Project Report On

In Silico Structural and Functional Annotation of Hypothetical Protein from *Mycobacterium Tuberculosis*

A Project Paper, submitted to the Department of Pharmacy, Daffodil International University to complete the course of B.Pharm.

Submitted To

Department of pharmacy

Faculty of Allied Health Science

Daffodil International University

Submitted By

ID: 191-29-1438

Batch: 21B

Department of Pharmacy

Daffodil International University

Date of Submission: April, 2023

APPROVAL

Both the presentation and the subject matter of the project In silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis, which was submitted to the Department of Pharmacy, Faculty of Allied Health Sciences, Daffodil International University, have received praise. It has been acknowledged as meeting a portion of the criteria for the Bachelor of Pharmacy degree.

Board of Examiners:

………………………………………..

Professor Dr. Muniruddin Ahmed

Head of the department of Pharmacy

Faculty of Allied Health Science

Daffodil International University

…………………………….	**Internal Examiner-I**
…………………………….	**Internal Examiner-II**
…………………………….	**Internal Examiner-III**

CERTIFICATE

In order to prove that the research findings contained in this project are unique and have never previously been submitted in their whole for a degree from this institution, we thus testify to their originality. The foundation for the whole current work, which has been provided as a project work for the degree of Bachelor of Pharmacy, is the author's (Mayanur Islam Nila, ID: 191-29-1438) individual research results.

Supervised By:

Mohammad Touhidul Islam

Lecturer (Senior Scale)

Department of Pharmacy

Faculty of Allied Health Sciences

Daffodil International University

DECLARATION

I hereby declare that, this project report is done by me under the supervision of **Mohammad Touhidul Islam,** Lecturer (Senior Scale), Department of Pharmacy, Daffodil International University, impartial fulfillment of the requirements for the degree of Bachelor of Pharmacy. I am declaring that this project is my original work. I also declare that neither this project nor any part therefore has been submitted elsewhere for the award of Bachelor or any degree.

Submitted By:

Name: Mayanur Islam Nila

ID: 191-29-1438

Batch: 21

DEDICATION

I would like to dedicate my work to the

Almighty GOD, My Beloved Parents

And

My supervisor.

ACKNOWLEDGEMENT

As a human being, I want to begin by thanking God for giving me the opportunity to complete this assignment successfully.

I want to thank my supervisor, Mr. Mohammad Touhidul Islam, Lecturer (Senior Scale), Department of Pharmacy, Daffodil International University, for his wonderful guidance in repairing this project and ensuring its success. His suggestions and guidance were really helpful in getting the job finished.

I would like to express my heartfelt thanks to Professor Dr. Muniruddin Ahamed, Head of the Pharmacy Department at Daffodil International University, for providing me the opportunity to finish my project work.

I also want to thank the other faculty members from the pharmacy department of Daffodil International University for working with me so well.

Then, I'd want to thank my parents and friends for their assistance, counsel, and inspiration at different phases of the project's completion. Not least, I want to express my gratitude to my students for their constant support.

Abstract

One of the oldest illnesses known to affect humans, tuberculosis (TB), kills two million people annually and is still the main cause of death. However, Mycobacterium tuberculosis, which is prevalent in aerosol droplets, deposits on lung alveolar surfaces, causing TB to be largely a pulmonary illness. Despite the fact that TB may damage the bone, the neurological system, and several other organ systems, its primary symptom is lung illness. There are a number of potential consequences as the illness advances from this point , most of which are determined by the host immune system's response. Researchers are identifying targets in M. TB that will aid in the development of these urgently required anti-tubercular medications using the whole M. tuberculosis genome structure as well as innovative genetic and physiological approaches.

In order to comprehend the function that a putative protein (OHO18223.1) discovered in Mycobacterium TB plays in the organism, the current study aims to investigate the structural and functional evaluation of the protein. Using an in-silico technique, homology modeling was utilized to build the 3D structure, and P fam, Interpro, and Uniport were used to profile the functions. The putative protein's primary and secondary structures were examined, and they revealed that it was stable, confined within the cytoplasm, and included a significant number of random coils. The 3D models that the SWISS-MODEL server predicted are 62.17 percent comparable to the template that received the best score. The SAVES v6.0 server and Prosa-web were used to assess the 3D model quality. The protein model has remarkable quality, as shown by the Ramachandran plot-based overall quality assessment results from Prosa-web, Verify3D, and Procheck. According to the websites P fam, Interpro, and Uniport, the putative protein is a cytoplasmic protein that breaks down unwanted, damaged, and improperly folded intracellular proteins, hence preserving the quality of protein in the cell. Finally, we proposed that more research into the structure and function of the protein may result in brand-new therapies.

Keywords: Function prediction; Structure assessment; *Mycobacterium tuberculosis*; Hypothetical protein; Homology modeling; novel Drug.

Table of Content

Chapter One: Introduction

Chapter Two: Purpose of my study

Chapter Three: Materials and Methods

Chapter Four: Results

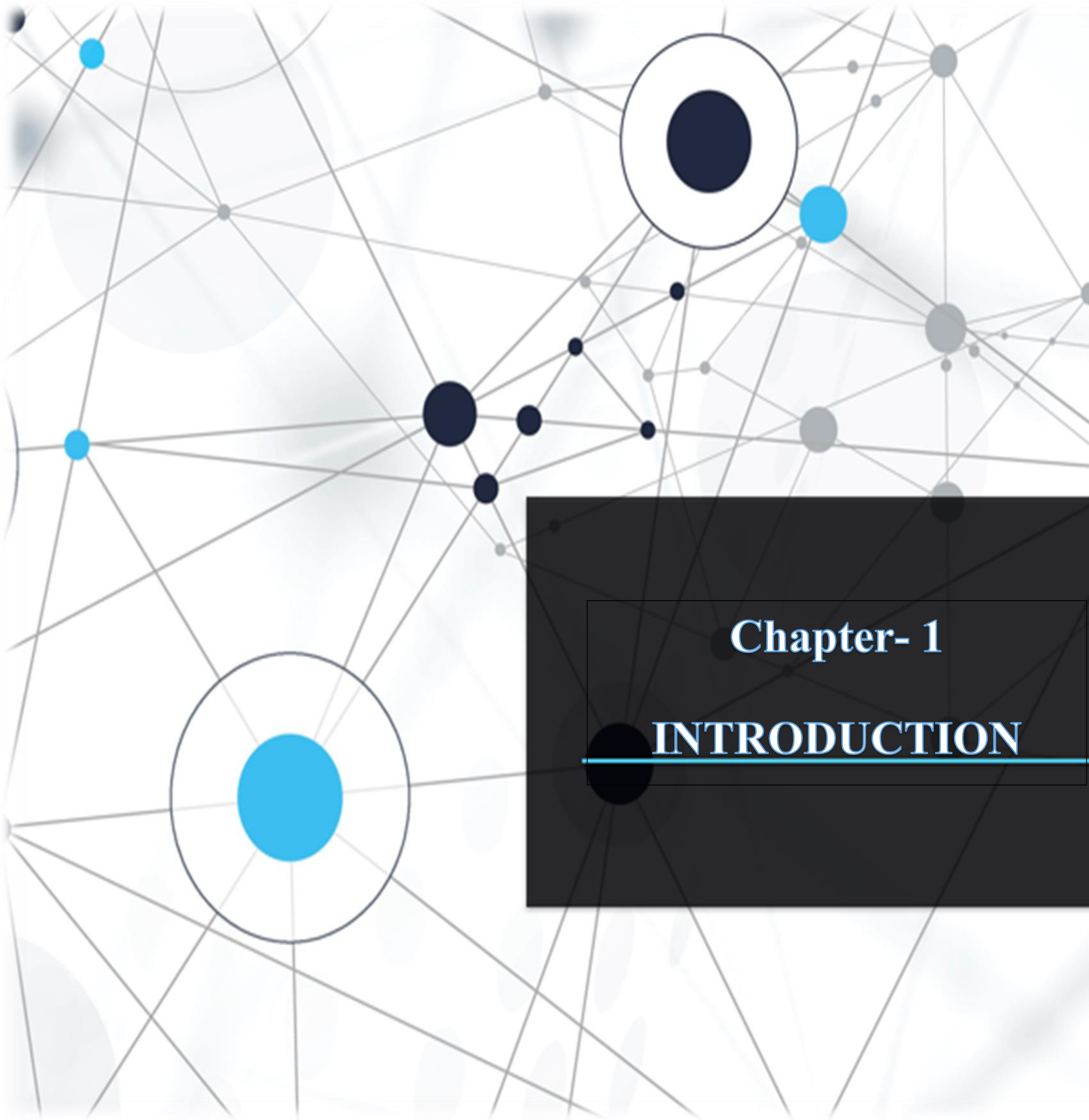
Chapter Five: Discussion

Chapter Six: Conclusion

Chapter Seven: References

List of Figure

List of Table



Chapter- 1

INTRODUCTION

Introduction

There is a lack of functional labelling for many recently decoded proteins despite the ever-increasing quantities of biological data, including raw data, such as genomic sequences, and functional genomic data from high-throughput studies. For instance, up to 40% of recognized proteins in the majority of bacterial genes are classified as "uncharacterized," "unknown," or "hypothetical" proteins [1]. HPs are predicted proteins, proteins that have not been demonstrated to exist by experimental protein chemistry data but are expected from nucleic acid patterns [2]. Scientists have used several computational techniques to develop several methods for forecasting protein function. These techniques have been made possible by utilizing series homology, phylogenetic analysis, interactions between proteins, interactions between proteins and ligands, similarities between active site residues, commonly conserved domains and motifs, phosphorylation sites, and patterns of gene expression. The conventional method of figuring out function, however, relies on sequence similarity and programs like BLAST, FASTA, and PSI-BLAST. There is no scientific or biological evidence for the existence of HPs, which are hypothetical proteins predicted from nucleic acid sequences. Moreover, these proteins share little in common with known, annotated proteins.

Only a small number of HPs have managed to endure over time, and they can be found in various evolutionary lineages. While HPs constitute a considerable portion of genes in sequenced microbial genomes, their complete functional characterization and detailed protein chemistry remain incomplete [3]. In HPs, there are two categories. Uncharacterized protein families (UPFs) make up one class, whereas domains of uncertain function make up the other (DUFs). Although their experimental structures have not been described or connected to any known genes, unknown proteins do indeed exist. Despite being proteins that have been experimentally identified, DUFs lack any structural or functional domains. They might have transmembrane sections or coiled-coil features that prevent the function from being assigned. Studying the function of proteins with unidentified roles has several advantages, including the evaluation of novel domains and motifs, the identification of extra protein pathways and cascades, and the finding of novel conformational orientations of 3-dimensional structures. Potential pharmacological targets may be discovered in these novel domains. Moreover, high throughput techniques, such as systematic synthetic lethal analysis, are employed, both microarray gene expression patterns and protein complex identification by mass spectrometry are useful. You can also use phylogenetic analysis of proteins across numerous genes to determine function. Genes with comparable roles are apt to be co-expressed, which is why the wide, common technique of grouping gene-expression patterns [4] is so popular. Many protein domains engage in an organism's biochemical processes and have unclear roles; however, they may also be detrimental. The function of a protein can rarely alter due to changes such as insertions, deletions, and replacements. The study's primary aim is to find and categories a protein region with an unknown function using computational technologies. The 2020 Global Tuberculosis Survey revealed that there were 10 million new cases of tuberculosis (TB) in 2019. Out of these, 1.2 million were people who did not have HIV, while the remaining 2,08,000 were people who were HIV-positive. The BCG vaccine is the only one that has been developed to prevent tuberculosis, but it does not work very well in adults.

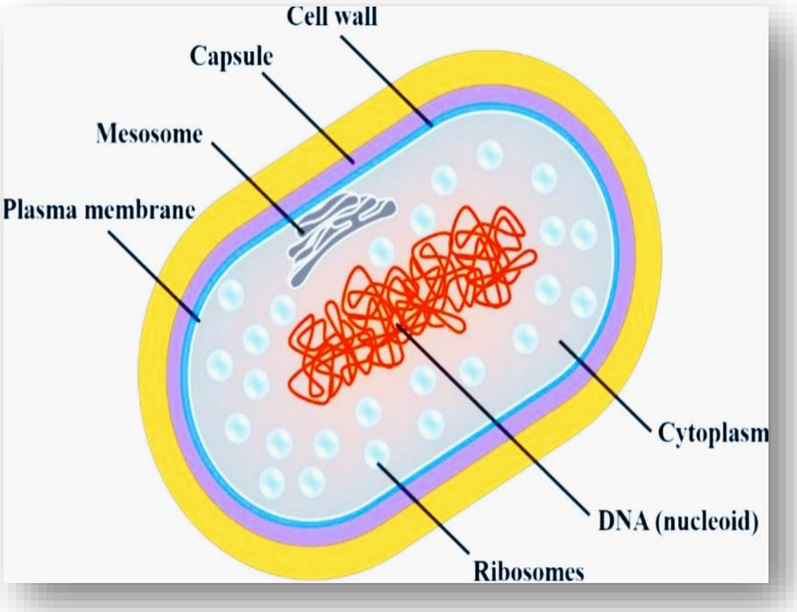
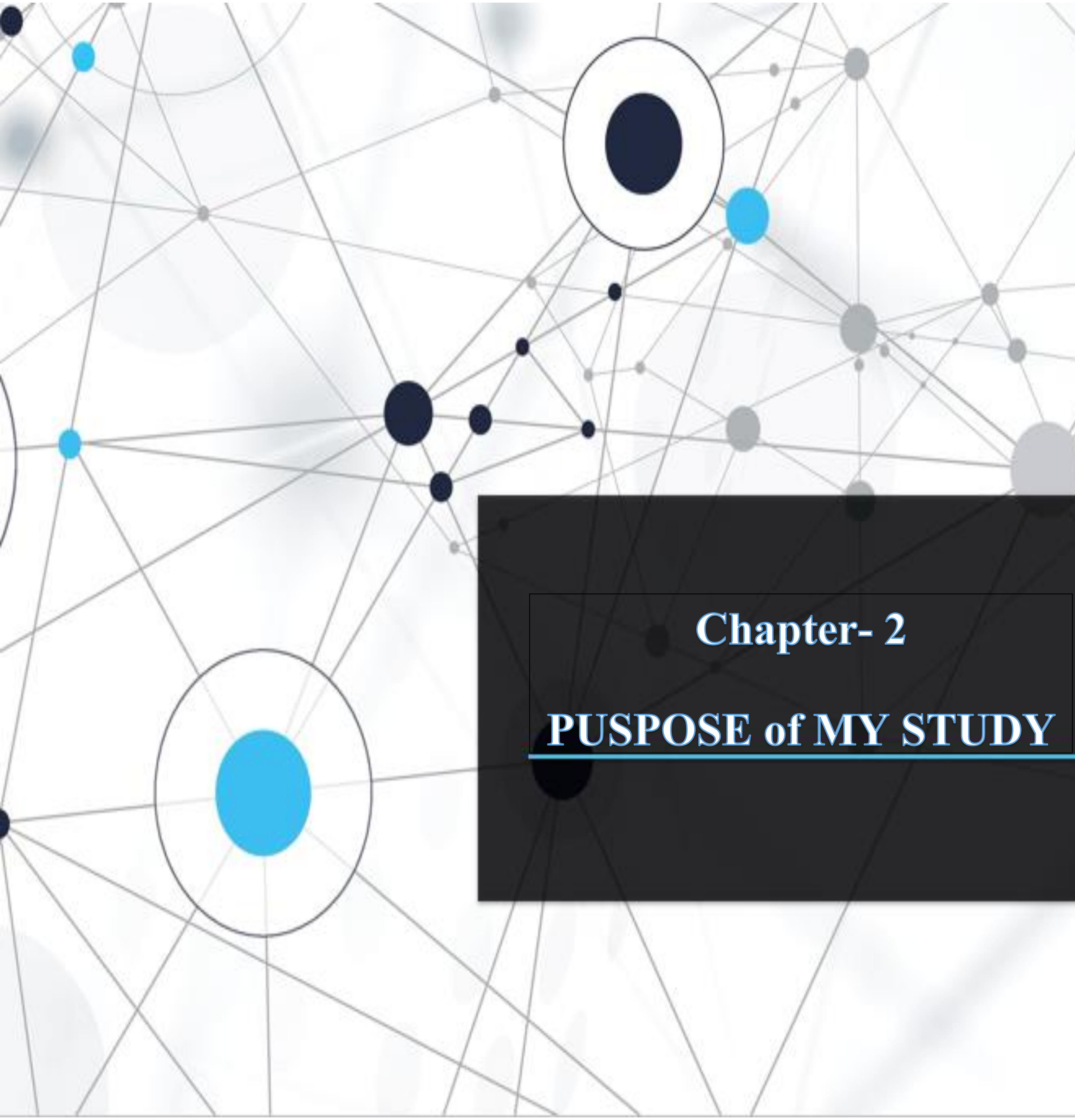


Figure 1.1 Cell structure of *Mycobacterium Tuberculosis* [8]

Despite the fact that many different immunizations are still in the testing stage, there is no evidence to suggest that any of them are particularly effective. Even though it was one of the most promising potential vaccines, MVA85A was unable to offer HIV patients in a clinical study any significant protection from the illness. These proteins are predicted to be essential for MTB survival and growth in their chosen environments as well as for mediating interactions between mycobacterium and host cells, according to some annotation predictions for these proteins, which are expressed in response to the shifting microenvironments that the pathogen encounters [5]. Because they are a part of a vastly expanded family of protein families, the functions of these proteins are essential for expanding our understanding of this species.

Moreover, Mycobacterium TB shows signs of treatment resistance, which makes the condition harder to treat [6]. More than half of M. tuberculosis's first complete genome, which was decoded in 1998, appears to contain potential proteins. Because the species has coevolved with other creatures and people, members of the M. TB complex have distinct harmful impacts despite having 99% of their DNA code. It has been explicitly hypothesised that a number of these "hypothetical" proteins, such as those in the ProGlu or Pro-Pro-Glu (PE/PPE) family, play a crucial role in the internal lifestyle of Mycobacterium TB and may aid in its survival under a variety of conditions. To anticipate the functions of many of the proteins that Mycobacterium TB is expected to generate, we created a functional interaction network for its proteins. In this research, we compared the network properties of hypothesised proteins to those of known proteins in order to identify statistically important classifications. Based on the expected molecular roles of these proteins, we also performed functional enrichment analysis on them. Numerous membrane-anchored genes implicated in lipid metabolism were found among the infectious mycobacterial genes that were examined. On the pathogenesis and pathogenicity of the bacteria, potential proteins that has high GC are believed to have a significant biochemical effect [7]. Expasy, Prot Param, NCBI CDD blast, P farm, Clustal omega, Sopma, Psipred, Procheck, SWISS-Model, Raptor X, Phyre 2, and ProSA -web server were among the computational tools that could be used to analyse the molecular and functional analysis of a new putative protein (BBW91 17915) in the current study. The outcomes of the in silico experiment were compared using three positive controls and three negative controls, respectively.



Chapter- 2

PUSPOSE of MY STUDY

Purpose of my study

The goal of my research in 2019, there were ten million newly diagnosed instances of tuberculosis (TB), among them about 1.2 million people tested HIV- negative and 2.08,000 tested positive, according to the Global Tuberculosis Report, 2020. BCG, the as-yet-uncreated anti-TB antibody, isn't particularly effective against tuberculosis in adults. Despite the fact that many antibodies are still in the testing phase, notable survival has not yet been described. Additionally, Mycobacterium TB looks resistant to frequently prescribed medications, making the disease more difficult to treat [9].

Finding new therapeutic targets for Mycobacterium tuberculosis is made easier with the help of computational drug discovery research. His investigation of the MTBC genome using SMRT sequencing reveals that the genome is 99% identical and well conserved. Only 7 secreted proteins have been described to date, despite the fact that 51 M. tuberculosis secreted proteins have been identified using computational techniques. Bioinformatics methods have been used to study various groups of hypothetical proteins, including enzymes, transporters, receptors, and structural proteins. Few in silico methods have been used to forecast vaccine potential, ribosome binding sites, GTP binding regions, or any other properties of M. TB proteins, but many M. tuberculosis proteins have received examination of mutations, useful research, and structural forecast. To create new, highly immunogenic multi-epitope subunit vaccines or drugs against tuberculosis, reverse vaccine trials are being done.

2.1 This study's objective is to:

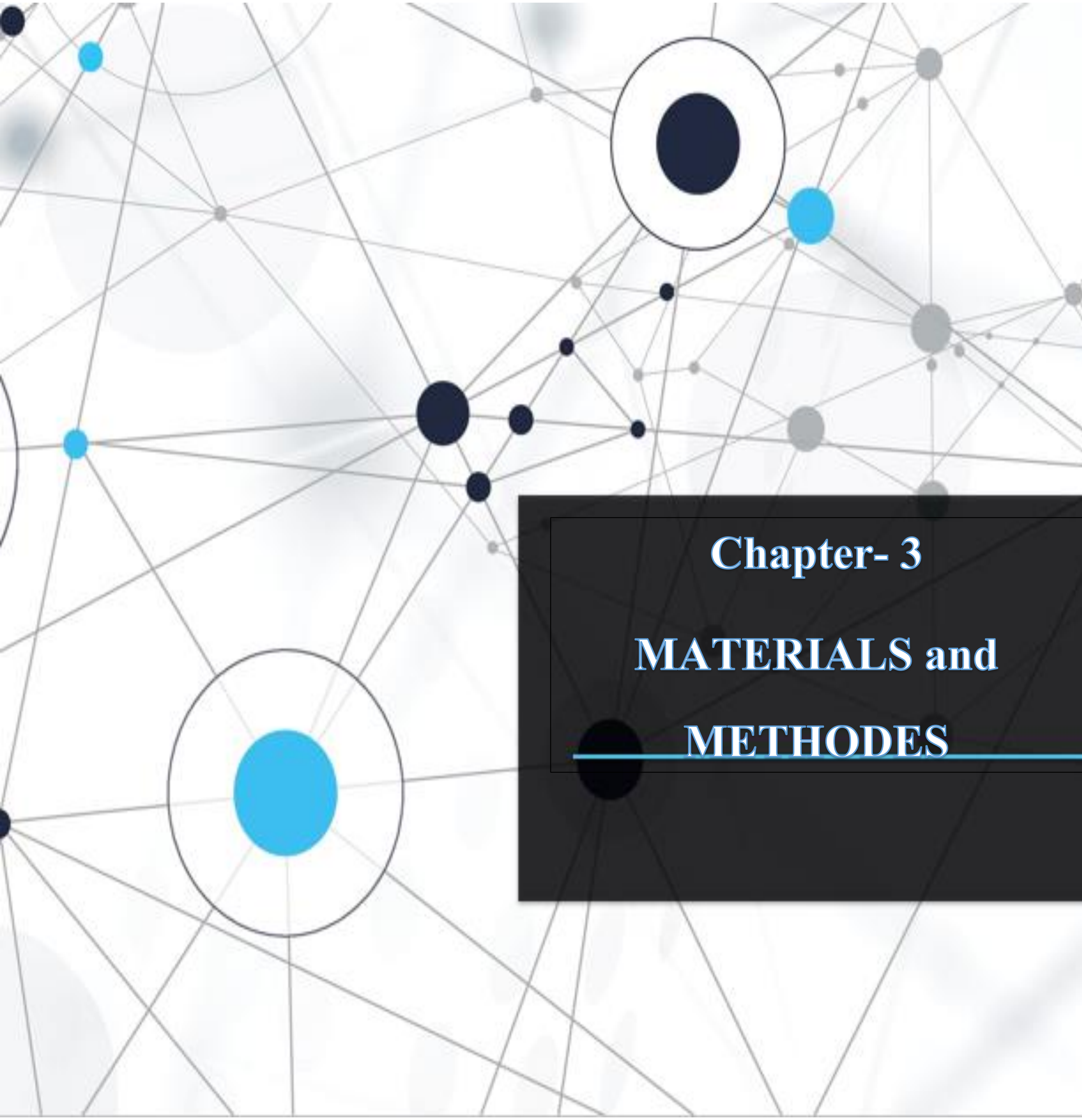
-Discovery of new biotechnologically significant proteins as well as novel drug design.

- Prediction of the structure of hypothetical proteins.

- Determination of hypothetical proteins' functional roles.

- Describe the physiochemical characteristics of hypothetical proteins.

The significance of this entire work lies in its ability to shed light on the structure and function of the targeted *Mycobacterium tuberculosis* proteins.



Chapter- 3

MATERIALS and METHODES

Materials and Methods

This inquiry involves two successive phases. Homology Search, Physiochemical Parameters, Subcellular Localization and Stability Prediction, Identification and Retrieval of Unique HPs from Internet Databases, were all included in PHASE I. PHASE II involved the structural and functional analysis of the selected proteins, which included identifying their interactions with other proteins, classifying protein families, looking for signal peptides, homology modeling, and validating the 3D structures of the proteins. A list of the bioinformatics sources and instruments used to functionally annotate Mycobacterium tuberculosis (Accession No. **OHO18223.1**) HPs can be found in Supplementary Table 1.1

Table 1.1 The datasets and bioinformatics tools used in this study's structural and functional investigation of the HP are listed below.

Tools/Servers	URL	Function
A)Sequence Retrieval- Download FASTA of Target Protein **1. NCBI**	https:// www.ncbi.nlm.nih.gov/	Sequence retrieval or selection of the hypothetical protein and download FASTA of target protein
B) Determination of physicochemical Properties **1. ExPASy – ProtParam**	http:// web.expasy.org/protparam	To analyze physicochemical characterization of Hypothetical Protein
C) Subcellular Localization **1. PSLpred** **2. SOSUI-GramN**	https:// webs.iiitd.edu.in/raghava/pslpred/submit.html https:// harrier.nagahama-i-bio.ac.jp/sosui/sosuigramn/sosuigramnsubmit.htm	Gram-negative bacteria's five primary subcellular localizations predicting.
D) Function Predictions **1. Conserved domain database** **2. INTERPRO**	http:// www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi	A method for looking for functional regions in a series and Family relationship identification.

