ScienSage

Journal of Advanced Scientific Research

Available online through http://www.sciensage.info

ISSN
0976-9595
Research Article

IN SILICO STRUCTURAL, FUNCTIONAL AND PHYLOGENETIC ANALYSIS OF PNEUMONIA VIRUS SURFACE GLYCOPROTEIN FROM A SEAFOOD MARKET OF CHINA

Sisir Ghosh

Department of Botany, Sreegopal Banerjee College, Bagati, Mogra, Hooghly, West Bengal, India

**Corresponding author: sisirghosh1981@gmail.com*

ABSTRACT

Severe acute respiratory syndrome (SARS) caused by SARS related coronaviruses (SARSr-CoVs) has a great diversity in forming the zoonotic epidemic diseases in different countries of the World. Some viruses from sea food market of China causing pneumonia are associated with the 2019 novel coronavirus in rural areas of the different parts of the country. The pathogens have the capacity to transmit from person to person. The surface glycoprotein of the viruses plays a key role for zoonotic infections among the hosts. The glycoprotein helps the virus for adsorption to the host cells and helps disease development. *In silico* molecular characteristics and proteomic structure of the said surface glycoprotein from pneumonia virus was studied in the present research. It was found that the surface glycoprotein of the pneumonia virus was about 141 kDa stable leucine rich acidic proteins. Phylogenetic analysis of the protein indicated the similarity with spike glycoprotein of bat coronavirus. Attempts were made to describe the physicochemical structure and function of the surface glycoprotein of the virus by the computational analysis of the data collected from different software tools.

Keywords: SARSr-CoVs, Coronavirus, *In silico*, Protein model, Surface glycoprotein

1. INTRODUCTION

Emergence of zoonotic infectious diseases throughout the World is a very serious threat to human population now-a-days. The nature of these specific infections is thought to be the association and interactions of the human beings with various wild animals [1]. Many wild animals like bat (mammals under Chiroptera order) are act as vectors of the infectious diseases caused by viruses. The outbreak of Severe Acute Respiratory Syndrome (SARS) caused by coronavirus (SARS-CoV) in the very beginning of the twenty first century affected many countries of the World infected millions of people and as a result caused death of thousands of infected patients depicted by World Health Organization. The endemic Middle East Respiratory Syndrome coronavirus (MERS-CoV) was also spilled over from bats and caused infected and death of a number of people throughout the World during 2019 [2, 3]. The increasing number of bat coronaviruses [4, 5] and its distribution in rural regions of China [6] was discovered earlier. SARS-related coronaviruses (SARSr-CoVs) from bat has a great diversity and very low species specificity according to their phylogenetic and pathogenesis studies [7- 9]. Serological evidence of the bat-borne SARSr-CoVs transmission to the human beings was also studied [1]. But no *in silico* report of characterization about the surface glycoprotein of such viruses from any host have been reported earlier.

Coronaviruses (and toroviruses) are classified together on the basis of the crown or halo-like appearance of the envelope glycoproteins, and on characteristic features of chemistry and replication. Most human coronaviruses fall into one of two serotypes: OC43-like and 229E-like [10].

Coronaviruses comprise the subfamily Orthocorona - virinae in the family *Coronaviridae*, in the order *Nidovirales* [11].

The present study emphasises the *in silico* analysis of molecular proteomic structure of the surface glycoprotein of a pneumonia virus from seafood market of China [12]. This pneumonia virus associated with the 2019 novel coronavirus has the capacity of transmission from person to person and casing diseases [12]. The *in silico* research of this surface glycoprotein will definitely help the future researchers to analyze the molecular proteomic structure development of the protein in dry as well as wet laboratory experiments. Further this *in silico* study might be helpful on the way to vaccine preparation to control the zoonotic epidemic diseases caused by SARSr-CoVs.

2. MATERIAL AND METHODS

2.1. Retrieval of the amino acid sequence

Amino acid sequence (1273 aa) of surface glycoprotein of Wuhan seafood market pneumonia virus [12] was selected for the study and was retrieved from National Centre for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/) database in Fast Alignment (FASTA) format. The sequence (QHN73810.1) was selected for reference sequence and was used for investigation including the model preparation.

2.2. Analysis of the physicochemical characteristics

The physicochemical characteristics of the protein after analysing the primary sequence of surface glycoprotein of Wuhan seafood market pneumonia virus was performed using the ExPASyProtParam tool [13]. With the help of this tool the molecular weight of the protein and the amino acid composition were determined.

