In-Silico Structural prediction and functional annotation of uncharacterized protein Q7TLC7 (Y14_SARS) presence in SARS-COV2

Ashiqur Rahman[1] Rumana Akter Ruma[2]

1. Dept. of Biochemistry and Microbiology , North South University

2. Dept. of Biotechnology and Genetic Engineering , Islamic University

Abstract

Protein structure prediction strategies point to predict the structures of proteins from their amino acid sequences, utilizing different computational calculations. Basically, the prediction of 3D structure of a protein from its amino acid sequence is one of the foremost critical indecisive issues in computational biology. This paper endeavors to grant a comprehensive presentation of the foremost later exertion and advance on protein structure prediction. Taking after the common flowchart of structure prediction, related concepts and strategies are displayed and experimentally established. In addition, brief presentations are made to a few widely-used prediction methods and the community-wide basic appraisal of protein structure prediction experiments. Here we can see Unknown proteins exist but have not been characterized or connected to known qualities. Domains of unknown function are experimentally distinguished proteins with no known functional or structural domain. In this paper, the examination and characterization of the likely useful perspectives of a hypothetical protein: coronavirus Q7TLC7 (Y14_SARS) was performed utilizing different computational strategies and tools. As the protein tertiary structure not accessible within the Protein Data Bank, the basic demonstrate expectation with its structural and functional annotation is well explained in this paper which consequently, gives an understanding into this hypothetical protein (Q7TLC7) .In this manner, unleashes an opportunity for medicate and immunization focusing on against the disease by COVID19 investigation will be of significance in understanding the mechanism of the infections and will demonstrate to be advantageous within the revelation of new drugs.

Keyword : computational biology , hypothetical protein , functional annotation , COVID19 , infections .

Introduction

SARS CoV was discovered in 2003 as a result of a global outbreak of an atypical type of pneumonia known as Severe Acute Respiratory Syndrome (SARS) (Wang et al., 2016). People who are infected with the SARS CoV-2 virus certainly experienced diffuse alveolar damage which could lead to acute respiratory distress syndrome & death. SARS CoV-2 is a positive sense, single stranded RNA coronavirus with large genome sizes that is enclosed by an enveloped structure, classified into four genera: alpha coronavirus, beta coronavirus, delta coronavirus & gamma coronavirus among these alpha & beta coronavirus infecting humans & it's a main source of disaster in the 21st century (Seah et al., 2020; Sen et al., 2020). SARS CoV-2 also known as COVID-19, was first recognized in the wet market of Wuhan, China (Khan et al., 2021). COVID-19 is the name given to this disease by the WHO (World Health Organization) , an acronym that stands for "C- corona", "VI- viruses", "D- disease" & "19- the year 2019" (Zhou et al., 2020). COVID-19, a Coronavirus outbreak, first appeared in China at the end of 2019 and rapidly spread throughout the world. By March 2020, several countries across the world had implemented a lockdown and implemented stay-at-home and work-from-home policies as a result of the rapid spread (https://doi.org/10.1016/genrep.2021.101064). A significant number of SARS-CoV-2 ORFs have unknown or poorly understood features, SARS-CoV has its own collection of accessory proteins (Baruah et al., 2020). An uncharacterized accessory protein of Q7TLC7 present in SARS-CoV-2 of Human SARS CoV (Giri et al., 2020). However, physicochemical characterizations of the tertiary structure with ligand binding active sites have yet to be published. Therefore, the present study reports by using various bioinformatic methods, an attempt was made to predict the structure and biological function & evolutionary analysis of an uncharacterized protein (Q7TLC7). It is exceptionally important to do the functional annotation of the hypothetical proteins which is included in disease, drug resistance and fundamental biosynthetic for the improvement of the strong antiviral between the infectious agents. To form them as the potential targets of antimicrobial drugs, the enhancement of understanding of this protein is remarkably vital. Here all the descriptions and results were recorded within the table.

2. MATERIALS AND METHODS

Selection of the hypothetical protein

Speculative proteins were looked at the protein database of NCBI utilizing the keyword, "hypothetical protein," and the resultant hits were haphazardly chosen to ponder the close relatives utilizing blast programs. To predict the work of the query protein, a similitude look was performed utilizing NCBI impact tools to recognize proteins that will have structural similarity with that of the hypothetical protein.

2.1 Sequence retrieval

The amino acid sequence of Q7TLC7 obtained from the National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov) (Johnson et al., 2008) with the PUBMED ID 12730501. Then the sequence was stored as a FASTA format sequence.

2.2 Physico-chemical properties analysis

We used a web-based server named Protparam tool of ExPASy (https://web.expasy.org/protparam) (Gasteiger et al., 2003) for the determination of the physicochemical properties of the uncharacterized protein.

2.3 Functional Annotation Prediction

The CD Search tool of NCBI was used for domain prediction. The CD Search tool (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) (Marchler-Bauer et al., 2005) predicted a domain of the protein Q7TLC7.