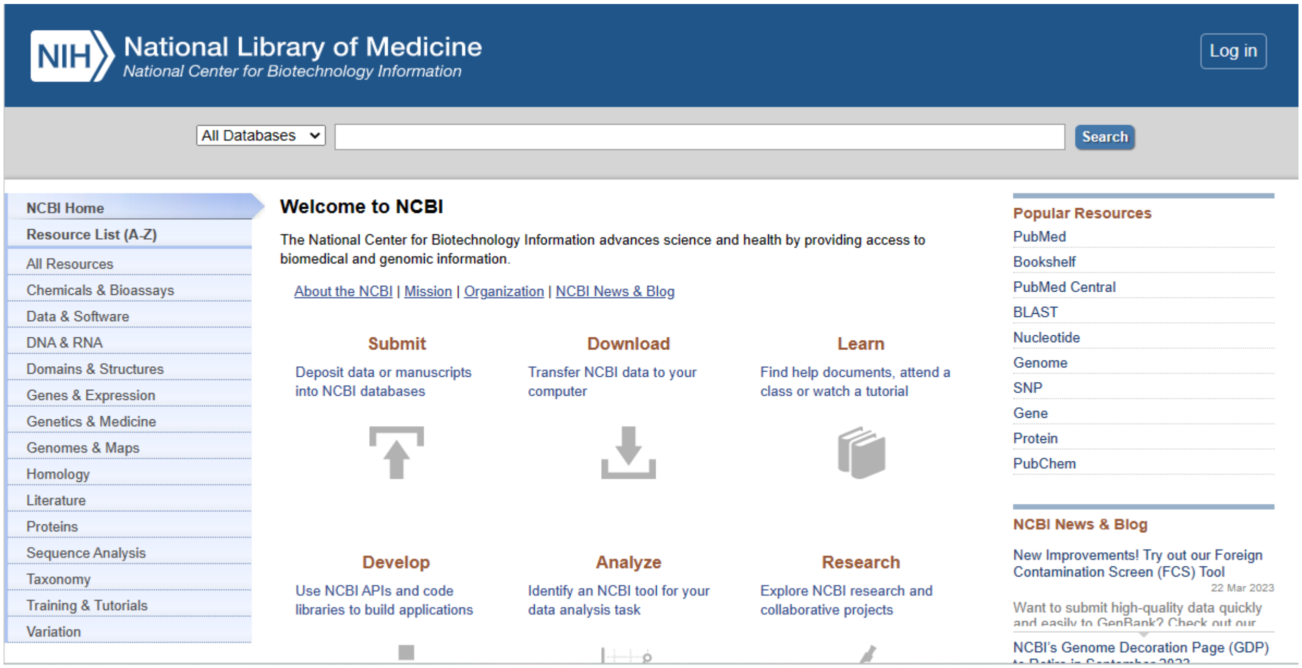
3. Pfam	http:// www.ebi.ac.uk/interpro/ http:// pfam.xfam.org/	
E) Homology Searching **1. PBLAST**	https:// blast.ncbi.nlm.nih.gov/Blast.cgi#	Searches for sequence similarity
F) Multiple Sequence Alignment and Phylogenetic Tree Prediction **1. Clustal Omega**	https:// www.ebi.ac.uk/Tools/services/web_clustalo/toolform.ebi	Accomplishes multiple sequence alignment and phylogenetic tree prediction
G) Secondary Structure Prediction **1. PSIPRED** **2. SOPMA**	http:// bioinf.cs.ucl.ac.uk/psipred https:// npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_sopma.html	Accurately forecasts amino acids for a three-state secondary structure description of the secondary structure.
H) Tertiary Structure Prediction **1. SWISS-Model** **2. RaptorX** **3. Phyre2**	https:// swissmodel.expasy.org/interactive/7hXdRN/models/ http:// raptorx6.uchicago.edu/ContactMap/ http:// www.sbg.bio.ic.ac.uk/~phyre2/html/page.cgi?id=index	Evaluate the model accuracy in the database, search for different blueprint designs and then dynamically build models.
I) 3DStructure Validation **1. PROCHECK** **2. ProSA-web** **3. Verify3D** **4. ERRAT**	https:// prosa.services.came.sbg.ac.at/prosa.php https:// saves.mbi.ucla.edu/	Validate using the structural analysis and verification service.

PHASE I

3.1 Sequence Retrieval or Selection of the Hypothetical Protein

Uncharacterized Mycobacterium TB proteins were selected from the NCBI protein database (https:// www.ncbi.nlm.nih.gov/) using the search term "hypothetical proteins." Using blast analysis, it was possible to compare the degree of resemblance to other alleged fictitious proteins.

In this study, a putative Mycobacterium tuberculosis protein with 255 amino acid residues, accession number **OHO18223.1**, was employed. For further research, the protein's main sequence in FASTA format was obtained.



Welcome to NCBI

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

Submit: Deposit data or manuscripts into NCBI databases

Download: Transfer NCBI data to your computer

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

Develop: Use NCBI APIs and code libraries to build applications

Analyze: Identify an NCBI tool for your data analysis task

Research: Explore NCBI research and collaborative projects

NCBI News & Blog

New Improvements! Try out our Foreign Contamination Screen (FCS) Tool

22 Mar 2023

Want to submit high-quality data quickly and easily to GenBank? Check out our

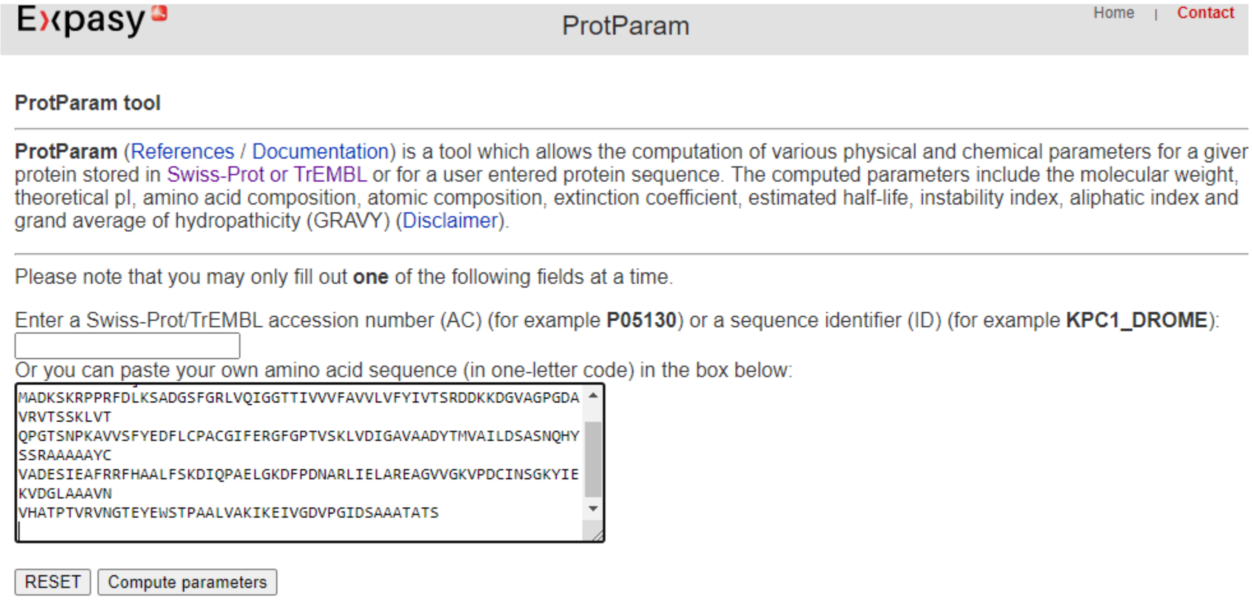
NCBI's Genome Decoration Page (GDP)

Figure 3.1.1: National Center for Biotechnology Information used for sequence retrieval

3.2 Physiochemical Properties

The Protparam server from Expasy was used to analyze physicochemical characterization (http:// us.expasy.org/tools/protparam).

Molecular weight, predicted pI, the ratio of amino acids to atoms, the extinction coefficient, the expected half-life, and the instability index, (which measures the integrity of the protein structure; a score of 40 indicates that the protein is stable), aliphatic index, and GRAVY parameters were calculated by ProtParam (if this score is negative, it is mine soluble in water; if it is positive, it is not soluble in water). To understand protein characteristics, use this resource.



ProtParam tool

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

Please note that you may only fill out **one** of the following fields at a time.

Enter a Swiss-Prot/TrEMBL accession number (AC) (for example **P05130**) or a sequence identifier (ID) (for example **KPC1_DROME**):

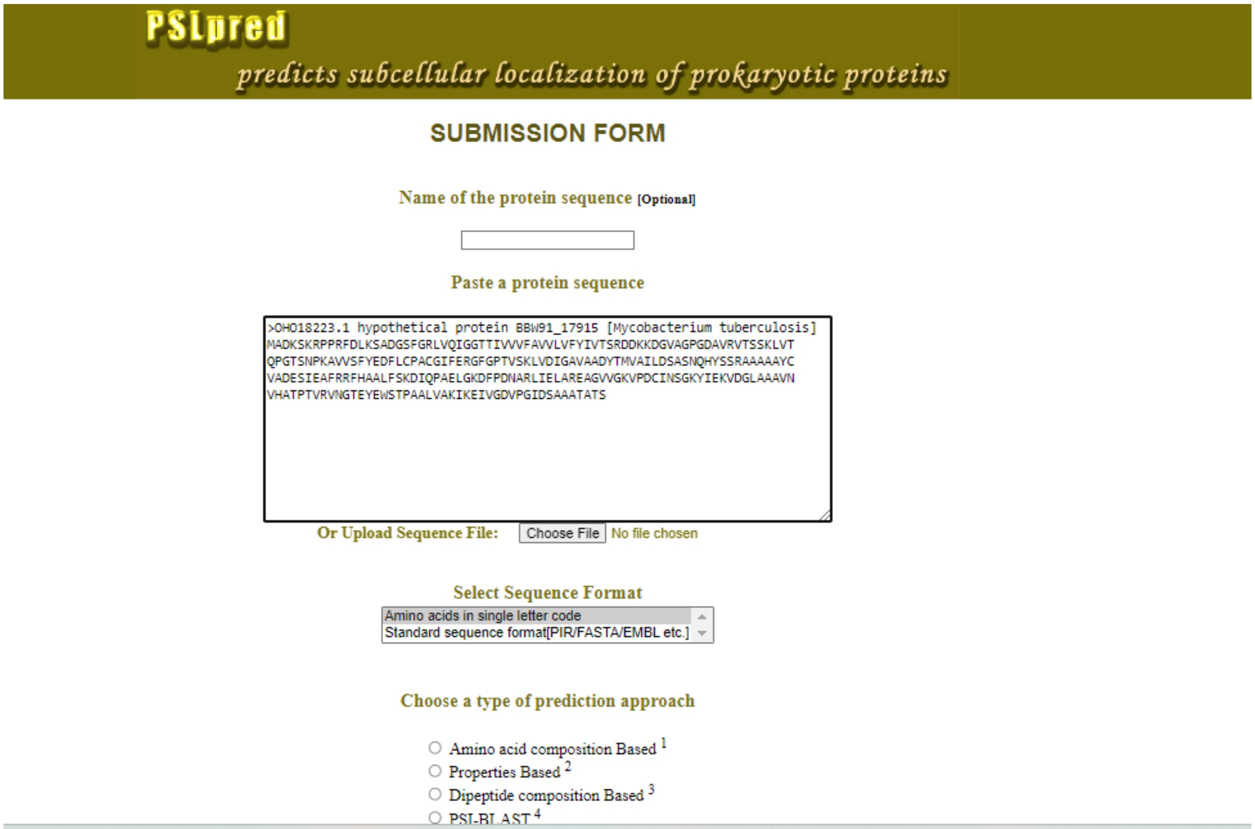
Or you can paste your own amino acid sequence (in one-letter code) in the box below:

```
MADKSKRPPRFDLKSADGSFGRLVQIGGTTIVVVFAVVLVFYIVTSRDDKKDGVAGPGDA
VRVTSSKLVT
QPGTSNPKAVVSFYEDFLCPACGIFERGFGPTVSKLVDIGAVAADYTMVAILDSASNQHY
SSRAAAAAYC
VADESIEAFRRFHAALFSKDIQPAELGKDFPDNARLIELAREAGVVGKVPDCINSGKYIE
KVDGLAAAVN
VHATPTVRVNGTEYEWSTPAALVAKIKEIVGDVPGIDSAAATATS
```

Figure 3.2.1 ExPASy's ProtParam tool

3.3 Subcellular Localization and Solubility Prediction

Gram-negative bacteria's five primary subcellular localizations are predicted using PSLpred (https:// webs.iiitd.edu.in/raghava/pslpred/submit.html), an SVM -based method (cytoplasm, inner membrane, outer membrane, extracellular, and periplasm). This method employs a variety of SVMs based on the composition of 33 physicochemical characteristics, PSI-Blast evolutionary data, and the dipeptide makeup, protein-repelling characteristics, and features of amino acid composition. The overall prediction accuracy for these SVM modules is 6%, 86%, 83%, and 68%, respectively. Moreover, a hybrid approach-based 3VM module has been developed based on all of the aforementioned characteristics of a protein. With a 91% total accuracy rate, the hybrid module outperforms all other methods currently in use for determining the intracellular location of bacterial proteins.



SUBMISSION FORM

Name of the protein sequence [Optional]

Paste a protein sequence

```
>OHO18223.1 hypothetical protein BBW91_17915 [Mycobacterium tuberculosis]
MADKSKRPPRFDLKSADGSFGRLVQIGGTTIVVVFAVVLVFYIVTSRDDKKDGVAGPGDAVRVTSSKLVT
QPGTSNPKAVVSFYEDFLCPACGIFERGFGPTVSKLVDIGAVAADYTMVAILDSASNQHYSSRAAAAAYC
VADESIEAFRRFHAALFSKDIQPAELGKDFPDNARLIELAREAGVVGKVPDCINSGKYIEKVDGLAAAVN
VHATPTVRVNGTEYEWSTPAALVAKIKEIVGDVPGIDSAAATATS
```

Or Upload Sequence File: Choose File No file chosen

Select Sequence Format

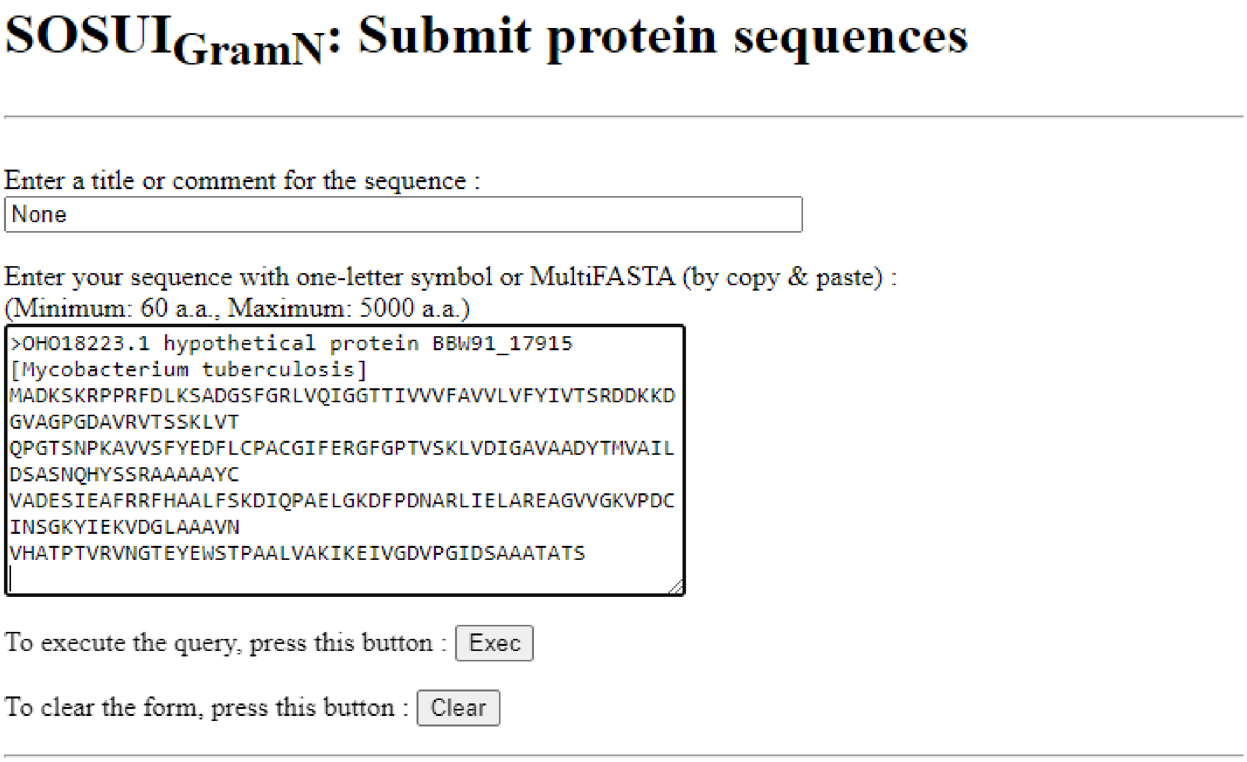
Amino acids in single letter code
Standard sequence format[PIR/FASTA/EMBL etc.]

Choose a type of prediction approach

○ Amino acid composition Based [1]
○ Properties Based [2]
○ Dipeptide composition Based [3]
○ PSI-BLAST [4]

Figure 3.3. 1 PSLpred subcellular localization predictor

Gram-negative bacteria's distribution of proteins within cells is predicted with high efficiency using SOSUI -GramN (https:// harrier.nagahama-i-bio.ac.jp/sosui/sosuigramn/sosuigramn submit.htm). The method only employs the molecular properties of the N- and C-terminal signal sequences as well as the complete sequence and does not require the sequence similarity information of any known sequences. For the localization of extracellular, outer, periplasmic, inner, and cytoplasmic mediums, the prediction system's accuracy was 92.3%, 89.4%, 86.4%, 97.5%, and 93.5%, respectively.



SOSUI_{GramN}: Submit protein sequences

Enter a title or comment for the sequence :

None

Enter your sequence with one-letter symbol or MultiFASTA (by copy & paste) :
(Minimum: 60 a.a., Maximum: 5000 a.a.)

```
>OHO18223.1 hypothetical protein BBW91_17915
[Mycobacterium tuberculosis]
MADKSKRPPRFDLKSADGSFGRLVQIGGTTIVVVFAVVLVFYIVTSRDDKKD
GVAGPGDAVRVTSSKLVT
QPGTSNPKAVVSFYEDFLCPACGIFERGFGPTVSKLVDIGAVAADYTMVAIL
DSASNQHYSSRAAAAAYC
VADESIEAFRRFHAALFSKDIQPAELGKDFPDNARLIELAREAGVVGKVPDC
INSGKYIEKVDGLAAAVN
VHATPTVRVNGTEYEWSTPAALVAKIKEIVGDVPGIDSAAATATS
```

To execute the query, press this button : Exec

To clear the form, press this button : Clear

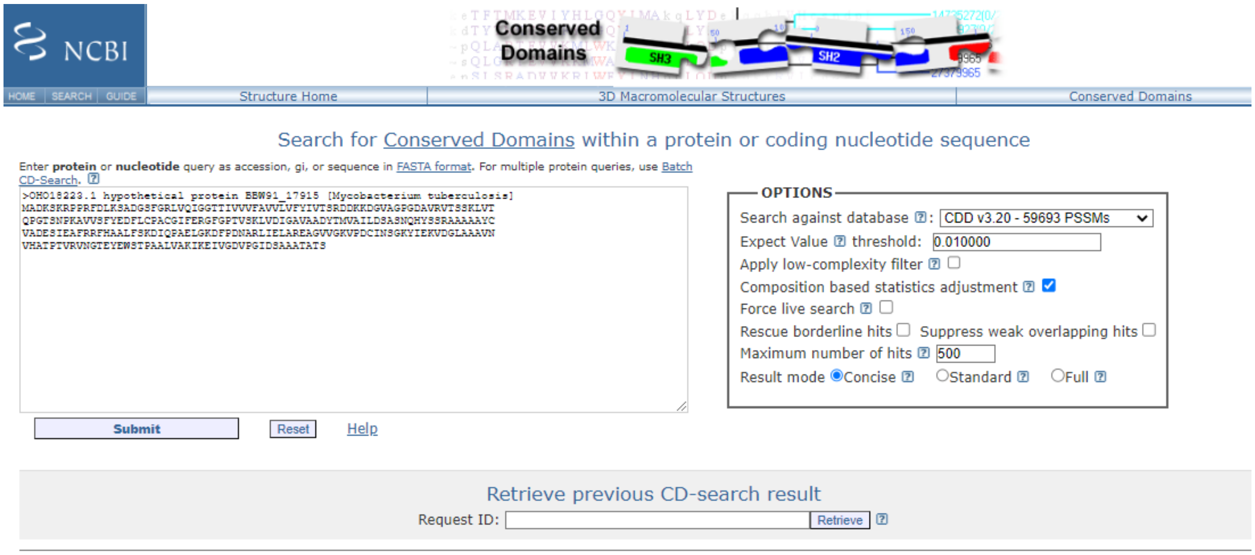
Figure 3.3. 2 SOSUI-GramN subcellular localization predictor

<u>PHASE- II</u>

3.4 Conserved domains search and Function prediction by Conserved domains analysis

In proteins' chemical development, protein domains can be grouped into an evolutionary taxonomy and viewed as units. This list is attempted to be assembled, and similar domain models are arranged in a hierarchical manner by the Conserved Domain Database (CDD) at NCBI.

The CDD program is used to do the conserved domain analysis, and it may be found online at (www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). P fam (http:// pfamlegacy.xfam.org/), Interpro Scan (http:// www.ebi.ac.uk/Tools/services/web/toolform.ebi?tool=iprscan5]), and other tools are used for domain analysis as well as Uniport (www.uniprot.org/blast).



Search for Conserved Domains within a protein or coding nucleotide sequence

Enter **protein** or **nucleotide** query as accession, gi, or sequence in FASTA format. For multiple protein queries, use Batch CD-Search.

```
>OHO18223.1 hypothetical protein BBW91_17915 [Mycobacterium tuberculosis]
MADKSKRPPRFDLKSADGSFGRLVQIGGTTIVVVFAVVLVFYIVTSRDDKKDGVAGPGDAVRVTSSKLVT
QPGTSNPKAVVSFYEDFLCPACGIFERGFGPTVSKLVDIGAVAADYTMVAILDSASNQHYSSRAAAAAYC
VADESIEAFRRFHAALFSKDIQPAELGKDFPDNARLIELAREAGVVGKVPDCINSGKYIEKVDGLAAAVN
VHATPTVRVNGTEYEWSTPAALVAKIKEIVGDVPGIDSAAATATS
```

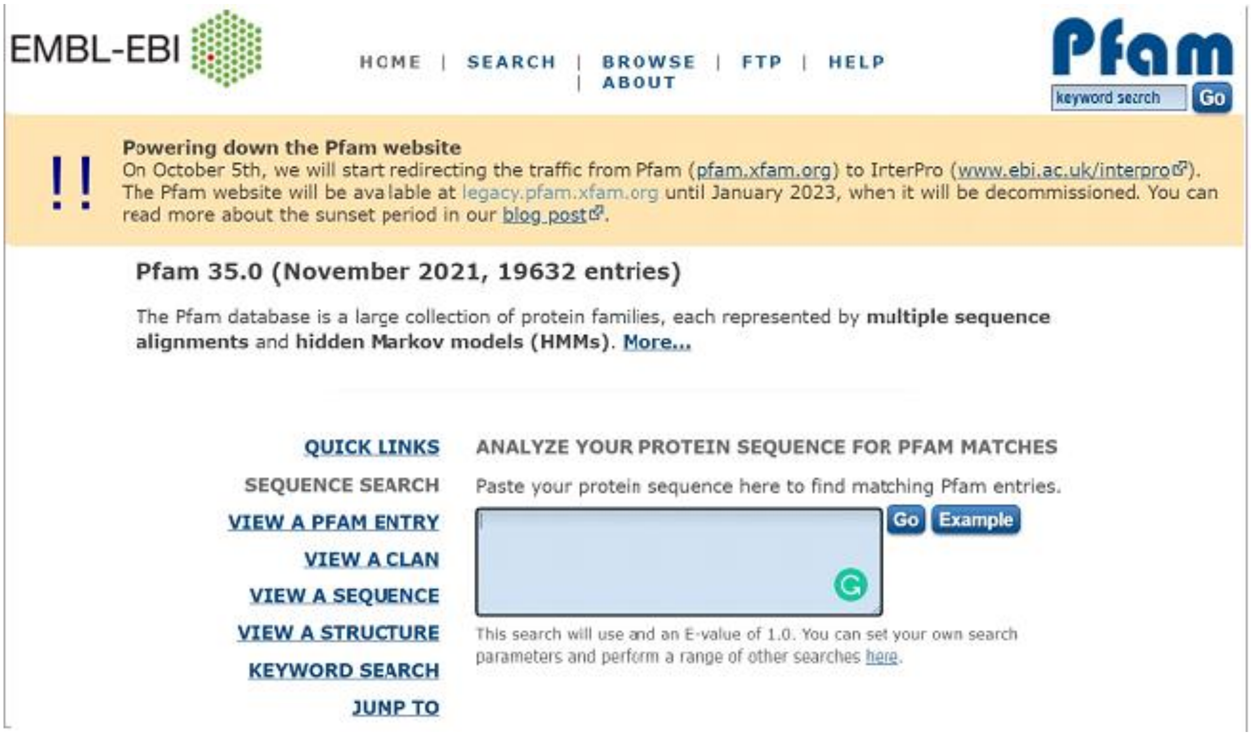
OPTIONS

Search against database: CDD v3.20 - 59693 PSSMs
Expect Value threshold: 0.010000
Apply low-complexity filter □
Composition based statistics adjustment ☑
Force live search □
Rescue borderline hits □ Suppress weak overlapping hits □
Maximum number of hits 500
Result mode ● Concise ○ Standard ○ Full

Retrieve previous CD-search result

Request ID:

Figure 3.4.1: National Center for Biotechnology Information sever for search conserved domain



!! **Powering down the Pfam website**
On October 5th, we will start redirecting the traffic from Pfam (pfam.xfam.org) to InterPro (www.ebi.ac.uk/interpro). The Pfam website will be available at legacy.pfam.xfam.org until January 2023, when it will be decommissioned. You can read more about the sunset period in our blog post.

Pfam 35.0 (November 2021, 19632 entries)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. More...

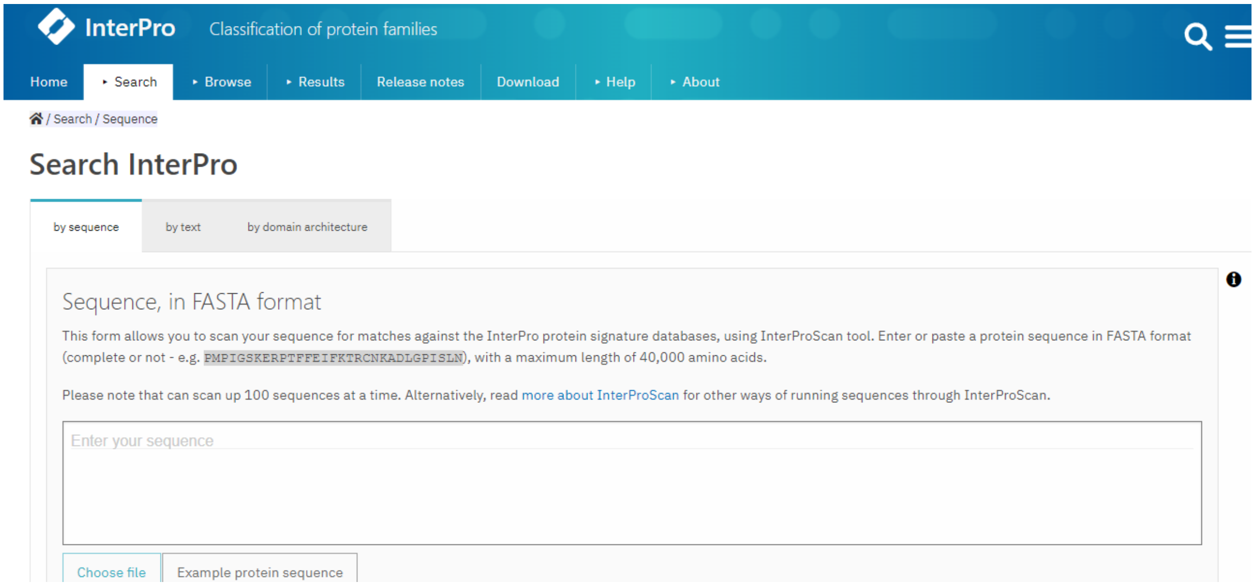
ANALYZE YOUR PROTEIN SEQUENCE FOR PFAM MATCHES

Paste your protein sequence here to find matching Pfam entries.

This search will use and an E-value of 1.0. You can set your own search parameters and perform a range of other searches here.

Figure 3.4.2: EMBL-EBI Pfam server for Function prediction

Functional analysis of proteins is provided by Interpro (https:// www.ebi.ac.uk/interpro), which classifies proteins into families and forecasts their domains and important sites. In order to categorize proteins in this way, the Interpro collaboration uses prediction models, known as signatures, supplied by several databases (referred to as member databases) [10].



