An *In silico* Approach for Structural and Functional Annotation of Uncharacterized Protein Rv0986 present in *Mycobacterium tuberculosis*

Abu Saim Mohammad Saikat[1]*, Md. Lutful Kabir[1], and Abul Bashar Ripon Khalipha[2]

[1]Dept. of Biochemistry and Molecular Biology, Bangabandhu Sheikh Mujibur Rahman Science and Technology University (BSMRSTU), Gopalganj, Bangladesh; [2]Dept. of Pharmacy, Bangabandhu Sheikh Mujibur Rahman Science and Technology University (BSMRSTU), Gopalganj, Bangladesh.

*Correspondence: asmsaikat.bmb@gmail.com (Abu Saim Mohammad Saikat, Dept. of Biochemistry and Molecular Biology, BSMRSTU, Gopalganj, Bangladesh.

ABSTRACT

Tuberculosis (TB) is an ancient infectious disease caused by *Mycobacterium tuberculosis* (MTB). MTB is a human pathogen. Surprisingly, TB has become the top disease for its death rate worldwide. The uncharacterized protein Rv0986 is closely related to the transporters of the ATP-binding cassette domain, therefore, take part in the export of macrolide as well as a lipoprotein. Furthermore, it is associated with cell division protein. Hence, the protein has a significant role in mycobacterial infection. But, so far, the uncharacterized protein Rv0986 is not elaborated. As a result, in this study, the structural and functional annotation of the protein is described through *in silico* approach. The predicted tertiary structures of the protein generated by Swiss Model, Modeller, and Phyre2, and documented by the Ramachandran Plot analysis with PROCHECK, Verify 3D, and Swiss-Model Interactive Workplace. Z-score also applied for the overall structural assessment. This study will unleash the importance of the uncharacterized protein present in MTB, therefore, it provides an opportunity for drug and vaccine targeting against infection by MTB.

Keywords: *Mycobacterium tuberculosis*, Homology modeling, Protein Rv0986, *In silico*, and Ligand binding site.

1. INTRODUCTION

Mycobacterium tuberculosis is a human pathogen caused the most lethal disease tuberculosis. MTB is a rod-like, acid-fast, Gram-positive organism. Amazingly, TB is one of the most lethal diseases and ranked over HIV/AIDS globally. Tuberculosis spreads from infected people with MTB bacteria into the air, i.e. by coughing. In general, pulmonary TB causes by infection with lungs by MTB, and infection with MTB of other sites of the body causes extra pulmonary TB. It is estimated that about 10 million people are in new cases (range: 9.0-11.1 million), and about 1.2 million deaths by MTB infection (range, 1.1-1.3 million) even though people with HIV-negative in 2018. Therefore,

people are at high risk for infection with MTB worldwide (WHO;*Global tuberculosis report 2019*; Ellison, 2019; Duarte *et al.*, 2017). The protein Rv0986 is an ABC transporter ATP-binding protein of *M. tuberculosis* (strain ATCC 25618/H37Rv. Rv0986 has the ATPase catalytic subunit of an ABC transporter complex. This complex protein is deeply associated with coupling the energy of ATP hydrolysis which the main function is to the import of one or more from a variety of substrates like macrolide, lipoproteins, and hemin. As a result, Rv0986 is related to MTB infection (Wu *et al.*, 2019; Holland and Holland, 2005; Beis, 2015). But, the protein structural and functional annotation with ligand binding active sites not explored.

Therefore, the tertiary structure of the functional protein Rv0986 with detailed physicochemical characterization is predicted in this study by an *in silico* approach.

2. MATERIALS AND METHODS

2.1 Sequence Retrieval

The amino acid sequence of the functional protein Rv0986 collected from the UniProtKB server (Alex, 2019) following the accession ID of P9WQK1. There was no tertiary structure available of the protein Rv0986 in the Protein Data Bank. Hence, the tertiary structure modeling of the protein Rv0986 undertook by using the 248 amino acid long protein sequence.

