Research Article

An *in silico* approach for structural and functional annotation of matrix protein of *Nipah henipavirus*: A protein functional analysis

Apurbo Kumar Paul[1], Abu Saim Mohammad Saikat[2]*

[1]Department of Materials Science and Engineering, University of Rajshahi, Rajshahi 6205, Bangladesh; apurbomse@protonmail.com**;** https://orcid.org/0000-0002-9485-2778

[2]Department of Biochemistry and Molecular Biology, Life Science Faculty, Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Gopalganj 8100, Bangladesh; asmsaikat.bmb@gmail.com; https://orcid.org/0000-0002-6850-1245

*Correspondence should be addressed to asmsaikat.bmb@gmail.com (Abu Saim Mohammad Saikat).

Abstract

Nipah henipavirus is an emerging RNA virus with increased mortality threatening global security. In South and Southeast Asia, the Nipah virus has caused numerous disease outbreaks. The matrix protein in *Nipah henipavirus* has an important role, in connecting the viral envelope with the virus core. For virus assembly, linking the viral envelope with the virus core are very crucial. Through functional and structural explanation evaluations, bioinformatics strategies can help us better understanding of the protein. This investigation aims to allocate the structural and functional annotation of protein. Moreover, the investigation attributes physicochemical parameters, three-dimensional structure, and functional annotation of the protein (QBQ56721.1) applying an *in silico* perspective. The *in silico* analysis confirmed the protein's hydrophilic nature, with a secondary structure dominated by alpha (α) helices. Based on several quality assessment methodologies, the tertiary-structure model of the protein has been shown to be reasonably consistent. The functional explanation suggested the protein as a structural protein connected to the viral envelope with the virus core, a protein required for virus assembly. This investigation unleashes the significance of the matrix protein (QBQ56721.1) as a functional protein required for *Nipah henipavirus*.

Keywords: *Nipah henipavirus*; Matrix protein; Functional annotation; Protein characterization;

Introduction

Nipah henipavirus is a bat-borne virus that can infect humans and other animals.[1]. Nipah virus infection is a zoonotic disease that is spread from animals to humans [2]. It can also be spread through contaminated food or from person to person [2]. It causes a variety of symptoms in infected individuals, ranging from asymptomatic (subclinical) infection to acute respiratory sickness and deadly encephalitis [2]. Nipah virus is a member of the Paramyxoviridae family and the genus Henipavirus along with the Hendra virus, which has also caused disease outbreaks[3].The Nipah virus genome is a single (nonsegmented) negative-sense, single-stranded RNA with a length of over 18 kb, far longer than that of other paramyxoviruses [1][4]. The Nipah virus was initially detected in pigs and pig farmers in Peninsular Malaysia in 1998 [5]. Infection outbreaks of the Nipah virus have been documented in Malaysia, Singapore, Bangladesh, and India [6]. The highest rates of death from Nipah virus infection have been reported in Bangladesh, where outbreaks are most common in the winter [6]. Consumption of fruits or fruit products (such as raw date palm juice) contaminated with urine or saliva from infected fruit bats was the most likely cause of infection in later outbreaks in Bangladesh and India [2]. About 700 human cases of Nipah virus had been reported as of May 2018, with 50 to 75 percent of those affected dying [5]. In the Indian state of Kerala, an epidemic of the disease led to 17 deaths in May 2018 [7].

A study of the proteins using bioinformatics tools allows researchers to assess their three-dimensional structural conformation, classify new domains, explore certain pathways to gain a better understanding of our evolutionary tree, uncover more clusters, and assign roles to the proteins [8]. This knowledge can also be used to develop successful pharmacological methods and aid in the development of new drugs to treat a wide range of diseases[9][10].

2. Methodology

2.1. Protein Selection and Sequence Retrieval.

The amino acid (aa) sequence of the matrix protein found in *Nipah henipavirus* was obtained in FASTA format from the NCBI database (https://www.ncbi.nlm.nih.gov/).

2.2. Physicochemical Characterization.

The amino acid sequence composition, instability index, aliphatic index, GRAVY (assessment of hydrophobicity or hydrophilicity of a protein), and extinction coefficients were all measured using the ExPASy server ProtParam tool [11]. The theoretical isoelectric point (pI) of the QBQ56721.1 protein was also measured using SMS Suite (v2.0) [12].

