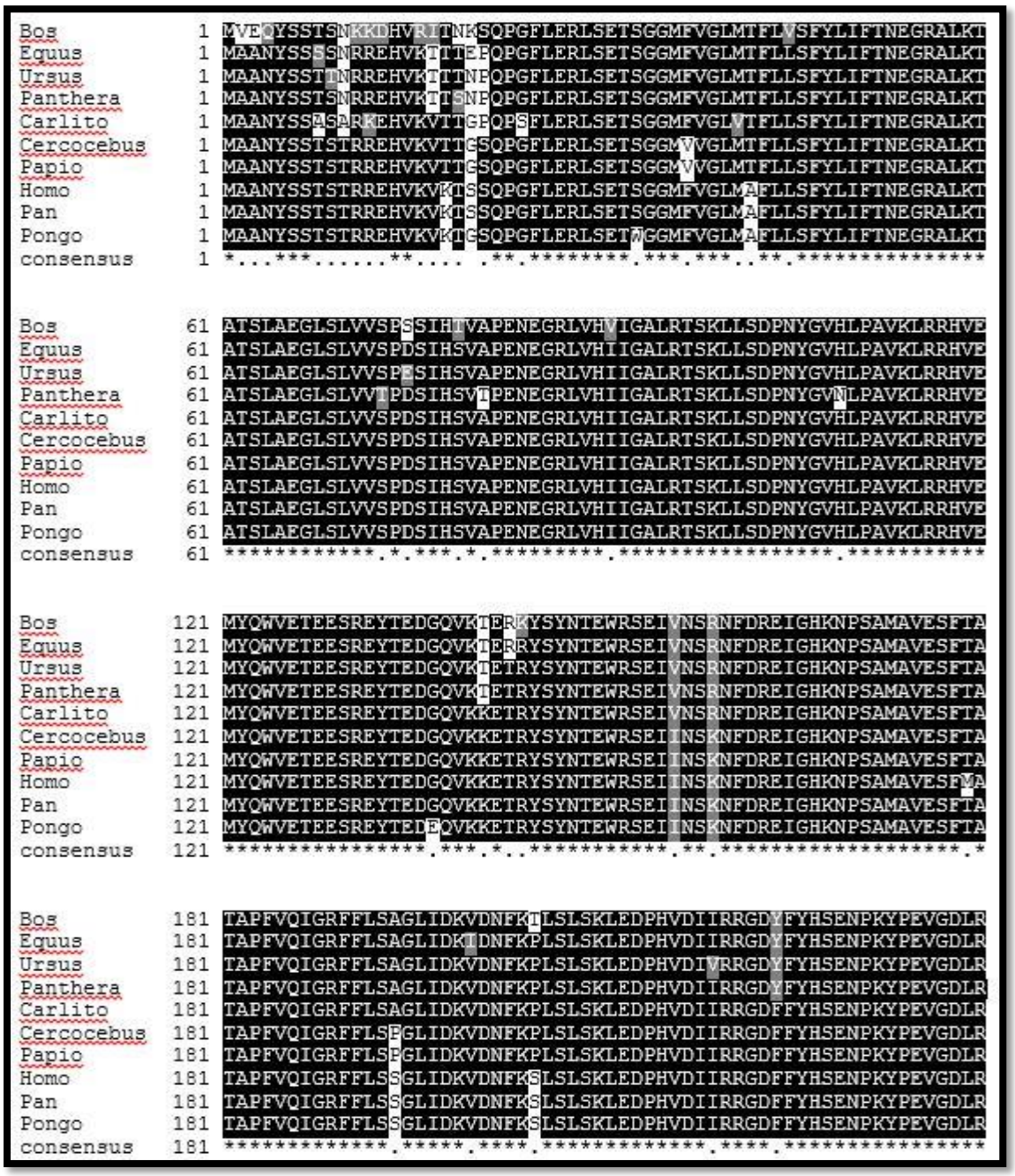


Figure 5.6c: An example of the corresponding alignment of a single sequence

5.1.2.2 Multiple Sequence Alignment (MSA) results using Clustal Omega and BoxShade

Multiple sequence alignment of the selected protein sequences was performed using Clustal Omega. To convert the results into a publishable format, the alignment file was downloaded and the alignments were pasted into the BoxShade server. The output file generated was downloaded where identical or similar amino acid sequences were shaded (Figure 5.7).



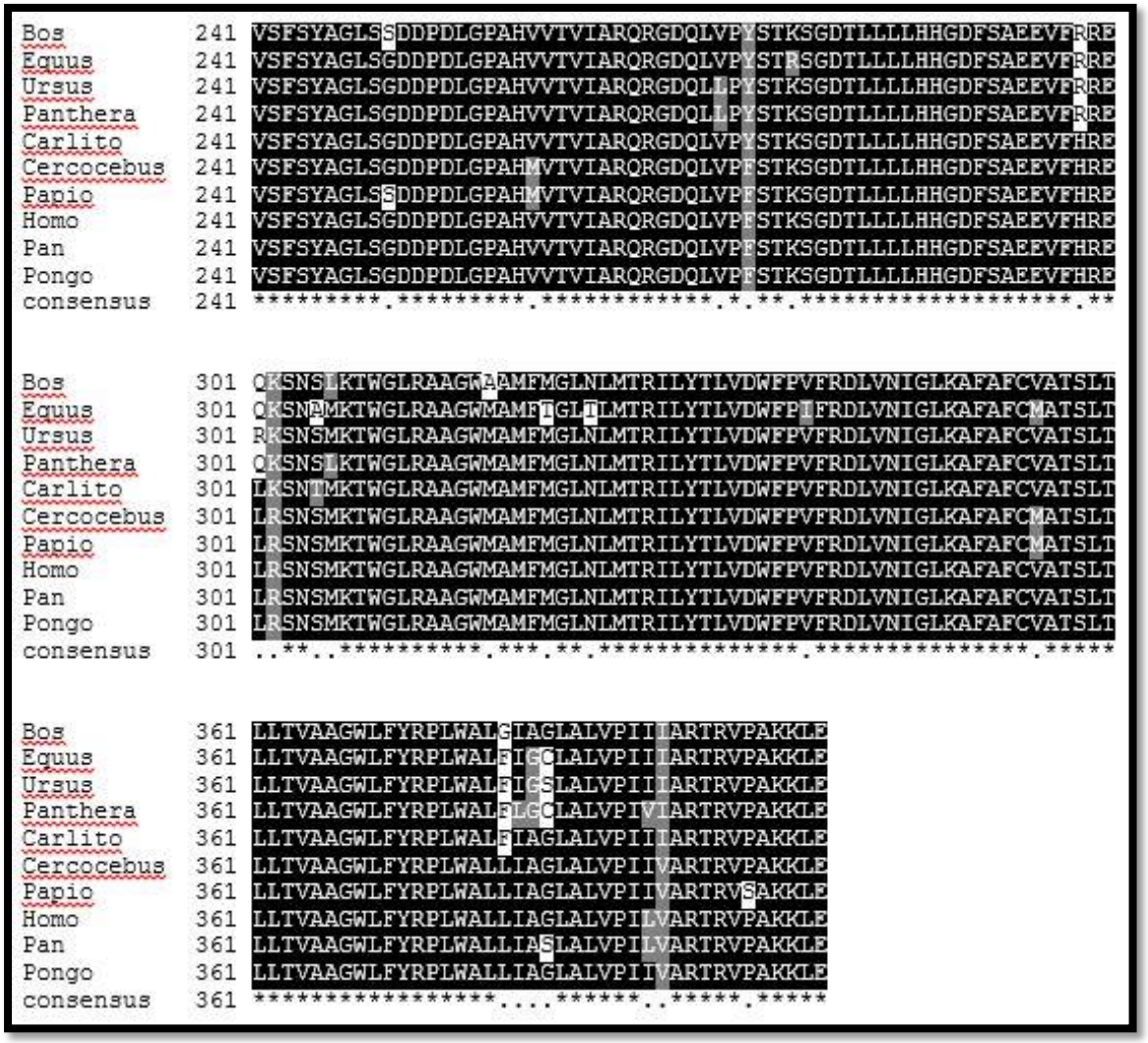


Figure 5.7: The publishable format of the multiple sequence alignment showing the conservation of amino acid residues of the homologous sequences selected where Homo is the query sequence generated by BoxShade. Identical residues are shaded black.

The interpretation of a multiple sequence alignment depends mainly on its appearance. The protein sequences were aligned to show the similarities and differences between each amino acid residue of the selected protein sequences. Different types of shades were applied to identical or similar residues, where each column shaded has certain level of conservation. Two types of shading were used, whereby black shade was used to represent identical amino acids and gray shade indicated similar amino acids. The last row shows consensus sequences which are short sequences of amino acids found in protein motifs that share a common sequence among families of similar proteins. The star '*' indicates an entirely conserved column and the period '**.**' indicate columns where the size or hydropathy has been preserved during the course of evolution.

Protein structures contain surface loops (softer portions of protein that connect its more rigid portions) that evolve rapidly and core regions that act as support walls for the protein. These support walls evolve less rapidly than the loops on the surface. In the multiple sequence alignment the gap-free blocks correspond to the core regions and gap-rich regions correspond to the loops. Thus surface loops usually do not have high conservation of residues as they evolve rapidly. On the other hand, the core of the protein is more stable and is expected to have conserved columns of residues. As observed in figure 5.7, the amino acid sequences were highly conserved from the beginning to the end of the sequence and showed gap-free regions. Thus it can be said that the selected protein sequences were highly similar and mainly contained core regions with very few or no surface loops.

5.1.3 Phylogenetic

Interpreting a phylogenetic tree is a lot similar to reading a family tree. The tree can be broken down into 'nodes', where lineage splits (speciation) leading to branching on a phylogeny and single ancestral lineage divides into two or more lineages and 'tips' or 'leaves' of branches that represent descendants of that ancestor. Moving forward from nodes to tips represent moving forward in time. The phylogenetic tree was constructed using the 'one click' method keeping all the parameters in default (Figure 5.8).

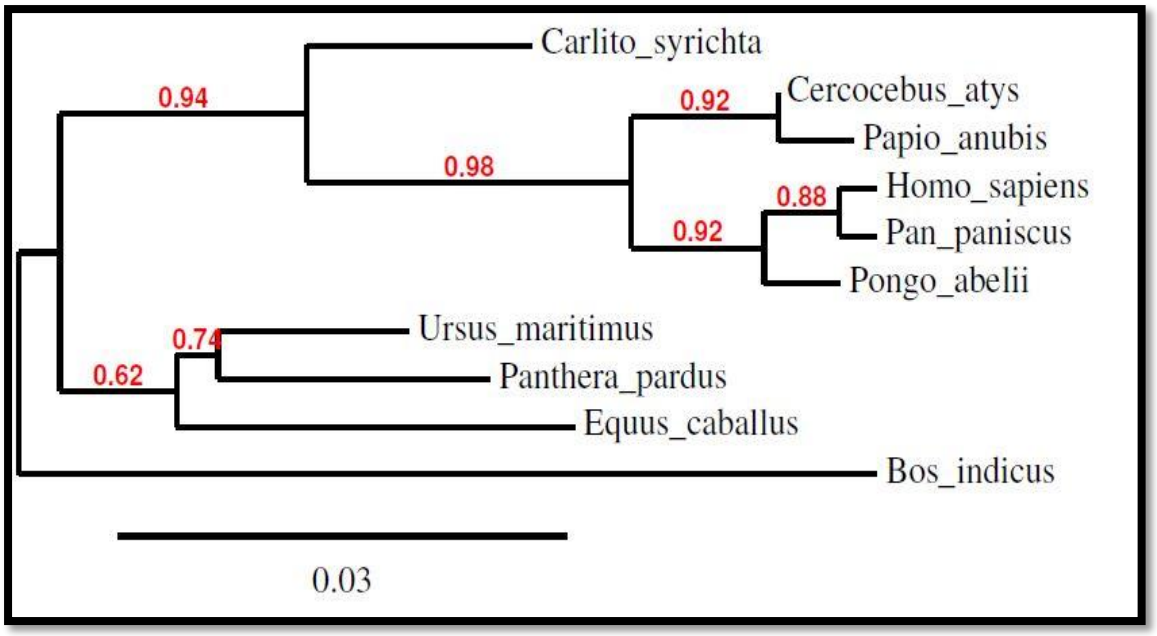


Figure 5.8: Phylogenetic tree obtained from Phylogeny.fr

The horizontal lines represent the branches and show changes in evolutionary lineage over time. The longer the horizontal lines, the larger the amount of change. The bar at the bottom of the figure provides a scale for this. In this case the line segment with the number '0.03' showed the length of the branch that represents an amount genetic change of 0.03. Here the protein of *Bos_indicus* is more distantly related to all the other proteins. The numbers next to each node gave a measure of support for the node. The numbers are usually between 0 and 1 where 1 represents maximal support. These were calculated by different statistical approaches like 'bootstrapping' and 'Bayesian posterior probabilities'.