2.2 Physicochemical Characterization

We used two web-based servers for the determination of the physicochemical properties of the uncharacterized protein Rv0986 present in MTB. ExPasy's ProtParam tool (ProtParam, 2017) used to predict the composition of amino acid, aliphatic and instability index, extinction coefficients, and GRAVY - a grand average of hydropathicity. Also, the isoelectric point (pI) predicted by the Sequence Manipulation Suite (SMS) tool of version 2 (Martin, 2016).

2.3 Secondary Structure Prediction

The self-optimized prediction method with alignment (SOPMA) (Combet *et al*., 2000) and SPIPRED (David, 1999) programs used for the secondary structure prediction of the protein Rv0986. DISOPRED tool (Vikram and Kumar, 2018) utilized for disorder prediction.

2.4 Tertiary Structure Modeling and Validation

Three web-based servers including Modeller (Zimmermann *et al*., 2018) following the HHpred (Zimmermann *et al*., 2018), Swiss Model (Arnold *et al*., 2020), and Phyre2 (Kelley *et al*., 2015) applied for tertiary structure prediction of the protein Rv0986 present in MTB.

In the homology protein modeling process, the most suitable template selected for final tertiary structures prediction. The Ramachandran map analysis following the PROCHECK and the Verify 3D server (https://servicesn.mbi.ucla.edu/Verify3D/) app-lied for

the modeled tertiary structure quality documentation. Also, theSwiss-Model Interactive Work place (https://swissmodel.expasy.org/assess) followed for the structural quality assessment of the predicted protein structures. Also, the Prosa-web (Wiederstein and Sippl, 2007) used to obtain the Z-scores which was followed for the overall tertiary structural quality assessment experiment.

3. RESULTS AND DISCUSSION

3.1 Physicochemical Characterization

The protein Rv0986 is 248 amino acids long and the sequence retrieved in the FASTA format. This protein sequence applied as a query sequence to detect the physicochemical characteristics of the protein Rv0986 present in MTB. As the instability index is 32.44 which is less than 40, the protein is stable (Guruprasad *et al*., 1990; Islam *et al*., 2020). The theoretical isoelectric point (pI) indicates that the protein is acidic (pI 5.63, 5.72*).

The molecular weight of the protein Rv0986 is 27373.11 Da. The higher aliphatic index value of the protein is 95.52 indicating as a positive factor for increased thermos-stability for a wide temperature range (Gill, and Hippel, 1989). The hydrophilic nature of the protein and the possibility of better interaction with water (Ikai, 1980; Shahen *et al*., 2019) were indicated by the GRAVYindices value of -0.265 (**Table 1**).

3.2 Secondary Structure Prediction

SOPMA considered the default parameters including, window width: 17; similarity threshold: 8; division factor: 4, for secondary structure prediction of the protein Rv0986 present in MTB.

SOPMA utilized 248 amino acid-long protein sequences (sub-database) as well as 33 aligned proteins; therefore, it predicted 32.66% of the residues as random coils in comparison to alpha-helix of 47.18%, Beta turn of 5.24%, and the extended strand of 14.92% (**Table 2**). The PSIPRED program showed higher confidence in the prediction of the helix, strand, and coil (**Fig 1**).

Table 1: Physicochemical Parameters.

Physicochemical Parameters	Values
Number of amino acids	248
Molecular weight	27373.11
Theoretical isoelectric point (pI)	5.63, 5.72*
Aliphatic index	95.52
The instability index (II)	32.44
Total number of negatively charged residues (Asp + Glu)	30
Total number of positively charged residues (Arg + Lys)	26
Grand average of hydropathicity (GRAVY)	-0.265

*pI determined by SMS Version2

Table 2: Secondary Structure Elements.

