Structural and Functional Annotation of Uncharacterized Protein NCGM946K2_146 of *Mycobacterium tuberculosis*: An *In-Silico* Approach

Abu Saim Mohammad Saikat [1,*]

[1] Department of Biochemistry and Molecular Biology, Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Gopalganj, Bangladesh

[*]Correspondence: asmsaikat.bmb@gmail.com

ABSTRACT: The human pathogen *Mycobacterium tuberculosis* (MTB) is indeed one of the renowned important longtime infectious diseases that cause tuberculosis (TB). Interestingly, MTB infection has become one of the world's leading causes of human death. In trehalose synthase, the protein NCGM 946K2 146 found in MTB has an important role. For carbohydrate transport and metabolism, trehalose synthase is required. The protein is not clarified yet, however. In this research, an *in silico* approach was therefore formulated for functional and structural documentation of the uncharacterized protein NCGM946K2 146. Three different servers, including the Modeller, the Phyre2, and the Swiss Model, were used to evaluate the predicted tertiary structure. The top materials are selected using structural evaluations conducted with the analysis of Ramachandran Plot, Swiss-Model Interactive Workplace, Prosa-web, Verify 3D, and Z scores. This analysis aimed to uncover the value of the NCGM946K2 146 protein of MTB. This research will, therefore, improve our pathogenesis awareness and give us a chance to target the protein compound.

KEYWORDS: *Mycobacterium tuberculosis*; protein NCGM946K2_146; Homology Modeling; Ligand binding site; Tuberculosis

INTRODUCTION

Mycobacterium tuberculosis (MTB) is an antiquity bacterial species which is a rod-like, acid-fast, and Gram-positive organism responsible for one of the most lethal diseases (ranking above HIV/AIDS) – tuberculosis (TB). Typically, TB spreads from an individual infected with MTB via the air, such as by coughing. Pulmonary

TB is the infection of the lungs, and extrapulmonary TB is the infection of other sites of the body. It was reported in 2018 that almost 10 million tuberculosis patients (range 9.0–11.1 million) died from HIV-negative deaths in 2018, creating a high risk for tuberculosis spreading and global growth [1]. MBT has the second-largest bacterial genome sequence on tap with 3924 open reading frames. Indeed, multi-gene families and duplicate housekeeping genes have multiple, repetitive DNAs, particularly insertion sequences present in MTB [2]. The CD-search tool [3] predicted a domain of the protein NCGM946K2_146 and was described as a functional protein. Moreover, the uncharacterized protein NCGM946K2 146 from MTB is structurally, and functionally is not reported. The analysis, therefore, explains the comprehensive physicochemical characterization and the predicted functionally annotated tertiary structure.

MATERIALS AND METHODS

Sequence retrieval

The amino acid sequence of NCGM946K2_146 was retrieved in FASTA format from the National Center for Biotechnology Information (NCBI) [4] with the accession ID of BAW10952. However, up to this point, NCGM946K2_146 is not accessible in the Protein Data Bank (PDB) as a tertiary structure of the unknown protein. The structural patterns of this protein subsequently began using the protein NCGM946K2 146 with a 455 amino acid long chain.

Physicochemical Characterization

The ExPASy server ProtParam method has been used to measure the amino acid sequence composition, the instability index, the aliphatic index, the GRAVY, and extinction coefficients as well [5]. Moreover, the SMS Suite (v2.0) for the measurement of the theoretical isoelectric point (pI) of the NCGM946K2 146 protein was performed [6].

Functional Annotation Prediction

The CD Search tool of NCBI [3] is used for domain prediction. The CD Search tool (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) predicted a domain of the protein NCGM946K2_146.

Secondary Structure Prediction

The secondary structure NCGM946K2_146 has been predicted by the SPIPRED software (server-based), and the SOPMA framework was used for elements prediction [7] [8].

Tertiary Structure Modeling and Validation

Currently, there is no experimentally concluded tertiary structure available for NCGM946K2_146 of MTB in the Protein Data Bank (PDB). Consequently, the tertiary structures of the protein modeled by utilizing three different programs, including Modeller [9] with the HHpred tool [10], Phyre2 [11], and Swiss-Model server [12]. Predicted tertiary structures obtained from the three different servers, including the Modeller, the Phyre2, and the Swiss-Model servers, were executed for structural quality assessment experiments. The Ramachandran map evaluation by PROCHECK [13], Swiss-Model Interactive Workplace (https://swissmodel.expasy.org/assess), and Verify 3D [14] utilized for modeled protein structure quality documentation. Additionally, Z-scores derived from the Prosa-web (https://prosa.services.came.sbg.ac.at/prosa.php) demanded consistency validation of the complete tertiary model.