2.4 Secondary structure prediction

The retrieved sequence was used for the prediction of secondary structure elements of the protein by the Self-Optimized Prediction Method with Alignment (SOPMA https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_sopma.html) (Geourjon & Deléage, 1995)) &

the PSIPRED tool (http://bioinf.cs.ucl.ac.uk/psipred/) (McGuffin et al., 2000) used to predict the secondary structure of Q7TLC7.

2.5 Tertiary structure modeling

The tertiary structure of the protein was predicted by the servers including Moeller following the HHpred tool (https://toolkit.tuebingen.mpg.de) (Zimmermann, et al., 2018).

2.6 Model quality assessment

Finally the quality of the predicted tertiary structure was assessed by PROCHECK (https://servicesn.mbi.ucla.edu/PROCHECK/) (Laskowski et al., 1993) & ERRAT Structure Evaluation server (https://servicesn.mbi.ucla.edu/ERRAT/) (Colovos & Yeates, 1993) & PROVE.

2.7 Active site prediction

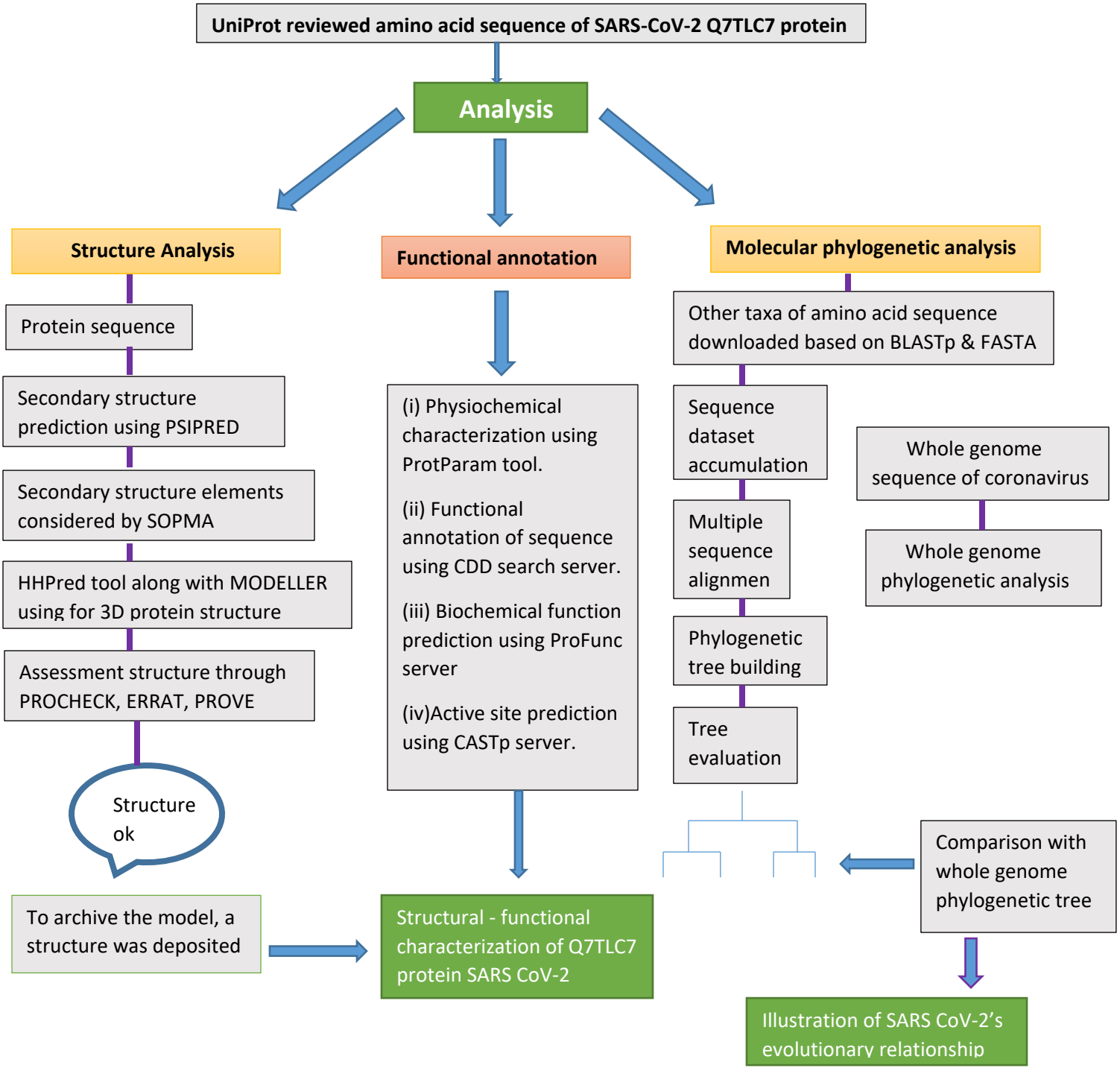
We used CASTp server (sts.bioe.uic.edu/castp/index.html?3igg) (Binkowski et al., 2003) for identifying the active site of our hypothetical protein. Basically, The CASTp server takes protein structures within the PDB arrange and a test sweep as input for topographic computation. Through the instinctive interface, inputting our claim protein structures (PDB format) to ask customized computation.