Search InterPro

Sequence, in FASTA format

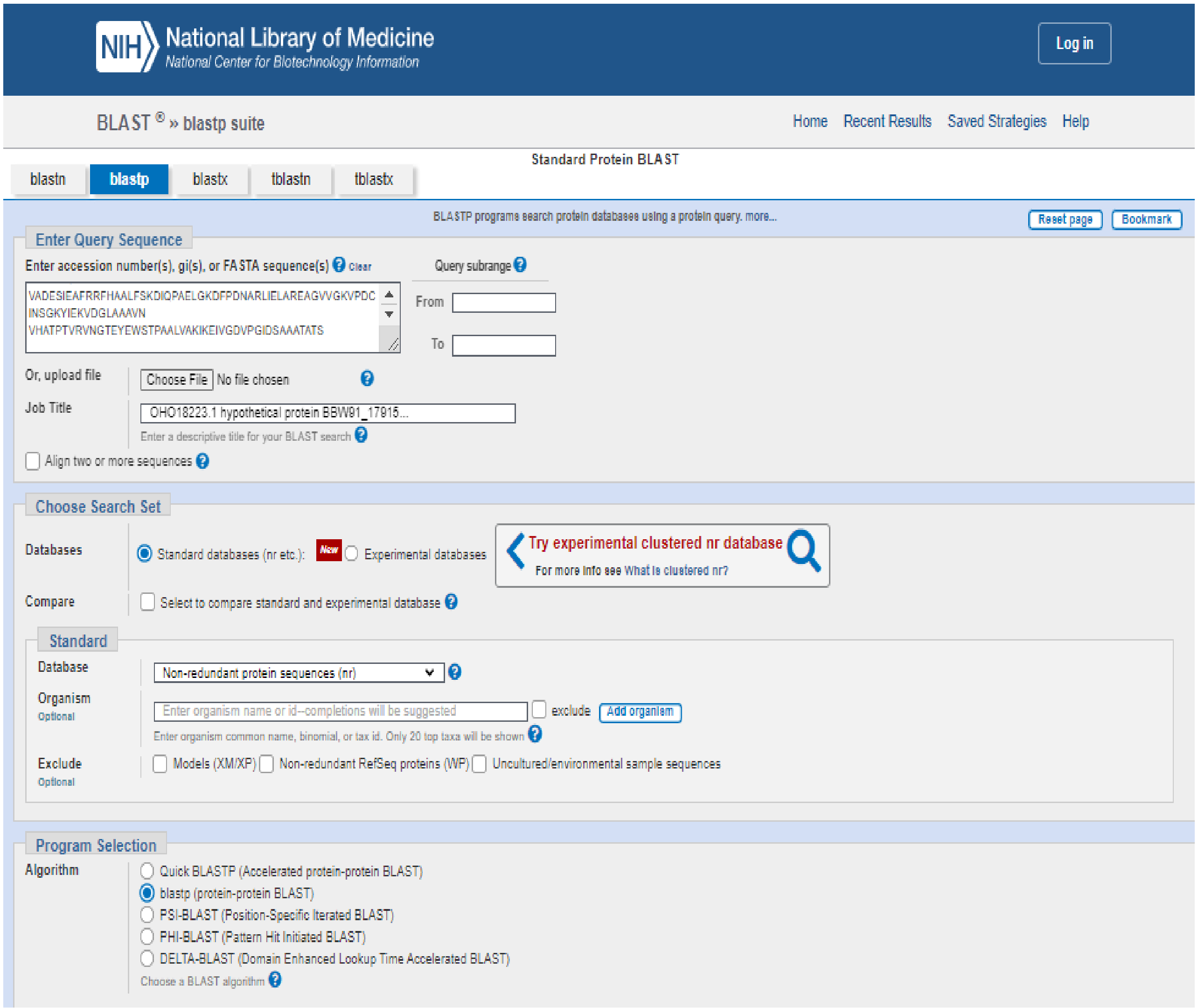
This form allows you to scan your sequence for matches against the InterPro protein signature databases, using InterProScan tool. Enter or paste a protein sequence in FASTA format (complete or not - e.g. PMPIGSKERPTFFEIFKTRCNKADLGPISLN), with a maximum length of 40,000 amino acids.

Please note that can scan up 100 sequences at a time. Alternatively, read more about InterProScan for other ways of running sequences through InterProScan.

Figure 3.4.3 InterPro Classification of protein families

3.5 Homology Searching

The most popular and effective method for characterizing newly found genomes is to search for nucleotide homology, usually with BLAST. By finding excess similarity, a numerically significant similarity that indicates common lineage, sequence similarity searches can discover "homologous" proteins or genes. More than 80% of metagenomics sequence samples commonly display substantial homology to proteins in sequence databases because of the extent of contemporary protein sequence databases. To improve accessibility and performance, BLAST's public interface at the NCBI site (http:// www. ncbi.nlm.nih.gov/blast) was recently rebuilt.



Standard Protein BLAST

BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

```
VADESIEAFRRFHAALFSKDIQPAELGKDFPDNARLIELAREAGVVGKVPDC
INSGKYIEKVDGLAAAVN
VHATPTVRVNGTEYEWSTPAALVAKIKEIVGDVPGIDSAAATATS
```

Query subrange
From
To

Or, upload file: Choose File No file chosen

Job Title: OHO18223.1 hypothetical protein BBW91_17915...
Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Databases: ● Standard databases (nr etc.): New ○ Experimental databases

Try experimental clustered nr database
For more info see What is clustered nr?

Compare: □ Select to compare standard and experimental database

Standard

Database: Non-redundant protein sequences (nr)

Organism Optional: Enter organism name or id--completions will be suggested □ exclude Add organism
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude Optional: □ Models (XM/XP) □ Non-redundant RefSeq proteins (WP) □ Uncultured/environmental sample sequences

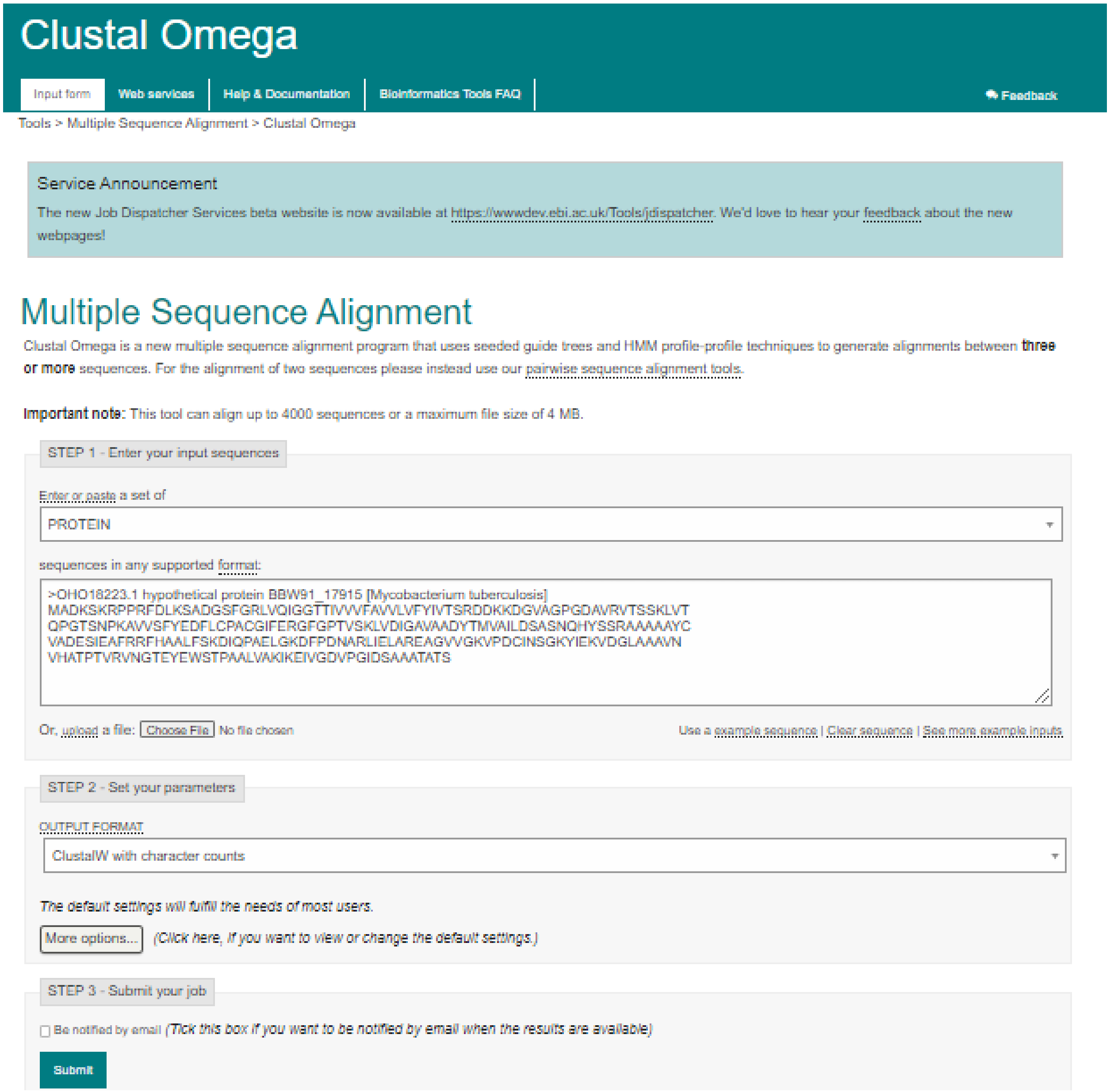
Program Selection

Algorithm:
○ Quick BLASTP (Accelerated protein-protein BLAST)
● blastp (protein-protein BLAST)
○ PSI-BLAST (Position-Specific Iterated BLAST)
○ PHI-BLAST (Pattern Hit Initiated BLAST)
○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
Choose a BLAST algorithm

Figure 3.5.1: NCBI BLAST Tool Webpage for Protein blast

3.6 Multiple nucleotide matching and Predicted Phylogenetic tree

Multiple sequence alignment is typically accomplished using Clustal series programs. The use of the mBed method for building guide-trees allows it to manage exceptionally large quantities (many tens of thousands) of DNA/RNA or protein sequences. With the help of planted guidance trees and HMM profile-profile methods, the brand-new multiple sequence alignment software Clustal Omega creates arrangements involving three or more [11]. This method has been one of the most widely used in bioinformatics for decades because it is a prerequisite for the majority of phylogenetic or comparative analysis of homologous genes or proteins.



Service Announcement

The new Job Dispatcher Services beta website is now available at https://wwwdev.ebi.ac.uk/Tools/jdispatcher. We'd love to hear your feedback about the new webpages!

Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between **three or more** sequences. For the alignment of two sequences please instead use our pairwise sequence alignment tools.

Important note: This tool can align up to 4000 sequences or a maximum file size of 4 MB.

STEP 1 - Enter your input sequences

Enter or paste a set of

PROTEIN

sequences in any supported format:

```
>OHO18223.1 hypothetical protein BBW91_17915 [Mycobacterium tuberculosis]
MADKSKRPPRFDLKSADGSFGRLVQIGGTTIVVVFAVVLVFYIVTSRDDKKDGVAGPGDAVRVTSSKLVT
QPGTSNPKAVVSFYEDFLCPACGIFERGFGPTVSKLVDIGAVAADYTMVAILDSASNQHYSSRAAAAAYC
VADESIEAFRRFHAALFSKDIQPAELGKDFPDNARLIELAREAGVVGKVPDCINSGKYIEKVDGLAAAVN
VHATPTVRVNGTEYEWSTPAALVAKIKEIVGDVPGIDSAAATATS
```

Or, upload a file: Choose File No file chosen

Use a example sequence | Clear sequence | See more example inputs

STEP 2 - Set your parameters

OUTPUT FORMAT

ClustalW with character counts

The default settings will fulfil the needs of most users.

More options... *(Click here, if you want to view or change the default settings.)*

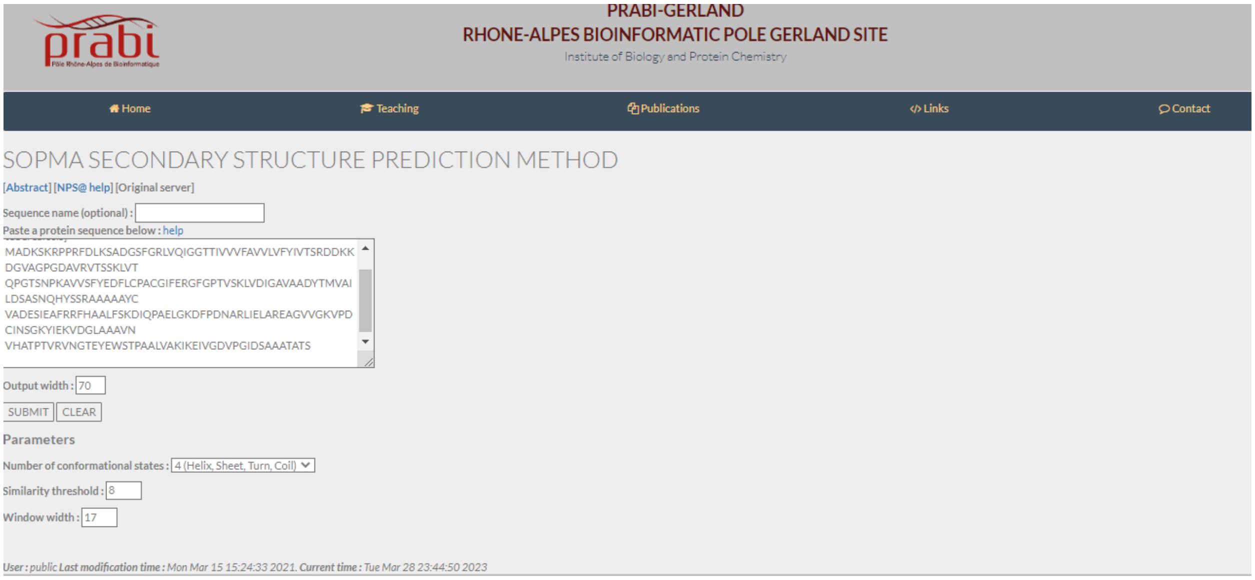
STEP 3 - Submit your job

□ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

Figure 3.6.1 Clustal Omega tools Used for Multiple Sequence Alignment

3.7 Secondary Structure Determination

The URL for this page is (https:// npsa-prabi.ibcp.fr/cgi-bin/npsa automat.pl?page=/NPSA/npsa sopma.html). Alignment-Based Self-Adjusted Forecast Technique It has been proposed that SOPMA could improve predictions of protein secondary structure. In a collection of 126 sequences of non-homologous (less than 25% identical) proteins, SOPMA correctly forecasts 69.5% of using amino acids to describe the tertiary structure in three states (-helix, -sheet, and coil) [12]. Users can enter a protein sequence, make a projection of their choice, and receive the prediction results both textually by e-mail and visually on the web using the PSIPRED technique for forecasting protein shape (http:// bioinf.cs.ucl.ac.uk/psipred/).


PRABI-GERLAND
RHONE-ALPES BIOINFORMATIC POLE GERLAND SITE
Institute of Biology and Protein Chemistry
Home
Teaching
Publications
Links
Contact

SOPMA SECONDARY STRUCTURE PREDICTION METHOD

[Abstract] [NPS@ help] [Original server]

Sequence name (optional) :

Paste a protein sequence below : help

```
MADKSKRPPRFDLKSADGSFGRLVQIGGTTIVVVFAVVLVFYIVTSRDDKK
DGVAGPGDAVRVTSSKLVT
QPGTSNPKAVVSFYEDFLCPACGIFERGFGPTVSKLVDIGAVAADYTMVAI
LDSASNQHYSSRAAAAAYC
VADESIEAFRRFHAALFSKDIQPAELGKDFPDNARLIELAREAGVVGKVPD
CINSGKYIEKVDGLAAAVN
VHATPTVRVNGTEYEWSTPAALVAKIKEIVGDVPGIDSAAATATS
```

Output width : 70

SUBMIT CLEAR

Parameters

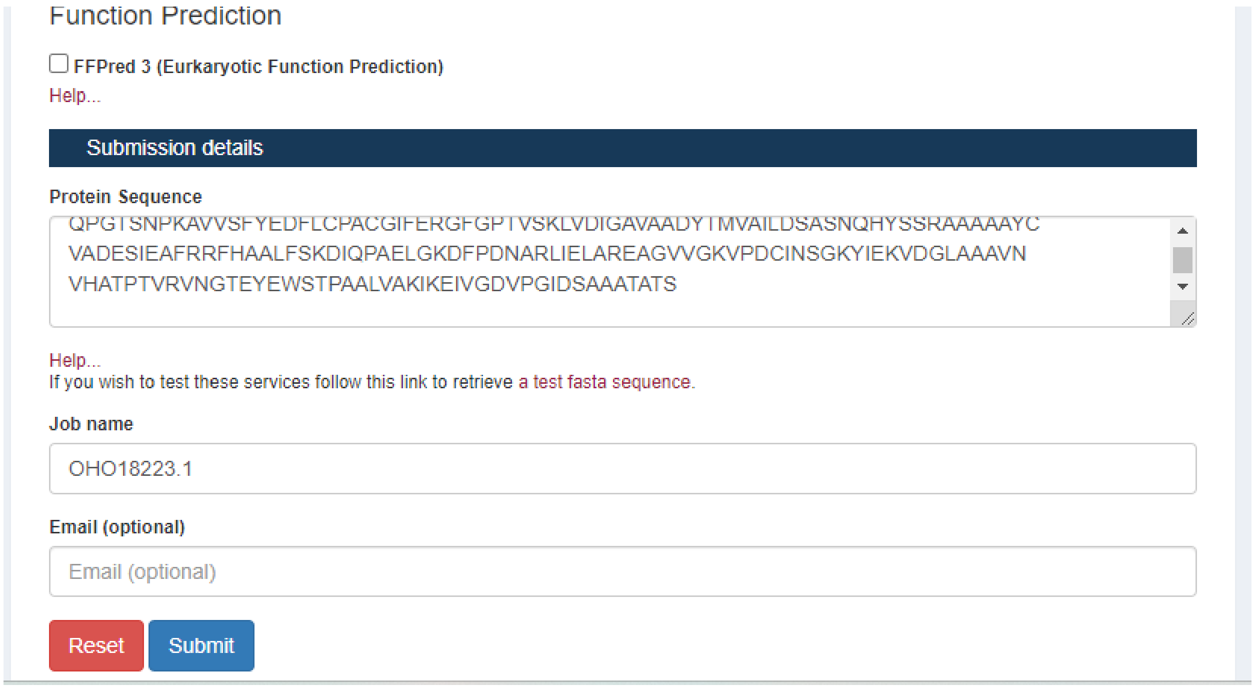
Number of conformational states : 4 (Helix, Sheet, Turn, Coil)

Similarity threshold : 8

Window width : 17

User : public Last modification time : Mon Mar 15 15:24:33 2021. Current time : Tue Mar 28 23:44:50 2023

Figure 3.7. 1 SOMPA Server used for secondary structure prediction



Function Prediction

☐ FFPred 3 (Eurkaryotic Function Prediction)

Help...

Submission details

Protein Sequence

```
QPGTSNPKAVVSFYEDFLCPACGIFERGFGPTVSKLVDIGAVAADYTMVAILDSASNQHYSSRAAAAAYC
VADESIEAFRRFHAALFSKDIQPAELGKDFPDNARLIELAREAGVVGKVPDCINSGKYIEKVDGLAAAVN
VHATPTVRVNGTEYEWSTPAALVAKIKEIVGDVPGIDSAAATATS
```

Help...

If you wish to test these services follow this link to retrieve a test fasta sequence.

Job name

OHO18223.1

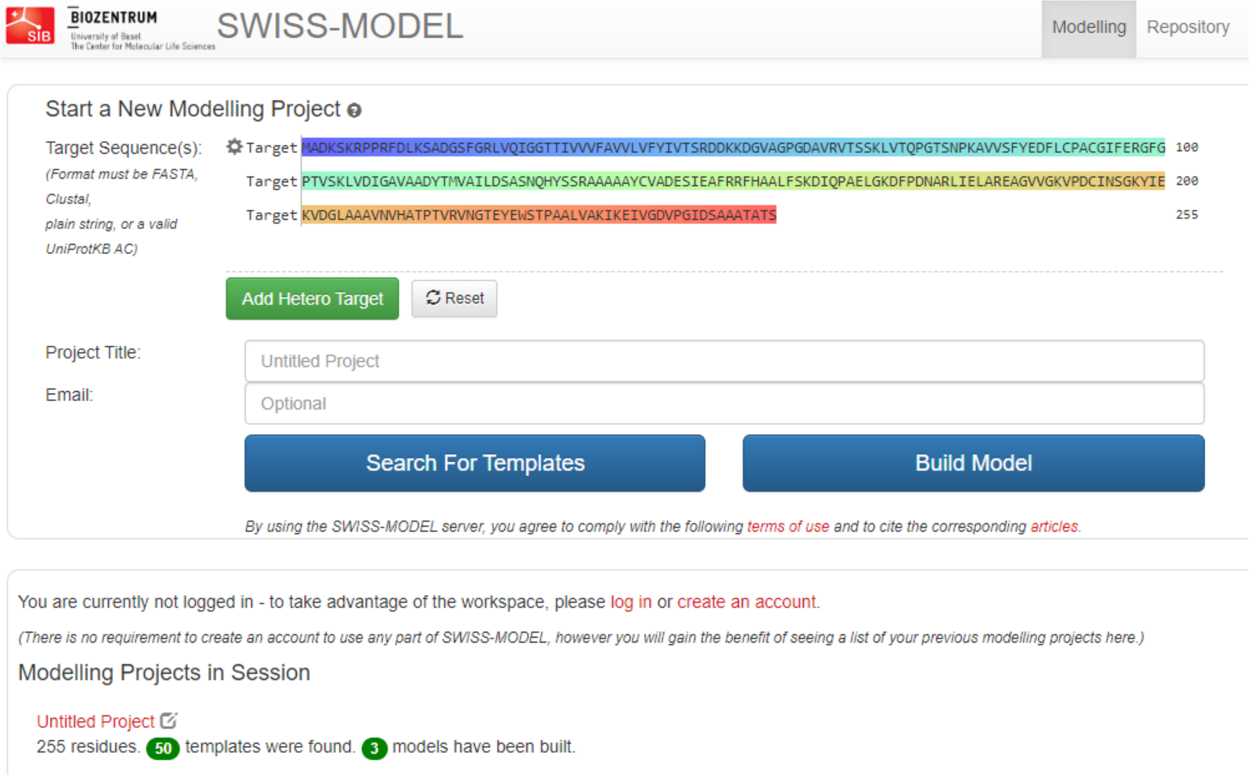
Email (optional)

Email (optional)

Figure 3.7. 2: PSIPRED server tool for Secondary structure prediction

3.8 Three-dimensional Structure Prediction

Users of the SWISS-MODEL Repository can evaluate the model accuracy in the database, search for different design layouts, and interactively construct models using SWISS- MODEL server (http:// swissmodel.expasy.org/workspace/). The first completely automatic protein homology modelling tool was SWISS-MODEL and it has advanced continuously over the past 25 years (25–30). The modelling of homo- and heterogenic compounds using the amino acid patterns of the contact partners as a beginning point has recently been added to its modelling capabilities.



Start a New Modelling Project

Target Sequence(s): (Format must be FASTA, Clustal, plain string, or a valid UniProtKB AC)

```
Target MADKSKRPPRFDLKSADGSFGRLVQIGGTTIVVVFAVVLVFYIVTSRDDKKDGVAGPGDAVRVTSSKLVTQPGTSNPKAVVSFYEDFLCPACGIFERGFG 100
Target PTVSKLVDIGAVAADYTMVAILDSASNQHYSSRAAAAAYCVADESIEAFRRFHAALFSKDIQPAELGKDFPDNARLIELAREAGVVGKVPDCINSGKYIE 200
Target KVDGLAAAVNVHATPTVRVNGTEYEWSTPAALVAKIKEIVGDVPGIDSAAATATS                                              255
```

Project Title: Untitled Project

Email: Optional

By using the SWISS-MODEL server, you agree to comply with the following terms of use and to cite the corresponding articles.

You are currently not logged in - to take advantage of the workspace, please log in or create an account.

(There is no requirement to create an account to use any part of SWISS-MODEL, however you will gain the benefit of seeing a list of your previous modelling projects here.)

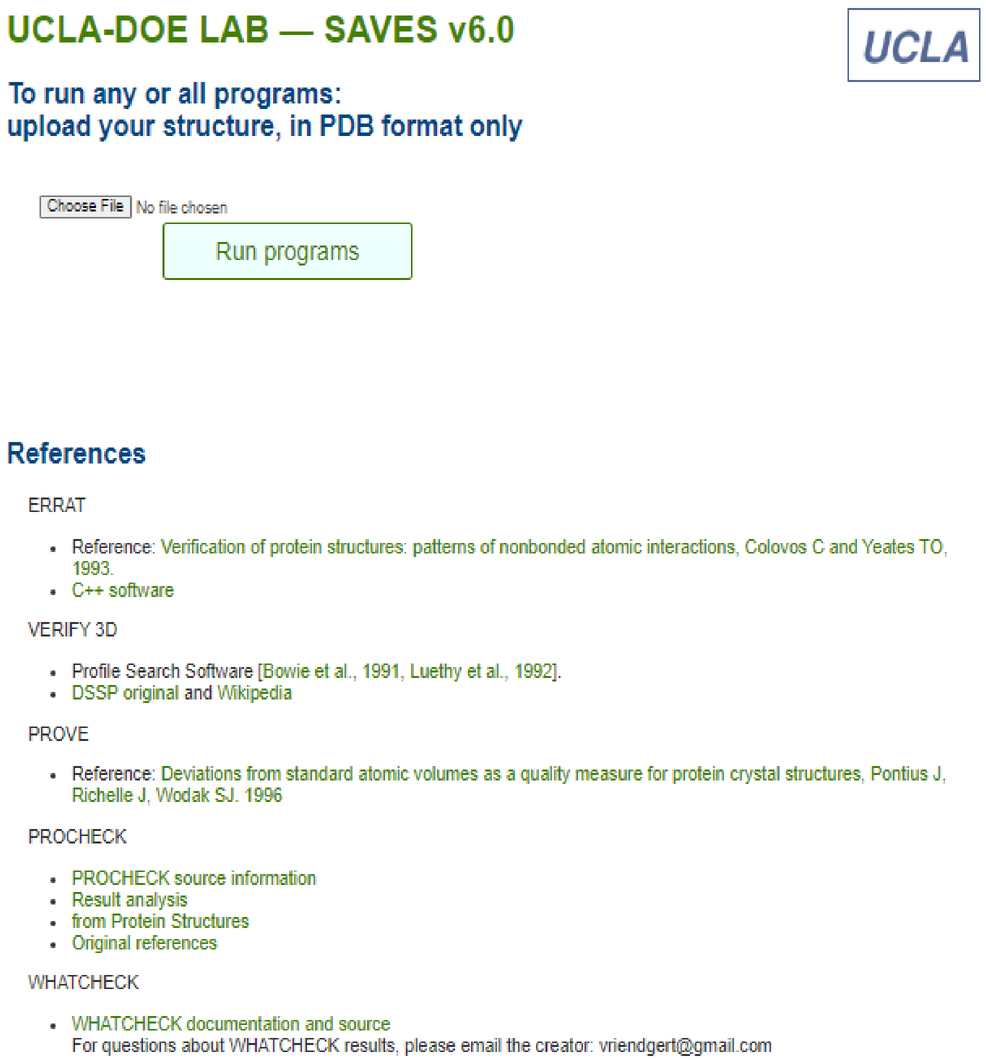
Modelling Projects in Session

Untitled Project
255 residues. 50 templates were found. 3 models have been built.