Sub-cellular localization

The subcellular localization predictor tool CELLO v. 2.5 (http://cello.life.nctu.edu.tw/) was executed for sub-cellular localization of the protein NCGM946K2_146 present in MTB. This tool is performed for the amino acid comp., N-peptide comp., physicochemical comp., neighboring seq. Comp. and partitioned seq. Comp. of the protein.

RESULTS

Physicochemical Characterization

The protein NCGM946K2_146 contains a 455 amino acid long sequence with a molecular weight of 49889.11 Da. Extinction coefficients (all pairs of Cys residues form cystines) is 69455, and extinction coefficients (all Cys residues are reduced) is 69330. The protein is acidic (pI 5.06, 4.82*) for containing the total number of negatively charged residues (Asp + Glu) is 61, and the total number of positively charged residues (Arg + Lys) is 46. The aliphatic index, instability index, and GRAVY are 88.62, 38.05, -0.185, respectively (Table 1).

Functional Annotation Prediction

A domain of a protein is a conserved part of a given protein sequence that has a function, and it exists independently of the rest of the protein chain [15]. The CD Search tool predicted a domain (accession ID of COG3281). This predicted domain is associated with trehalose synthase, an enzyme related to carbohydrate transport and metabolism.

Secondary Structure Prediction

SOPMA considered the default parameters for the secondary structure modeling. SOPMA predicted 33.85 percent of residues as random coils in comparison to alpha-helix (51.21 percent), extended strand (11.21 percent), and Beta turn (3.74 percent) by utilizing 455 aa long protein sequence and 51 aligned proteins as well (Table 2).

The PSIPRED program is showing the higher confidence of the prediction of the helix, strand, and coil (Figure 1). The amino acid composition obtained from the ExPASy ProtParam Tool showed in Table 3.

Binding Sites (Protein-Protein, and Protein-Polynucleotide)

The predict protein server executed for binding sites prediction of the protein NCGM946K2_146. It showed there were 17 different active protein binding sites including the position of viz: 1-2; 5; 18-19; 42-44; 87-88; 116; 164; 197-199; 237-238; 271-273; 280; 296; 355; 381-383, 388; 390; and 395 (Figure 2).

Sub-cellular localization

The Sub-cellular Localization Predator tool CELLO v. 2.5 [16], applied for subcellular localization of the protein NCGM946K2_146 of MTB. This tool predicted the localization of amino acid comp., N-peptide comp., physicochemical comp., neighboring seq. Comp. are cytoplasmic and for partitioned seq. comp. values of 0.931, 0.825, 0.817, 0.820 and 0.577, respectively. CELLO prediction values, including the reliabilities for cytoplasmic, membrane, extracellular, and cell wall, are of 3.791, 1.016, 0.177, and 0.0017, respectively (Table 4).

Modeling and Validation of Tertiary Structures

Three different tools, for example, the Modeller, the Phyre2, and the Swiss Model servers utilized for tertiary structures prediction of the protein NCGM946K2_146. The tertiary structure modeling is executed through Modeller [10] by selecting the most suitable template (chosen from 250 suitable models) for protein modeling. The target template (4O7O_B) was selected based on the probability rate (100%), the E-Value of 2.2e-59, SS of 58.1, Cols of 455, and the target length of 455 (data not showed). The modeled three-dimensional structure of the protein is stored in PDB format (Figure 3). Similarly, the Phyre2 server applied for tertiary structure prediction. The template (c4o7oB) was selected based on the confidence value of 100.0% and coverage of 99%. Besides, the Swiss Model server predicted tertiary structure of the protein based on the most favored template (4o7p.1.B). This template bears the values of GMQE, QSQE, Identity score of 0.99, 0.74, and 99.78, respectively.