5.1.4 Amino Acid Composition

The amino acid composition of the transmembrane protein 43 computed using Pepstats was tabulated (Table 1.2).

PEPSTATS of Homo from 1 to 400		
Molecular weight = 44875.55 Da		**Residues** = 400
Average residue weight = 112.189 Da		**Charge** = 7.0
Residue	**Number**	**Mole %**
A = Ala	30	7.500
B = Asx	0	0.000
C = Cys	1	0.250
D = Asp	18	4.500
E = Glu	25	6.250
F = Phe	23	5.750
G = Gly	27	6.750
H = His	12	3.000
I = Ile	16	4.000
K = Lys	18	4.500
L = Leu	48	12.000
M = Met	11	2.750
N = Asn	13	3.250
P = Pro	17	4.250
Q = Gln	6	1.500
R = Arg	26	6.500
S = Ser	36	9.000
T = Thr	23	5.750
V = Val	31	7.750
W = Trp	7	1.750
X = Xaa	0	0.000
Y = Tyr	12	3.000
Z = Glx	0	0.000

Table 1.2: Amino acid composition of the transmembrane protein 43 obtained by the Pepstats analysis tool

The protein in this study contains 400 amino acid residues with a total molecular weight of approximately 44876 Da. The amino acid composition of the protein showed that the most abundant amino acid is leucine which accounts for ~12% of the protein's primary structure whereas cysteine is the least common amino acid and consists of 0.25% of its primary structure. Leucine is a branched, aliphatic α-amino acid containing an aliphatic isobutyl side chain which makes leucine hydrophobic in nature. Because of its hydrophobic nature, leucine is generally buried in folded proteins. Thus the abundant number of leucine residues discovered in the transmembrane protein 43 (TMEM 43) make up the core of the protein along with other hydrophobic residues alanine, isoleucine, methionine, phenylalanine, valine, proline, glycine. While hydrophobic amino acid residues build up the core, polar and charged amino acids preferentially cover the surface of the molecule and are in contact with solvent due to their ability to form hydrogen bonds. Polar amino acids include cysteine, serine, threonine, asparagine, glutamine, histidine and tyrosine while arginine, lysine, aspartic acid and glutamic acid are charged. Cysteine residues are capable of formation of disulfide bonds which plays a role in stability and folding of the structure. The very low amounts of cysteine residues indicated that the protein gains its stability from other interactions as chances of disulfide bond formation are very low.

5.1.5 Protein Characteristic Analysis

The various physicochemical properties of the selected protein sequence were computed by ProtParam tool and tabulated (Table 1.3). The parameters computed were molecular weight, theoretical pI, amino acid composition, atomic composition, estimated half-life, instability index, aliphatic index and grand average of hydropathicity (GRAVY). The half-life is the predicted time taken for half of the protein to disappear after its synthesis inside a cell. Aliphatic index of a protein is the relative volume taken up by aliphatic side chains (alanine, valine, isoleucine, and leucine). Instability index gives an estimation of a protein's stability inside a test tube; a value above 40 indicates the protein may be unstable. Lastly, the GRAVY value of a protein indicates its hydrophobicity, whereby increasing positive score indicates a greater hydrophobicity.

ProtParam Results		
Number of amino acids: 400		**Molecular weight:** 44875.55
Theoretical pI: 7.86		
Amino acid composition:		
Total number of negatively charged residues (Asp + Glu): **43** Total number of positively charged residues (Arg + Lys): **44**		
Atomic composition:		
Carbon Hydrogen Nitrogen Oxygen Sulfur	C H N O S	2033 3168 546 577 12
Formula: $C_{2033}H_{3168}N_{546}O_{577}S_{12}$		
Total number of atoms: 6336		
Estimated half-life:		
The N-terminal of the sequence considered is M (Met). The estimated half-life is: 30 hours (mammalian reticulocytes, in vitro). >20 hours (yeast, in vivo). >10 hours (Escherichia coli, in vivo).		
Instability index:		
The instability index (II) is computed to be 40.65 This classifies the protein as unstable.		
Aliphatic index: 92.38		
Grand average of hydropathicity (GRAVY): -0.053		

Table 1.3: The parameters computed by the ProtParam program from the transmembrane protein 43 sequence in determination of the protein's physico-chemical properties.

The Isoelectric Point (pI) of the protein was calculated to be ~7.86. The pI was at a higher pH because the basic side chain introduces an "extra" positively charged residue. So the neutral form exists under more basic condition when the extra positive charge has been neutralized. At this point, the protein is least soluble, and therefore unstable. The instability index for the protein is greater than 40, classifying it as unstable. The protein had a negative GRAVY score, which means they are hydrophilic in nature. Lastly the Aliphatic index (Ai) value was quite high, which suggested that the protein may remain stable over a wide range of temperatures.

5.1.6 Prediction of Transmembrane Segments

5.1.6.1 Prediction via ProtScale

The presence of a transmembrane segment in the protein LUMA was confirmed by plotting the hydropathy index. The hydropathy index was plotted using the linear sequence of the protein beginning at the amino terminus. A 19 amino acid window was used to plot the index, starting with the first amino acid. The resulting plot revealed the relative hydrophobicity of segments of the protein (Figure 5.9). The image seen in figure 5.9 is the hydrophobicity plot returned by ProtScale using the Kyte & Doolittle Scale (Kyte & Doolittle, 1982). Since a large window size of 19 was selected for finding transmembrane domains, values above 1.6 were considered to be significant (Kyte and Doolittle, 1982).

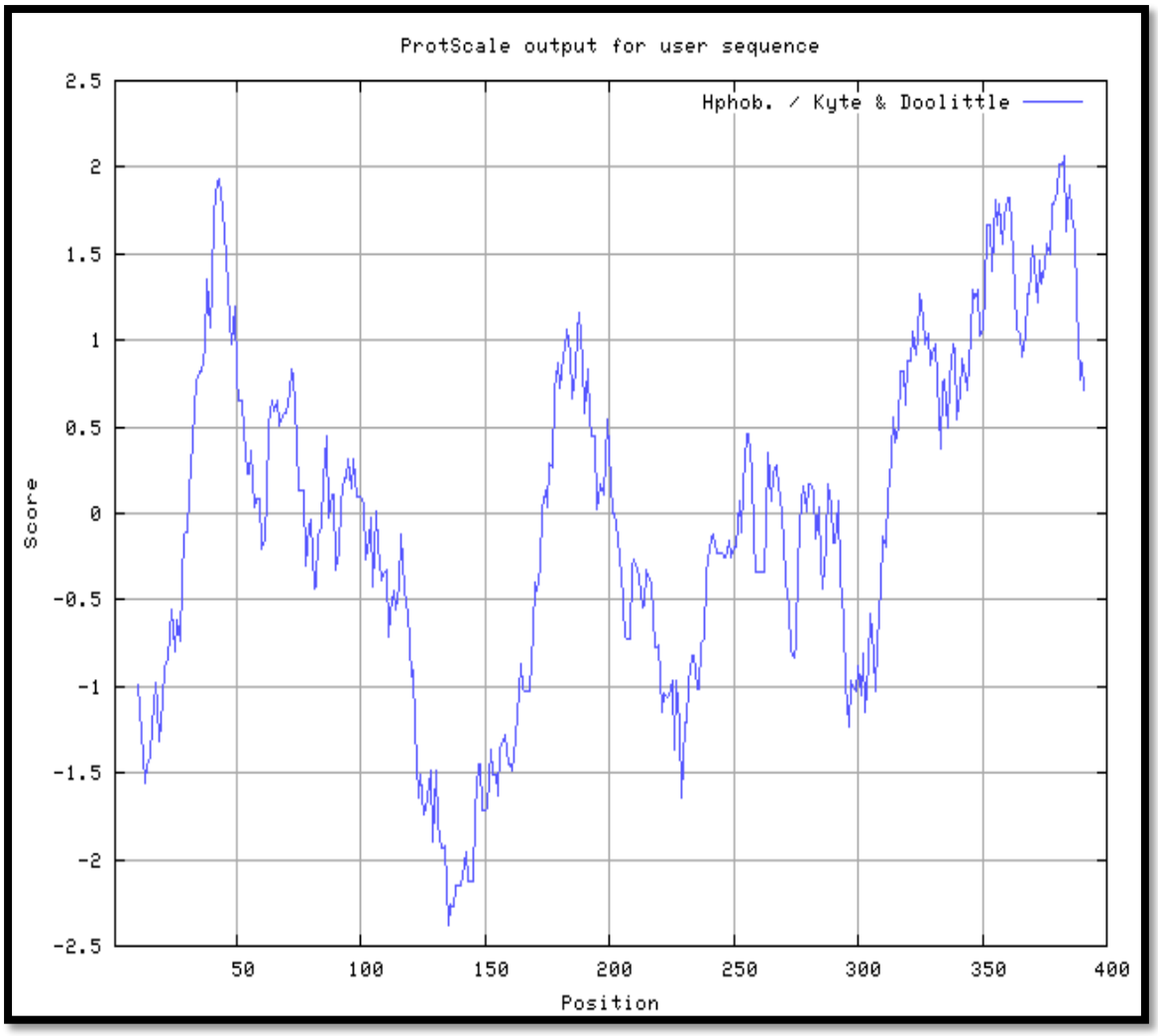


Figure 5.9: ProtScale output of transmembrane protein 43

The peaks in the plot are predicted to be the potential transmembrane regions present within the protein over a span of 400 amino acids. The higher the peak, the higher is the hydrophobicity of the region which indicates those regions are buried in the non-polar phase of the lipid membrane, which can therefore be said to be transmembrane regions. It can be seen from the plot that there are four peaks with significant score above the threshold value. Thus it can be concluded that there are four transmembrane regions within the protein. The highest score was observed for the last peak which means it is the most hydrophobic and this region also contains the most number of amino acids than the other three peaks because the base of the peak was wider than the other three peaks.

5.1.6.2 Prediction via TMHMM

The TMHMM is much more advanced with more detailed and better graphical representation of the predicted transmembrane regions. The graphical representation produced by TMHMM revealed the putative transmembrane regions within the target protein (Figure 6.0a)