Figure 3.8.1 SWISS-MODEL server for predicting 3D structure of Hypothetical protein

3.9 Tertiary Structure Validation and Model Quality Assessment

The modelled **OHO18223.1** was validated using the structural analysis and verification service **SAVES v6.0** (https:// saves.mbi.ucla.edu/). The principal structural model was checked using the **ERRAT**, Verify3D, and **PROCHECK** tools, which are all a part of the structural assessment and confirmation server SAVESv6.0 [13], to look for problems with the structure's three-dimensional representation. ProSA, also known as protein structure analysis, is a method used to improve and verify experimental protein structures as well as forecast and model protein structure. Also, the **PROCHECK** software tool provides a complete analysis of a protein structure's stereochemistry. Its outputs contain a complete residue-by-residue listing as well as a variety of graphs in PostScript format.



UCLA-DOE LAB — SAVES v6.0

To run any or all programs:
upload your structure, in PDB format only

References

ERRAT

- Reference: Verification of protein structures: patterns of nonbonded atomic interactions, Colovos C and Yeates TO, 1993.
- C++ software

VERIFY 3D

- Profile Search Software [Bowie et al., 1991, Luethy et al., 1992].
- DSSP original and Wikipedia

PROVE

- Reference: Deviations from standard atomic volumes as a quality measure for protein crystal structures, Pontius J, Richelle J, Wodak SJ. 1996

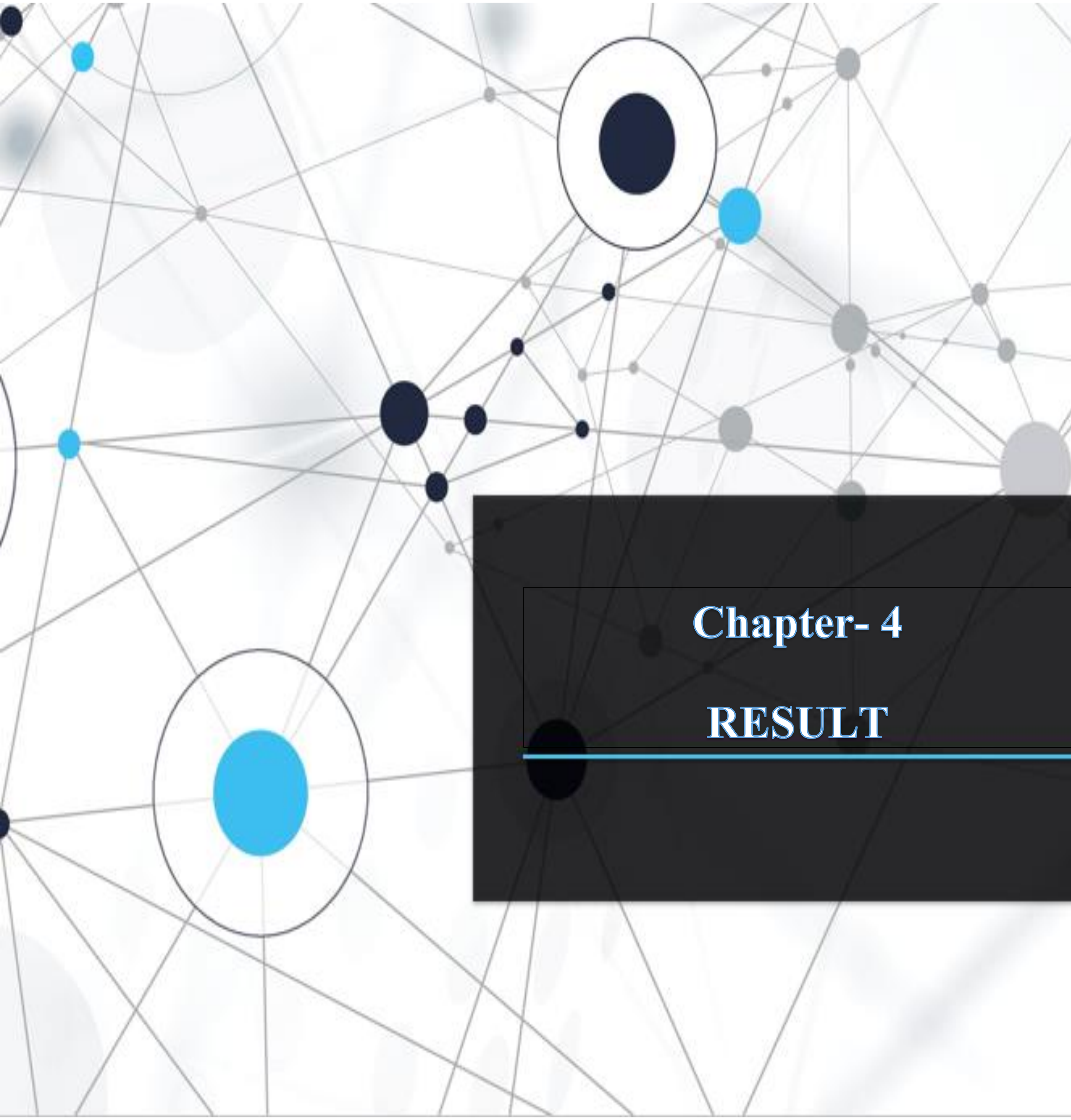
PROCHECK

- PROCHECK source information
- Result analysis
- from Protein Structures
- Original references

WHATCHECK

- WHATCHECK documentation and source
  For questions about WHATCHECK results, please email the creator: vriendgert@gmail.com

Figure 3.9. 1 SAVESv6.0 was analyzed in order to locate faults in the 3D structure



Chapter- 4

RESULT

Result

4.1 Sequence retrieval

A hypothetical protein sequence was chosen at random from the NCBI database using search term **"Hypothetical Protein of Mycobacterium Tuberculosis"** and the resulting protein sequence is shown below.

GenBank: **OHO18223.1**

>OHO18223.1 hypothetical protein BBW91_17915 [Mycobacterium tuberculosis]

MADKSKRPPRFDLKSADGSFGRLVQIGGTTIVVVFAVVLVFYIVTSRDDKKDGVAGPGD
AVRVTSSKLVTQPGTSNPKAVVSFYEDFLCPACGIFERGFGPTVSKLVDIGAVAADYTM
VAILDSASNQHYSSRAAAAAYCVADESIEAFRRFHAALFSKDIQPAELGKDFPDNARLIE
LAREAGVVGKVPDCINSGKYIEKVDGLAAAVNVHATPTVRVNGTEYEWSTPAALVAKI
KEIVGDVPGIDSAAATATS

4.2 Physicochemical properties analysis

For the hypothetical protein **OHO18223.1**, the ProtParam tool evaluated a number of physicochemical properties, which are shown in the table below.

Table 4.2.1 physicochemical properties of hypothetical protein

Accession number	Number of amino acids	Molecular weight	Estimated halflife	Theoretical pI	Asp + Glu	Arg + Lys	Aliphatic index	Instability index	Grand average of hydropathicity (GRAVY)
OHO18223.1	255	26861.54	30 hr	5.68	30	27	89.53	26.64	0.097

It was anticipated that the protein would have 255 amino acids, have a 26861.54 molecular weight and have a GRAVY score of 0.097, which indicates that it is water soluble. With a calculated instability index (II) of 26.64, the protein falls under the steady category. The protein is fragile if the protein instability score is higher than 40. The aliphatic index as determined is 89.53.

Alanine (14.1%) contained the highest concentration of amino acid residue. The sequence has

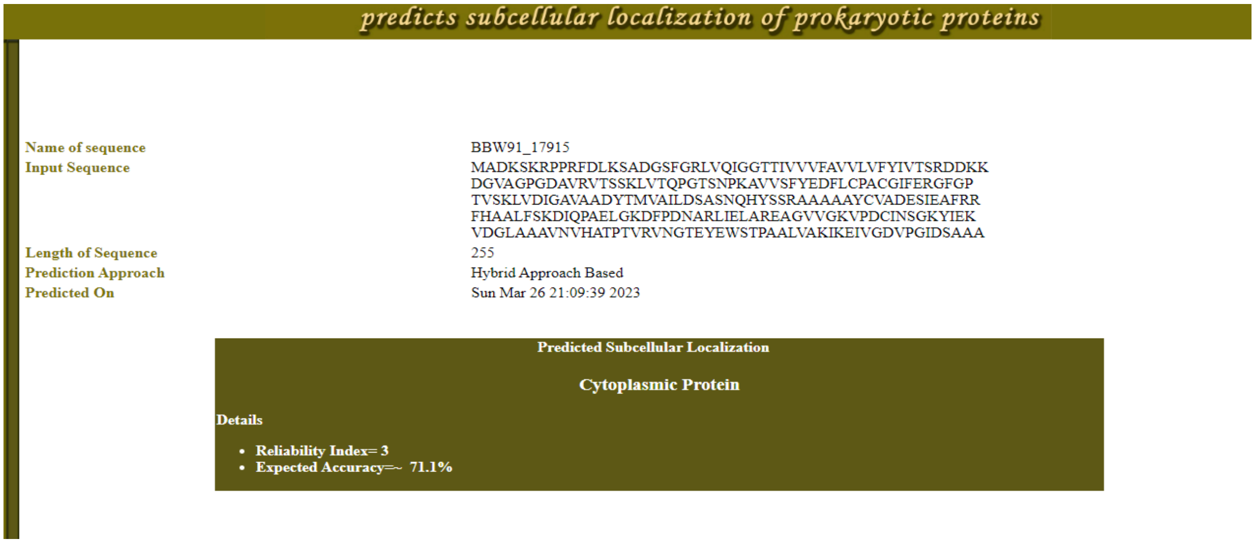
Total amount of molecules that are negatively charged (Asp+Glu): 30

Total amount of molecules that are positively charged (Arg+Lys): 27

4.3 Subcellular Localization Analysis

The hypothetical protein's subcellular location reveals details about its biological function. It should be beneficial in the process of creating a medication to ward off the target protein.

With and estimated accuracy of around 71.1% the PSLpred server suggested a cytoplasmic protein—a protein found within the cytoplasm of a cell—as the target protein's subcellular location.



predicts subcellular localization of prokaryotic proteins

Name of sequence	BBW91_17915
Input Sequence	MADKSKRPPRFDLKSADGSFGRLVQIGGTTIVVVFAVVLVFYIVTSRDDKK DGVAGPGDAVRVTSSKLVTQPGTSNPKAVVSFYEDFLCPACGIFERGFGP TVSKLVDIGAVAADYTMVAILDSASNQHYSSRAAAAAYCVADESIEAFRR FHAALFSKDIQPAELGKDFPDNARLIELAREAGVVGKVPDCINSGKYIEK VDGLAAAVNVHATPTVRVNGTEYEWSTPAALVAKIKEIVGDVPGIDSAAA
Length of Sequence	255
Prediction Approach	Hybrid Approach Based
Predicted On	Sun Mar 26 21:09:39 2023

Predicted Subcellular Localization

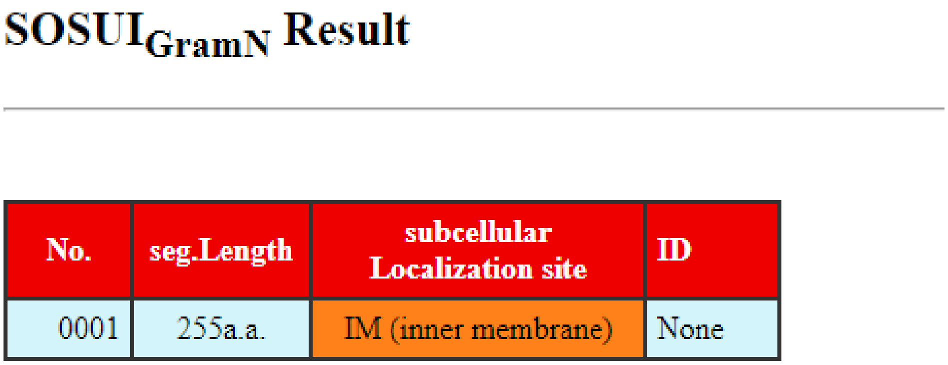
Cytoplasmic Protein

Details

- Reliability Index= 3
- Expected Accuracy=~ 71.1%

Figure 4.3.1 PSLpred subcellular localization result

The SOSUIGramN program states that the hypothetical is localized within the inner membrane of the cell.



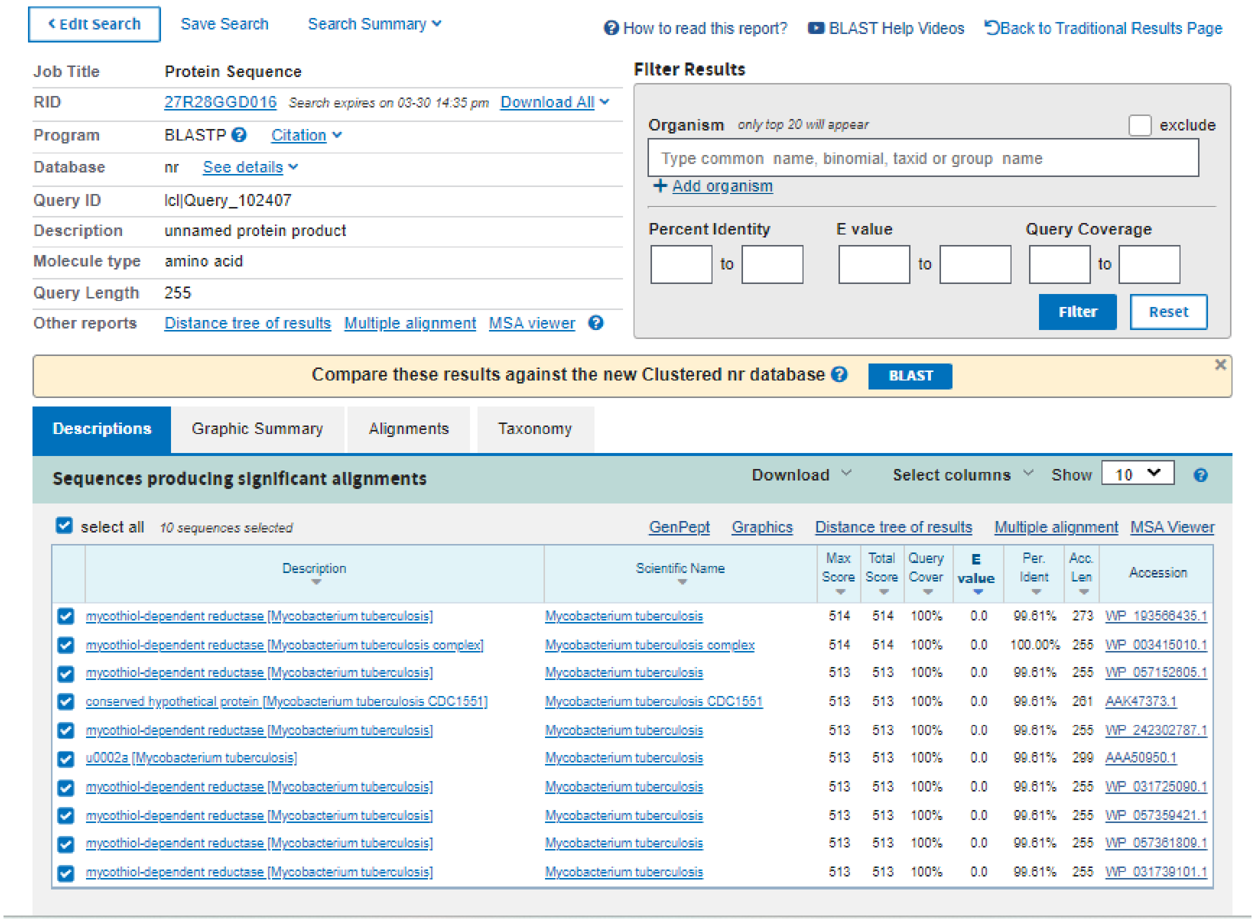
SOSUI$_{\text{GramN}}$ Result

No.	seg.Length	subcellular Localization site	ID
0001	255a.a.	IM (inner membrane)	None

Figure 4.3.2 SOSUI-GramN subcellular localization result

4.4 Homology Searching

The search for the homology of Hypothetical protein using the well-known BLASTp tools from NCBI. The top 10 results from a homology search are shown in the table below.



Job Title	Protein Sequence
RID	27R28GGD016 Search expires on 03-30 14:35 pm Download All
Program	BLASTP Citation
Database	nr See details
Query ID	lcl|Query_102407
Description	unnamed protein product
Molecule type	amino acid
Query Length	255
Other reports	Distance tree of results Multiple alignment MSA viewer

Sequences producing significant alignments

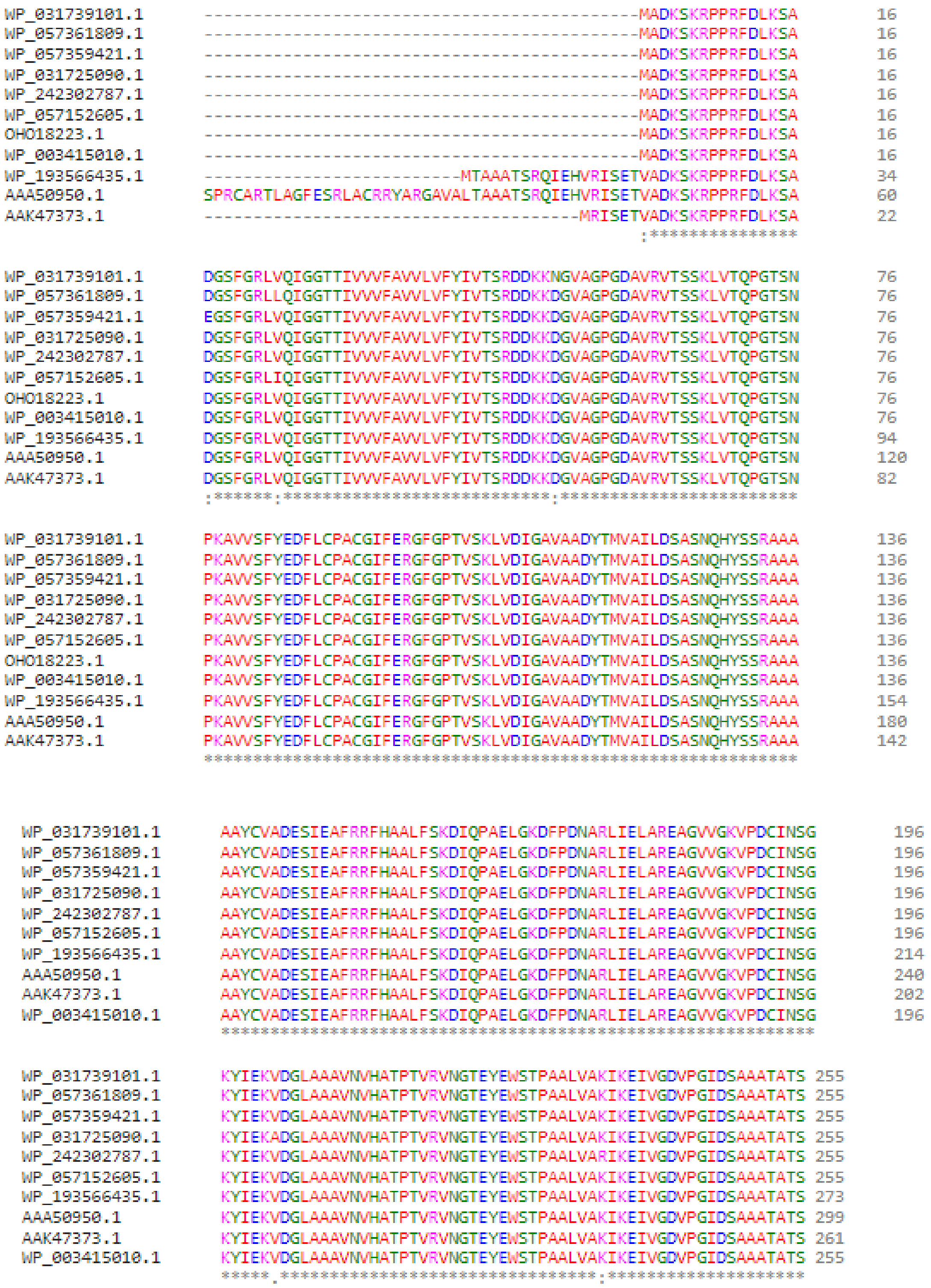
☑ select all 10 sequences selected

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	mycothiol-dependent reductase [Mycobacterium tuberculosis]	Mycobacterium tuberculosis	514	514	100%	0.0	99.61%	273	WP_193566435.1
☑	mycothiol-dependent reductase [Mycobacterium tuberculosis complex]	Mycobacterium tuberculosis complex	514	514	100%	0.0	100.00%	255	WP_003415010.1
☑	mycothiol-dependent reductase [Mycobacterium tuberculosis]	Mycobacterium tuberculosis	513	513	100%	0.0	99.61%	255	WP_057152605.1
☑	conserved hypothetical protein [Mycobacterium tuberculosis CDC1551]	Mycobacterium tuberculosis CDC1551	513	513	100%	0.0	99.61%	261	AAK47373.1
☑	mycothiol-dependent reductase [Mycobacterium tuberculosis]	Mycobacterium tuberculosis	513	513	100%	0.0	99.61%	255	WP_242302787.1
☑	u0002a [Mycobacterium tuberculosis]	Mycobacterium tuberculosis	513	513	100%	0.0	99.61%	299	AAA50950.1
☑	mycothiol-dependent reductase [Mycobacterium tuberculosis]	Mycobacterium tuberculosis	513	513	100%	0.0	99.61%	255	WP_031725090.1
☑	mycothiol-dependent reductase [Mycobacterium tuberculosis]	Mycobacterium tuberculosis	513	513	100%	0.0	99.61%	255	WP_057359421.1
☑	mycothiol-dependent reductase [Mycobacterium tuberculosis]	Mycobacterium tuberculosis	513	513	100%	0.0	99.61%	255	WP_057361809.1
☑	mycothiol-dependent reductase [Mycobacterium tuberculosis]	Mycobacterium tuberculosis	513	513	100%	0.0	99.61%	255	WP_031739101.1

Figure 4.4.1 Top 10 homology results from BLASTp tools

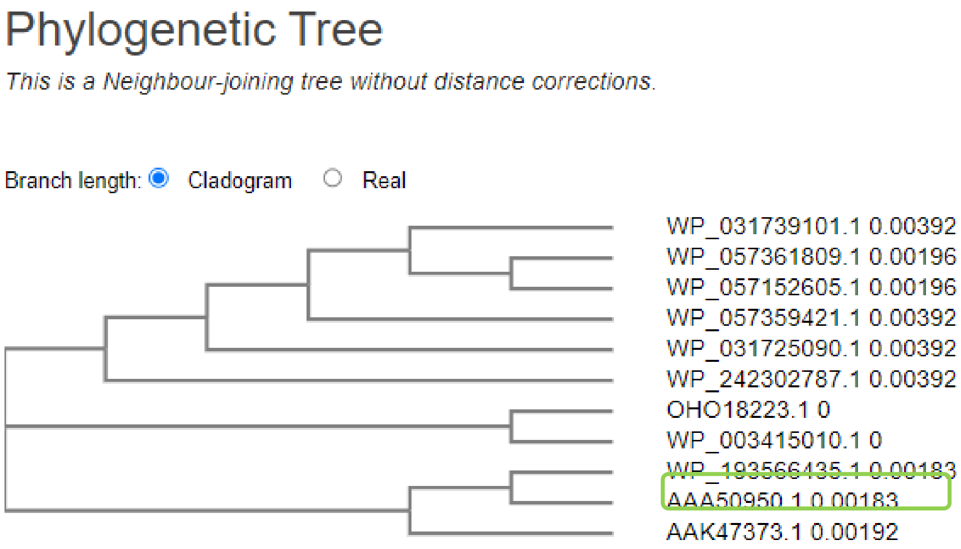
4.5 Multiple sequence alignment and phylogenetic Tree prediction

In the figure below the top 10 multiple sequence alignment and Phylogenetic analysis shown. Here out target protein **OHO18223.1** is more similar to **WP_003415010.1**



```
WP_031739101.1      --------------------------------------------MADKSKRPPRFDLKSA        16
WP_057361809.1      --------------------------------------------MADKSKRPPRFDLKSA        16
WP_057359421.1      --------------------------------------------MADKSKRPPRFDLKSA        16
WP_031725090.1      --------------------------------------------MADKSKRPPRFDLKSA        16
WP_242302787.1      --------------------------------------------MADKSKRPPRFDLKSA        16
WP_057152605.1      --------------------------------------------MADKSKRPPRFDLKSA        16
OHO18223.1          --------------------------------------------MADKSKRPPRFDLKSA        16
WP_003415010.1      --------------------------------------------MADKSKRPPRFDLKSA        16
WP_193566435.1      --------------------------MTAAATSRQIEHVRISETVADKSKRPPRFDLKSA        34
AAA50950.1          SPRCARTLAGFESRLACRRYARGAVALTAAATSRQIEHVRISETVADKSKRPPRFDLKSA        60
AAK47373.1          --------------------------------------MRISETVADKSKRPPRFDLKSA        22
                                                                :***************

WP_031739101.1      DGSFGRLVQIGGTTIVVVFAVVLVFYIVTSRDDKKNGVAGPGDAVRVTSSKLVTQPGTSN        76
WP_057361809.1      DGSFGRLLQIGGTTIVVVFAVVLVFYIVTSRDDKKDGVAGPGDAVRVTSSKLVTQPGTSN        76
WP_057359421.1      EGSFGRLVQIGGTTIVVVFAVVLVFYIVTSRDDKKDGVAGPGDAVRVTSSKLVTQPGTSN        76
WP_031725090.1      DGSFGRLVQIGGTTIVVVFAVVLVFYIVTSRDDKKDGVAGPGDAVRVTSSKLVTQPGTSN        76
WP_242302787.1      DGSFGRLVQIGGTTIVVVFAVVLVFYIVTSRDDKKDGVAGPGDAVRVTSSKLVTQPGTSN        76
WP_057152605.1      DGSFGRLIQIGGTTIVVVFAVVLVFYIVTSRDDKKDGVAGPGDAVRVTSSKLVTQPGTSN        76
OHO18223.1          DGSFGRLVQIGGTTIVVVFAVVLVFYIVTSRDDKKDGVAGPGDAVRVTSSKLVTQPGTSN        76
WP_003415010.1      DGSFGRLVQIGGTTIVVVFAVVLVFYIVTSRDDKKDGVAGPGDAVRVTSSKLVTQPGTSN        76
WP_193566435.1      DGSFGRLVQIGGTTIVVVFAVVLVFYIVTSRDDKKDGVAGPGDAVRVTSSKLVTQPGTSN        94
AAA50950.1          DGSFGRLVQIGGTTIVVVFAVVLVFYIVTSRDDKKDGVAGPGDAVRVTSSKLVTQPGTSN        120
AAK47373.1          DGSFGRLVQIGGTTIVVVFAVVLVFYIVTSRDDKKDGVAGPGDAVRVTSSKLVTQPGTSN        82
                    :******:***************************:************************

WP_031739101.1      PKAVVSFYEDFLCPACGIFERGFGPTVSKLVDIGAVAADYTMVAILDSASNQHYSSRAAA        136
WP_057361809.1      PKAVVSFYEDFLCPACGIFERGFGPTVSKLVDIGAVAADYTMVAILDSASNQHYSSRAAA        136
WP_057359421.1      PKAVVSFYEDFLCPACGIFERGFGPTVSKLVDIGAVAADYTMVAILDSASNQHYSSRAAA        136
WP_031725090.1      PKAVVSFYEDFLCPACGIFERGFGPTVSKLVDIGAVAADYTMVAILDSASNQHYSSRAAA        136
WP_242302787.1      PKAVVSFYEDFLCPACGIFERGFGPTVSKLVDIGAVAADYTMVAILDSASNQHYSSRAAA        136
WP_057152605.1      PKAVVSFYEDFLCPACGIFERGFGPTVSKLVDIGAVAADYTMVAILDSASNQHYSSRAAA        136
OHO18223.1          PKAVVSFYEDFLCPACGIFERGFGPTVSKLVDIGAVAADYTMVAILDSASNQHYSSRAAA        136
WP_003415010.1      PKAVVSFYEDFLCPACGIFERGFGPTVSKLVDIGAVAADYTMVAILDSASNQHYSSRAAA        136
WP_193566435.1      PKAVVSFYEDFLCPACGIFERGFGPTVSKLVDIGAVAADYTMVAILDSASNQHYSSRAAA        154
AAA50950.1          PKAVVSFYEDFLCPACGIFERGFGPTVSKLVDIGAVAADYTMVAILDSASNQHYSSRAAA        180
AAK47373.1          PKAVVSFYEDFLCPACGIFERGFGPTVSKLVDIGAVAADYTMVAILDSASNQHYSSRAAA        142
                    ************************************************************


WP_031739101.1      AAYCVADESIEAFRRFHAALFSKDIQPAELGKDFPDNARLIELAREAGVVGKVPDCINSG        196
WP_057361809.1      AAYCVADESIEAFRRFHAALFSKDIQPAELGKDFPDNARLIELAREAGVVGKVPDCINSG        196
WP_057359421.1      AAYCVADESIEAFRRFHAALFSKDIQPAELGKDFPDNARLIELAREAGVVGKVPDCINSG        196
WP_031725090.1      AAYCVADESIEAFRRFHAALFSKDIQPAELGKDFPDNARLIELAREAGVVGKVPDCINSG        196
WP_242302787.1      AAYCVADESIEAFRRFHAALFSKDIQPAELGKDFPDNARLIELAREAGVVGKVPDCINSG        196
WP_057152605.1      AAYCVADESIEAFRRFHAALFSKDIQPAELGKDFPDNARLIELAREAGVVGKVPDCINSG        196
WP_193566435.1      AAYCVADESIEAFRRFHAALFSKDIQPAELGKDFPDNARLIELAREAGVVGKVPDCINSG        214
AAA50950.1          AAYCVADESIEAFRRFHAALFSKDIQPAELGKDFPDNARLIELAREAGVVGKVPDCINSG        240
AAK47373.1          AAYCVADESIEAFRRFHAALFSKDIQPAELGKDFPDNARLIELAREAGVVGKVPDCINSG        202
WP_003415010.1      AAYCVADESIEAFRRFHAALFSKDIQPAELGKDFPDNARLIELAREAGVVGKVPDCINSG        196
                    ************************************************************

WP_031739101.1      KYIEKVDGLAAAVNVHATPTVRVNGTEYEWSTPAALVAKIKEIVGDVPGIDSAAATATS 255
WP_057361809.1      KYIEKVDGLAAAVNVHATPTVRVNGTEYEWSTPAALVAKIKEIVGDVPGIDSAAATATS 255
WP_057359421.1      KYIEKVDGLAAAVNVHATPTVRVNGTEYEWSTPAALVAKIKEIVGDVPGIDSAAATATS 255
WP_031725090.1      KYIEKADGLAAAVNVHATPTVRVNGTEYEWSTPAALVAKIKEIVGDVPGIDSAAATATS 255
WP_242302787.1      KYIEKVDGLAAAVNVHATPTVRVNGTEYEWSTPAALVARIKEIVGDVPGIDSAAATATS 255
WP_057152605.1      KYIEKVDGLAAAVNVHATPTVRVNGTEYEWSTPAALVAKIKEIVGDVPGIDSAAATATS 255
WP_193566435.1      KYIEKVDGLAAAVNVHATPTVRVNGTEYEWSTPAALVAKIKEIVGDVPGIDSAAATATS 273
AAA50950.1          KYIEKVDGLAAAVNVHATPTVRVNGTEYEWSTPAALVAKIKEIVGDVPGIDSAAATATS 299
AAK47373.1          KYIEKVDGLAAAVNVHATPTVRVNGTEYEWSTPAALVAKIKEIVGDVPGIDSAAATATS 261
WP_003415010.1      KYIEKVDGLAAAVNVHATPTVRVNGTEYEWSTPAALVAKIKEIVGDVPGIDSAAATATS 255
                    *****.********************************:********************
```