The Ramachandran Map Analysis by PROCHECK, the Verify 3D tool, and the Swiss-Model Interactive Workplace tools applied for structural assessment of the modeled tertiary structures obtained from the Modeller, the Phyre2, and the Swiss Model servers. In case of predicted tertiary structure by Modeller, the assessment experiment executed by the Ramachandran Map (PROCHECK) indicating 95.2% of the total residues (376) found in the core [A, B, L]; 4.8% of residues were found in the additional allowed regions [a,b,l,p]; and there was no residue found in the generously allowed parts [~a,~b,~l,~p] and the graciously allowed regions [~a,~b,~l,~p] (Table 5). The number of non-glycine, as well as the number of non-proline residues, was 395, which is 100%; the end-residues (excl. Gly and Pro) were 2; the glycine residues and proline residues were 33 and 25, respectively among the total residues of 455 (Figure 6). The Verify 3D documented the predicted structure by Modeller as an excellent three-dimensional structure (Figure 4). The Swiss-Model Interactive Workplace calculated the

MolProbity Score of 2.42, Ramachandran favored of 97.13%, and the QMEAN (Qualitative Model Energy Analysis), Cβ, the value of All Atom, measurement of solvation, and the torsion values of -1.05, -1.55, -0.71, 0.20 and -0.85, respectively (data not showed). On the other hand, the Phyre2 server modeled of the 451 residues (99% of the target sequence) with 100.0% confidence by the single highest scoring template. The Ramachandran Map analysis report of the designed tertiary (3D) structure by the Phyre2 program explained 93.6% of the residues were found in the most favored regions [A, B, L]; 6.1% were in the additional allowed parts [a,b,l,p]; 0.3% were found in the disallowed regions, and there was no aa residue found in the generously allowed areas (Table 5). The Verify 3D validated this predicted tertiary structure (Figure 4). The Swiss-Model Interactive Workplace calculated the parameters of this predicted three-dimensional structure by Phyre2 server as the MolProbity Score of 2.37, Ramachandran favored region of 96.66%, and the QMEAN, the Cβ, the value of the all-atom, solvation value, and the torsion values of -1.23, -1.20, 0.18, 0.37, and -1.2, respectively (data not showed). In contrast, the Ramachandran Map analysis report of the designed tertiary structure (3D) by the Swiss Model server showed 94.2% of the residues were found in most favored regions [A, B, L]; 5.3% were present in the additional allowed areas [a,b,l,p]; 0.1% were found in the disallowed regions, and 0.4% were available in the generously allowed regions (Table 5). The Verify 3D program showed this modeled structure quality as good (Figure 4). The Swiss-Model Interactive Workplace predicted the MolProbity Score of 1.89, Ramachandran favored of 97.21%, and the QMEAN, the Cβ, the value of all-atom, solvation value, and the torsion values of -1.22, -1.12, 0.25, -0.14, and -0.99, respectively (data not showed).

Furthermore, the Prosa-web server [17] used for standard bond angles detection in the modeled tertiary structures of the protein NCGM946K2_146. Z-score for the shaped tertiary structures from the three servers, including the Modeller, the Phyre2, and the Swiss-Model were -9.13, -9.38, and -8.04, respectively (data not showed).

DISCUSSION

The amino acid (aa) sequence of the uncharacterized protein NCGM946K2_146 of MTB was retrieved in FASTA format and used as a query sequence for the determination of physicochemical parameters. The theoretical isoelectric point (5.06, 4.82*) indicates the acidic nature of the protein (Table 1). The instability index of the protein NCGM946K2_146 is 38.05 (<40) reported its stability [18]. The secondary structure elements (Table 2)