In expansion, the CASTp server too gives engraves of topographic highlights. These results can be straightforwardly downloaded from CASTp server, which can be visualized utilizing either PymoL or BIOVIA Discovery Visualizer 2020.

2.8 Molecular phylogenetic analysis

Multiple sequence alignment (Figure 4) was considered the FASTA sequences of the uncharacterized protein Q7TLC7 (Y14_SARS) and the homologous annotated proteins. In order to confirm homology assessment between the proteins, down to the complex and subunit level, phylogenetic analysis was additionally performed. Phylogenetic tree was constructed based on the alignment and BLAST result give the similar concept about the protein is shown in Figure 5. The distances between branches are also included.

Result and discussion



3. RESULTS AND DISCUSSION

3.1 Sequence & similarity information

The BLASTp result against non-redundant and SwissProt database shown in Table 1 & 2.

Table 1: Similar protein obtained from Non-redundant Uniprot KB/SwissProt sequences.

Protein ID	Organism	Protein name	Identity	Score	e-value
YP_009825063	SARS coronavirus Tor2	ORF9a protein	100	140	5e-42
ACZ71988.1	SARS coronavirus ExoN1	hypothetical orf14 protein	98.57	139	2e-41
AHX37568.1	Rhinolophus affinis coronavirus	protein 14	94.29	132	6e-39
ACU31041.1	SARS coronavirus Rs_672/2006	Hypothetical protein	94.29	132	1e-38
ANA96037.1	Bat coronavirus	9b protein	84.29	117	1e-32

Table 2: Similar protein obtained from UniProt database

Entry Name	**Organism**	**Protein name**	**Identity**	**Score**	**e-value**

Q7TLC7.1	Severe acute respiratory syndrome related-coronavirus	Uncharacterized protein 14	100	140	7e-45
Q31515.1	Bat SARS CoV Rp3/2004	Uncharacterized protein 14	91.43	127	1e-39
PODTD3.1	Severe acute respiratory syndrome coronavirus2	Uncharacterized protein 14	84.21	100	9e-29

3.2 Physico-chemical characterization

A FASTA format sequence of Q7TLC7 of SARS CoV-2 (severe acute respiratory syndrome coronavirus 2) was restored which is used as a query sequence for the characterization of physicochemical properties (Saim et al., 2020). The protein contains 70 amino acid with a molecular weight of 7852.33 Da & the calculated theoretical pI was 6.25. The molecular formula of the protein was identified as $C_{356}H_{573}N_{93}O_{96S5}$. 6 & 7 were the total number of positively charged residues (Arg+Lys) & the total number of negatively charged residues (Asp+Glu) periodically. High Extinction coefficient values 8730 indicates the presence of Cys, Trp, and Tyr residues (Gill & Hippel, 1989). Higher values of Aliphatic Index of the query protein was 119.86 which gives an indication of proteins stability over a wide temperature range (Ikai, 1980). The protein is stable in nature because its instability index (26.67) is lower than 40 (Gamage et al., 2019). The grand average of hydropathicity (GRAVY) was positive which indicates the protein is polar due to its higher GRAVY indices value 0. 310 (Fernández-Fernández & Corpas, 2016). Protein half-life computed was found to be 30 hours (mammalian reticulocytes, in vitro), >20 hours (yeast, in vivo), >10 hours (Escherichia coli, in vivo). All results are shown in Table 1. The amino acid composition showed in Table 2, which obtained from the ExPASy ProtParam Tool.

Table 3: Physico-chemical properties analysis of the hypothetical protein.

No of Amino acid	MW	pI	(Asp+Glu)	(Arg+Lys)	Ext. Coefficient (all Cys residues form cysteines)	Ext. Coefficient (all Cys residues are reduced)	Aliphatic index (AI)	Instability index (II)	Grand average of hydropathicity (GRAVY)
70	7852.33	6.25	7	6	8730	8480	119.86	26.67	0.310