2.3. Functional Annotation Prediction.

The conserved domain in the protein QBQ56721.1 was predicted using the NCBI platform's CD-search tool [13]. The ExPASy software's ScanProsite tool (https:// prosite.expasy.org/scanprosite/), Pfam tool [14] were used to determine protein motifs. The evolutionary relationships of the protein QBQ56721.1 were assigned by the SuperFamily program [15].

2.5. Secondary Structural Assessment.

The self-optimized prediction method with alignment (SOPMA) is used to predict secondary structure elements [16] . The secondary structure was predicted using the SPIPRED v.4.0 [17] algorithm.

2.6. Three-Dimensional Structure Prediction and Validation.

With Modeller [18], HHpred [19] predicted the three-dimensional (tertiary) structure. The most suitable template (HHpred ID: 6BK6 A) was chosen for creating the tertiary structure, with probability, E value, aligned cols, and goal lengths of 100, 2.4e-116, 342, and 372 correspondingly. To predict the Ramachandran plot and validate the expected tertiary structure, the PROCHECK [20] tool of the SAVES v.6.0 program (https://saves.mbi.ucla.edu/) was used.

3. RESULTS AND DISCUSSION

3.1. Sequence Retrieval.

The amino acid (aa) sequence of *Nipah henipavirus* protein (QBQ56721.1) was obtained from the NCBI database. The 352-amino-acid-long protein sequence was

used to model the tertiary structure of the protein QBQ56721.1. Table 1 provides additional information on the protein (QBQ56721.1).

Table 1: Protein retrieval.

Protein individualities	**Protein information**
Locus	QBQ56721
Amino acid	352 aa
Definition	matrix protein [*Nipah henipavirus*]
Accession	QBQ56721
Version	QBQ56721.1
Source	*Nipah henipavirus*
Organism	*Nipah henipavirus*
FASTA sequence	MEPDIKSISSESMEGVSDFSPSSWENGGYLDKVEP EIDENGSMIPKYKIYTPGANERKYNNYMYLICYGF VEDVERTPETGKRKKIRTIAAYPLGVGKSASHPQD LLEELCSLKVTVRRTAGSTEKVVFGSSGPLNHLVP WKKVLTGGSIFNAVKVCRNVDQIQLDKHQALRIF FLSITKLNDSGIYMIPRTMLEFRRNNAIAFNLLVYL KIDADLSKMGIQGSLDKDGFKVASFMLHLGNFVR RAGKYYSVDYCRRKIDRMKLQFSLGSIGGLSLHIK INGVISKRLFAQMGFQKNLCFSLMDINPWLNRLT WNNSCEISRVAAVLQPSVPREFMIYDDVFIDNTGR ILKG

3.2. Physicochemical Properties.

The amino acid sequence of QBQ56721.1, which is found in *Nipah henipavirus*, was obtained in FASTA format and utilized as a query sequence for physicochemical parameter measurement. The protein is stable because its instability index is 30.59, which is less than 40. [21]. The theoretical isoelectric point (pI) of the protein (pI 9.31, 9.65*) indicates that it is basic [22]. The molecular weight, aliphatic index, instability index, and GRAVY are 39847.16 Dalton, 89.69, 30.59 and -0.212 respectively (Table 2). The protein's higher aliphatic index value of 89.69 indicates increased thermos-stability over a wide temperature range, which is a favorable factor [23]. The GRAVY index value of -0.212 suggested the protein's hydrophilic character and hence the prospect of more water interaction [24].

Table 3: Physicochemical parameters.

Parameters	Value
Molecular weight	39847.16
Theoretical pI	9.31, 9.65*
Total number of negatively charged residues (Asp+Glu)	36
Total number of positively charged residues (Arg+Lys)	48
Formula	$C_{1787}H_{2831}N_{485}O_{510}S_{18}$
Total number of atoms	5631
The estimated half-life	(a) 30 hours (mammalian reticulocytes, in vitro). (b) >20 hours (yeast, in vivo). (c) >10 hours (Escherichia coli, in vivo).
Instability index (II)	30.59
Aliphatic index	89.69
Grand average of hydropathicity (GRAVY)	-0.212

∗pI calculated by the SMS v.2.0.