and the amino acid composition (Table 3) reveals the characteristics of the protein NCGM946K2_146. The protein-protein and the protein-polynucleotide binding site characteristics (Figure 2) are essential for designing small molecules that modulate protein functions, and also for drug and vaccine targeting opportunities [19] [20]. The subcellular location analysis report of a protein provides an inside into the role of the protein [21]. The CELLO v. 2.5 predicted the subcellular location of the uncharacterized protein NCGM946K2_146 of MTB as cytoplasmic (Table 4). The modeled tertiary structures obtained from the three different servers – the Modeller, the Phyre2, and the Swiss Model servers, are compared for structure quality assessment through Ramachandran Map Analysis, Verify 3D analysis, and Interactive Workplace analysis of the Swiss-Model server (**Table 5**). Amino acid residues found in the most favored areas by the Modeller, the Phyre2, and the Swiss Model programs were 95.2%, 93.6%, and 94.2%, respectively. In the case of Modeller, there was no residue in the disallowed regions. Nevertheless, there 0.3% and 0.04% of residues were present in the forbidden areas by the Phyre2, and the Swiss Model, accordingly. The Swiss-Model Interactive Workplace predicted the Ramachandran favored are of the Modeller, the Phyre2, and by the Swiss Model were found as of 97.13%, 96.66%, 97.21%, respectively. The Verify 3D tool used for structural quality assessment indicated that the three programs, including the Modeller, the Phyre2, and the Swiss Model scored 86.81%, 89.36%, and 88.78%, respectively (Figure 4). Z-scores obtained from the Prosa-web indicating the 'degree of nativeness' of the predicted tertiary structures. In this analysis, all the three, i.e., the Modeller, the Phyre2, and the Swiss-Model servers, are documenting the similar values of Z-scores. Therefore, this comparison showed that the models generated by Modeller were more acceptable when compared to the Phyre2 and the Swiss Model. The CD-search tool [3] predicted a domain of the protein and was described as a functional protein. The predicted domain contains trehalose synthase. Trehalose synthase is an enzyme related to carbohydrate transport and metabolism. Trehalose is present in the cytoplasm of MTB as a free disaccharide and a mixture of cell-wall glycolipids. [22].

CONCLUSIONS

The structural, as well as the functional annotation of NCGM946K2 146, which is located in MTB, was documented in this study with the predicted ligand-binding active sites present in *M. tuberculosis.* The arrangement of amino acid sequences in the desired region was determined by assessing the protein structure.

Regarding understanding protein operations, the physicochemical parameters also functional enrichment estimation is beneficial. The secondary assumption and evaluation structures verified that alpha-helix, random spiral, extended strand, and beta turns were predominant in most sequences. Three different servers, including the Modeller, the Phyre2, and the Swiss Model servers, have assessed the assumed tertiary structures. PROCHECK for Ramachandran Map Analysis, the Verify 3D, the Swiss-Model Interactive Workplace server, and the Z-scores from Prosa-Web used as protein structure evaluation tools. The results showed that the Modeller is appropriate *in silico* documentation for the modeled protein NCGM946K2_146 from the three separate servers. This study would provide an opportunity to design effective therapeutic drugs against the protein of *M. tuberculosis*.

ACKNOWLEDGEMENT

None.

CONFLICTS OF INTEREST

None declared.

REFERENCES

[1] WHO | Global tuberculosis report 2019 (2019) URL https://www.who.int/tb/publications/global_report/en/ (accessed March 10, 2020).

[2] Yamurai Bishi L, Chaitanya Vedithi S, L. Blundell T, Mugumbate G. Computational Deorphaning of Mycobacterium tuberculosis Targets. Drug Discov. Dev. - New Adv., IntechOpen; 2019. https://doi.org/10.5772/intechopen.82374.

[3] CDD/SPARCLE: the conserved domain database in 2020. Nucleic acids research 48, 48(D1), D265-D268.

[4] NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2014;42(Database issue):D7-D17. doi:10.1093/nar/gkt1146

[5] Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, et al. Protein Identification and Analysis Tools on the ExPASy Server. Proteomics Protoc. Handb., Humana Press; 2005, p. 571–607. https://doi.org/10.1385/1-59259-890-0:571.

[6] Martin L, Garrity DM, Yao T. Genomics and transcriptomics of the molting gland (Y-organ) in the blackback land crab, Gecarcinus lateralis. Colorado State University. Libraries; 2016.

[7] Combet C, Blanchet C, Geourjon C, Deléage G. NPS@: Network protein sequence analysis. Trends Biochem Sci 2000;25:147–50. https://doi.org/10.1016/S0968-0004(99)01540-6.

[8] Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202. https://doi.org/10.1006/jmbi.1999.3091.

[9] Webb B, Sali A. Comparative Protein Structure Modeling Using MODELLER. Curr Protoc Protein Sci 2016;86:2.9.1-2.9.37. https://doi.org/10.1002/cpps.20.

[10] Zimmermann, Lukas, Andrew Stephens, Seung-Zin Nam, David Rau, Jonas Kübler, Marko Lozajic, Felix Gabler, Johannes Söding, Andrei N. Lupas, and Vikram Alva. "A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core." Journal of molecular biology 430, no. 15 (2018): 2237-2243.