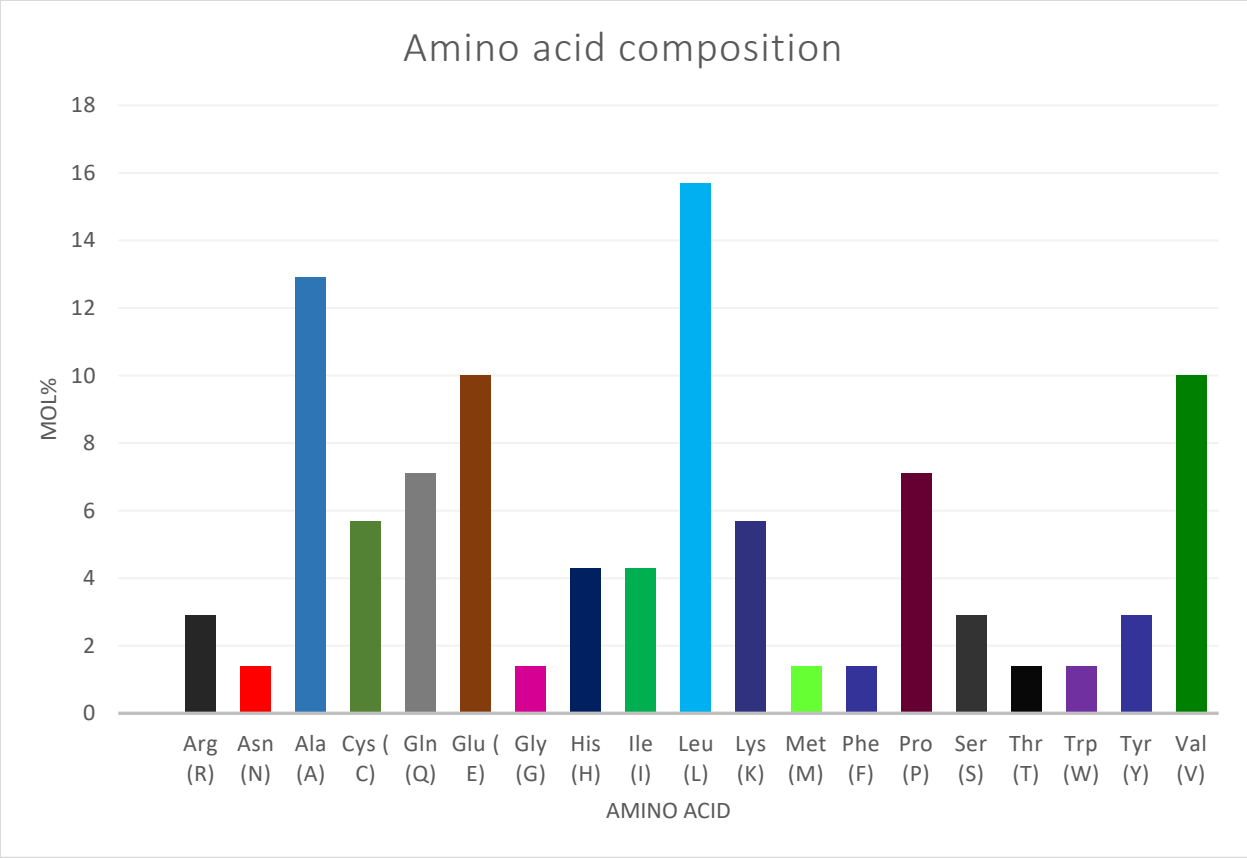


Fig2: Amino acid composition.

3.3 Secondary structure analysis

The SOPMA server was used for the secondary structure prediction, the default setup (window width of 17, similarity threshold of 8, and the number of states of 4) appeared the proportions of alpha helix, beta turn, extended strand, and the random coil of protein as 81.43%, 1.43%, 1.43%, and 15.71% periodically (**Table 5**). PSIPRED is showing the higher confidence of the prediction of the helix, strand, and coil (**Fig 3**).

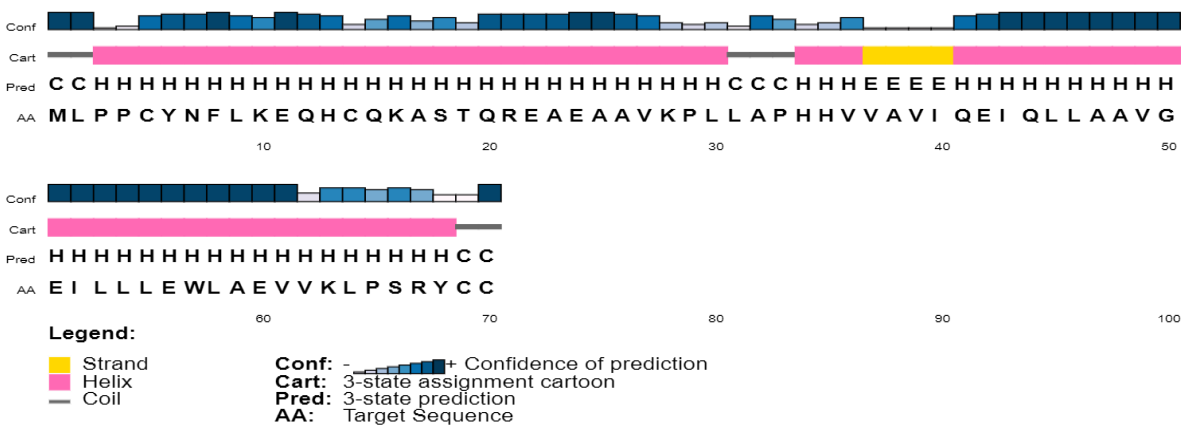


Fig 3: Predicted Secondary structure

Table 5: Secondary structure elements

Secondary Structure Elements	Values%
Alpha helix (Hh)	81.43
310 helix (Gg)	0.00
Pi helix (Ii)	0.00
Beta bridge (Bb)	0.00
Extended strand (Ee)	1.43
Beta turn (Tt)	1.43
Bend region (Ss)	0.00
Random coil (Cc)	15.71
Ambiguous states	0.00
Other states	0.00

.