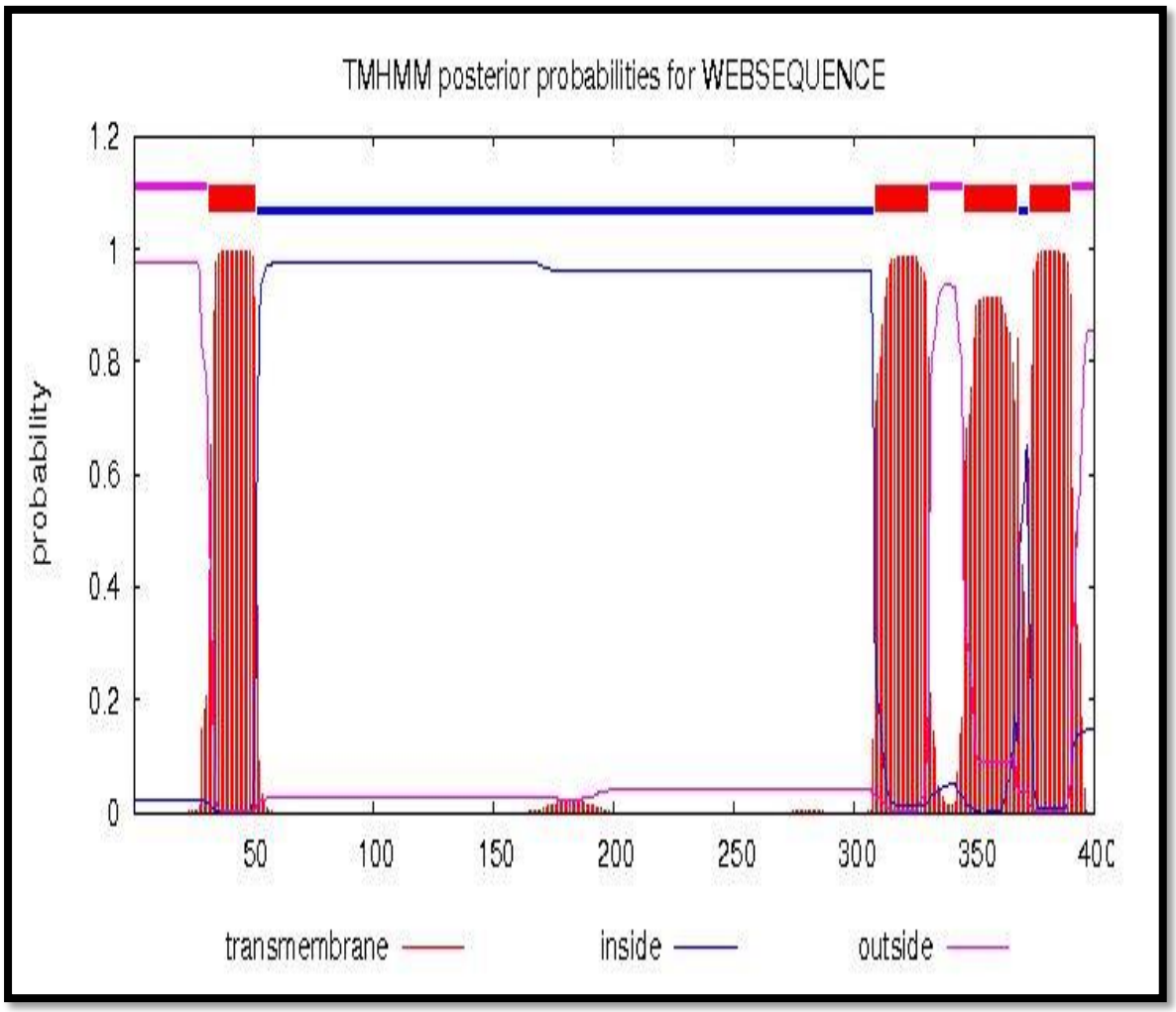
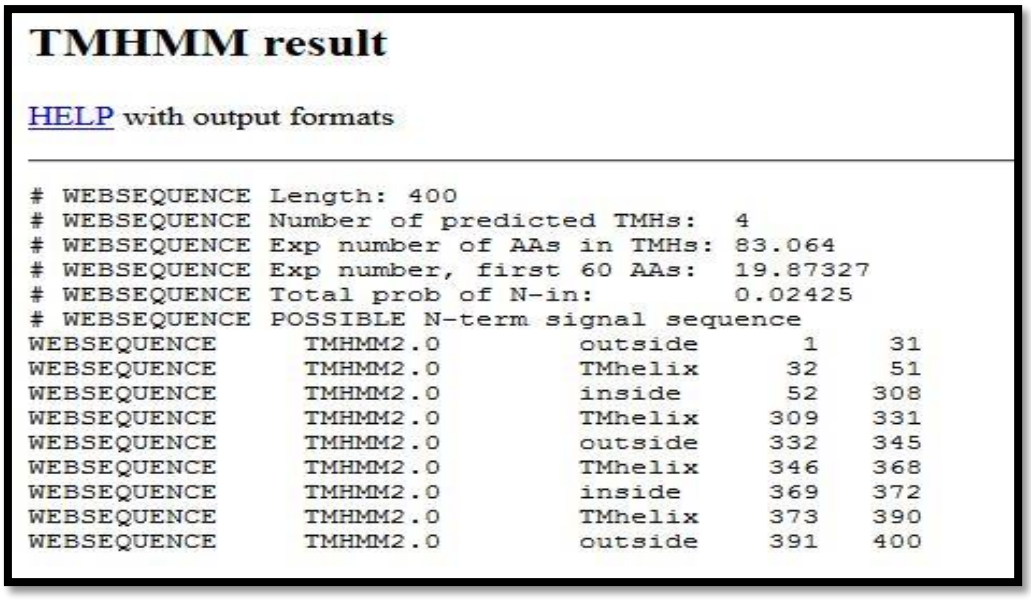


Figure 6.0a: TMHMM output of TMEM 43 (Graphical representation)



TMHMM result

HELP with output formats

```
# WEBSEQUENCE Length: 400
# WEBSEQUENCE Number of predicted TMHs:  4
# WEBSEQUENCE Exp number of AAs in TMHs: 83.064
# WEBSEQUENCE Exp number, first 60 AAs:  19.87327
# WEBSEQUENCE Total prob of N-in:        0.02425
# WEBSEQUENCE POSSIBLE N-term signal sequence
WEBSEQUENCE     TMHMM2.0        outside      1    31
WEBSEQUENCE     TMHMM2.0        TMhelix     32    51
WEBSEQUENCE     TMHMM2.0        inside      52   308
WEBSEQUENCE     TMHMM2.0        TMhelix    309   331
WEBSEQUENCE     TMHMM2.0        outside    332   345
WEBSEQUENCE     TMHMM2.0        TMhelix    346   368
WEBSEQUENCE     TMHMM2.0        inside     369   372
WEBSEQUENCE     TMHMM2.0        TMhelix    373   390
WEBSEQUENCE     TMHMM2.0        outside    391   400
```

Figure 6.0b: TMHMM output of TMEM 43 (Long output format)

It can be seen from figure 6.0b, the number of predicted transmembrane helices is 4 which confirms the results from ProtScale. The expected number of amino acids in transmembrane helices is 83.064 which is larger than 18. This very likely confirms it to be a transmembrane protein. The expected number of amino acids in transmembrane helices in the first 60 amino acids of the protein is 19.87327 which is more than 10 meaning the predicted transmembrane helix in the N-term could be a signal peptide. The total probability that the N-term is on the cytoplasmic side of the membrane is 0.02425. The region of the first transmembrane helix is from 32 to 51 amino acids, the second transmembrane helix is from 309 to 331, third helix is from 346 to 368 and the fourth helix is from 373 to 390. Also, the graphical representation in figure 6.0a shows peaks indicating four transmembrane domains. The blue lines indicate the region of the protein that is inside the membrane whereas the pink lines indicate the regions that are outside the membrane. The red lines indicate the transmembrane regions.

Most of the sequence consists of a hydrophilic region that is between the first and second transmembrane domain from 52 to 308 amino acids. This large region is predicted to be natively unfolded. This hydrophilic domain is most likely to reside in the ER lumen. The sequences of all four predicted transmembrane domains are predicted to form alpha helices of sufficient length to span the membrane.

5.1.7 Prediction of Molecular Structure

5.1.7.1 Results from SOPMA

The secondary structure of protein speaks about the interaction of the residues involved in hydrogen bonding in the polypeptide chain. The secondary structure was proposed by Linus Pauling and Robert Corey, who discovered the two structural features making up the backbone of the protein: alpha helix and beta sheet (Eisenberg, 2003). Apart from these two structures, the beta turns are also present in the secondary structure which together form as an intermediate which finally folds into a three dimensional structure. The self-optimized prediction method with alignment (SOPMA) was used to attain numerical values for the amount of alpha-helices, beta sheets and coils present within the polypeptide chain. The results from SOPMA are shown in figure 6.1 and 6.2.

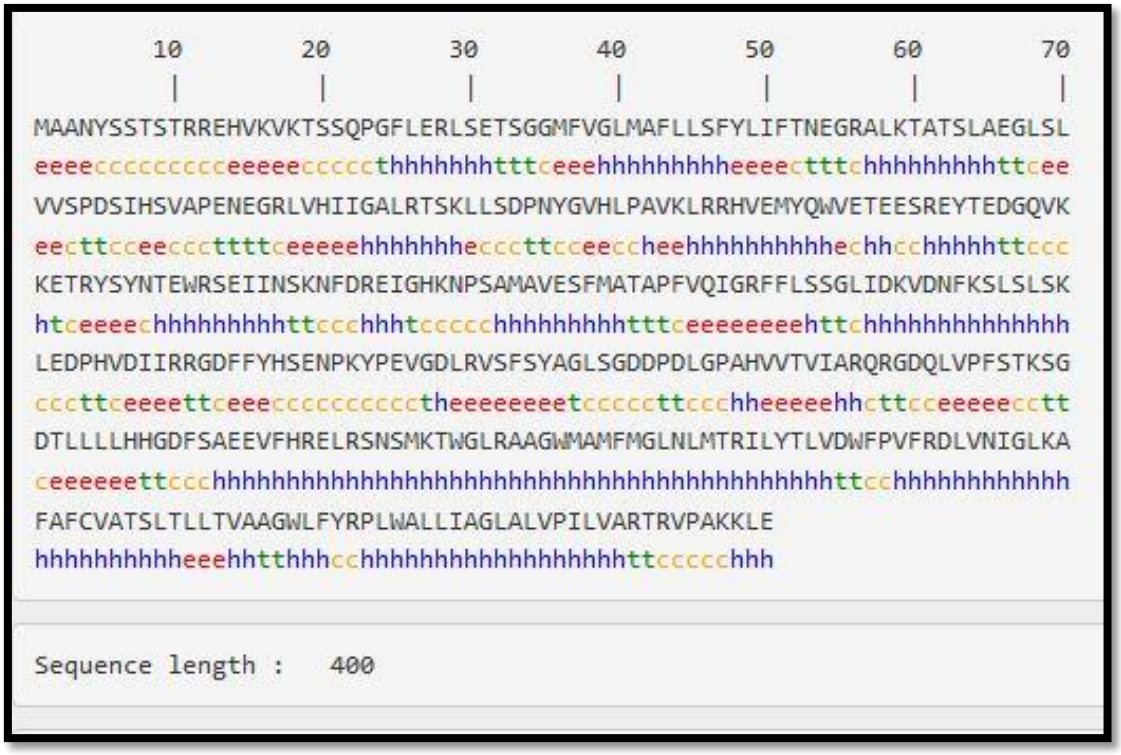


Figure 6.1: Results from SOPMA

```
SOPMA :
   Alpha helix      (Hh) :   182 is  45.50%
   3₁₀ helix        (Gg) :     0 is   0.00%
   Pi helix         (Ii) :     0 is   0.00%
   Beta bridge      (Bb) :     0 is   0.00%
   Extended strand  (Ee) :    79 is  19.75%
   Beta turn        (Tt) :    48 is  12.00%
   Bend region      (Ss) :     0 is   0.00%
   Random coil      (Cc) :    91 is  22.75%
   Ambiguous states (?)  :     0 is   0.00%
   Other states          :     0 is   0.00%
```

Figure 6.1 (continued): Results from SOPMA

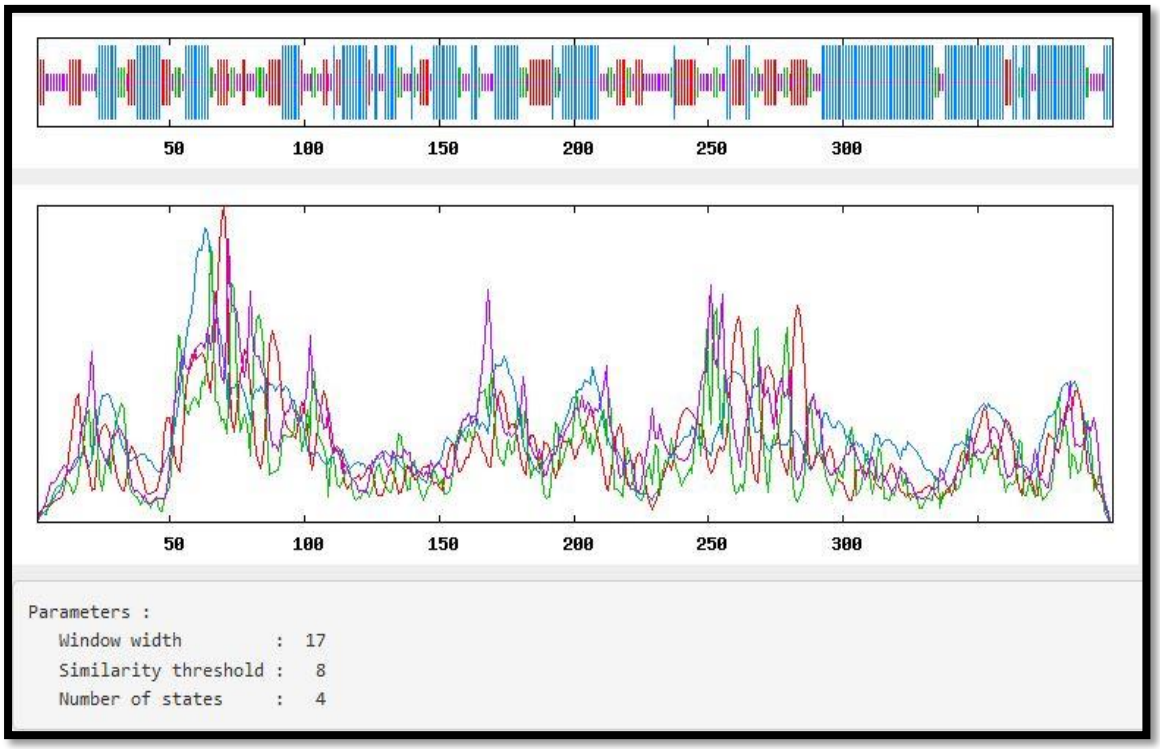


Figure 6.2: Graphs generated by SOPMA. The first one is to visualize the prediction and the second contains score curves for all predicted states. It also shows the parameters such as window width, number of states etc. that are used for the prediction.