[11] Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc 2015;10:845–58. https://doi.org/10.1038/nprot.2015.053.

[12] Arnold, K., Bordoli, L., Kopp, J., & Schwede, T. (2006). The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics, 22(2), 195-201.

[13] Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. J Mol Biol 1963;7:95–9. https://doi.org/10.1016/S0022-2836(63)80023-6.

[14] Lüthy R, Bowie JU, Eisenberg D. Assessment of protein models with three-dimensional profiles. Nature 1992;356:83–5. https://doi.org/10.1038/356083a0.

[15] Goodacre, N. F., Gerloff, D. L., & Uetz, P. (2014). Protein domains of unknown function are essential in bacteria. MBio, 5(1).

[16] Yu C-S, Lin C-J, Hwang J-K. Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n -peptide compositions . Protein Sci 2004;13:1402–6. https://doi.org/10.1110/ps.03479604.

[17] Wiederstein, M., & Sippl, M. J. (2007). ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic acids research, 35(suppl_2), W407-W410.

[18] Guruprasad K, Reddy BVB, Pandit MW. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. Protein Eng Des Sel 1990;4:155–61. https://doi.org/10.1093/protein/4.2.155.

[19] Sotriffer, C., & Klebe, G. (2002). Identification and mapping of small-molecule binding sites in proteins: computational tools for structure-based drug design. Il Farmaco, 57(3), 243-251.

[20] Tripathi A, Kellogg GE. A novel and efficient tool for locating and characterizing protein cavities and binding sites. Proteins Struct Funct Bioinforma 2010;78:825–42. https://doi.org/10.1002/prot.22608.

[21] Scott MS, Calafell SJ, Thomas DY, Hallett MT. Refining protein subcellular localization. PLoS Comput Biol 2005;1:0518–28. https://doi.org/10.1371/journal.pcbi.0010066.

[22] De Smet KAL, Weston A, Brown IN, Young DB, Robertson BD. Three pathways for trehalose biosynthesis in mycobacteria. Microbiology 2000;146:199–208.

https://doi.org/10.1099/00221287-146-1-199.