Figure 4.5.1: Top 10 multiple sequence alignment obtained by using Clustal Omega tools



Phylogenetic Tree

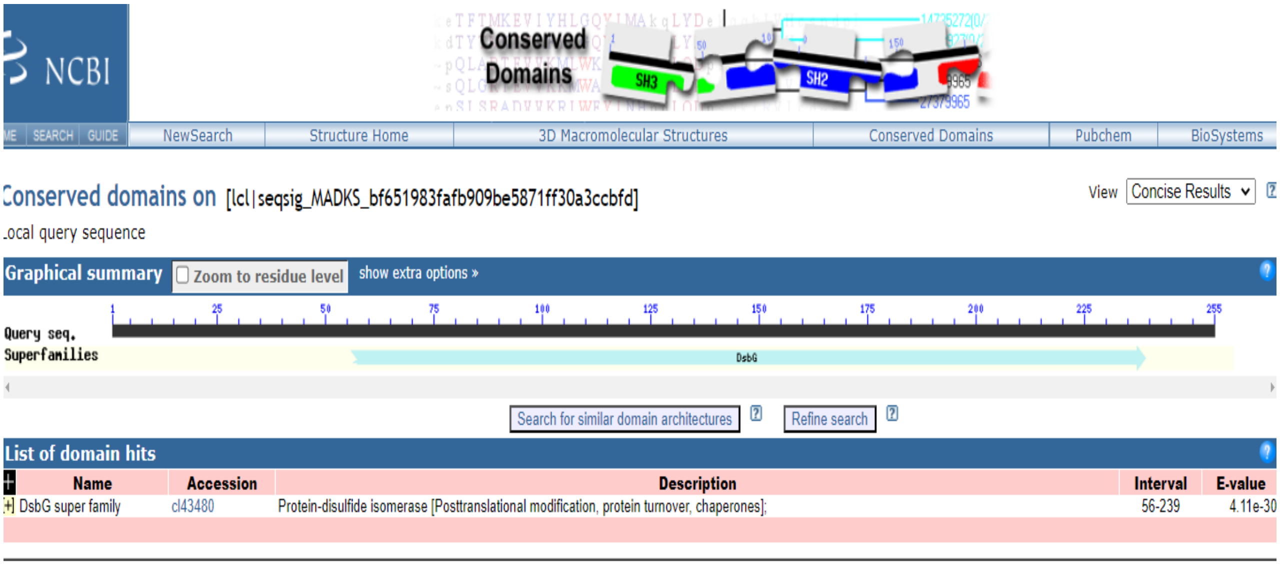
This is a Neighbour-joining tree without distance corrections.

Branch length: ● Cladogram ○ Real

Figure 4.5.2 Phylogenetic Tree Prediction

4.6 Protein family and Function prediction

The NCBI-CD Search predictions suggest the conserved domains and likely function of our target protein. The domain protein from the **DsbG super family** that is involved in protein-disulfide isomerase [posttranslational modification, protein turnover and chaperone] was discovered in the CD search with an Evalue of 4.11e-30 and an interval of 56-239.



Conserved domains on [lcl|seqsig_MADKS_bf651983fafb909be5871ff30a3ccbfd]

Local query sequence

View Concise Results

Graphical summary ☐ Zoom to residue level show extra options »

List of domain hits

	Name	Accession	Description	Interval	E-value
[+]	DsbG super family	cl43480	Protein-disulfide isomerase [Posttranslational modification, protein turnover, chaperones];	56-239	4.11e-30

Figure 4.6.1 Conserved domain result

The experiment is also done by Interpro server which has also shown us the same result.

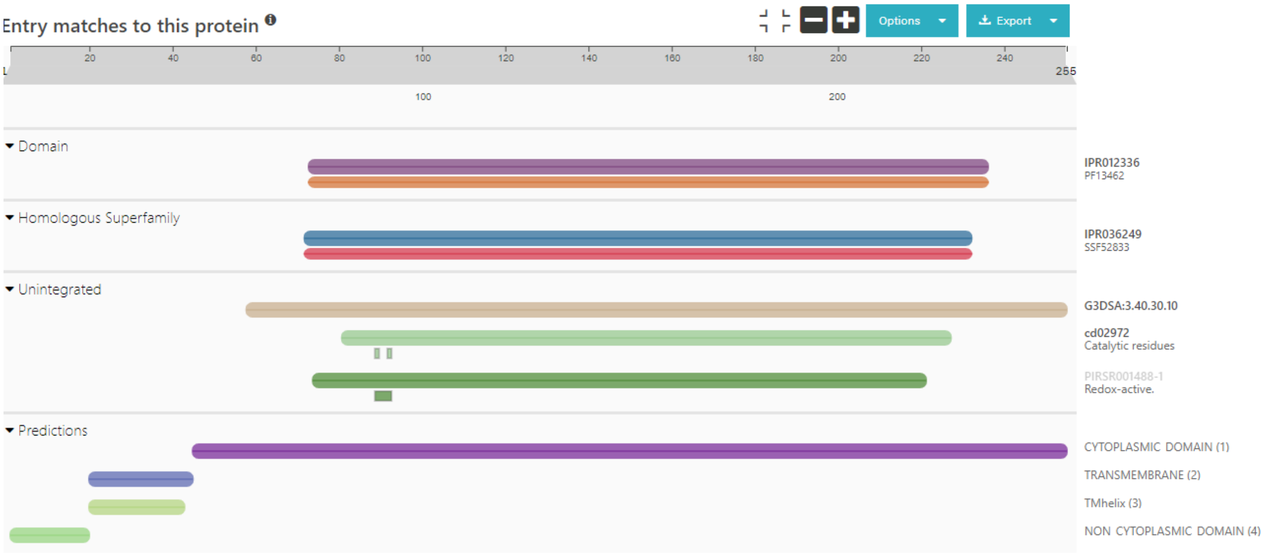


Figure 4.6.2 InterPro server result

Function of Protein-Disulfide Isomerase (PDI) in Prokaryotes:

❖ Protein folding and regular oxidative folding

The oxidoreductase and isomerase activities of protein disulfide-isomerase are influenced by the type of substrate that the enzyme binds to as well as changes in its redox state. These kinds of mechanisms enable protein oxidative folding. Disulfide bridges are formed when reduced cysteine residues in developing proteins are oxidized, stabilizing the protein and allowing it to form native structures, notably tertiary and quaternary structures. Protein folding is the physical transformation of a protein chain into its native three-dimensional structure, often known as a **"folded" conformation, which enables the protein to function physiologically** [14]. Protein folding in bacteria is closely controlled genetically, transcriptionally, and at the level of the protein structure because only folded proteins are usually effective. Moreover, essential cellular machinery supports polypeptide folding to guard against misfolding and ensure the creation of useful structures [15]. Proteins are selectively folded by PDI in the ER. At the active site (CGHC motif) of protein disulfide-isomerase, a cysteine residue in a stretched protein joins with another cysteine residue to create a mixed disulfide. The two cysteine residues in the protein disulfide-active isomerase's site are subsequently reduced as a result of a second cysteine residue in the substrate creating a stable disulfide bridge.

- **Redox signaling**

A redox chaperone called PDI is involved in the control of vascular NADPH oxidase, phagosome formation, and host cell absorption of bacteria and viruses. In addition to being related to phagocyte NADPH oxidase, PDI is also essential for L. chagasi to successfully infect macrophages . Many pathogens, including *Mycobacterium tuberculosis, Salmonella typhimurium, Toxoplasma gondii,* and *particularly Leishmania spp.* and *Trypanosoma cruzi,* can pass through or live in the phagosome [16].

- **Immune system**

PDI is a soluble homodimer without a transmembrane domain, like other ER chaperones. Moreover, it contains the KDEL C-terminal sequence, which binds to certain receptors in COP vesicles that migrate through the ER-Golgi region and recycles proteins back to the ER [17]. Antigenic peptides are helped to bind onto MHC class I molecules by protein disulfide-isomerase. These molecules (MHC I) are connected to the display of peptides during the immunological reaction by antigen-presenting cells.

- **Chaperone activity**

Another crucial role that protein disulfide-isomerase plays is chaperoning. One of the main functions of molecular chaperones is to prevent the buildup of proteins that are misfolded. Chaperones are a class of proteins that aid in the normal and stressed folding of proteins in the cell. They are both capable of recognizing and interacting with foreign proteins, which prevents unwanted protein aggregation [18]. Molecular chaperones play a variety of functions in bacterial cells, including assisting in protein release, repairing proteins that have been harmed or misfolded by stresses like a heat shock, as well as preventing protein aggregation in reaction to a heat shock. Moreover, they support protein folding during and after translation of newly produced proteins.

4.7 Secondary Structure determination

Using SOPMA, the protein's primary and secondary structural characteristics were identified. According to these calculations, the protein contains 35.29% alpha helix, 20.00% extended stands, 7.06% beta turns, and 37.65% random coils. The PRISPRED website, which presents a summary of comparable results, also double-checked the protein structural results. The putative protein's typical secondary structure (**OHO18223.1**) is displayed below.

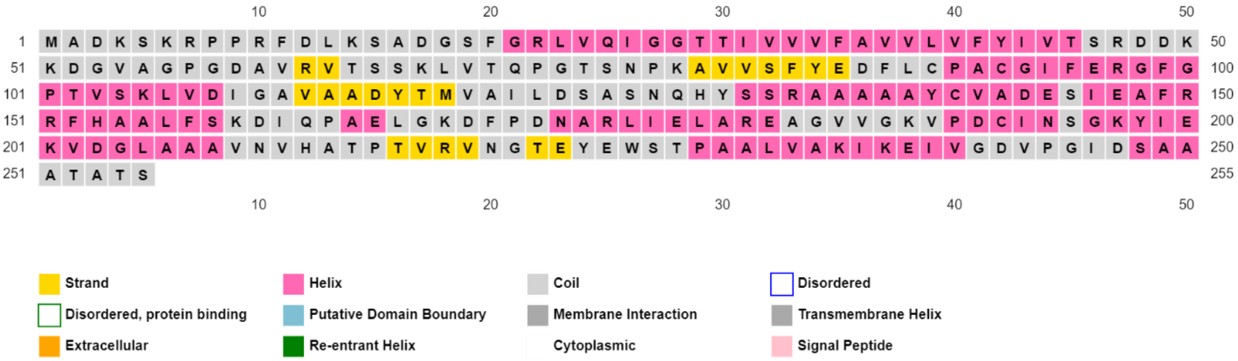
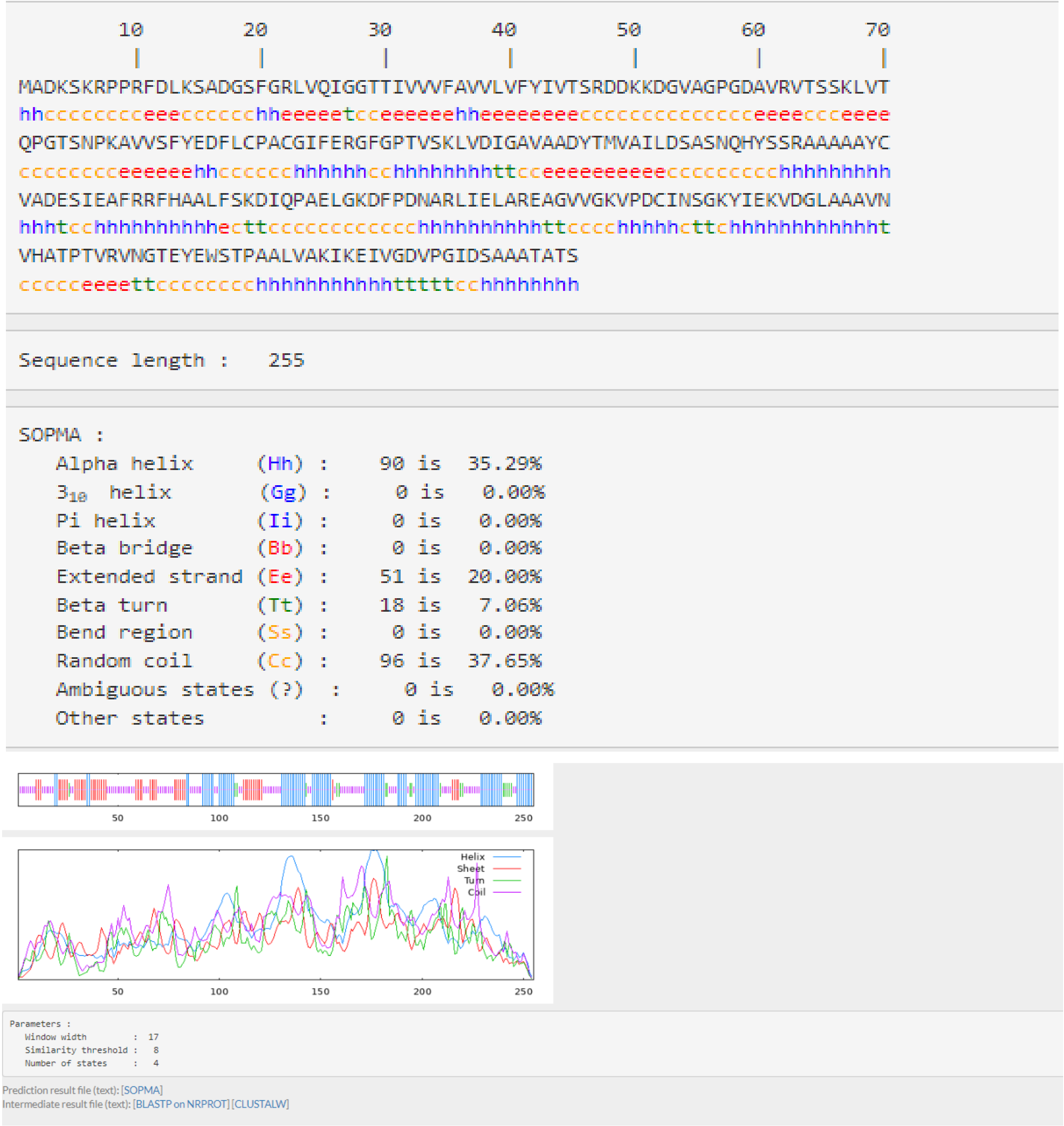

10 20 30 40 50
1 M A D K S K R P P R F D L K S A D G S F G R L V Q I G G T T I V V V F A V V L V F Y I V T S R D D K 50
51 K D G V A G P G D A V R V T S S K L V T Q P G T S N P K A V V S F Y E D F L C P A C G I F E R G F G 100
101 P T V S K L V D I G A V A A D Y T M V A I L D S A S N Q H Y S S R A A A A A Y C V A D E S I E A F R 150
151 R F H A A L F S K D I Q P A E L G K D F P D N A R L I E L A R E A G V V G K V P D C I N S G K Y I E 200
201 K V D G L A A A V N V H A T P T V R V N G T E Y E W S T P A A L V A K I K E I V G D V P G I D S A A 250
251 A T A T S 255
10 20 30 40 50
Strand
Helix
Coil
Disordered
Disordered, protein binding
Putative Domain Boundary
Membrane Interaction
Transmembrane Helix
Extracellular
Re-entrant Helix
Cytoplasmic
Signal Peptide

Figure 4.7.1. The positions of the alpha helix, beta turn, extended strand, and random coil are shown individually in this picture.



```
        10        20        30        40        50        60        70
         |         |         |         |         |         |         |
MADKSKRPPRFDLKSADGSFGRLVQIGGTTIVVVFAVVLVFYIVTSRDDKKDGVAGPGDAVRVTSSKLVT
hhcccccccceeeccccccchheeeeetcceeeeeeehheeeeeeeeecccccccccccccceeeeccceeee
QPGTSNPKAVVSFYEDFLCPACGIFERGFGPTVSKLVDIGAVAADYTMVAILDSASNQHYSSRAAAAAYC
cccccccceeeeeehhcccccchhhhhhcchhhhhhhhttcceeeeeeeeeeccccccccchhhhhhhhh
VADESIEAFRRFHAALFSKDIQPAELGKDFPDNARLIELAREAGVVGKVPDCINSGKYIEKVDGLAAAVN
hhhtcchhhhhhhhhhecttcccccccccccchhhhhhhhhhttcccchhhhhcttchhhhhhhhhhhht
VHATPTVRVNGTEYEWSTPAALVAKIKEIVGDVPGIDSAAATATS
ccccceeeettcccccccchhhhhhhhhhhtttttcchhhhhhhh

Sequence length :   255

SOPMA :
   Alpha helix      (Hh) :    90 is  35.29%
   3_10 helix        (Gg) :     0 is   0.00%
   Pi helix         (Ii) :     0 is   0.00%
   Beta bridge      (Bb) :     0 is   0.00%
   Extended strand  (Ee) :    51 is  20.00%
   Beta turn        (Tt) :    18 is   7.06%
   Bend region      (Ss) :     0 is   0.00%
   Random coil      (Cc) :    96 is  37.65%
   Ambiguous states (?)  :     0 is   0.00%
   Other states          :     0 is   0.00%
```

```
Parameters :
   Window width         :  17
   Similarity threshold :   8
   Number of states     :   4
```

Prediction result file (text): [SOPMA]
Intermediate result file (text): [BLASTP on NRPROT] [CLUSTALW]

Figure 4.7.2 Secondary structure of hypothetical protein

4.8 Three-dimensional structure prediction

Using the template with the highest rating, the SWISS-MODEL server completed a 3D prediction with a 100% identification rate.

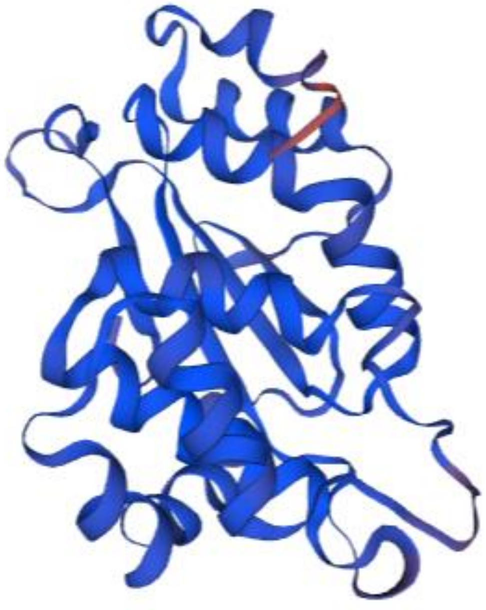
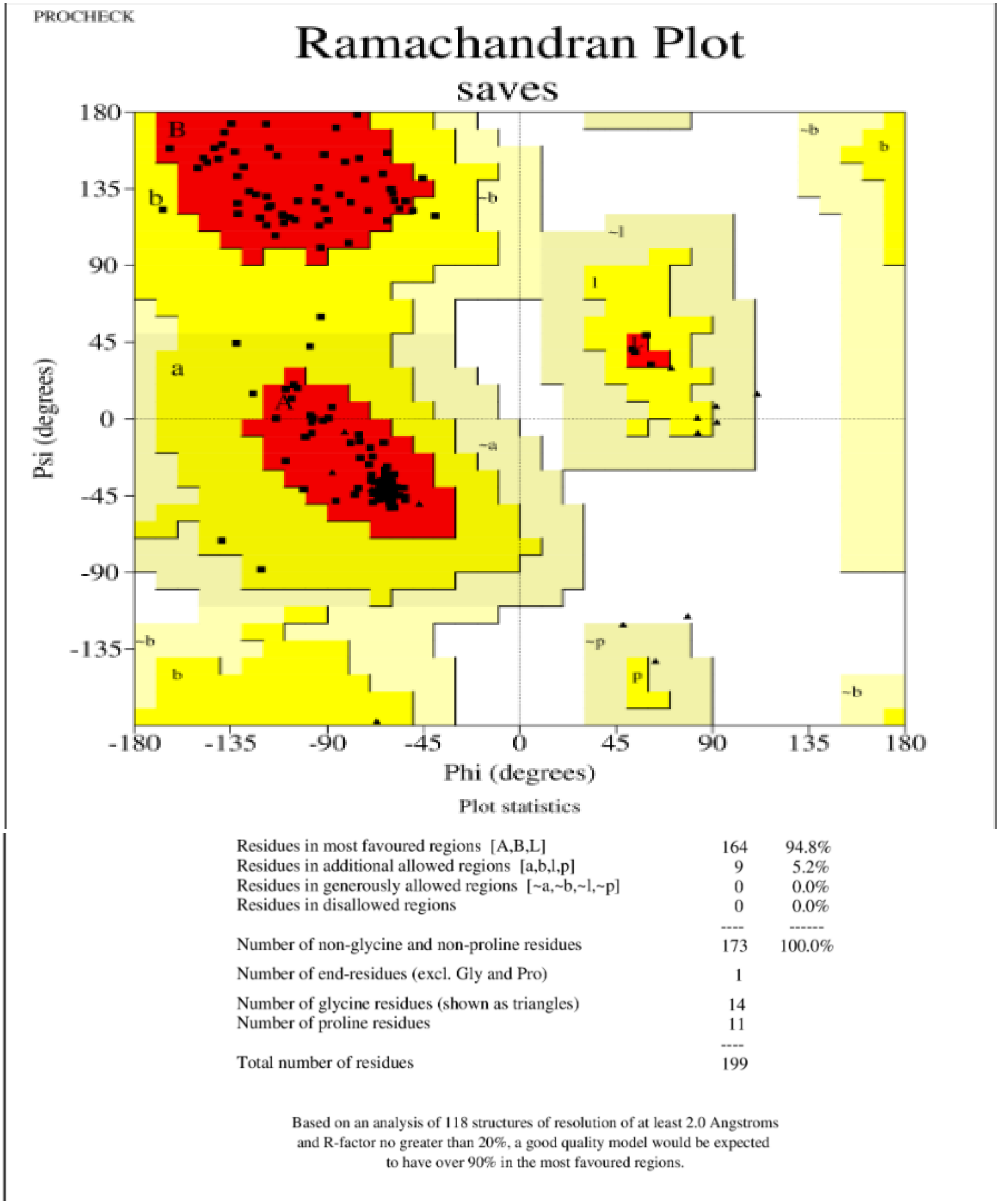


Figure 4.8.1. 3D structure of Hypothetical protein

4.9 Tertiary Structure Validation and model Quality Assessment

The quality of the 3D models was evaluated using Prosa -web and the Saves v6.0 server.

In the Prosa -web server, z-score was more common in the NMR than the x-ray regions. If the z-score is within the range of values frequently reported for genuine proteins of a similar size, the model is of excellent quality; otherwise, the structure is incorrect. The Prosa -web plot graph shows that the z-scores for PDB protein structures vary from -1 to -13. A value more that 90% inside the permitted zone is shown on the rating from the statistics of the Ramachandran plot, with 94.8% of residues in the intended locations.