The analysis revealed that the α-helices are dominant amongst the secondary structures followed by the coils, extended strands/β-sheets and β-turns. The data revealed that 45.50% of the target protein's secondary structure is composed of α-helix with 22.75% of the protein in random coils. β-sheets and β-turns accounted for only 19.75% and 12.00% of the protein's secondary structure respectively.

Based on the results, which show that the protein has a high a percentage of α-helices, it was predicted that stability of the protein is attained from the hydrogen bonds which is the most important feature of α-helix. Furthermore, it is common for transmembrane proteins to have a high percentage of α-helices in their secondary structure as the internal hydrogen bonds that form the backbone of the molecule can be satisfied by the helical structure. None of the polar groups are left open to the membrane so long as the sidechains are hydrophobic. Hydrogen bonds are vital for protein's stabilization, folding and function (Adamian & Liang, 2002; Curran & Engelman, 2003; Senes, Ubarretxena-Belandia, & Engelman, 2001). Therefore, it can be said that the protein achieves its stability mainly from its alpha helical structure.

5.1.7.2 Results from SwissModel

SwissModel computes modeling requests using a comparative modeling engine called ProMod3. The target sequence is searched with BLAST and HHBlits against the primary amino acid sequence contained in the SWISS-MODEL template library (SMTL). For each identified template, the template's quality is then predicted from features of the target-template alignment. The templates with the highest quality is then selected for model building. Models are then built based on the target-template alignment using ProMod3. The models generated are assessed in light of QMEAN scoring function (Benkert, Biasini, & Schwede, 2011) which gives a measure of the level of nativeness of the structural features discovered in the model and shows whether the model is analogous to experimental structures. Higher QMEAN score demonstrates better understanding between the model structure and experimental structures of comparative size. A score of -4.0 or below indicates that the models are of low quality. The results from SwissModel are shown in figure 6.3 and 6.4.

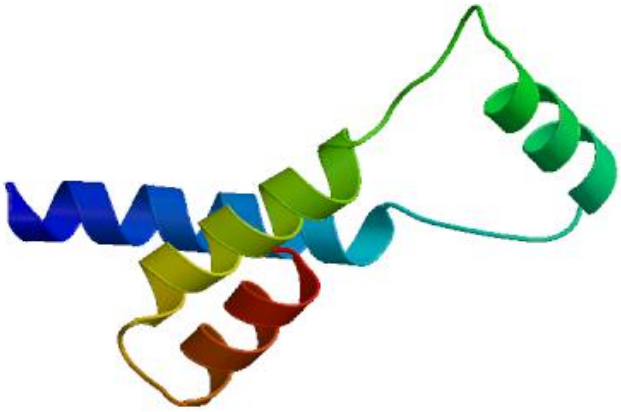


Figure 6.3a: Model 1 from SwissModel

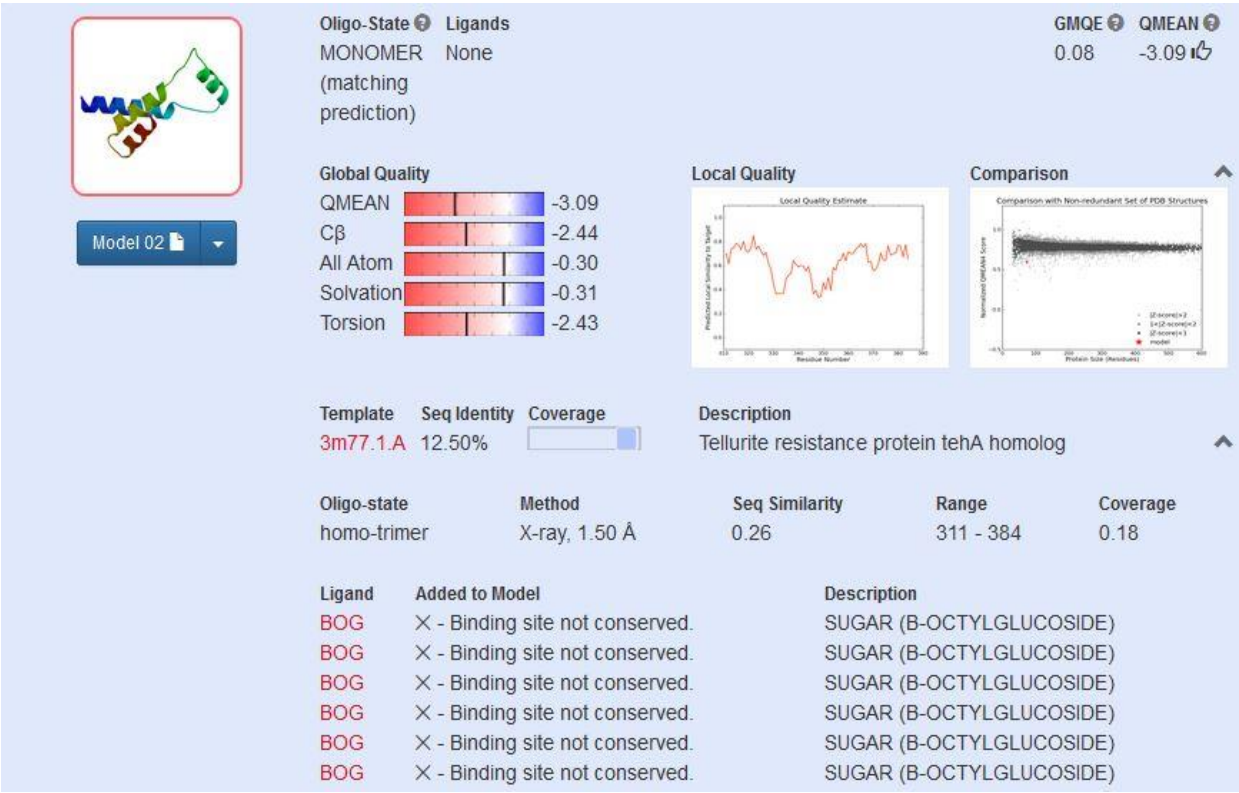


Figure 6.3b: SwissModel results for Model 1

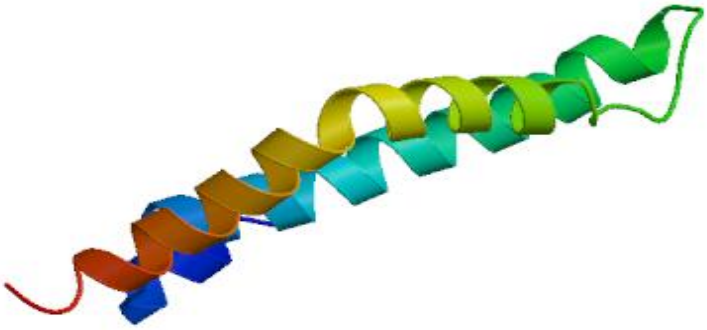


Figure 6.4a: Model 2 from SwissModel

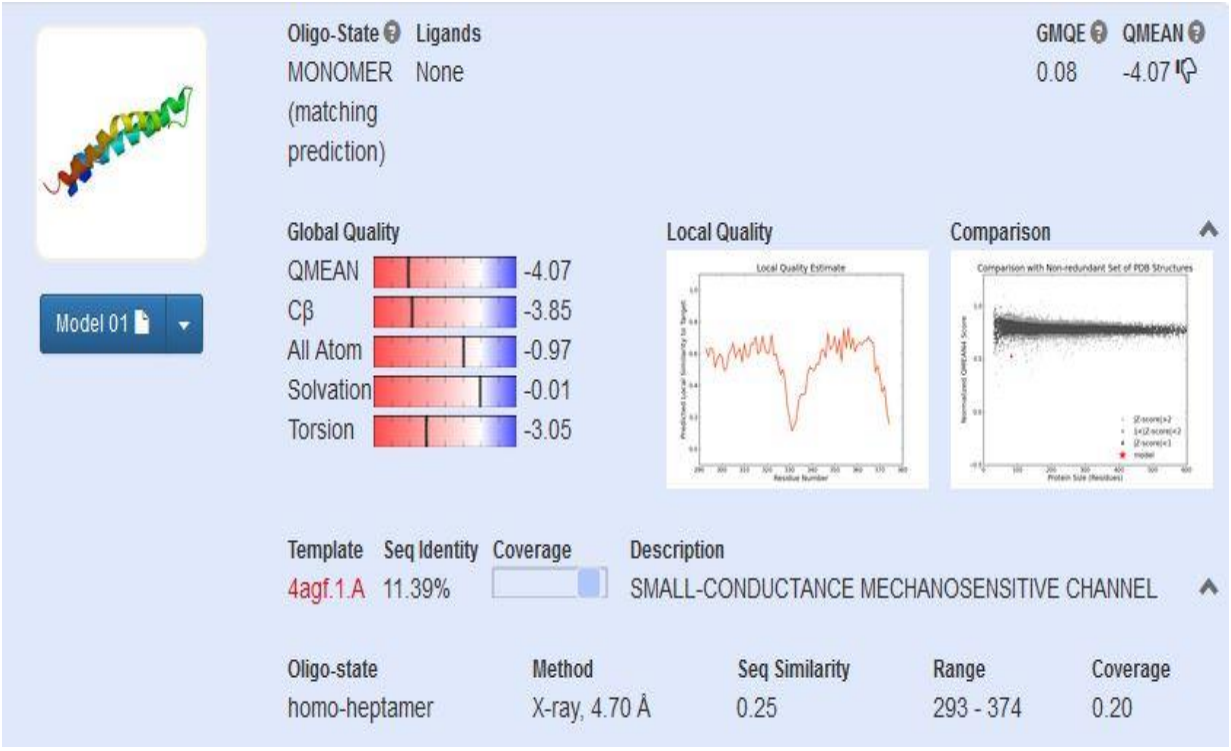


Figure 6.4b: SwissModel results for Model 2

SwissModel generated two probable final 3-D structure models of the transmembrane protein 43 using a total of 15 templates for the homology modeling. The final model was presented based on the QMEAN model quality. Information about the oligomeric state, as well as bound ligands and cofactors were also provided. Model 1 had a QMEAN score of -3.09 while model 2 had a score of -4.07. The oligomeric state of both the models generated was monomer. None of the models had any ligands that may be able to bind with the protein. Since the model 2 had a score below -4.0, this was an indication that the model was of very low quality. Thus, model 1 was believed to be the predicted structure of the protein.

5.1.7.3 Results from I-TASSER

I-TASSER modeling begins with the identification of structure templates by LOMETS from the PDB library. LOMETS is a meta-server threading approach containing multiple threading programs, where each threading program can generate tens of thousands of template alignments. Once the templates are generated they are selected in terms of the maximum significance in the threading alignments. Z-score is used to measure the significance of the templates. Normally, the template with the maximum Z-score is selected from each threading program. A list of the top ranking templates used by I-TASSER to generate the models that were attained is shown in table 1.4

I-TASSER reported up to five models each having a distinct C-score (Figure 6.5). The confidence of each model was quantitatively measured by C-score. Model 1 had the best C-score of -3.50. A C-score is typically in the range of [-5, 2], where a C-score of a higher value signifies a model with a higher confidence and vice-versa. The C-score of each model is tabulated in table 1.5.