2.3. Prediction of secondary structure

The secondary structure (i.e. the number of helices, sheets, turns and coils in the amino acid sequence) of the protein was predicted using CFSSP: Chou and Fasman secondary structure prediction server (www.biogem.org/tool/choufasman) [14, 15].

2.4. Determination of homology protein model

To determine the homology protein model of surface glycoprotein of Wuhan seafood market pneumonia virus SWISS-MODEL (ProMod3 version 1.0.2) workspace [16] was used. SAVES server (http://services.mbi.ucla.edu/SAVES/) was used for evaluation of build model. Pdb file for model preparation was used for constructing the Ramachandran Plot in RAMPAGE (http://morded.bioc.cam.ac.uk/rapper/rampage.php)

2.5. Identification of the functional motifs

To identify the functional motifs in Wuhan seafood market pneumonia virus surface glycoprotein sequence the motif search tool (http://www.genome.jp/tools/motif/) was used.

2.6. Construction of the Phylogenetic tree

A Phylogenetic tree was constructed using tool from NCBI (http://www.ncbi.nlm.nih.gov/) after selecting the similar protein sequences by Basic local alignment search tool (BLAST) search using NCBI-BLAST to find the evolutionary relationships among selected proteins of the same organism.

3. RESULTS

3.1. Retrieval of amino acid sequence

The amino acid sequence of surface glycoprotein of Wuhan seafood market pneumonia virus [12] was retrieved from NCBI (Fig. 1). The sequence selected for the study contained 1273 continuous stretches of amino acids and download it in the FASTA format. The accession number of the sequence was QHN73810.1 (Fig. 1).

3.2. Selection of the reference sequence

The reference sequences were only selected for downloading the similar sequences of surface glycoprotein of the virus using BLAST search. The selected sequences were downloaded in the FASTA format for further study.

3.3. Construction of Phylogenetic tree from the similar sequences

Phylogenetic tree (Fig. 2) from the similar amino acid sequences was constructed using NCBI software tool. It was revealed that the surface glycoprotein isolated from the pneumonia virus was showed closed relationship with spike glycoprotein of bat coronavirus (Fig. 2).

3.4. Physicochemical characteristics

The Physicochemical characteristics of the surface glycoprotein was analysed with the help of ExPASyProtParam tool. The characteristics determined from the tool were listed in a tabulated manner (Table 1).

3.5. Prediction of secondary structure

Secondary structure was predicted with the help of Chou and Fasman Secondary Structure Prediction Server. It was found that the surface glycoprotein has the secondary elements of alpha helix 56.9%, sheets 56.2% and turns 11.1% (Fig. 3).

3.6. Model preparation and evaluation

Tertiary homology model was built after target-template alignment using the best match template (4iv9.1.A) taken from SWISS server (Fig. 4-6). The data taken from the Ramachandran plot revealed that number of residues in favoured region was more than 80%, in allowed region was 2.0% and in outlier region

was 0.2% (Fig. 7). From the SAVES server it was found that the overall quality factor of the pdb model for both the chain was above 80.00% (Fig. 7). It was also found that about 66.49% of the residues have the average 3D-1D score >= 0.2 or at least 80% of the amino acids have scored >= 0.2 in 3D/1D profile (i.e. Pass) (Fig. 7).

3.7. Motif functional analysis

There were 09 different protein functional motifs found from the motif search (Fig 8). Among these Corona_S2 (PF01601, Coronavirus S2 glycoprotein) at position 686.1270 with Independent E-value 1.4e-266 was maximum.