Table- 6: Predicted functions of SARS-CoV-2 Q7TLC7 protein with ProFunc score.

		Gene Ontology (GO) Terms		
Protein Name	**Protein name terms**	**Cellular component**	**Biological process**	**Biochemical function**
Q7TLC7	domain (1.26) human (0.91) between (0.83) alpha-catenin (0.51) smba (0.50) c-di-gmp (0.50) heterocomplex between human keratin (0.50) between human keratin coil (0.50)-	cytoplasm (1.52) cell (1.52) cell part (1.52) intracellular (1.52)	cellular process (1.52) cellular metabolic process (1.15) metabolic process (1.15) biological regulation (0.87)	cellular process (1.52) cellular metabolic process (1.15) metabolic process (1.15) biological regulation (0.87)

3.4 Three-dimensional structure analysis

HHpred server predicted 3D structure of the protein. The query sequence of Q7TLC7 was inserted as FASTA format to HHpred & the highest scoring template was selected (4O6K_B) among the 100 number of hits with the probability rate 27.59, E-value 130, Score 19.1, SS 3.4, Aligned Cols 22 and the Target length 144 (data not shown). 4O6K_B is the crystal structure of *Danio rerio.* Finally 3D structure of the query protein was stored in PDB format prdicted by Modeller.

Fig: Structure of Q7TLC7 predicted by Modeller**.**

3.5 Validation of predicted protein structure

Validation of the predicted three-dimensional model was assessed by PROCHECK program (fig:3) through Ramachandra plot analysis where residues in the most favored regions covered 100% which is the quality of a valid model. Afterwards, the predicted model of 3D structure for the target sequence was verified by structure validation server ERRAT & PROVE. The overall quality factor was 85.7143 predicted by ERRAT server which indicates a good model. The PROVE resulted Z-score mean was -0.256 & Z-score RMS was 1.540.

(a)

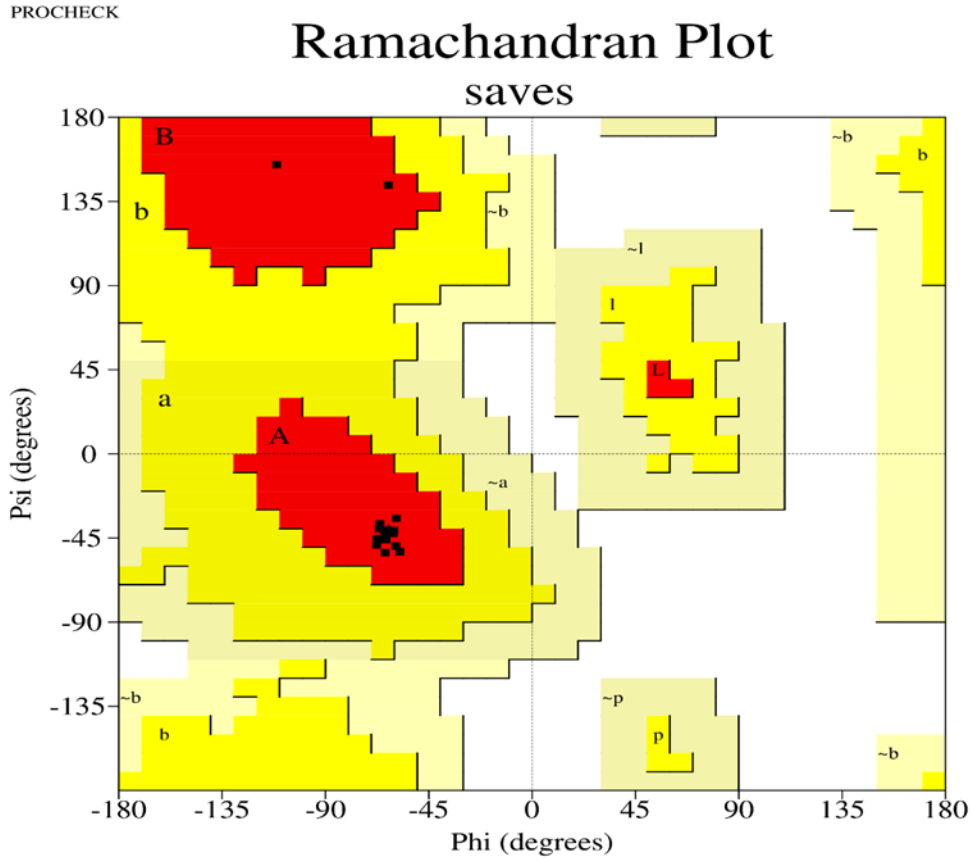
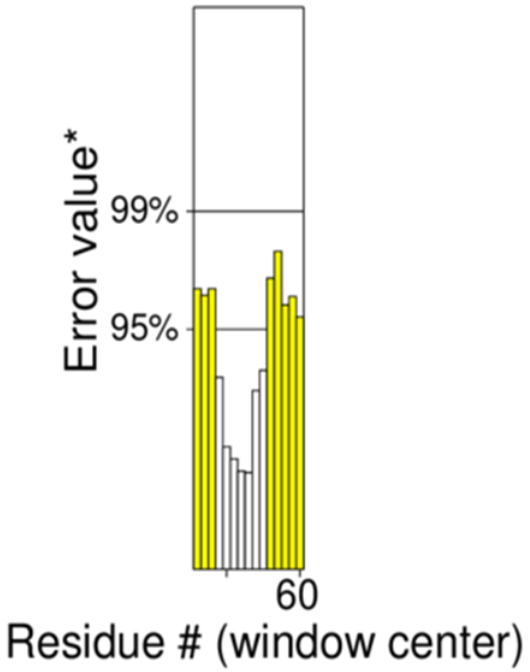