3.4. Functional Annotation of QBQ56721.1.

The NCBI CDD tool identifies the domain that appears in identical protein sequences. RPS-BLAST is used by CD-Search to compare a test sequence to position-specific rating datasets compiled from conserved domain (CD) alignments in the CD protein cluster. The CD search engine identified a conserved domain in the protein QBQ56721.1 as a viral matrix protein (matrix, accession no. pfam00661). Viral matrix proteins are structural proteins that connect the viral envelope and the virus core [25]. The matrix protein play an important role in virus assembly and linking the viral envelope with the virus core [25]. It's possible that they are found in Morbillivirus, Paramyxovirus, and Pneumovirus [25]. A motif was also predicted by the Pfam software at locations 16–349 (Pfam ID: PF00661; Viral matrix protein; e value of 1:7 10146). Protein motifs are small regions of three-dimensional protein structure or amino acid sequence that are shared by multiple proteins[26]. Motifs are distinct regions of protein structure that may or may not be defined by a distinct chemical or biological function[26].

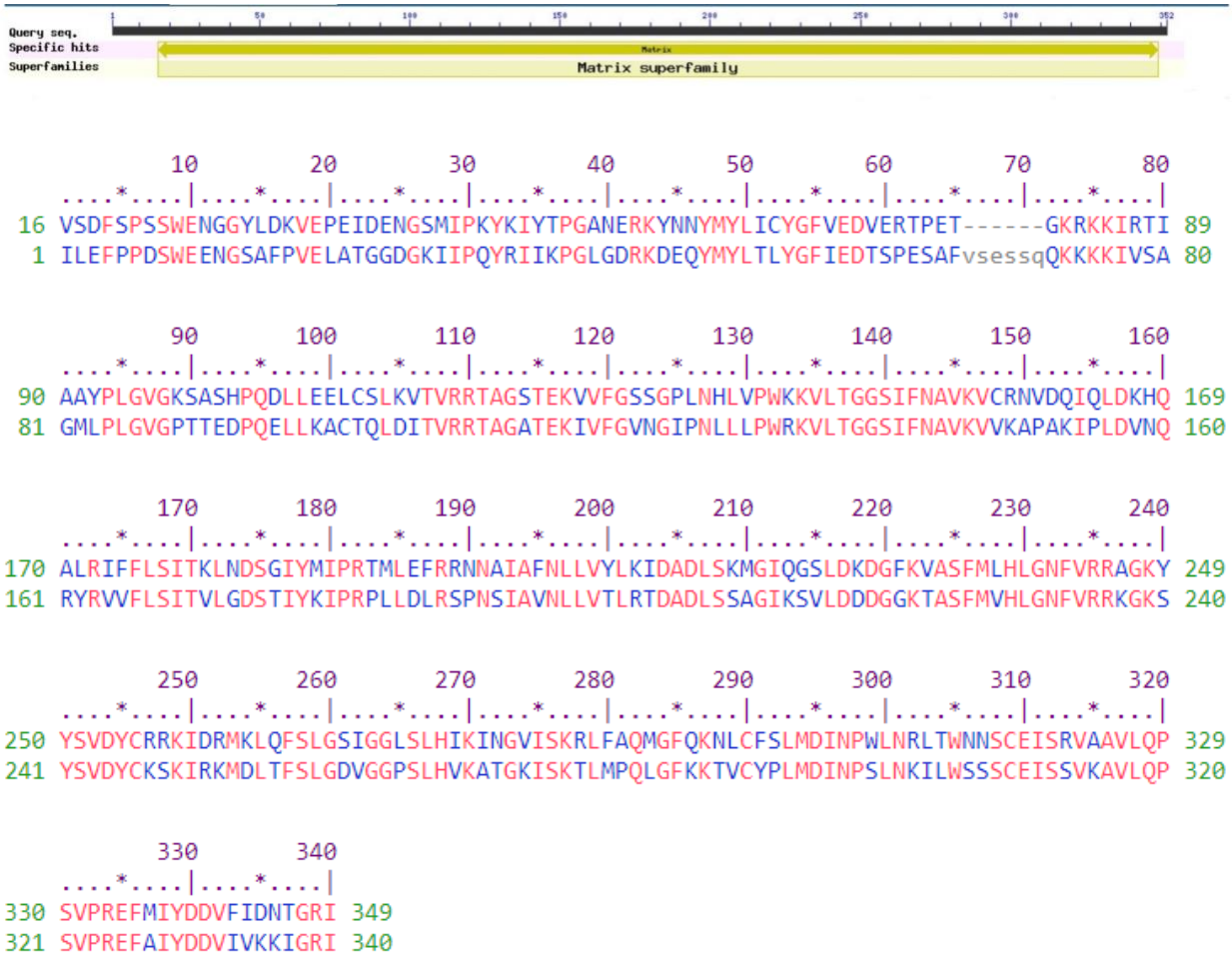

Figure 1: Functional annotation of the protein QBQ56721.1. The graphical summary represents the conserved domains identified inthe query sequence. The aligned sequences represent the conserved domains identified on the query sequence by comparing with theconserved protein domain family, matrix (CDD accession no. pfam00661). The Pfam software predicted a motif at 16–349 (accession no. PF00661) as viral matrix protein. The SuperFamily program predicted the protein as a member of the matrix superfamily.