Fig. 1: Source of the sequence of the surface glycoprotein

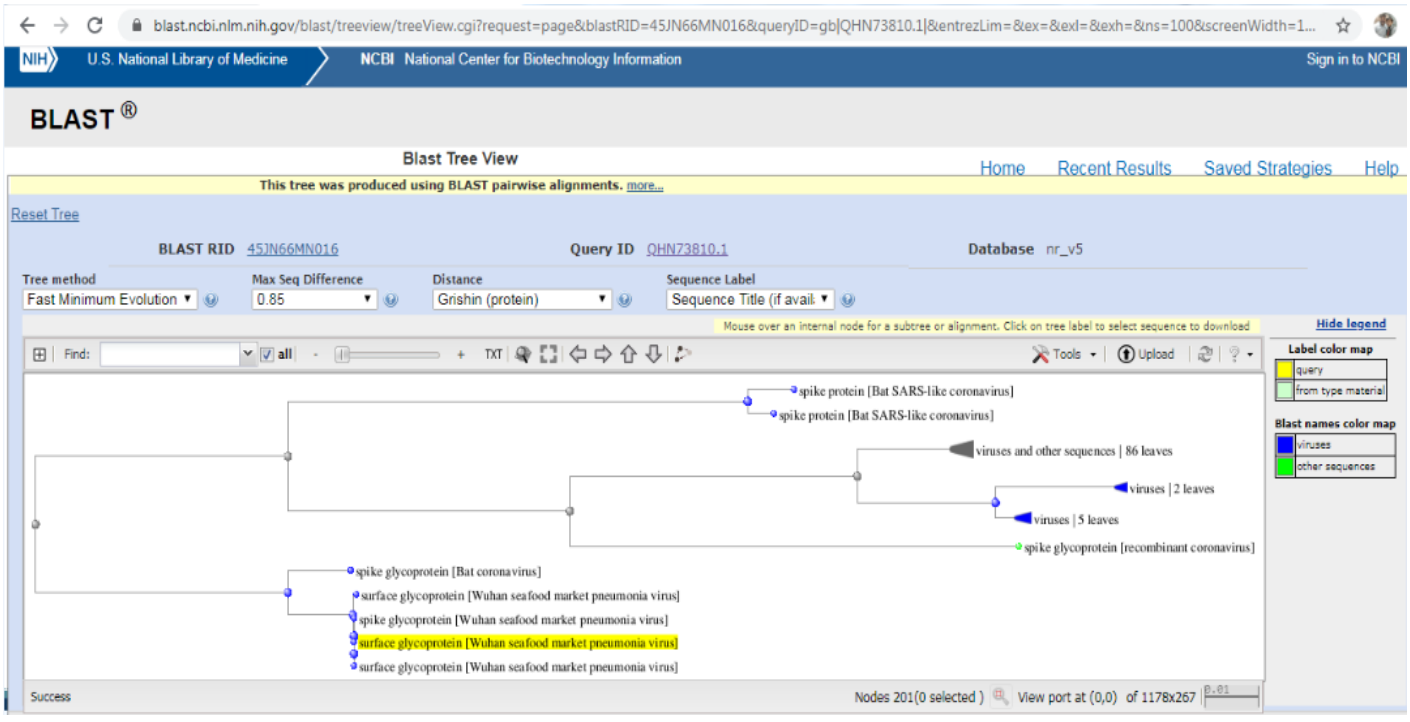
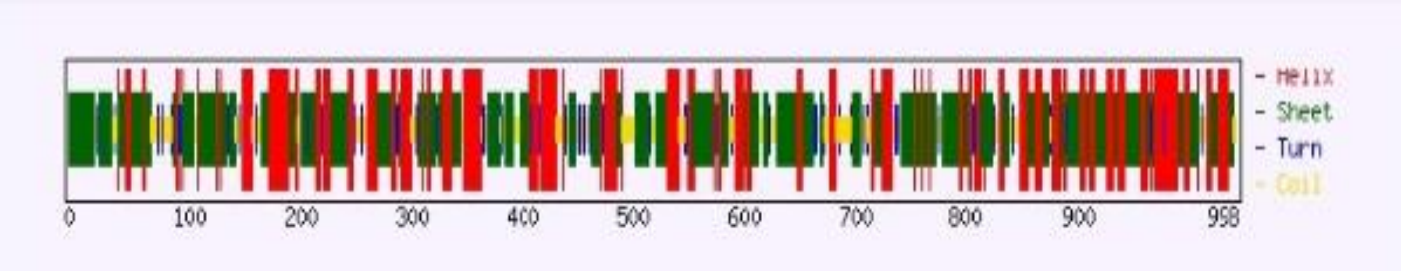


Fig. 2: Phylogenetic tree prepared using NCBI software tool

Table 1: Physicochemical characteristics of the surface glycoprotein

Parameters	Results
Total number of amino acids	1273
Molecular weight	141178.47
Theoretical pI	6.24
Total number of negatively charged residues	110
Total number of positively charged residues	103
Formula	$C_{6336}H_{9770}N_{1656}O_{1894}S_{54}$
Total number of atoms	19710
Extinction coefficient	148960
Instability index	33.01
Aliphatic index	84.67
Grand average of hydropathicity (GRAVY)	-0.079