Rank	PDB Hit	Iden 1	Iden 2	Cov	Norm. Z-score
1	5n8oA	0.11	0.21	0.96	0.82
2	3jafA	0.11	0.16	0.69	0.89
3	5k47A	0.08	0.14	0.80	0.90
4	3puqA	0.09	0.19	0.90	0.89
5	5givE	0.12	0.08	0.38	0.53
6	5k47A	0.09	0.14	0.84	0.65
7	4ke4A	0.09	0.21	0.90	0.94
8	5a63C	0.12	0.12	0.21	0.81
9	4pirA	0.13	0.20	0.59	0.69
10	5lp2B	0.00	0.15	0.86	0.52

Note:

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

Table 1.4: The top 10 threading templates generated by LOMETS threading program

I-TASSER Models	C-score
Model 1	-3.50
Model 2	-3.85
Model 3	-4.11
Model 4	-3.51
Model 5	-3.99

Table 1.5: The C-score of the 5 models generated by I-TASSER

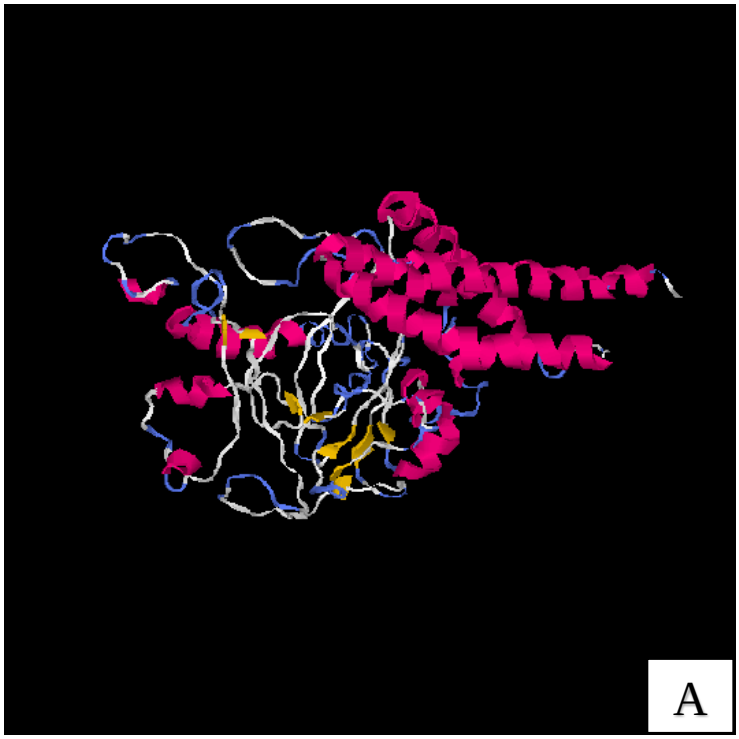
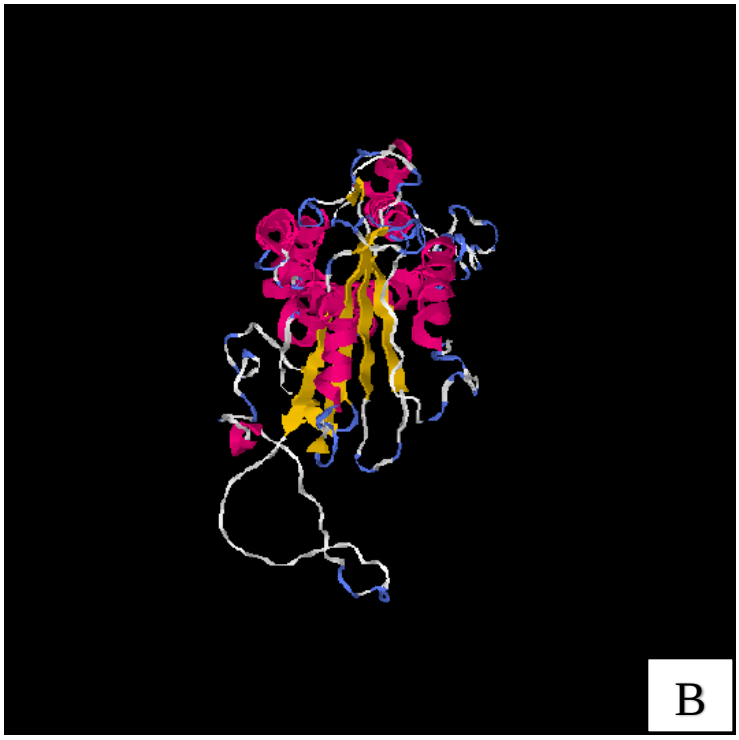


Figure 6.5: The final models generated by I-TASSER. (A) Model 1, (B) Model 2, (C) Model 3, (D) Model 4, (E) Model 5



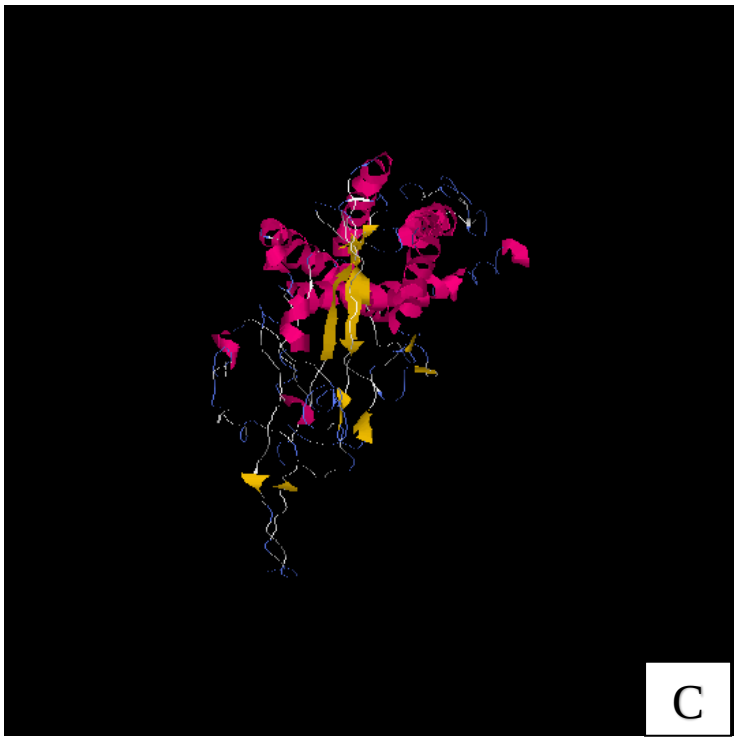
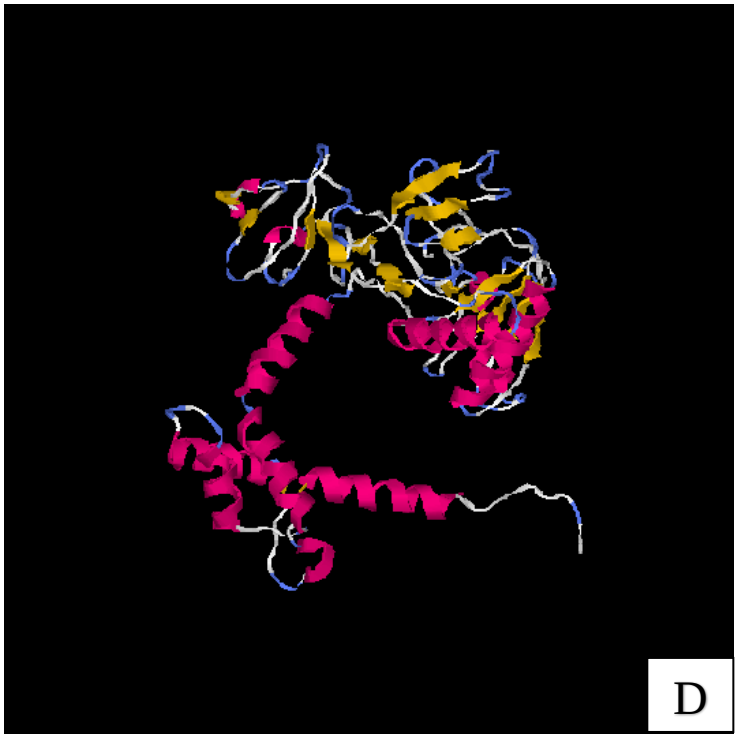


Figure 6.5 (continued): The final models generated by I-TASSER. (A) Model 1, (B) Model 2, (C) Model 3, (D) Model 4, (E) Model 5



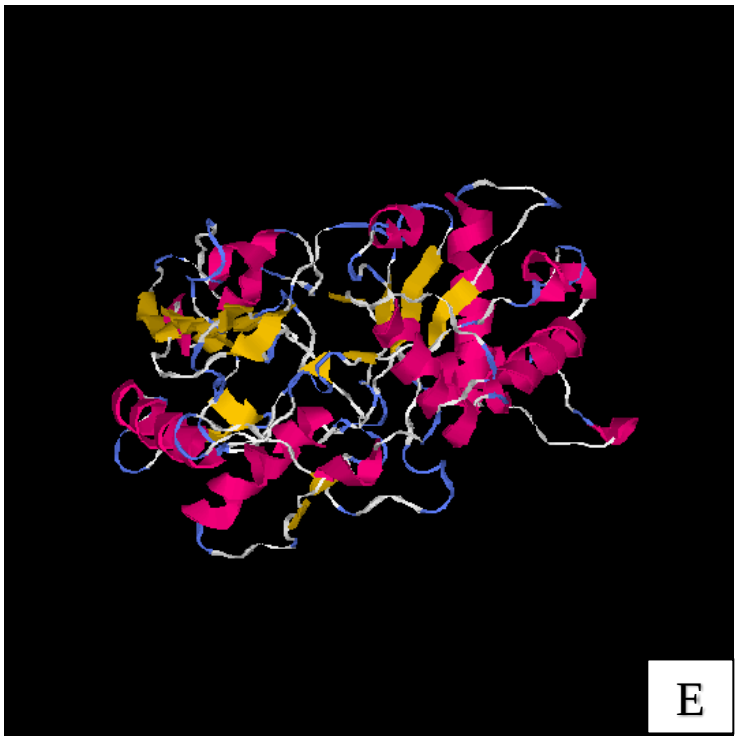


Figure 6.5 (continued): The final models generated by I-TASSER. (A) Model 1, (B) Model 2, (C) Model 3, (D) Model 4, (E) Model 5

After the structure assembly simulation, I-TASSER uses the TM-align structural alignment program to match the first I-TASSER model to all structures in the PDB library. A list of the top 10 proteins from the PDB that have the closest structural similarity, i.e. the highest TM-score, to the predicted I-TASSER model was also generated (Table 1.6). Due to the structural similarity, these proteins often have similar function to the target.

Rank	**PDB Hit**	**TM-score**	**RMSD[a]**	**IDEN[a]**	**Cov**
1	3puqA	0.815	2.89	0.080	0.905
2	3kv6D	0.707	4.08	0.080	0.873
3	3puaA	0.680	3.76	0.098	0.818
4	3kv4A	0.676	4.40	0.066	0.850
5	3kvbA	0.668	4.10	0.084	0.830
6	4do0A	0.666	3.95	0.076	0.815
7	2yu2A	0.656	3.90	0.088	0.795
8	3avsA	0.569	4.72	0.083	0.750
9	2xueA	0.565	4.61	0.074	0.740
10	5cehA	0.5458	5.24	0.063	0.757

Note:

1. **RMSDa** is the RMSD between residues that are structurally aligned by TM-align
2. **IDENa** is the percentage sequence identity in the structurally aligned region.
3. **Cov** represents the coverage of the alignment by TM-align and is equal to the

Table 1.6: Top ten identified structural analogs in PDB provided by I-TASSER

To infer the function of the target protein the data in the section 'Predicted function using COACH' was used. This is because COACH is extensively trained to derive biological functions from multi-source of sequence and structural features which have on average a higher accuracy than the function annotations derived only from the global structure comparison. The biological annotations of the target protein were reported by COACH based on the I-TASSER structure prediction (Table 1.7 & 1.8). COACH is a meta-server approach that combines multiple function annotation results from the COFACTOR, TM-SITE and S-SITE programs.

Rank	C-score	Cluster size	PDB Hit	Lig Name	Ligand Binding Site Residues
1	0.08	3	3n9nA	PEPTIDE	6,171,172,209,214,217,219,227,228,230,231,256
2	0.08	3	5fzaA	P6B	318,319,322,326,357,358,359
3	0.08	3	2axtH	CLA	359,362
4	0.05	2	1aijR	BPH	353,356,357,361
5	0.03	1	5fz3A	7SI	107,110,111,112,113,115

Note:

1. C-score is the confidence score of the prediction. C-score ranges [0-1], where a higher score indicates a more reliable prediction
2. Cluster size is the total number of templates in a cluster.
3. Lig Name is name of possible binding ligand.

Table 1.7: Ligand binding sites

Rank	Cscore^{EC}	PDB Hit	TM-score	RMSD^{a}	IDEN^{a}	Cov	EC Number	Active Site Residue
1	0.098	3kv4A	0.676	4.40	0.066	0.850	1.14.11.27	217
2	0.097	2yu1A	0.654	3.98	0.082	0.797	1.14.11.27	NA
3	0.065	2ilmA	0.417	5.16	0.037	0.575	1.14.11.16	NA
4	0.065	1gp6A	0.396	5.69	0.039	0.570	1.14.11.19	NA
5	0.064	2zuwC	0.382	6.92	0.060	0.623	2.4.1.211	NA

Note:

1. Cscore^{EC} is the confidence score for the EC number prediction. Cscore^{EC} values range in between [0-1]; where a higher score indicates a more reliable EC number prediction.
2. TM-score is a measure of global structural similarity between query and template protein.
3. RMSD^{a} is the RMSD between residues that are structurally aligned by TM-align.
4. IDEN^{a} is the percentage sequence identity in the structurally aligned region.
5. Cov represents the coverage of global structural alignment and is equal to the number of structurally aligned residues divided by length of the query protein

Table 1.8: Enzyme Commission (EC) numbers and active sites

I-TASSER showed five ligands of the protein along with their respective ligand binding sites on the protein. The server also predicted the active site residues for the query protein along with the predicted EC (enzyme commission) numbers for the query protein.