The CDD technique also confirmed the presence of viral matrix protein at the 17–348 position. The lone member of the superfamily cl02918 is the viral matrix protein (CDD no. pfam00661). A protein superfamily is a group of proteins made up of one or more protein families [27]. The set of all superfamilies must be a partitioning of the set of all protein sequences or subsequences defined by the protein families relationship, and each superfamily must be closed under transitivity [27].The protein QBQ56721.1 (Figure 1) was predicted to be closely related to the matrix superfamily by the SuperFamily tool (e value of 0).

3.6. Secondary Structure Inquiry.

Protein structure and function is inextricably linked. Helix, coil, sheet, and turn are secondary structural elements that have a fantastic association with protein function, structure, and engagement [28]. The secondary-structural element of the protein (QBQ56721.1) was predicted by the SOPMA software when the alpha helix (Hh), extended strand (Ee), beta turn (Tt), and random coil (Cc) were 60 (17.05%), 87 (24.72%), 20 (5.68%), and 185 (52.56%), respectively (Table 5). The PSIPRED software predicts the helix, strand, and coil of the matrix protein (QBQ56721.1) with more confidence (Figure 1). Table 4 shows the amino acid composition obtained from the ExPASy service's ProtParam Tool.

Table 4: Amino acid composition.

Amino Acids	Percentage (%)
Ala (A)	5.2%
Arg (R)	2.1%
Asn (N)	9.4%
Asp (D)	5.7%
Cys (C)	0.2%
Gln (Q)	5.2%
Glu (E)	8.5%
Gly (G)	4.8%
His (H)	1.3%
Ile (I)	6.3%
Leu (L)	9.2%
Lys (K)	9.1%
Met (M)	2.0%
Phe (F)	4.0%
Pro (P)	3.0%
Ser (S)	6.9%
Thr (T)	6.9%
Trp (W)	0.6%
Tyr (Y)	3.0%
Val (V)	6.7%

Table 5: Secondary structural elements.

Secondary Structure Elements	Values (%)
Alpha helix (Hh)	60 (17.05)
3_{10} helix (Gg)	0 (0.00)
Pi helix (Ii)	0 (0.00)
Beta bridge (Bb)	0 (0.00)
Extended strand (Ee)	87 (24.72)
Beta turn (Tt)	20 (5.64)
Bend region (Ss)	0 (0.00)
Random coil (Cc)	185 (52.56)
Ambiguous states (?)	0 (0.00)
Other states	0 (0.00)