Plot statistics

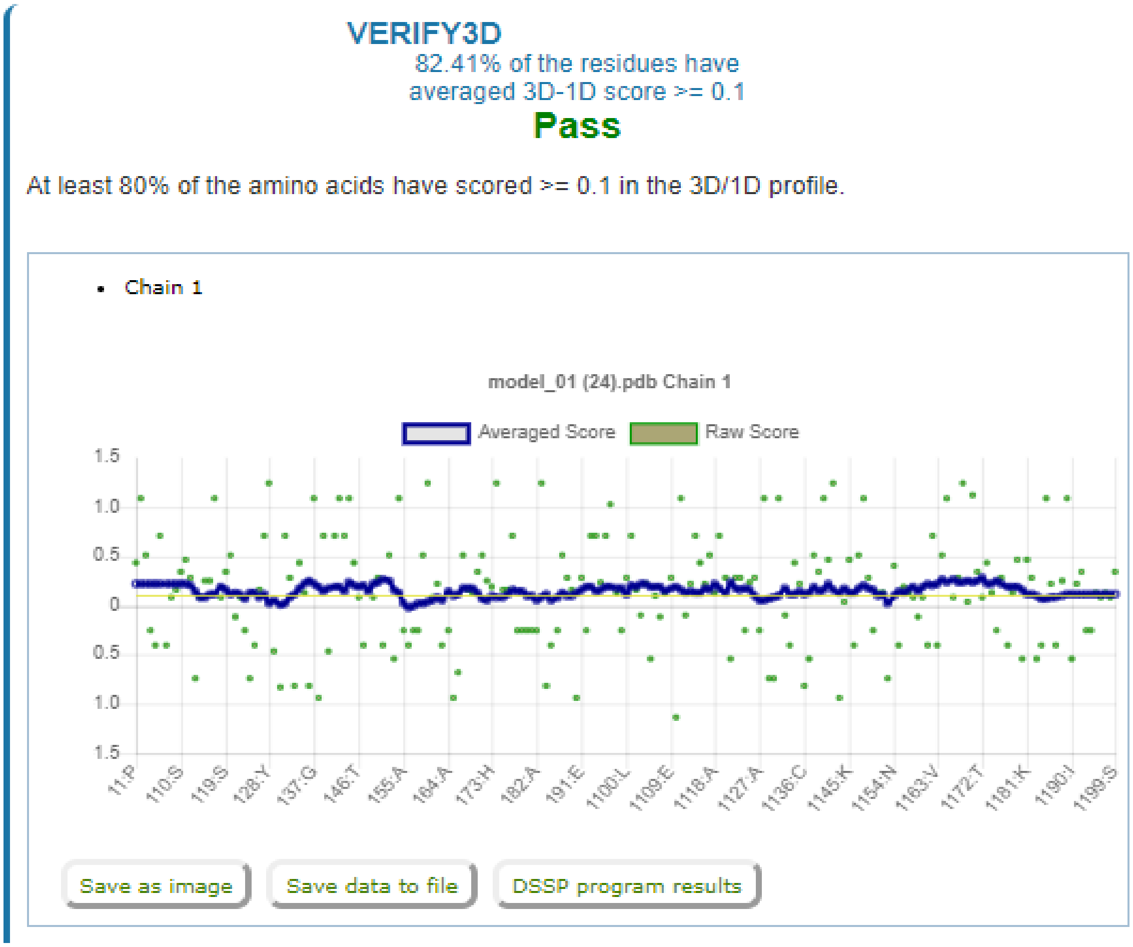
Residues in most favoured regions [A,B,L]	164	94.8%
Residues in additional allowed regions [a,b,l,p]	9	5.2%
Residues in generously allowed regions [~a,~b,~l,~p]	0	0.0%
Residues in disallowed regions	0	0.0%
	----	------
Number of non-glycine and non-proline residues	173	100.0%
Number of end-residues (excl. Gly and Pro)	1	
Number of glycine residues (shown as triangles)	14	
Number of proline residues	11	

Total number of residues	199	

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.

Figure 4.9.1 Ramachandran plot statistics

The average 3D-1D score of 82.41% of the remnants was superior to or equivalent to 0.1, demonstrating the moral uprightness and suitability of these structures. A verify3D score of higher than 80% denotes a high caliber model [19].



VERIFY3D

82.41% of the residues have averaged 3D-1D score >= 0.1

Pass

At least 80% of the amino acids have scored >= 0.1 in the 3D/1D profile.

- Chain 1

model_01 (24).pdb Chain 1

Figure 4.9.2 Verification of a 3d model via Verify3D 82.41% of the residues have averaged 3D-ID score >=0.1

The model’s projected Z -score of -6.07. The range of high-quality model’s z -scores for protein structures that have been experimentally confirmed.

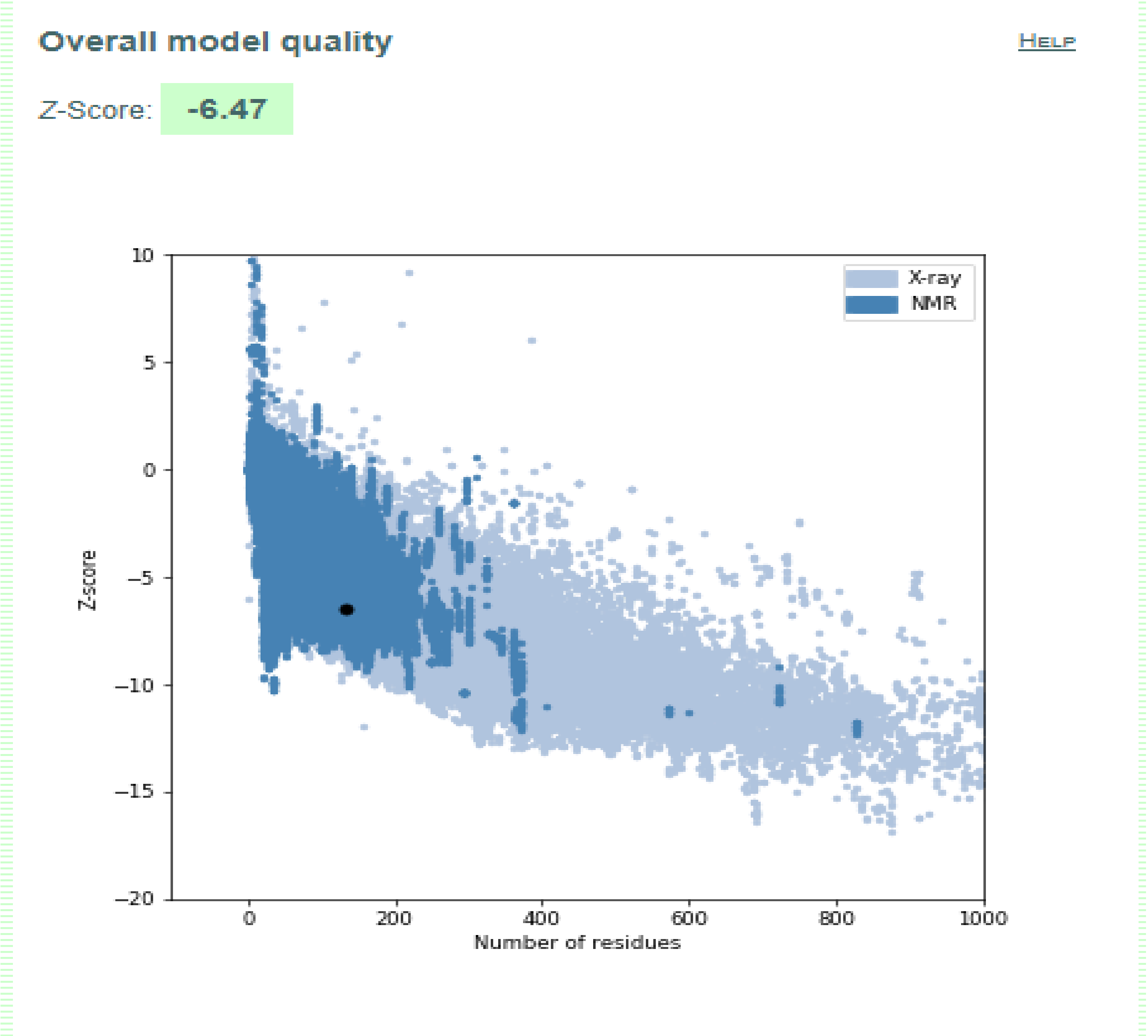
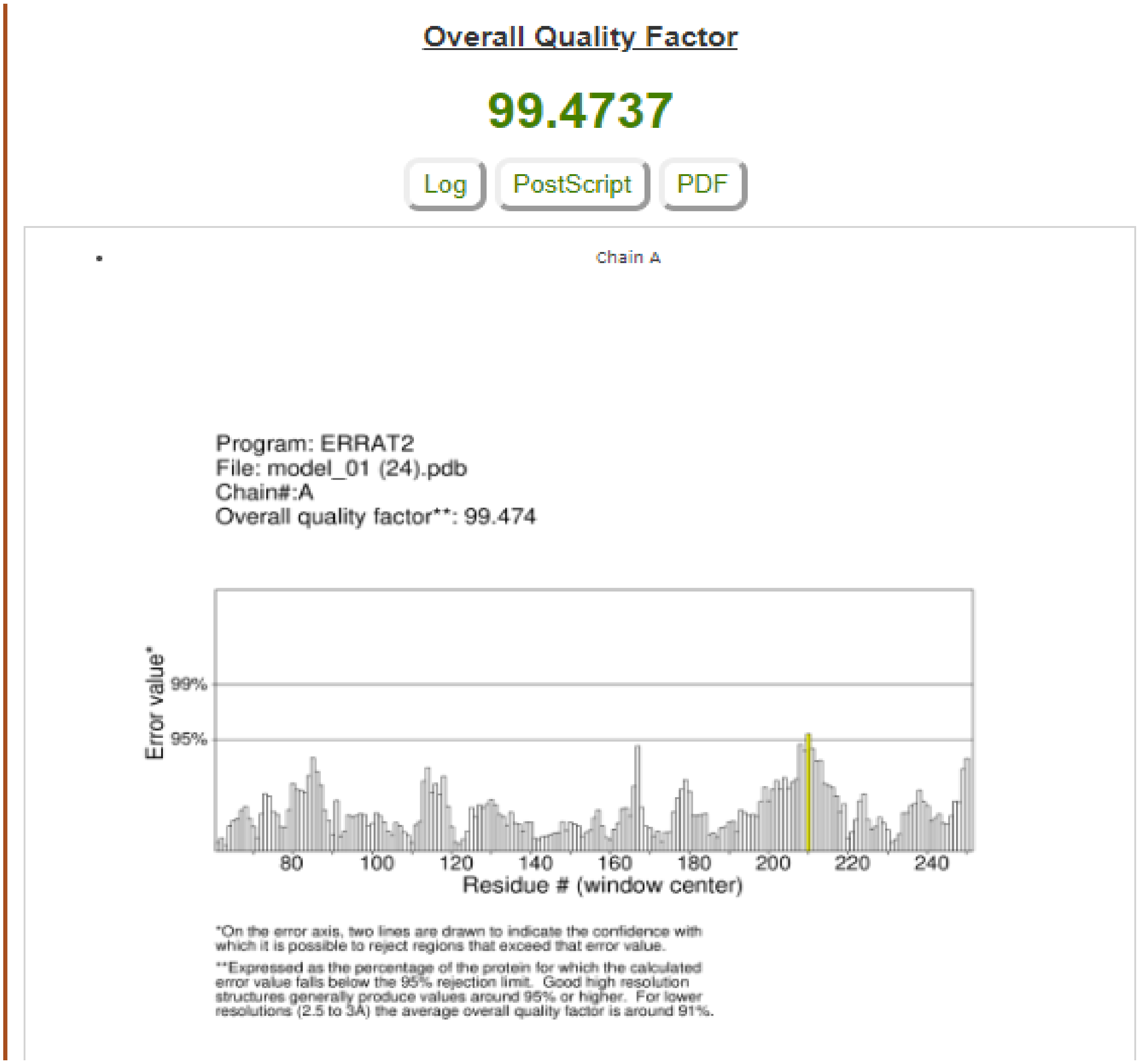


Figure 4.9.3 The link between HP and other proteins is shown using the z-score, with HP showing as black dots and other proteins as grey. Score for z: -6.47 (Prossa server)

Based on frequent atomic connections, Errat looked into the statistical distribution of non-bonded lings among various atom types. The overall quality factor for this model was 99.4737%, which is regarded as exceptional. Use of high-resolution models with Errat values of 95% or higher is advocated [20].



Overall Quality Factor

99.4737

Chain A

Program: ERRAT2
File: model_01 (24).pdb
Chain#:A
Overall quality factor**: 99.474

*On the error axis, two lines are drawn to indicate the confidence with which it is possible to reject regions that exceed that error value.

**Expressed as the percentage of the protein for which the calculated error value falls below the 95% rejection limit. Good high resolution structures generally produce values around 95% or higher. For lower resolutions (2.5 to 3A) the average overall quality factor is around 91%.

Figure 4.9.4: The overall quality factor plot (Errat) of the original model is 99.474% accurate. Red denoted difficult, yellow denoted less problematic, and gray denoted non-problematic.

The model was of good quality, according to the results of the overall quality assessment tests conducted by Errat, Prosa-web, verify3D, and Procheck (Ramachandran Plot)

Chapter-5

DISCUSSION

Discussion

Sequence taken from the NCBI database that is fictitious. The protein's physicochemical properties were then examined using ExPASy's ProtParam tool. It was determined that the protein had 255 amino acids, a molecular weight of 26861.54, a theoretical pI of 5.68, and a grand average of hydropathicity (GRAVY): 0.097 value is positive, indicating that the protein is water soluble. The ability of PSIPRED to get an average Q3 score of 76.5% using a rigorous cross validation approach has been demonstrated [21]. The **PSLpred** server predicted a cytoplasmic protein (a protein located within the cytoplasm of a cell) as the subcellular location of the target protein, with an estimated accuracy of about 71.1%. The domain protein from the **DsbG super family** that is involved in protein-disulfide isomerase [posttranslational modification, protein turnover, chaperones] was discovered in the CD search with an E-value of 4.11e-30 and an interval of 56-239.

Using SOPMA, the protein's primary and secondary structural characteristics were identified. According to these calculations, the protein contains 35.29% alpha helix, 20.00% extended strands, 7.06% beta turns, and 37.65% random coils. Using the template with the highest rating, the SWISS-MODEL server completed a 3D prediction with a 100% identification rate. The quality of the 3D models was evaluated using Prosa -web and the Saves v6.0 server. All three analyses—Errat, Verify3d, and Ramachandran plot—are available on the Saves v6.0 server.

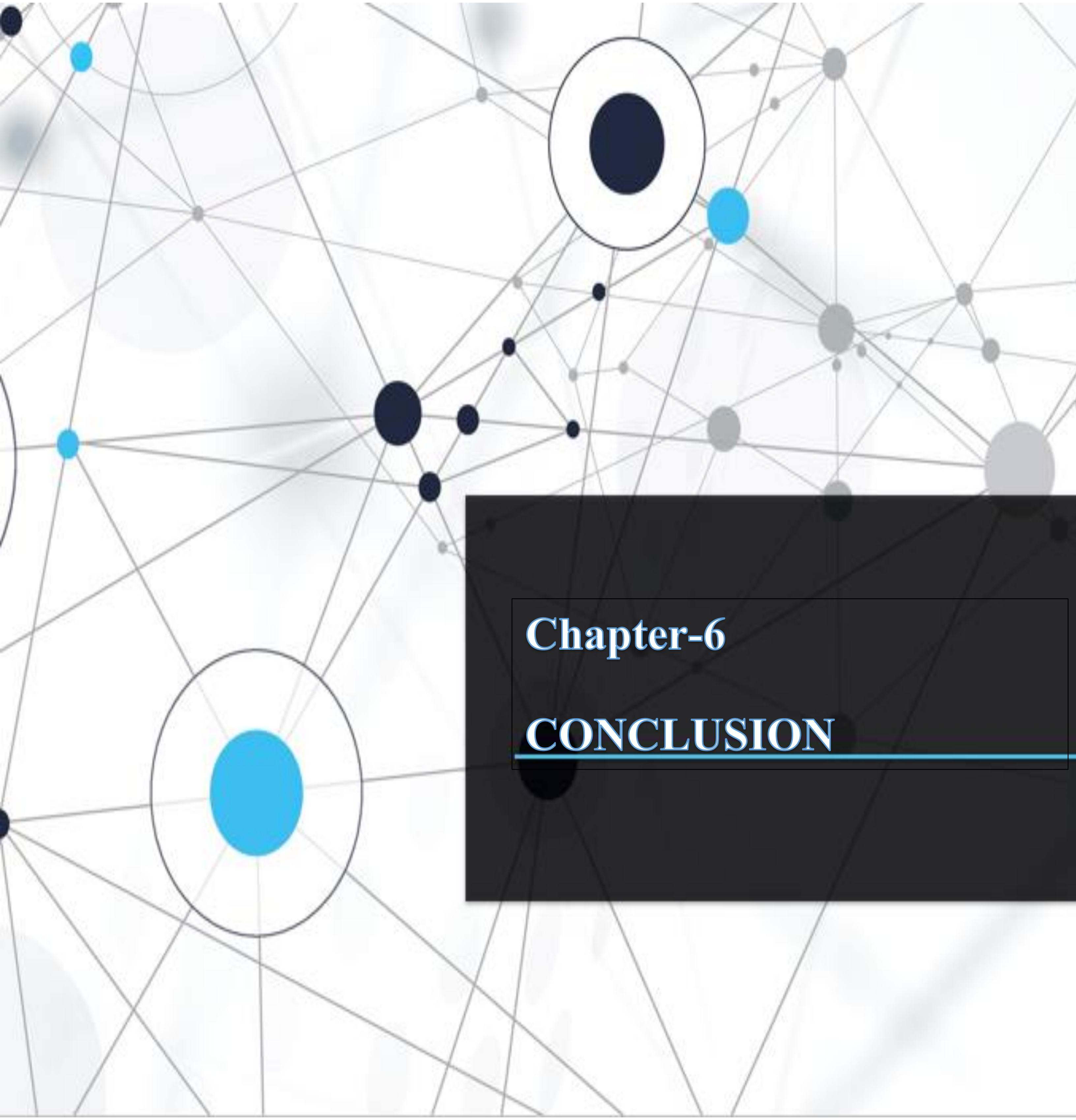
On the Prosa-web server, our z-score was more common in the NMR than the x-ray regions. If the z-score is within the range of values normally seen for genuine proteins of a similar size, the model is of outstanding quality; otherwise, the structure is incorrect. The z-scores for PDB protein structures range from -1 to -13, according to the Prosa-web plot graph.

The excellent level of quality of Prosa is demonstrated by the model's projected z-score of -6.07. the range of high-quality models' z-scores for protein structures that have been experimentally confirmed. A value of more than 90% inside the permitted zone is shown on the Ramachandran map, suggesting a high-quality model [19]. The model received an excellent rating from the Ramachandran Plot Statistics, with 94.8% of residues in the intended locations. Tool check3D, The average 3D-1D score of 82.41% of the residues was greater than or equal to 0.1, demonstrating the moral uprightness and suitability of these structures. A Verify3D score of higher than 80%

denotes a high caliber model [19]. Based on common atomic affiliations, Errat explored the statistical distribution of non-bonded links across a wide range of atom categories. The overall quality factor for Errat was 99.4737%, indicating that the model was excellent. A high-resolution model with an Errat value of 95% or higher is recommended [22].

The model was of good quality, according to the results of the overall quality assessment tests conducted by Errat, Prosa-web, Verify3D, and Procheck (Ramachandran Plot).

Here, **WP 003415010** and our target protein **OHO18223.1** are similar. 1 more closely; as a result, their functions can complement one another. The enzyme protein disulfide isomerase produces native disulfide pairs in secretory proteins (PDI). The essential enzyme PDI has the ability to disulfide isomerase. Protein Disulfide Isomerase is a redox protein from the thioredoxin class (PDI). Disulfide isomerase, thiol disulfide ox reductase, and redox dependent chaperone are the three enzymatic activities of PDI. The endoplasmic reticulum interior was where PDI was first discovered; it was then discovered in the cytoplasm and on cell membranes. This study will give a summary of current research linking the structural characteristics of PDI to its range of catalytic roles, as well as its physiological and clinical functions related to redox regulation and folded proteins [23]. The significance of this entire study lies in its ability to explain the structure and function of our potential target protein from *mycobacterium tuberculosis*.



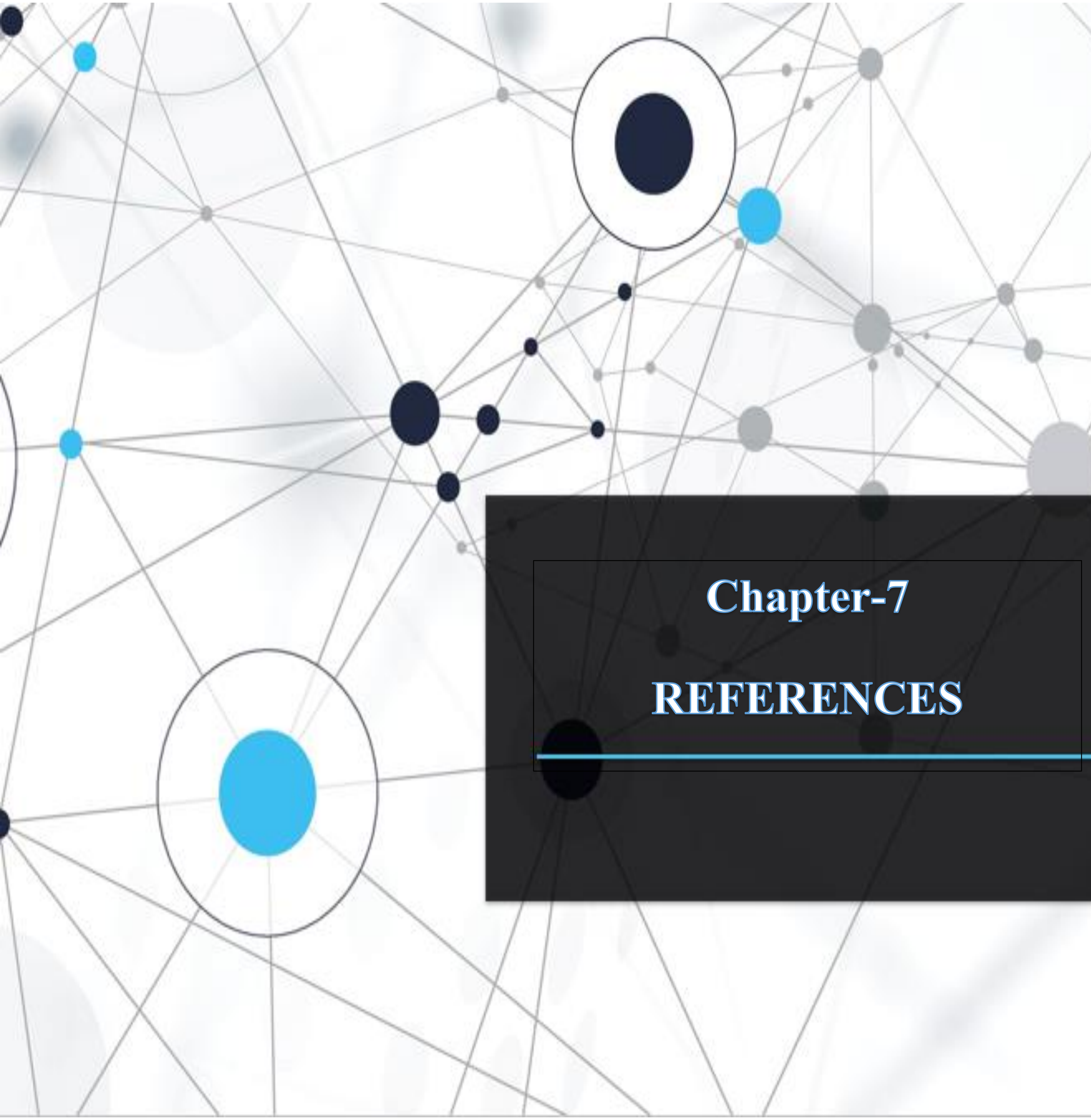
Chapter-6

CONCLUSION

Conclusion

Based on the findings of the overall evaluation of model quality and resemblance, the best projected models from SWISSMODEL were chosen. Future protein studies that require knowledge of protein structure dynamics can use these generated models as a reference.

This study's objectives were to determine the 3D structure of a hypothetical protein from the mycobacterium TB (Accession No. **OHO18223.1**), as well as to make recommendations for its possible applications. The most promising projected models from SWISSMODEL were selected based on the results of the overall model quality and similarity evaluation. Further protein research that requires knowledge of protein structure can use these developed models as a guide. Researchers identified a novel, cytoplasmic, stable protein with the multifunctional Protein Disulfide Isomerase (PDI) domain required for *Mycobacterium tuberculosis*. The functioning cell will die if we can prevent this protein from functioning in the cell. As a result, we can stop important processes, which may lead to the development of an anti-tubercular drug in the future. Further research into the structure and function of this protein may aid in the development of inventive medicines.



Chapter-7

REFERENCES

References

1. Enault F, Suhre K, Claverie JM. Phydbac" Gene Function Predictor": a gene annotation tool based on genomic context analysis. BMC bioinformatics. 2005 Dec;6(1):1-0.
2. Lubec G, Afjehi-Sadat L, Yang JW, John JP. Searching for hypothetical proteins: theory and practice based upon original data and literature. Progress in neurobiology. 2005 Sep 1;77(1-2):90-127.
3. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences. 1998 Dec 8;95(25):14863-8.
4. Banu S, Honoré N, Saint-Joanis B, Philpott D, Prévost MC, Cole ST. Are the PE-PGRS proteins of Mycobacterium tuberculosis variable surface antigens?. Molecular microbiology. 2002 Apr;44(1):9-19.
5. Mortimer TD, Weber AM, Pepperell CS. Signatures of selection at drug resistance loci in Mycobacterium tuberculosis. MSystems. 2018 Feb 27;3(1):e00108-17.
6. Chakaya J, Khan M, Ntoumi F, Aklillu E, Fatima R, Mwaba P, Kapata N, Mfinanga S, Hasnain SE, Katoto PD, Bulabula AN. Global Tuberculosis Report 2020–Reflections on the Global TB burden, treatment and prevention efforts. International journal of infectious diseases. 2021 Dec 1;113:S7-12.
7. Huang W. Computational methods for identifying and characterizing the human gene regulatory regions and cis-elements. North Carolina State University; 2005.
8. Gasteiger E, Hoogland C, Gattiker A, Duvaud SE, Wilkins MR, Appel RD, Bairoch A. Protein identification and analysis tools on the ExPASy server. Humana press; 2005.
9. Imai K, Hayat S, Sakiyama N, Fujita N, Tomii K, Elofsson A, Horton P. Localization prediction and structure-based in silico analysis of bacterial proteins: with emphasis on outer membrane proteins. Data Mining for Systems Biology: Methods and Protocols. 2013:115-40.
10. Geourjon C, Deleage G. SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. Bioinformatics. 1995 Dec 1;11(6):681-4.
11. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics. 2000 Apr 1;16(4):404-5.

12. Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Cassarino TG, Bertoni M, Bordoli L, Schwede T. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic acids research. 2014 Jul 1;42(W1):W252-8.
13. Dym O, Eisenberg D, Yeates TO. Detection of errors in protein models.
14. Alberts B, Bray D, Hopkin K, Johnson AD, Lewis J, Raff M, Roberts K, Walter P. Essential cell biology. Garland Science; 2015.
15. Navarro S, Villar-Piqué A, Ventura S. Selection against toxic aggregation-prone protein sequences in bacteria. Biochimica et Biophysica Acta (BBA)-Molecular Cell Research. 2014 May 1;1843(5):866-74.
16. Bertholet S, Goldszmid R, Morrot A, Debrabant A, Afrin F, Collazo-Custodio C, Houde M, Desjardins M, Sher A, Sacks D. Leishmania antigens are presented to CD8+ T cells by a transporter associated with antigen processing-independent pathway in vitro and in vivo. The Journal of Immunology. 2006 Sep 15;177(6):3525-33.
17. Stolf BS, Smyrnias I, Lopes LR, Vendramin A, Goto H, Laurindo FR, Shah AM, Santos CX. Protein disulfide isomerase and host-pathogen interaction. TheScientificWorldJOURNAL. 2011 Oct 18;11:1749-61.
18. Beissinger M, Buchner J. How chaperones fold proteins. Biological chemistry. 1998 Mar 1;379(3):245-59.
19. Haron FN, Azazi A, Chua KH, Lim YA, Lee PC, Chew CH. RESEARCH ARTICLE In silico structural modeling and quality assessment of Plasmodium knowlesi apical membrane antigen 1 using comparative protein models. Tropical Biomedicine. 2022;39(3):394-401.
20. Singh R, Gurao A, Rajesh C, Mishra SK, Rani S, Behl A, Kumar V, Kataria RS. Comparative modeling and mutual docking of structurally uncharacterized heat shock protein 70 and heat shock factor-1 proteins in water buffalo. Veterinary world. 2019 Dec;12(12):2036.
21. Tran NT, Jakovlić I, Wang WM. In silico characterisation, homology modelling and structure-based functional annotation of blunt snout bream (Megalobrama amblycephala) Hsp70 and Hsc70 proteins. Journal of Animal Science and Technology. 2015 Dec;57:1-9.

22. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics. 2000 Apr 1;16(4):404-5.
23. Ali Khan H, Mutus B. Protein disulfide isomerase a multifunctional protein with multiple physiological roles. Frontiers in chemistry. 2014 Aug 26;2:70.

Turnitin Originality Report

Processed on: 09-Apr-2023 09:04 +06
ID: 2059313092

Word Count: 8089

Submitted: 1

191-29-1438 By Mayanur Islam Nila

Similarity Index	Similarity by Source	
22%	Internet Sources:	18%
	Publications:	17%
	Student Papers:	12%

7% match (Internet from 12-Feb-2023)
https://www.researchgate.net/figure/Bos-taurus-Hsc70-PDB-ID-1yuw-A-structural-analog-backbone-trace-superimposed- upon_fig1_286400922

1% match (Internet from 11-Oct-2022)
http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/20.500.11948/1864/P05742.pdf?isAllowed=y&sequence=2

1% match (Internet from 22-Sep-2022)
https://downloads.hindawi.com/journals/bmri/2022/4302625.pdf

1% match (Internet from 18-Jan-2020)
https://www.mdpi.com/1422-0067/13/6/7283/htm

1% match (P. Bharat Siva Varma, Yesu B. Adimulam, Subrahmaniam Kodukula. "In silico functional annotation of ahypothetical protein from Staphylococcus aureus", Journal of Infection and Public Health, 2015)
P. Bharat Siva Varma, Yesu B. Adimulam, Subrahmaniam Kodukula. "In silico functional annotation of a hypothetical proteinfrom Staphylococcus aureus", Journal of Infection and Public Health, 2015

1% match (Internet from 30-Mar-2023)
https://msptm.org/files/Vol39No3/tb-39-3-009-Haron-F-N.pdf

1% match (Kaviya Parambath Kootery, Suma Sarojini. "Structural and functional characterization of a hypothetical proteinin the RD7 region in clinical isolates of Mycobacterium tuberculosis — an in silico approach to candidate vaccines", Journalof Genetic Engineering and Biotechnology, 2022)
Kaviya Parambath Kootery, Suma Sarojini. "Structural and functional characterization of a hypothetical protein in the RD7 region in clinical isolates of Mycobacterium tuberculosis — an in silico approach to candidate vaccines", Journal of GeneticEngineering and Biotechnology, 2022

< 1% match (Internet from 22-Feb-2023)
https://www.researchgate.net/publication/323527127_Wheat_Oxygen_Evolving_Enhancer_Protein_Identification_and_Characterization_of_Binding_Metalloprotein_of_Photosynthetic_Pathway_Involved_in_Regulating_Photosytem_II_Integrity_and_Network_of_Antioxid

< 1% match (Internet from 11-Oct-2022)
http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/7182/171-29-1050%20%2821%25%29%20clearence.pdf?isAllowed=y&sequence=1

< 1% match (Internet from 11-Oct-2022)
http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/20.500.11948/1861/P05724.pdf?isAllowed=y&sequence=2

< 1% match (Internet from 25-Oct-2022)
http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/4327/P14714%20%2820_%29.pdf?isAllowed=y&sequence=1

< 1% match (Internet from 11-Oct-2022)
http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/8497/171-29-1029.pdf?isAllowed=y&sequence=1

< 1% match (Internet from 11-Oct-2022)
http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/8060/171-29-

014%20%2818%25%29.pdf? isAllowed=y&sequence=1

< 1% match (Internet from 11-Oct-2022)
http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/7188/171-29-1086%20%2817%25%29.pdf? isAllowed=y&sequence=1

< 1% match (Internet from 11-Oct-2022)
http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/7185/171-29-1022%20%2817%25%29%20clearance.pdf?isAllowed=y&sequence=1

< 1% match (Internet from 11-Oct-2022)
http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/20.500.11948/1870/P05751.pdf?isAllowed=y&sequence=2

< 1% match (Internet from 12-Nov-2022)
https://downloads.hindawi.com/journals/tswj/2011/289182.pdf

< 1% match (Internet from 02-Mar-2023)
https://www.mdpi.com/2079-6382/12/2/339

< 1% match (student papers from 13-Nov-2022)

Submitted to The Hong Kong Polytechnic University on 2022-11-13

< 1% match (student papers from 22-Nov-2022)

Submitted to The Hong Kong Polytechnic University on 2022-11-22

< 1% match (student papers from 03-Jun-2022)
Submitted to Daffodil International University on 2022-06-03

< 1% match (student papers from 09-Apr-2018)
Class: Article 2018
Assignment: Journal Article
Paper ID: 943543955

< 1% match (student papers from 16-Apr-2018)
Class: April 2018 Project Report
Assignment: Student Project
Paper ID: 947570503

< 1% match (student papers from 21-Aug-2022)
Submitted to Adtalem Global Education on 2022-08-21

< 1% match (Internet from 22-Feb-2022)
https://discovery.researcher.life/download/article/5a5a9450ef5d37538417f17224a5e35c/full-text

< 1% match (Internet from 22-May-2009)
http://www.ist.temple.edu/~vucetic/cis595_spring2002/presentations/xioaoyong.txt

< 1% match (Internet from 17-Jun-2022)
https://www.frontiersin.org/articles/10.3389/fpls.2022.911761/full

< 1% match (Internet from 09-Jan-2023)
https://www.frontiersin.org/articles/10.3389/fgene.2020.00788/full

< 1% match (Mazandu, Gaston K., and Nicola J. Mulder. "Function Prediction and Analysis of Mycobacterium tuberculosis Hypothetical Proteins", International Journal of Molecular Sciences, 2012.)
Mazandu, Gaston K., and Nicola J. Mulder. "Function Prediction and Analysis of Mycobacterium tuberculosis Hypothetical Proteins", International Journal of Molecular Sciences, 2012.

< 1% match (Internet from 17-Aug-2022)
https://rjptonline.org/HTML_Papers/Research%20Journal%20of%20Pharmacy%20and%20Technology_PID_2021-14-12-10.html

< 1% match (Internet from 13-Mar-2023)
https://www2.mdpi.com/2073-4425/14/3/667/htm

< 1% match (Internet from 08-May-2009)
http://www.bioinformation.net/002/009000022008.pdf

< 1% match (Internet from 22-Mar-2023)
https://www.science.gov/topicpages/g/glutathione+disulfide+gssg

< 1% match (Internet from 15-Dec-2022)
https://ebin.pub/bioinformatics-and-functional-genomics-third-edition-1118581784-978-1-118-58178-0-9781118581698-1118581695-9781118581766-1118581768.html

< 1% match (student papers from 08-Aug-2021)
Submitted to Coventry University on 2021-08-08

< 1% match (Md. Foyzur Rahman, Rubait Hasan, Mohammad Shahangir Biswas, Jamiatul Husna Shathi et al. "A bioinformatics approach to characterize a hypothetical protein Q6S8D9_SARS of SARS-CoV", Genomics & Informatics, 2023)
Md. Foyzur Rahman, Rubait Hasan, Mohammad Shahangir Biswas, Jamiatul Husna Shathi et al. "A bioinformatics approach to characterize a hypothetical protein Q6S8D9_SARS of SARS-CoV", Genomics & Informatics, 2023

< 1% match (Haron F.N.. "In silico structural modeling and quality assessment of Plasmodium knowlesi apical membrane antigen 1 using comparative protein models", Tropical Biomedicine, 2022)
Haron F.N.. "In silico structural modeling and quality assessment of Plasmodium knowlesi apical membrane antigen 1 using comparative protein models", Tropical Biomedicine, 2022

< 1% match (student papers from 09-Jan-2022)
Submitted to De Montfort University on 2022-01-09

< 1% match (Ming Zhong, Ghanbar Mahmoodi Chalbatani, Meifang Deng, Qiuyi Li et al. "Functional Characterization and Development of Novel Human Kinase Insert Domain Receptor Chimeric Antigen Receptor T-cells for Immunotherapy of Non-Small Cell Lung Cancer", European Journal of Pharmaceutical Sciences, 2022)
Ming Zhong, Ghanbar Mahmoodi Chalbatani, Meifang Deng, Qiuyi Li et al. "Functional Characterization and Development of Novel Human Kinase Insert Domain Receptor Chimeric Antigen Receptor T-cells for Immunotherapy of Non-Small Cell Lung Cancer", European Journal of Pharmaceutical Sciences, 2022

< 1% match (Internet from 17-Aug-2022)
https://www.iosrjournals.org/iosr-jpbs/papers/Vol16-issue1/Series-4/C1601042127.pdf

< 1% match ("Prediction of Protein Structures, Functions, and Interactions", Wiley, 2008)

"Prediction of Protein Structures, Functions, and Interactions", Wiley, 2008

< 1% match (Lincon Mazumder, Md. Rakibul Hasan, Kanij Fatema, Md. Zahirul Islam, Sanjida Khanam Tamanna. "Structural and Functional Annotation and Molecular Docking Analysis of a Hypothetical Protein from Neisseria gonorrhoeae: An In-Silico Approach", BioMed Research International, 2022)
Lincon Mazumder, Md. Rakibul Hasan, Kanij Fatema, Md. Zahirul Islam, Sanjida Khanam Tamanna. "Structural and Functional Annotation and Molecular Docking Analysis of a Hypothetical Protein from Neisseria gonorrhoeae: An In-Silico Approach", BioMed Research International, 2022

< 1% match (Md. Fazley Rabbi, Saiwda Asma Akter, Md. Jaimol Hasan, Al Amin. " In Silico Characterization of a Hypothetical Protein from ATCC 12039 Reveals a Pathogenesis-Related Protein of the Type-VI Secretion System ", Bioinformatics and Biology Insights, 2021)
Md. Fazley Rabbi, Saiwda Asma Akter, Md. Jaimol Hasan, Al Amin. " In Silico Characterization of a Hypothetical Protein from ATCC 12039 Reveals a Pathogenesis-Related Protein of the Type-VI Secretion System ", Bioinformatics and Biology Insights, 2021

< 1% match (Seyyed Soheil Rahmatabadi, Keivan Mobini, Soudabeh Askari, Javad Najafian, Keyvan Karami, Bijan Soleymani, Ali Mostafaie. " In silico characterization of properties from ", Journal of Molecular Recognition, 2022)
Seyyed Soheil Rahmatabadi, Keivan Mobini, Soudabeh Askari, Javad Najafian, Keyvan Karami, Bijan Soleymani, Ali Mostafaie. " In silico characterization of properties from ", Journal of Molecular Recognition, 2022

< 1% match (Sofía Fraile, María Briones, Mónica Revenga-Parra, Víctor de Lorenzo, Encarnación Lorenzo, Esteban Martínez-García. " Engineering Tropism of toward Target Surfaces through Ectopic Display of Recombinant Nanobodies ", ACS Synthetic Biology, 2021)
Sofía Fraile, María Briones, Mónica Revenga-Parra, Víctor de Lorenzo, Encarnación Lorenzo, Esteban Martínez-García. " Engineering Tropism of toward Target Surfaces through Ectopic Display of Recombinant Nanobodies ", ACS Synthetic Biology, 2021

< 1% match ()
Prerna Goel, Tanya Panchal, Nandini Kaushik, Ritika Chauhan, Sandeep Saini, Vartika Ahuja, Chander Jyoti Thakur. "In silico functional and structural characterization revealed virulent proteins of strain SCHU4 ", Molecular Biology Research Communications

< 1% match (Internet from 24-Oct-2022)
https://link.springer.com/article/10.1007/s00239-009-9299-1?code=ea501aa3-b6b3-403f-a32b-0df9b0f897c3&error=cookies_not_supported

< 1% match ("Tuberculosis", Springer Science and Business Media LLC, 2023)
"Tuberculosis", Springer Science and Business Media LLC, 2023

< 1% match (Md. Saiful Islam, Shah Md. Shahik, Md. Sohel, Noman I. A. Patwary, Md. Anayet Hasan. " Structural and Functional Annotation of Hypothetical Proteins of O139 ", Genomics & Informatics, 2015)
Md. Saiful Islam, Shah Md. Shahik, Md. Sohel, Noman I. A. Patwary, Md. Anayet Hasan. " Structural and Functional Annotation of Hypothetical Proteins of O139 ", Genomics & Informatics, 2015

< 1% match (Internet from 15-Oct-2021)
https://genominfo.org/journal/view.php?doi=10.5808%2FGI.2020.18.3.e28

< 1% match (Internet from 31-Aug-2022)
https://mdpi-res.com/d_attachment/ijms/ijms-23-09575/article_deploy/ijms-23-09575.pdf?version=1661335074

< 1% match (Internet from 06-Aug-2021)
https://Scholar.ufs.ac.za/bitstream/handle/11660/1798/TheronCW.pdf

< 1% match (Internet from 29-Sep-2022)
https://vdoc.pub/documents/medical-imaging-in-clinical-applications-algorithmic-and-computer-based-approaches-297fn39vp9h0

Project Report On In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis A Project Paper, submitted to the Department of Pharmacy, Daffodil International University to complete the course of B.Pharm. Submitted To Department of pharmacy Faculty of Allied Health Science Daffodil International University Submitted By ID: 191-29-1438 Batch: 21B Department of Pharmacy Daffodil International University Date of Submission: April, 2023 APPROVAL Both the presentation and the subject matter of the project In silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis, which was submitted to the Department of Pharmacy, Faculty of Allied Health Sciences, Daffodil International University, have received praise. It has been acknowledged as meeting a portion of the criteria for the Bachelor of Pharmacy degree. Board of Examiners: ………………………………….. Professor Dr. Muniruddin Ahmed Head of the department of Pharmacy Faculty of Allied Health Science Daffodil International University …………………………. Internal Examiner-I ……………………………….Internal Examiner-II ……………………………. Internal Examiner-III CERTIFICATE In order to prove that the research findings contained in this project are unique and have never previously been submitted in their whole for a degree from this institution, we thus testify to their originality. The foundation for the whole current work, which has been provided as a project work for the degree of Bachelor of Pharmacy, is the author's (Mayanur Islam Nila, ID: 191-29-1438) individual research results. Supervised By:__________________________Mohammad Touhidul Islam Lecturer (Senior Scale) Department of Pharmacy Faculty of Allied Health Sciences Daffodil International University DECLARATION I hereby declare that, this project report is done by me under the supervision of Mohammad Touhidul Islam, Lecturer (Senior Scale), Department of Pharmacy, Daffodil International University, impartial fulfillment of the requirements for the degree of Bachelor of Pharmacy. I am declaring that this project is my original work. I also declare that neither this project nor any part therefore has been submitted elsewhere for the award of Bachelor or any degree. Submitted By: Name: Mayanur Islam Nila ID: 191-29-1438 Batch: 21 DEDICATION I would like to dedicate my work to the Almighty GOD, My Beloved Parents And My supervisor. ACKNOWLEDGEMENT As a human being, I want to begin by thanking God for giving me the opportunity to complete this assignment successfully. I want to thank my supervisor, Mr. Mohammad Touhidul Islam, Lecturer (Senior Scale), Department of Pharmacy, Daffodil International University, for his wonderful guidance in repairing this project and ensuring its success. His suggestions and guidance were really helpful in getting the job finished. I would like to express my heartfelt thanks to Professor Dr. Muniruddin Ahamed, Head of the Pharmacy Department at Daffodil International University, for providing me the opportunity to finish my project work. I also want to thank the other faculty members from the pharmacy department of Daffodil International University for working with me so well. Then, I'd want to thank my parents and friends for their assistance, counsel, and inspiration at different phases of the project's completion. Not least, I want to express my gratitude to my students for their constant support. Abstract One of the oldest illnesses known to affect humans, tuberculosis (TB), kills two million people annually and is still the main cause of death. However, Mycobacterium tuberculosis, which is prevalent in aerosol droplets, deposits on lung alveolar surfaces, causing TB to be largely a pulmonary illness. Despite the fact that TB may damage the bone, the neurological system, and several other organ systems, its primary symptom is lung illness. There are a number of potential consequences as the illness advances from this point , most of which are determined by the host immune system's response. Researchers are identifying targets in M. TB that will aid in the development of these urgently required anti-tubercular medications using the whole M. tuberculosis genome structure as well as innovative genetic and physiological approaches. In order to comprehend the function that a putative protein (OHO18223.1) discovered in Mycobacterium TB plays in the organism, the current study aims to investigate the structural and functional evaluation of the protein. Using an in-silico technique, homology modeling was utilized to build the 3D structure, and P fam, Interpro, and Uniport were used to profile the functions. The putative protein's primary and secondary structures were examined, and they revealed that it was stable, confined within the cytoplasm, and included a significant number of random coils. The 3D models that the SWISS-MODEL server predicted are 62.17 percent comparable to the template that received the best score. The SAVES v6.0 server and Prosa-web were used to assess the 3D model quality. The protein model has remarkable quality, as shown by the Ramachandran plot- based overall quality assessment results from Prosa-web, Verify3D, and Procheck. According to the websites P fam, Interpro, and Uniport, the putative protein is a cytoplasmic protein that breaks down unwanted, damaged, and improperly folded intracellular proteins, hence preserving the quality of protein in the cell. Finally, we proposed that more research into the structure and function of the protein may result in brand-new therapies. Keywords: Function prediction; Structure assessment; Mycobacterium tuberculosis; Hypothetical protein; Homology modeling; novel Drug. Chapter One: Introduction S. No. 1 Table of Content Topic Page No. Introduction 15-17 Chapter Two: Purpose of my study S. No. Topic Page No. 2 Purpose of my study 19 Chapter Three: Materials and Methods S. No. Topic 3.1 Sequence Retrieval or Selection of the Hypothetical Protein Page No. 22-23 3.2 Physiochemical Properties 23-24 3.3 Subcellular localization and solubility prediction 24-26 3.4 Conserved domains search and Function prediction by Conserved domains analysis 26-28 3.5 Homology Searching 28-29 3.6 Multiple sequence alignment and Phylogenetic Tree Prediction 29-30 3.7 Secondary structure determination 30-31 3.8 Three-dimensional structure prediction 32 3.9 Tertiary Structure Validation and Model Quality Assessment 32-33 Chapter Four: Results S. No. Topic 4.1 Sequence Retrieval or Selection of the Hypothetical Protein Page No. 35 4.2 Physiochemical Properties 35 4.3 Subcellular localization and solubility prediction 36 4.4 Homology Searching 37 4.5 Multiple sequence alignment and Phylogenetic Tree Prediction 37-39 4.6 Protein family and Function prediction 39-40 4.7 Secondary structure determination 40-43 4.8 Three-dimensional structure prediction 44 4.9 Tertiary Structure Validation and Model Quality Assessment 44-48 Chapter Five: Discussion S. No. Topic Page No. 5 Discussion 50-51 Chapter Six: Conclusion S. No. 6 Topic Conclusion Page No. 53 Chapter Seven: References S. No. 7 Topic Page No. References 55 - 57 S. No. 1.1 3.1.1 3.2.1 List of Figure Figure name Cell structure of Mycobacterium Tuberculosis National Center for Biotechnology Information used for sequence retrieval Expasy's Protparam tool 3.3.1 Pslpred subcellular localization predictor 3.3.2 SOSUI- GramN subcellular localization predictor Page No. 16 23 24 25 26 3.4.1 National Center for Biotechnology Information sever for 27 search conserved domain 3.4.2 EMBL-EBI P fam server for Function prediction 27 3.4.3 Interpro Classification of protein families 28 3.5.1 NCBI Blast Tool Webpage for Protein blast 29 3.6.1 Clustal omega tools Used for Multiple Sequence Alignment 30 3.7.1 Sopma Server used for secondary structure prediction 31 3.7.2 Psipred server tool for Secondary structure prediction 31 3.8.1 SWISS-MODEL server for predicting 3D structure of Hypothetical protein 32 3.9.1 Saves v6.0 was analyzed in order to locate faults in the 3D structure 33 4.3.1 Pslpred subcellular localization result 36 4.3.2 SOSUI-GramN subcellular localization result 36 4.4.1 Top 10 homology results from Blastp tools 37 4.5.1 Top 10 multiple sequence alignment obtained by using Clustal omega tools 38 4.5.2 Phylogenetic Tree Prediction 39 4.6.1 Conserved domain result 39 4.6.2 Interpro server result 40 4.7.1 this figure shows individual position of alpha-helix, beta - 42 turn, extended strand, and random coil in sequence. 43 4.7.2 Secondary structure of hypothetical protein 4.8.1 3D structure of Hypothetical protein 44 4.9.1 Ramachandran Plot Statistics 45 4.9.2 Verification of a 3d model via Verify3D 82.41% of the residues have averaged 3D-ID score >=0.1 46 4.9.3 Based on z-score, the relationship between HP and other proteins is shown, with HP appearing as black dots and other proteins appearing as grey. z-score: -6.47 (Prosa server) 4.9.4 The original model's overall quality factor plot (ERRAT) has an accuracy of 99.474%. Gray showed non-problematic, red indicated troublesome, and yellow suggested less problematic zone. 47 48 List of Table S. No. Table name Page No. 1.1 List of bioinformatics tools and databases used in this study for structural and functional analysis of the HP 21-22 4.2.1 Physicochemical properties of hypothetical protein 35 Introduction There is a lack of functional labelling for many recently decoded proteins despite the ever- increasing quantities of biological data, including raw data, such as genomic sequences, and functional genomic data from high-throughput studies. For instance, up to 40% of recognized proteins in the majority of bacterial genes are classified as "uncharacterized," "unknown," or "hypothetical" proteins [1]. HPs are predicted proteins, proteins that have not been demonstrated to exist by experimental protein chemistry data but are expected from nucleic acid patterns [2]. Scientists have used several computational techniques to develop several methods for forecasting protein function. These techniques have been made possible by utilizing series homology, phylogenetic analysis, interactions between proteins, interactions between proteins and ligands, similarities between active site residues, commonly conserved domains and motifs, phosphorylation sites, and patterns of gene expression. The conventional method of figuring out function, however, relies on sequence similarity and programs like BLAST, FASTA, and PSI- BLAST. There is no scientific or biological evidence for the existence of HPs, which are hypothetical proteins predicted from nucleic acid sequences. Moreover, these proteins share little in common with known, annotated proteins. Only a small number of HPs have managed to endure over time, and they can be found in various evolutionary lineages. While HPs constitute a considerable portion of genes in sequenced microbial genomes, their complete functional characterization and detailed protein chemistry remain incomplete [3]. In HPs, there are two categories. Uncharacterized protein families (UPFs) make up one class, whereas domains of uncertain function make up the other (DUFs). Although their experimental structures have not been described or connected to any known genes, unknown proteins do indeed exist. Despite being proteins that have been experimentally identified, DUFs lack any structural or functional domains. They might have transmembrane sections or coiled-coil features that prevent the function from being assigned. Studying the function of proteins with unidentified roles has several advantages, including the evaluation of novel domains and motifs, the identification of extra protein pathways and cascades, and the finding of novel conformational orientations of 3-dimensional structures. Potential pharmacological targets may be discovered in these novel domains. Moreover, high throughput techniques, such as systematic synthetic lethal analysis, are employed, both microarray gene expression patterns and protein complex identification by mass spectrometry are useful. You can also use phylogenetic analysis of proteins across numerous genes to determine function. Genes with comparable roles are apt to be co- expressed, which is why the wide, common technique of grouping gene-expression patterns [4] is so popular. Many protein domains engage in an organism's biochemical processes and have unclear roles; however, they may also be detrimental. The function of a protein can rarely alter due to changes such as insertions, deletions, and replacements. The study's primary aim is to find and categories a protein region with an unknown function using computational technologies. The 2020 Global Tuberculosis Survey revealed that there were 10 million new cases of tuberculosis (TB) in 2019. Out of these, 1.2 million were people who did not have HIV, while the remaining 2,08,000 were people who were HIV- positive. The BCG vaccine is the only one that has been developed to prevent tuberculosis, but it does not work very well in adults. Figure 1.1 Cell structure of Mycobacterium Tuberculosis [8] Despite the fact that many different immunizations are still in the testing stage, there is no evidence to suggest that any of them are particularly effective. Even though it was one of the most promising potential vaccines, MVA85A was unable to offer HIV patients in a clinical study any significant protection from the illness. These proteins are predicted to be essential for MTB survival and growth in their chosen environments as well as for mediating interactions between mycobacterium and host cells, according to some annotation predictions for these proteins, which are expressed in response to the shifting microenvironments that the pathogen encounters [5]. Because they are a part of a vastly expanded family of protein families, the functions of these proteins are essential for expanding our understanding of this species. Moreover, Mycobacterium TB shows signs of treatment resistance, which makes the condition harder to treat [6]. More than half of M. tuberculosis's first complete genome, which was decoded in 1998, appears to contain potential proteins. Because the species has coevolved with other creatures and people, members of the M. TB complex have distinct harmful impacts despite having 99% of their DNA code. It has been explicitly hypothesised that a number of these "hypothetical" proteins, such as those in the ProGlu or Pro-Pro-Glu (PE/PPE) family, play a crucial role in the internal lifestyle of Mycobacterium TB and may aid in its survival under a variety of conditions. To anticipate the functions of many of the proteins that Mycobacterium TB is expected to generate, we created a functional interaction network for its proteins. In this research, we compared the network properties of hypothesised proteins to those of known proteins in order to identify statistically important classifications. Based on the expected molecular roles of these proteins, we also performed functional enrichment analysis on them. Numerous membrane-anchored genes implicated in lipid metabolism were found among the infectious mycobacterial genes that were examined. On the pathogenesis and pathogenicity of the bacteria, potential proteins that has high GC are believed to have a significant biochemical effect [7]. Expasy, Prot Param, NCBI CDD blast, P farm, Clustal omega, Sopma, Psipred, Procheck, SWISS-Model, Raptor X, Phyre 2, and ProSA -web server were among the computational tools that could be used to analyse the molecular and functional analysis of a new putative protein (BBW91 17915) in the current study. The outcomes of the in silico experiment were compared using three positive controls and three negative controls, respectively. Purpose of my study The goal of my research in 2019, there were ten million newly diagnosed instances of tuberculosis (TB), among them about 1.2 million people tested HIV- negative and 2.08,000 tested positive, according to the Global Tuberculosis Report, 2020. BCG, the as-yet-uncreated anti-TB antibody, isn't particularly effective against tuberculosis in adults. Despite the fact that many antibodies are still in the testing phase, notable survival has not yet been described. Additionally, Mycobacterium TB looks resistant to frequently prescribed medications, making the disease more difficult to treat [9]. Finding new therapeutic targets for Mycobacterium tuberculosis is made easier with the help of computational drug discovery research. His investigation of the MTBC genome using SMRT sequencing reveals that the genome is 99% identical and well conserved. Only 7 secreted proteins have been described to date, despite the fact that 51 M. tuberculosis secreted proteins have been identified using computational techniques. Bioinformatics methods have been used to study various groups of hypothetical proteins, including enzymes, transporters, receptors, and structural proteins. Few in silico methods have been used to forecast vaccine potential, ribosome binding sites, GTP binding regions, or any other properties of M. TB proteins, but many M. tuberculosis proteins have received examination of mutations, useful research, and structural forecast. To create new, highly immunogenic multi-epitope subunit vaccines or drugs against tuberculosis, reverse vaccine trials are being done. 2.1 This study's objective is to: -Discovery of new biotechnologically significant proteins as well as novel drug design. - Prediction of the structure of hypothetical proteins. - Determination of hypothetical proteins' functional roles. - Describe the physiochemical characteristics of hypothetical proteins. The significance of this entire work lies in its ability to shed light on the structure and function of the targeted Mycobacterium tuberculosis proteins. Materials and Methods This inquiry involves two successive phases. Homology Search, Physiochemical Parameters, Subcellular Localization and Stability Prediction, Identification and Retrieval of Unique HPs from Internet Databases, were all included in PHASE I. PHASE II involved the structural and functional analysis of the selected proteins, which included identifying their interactions with other proteins, classifying protein families, looking for signal peptides, homology modeling, and validating the 3D structures of the proteins. A list of the bioinformatics sources and instruments used to functionally annotate Mycobacterium tuberculosis (Accession No. OHO18223.1) HPs can be found in Supplementary Table 1.1 Table 1.1 The datasets and bioinformatics tools used in this study's structural and functional investigation of the HP are listed below. Tools/Servers A)Sequence Retrieval- Download FASTA of Target Protein 1. NCBI B) Determination of physicochemical Properties 1. ExPASy – ProtParam C) Subcellular Localization 1. PSLpred 2. SOSUI-GramN D) Function Predictions 1. Conserved domain database 2. INTERPRO URL https:// www.ncbi.nlm.nih.gov/ http:// web.expasy.org/protparam https :// webs.iiitd.edu.in/raghava/pslpred/sub mit.html https:// harrier.nagahama-i- bio.ac.jp/sosui/sosuigramn/sosuigramn submit.htm http:// www.ncbi.nlm.nih.gov/Structure/cdd/ wrpsb.cgi Function Sequence retrieval or selection of the hypothetical protein and download FASTA of target protein To analyze physicochemical characterization of Hypothetical Protein Gram-negative bacteria's five primary subcellular localizations predicting. A method for looking for functional regions in a series and Family relationship identification. 3. Pfam http:// www.ebi.ac.uk/interpro/ http:// pfam.xfam.org/ E) Homology Searching https:// 1. PBLAST blast.ncbi.nlm.nih.gov/Blast.cgi# Searches for sequence similarity F) Multiple Sequence Alignment and Phylogenetic Tree Prediction 1. Clustal Omega https:// www.ebi.ac.uk/Tools/services/web_cl ustalo/toolform.ebi Accomplishes multiple sequence alignment and phylogenetic tree prediction G) Secondary Structure Prediction 1. PSIPRED 2. SOPMA http:// bioinf.cs.ucl.ac.uk/psipred https:// npsa-prabi.ibcp.fr/cgi- bin/npsa_automat.pl?page=/NPSA/nps a_sopma.html Accurately forecasts amino acids for a three-state secondary structure description of the secondary structure. H) Tertiary Structure Prediction 1. SWISS-Model 2. RaptorX 3. Phyre2 https:// swissmodel.expasy.org/interactive/7h XdRN/models/ http:// raptorx6.uchicago.edu/ContactMap/ http:// www.sbg.bio.ic.ac.uk/~phyre2/html/p age.cgi?id=index Evaluate the model accuracy in the database, search for different blueprint designs and then dynamically build models. I) 3DStructure Validation 1. PROCHECK 2. ProSA-web 3. Verify3D 4. ERRAT https:// prosa.services.came.sbg.ac.at/prosa.ph p Validate using the structural analysis and verification service. https:// saves.mbi.ucla.edu/ PHASE I 3.1 Sequence Retrieval or Selection of the Hypothetical Protein Uncharacterized Mycobacterium TB proteins were selected from the NCBI protein database (https:// www.ncbi.nlm.nih.gov/) using the search term "hypothetical proteins." Using blast analysis, it was possible to compare the degree of resemblance to other alleged fictitious proteins. In this study, a putative Mycobacterium tuberculosis protein with 255 amino acid residues, accession number OHO18223.1, was employed. For further research, the protein's main sequence in FASTA format was obtained. Figure 3.1.1: National Center for Biotechnology Information used for sequence retrieval 3.2 Physiochemical Properties The Protparam server from Expasy was used to analyze physicochemical characterization (http:// us.expasy.org/tools/protparam). Molecular weight, predicted pI, the ratio of amino acids to atoms, the extinction coefficient, the expected half-life, and the instability index, (which measures the integrity of the protein structure; a score of 40 indicates that the protein is stable), aliphatic index, and GRAVY parameters were calculated by ProtParam (if this score is negative, it is mine soluble in water; if it is positive, it is not soluble in water). To understand protein characteristics, use this resource. Figure 3.2.1 ExPASy's ProtParam tool 3.3 Subcellular Localization and Solubility Prediction Gram-negative bacteria's five primary subcellular localizations are predicted using PSLpred (https:// webs.iiitd.edu.in/raghava/pslpred/submit.html), an SVM -based method (cytoplasm, inner membrane, outer membrane,