Table 7: Ramachandra plot analysis.

Ramachandran plot statistics		%
Residues in the most favored regions [A, B, L]	20	100%
Residues in the additional allowed regions [a, b, l, p]	0	0.0%
Residues in the generously allowed regions [a, b, l, p]	0	0.0%
Residues in the disallowed regions	0	0.0%
Number of non-glycine and non-proline residues	20	100%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown in triangles)	1	
Number of proline residues	0	
Total number of residues	23	

(b)



(c)

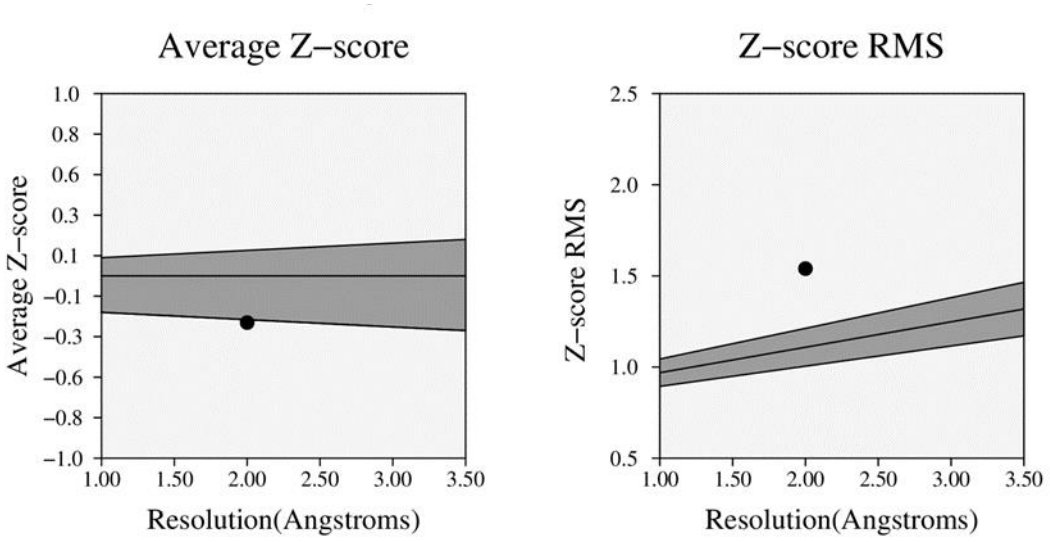


Figure 2: (a) Ramachandran plot of modelled structure validated by PROCHECK program.

(b) ERRAT score, (c) PROVE plot shows average Z score.

3.6 Active site of the hypotheticl protein

CASTp was used to determine the active site of the protein. In 70 amino acids resides, only 23 amino acids are potent active site. The best active site was found in areas with 9.623 & a volume of 4.336 amino acids.

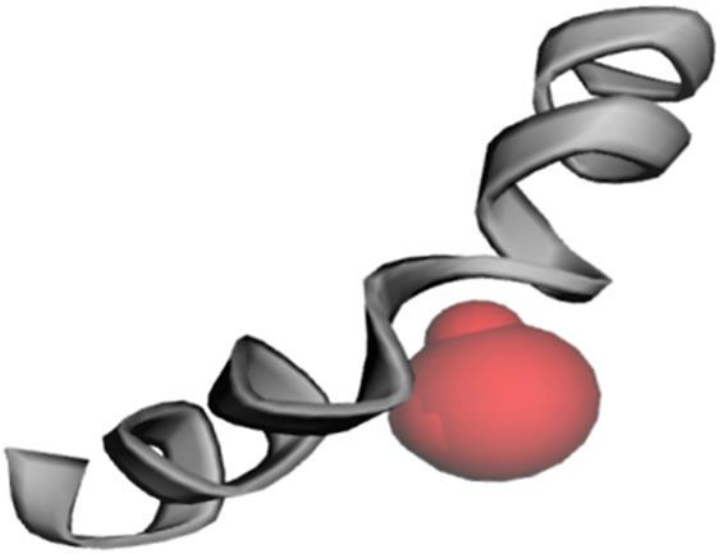


Fig 3: The location of active site in this protein.