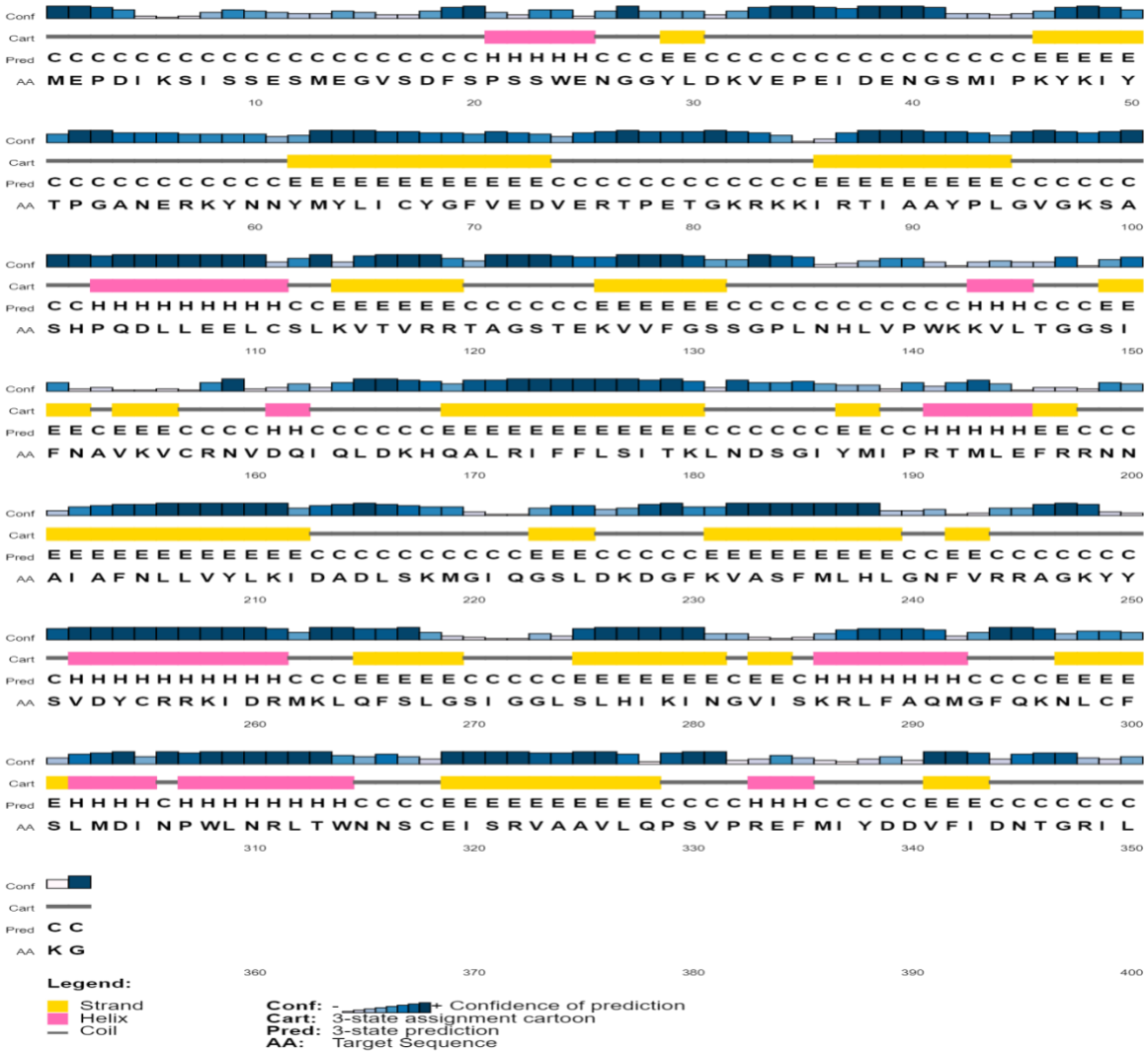

Figure 2 : Predicted secondary structure

3.7. Tertiary-Structure Prediction and Validation.

The target sequence of QBQ56721.1 in FASTA format was inserted into the HHpred Template Selection tool as input, and the most suitable template (6BK6 A) was selected with a probability rate of 100%, E-Value of 2.4e-116, Cols of 342, and the target length of 372, and finally stored the tertiary modeled protein structure in PDB format predicted by Modeller (Fig 3). The Ramachandran plot by PROCHECK (Fig 4) was used to assess the matrix protein's tertiary structure, which revealed that 92.4 percent of the total residues (342) were found in the core [A,B,L]; 6.3 percent of residues were in the additional allowed regions [a,b,l,p];

and 0.7 percent of residues were in the generously allowed regions [a,b,l,p]. The total number of non-glycine and non-proline residues was 301; the end-residues (excluding Gly and Pro) were 1; the glycine and proline residues were 27 and 13, respectively, out of 473 total residues (Table 7). Verify 3D; a tertiary structure evaluation tool was used to demonstrate that the anticipated tertiary structure passed the evaluation.

