Characterization and Structural Prediction of Proteins in SARS-CoV-2 Bangladeshi Variant Through Bioinformatics

Microbiology Insights
Volume 15: 1–14
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/11786361221115595
SAGE

Pinky Debnath[1], Umama Khan[2] and Md. Salauddin Khan[3]

[1]Chemical Biotechnology Department, Technical University of Munich, Straubing, Germany. [2]Biotechnology and Genetic Engineering Discipline, Khulna University, Bangladesh. [3]Statistics Discipline, Khulna University, Bangladesh.

ABSTRACT: The renowned respiratory disease induced by the severe acute respiratory syndrome-coronavirus-2 (SARS-CoV-2) has become a global epidemic in just less than a year by the first half of 2020. The subsequent efficient human-to-human transmission of this virus eventually affected millions of people worldwide. The most devastating thing is that the infection rate is continuously uprising and resulting in significant mortality especially among the older age population and those with health co-morbidities. This enveloped, positive-sense RNA virus is chiefly responsible for the infection of the upper respiratory system. The virulence of the SARS-CoV-2 is mostly regulated by its proteins such as entry to the host cell through fusion mechanism, fusion of infected cells with neighboring uninfected cells to spread virus, inhibition of host gene expression, cellular differentiation, apoptosis, mitochondrial biogenesis, etc. But very little is known about the protein structures and functionalities. Therefore, the main purpose of this study is to learn more about these proteins through bioinformatics approaches. In this study, ORF10, ORF7b, ORF7a, ORF6, membrane glycoprotein, and envelope protein have been selected from a Bangladeshi Corona-virus strain G039392 and a number of bioinformatics tools (MEGA-X-V10.1.7, PONDR, ProtScale, ProtParam, SCRIBER, NetSurfP v2.0, IntFOLD, UCSF Chimera, and PyMol) and strategies were implemented for multiple sequence alignment and phylogeny analysis with 9 different variants, predicting hydropathicity, amino acid compositions, protein-binding propensity, protein disorders, and 2D and 3D protein modeling. Selected proteins were characterized as highly flexible, structurally and electrostatically extremely stable, ordered, biologically active, hydrophobic, and closely related to proteins of different variants. This detailed information regarding the characterization and structure of proteins of SARS-CoV-2 Bangladeshi variant was performed for the first time ever to unveil the deep mechanism behind the virulence features. And this robust appraisal also paves the future way for molecular docking, vaccine development targeting these characterized proteins.

KEYWORDS: SARS-CoV-2, bioinformatics, Bangladeshi covid-19 variant, ORF proteins, membrane and envelope protein, structural prediction

RECEIVED: January 21, 2022. **ACCEPTED:** June 30, 2022.

TYPE: Original Research

FUNDING: The author(s) received no financial support for the research, authorship, and/or publication of this article.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Md. Salauddin Khan, Statistics Discipline, Khulna University, Khulna 9208, Bangladesh. Email: salauddinstat@ku.ac.bd

Introduction

In December 2019, the whole world was stunned by the outbreak of unknown cause pneumonia, which was originated from Wuhan, Hubei Province of China. And then, by January 7, 2020, Chinese scientists have screened a novel Coronavirus (CoV) mainly responsible for the infection of the upper respiratory system from patients in Wuhan.[1] The ensuing proficient human-to-human transmission of the virus ultimately affected millions of people worldwide. Since December 2021, there were more than 288.7 million confirmed cases and 5.45 million people have died around the world by this devastating virus. The most devastating thing is that the infection rate is continuously uprising which is resulting in significant mortality especially among the older age population and those with health co-morbidities.