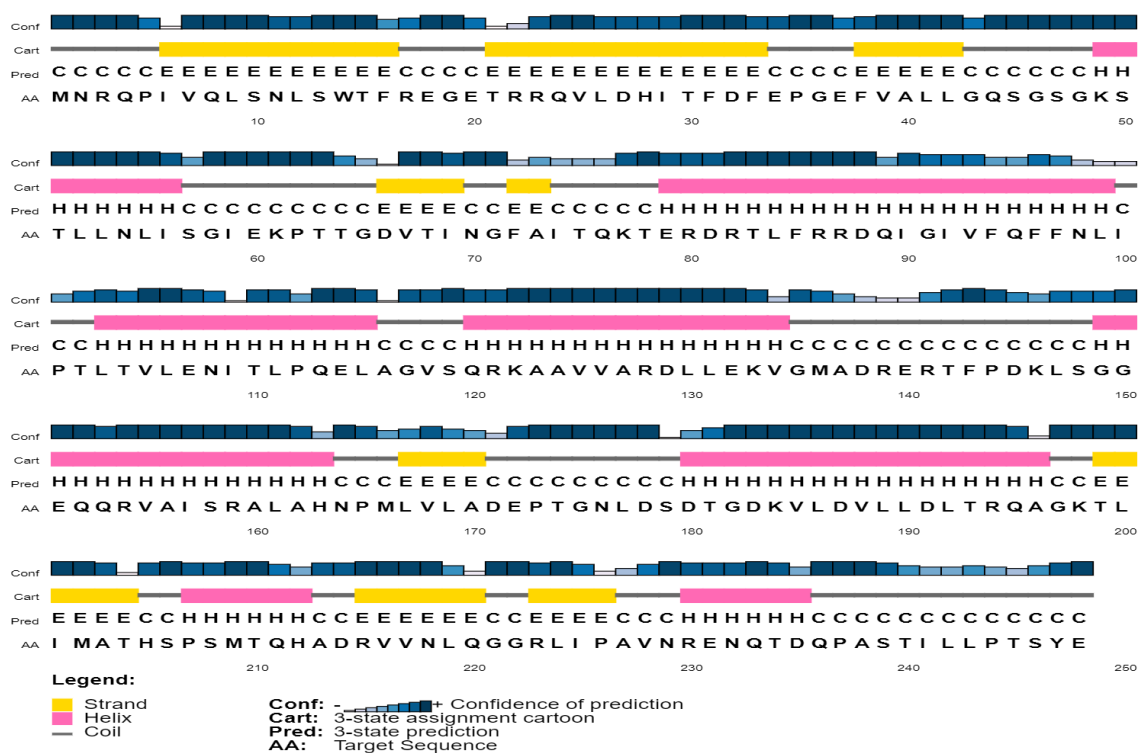| Secondary Structure Elements	Values (%)
Alpha helix (Hh)	47.18
3_{10} helix (Gg)	0.00
Pi helix (Ii)	0.00
Beta bridge (Bb)	0.00
Extended strand (Ee)	14.92
Beta turn (Tt)	5.24
Bend region (Ss)	0.00
Random coil (Cc)	32.66
Ambiguous states	0.00
Other states	0.00

Fig 1: Predicted Secondary Structure.

3.3 Protein-Protein and Protein-Polynucleotide Binding Sites

Predict protein server predicted the binding sites in the protein structure of the protein Rv0986 where 11 different protein binding sites identified at the positions of: 1-4, 17-22; 61-62; 84; 87-88; 120-122; 132-133; 139-140; 193; 212; 229-234; 246. There 4 different polynucleotide binding sites identified at the positions of: 45; 47; 49; and 49-50 (**Fig 2**).
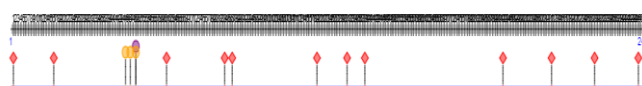

Fig 2: Protein-Protein and Protein-Polynucleotide Binding Sites.