5.1.8 Model Validation

5.1.8.1 Selection of best model between the five I-TASEER models

The five I-TASSER models were evaluated using the PROCHECK which generated Ramachandran plots (Figure 6.6) and plot statistics (Table 1.9) for the respective models. Ramachandran Plot supported model 5 as it had the highest number of residues in the most favored regions (63.6%) and lowest number of residues in the disallowed regions (2.5%). However, the G-Factor overall average was -0.81. 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. Although model 5 was slightly unusual, it was selected to be the best model as it had the most favored regions and least disallowed regions.

Validation		**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**
PROCHECK Ramachandran Plot	Most favored regions	60.2%	62.7%	57.1%	60.7%	63.6%
	Additional allowed regions	31.1%	28.2%	33.6%	31.6%	27.7%
	Generously allowed regions	5.10%	5.40%	6.80%	4.00%	6.20%
	Disallowed regions	3.70%	3.70%	2.50%	3.70%	2.50%
G-Factor Overall Average		-0.42	-0.75	-1.01	-0.70	-0.81

Table 1.9: Comparative values of PROCHECK, G-Factor between the three I-TASSER modeled proteins

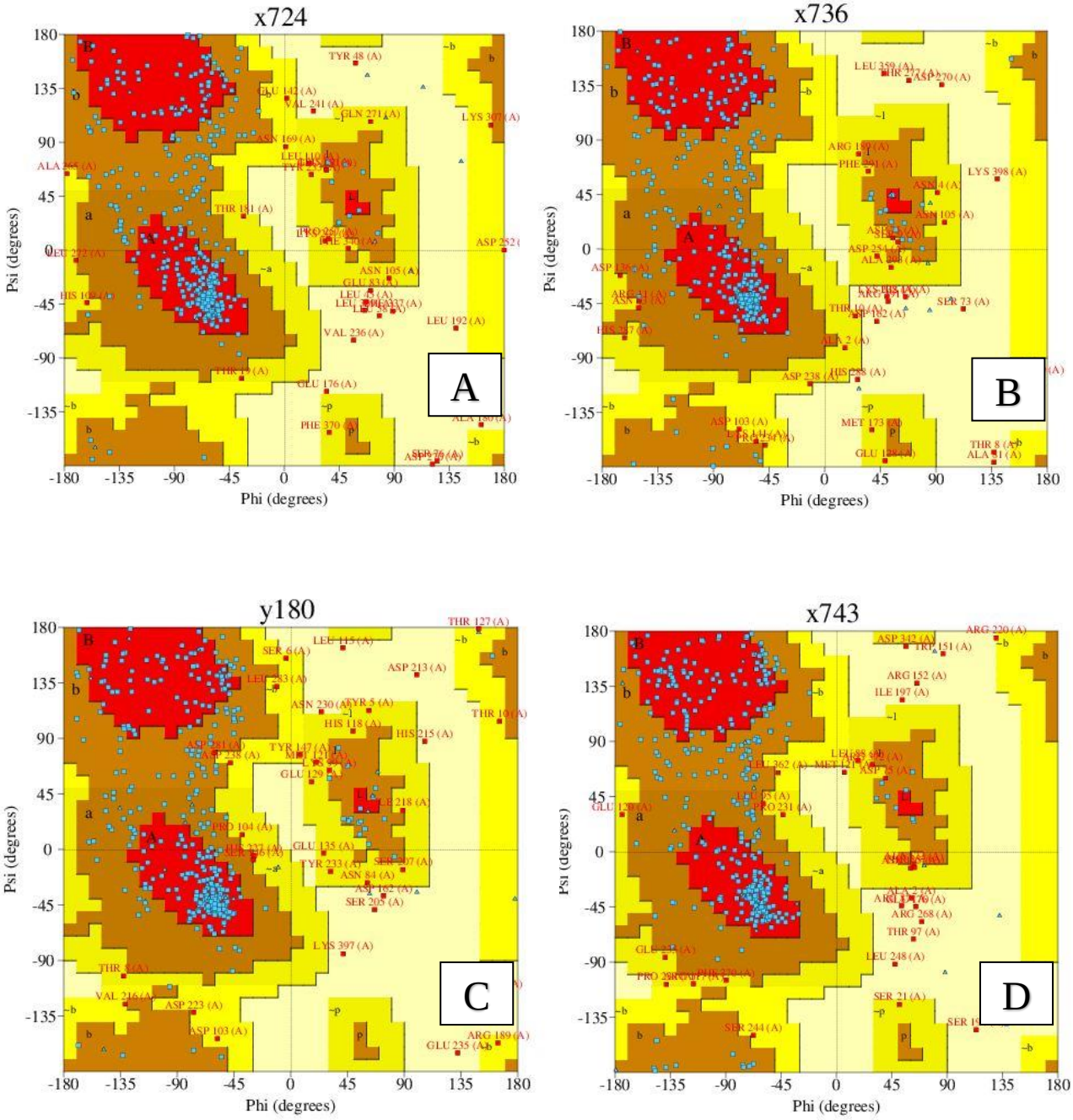


Figure 6.6: Ramachandran plots of models from I-TASSER produced by PROCHECK. (A) Model 1, (B) Model 2, (C) Model 3, (D) Model 4, (E) Model 5

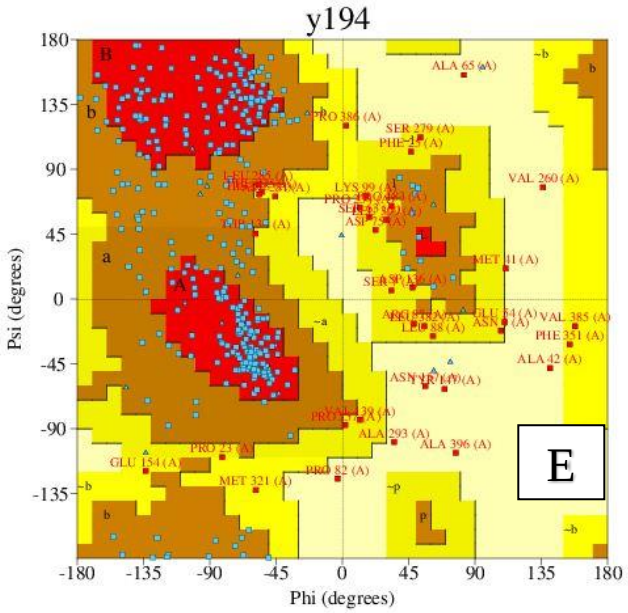


Figure 6.6 (continued): Ramachandran plots of models from I-TASSER produced by PROCHECK. (A) Model 1, (B) Model 2, (C) Model 3, (D) Model 4, (E) Model 5

The RMS distance values of the different planar groups in the structure were also considered to see if the amino acids are in the ideal position in the structure. The histograms (Figure 6.7) generated for each model showed the RMS distances from planarity for the different planar groups in the structure. The dashed lines indicate different ideal values for aromatic rings (Phe, Tyr, Trp, His) and for planar end-groups (Arg, Asn, Asp, Gln, Glu). The default values are 0.03Å and 0.02Å, respectively.

Model 1 had almost all the amino acids in the correct position except for glutamate which showed higher deviation from the optimum value, while most of the aromatic amino acids of model 2 crossed the ideal value. Thus the frequent red bars. Comparing model 3 and model 4, it was observed that both models had some amino acids out of their ideal position but model 4 had fewer abnormalities in the amino acid position. On the other hand, in model 5 the position of planar end-groups mostly (Asp and Glu) showed higher abnormalities than the aromatic rings and the frequency of amino acids appearing in an unconventional position was moderate.

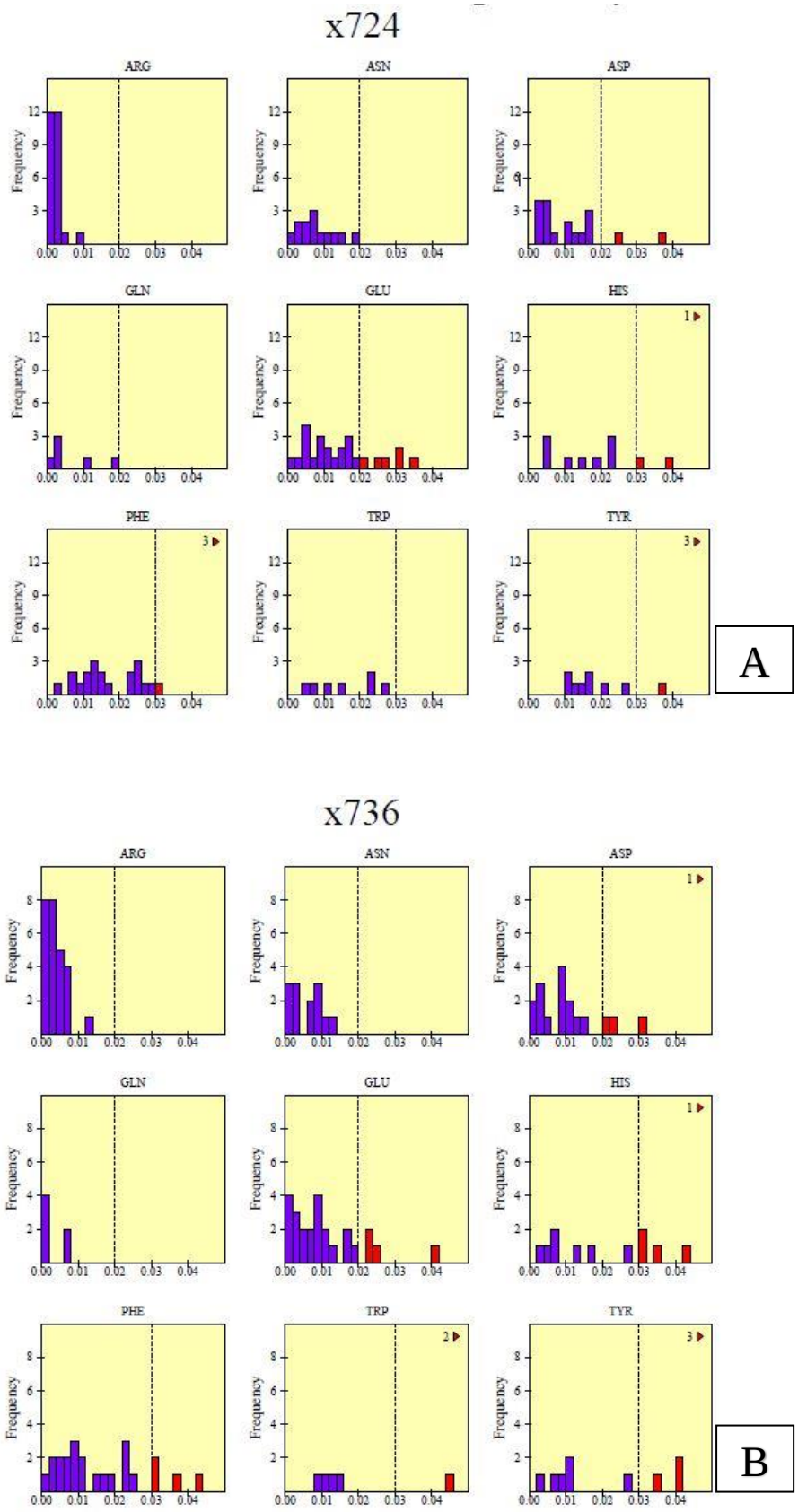
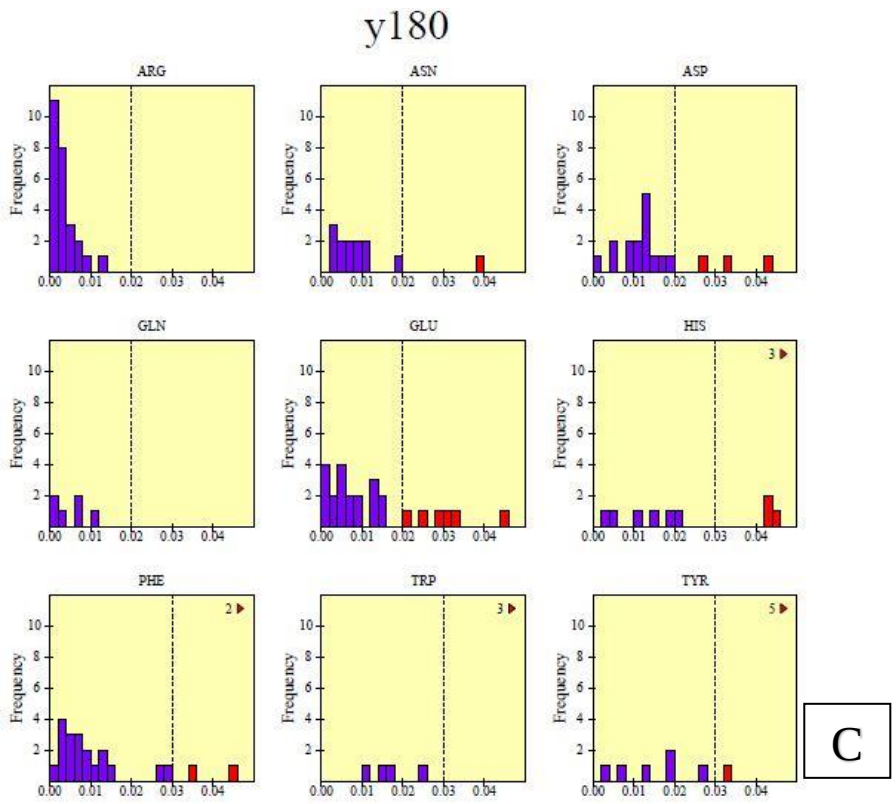