Corona viruses are very tiny in size (diameter, 65-125 nm) consist of a single-strand RNA as nucleic material.[2,3] Along with RNA, this particular virus consists of 12 different proteins such as nonstructural proteins (ORF1a and ORF1b) at the 5′ -end, structural proteins (spike surface glycoprotein [S], envelope [E], matrix [M], and nucleocapsid [N]) and multiple lineage-specific accessory proteins (ORF3a, ORF6, ORF7b, ORF8, and ORF10) at the 3′ -end.[4] Although these proteins are basically involved in host receptor recognition, attachment, and entry into host cells, very slight is recognized about these protein structures and specific functionalities. The ORF10 protein is found upstream of the 3′ -untranslated region (3′ -UTR), apparently encodes for a protein of 38 amino acids long.[5] The ORF7b protein is a presumed viral accessory protein encoded on subgenomic (sg) RNA 7,[6] whereas, ORF7a possessed a distinctive immunoglobulin (Ig)-like domain with a 15-a.a single peptide sequence at its N terminus, an 81-a.a luminal domain, a 21 a.a transmembrane domain, and a short C-terminal tail.[7] Also, the SARS-CoV ORF6 is characteristically between 42 and 63 amino acids in length and by transcribing into mRNA6 encodes SARS 6 protein.[8] Moreover, the membrane glycoprotein is found abundantly and plays the main role in virion assemble, morphogenesis, and, also, define the shape of the viral envelope.[9,10] Lastly, the envelope proteins are short-chain polypeptide with a single α-helical transmembrane domain that can produce homopentametric ion channels (IC).[11]

Though Bangladesh is not exempt from the severe outbreak of the Corona virus, a large number of Bangladeshi strains also have been identified. On 8 March 2020, in Bangladesh, SARS-CoV-2 was reported for the very first time. A new strain was acknowledged on January 2021, from a 50-year-old symptomatic male patient in Dhaka, Bangladesh (SARS-CoV-2 strain G039392) and the strain was found as 99.9% identical to the UK variant B.1.1.7.[12] In this study, 6 proteins (ORF10, ORF7b, ORF7a, ORF6, membrane glycoprotein, and envelope protein) were randomly selected from the SARS-CoV-2 Bangladeshi novel strain G039392 regarding characterization, so that more detailed studies could elucidate their structures and provide insights into the possible functions for the selected proteins.

These 6 proteins were selected due to their availability in online database portal. From 9 different countries, selected 6 proteins were chosen for analysis from 9 different variants. At the time of the analysis all required data were available for only those important virulent proteins of SARS-CoV-2 virus though there are other proteins like Spike and nuclear capture proteins which are very important for the variant definition. Therefore, the major purpose of this analysis is to learn more about these proteins like the assessment of amino acid (a.a) composition, the energy level of chemical bonds, hydropathicity, etc. through bioinformatics approaches which could provide insight into probing novel functions regarding virulence of Covid-19. Moreover, structural prediction of 2D and 3D SARS-CoV-2 protein models could give further way to docking molecular components which can optimize devastating viral properties of this particular virus. Thereby, this study utilizes strictly bioinformatics approaches to theoretically characterize, classify, and construct the putative structure of selected 6 proteins in SARS-CoV-2 Bangladeshi strain G039392.

Materials and Methods

Sequence alignment and phylogenetic analysis

The reference sequence corresponding to the selected 6 proteins (ORF10, ORF7b, ORF7a, ORF6, membrane glycoprotein, and envelope protein) in SARS-CoV-2 strain G039392, along with other 9 variants from 9 different countries, were acquired from NCBI's Protein Database. Sequences were aligned using MUSCLE on the MEGA-X-V10.1.7 software.[13,14] The neighbor-joining method was implemented by maintaining other default settings. Alignments reliability was measured by overall mean distance ($\leqslant$0.7 is reliable) and determined using p-distance substitution model.[15] The protein trees were constructed using the neighbor-joining method and visualized on MEGA-X-V10.1.7. The phylogeny trees were tested using the bootstrap method.