3.4 Structure Modeling and Validation

The HHpred tool was used for executing the Modeller for tertiary structure prediction. Protein sequence in FASTA format inserted into HHpred (Zimmermann *et al.*, 2018) and executed the Modeller (Webb and Sali, 2016). The most suitable template selected (3TUI_D) among the number of hits of 250 for tertiary structure prediction of the protein Rv0986. This template indicated the probability rate of 100%, SS of 27.5, E-Value of 6.6e-32, Cols of 239, and the target length of 366 (data not shown). The generated structure stored in PDB format (**Fig 3**). The Ramachandran Map was analysis by PROCHECK (**Table 3**) indication an excellent protein model quality. It showing that 95.2%of the total residues found in the core; 4.8% of residues in the additional allowed regions. But, there was no residue in both generously allowed regions and the disallowed regions (**Fig 5**). The number of non-glycine and non-proline residues (209) was 100%; 2 residues were as the end-residues (excl. Gly and Pro); the proline and glycine residues were 11 and 18, respectively, among the total residues (**Fig 3**). The Verify 3D tool documented the modeled tertiary protein structure (data not shown).The Swiss-Model Interactive Workplace used for the structural tertiary protein structure assessment experiment indication an excellent protein structure showing the MolProbity Score of 3.02; Ramachandran favored of 97.48% with the Qualitative Model Energy Analysis (QMEAN), Cβ, All Atom, solvation, and torsion values of -1.15, -2.72, -3.06, 0.18, and -0.61, respectively (data not shown).
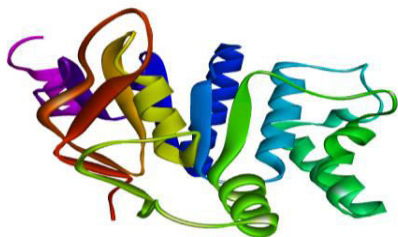
Similarly, the Phyre2 server was used for the prediction of the tertiary protein structure of the protein Rv0986. The tertiary structure predicted based on the top suitable template (c5ws4A) bearing the value of confidence of 100.0% and the coverage of 99% (**Fig 4a**). Phyre2 modeled 245 residues (99% of the protein sequence) with 100.0% confidence. Phyre2 also described the secondary structure parameter including the disordered of 16%,beta-strand of 25%, alpha-helix of 39% (data not shown).The Ramachandran Map followed by the PRO-CHECK indicating 87.9% of the residues in most favored regions and 11.2% as in additional allowed regions. There 0.5% of the total residues were in the disallowed regions, and 0.5% residues in the generously allowed regions (**Table 3**). This modeled the tertiary structure of the protein Rv0986 passed the Verify 3D protein structural quality assessment experiment (data not shown). The Swiss-Model Interactive Workplace validated the modeled protein structure by indicating the values of the MolProbity Score of 2.8, as well as the Ramachandran, favored of 91.80%; the QMEAN, Cβ, all-atom, solvation, and torsion values of -1.92, -1.81, -1.88, 0.54, and -1.81, respectively (data not shown).

Fig 4 (a): Structure of Rv0986 Predicted by Phyre2.

Fig 3: Structure of Rv0986 Predicted by Modeller

Fig 4 (b): Structure of Rv0986 Predicted by Swiss Model.

Table 3: Ramachandran Plot Analysis.

Server	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]	87.9
	Residues in additional allowed regions [a,b,l,p]	11.2
	Residues in generously allowed regions [~a,~b,~l,~p]	0.5
	Residues in disallowed regions	0.5
Swiss Model	Residues in most favored regions [A,B,L]	92.9
	Residues in additional allowed regions [a,b,l,p]	6.4
	Residues in generously allowed regions [~a,~b,~l,~p]	0.5
	Residues in disallowed regions	0.2

Likewise, the tertiary structure of the protein Rv0986 also modeled by the Swiss-Model server. The 3D structure modeled based on the most suitable template selected by comparing with the top five similar level templates (5lj6.1.A, 5lil.1.B, 5lil.1.A, 5lil.1.B, and 2ouk.5.A). The selected template parameters were including the Qualitative Model Energy Analysis (QMEAN) score of (-1.03), Global Model Quality Estimate (GMQE) score of 0.73, percentage of sequence identity of 39.50, and the coverage of 96%. Generated protein structure stored in PDB file format. The Ramachandran Map by PROCHECK showed that 92.9% of the total residues found in the, and 92.9% of residues in the additional allowed regions. There were 0.5% of residues in the generously allowed regions and 0.2% of the residues in the disallowed regions (**Table 3**). In the predicted protein structure, the number of non-glycine and non-proline residues was 423 which was 100% and 4 residues were asthe end-residues (excl. Gly and Pro). Among the total residues in the protein structure, the glycine residues and proline residues were 36 and 22, respectively (**Fig 4b**). Verify 3D validated the protein structural quality as an excellent structure of the protein (data not shown). The Swiss-Model Interactive Workplace validated the modeled tertiary structure of the protein Rv0986 indicating as the