extracellular, and periplasm). This method employs a variety of SVMs based on the composition of 33 physicochemical characteristics, PSI-Blast evolutionary data, and the dipeptide makeup, protein-repelling characteristics, and features of amino acid composition. The overall prediction accuracy for these SVM modules is 6%, 86%, 83%, and 68%, respectively. Moreover, a hybrid approach-based 3VM module has been developed based on all of the aforementioned characteristics of a protein. With a 91% total accuracy rate, the hybrid module outperforms all other methods currently in use for determining the intracellular location of bacterial proteins. Figure 3.3. 1 PSLpred subcellular localization predictor Gram-negative bacteria's distribution of proteins within cells is predicted with high efficiency using SOSUI -GramN (https:// harrier.nagahama-i-bio.ac.jp/sosui/sosuigramn/sosuigramn submit.htm). The method only employs the molecular properties of the N- and C-terminal signal sequences as well as the complete sequence and does not require the sequence similarity information of any known sequences. For the localization of extracellular, outer, periplasmic, inner, and cytoplasmic mediums, the prediction system's accuracy was 92.3%, 89.4%, 86.4%, 97.5%, and 93.5%, respectively. Figure 3.3. 2 SOSUI-GramN subcellular localization predictor PHASE- II 3.4 Conserved domains search and Function prediction by Conserved domains analysis In proteins' chemical development, protein domains can be grouped into an evolutionary taxonomy and viewed as units. This list is attempted to be assembled, and similar domain models are arranged in a hierarchical manner by the Conserved Domain Database (CDD) at NCBI. The CDD program is used to do the conserved domain analysis, and it may be found online at (www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). P fam (http:// pfamlegacy.xfam.org/), Interpro Scan (http:// www.ebi.ac.uk/Tools/services/web/toolform.ebi?tool=iprscan5]), and other tools are used for domain analysis as well as Uniport (www.uniprot.org/blast). Figure 3.4.1: National Center for Biotechnology Information sever for search conserved domain Figure 3.4.2: EMBL-EBI Pfam server for Function prediction Functional analysis of proteins is provided by Interpro (https:// www.ebi.ac.uk/interpro), which classifies proteins into families and forecasts their domains and important sites. In order to categorize proteins in this way, the Interpro collaboration uses prediction models, known as signatures, supplied by several databases (referred to as member databases) [10]. Figure 3.4.3 InterPro Classification of protein families 3.5 Homology Searching The most popular and effective method for characterizing newly found genomes is to search for nucleotide homology, usually with BLAST. By finding excess similarity, a numerically significant similarity that indicates common lineage, sequence similarity searches can discover "homologous" proteins or genes. More than 80% of metagenomics sequence samples commonly display substantial homology to proteins in sequence databases because of the extent of contemporary protein sequence databases. To improve accessibility and performance, BLAST's public interface at the NCBI site (http:// www. ncbi.nlm.nih.gov/blast) was recently rebuilt. Figure 3.5.1: NCBI BLAST Tool Webpage for Protein blast 3.6 Multiple nucleotide matching and Predicted Phylogenetic tree Multiple sequence alignment is typically accomplished using Clustal series programs. The use of the mBed method for building guide-trees allows it to manage exceptionally large quantities (many tens of thousands) of DNA/RNA or protein sequences. With the help of planted guidance trees and HMM profile-profile methods, the brand-new multiple sequence alignment software Clustal Omega creates arrangements involving three or more [11]. This method has been one of the most widely used in bioinformatics for decades because it is a prerequisite for the majority of phylogenetic or comparative analysis of homologous genes or proteins. Figure 3.6.1 Clustal Omega tools Used for Multiple Sequence Alignment 3.7 Secondary Structure Determination The URL for this page is (https:// npsa-prabi.ibcp.fr/cgi-bin/npsa automat.pl?page=/NPSA/npsa sopma.html). Alignment-Based Self-Adjusted Forecast Technique It has been proposed that SOPMA could improve predictions of protein secondary structure. In a collection of 126 sequences of non-homologous (less than 25% identical) proteins, SOPMA correctly forecasts 69.5% of using amino acids to describe the tertiary structure in three states (-helix, -sheet, and coil) [12]. Users can enter a protein sequence, make a projection of their choice, and receive the prediction results both 30 textually by e-mail and visually on the web using the PSIPRED technique for forecasting protein shape (http:// bioinf.cs.ucl.ac.uk/psipred/). Figure 3.7. 1 SOMPA Server used for secondary structure prediction Figure 3.7. 2: PSIPRED server tool for Secondary structure prediction 3.8 Three-dimensional Structure Prediction Users of the SWISS-MODEL Repository can evaluate the model accuracy in the database, search for different design layouts, and interactively construct models using SWISS-MODEL server (http:// swissmodel.expasy.org/workspace/). The first completely automatic protein homology modelling tool was SWISS-MODEL and it has advanced continuously over the past 25 years (25– 30). The modelling of homo- and heterogenic compounds using the amino acid patterns of the contact partners as a beginning point has recently been added to its modelling capabilities. Figure 3.8.1 SWISS-MODEL server for predicting 3D structure of Hypothetical protein 3.9 Tertiary Structure Validation and Model Quality Assessment The modelled OHO18223.1 was validated using the structural analysis and verification service SAVES v6.0 (https:// saves.mbi.ucla.edu/). The principal structural model was checked using the ERRAT, Verify3D, and PROCHECK tools, which are all a part of the structural assessment and confirmation server SAVESv6.0 [13], to look for problems with the structure's three-dimensional representation. ProSA, also known as protein structure analysis, is a method used to improve and verify experimental protein structures as well as forecast and model protein structure. Also, the PROCHECK software tool provides a complete analysis of a protein structure's stereochemistry. Its outputs contain a complete residue-by-residue listing as well as a variety of graphs in PostScript format. Figure 3.9. 1 SAVESv6.0 was analyzed in order to locate faults in the 3D structure Result 4.1 Sequence retrieval A hypothetical protein sequence was chosen at random from the NCBI database using search term "Hypothetical Protein of Mycobacterium Tuberculosis" and the resulting protein sequence is shown below. GenBank: OHO18223.1 >OHO18223.1 hypothetical protein BBW91_17915 [Mycobacterium tuberculosis]
MADKSKRPPRFDLKSADGSFGRLVQIGGTTIVVVFAVVLVFYIVTSRDDKKDGVAGPGD
AVRVTSSKLVTQPGTSNPKAVVSFYEDFLCPACGIFERGFGPTVSKLVDIGAVAADYTM
VAILDSASNQHYSSRAAAAAYCVADESIEAFRRFHAALFSKDIQPAELGKDFDPDNARLIE
LAREAGVVGKVPDCINSGKYIEKVDGLAAAVNVHATPTVRVNGTEYEWSTPAALVAKI KEIVGDVPGIDSAAATATS 4.2 Physicochemical properties analysis For the hypothetical protein OHO18223.1, the ProtParam tool evaluated a number of physicochemical properties, which are shown in the table below. Table 4.2.1 physicochemical properties of hypothetical protein Accession Number Molecular Estimated number of weight halflife amino acids OHO18223.1 255 26861.54 30 hr Theor Asp etical + pI Glu 5.68 30 Arg + Lys 27 Aliphatic Instability Grand average index index of hydropathici ty (GRAVY) 89.53 26.64 0.097 It was anticipated that the protein would have 255 amino acids, have a 26861.54 molecular weight and have a GRAVY score of 0.097, which indicates that it is water soluble. With a calculated instability index (II) of 26.64, the protein falls under the steady category. The protein is fragile if the protein instability score is higher than 40. The aliphatic index as determined is 89.53. Alanine (14.1%) contained the highest concentration of amino acid residue. The sequence has Total amount of molecules that are negatively charged (Asp+Glu): 30 Total amount of molecules that are positively charged (Arg+Lys): 27 4.3 Subcellular Localization Analysis The hypothetical protein's subcellular location reveals details about its biological function. It should be beneficial in the process of creating a medication to ward off the target protein. With and estimated accuracy of around 71.1% the PSLpred server suggested a cytoplasmic protein—a protein found within the cytoplasm of a cell—as the target protein's subcellular location. Figure 4.3.1 PSLpred subcellular localization result The SOSUIGramN program states that the hypothetical is localized within the inner membrane of the cell. Figure 4.3.2 SOSUI-GramN subcellular localization result 4.4 Homology Searching The search for the homology of Hypothetical protein using the well-known BLASTp tools from NCBI. The top 10 results from a homology search are shown in the table below. Figure 4.4.1 Top 10 homology results from BLASTp tools 4.5 Multiple sequence alignment and phylogenetic Tree prediction In the figure below the top 10 multiple sequence alignment and Phylogenetic analysis shown. Here out target protein OHO18223.1 is more similar to WP_003415010.1 Figure 4.5.1: Top 10 multiple sequence alignment obtained by using Clustal Omega tools Figure 4.5.2 Phylogenetic Tree Prediction 4.6 Protein family and

Function prediction The NCBI-CD Search predictions suggest the conserved domains and likely function of our target protein. The domain protein from the DsbG super family that is involved in protein-disulfide isomerase [posttranslational modification, protein turnover and chaperone] was discovered in the CD search with an Evalue of 4.11e-30 and an interval of 56-239. Figure 4.6.1 Conserved domain result The experiment is also done by Interpro server which has also shown us the same result. Figure 4.6.2 InterPro server result Function of Protein-Disulfide Isomerase (PDI) in Prokaryotes: ❖ Protein folding and regular oxidative folding The oxidoreductase and isomerase activities of protein disulfide-isomerase are influenced by the type of substrate that the enzyme binds to as well as changes in its redox state. These kinds of mechanisms enable protein oxidative folding. Disulfide bridges are formed when reduced cysteine residues in developing proteins are oxidized, stabilizing the protein and allowing it to form native structures, notably tertiary and quaternary structures. Protein folding is the physical transformation of a protein chain into its native three-dimensional structure, often known as a "folded" conformation, which enables the protein to function physiologically [14]. Protein folding in bacteria is closely controlled genetically, transcriptionally, and at the level of the protein structure because only folded proteins are usually effective. Moreover, essential cellular machinery supports polypeptide folding to guard against misfolding and ensure the creation of useful structures [15]. Proteins are selectively folded by PDI in the ER. At the active site (CGHC motif) of protein disulfide-isomerase, a cysteine residue in a stretched protein joins with another cysteine residue to create a mixed disulfide. The two cysteine residues in the protein disulfide-active isomerase's site are subsequently reduced as a result of a second cysteine residue in the substrate creating a stable disulfide bridge. ❖ Redox signaling A redox chaperone called PDI is involved in the control of vascular NADPH oxidase, phagosome formation, and host cell absorption of bacteria and viruses. In addition to being related to phagocyte NADPH oxidase, PDI is also essential for L. chagasi to successfully infect macrophages . Many pathogens, including Mycobacterium tuberculosis, Salmonella typhimurium, Toxoplasma gondii, and particularly Leishmania spp. and Trypanosoma cruzi, can pass through or live in the phagosome [16]. ❖ Immune system PDI is a soluble homodimer without a transmembrane domain, like other ER chaperones. Moreover, it contains the KDEL C-terminal sequence, which binds to certain receptors in COP vesicles that migrate through the ER-Golgi region and recycles proteins back to the ER [17]. Antigenic peptides are helped to bind onto MHC class I molecules by protein disulfide-isomerase. These molecules (MHC I) are connected to the display of peptides during the immunological reaction by antigen-presenting cells. ❖ Chaperone activity Another crucial role that protein disulfide-isomerase plays is chaperoning. One of the main functions of molecular chaperones is to prevent the buildup of proteins that are misfolded. Chaperones are a class of proteins that aid in the normal and stressed folding of proteins in the cell. They are both capable of recognizing and interacting with foreign proteins, which prevents unwanted protein aggregation [18]. Molecular chaperones play a variety of functions in bacterial cells, including assisting in protein release, repairing proteins that have been harmed or misfolded by stresses like a heat shock, as well as preventing protein aggregation in reaction to a heat shock. Moreover, they support protein folding during and after translation of newly produced proteins. 4.7 Secondary Structure determination Using SOPMA, the protein's primary and secondary structural characteristics were identified. According to these calculations, the protein contains 35.29% alpha helix, 20.00% extended stands, 7.06% beta turns, and 37.65% random coils. The PRISPRED website, which presents a summary of comparable results, also double-checked the protein structural results. The putative protein's typical secondary structure (OHO18223.1) is displayed below. Figure 4.7.1. The positions of the alpha helix, beta turn, extended strand, and random coil are shown individually in this picture. Figure 4.7.2 Secondary structure of hypothetical protein 4.8 Three-dimensional structure prediction Using the template with the highest rating, the SWISS-MODEL server completed a 3D prediction with a 100% identification rate. Figure 4.8.1. 3D structure of Hypothetical protein 4.9 Tertiary Structure Validation and model Quality Assessment The quality of the 3D models was evaluated using Prosa -web and the Saves v6.0 server. In the Prosa -web server, z-score was more common in the NMR than the x-ray regions. If the z-score is within the range of values frequently reported for genuine proteins of a similar size, the model is of excellent quality; otherwise, the structure is incorrect. The Prosa -web plot graph shows that the z-scores for PDB protein structures vary from -1 to -13. A value more that 90% inside the permitted zone is shown on the rating from the statistics of the Ramachandran plot, with 94.8% of residues in the intended locations. Figure 4.9.1 Ramachandran plot statistics The average 3D-1D score of 82.41% of the remnants was superior to or equivalent to 0.1, demonstrating the moral uprightness and suitability of these structures. A verify3D score of higher than 80% denotes a high caliber model [19]. Figure 4.9.2 Verification of a 3d model via Verify3D 82.41% of the residues have averaged 3D-ID score >=0.1 The model's projected Z -score of -6.07. The range of high-quality model's z -scores for protein structures that have been experimentally confirmed. Figure 4.9.3 The link between HP and other proteins is shown using the z-score, with HP showing as black dots and other proteins as grey. Score for z: -6.47 (Prossa server) Based on frequent atomic connections, Errat looked into the statistical distribution of non-bonded lings among various atom types. The overall quality factor for this model was 99.4737%, which is regarded as exceptional. Use of high-resolution models with Errat values of 95% or higher is advocated [20]. Figure 4.9.4: The overall quality factor plot (Errat) of the original model is 99.474% accurate. Red denoted difficult, yellow denoted less problematic, and gray denoted non-problematic. The model was of good quality, according to the results of the overall quality assessment tests conducted by Errat, Prosa-web, verify3D, and Procheck (Ramachandran Plot) Discussion Sequence taken from the NCBI database that is fictitious. The protein's physicochemical properties were then examined using ExPASy's ProtParam tool. It was determined that the protein had 255 amino acids, a molecular weight of 26861.54, a theoretical pI of 5.68, and a grand average of hydropathicity (GRAVY): 0.097 value is positive, indicating that the protein is water soluble. The ability of PSIPRED to get an average Q3 score of 76.5% using a rigorous cross validation approach has been demonstrated [21]. The PSLpred server predicted a cytoplasmic protein (a protein located within the cytoplasm of a cell) as the subcellular location of the target protein, with an estimated accuracy of about 71.1%. The domain protein from the DsbG super family that is involved in protein-disulfide isomerase [posttranslational modification, protein turnover, chaperones] was discovered in the CD search with an E-value of 4.11e-30 and an interval of 56- 239. Using SOPMA, the protein's primary and secondary structural characteristics were identified. According to these calculations, the protein contains 35.29% alpha helix, 20.00% extended strands, 7.06% beta turns, and 37.65% random coils. Using the template with the highest rating, the SWISS-MODEL server completed a 3D prediction with a 100% identification rate. The quality of the 3D models was evaluated using Prosa -web and the Saves v6.0 server. All three analyses— Errat, Verify3d, and Ramachandran plot—are available on the Saves v6.0 server. On the Prosa-web server, our z-score was more common in the NMR than the x-ray regions. If the z-score is within the range of values normally seen for genuine proteins of a similar size, the model is of outstanding quality; otherwise, the structure is incorrect. The z-scores for PDB protein structures range from -1 to -13, according to the Prosa-web plot graph. The excellent level of quality of Prosa is demonstrated by the model's projected z-score of -6.07. the range of high-quality models' z-scores for protein structures that have been experimentally confirmed. A value of more than 90% inside the permitted zone is shown on the Ramachandran map, suggesting a high-quality model [19]. The model received an excellent rating from the Ramachandran Plot Statistics, with 94.8% of residues in the intended locations. Tool check3D, The average 3D-1D score of 82.41% of the residues was greater than or equal to 0.1, demonstrating the moral uprightness and suitability of these structures. A Verify3D score of higher than 80% denotes a high caliber model [19]. Based on common atomic affiliations, Errat explored the 50 statistical distribution of non-bonded links across a wide range of atom categories. The overall quality factor for Errat was 99.4737%, indicating that the model was excellent. A high-resolution model with an Errat value of 95% or higher is recommended [22]. The model was of good quality, according to the results of the overall quality assessment tests conducted by Errat, Prosa-web, Verify3D, and Procheck (Ramachandran Plot). Here, WP 003415010 and our target protein OHO18223.1 are similar. 1 more closely;

as a result, their functions can complement one another. The enzyme protein disulfide isomerase produces native disulfide pairs in secretory proteins (PDI). The essential enzyme PDI has the ability to disulfide isomerase. Protein Disulfide Isomerase is a redox protein from the thioredoxin class (PDI). Disulfide isomerase, thiol disulfide ox reductase, and redox dependent chaperone are the three enzymatic activities of PDI. The endoplasmic reticulum interior was where PDI was first discovered; it was then discovered in the cytoplasm and on cell membranes. This study will give a summary of current research linking the structural characteristics of PDI to its range of catalytic roles, as well as its physiological and clinical functions related to redox regulation and folded proteins [23]. The significance of this entire study lies in its ability to explain the structure and function of our potential target protein from mycobacterium tuberculosis. Conclusion Based on the findings of the overall evaluation of model quality and resemblance, the best projected models from SWISSMODEL were chosen. Future protein studies that require knowledge of protein structure dynamics can use these generated models as a reference. This study's objectives were to determine the 3D structure of a hypothetical protein from the mycobacterium TB (Accession No. OHO18223.1), as well as to make recommendations for its possible applications. The most promising projected models from SWISSMODEL were selected based on the results of the overall model quality and similarity evaluation. Further protein research that requires knowledge of protein structure can use these developed models as a guide. Researchers identified a novel, cytoplasmic, stable protein with the multifunctional Protein Disulfide Isomerase (PDI) domain required for Mycobacterium tuberculosis. The functioning cell will die if we can prevent this protein from functioning in the cell. As a result, we can stop important processes, which may lead to the development of an anti-tubercular drug in the future. Further research into the structure and function of this protein may aid in the development of inventive medicines.

References 1. Enault F, Suhre K, Claverie JM. Phydbac" Gene Function Predictor": a gene annotation tool based on genomic context analysis. BMC bioinformatics. 2005 Dec;6(1):1-0. 2. Lubec G, Afjehi-Sadat L, Yang JW, John JP. Searching for hypothetical proteins: theory and practice based upon original data and literature. Progress in neurobiology. 2005 Sep 1;77(1-2):90-127. 3. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome- wide expression patterns. Proceedings of the National Academy of Sciences. 1998 Dec 8;95(25):14863-8. 4. Banu S, Honoré N, Saint-Joanis B, Philpott D, Prévost MC, Cole ST. Are the PE-PGRS proteins of Mycobacterium tuberculosis variable surface antigens?. Molecular microbiology. 2002 Apr;44(1):9-19. 5. Mortimer TD, Weber AM, Pepperell CS. Signatures of selection at drug resistance loci in Mycobacterium tuberculosis. MSystems. 2018 Feb 27;3(1):e00108-17. 6. Chakaya J, Khan M, Ntoumi F, Aklillu E, Fatima R, Mwaba P, Kapata N, Mfinanga S, Hasnain SE, Katoto PD, Bulabula AN. Global Tuberculosis Report 2020–Reflections on the Global TB burden, treatment and prevention efforts. International journal of infectious diseases. 2021 Dec 1;113:S7-12. 7. Huang W. Computational methods for identifying and characterizing the human gene regulatory regions and cis-elements. North Carolina State University; 2005. 8. Gasteiger E, Hoogland C, Gattiker A, Duvaud SE, Wilkins MR, Appel RD, Bairoch A. Protein identification and analysis tools on the ExPASy server. Humana press; 2005. 9. Imai K, Hayat S, Sakiyama N, Fujita N, Tomii K, Elofsson A, Horton P. Localization prediction and structure-based in silico analysis of bacterial proteins: with emphasis on outer membrane proteins. Data Mining for Systems Biology: Methods and Protocols. 2013:115-40. 10. Geourjon C, Deleage G. SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. Bioinformatics. 1995 Dec 1;11(6):681-4. 11. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics. 2000 Apr 1;16(4):404-5. 55 12. Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Cassarino TG, Bertoni M, Bordoli L, Schwede T. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic acids research. 2014 Jul 1;42(W1):W252-8. 13. Dym O, Eisenberg D, Yeates TO. Detection of errors in protein models. 14. Alberts B, Bray D, Hopkin K, Johnson AD, Lewis J, Raff M, Roberts K, Walter P. Essential cell biology. Garland Science; 2015. 15. Navarro S, Villar-Piqué A, Ventura S. Selection against toxic aggregation-prone protein sequences in bacteria. Biochimica et Biophysica Acta (BBA)-Molecular Cell Research. 2014 May 1;1843(5):866-74. 16. Bertholet S, Goldszmid R, Morrot A, Debrabant A, Afrin F, Collazo-Custodio C, Houde M, Desjardins M, Sher A, Sacks D. Leishmania antigens are presented to CD8+ T cells by a transporter associated with antigen processing-independent pathway in vitro and in vivo. The Journal of Immunology. 2006 Sep 15;177(6):3525-33. 17. Stolf BS, Smyrnias I, Lopes LR, Vendramin A, Goto H, Laurindo FR, Shah AM, Santos CX. Protein disulfide isomerase and host-pathogen interaction. TheScientificWorldJOURNAL. 2011 Oct 18;11:1749-61. 18. Beissinger M, Buchner J. How chaperones fold proteins. Biological chemistry. 1998 Mar 1;379(3):245-59. 19. Haron FN, Azazi A, Chua KH, Lim YA, Lee PC, Chew CH. RESEARCH ARTICLE In silico structural modeling and quality assessment of Plasmodium knowlesi apical membrane antigen 1 using comparative protein models. Tropical Biomedicine. 2022;39(3):394-401. 20. Singh R, Gurao A, Rajesh C, Mishra SK, Rani S, Behl A, Kumar V, Kataria RS. Comparative modeling and mutual docking of structurally uncharacterized heat shock protein 70 and heat shock factor-1 proteins in water buffalo. Veterinary world. 2019 Dec;12(12):2036. 21. Tran NT, Jakovlić I, Wang WM. In silico characterisation, homology modelling and structure-based functional annotation of blunt snout bream (Megalobrama amblycephala) Hsp70 and Hsc70 proteins. Journal of Animal Science and Technology. 2015 Dec;57:1-9. 22. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics. 2000 Apr 1;16(4):404-5. 23. Ali Khan H, Mutus B. Protein disulfide isomerase a multifunctional protein with multiple physiological roles. Frontiers in chemistry. 2014 Aug 26;2:70.

In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical

Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis In Silico Structural and Functional Annotation of Hypothetical Protein from Mycobacterium Tuberculosis 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 31 32 33 34 35 36 37 38

39 40 41 42 43 44 45 46 47 48 49 51 52 53 54 56 57