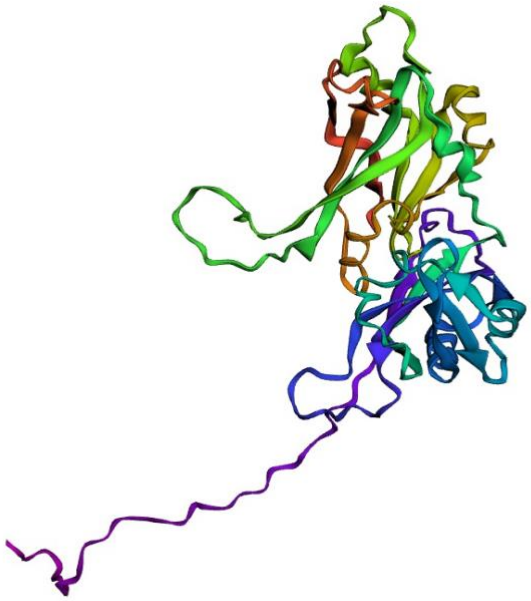

Figure 3: Predicted tertiary structure by HHpred tool employing the Modeller application.

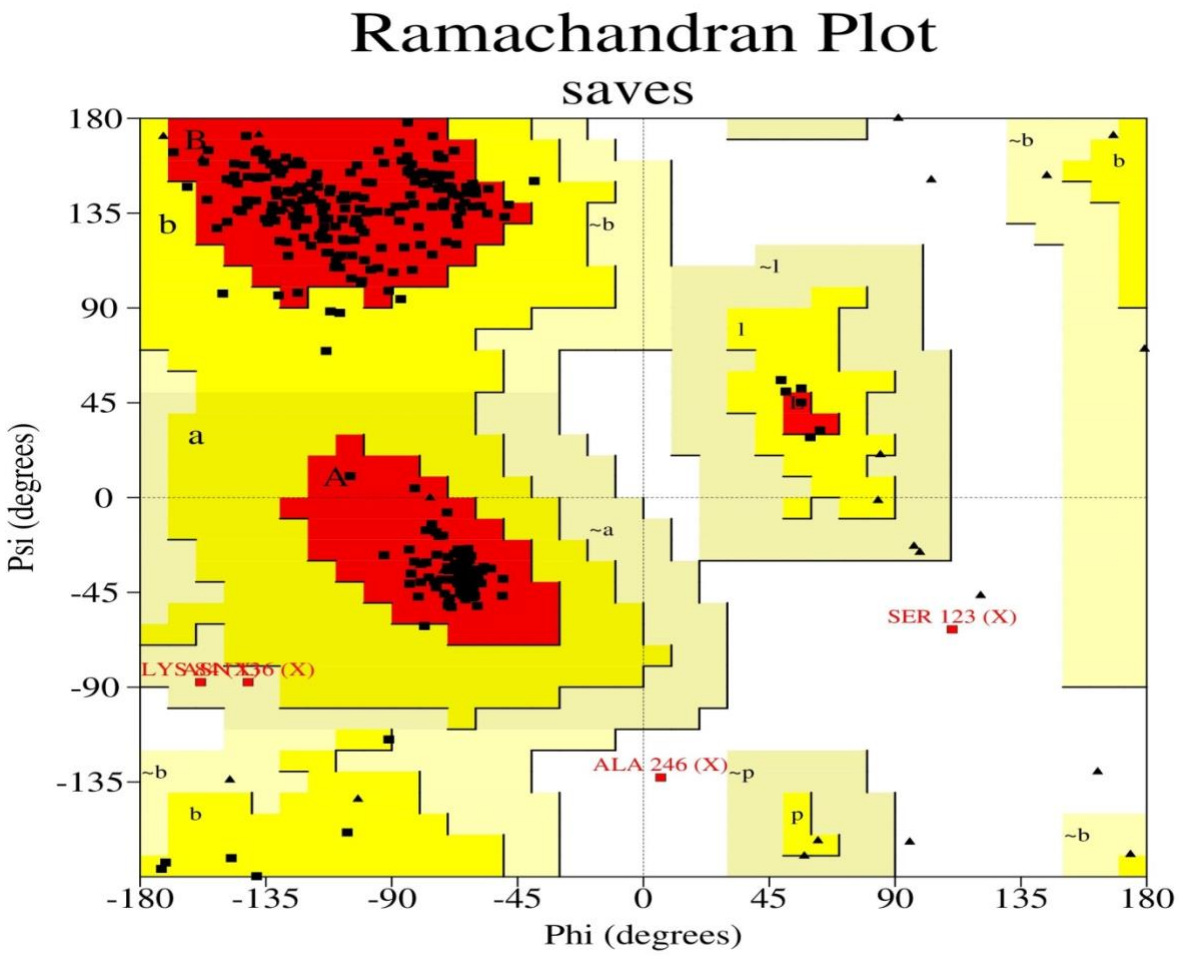

Figure 4 : The Ramachandran plot statistics of Modeller-predicted three-dimensional protein structure validated by the PROCHECK program.