Figure 6.7: Histograms showing RMS distances of planar atoms from best-fit plane (A) Model 1, (B) Model 2, (C) Model 3, (D) Model 4, (E) Model 5



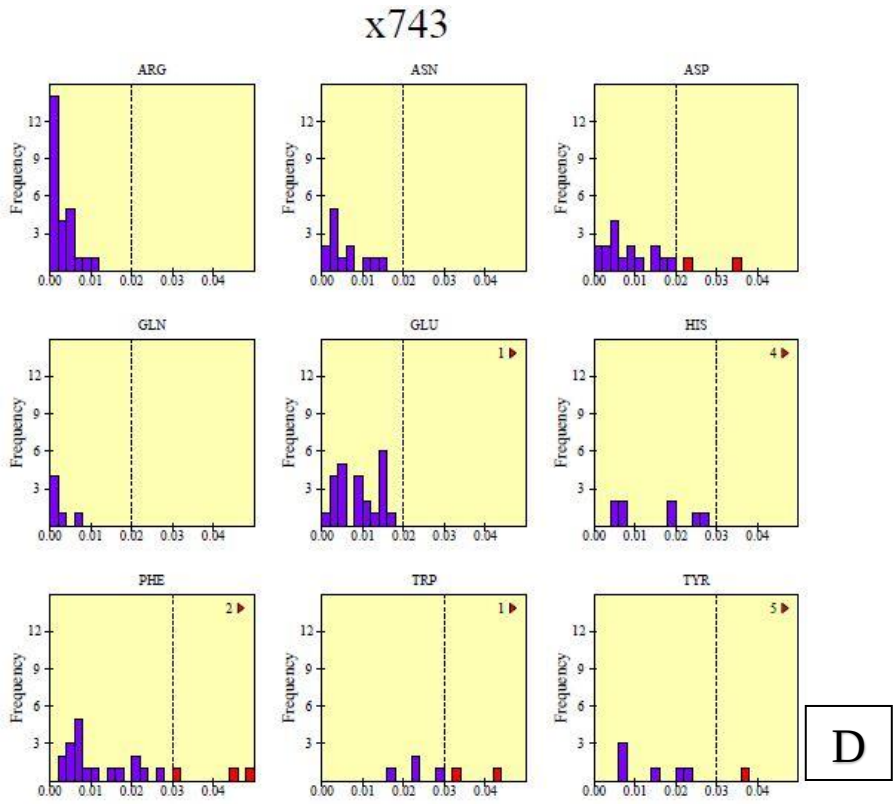


Figure 6.7 (continued): Histograms showing RMS distances of planar atoms from best-fit plane (A) Model 1, (B) Model 2, (C) Model 3, (D) Model 4, (E) Model 5

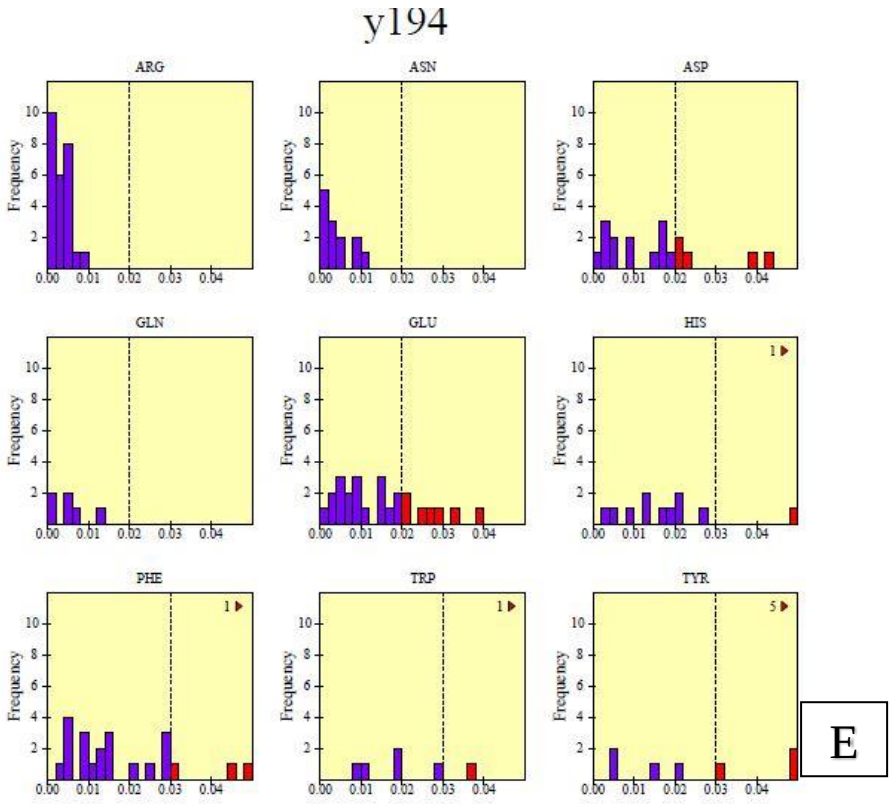


Figure 6.7 (continued): Histograms showing RMS distances of planar atoms from best-fit plane (A) Model 1, (B) Model 2, (C) Model 3, (D) Model 4, (E) Model 5

5.1.8.2 Selection of final model between models generated by I-TASSER and SwissModel

A final comparison was conducted between the I-TASSER model and the model from SwissModel (Table 2.0). Ramachandran plot for the model 1 from SwissModel was provided by PROCHECK (Figure 6.8)

The selected model from SwissModel (Model 1) indicated that 92.3% of the residues are in the most favored regions and 7.70% of the same are in the additional allowed regions. Surprisingly, none of the amino acids are in disallowed regions. These results revealed that the bulk of the amino acids are in the phi-psi distribution that is consistent with beta strands and right-handed alpha helix. It had a better G-Factor score of -0.70 than the model from I-TASSER (-0.75). All these results suggested that the model from SwissModel is more likely to be correct model. However, the Ramachandran plot analysis

of the model from SwissModel varied greatly compared to all the other five models from I-TASSER. More than 90% of the structure was in the most favored region with no disallowed region which was quite odd when compared to the rest of the models as the difference was very significant. The model generated by SwissModel seemed too accurate to be true and was considered unreliable. Thus, model 5 from I-TASSER was regarded as the most ideal model of the all the predicted models.

Validation		**I-TASSER (Model 5)**	**SwissModel (Model 1)**
PROCHECK Ramachandran Plot	Most favored regions	63.6%	92.3%
	Additional allowed regions	27.7%	7.70%
	Generously allowed regions	6.20%	0.00%
	Disallowed regions	2.50%	0.00%
G-Factor Overall Average		-0.75	-0.70

Table 2.0: Comparative values of PROCHECK, G-Factor between protein model from I-TASSER and SwissModel

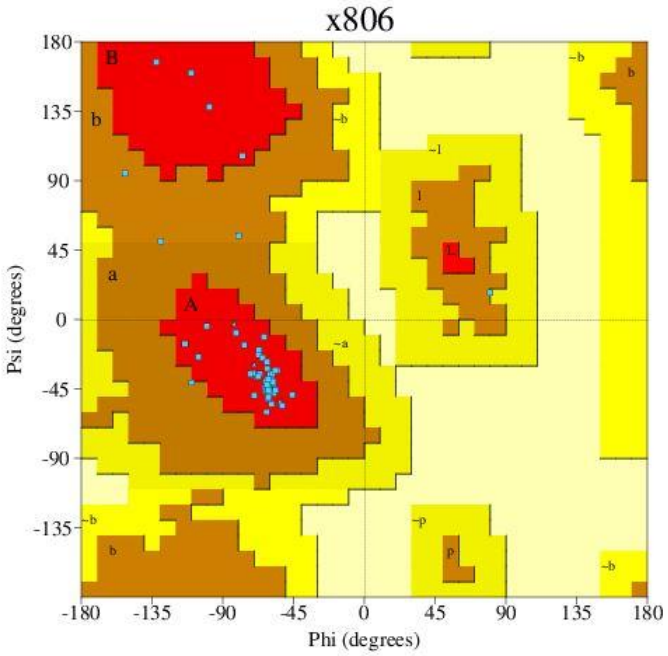


Figure 6.8: Ramachandran plot of model 1 from SwissModel produced by PROCHECK.

Chapter 6:

Discussion

6.1 Discussion

From the analysis of data and results obtained from this study, transmembrane Protein 43 (TMEM 43) was found to contain 400 amino acids with the accession ID: NP_077310.1.

The initial stages before constructing a molecular structure of the protein consisted of finding the homologous sequences of other species to the sequence of the desired protein. This is an important step because homologous sequences give us insight on the possible function of the protein and also may identify proteins with established three-dimensional structure that can serve as models for the structure of the protein of interest. The results allow prediction of the function from already identified functions of the homologous proteins, to discover evolutionary relationships or to find structural features. Following blast, the homologous sequences were selected based on E-value and identity values.

Both local alignment (blastp) and global alignment (Clustal omega) were used to find near and distant relationships of transmembrane Protein 43 with other species. Blastp found out the segments of the sequence that are conserved among other organism species, whereas Clustal omega enhanced this search and generated a more advanced result, displaying the similarities among the analogous sequences of amino acids of each of the compared proteins. The results from multiple sequence alignment can be used to deduce the protein domains that have similar functions as the protein sequences they were compared with. From the conservation results obtained from Clustal omega, it was observed that amino acids sequences of the desired protein are highly conserved with gap-free regions. This means that the protein mainly contains core regions with few or no surface loops.

Construction of a phylogenetic tree delivered a visual representation of how the different species selected by their homologous protein sequences are evolutionarily related. From the phylogenetic tree it was observed that the protein primarily formed three adjacent clades. One of these branches was exclusively for the protein from the species *Bos_inducus* (even toed unqulates) which has a different ancestral origin from all the other species. It was seen that TMEM 43 from *Homo_sapiens* is more closely related to the protein from *Pongo_abelii* (Sumatran orangutan) and shares a common ancestor. The

clade containing the species Ursus_maritimus (polar bear), Panthera_pardus (leopard), Equus_caballus (horse) are much more distantly related from *Homo_spaiens*.

Analysis of the characteristics of the protein revealed that the protein has a molecular weight of approximately 44876 Da. The amino acid composition of the protein showed the protein's primary structure consists mostly of leucine (~12%) which makes up the core of the protein. The hydrophobic nature of leucine explains why it is abundant in its core as it is a transmembrane protein and the core exposed to the inner nuclear membrane is also hydrophobic. On the other hand, low amounts of cysteine residues indicated the protein achieves its stability by other interactions as the chance of disulfide bond formation is very low. Furthermore, a negative GRAVY value indicated they are hydrophilic in nature. Lastly, the aliphatic index was quite high, thus they may remain stable over a wide range of temperatures.

Before deducing the structure of the protein, it was necessary to compute the number of transmembrane regions of the protein. Both ProtScale and TMHMM predicted the protein to have four transmembrane helices. These transmembrane domains are hydrophobic with the fourth domain being the most hydrophobic. The segment between the first and second transmembrane region is the longest region that is predicted to be natively unfolded and is most likely to reside in the endoplasmic reticulum.

The last step before deducing the 3D structure models was to compute the secondary structure of the protein. Prediction of the secondary structure of a protein from its primary structure is important to form protein structures. The results from SOPMA generated the secondary structure of the protein. The analysis revealed that α-helices are more prevalent (45.50%) amongst the secondary structures. This suggested that the protein gains its stability from the hydrogen bonds which are common in α-helices. Needless to mention, most transmembrane proteins contain higher a percentage of α-helices.