3.7 Evolutionary analysis by Maximum Likelihood method

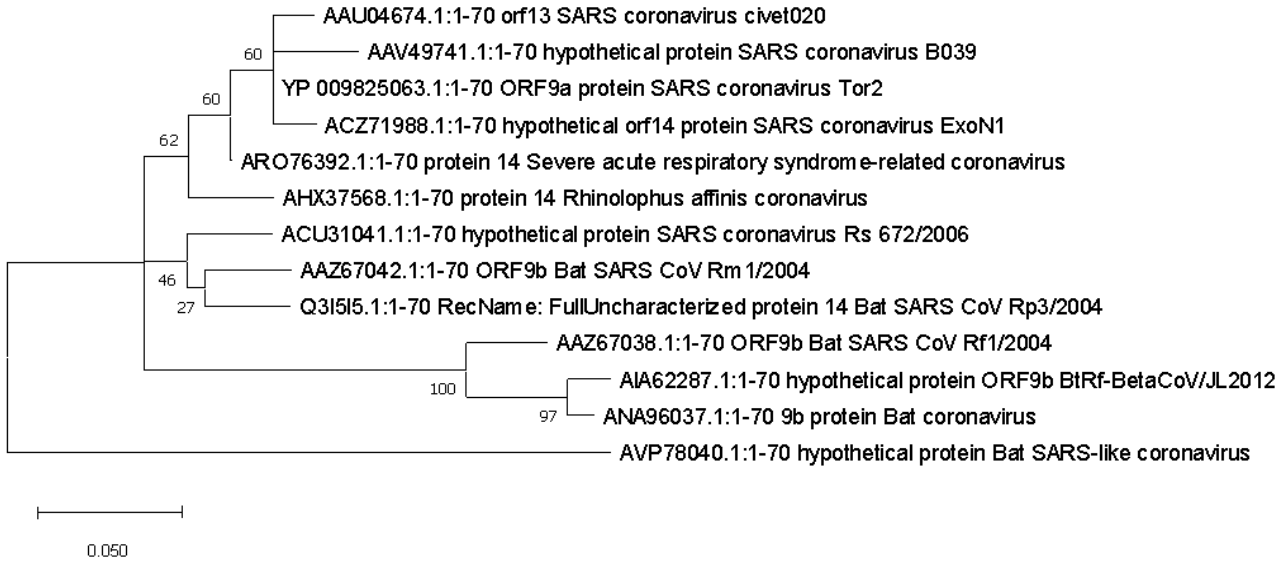


Fig: The evolutionary history was inferred by using the Maximum Likelihood method and JTT matrix-based model. The tree with the highest log likelihood (-469.13) is shown. There was a total of 70 positions in the final dataset. Evolutionary analyses were conducted in MEGA X.

Conclusion

Studying the functions of hypothetical protein is significant since it encourages the advance comprehension of its part in biochemical/physiological pathways and the recognizable proof of novel classes of therapeutic targets .Expectancy is this study will support and hone our information in pathogenesis, it will improve our pathogenesis awareness and give us a chance to target the protein compound . The recognized protein uncovered a few characteristics such as , Sequence similarity data, Gene ontology, the evolutionary history etc So, broadly redaction of invitro research must be carried out to tentatively approve the conceivable outcomes appeared here and to discover out the proteins' part in life science research .

Reference

Baruah, C., Devi, P., & Sharma, D. K. (2020). Sequence Analysis and Structure Prediction of SARS-CoV-2 Accessory Proteins 9b and ORF14: Evolutionary Analysis Indicates Close Relatedness to Bat Coronavirus. *BioMed Research International*, *2020*. https://doi.org/10.1155/2020/7234961

Binkowski, T. A., Naghibzadeh, S., & Liang, J. (2003). CASTp: Computed Atlas of Surface Topography of proteins. *Nucleic Acids Research*, *31*(13), 3352–3355. https://doi.org/10.1093/nar/gkg512

Colovos, C., & Yeates, T. O. (1993). Verification of protein structures: Patterns of nonbonded atomic interactions. *Protein Science*, *2*(9), 1511–1519. https://doi.org/10.1002/pro.5560020916

Fernández-Fernández, Á. D., & Corpas, F. J. (2016). In Silico Analysis of Arabidopsis thaliana Peroxisomal 6-Phosphogluconate Dehydrogenase. *Scientifica*, *2016*. https://doi.org/10.1155/2016/3482760

Gamage, D. G., Gunaratne, A., Periyannan, G. R., & Russell, T. G. (2019). Applicability of Instability Index for In vitro Protein Stability Prediction. *Protein & Peptide Letters*, *26*(5), 339–347. https://doi.org/10.2174/0929866526666190228144219