Table 4: Ramachandran plot statistics of the modeled protein.

Ramachandran plot statistics	Value (%)
Residues in the most favored regions [*A, B, L*]	278 (92.4)
Residues in additional allowed regions [*a, b, l, p*]	19(6.3)
Residues in generously allowed regions [*~ a, ~ b, ~ l, ~ p*]	2 (0.7)
Residues in disallowed regions	2 (0.7)
Number of nonglycine and nonproline residues	301
Number of end residues (excl. Gly and	1

Pro)	
Number of glycine residues (shown as triangles)	27
Number of proline residues	13
Total number of residues	342

4. CONCLUSION

Understanding how proteins function is vital for describing how they work, and this protein is critical for virus assembly. With the virus core, matrix protein binds to the viral envelope. This research reveals the protein's fundamental features, such as its hydrophilic nature and functional annotation, in relation to its tertiary structure. As a result, the outcomes of this study demonstrate the efficacy and scope of future research on the matrix protein using the bioinformatics methodologies used in this investigation. This research will strengthen and sharpen our understanding of pathophysiology, allowing for the development of drugs and vaccines to combat Nipah virus infection.

ACKNOWLEEMENT
None.

CONFLICT OF INTEREST
The authors declared that there is no conflict of interest.

FUNDING
The authors received no external funding.

Reference

[1] Aditi and M. Shariff, “Nipah virus infection: A review,” *Epidemiol. Infect.*, vol. 147, pp. 1–6, 2019, doi: 10.1017/S0950268819000086.
[2] “Nipah virus infection,” *World Health Organization (WHO).*, 2020. https://www.who.int/health-topics/nipah-virus-infection#tab=tab_1 (accessed Oct. 14, 2020).
[3] B. A. Clayton, L. F. Wang, and G. A. Marsh, “Henipaviruses: An Updated

Review Focusing on the Pteropid Reservoir and Features of Transmission," *Zoonoses Public Health*, vol. 60, no. 1, pp. 69–83, 2013, doi: 10.1111/j.1863-2378.2012.01501.x.

[4] M. Amaya and C. C. Broder, "Vaccines to Emerging Viruses: Nipah and Hendra," *Annu. Rev. Virol.*, vol. 7, pp. 447–473, 2020, doi: 10.1146/annurev-virology-021920-113833.

[5] B. T. Eaton, C. C. Broder, D. Middleton, and L. F. Wang, "Hendra and Nipah viruses: Different and dangerous," *Nat. Rev. Microbiol.*, vol. 4, no. 1, pp. 23–35, 2006, doi: 10.1038/nrmicro1323.

[6] M. S. Chadha *et al.*, "Nipah virus-associated encephalitis outbreak, Siliguri, India," *Emerg. Infect. Dis.*, vol. 12, no. 2, pp. 235–240, 2006, doi: 10.3201/eid1202.051247.

[7] H. Times, "Nipah virus outbreak: Death toll rises to 14 in Kerala, two more cases identified," *The Hindustan Times*, pp. 1–2, 2018.

[8] C. L. Mills, P. J. Beuning, and M. J. Ondrechen, "Biochemical functional predictions for protein structures of unknown or uncertain function," *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 182–191, Jan. 2015, doi: 10.1016/J.CSBJ.2015.02.003.

[9] W. F. De Azevedo *et al.*, "Docking and small angle X-ray scattering studies of purine nucleoside phosphorylase," *Biochem. Biophys. Res. Commun.*, vol. 309, no. 4, pp. 923–928, Oct. 2003, doi: 10.1016/J.BBRC.2003.08.093.

[10] W. F. De Azevedo, F. Canduri, V. Fadel, L. G. V. L. Teodoro, V. Hial, and R. A. S. Gomes, "Molecular Model for the Binary Complex of Uropepsin and Pepstatin," *Biochem. Biophys. Res. Commun.*, vol. 287, no. 1, pp. 277–281, Sep. 2001, doi: 10.1006/BBRC.2001.5555.