Both I-TASSER and SwissModel generated probable models of the structure. SwissModel generated two probable models of protein which were evaluated based on QMEAN scoring function. A higher QMEAN score indicates better quality of the predicted model, which is why model 1 from SwissModel is the most likely model for the query protein. However the ligand binding site for the sugar (B-OCTYLGLUCOSIDE) is

not conserved. On the other hand, I-TASSER reported up to five probable models. Of the five models, model 1 had the best C-score (-3.50) but C-score alone is not enough to assert that it is most accurate model. To predict the biological function of the protein, I-TASSER server matched the predicted 3D models to the proteins in three independent libraries which consisted of proteins of known enzyme classification (EC) number, gene ontology (GO), and ligand-binding sites. The final result of function prediction was deduced from the consensus of top structural matches with the function score calculated based on the confidence score of the I-TASSER structural models, the structural similarity between model and templates as evaluated by TM-score, and the sequence identity in the structurally aligned (Roy et al., 2010; Yang et al., 2015).

However, further analysis was done by PROCHECK to test the reliability of the models and to choose the best model generated by I-IASSER and SwissModel. Ramachandran plot analysis was performed and out of the five models from I-TASSER, model 5 appeared to have the best topology compared to any other model as it contained the most favored regions and least disallowed regions. When plot analysis of this model was compared to the result of model 1 from SwissModel the difference between the results was eye-catching. The results differed greatly from all the I-TASSER models. The structure seemed too correct to be true as it had over 90% correct topology with no disallowed regions. Thus, it was considered unreliable and the model 5 from I-TASSER was chosen to be the best-fit for the structure of transmembrane Protein 43.

Chapter 7:

Conclusion

7.1 Conclusion

A transmembrane protein of the inner nuclear membrane, known as transmembrane protein 43 (TMEM 43), was selected for research in this project. The aim behind choosing this protein was to work on a comparatively novel protein whose structure is yet to be determined. Using bioinformatics the process of these were made labor efficient and time efficient. The experiment was limited to only analysis of a very specific protein, the amino acid for which was already discovered. Transmembrane protein 43 has an important role in maintaining the nuclear envelope structure by organizing protein complexes at the inner nuclear membrane. Interestingly, the remarkably highly conserved protein LUMA is not expressed in the skeletal muscle. The protein has four transmembrane domains that spans the inner nuclear membrane. Although the functional significance of these domains in TMEM 43 is still unknown, it was confirmed that the p.S358L mutation that causes ARVC type 5 occurs within the third of the protein's four transmembrane spanning domains and is predicted to disrupt the transmembrane helix (Siragam et al., 2014). On the other hand, according to (Bengtsson & Otto, 2007) protein LUMA dysfunction could also be involved in the molecular mechanism leading to muscular dystrophy (EDMD). Thus it has become imperative to invest time in researching about the protein and possibly develop ways to treat or manage these genetic conditions. Looking to the future, the functions and characteristics of this protein still needs to be validated by further wet lab research.

Bibliography

1. Adamian, L., & Liang, J. (2002). Interhelical hydrogen bonds and spatial motifs in membrane proteins: Polar clamps and serine zippers. *Proteins: Structure, Function and Genetics*, *47*(2), 209–218. https://doi.org/10.1002/prot.10071
2. Ahram, M., Litou, Z. I., Fang, R., & Al-Tawallbeh, G. (2006). Estimation of membrane proteins in the human proteome. *In Silico Biol*, *6*(5), 379–386. https://doi.org/2006060036 [pii]
3. Aloy, P., & Russell, R. B. (2006). Structural systems biology: modelling protein interactions. *Nat Rev Mol Cell Biol*, *7*(3), 188–197. https://doi.org/10.1038/nrm1859
4. Arnold, K., Bordoli, L., Kopp, J., & Schwede, T. (2006). The SWISS-MODEL workspace: A web-based environment for protein structure homology modelling. *Bioinformatics*, *22*(2), 195–201. https://doi.org/10.1093/bioinformatics/bti770
5. Bengtsson, L., & Otto, H. (2007). LUMA interacts with emerin and influences its distribution at the inner nuclear membrane. *Journal of Cell Science*, *4*(124), 538–546. https://doi.org/10.1242/jcs.019281
6. Benkert, P., Biasini, M., & Schwede, T. (2011). Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics*, *27*(3), 343–350. https://doi.org/10.1093/bioinformatics/btq662
7. Berk, J. M., Tifft, K. E., & Wilson, K. L. (2013). The nuclear envelope LEM-domain protein emerin. *Nucleus*, *4*(4), 298–314. https://doi.org/10.4161/nucl.25751
8. Berman, H. M. (2000). Westbrook, J. Feng, Z. Gilliland, G. Bhat, T. N. Weissig, H. Shindyalov, I. N. Bourne, P. E. The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242.
9. Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., … Schwede, T. (2014). SWISS-MODEL: Modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Research*, *42*(W1). https://doi.org/10.1093/nar/gku340
10. Curran, A. R., & Engelman, D. M. (2003). Sequence motifs, polar interactions and conformational changes in helical membrane proteins. *Current Opinion in Structural Biology*. https://doi.org/10.1016/S0959-440X(03)00102-7

11. Dreger, M., Bengtsson, L., Schoneberg, T., Otto, H., & Hucho, F. (2001). Nuclear envelope proteomics: novel integral membrane proteins of the inner nuclear membrane. *Proc Natl Acad Sci U S A*, *98*(21), 11943–11948. https://doi.org/10.1073/pnas.211201898
12. Eisenberg, D. (2003). The discovery of the alpha-helix and beta-sheet, the principal structural features of proteins. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(20), 11207–10. https://doi.org/10.1073/pnas.2034522100
13. Ellenberg, J., Siggia, E. D., Moreira, J. E., Smith, C. L., Presley, J. F., Worman, H. J., & Lippincott-Schwartz, J. (1997). Nuclear membrane dynamics and reassembly in living cells: targeting of an inner nuclear membrane protein in interphase and mitosis. *J Cell Biol*, *138*, 1193–1206. Retrieved from file:///C:/PDF/Ellenberg1997.pdf%5Cnhttp://www.jcb.org/cgi/content/abstract/138/6/1193
14. Elofsson, A., & von Heijne, G. (2007). Membrane protein structure: prediction versus reality. *Annual Review of Biochemistry*, *76*, 125–140. https://doi.org/10.1146/annurev.biochem.76.052705.163539
15. Foisner, R., & Gerace, L. (1993). Integral membrane proteins of the nuclear envelope interact with lamins and chromosomes, and binding is modulated by mitotic phosphorylation. *Cell*, *73*(7), 1267–1279. https://doi.org/10.1016/0092-8674(93)90355-T
16. Franke, W. W., Dörflinger, Y., Kuhn, C., Zimbelmann, R., Winter-Simanowski, S., Frey, N., & Heid, H. (2014). Protein LUMA is a cytoplasmic plaque constituent of various epithelial adherens junctions and composite junctions of myocardial intercalated disks: A unifying finding for cell biology and cardiology. *Cell and Tissue Research*, *357*(1), 159–172. https://doi.org/10.1007/s00441-014-1865-1
17. Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., & Bairoch, A. (2005). Protein Identification and Analysis Tools on the ExPASy Server. In *The Proteomics Protocols Handbook* (pp. 571–607). https://doi.org/10.1385/1592598900
18. Geourjon, C., & Deléage, G. (1995). Sopma: Significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Bioinformatics*, *11*(6), 681–684. https://doi.org/10.1093/bioinformatics/11.6.681

19. Gruenbaum, Y., Margalit, A., Goldman, R. D., Shumaker, D. K., & Wilson, K. L. (2005). The nuclear lamina comes of age. *Nat Rev Mol Cell Biol*, *6*(1), 21–31. https://doi.org/nrm1550 [pii]\r10.1038/nrm1550
20. Guex, N., Peitsch, M. C., & Schwede, T. (2009). Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective. *Electrophoresis*, *30*(SUPPL. 1). https://doi.org/10.1002/elps.200900140
21. Kiefer, F., Arnold, K., Künzli, M., Bordoli, L., & Schwede, T. (2009). The SWISS-MODEL Repository and associated resources. *Nucleic Acids Research*, *37*(SUPPL. 1). https://doi.org/10.1093/nar/gkn750
22. Kyte, J., & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, *157*(1), 105–132. https://doi.org/10.1016/0022-2836(82)90515-0
23. Lusk, C. P., Blobel, G., & King, M. C. (2007). Highway to the inner nuclear membrane: rules for the road. *Nature Reviews. Molecular Cell Biology*, *8*(5), 414–20. https://doi.org/10.1038/nrm2165
24. Merner, N. D., Hodgkinson, K. A., Haywood, A. F. M., Connors, S., French, V. M., Drenckhahn, J. D., … Young, T. L. (2008). Arrhythmogenic Right Ventricular Cardiomyopathy Type 5 Is a Fully Penetrant, Lethal Arrhythmic Disorder Caused by a Missense Mutation in the TMEM43 Gene. *American Journal of Human Genetics*, *82*(4), 809–821. https://doi.org/10.1016/j.ajhg.2008.01.010
25. Müller, D. J., Wu, N., & Palczewski, K. (2008). Vertebrate membrane proteins: structure, function, and insights from biophysical approaches. *Pharmacological Reviews*, *60*(1), 43–78. https://doi.org/10.1124/pr.107.07111
26. Ohba, T., Schirmer, E. C., Nishimoto, T., & Gerace, L. (2004). Energy- and temperature-dependent transport of integral proteins to the inner nuclear membrane via the nuclear pore. *Journal of Cell Biology*, *167*(6), 1051–1062. https://doi.org/10.1083/jcb.200409149
27. Östlund, C., Sullivan, T., Stewart, C. L., & Worman, H. J. (2006). Dependence of diffusional mobility of integral inner nuclear membrane proteins on A-type lamins. *Biochemistry*, *45*(5), 1374–1382. https://doi.org/10.1021/bi052156n
28. Rajkumar, R., Sembrat, J. C., McDonough, B., Seidman, C. E., & Ahmad, F. (2012). Functional effects of the TMEM43 Ser358Leu mutation in the pathogenesis of arrhythmogenic right ventricular cardiomyopathy. *BMC Medical Genetics*, *13*(1), 21. https://doi.org/10.1186/1471-2350-13-21

29. Ramachandran, S., & Dokholyan, N. V. (2012). Homology Modeling: Generating Structural Models to Understand Protein Function and Mechanism. In *Computational Modeling of Biological Systems* (pp. 97–116). https://doi.org/10.1007/978-1-4614-2146-7_5
30. Remm, M., & Sonnhammer, E. (2000). Classification of transmembrane protein families in the Caenorhabditis elegans genome and identification of human orthologs. *Genome Research*, *10*(11), 1679–1689. https://doi.org/10.1101/gr.GR-1491R
31. Roy, A., Kucukural, A., & Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols*, *5*(4), 725–738. https://doi.org/10.1038/nprot.2010.5
32. Šali, A., & Blundell, T. L. (1993). Comparative Protein Modelling by Satisfaction of Spatial Restraints. *Journal of Molecular Biology*, *234*(3), 779–815. https://doi.org/10.1006/jmbi.1993.1626
33. Schirmer, E. C., Florens, L., Guan, T., Yates, J. R., & Gerace, L. (2003). Nuclear membrane proteins with potential disease links found by subtractive proteomics. *Science (New York, N.Y.)*, *301*(5638), 1380–2. https://doi.org/10.1126/science.1088176
34. Schirmer, E. C., & Foisner, R. (2007). Proteins that associate with lamins: Many faces, many functions. *Experimental Cell Research*. https://doi.org/10.1016/j.yexcr.2007.03.012
35. Schmidt, A., & Lamzin, V. S. (2002). Veni, vidi, vici - Atomic resolution unravelling the mysteries of protein function. *Current Opinion in Structural Biology*. https://doi.org/10.1016/S0959-440X(02)00394-9
36. Senes, A., Ubarretxena-Belandia, I., & Engelman, D. M. (2001). The Calpha ---H...O hydrogen bond: a determinant of stability and specificity in transmembrane helix interactions. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(16), 9056. https://doi.org/10.1073/pnas.161280798
37. Siragam, V., Cui, X., Masse, S., Ackerley, C., Aafaqi, S., Strandberg, L., … Hamilton, R. M. (2014). TMEM43 mutation p.s358L alters intercalated disc protein expression and reduces conduction velocity in arrhythmogenic right ventricular cardiomyopathy. *PLoS ONE*, *9*(10). https://doi.org/10.1371/journal.pone.0109128

38. Toyofuku, T., Akamatsu, Y., Zhang, H., Kuzuya, T., Tada, M., & Hori, M. (2001). c-Src regulates the interaction between connexin-43 and ZO-1 in cardiac myocytes. *Journal of Biological Chemistry*, *276*(3), 1780–1788. https://doi.org/10.1074/jbc.M005826200
39. Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., & Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction. *Nat Meth*, *12*(1), 7–8. https://doi.org/10.1038/nmeth.3213\rhttp://www.nature.com/nmeth/journal/v12/n1/abs/nmeth.3213.html#supplementary-information