MolProbity Score of 1.53 and the Ramachandran favored 96.88%. The values for the predicted tertiary structure are including the QMEAN (Qualitative Model Energy Analysis), Cβ, All Atom, solvation, and torsion values of -1.03, -2.87, -1.07, 0.10, and -0.53, respectively (data not shown).

The Prosa-web (Wiederstein and Sippl, 2007) server used for the overall tertiary structure quality assessment. The Prosa-web calculated the Standard Bond Angles and Z-scores. The Z-scores used to estimate the 'degree of nativeness' of the predicted structures (Nas *et al.*, 2020). The Z-scores for the modeled tertiary structures by the three servers including Modeller, Phyre2, and Swiss-Model were -7.42, -7.31, and -7.37, respectively. In this paper, Modeller, Phyre2, and Swiss-Model servers are presenting similar values.

4. CONCLUSIONS

In this study, it is concluded that the protein Rv0986 present in *Mycobacterium tuberculosis* is a functional protein, thereby, associated with microbial infection in humans. As the protein tertiary structure not available in the Protein Data Bank, the structural model prediction with its structural and functional annotation is well explained in this paper. Hence, provides an insight

into this uncharacterized protein Rv0986. This study will bolster and sharpen our knowledge in pathogenesis, therefore, unleashes an opportunity for drug and vaccine targeting against the infection by MTB.
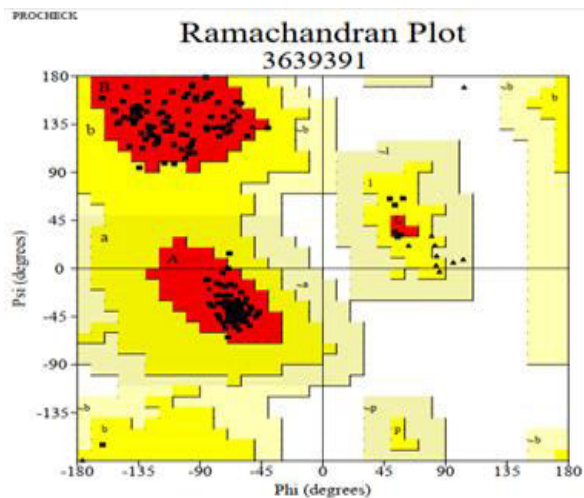

Fig 5: Ramachandran Plot Analysis.

5. ACKNOWLEDGEMENTS

6. CONFLICTS OF INTETEST

The author declares no conflict of interest.

7. REFERENCES

1. Alex Bateman (2019). 'UniProt: a worldwide hub of protein knowledge', *Nucleic Acids Research*, **47**(D1), D506–D515
   https://doi.org/10.1093/nar/gky1049

2. Arnold, K. *et al.* (2020). 'The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling', *Bioinformatics*, **22**(2), 195–201
   https://doi.org/10.1093/bioinformatics/bti770

3. Beis K. (2015). 'Structural basis for the mechanism of ABC transporters', *Biochemical Society Transactions*. Portland Press Ltd, **43**, 889-893.
   https://doi.org/10.1042/BST20150047

4. Combet C, Blanchet C, Geourjon C, Deléage G. (2000). 'NPS@: network protein sequence analysis' (2020), COMPUTER CORNER, **25**(3), 147-150.
   https://doi.org/10.1016/S0968-0004(99)01540-6

5. David T. Jones, (1999). 'Protein secondary structure prediction based on position-specific scoring matrices', *Journal of Molecular Biology,* **292**(2), 195-202
   https://doi.org/10.1006/jmbi.1999.3091