Total Residues: H: 568 E: 561 T: 111
Percent: H: 56.9 E: 56.2 T: 11.1

Fig. 3: Secondary structure of the surface glycoprotein

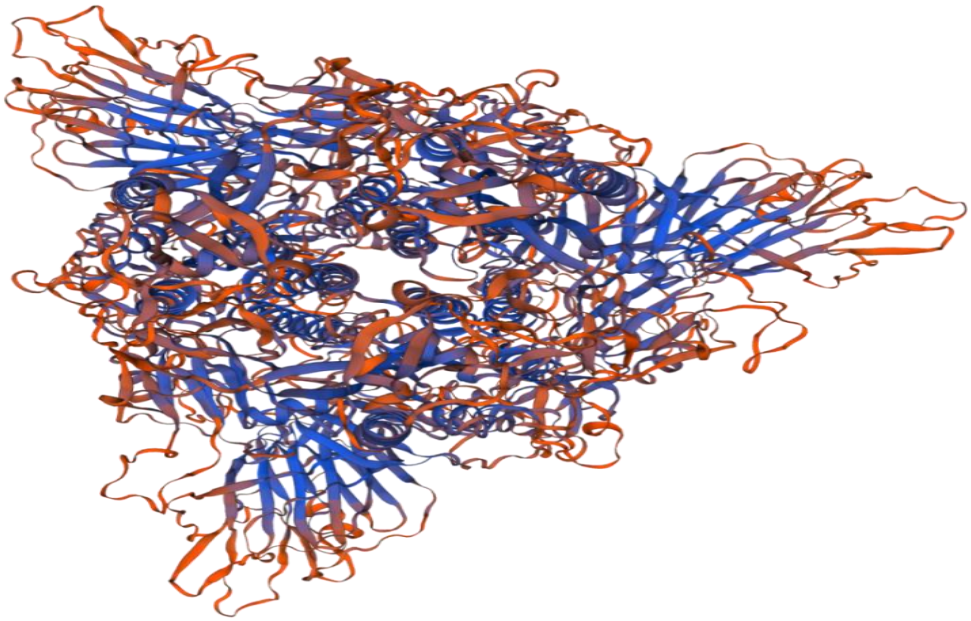


Fig. 4: Tertiary protein model of the selected protein

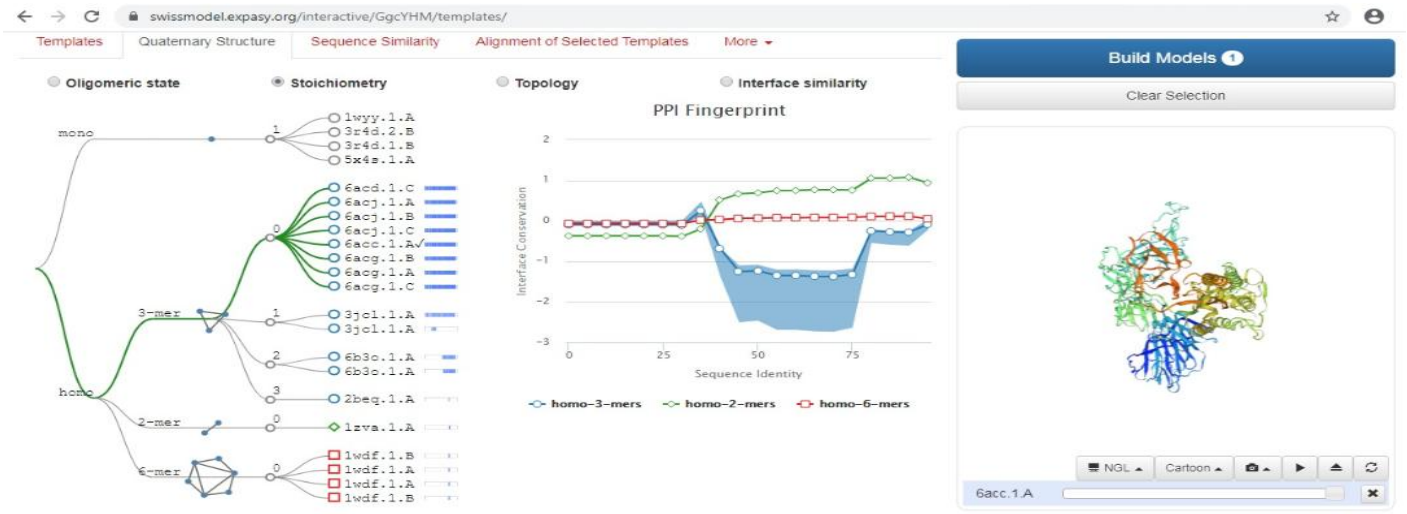


Fig. 5: Structure preparation from SWISS Model

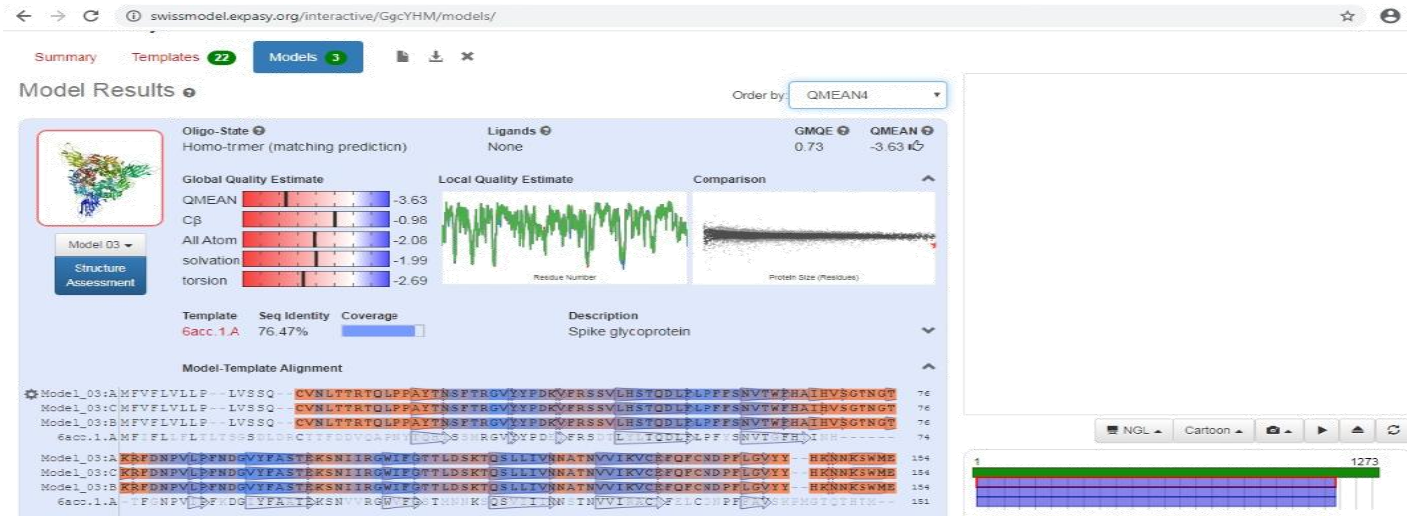


Fig. 6: Evaluation of protein model from QMEAN and SAVES server

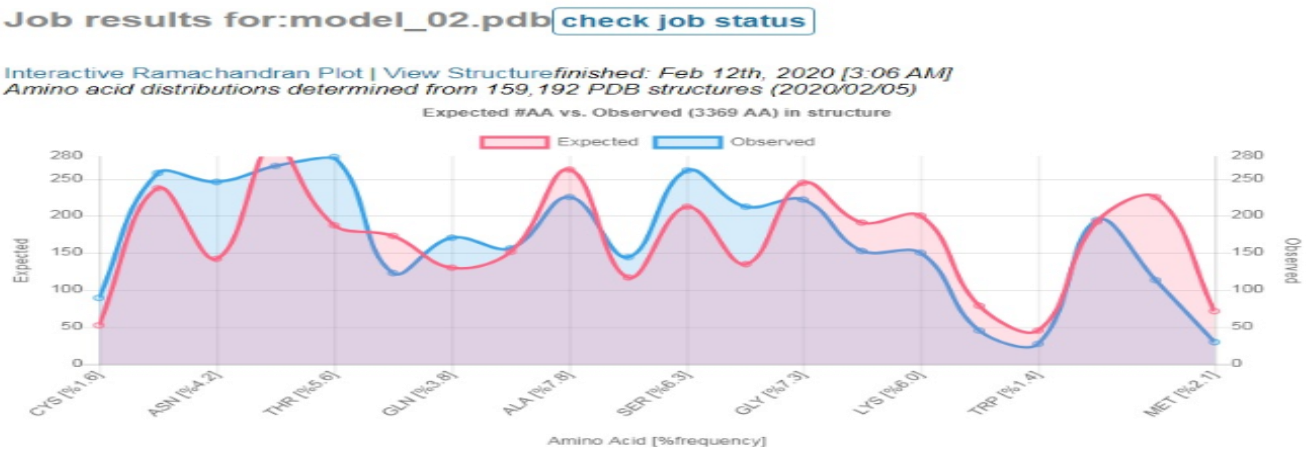
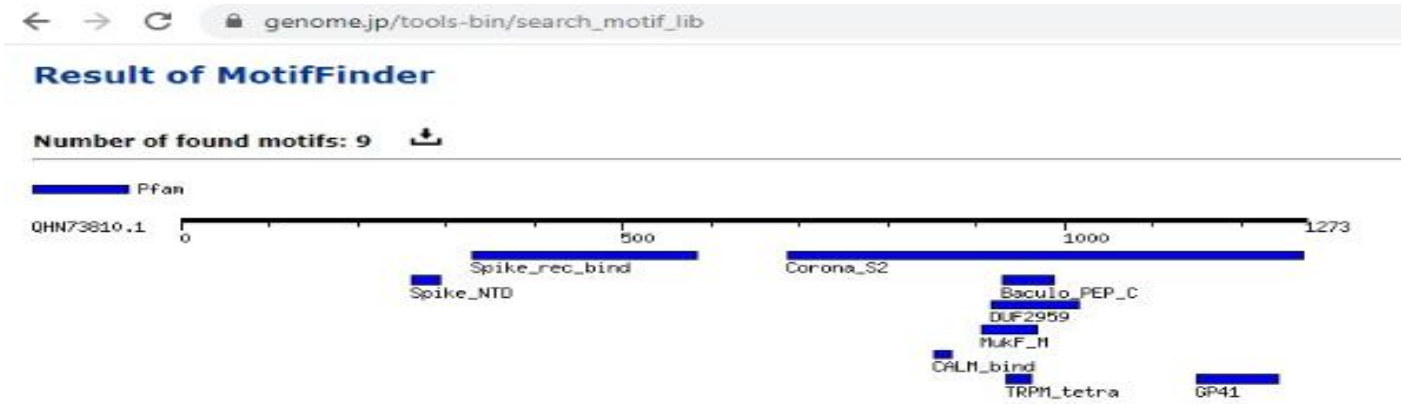


Fig. 7: Evaluation of overall quality factor from SAVES server



Pfam (9 motifs)

Pfam	Position(Independent E-value)		Description