Protein characterization

Phosphorylation sites were detected by DEPP server of PONDR®.[16] ProtScale was used to generate hydrophobicity plot and ProtParam to determine the grand average of hydrophobicity (GRAVY).[17] Also, the protein disorder predictions were performed using PONDR® (Predictor of Natural Disordered Regions) VLS2, XL1.[18] Moreover, amino acid compositions and aliphatic index were analyzed employing ProtParam. Finally, the protein-binding propensities of the interacting residues were evaluated using SCRIBER.[19]

Protein secondary structure prediction, 3D modeling, evolution, and validation

NetSurfP v2.0 server was employed to predict the protein secondary structure.[20] The predictions of transmembrane helix (TH) were performed by TMHMM[21] and Phobius[22] by averaging the predictions and the most constant range of scores were utilized for analysis. The web server IntFOLD was used to make use of an ab initio modeling for constructing the selected proteins.[23] According to the IntFOLD's quality and confidence scoring, the models were evaluated and utilizing the 3Drefine web-server, the best model was then refined.[24] The maximum QMEAN *Z*-score[25] and Ramachandran plot[26] were considered as most favorable among the 5 generated post-refinement models. Both UCSF Chimera and PyMol were used to visualize the most favorable 3D protein model.[27,28] The hydrophobicity surfaces were created according to the Kyte-Doolittle scale.[29]

Results

Sequence alignment and phylogenetic analysis

In phylogenetic tree, the overall mean distance of ORF10 is 0.01, which is corresponding to almost 99.9% identity for the entire alignment (Figure 1G). The ORF10 protein from the strain of Spain has shown difference at the 30th position which is Leu (L) rather than Val (V) (Figure 1A). So, the height of the conserved region is from 1 to 29 residues. While, for ORF7b, ORF7a, and ORF6, membrane glycoprotein, envelope protein, the mean distance is 0.00 along with 100% conserved regions which are correspondences for the entire alignment (Figure 1).

Phosphorylation

Single phosphorylation site (phosphorylated serine) was identified in ORF7a consist of 14.29%, whereas other proteins did not show any phosphorylation sites (Table 1).

Hydropathicity

In ORF10 protein, the hydrophobicity plot exposed 2 hydrophobic regions spanning residues 3 to 20 and 28 to 35 along with 2 hydrophilic regions; residues 21 to 27 and a residue of 36 (Figure 2A).There are single hydrophilic and hydrophilic regions spanning residues 3 to 32 and 33 to 41, respectively in ORF7b protein (Figure 2B). In case of ORF7a protein,

hydrophobic regions are 3 to 16, 25, 28 to 31, 47 and 48, 54 to 61, 63 to 67, 69, 72, 84 to 88, 98 to 115 and the hydrophilic regions are 17 to 24, 26 and 27, 32 to 46, 49 to 53, 62, 68, 70 and 71, 76 to 83, 89 to 97, 116 to 119 with the neutral regions 73, 75 (Figure 2C). In ORF6 protein, the hydrophobic regions are 3 to 7, 9 to 20, 23 to 28, 31 to 39, 42, hydrophilic regions are 8, 21 and 22, 29, 40 and 41, 43 to 59 with a neutral position 30 (Figure 2D). Moreover, for membrane glycoprotein, 8 to 11, 15, 21 to 39, 46 to 71, 74 and 75, 77 to 102, 104, 110, 117 to 122, 124, 126 to 132, 134, 137 to 146, 149 to 151, 168 to 171, 182