6. Duarte T. A. *et al.* (2017). 'A systematic review of East African-Indian family of *Mycobacterium tuberculosis* in Brazil', *Brazilian J. of Infectious Diseases*, **21**(3), 317–324.
   https://doi.org/10.1016/j.bjid.2017.01.005

7. Ellison A. L. (2018). 'Exploring the role of the striengent respose in by'.

8. Gill C, Hippel P. H. von. (1989). 'Calculation of protein extinction coefficients from amino acid sequence data', *Analytical Biochemistry*, **182**(2), 319-326.
   https://doi.org/10.1016/0003-2697(89)90602-7

9. Guruprasad K. *et al.* (1990). 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 Engineering, Design and Selection*, **4**(2), 155–161.
   https://doi.org/10.1093/protein/4.2.155

10. Holland K. A. and Holland, I. B. (2005) 'Adventures with ABC-proteins: Highly conserved ATP-dependent transporters', *Acta Microbiologica et Immunologica Hungarica*, **52**(3-4), 309–322.
    https://doi.org/10.1556/amicr.52.2005.3-4.4

11. Ikai A. (1980) *Thermostability and Aliphatic Index of Globular Proteins*, *The Journal of Biochemistry,* **88**(6), 1895–1898.
    https://doi.org/10.1093/oxfordjournals.jbchem.a133168

12. Islam S, Saleh MA, ShahriarZaman. (2020). Molecular identification and characterization of smartphone screen associated pathogenic bacteria. *Am. J. Pure Appl. Sci*., **2**(3), 94-103.
    https://doi.org/10.34104/ajpab.020.0940103

13. Kelley L. A. *et al.* (2015) 'The Phyre2 web portal for protein modeling, prediction and analysis', *Nature Protocols*. *Nature Publishing Group*, **10**(6), 845–858.
    https://doi.org/10.1038/nprot.2015.053

14. Martin L. (2016) 'Genomics and transcriptomics of the molting gland (Y-organ) in the blackback land crab, *Gecarcinus lateralis*'. Colorado State University, M.Sc Thesis, 1-97. http://hdl.handle.net/10217/176669

15. Nas FS, Yahaya A, and Ali M. (2020). Prevalence of *Trichomonas vaginalis* among pregnant women attending ante-natal care in Kano, Nigeria. *Eur. J. Med. Health Sci*., **2**(2), 39-45. https://doi.org/10.34104/ejmhs.020.39045

16. ProtParam, E. (2017). 'ExPASy-ProtParam tool'.

17. Shahen MZ, Mahmud S, Uddin ME and Alam MS. (2019). Effect of antibiotic susceptibility and inhibitory activity for the control of growth and survival of microorganisms of extracts of *Calendula officinalis*, *Eur. J. Med. Health Sci.* **1**(1), 1-9. https://doi.org/10.34104/ejmhs.019

18. Vikram, T., Kumar P. (2018). 'Analysis of Hepatitis E virus (HEV) X-domain structural model', *Bioinformation*. **14**(7): 398-403. https://doi.org/10.6026/97320630014398

19. Webb B. and Sali, A. (2016) 'Comparative Protein Structure Modeling Using MODELLER', *Current Protocols in Protein Science*. **86**(1), 2.9.1-2.9.37. https://doi.org/10.1002/cpps.20

20. Wiederstein M, and 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. https://doi.org/10.1093/nar/gkm290

21. WHO*;* Global tuberculosis report 2019 (2019). Tuberculosis (TB), 1-297. https://www.who.int/tb/publications/global_report/en/

22. Wu C. *et al.* (2019) 'Insect ATP-binding cassette (ABC) transporters: Roles in xenobiotic detoxification and Bt insecticidal activity', *Int. J. Mol. Sci*., **20**(11), 2829. https://doi.org/10.3390/ijms20112829

23. Zimmermann, L. *et al.* (2018) 'A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core', *J. of Molecular Biology*. Academic Press, 430(15), 2237–2243. https://doi.org/10.1016/j.jmb.2017.12.007