Corona_S2	686..1270(1.4e-266)	Detail	PF01601, Coronavirus S2 glycoprotein
Spike_rec_bind	330..583(6.6e-75)	Detail	PF09408, Spike receptor binding domain
Spike_NTD	262..294(0.0014)	Detail	PF16451, Spike glycoprotein N-terminal domain
Baculo_PEP_C	931..989(0.11)	Detail	PF04513, Baculovirus polyhedron envelope protein, PEP, C terminus
DUF2959	917..1015(0.057)	Detail	PF11172, Protein of unknown function (DUF2959)
MukF_M	906..969(0.22)	Detail	PF17192, MukF middle domain
CALM_bind	853..871(0.26)	Detail	PF16025, Calcium-dependent calmodulin binding
TRPM_tetra	934..962(0.21)	Detail	PF16519, Tetramerisation domain of TRPM
GP41	1149..1241(0.26)	Detail	PF00517, Retroviral envelope protein

Fig. 8: Motif search result showing functional motif for the selected protein

Number of amino acids: 1273

Molecular weight: 141178.47

Theoretical pI: 6.24

Amino acid composition: CSV format

Ala (A)	79	6.2%
Arg (R)	42	3.3%
Asn (N)	88	6.9%
Asp (D)	62	4.9%
Cys (C)	40	3.1%
Gln (Q)	62	4.9%
Glu (E)	48	3.8%
Gly (G)	82	6.4%
His (H)	17	1.3%
Ile (I)	76	6.0%
Leu (L)	108	8.5%
Lys (K)	61	4.8%
Met (M)	14	1.1%
Phe (F)	77	6.0%
Pro (P)	58	4.6%
Ser (S)	99	7.8%
Thr (T)	97	7.6%
Trp (W)	12	0.9%
Tyr (Y)	54	4.2%
Val (V)	97	7.6%
Pyl (O)	0	0.0%
Sec (U)	0	0.0%
(B)	0	0.0%
(Z)	0	0.0%
(X)	0	0.0%