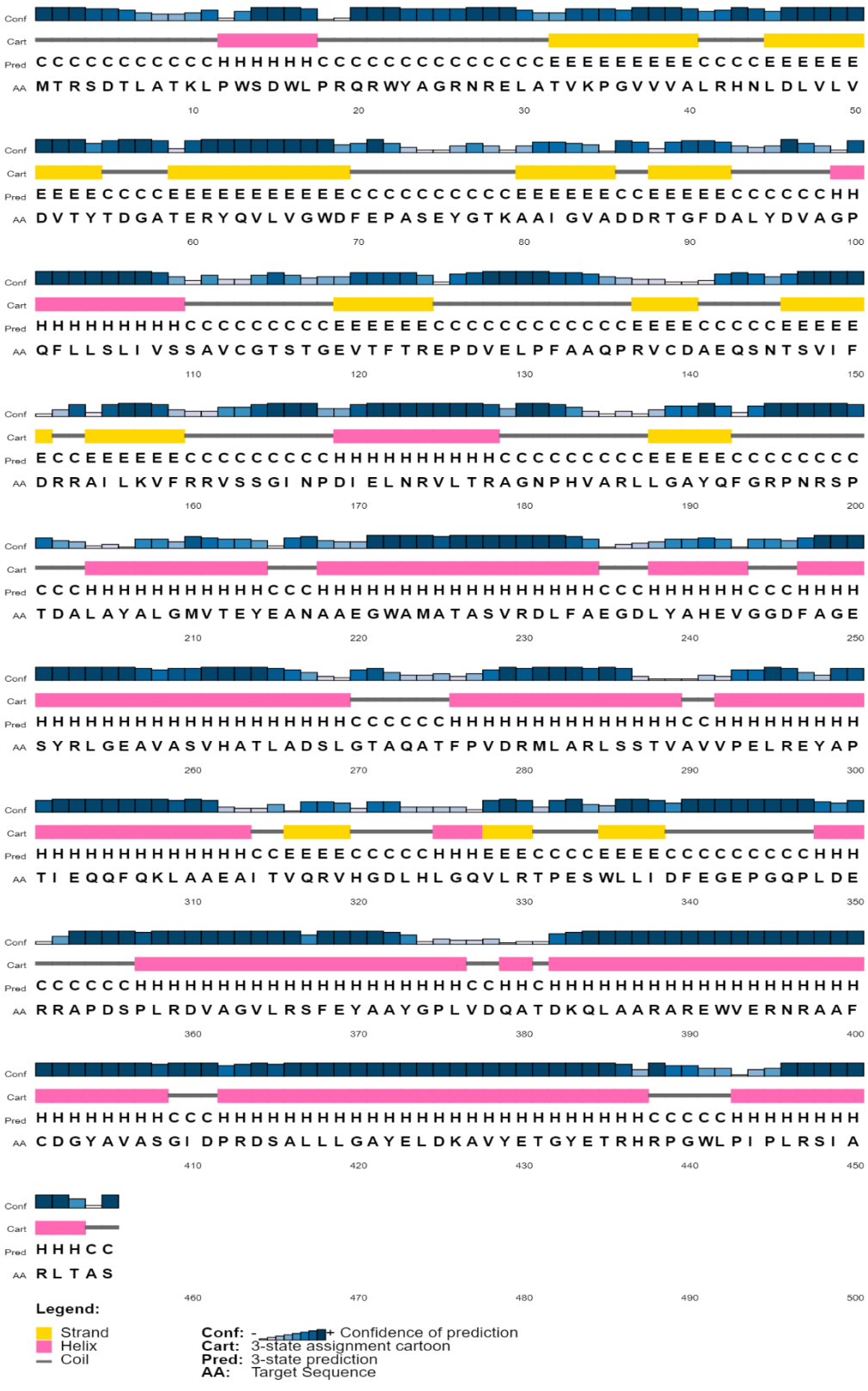

Figure 1 Predicted Secondary Structure

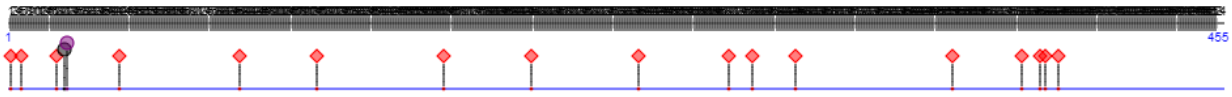

Figure 2 Protein-Protein and Protein-Polynucleotide Binding Sites

The 'red triangles' and the 'colorful circled shapes' indicate the protein-protein, and protein-polynucleotide binding sites, respectively.