[11] M. R. Wilkins *et al.*, "Protein identification and analysis tools in the ExPASy server.," *Methods Mol. Biol.*, vol. 112, pp. 531–552, 1999, doi: 10.1385/1-59259-584-7:531.

[12] Martin; L.; Garrity; D.M.; Yao; T., "Genomics and Transcriptomics of the Molting Gland (Y-Organ) in the Blackback Land Crab, Gecarcinus Lateralis;," Colorado State University, 2016.

[13] S. Lu *et al.*, "CDD/SPARCLE: The conserved domain database in 2020," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D265–D268, 2020, doi: 10.1093/nar/gkz991.

[14] R. D. Finn *et al.*, "Pfam: The protein families database," *Nucleic Acids Res.*, vol. 42, no. D1, pp. 222–230, 2014, doi: 10.1093/nar/gkt1223.

[15] D. Wilson, M. Madera, C. Vogel, C. Chothia, and J. Gough, "The SUPERFAMILY database in 2007: Families and functions," *Nucleic Acids Res.*, vol. 35, no. SUPPL. 1, pp. 308–313, 2007, doi: 10.1093/nar/gkl910.

[16] C. Combet, C. Blanchet, C. Geourjon, and G. Deléage, "NPS@: Network

protein sequence analysis," *Trends Biochem. Sci.*, vol. 25, no. 3, pp. 147–150, 2000, doi: 10.1016/S0968-0004(99)01540-6.
[17] D. W. A. Buchan and D. T. Jones, "The PSIPRED Protein Analysis Workbench: 20 years on," *Nucleic Acids Res.*, vol. 47, no. W1, pp. W402–W407, 2019, doi: 10.1093/nar/gkz297.
[18] B. Webb and A. Sali, "Comparative protein structure modeling using MODELLER," *Curr. Protoc. Bioinforma.*, vol. 2016, no. June, pp. 5.6.1-5.6.37, 2016, doi: 10.1002/cpbi.3.
[19] L. Zimmermann *et al.*, "A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core," *J. Mol. Biol.*, vol. 430, no. 15, pp. 2237–2243, 2018, doi: 10.1016/j.jmb.2017.12.007.
[20] and J. M. T. R. A. Laskowski, M. W. Mac Arthur, "International tables for crystallography," p. 284, 1983.
[21] K. Guruprasad, B. V. B. Reddy, and M. W. Pandit, "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.*, vol. 4, no. 2, pp. 155–161, 1990, doi: 10.1093/protein/4.2.155.
[22] X. Xia, "Protein isoelectric point," *Bioinforma. Cell*, pp. 207–219, 2007, doi: 10.1007/978-0-387-71337-3_10.
[23] S. C. Gill and P. H. von Hippel, "Calculation of protein extinction coefficients from amino acid sequence data [published erratum appears in Anal Biochem 1990 Sep;189(2):283]," *Anal Biochem*, vol. 182, no. 2, pp. 319–326, 1989.
[24] I. Atsushi, "Thermostability and aliphatic index of globular proteins," *J. Biochem.*, vol. 88, no. 6, pp. 1895–1898, 1980.
[25] A. J. Battisti *et al.*, "Structure and assembly of a paramyxovirus matrix protein," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, no. 35, pp. 13996–14000, 2012, doi: 10.1073/pnas.1210275109.
[26] G. Bergtrom, "Protein Domains, Motifs, and Folds in Protein Structure.," 2021. https://bio.libretexts.org/@go/page/16427 (accessed Aug. 20, 2021).
[27] E. Lindahl and A. Elofsson, "Identification of related proteins on family, superfamily and fold level," *J. Mol. Biol.*, vol. 295, no. 3, pp. 613–625, Jan. 2000, doi: 10.1006/JMBI.1999.3377.
[28] H. B. Uchôa, G. E. Jorge, N. J. Freitas Da Silveira, J. C. Camera, F. Canduri, and W. F. De Azevedo, "Parmodel: A web server for automated comparative modeling of proteins," *Biochem. Biophys. Res. Commun.*, vol. 325, no. 4, pp. 1481–1486, 2004, doi: 10.1016/j.bbrc.2004.10.192.