Fig. 9: Primary parameters (a) of the selected protein using ExPasyProtParam tool

Atomic composition:

Carbon	C	6336
Hydrogen	H	9770
Nitrogen	N	1656
Oxygen	O	1894
Sulfur	S	54

Formula: $C_{6336}H_{9770}N_{1656}O_{1894}S_{54}$
Total number of atoms: 19710

Extinction coefficients:

Extinction coefficients are in units of M^{-1} cm^{-1}, at 280 nm measured in water.

Ext. coefficient 148960
Abs 0.1% (=1 g/l) 1.055, assuming all pairs of Cys residues form cystines

Ext. coefficient 146460
Abs 0.1% (=1 g/l) 1.037, assuming all Cys residues are reduced

Estimated half-life:

The N-terminal of the sequence considered is M (Met).

The estimated half-life is: 30 hours (mammalian reticulocytes, in vitro).
>20 hours (yeast, in vivo).
>10 hours (Escherichia coli, in vivo).

Instability index:

The instability index (II) is computed to be 33.01
This classifies the protein as stable.

Aliphatic index: 84.67

Grand average of hydropathicity (GRAVY): -0.079

Fig. 10: Primary parameters (b) of the selected protein using ExPasyProtParam tool

4. DISCUSSION

The surface glycoprotein of pneumonia virus associated with 2019 novel coronavirus from a seafood market of China [12] is very important in its pathogenecity and transmission from person to person. There were a few *in silico* report of characterization of different surface glycoprotein of other viruses. Again, in silico physicochemical characterization of the surface glycoprotein of pneumonia virus related with SARSr-CoVs was yet not been studied so far, that's why the present work has been selected for the study.

The amino acid sequence (QHN73810.1) of the surface glycoprotein of pneumonia virus (Fig. 1) was retrieved in the FASTA format from NCBI. The sequence studied for the further experiment was 1273 amino acids in length. A number of similar protein sequences of different strains of the same organism were also retrieved using NCBI-BLAST to study the phylogenetic relationship (Fig. 2) among the different sequences of the same protein. It was found that the surface glycoprotein isolated from other viruses showed the closed relationship with the spike glycoprotein of bat coronavirus. Similar type of phylogenetic assessment of different other proteins were performed by other authors [17-19].

ExPASyProtParam tool was used for determining the overall theoretical nature of the protein (Fig. 9, 10). To analyse the physicochemical characteristics (Table 1) of the protein, it was found that the protein contained maximum amount of Leucine (8.5%) and minimum amount of Tryptophan (0.9%) with other amino acids (Fig. 9). The molecular weight of the protein was 141178.47 with the isoelectric point 6.24 (Fig. 9). As the isoelectric point of the protein was below the neutral pH, hence this protein was said to be acidic. The instability index of the protein was 33.01 (Fig. 10), which was below 40 indicated the stability of the protein. Thermostability of the protein was confirmed with the higher range (84.67) of aliphatic index like other researcher [20]. Thermostable proteins were also reported by many authors [19, 21]. The extinction coefficient was calculated as 148960 $M^{-1}cm^{-1}$ in water at 280 nm assuming all pairs of Cys residues from cystines. The molecular weight of the protein was near about 141 kDa. The GRAVY as calculated (Fig. 10) from EXPASy tool was very low (-0.079), indicated better interaction of the protein with water molecules [17, 19].

From the computational investigation secondary structure (Fig 3) was predicted. From the Chou and Fasman secondary structure prediction server it was found that the content of alpha helices was very high, which indicated the thermostable nature of the protein [17, 19, 22].

Ramachandran plot (Fig. 7) revealed that number of residues in favoured region was more than 80%. More than 80% of the residues in favoured region indicated to have a good quality model [23]. The overall quality factor in the pdb model from the SAVES server was more than 81% (Fig. 7), which indicated a good quality with high resolution model [24].

Among the 09 different motifs found from the motif search, corona_S2 (PF01601, Corona S2 glycoprotein) was present at the top position (Fig. 8).