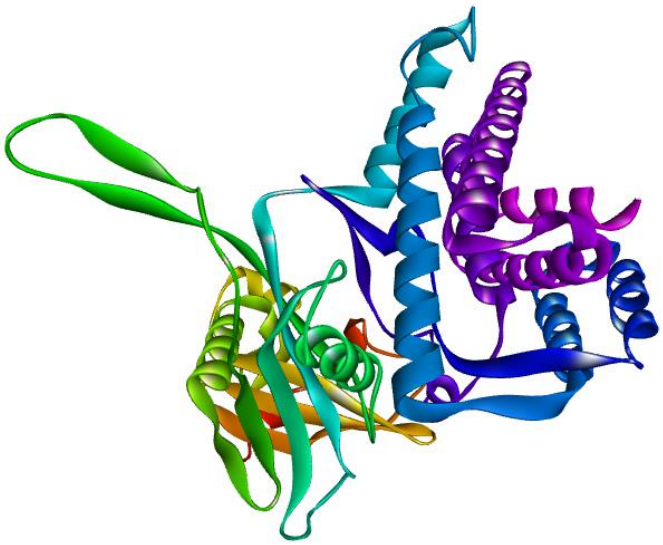

Figure 3 Structure of NCGM946K2_146 Predicted by Modeller

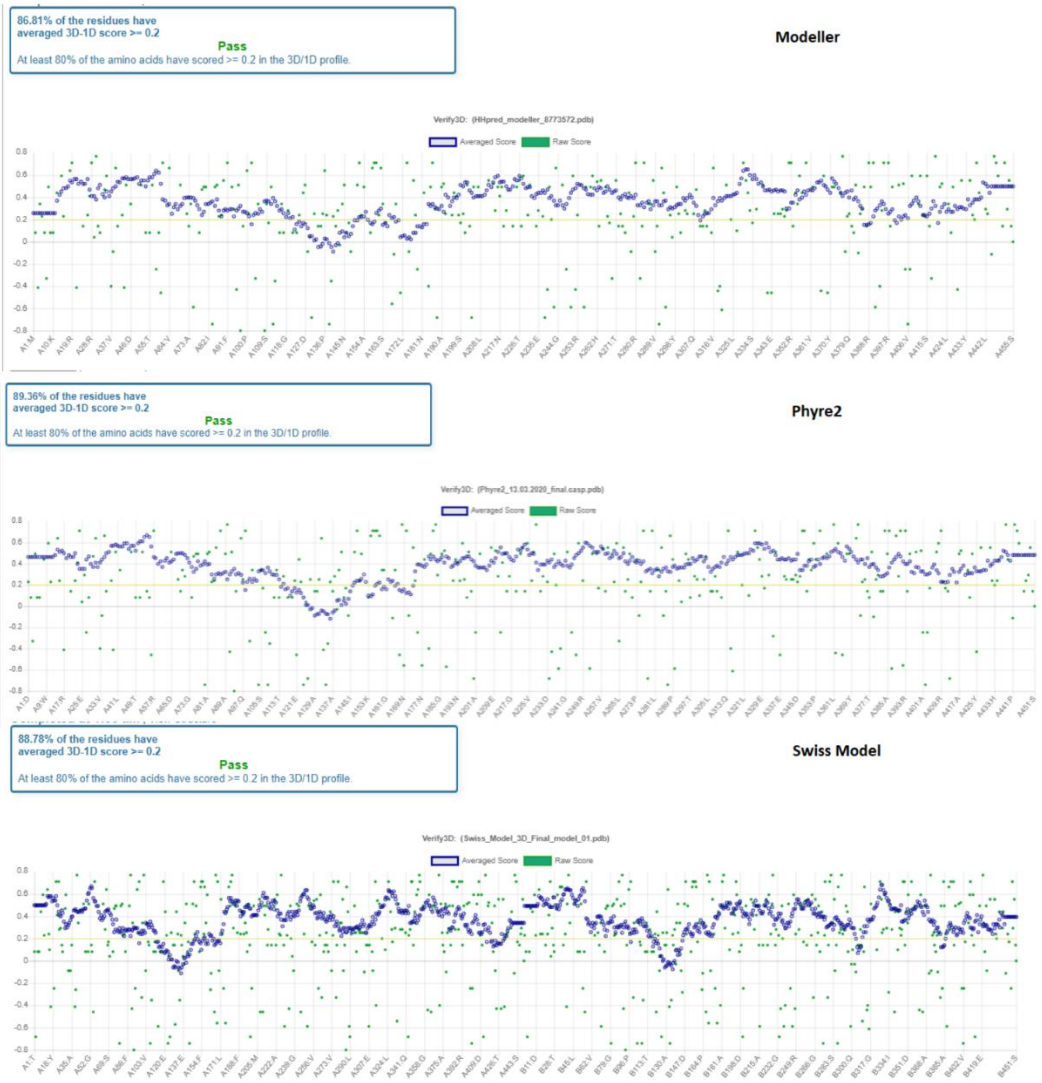

Figure 4 Verify 3D Results

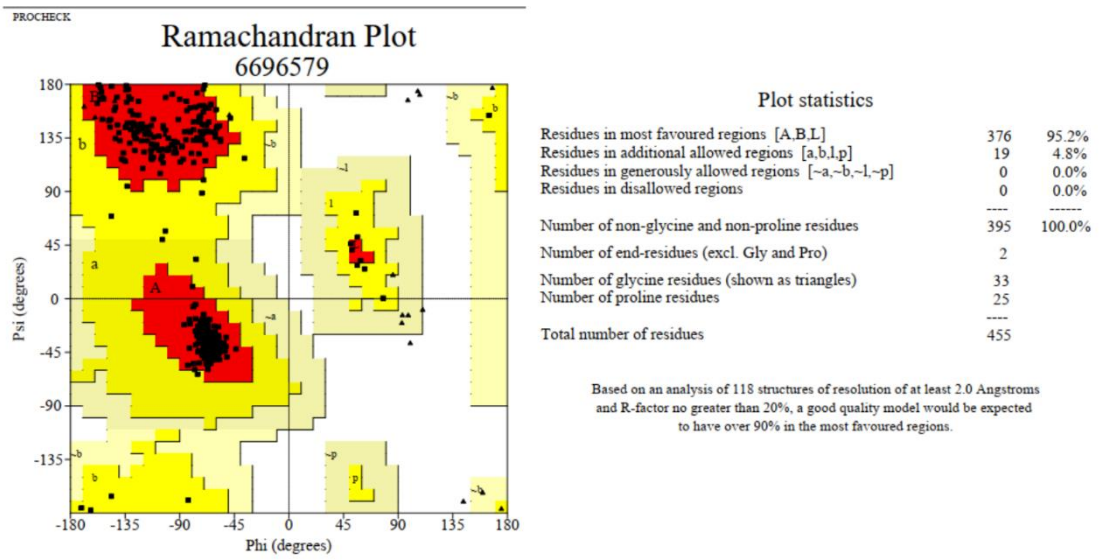

Figure 5 Ramachandran Plot Analysis of Modeller Predicted Protein Structure

Table 1 Physicochemical Parameters

Physio-chemical Parameters	Values
Number of amino acids	455
Molecular weight	49889.11
Theoretical Isoelectric Point (pI)	5.06, 4.82*
Aliphatic index	88.62
Instability index	38.05
Extinction coefficients (All pairs of Cys residues form cystines)	69455
Extinction coefficients (all Cys residues are reduced)	69330
Total number of negatively charged residues (Asp + Glu)	61
Total number of positively charged residues (Arg + Lys)	46
Grand average of hydropathicity (GRAVY)	-0.185

*pI determined by SMS Version 2

Table 2 Secondary Structure Elements

Secondary Structure Elements	Values (%)
Alpha helix (Hh)	51.21
3_{10} helix (Gg)	0.00
Pi helix (Ii)	0.00
Beta bridge (Bb)	0.00
Extended strand (Ee)	11.21
Beta turn (Tt)	3.74
Bend region (Ss	0.00
Random coil (Cc)	33.85
Ambiguous states	0.00
Other states	0.00

Table 3 Amino Acid Composition

S. No.	Amino Acids	No. of Amino Acids	Percentage (%)
1	Ala (A)	60	13.2
2	Arg (R)	39	8.6
3	Asn (N)	9	2.0
4	Asp (D)	31	6.8
5	Cys (C)	3	0.7
6	Gln (Q)	15	3.3
7	Glu (E)	30	6.6
8	Gly (G)	33	7.3