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D., & Bairoch, A. (2003). ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Research*, *31*(13), 3784–3788. https://doi.org/10.1093/nar/gkg563

Geourjon, C., & Deléage, G. (1995). Sopma: Significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Bioinformatics*, *11*(6), 681–684. https://doi.org/10.1093/bioinformatics/11.6.681

Giri, R., Bhardwaj, T., Shegane, M., Gehi, B., Kumar, P., Gadhave, K., Oldfield, C., & Uversky, V. (2020). *When Darkness Becomes a Ray of Light in the Dark Times: Understanding the COVID-19 via the Comparative Analysis of the Dark Proteomes of SARS-CoV-2, Human SARS and Bat SARS-Like Coronaviruses*. 1–63. https://doi.org/10.1101/2020.03.13.990598

Ikai, A. (1980). Thermostability and aliphatic index of globular proteins. *Journal of Biochemistry*, *88*(6), 1895–1898. https://doi.org/10.1093/oxfordjournals.jbchem.a133168

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., & Madden, T. L. (2008). NCBI BLAST: a better web interface. *Nucleic Acids Research*, *36*(Web Server issue), 5–9. https://doi.org/10.1093/nar/gkn201

Khan, A. G., Kamruzzaman, M., Rahman, M. N., Mahmood, M., & Uddin, M. A. (2021). Quality of life in the COVID-19 outbreak: influence of psychological distress, government strategies, social distancing, and emotional recovery. *Heliyon*, *7*(3), e06407. https://doi.org/10.1016/j.heliyon.2021.e06407

Laskowski, R. A., MacArthur, M. W., Moss, D. S., & Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, *26*(2), 283–291. https://doi.org/10.1107/s0021889892009944

Marchler-Bauer, A., Anderson, J. B., Cherukuri, P. F., DeWeese-Scott, C., Geer, L. Y., Gwadz, M., He, S., Hurwitz, D. I., Jackson, J. D., Ke, Z., Lanczycki, C. J., Liebert, C. A., Liu, C., Lu, F., Marchler, G. H., Mullokandov, M., Shoemaker, B. A., Simonyan, V., Song, J. S., … Bryant, S. H. (2005). CDD: A Conserved Domain Database for protein classification. *Nucleic Acids Research*, *33*(DATABASE ISS.), 192–196. https://doi.org/10.1093/nar/gki069

McGuffin, L. J., Bryson, K., & Jones, D. T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics*, *16*(4), 404–405. https://doi.org/10.1093/bioinformatics/16.4.404

Saim, A., Saikat, M., Sheikh, B., Rahman, M., Ripon, A. B., Sheikh, B., & Rahman, M. (2020). Structure Prediction, Characterization, and Functional Annotation of Uncharacterized Protein BCRIVMBC126_02492 of Bacillus cereus: An In Silico Approach. *American Journal of Pure and Applied Biosciences*, *July*, 104–111. https://doi.org/10.34104/ajpab.020.01040111

Seah, I., Su, X., & Lingam, G. (2020). Revisiting the dangers of the coronavirus in the ophthalmology practice. *Eye (Basingstoke)*, *34*(7), 1155–1157. https://doi.org/10.1038/s41433-020-0790-7

Sen, D., Debnath, P., Debnath, B., Bhaumik, S., & Debnath, S. (2020). Identification of potential inhibitors of SARS-CoV-2 main protease and spike receptor from 10 important spices through structure-based virtual screening and molecular dynamic study. *Journal of Biomolecular Structure and Dynamics*, *0*(0), 1–22. https://doi.org/10.1080/07391102.2020.1819883

Wang, F., Chen, C., Tan, W., Yang, K., & Yang, H. (2016). Structure of Main Protease from Human Coronavirus NL63: Insights for Wide Spectrum Anti-Coronavirus Drug Design. *Scientific Reports*, *6*(March), 1–12. https://doi.org/10.1038/srep22677

Zhou, P., Yang, X. Lou, Wang, X. G., Hu, B., Zhang, L., Zhang, W., Si, H. R., Zhu, Y., Li, B., Huang, C. L., Chen, H. D., Chen, J., Luo, Y., Guo, H., Jiang, R. Di, Liu, M. Q., Chen, Y., Shen, X. R., Wang, X., … Shi, Z. L. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, *579*(7798), 270–273. https://doi.org/10.1038/s41586-020-2012-7

Gill, S. C., & Hippel, P. H. (1989, November 1). Calculation of protein extinction coefficients from amino acid sequence data. *ScienceDirect, 182*(2), 319-326. Retrieved from https://doi.org/10.1016/0003-2697(89)90602-7

Zimmermann, L., Andrew, Seung-Zin, David, Jonas, Marko, . . . Vikram. (2018, July 20). A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *430*(15), 2237-2243. Retrieved from https://doi.org/10.1016/j.jmb.2017.12.007