From the present study it may be concluded that detailed in silico investigation of the surface glycoprotein of pneumonia virus from sea food market is a leucine rich, acidic, with high aliphatic index, thermostable protein having the molecular weight of about 141 kDa and which is very similar with bat coronavirus.

This in silico research will help the future researchers in dry as well as wet laboratory on the way to further research and may help in vaccine preparation to control the epidemic disease caused by the SARSr-CoVs.

5. ACKNOWLEDGEMENT

The author is grateful to the Head, Department of Botany and Principal of Sreegopal Banerjee College for giving the laboratory facility.

6. REFERENCES

1. Li H, Mendelsohn E, Zong C, Zhang W, Hagan E, Wang N, et al. *Biosafety and Health*, 2019; **1**:84-90.
2. World Health Organization. Middle East Respiratory Syndrome Coronavirus (MERSCoV), 2019; https://www.who.int/ emergencies/mers-cov/en/
3. Anthony SJ, Gilardi K, Menachery VD, Goldstein T, Ssebide B, Mbabazi R, et al. *mBio*, 2017; **8**.
4. Han H-J, Wen H-l, Zhou C-M, Chen F-F, Luo L-M, Liu J-w, Yu X-J. *Virus Res.*, 2015; **205**:1-6.
5. Hu B, Zeng L-P, Yang X-L, Ge X-Y, Zhang W, Li B, et al. *PLoS Pathog.*, 2017; **13**: e1006698.
6. Luo J, Jiang T, Lu G, Wang L, Wang J, Feng J. *Oryx*, 2013; **47**:526-531.
7. Wang LF, Anderson DE. *Curr. Opin. Virol.*, 2019; **34**:79-89.
8. Cui J, Li F, Shi ZL. *Nat. Rev. Microbiol.*, 2019; **17**:181-192.
9. Fan Y, Zhao K, Shi ZL, Zhou P. *Viruses*, 2019; **11**.
10. Tyrrell DAJ, Myint SH. *Medical Microbiology*. 4th edition. Editor: Samuel Baron, Galveston (TX): University of Texas Medical Branch at Galveston, 1996; ISBN-10: 0-9631172-1-1.
11. de Groot RJ, Baker SC, Baric R, Enjuanes L, Gorbalenya AE, Holmes KV, et al. International Committee on Taxonomy of Viruses, International Union of Microbiological Societies. Virology Division (eds.). *Ninth Report of the International Committee on Taxonomy of Viruses*. Oxford: Elsevier, 2011; pp. 806–828. ISBN 978-0-12-384684-6.
12. Chan JF-W, Yuan S, Kok KH, To KK-W, Chu H, Yang J, et al. *The Lancet*, 2020; **395(10223)**: 514-523.
13. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A. Protein identification and analysis tools on the ExPASy server. In: Walker, John M (Ed) The Proteomics Protocols Handbook. Humana Press, 2005; pp. 571-607
14. Chou PY, Fasman GD. *Biochemistry*, 1974; **13**:211-222.
15. Chou PY, Fasman GD. *Biochemistry*, 1974; **13**: 222-245.
16. Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, et al. *Nucl. Acids Res.*, 2014; **42**:W252-W258.
17. Verma A, Singh VK, Gaura S. *Comp. Biol. Chem.*, 2016; **60**:53-58.
18. Pramanik K, Ghosh PK, Ray S, Sarkar A, Mitra S, Maiti TK. *J. Gen. Eng. Biotech.*, 2017; **15**:527-537.
19. Pramanik K, Kundu S, Banerjee S, Ghosh PK, Maiti TK. *3 Biotech*, 2018; **8**:262.
20. Niño-Gómez DC, Rivera-Hoyos CM, Morales-Álvarez ED, Reyes-Montaño EA, Vargas-Alejo NE et al. *Enzyme Res.*, 2017; **9746191:**1-23.
21. Kalsi HK, Singh R, Dhaliwal HS, Kumar V. *3 Biotech*, 2016: 6 (1).
22. Kumar S, Tsai C-J, Nussinov R. *Protein Eng.*, 1999; **13**:179-191.
23. Yadav PK, Singh G, Gautam B, Singh S, Yadav M et al. *Bioinformation,* 2013; **9(3)**:158-164.
24. Benkert P, Kunzli M, Schwede T. *Nucleic Acids Res.*, 2009; **37**:510-514.