9	His (H)	7	1.5
10	Ile (I)	12	2.6
11	Leu (L)	47	10.3
12	Lys (K)	7	1.5
13	Met (M)	4	0.9
14	Phe (F)	15	3.3
15	Pro (P)	25	5.5
16	Ser (S)	25	5.5
17	Thr (T)	29	6.4
18	Trp (W)	8	1.8
19	Tyr (Y)	17	3.7
20	Val (V)	39	8.6

Table 4 Subcellular Localization Analysis Report

SVM	Localization	Reliability
Amino Acid Comp.	Cytoplasmic	0.931
N-peptide Comp.	Cytoplasmic	0.825
Partitioned seq. Comp.	Membrane	0.577
Physicochemical Comp.	Cytoplasmic	0.817
Neighboring seq. Comp.	Cytoplasmic	0.820
CELLO Prediction	Cytoplasmic	3.791*
	Membrane	1.016
	Extracellular	0.177
	Cell Wall	0.017

*CELLO predicted the subcellular location of the protein as cytoplasmic.

Table 5 Ramachandran Plot Analysis

Servers	Ramachandran Plot Calculation	Value (%)
Modeller	Residues in most favored regions [A,B,L]	95.2
	Residues in additional allowed regions [a,b,l,p]	4.8
	Residues in generously allowed regions [~a,~b,~l,~p]	0.0
	Residues in disallowed regions	0.0
Phyre2	Residues in most favored regions [A,B,L]	93.6
	Residues in additional allowed regions [a,b,l,p]	6.1
	Residues in generously allowed regions [~a,~b,~l,~p]	0.0
	Residues in disallowed regions	0.3
Swiss Model	Residues in most favored regions [A,B,L]	94.2
	Residues in additional allowed regions [a,b,l,p]	5.3
	Residues in generously allowed regions [~a,~b,~l,~p]	0.4
	Residues in disallowed regions	0.1