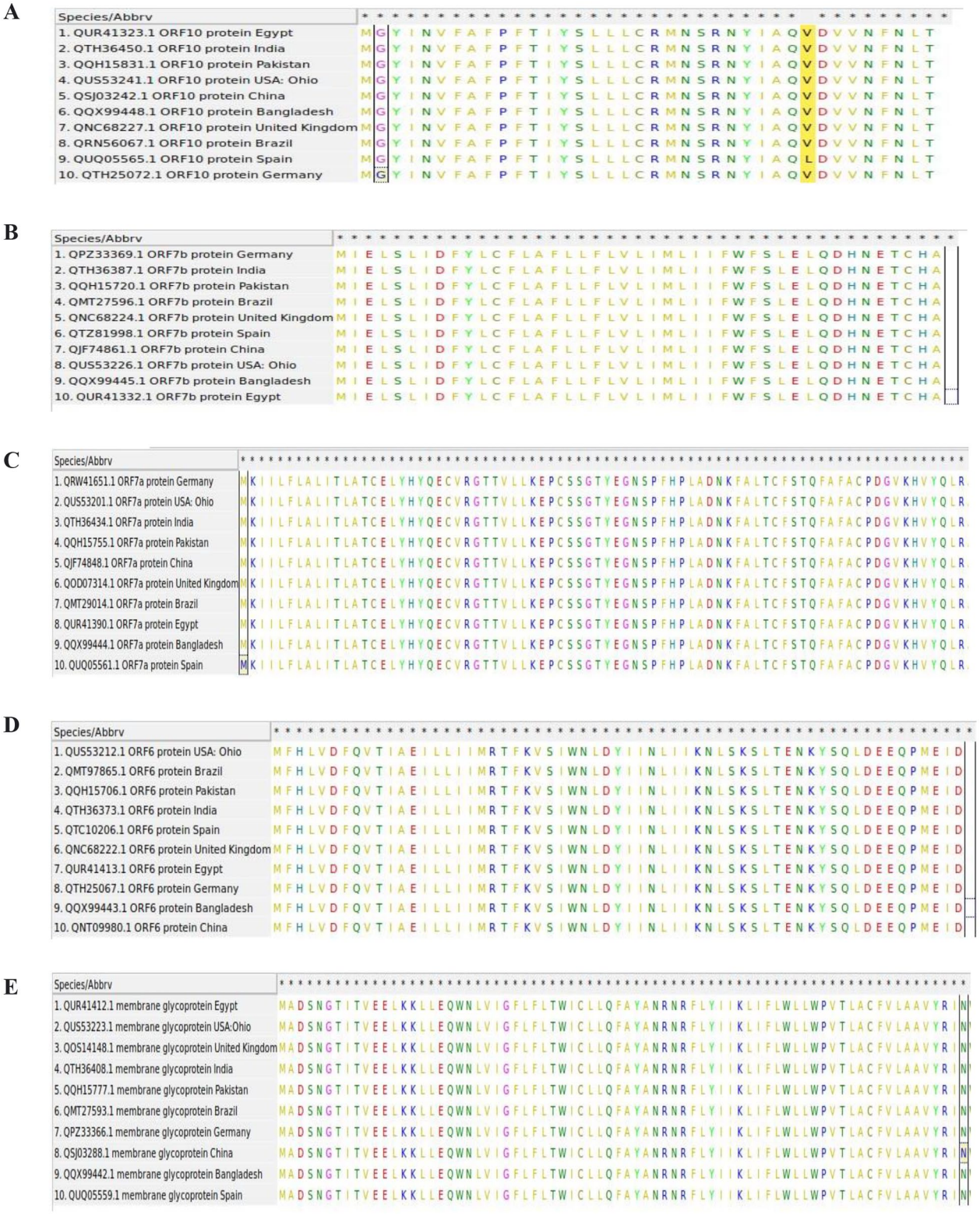

Figure 1. *(Continued)*

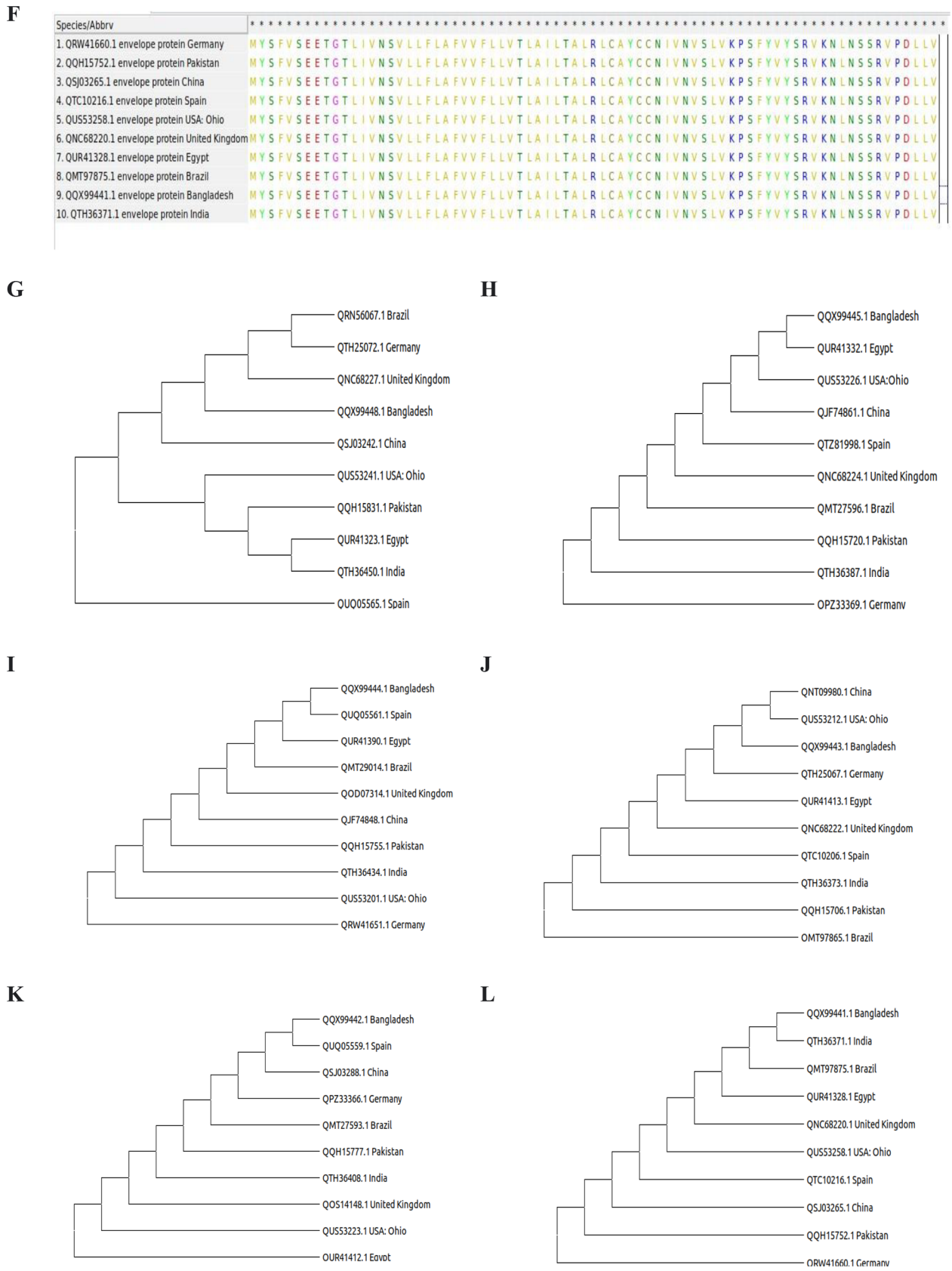

Figure 1. Multiple sequence alignment and phylogenetic analysis of 6 different protein sequences depicting evolutionary relationships with SARS-CoV-2 varieties of 10 different countries (A–F). Sequence alignment of ORF10, ORF7b, ORF7a, and ORF6, membrane glycoprotein, and envelope protein respectively. (G–L) Neighbor joining phylogenetic tree of ORF10, ORF7b, ORF7a, and ORF6, membrane glycoprotein, and envelope protein respectively. All proteins appear closely related to the proteins of different variants of different countries.

Table 1. Phosphorylation sites of the selected proteins.

PROTEIN NAME	NUMBER OF PHOSPHORYLATED SERINES	NUMBER OF PHOSPHORYLATED THREONINES	NUMBER OF PHOSPHORYLATED TYROSINES
ORF10 protein	0 out of 2 (0.00%)	0 out of 2 (0.00%)	0 out of 3 (0.00%)
ORF7b protein	0 out of 2 (0.00%)	0 out of 1 (0.00%)	0 out of 1 (0.00%)
ORF7a protein	1 out of 7 (14.29%)	0 out of 10 (0.00%)	0 out of 5 (0.00%)
ORF6 protein	0 out of 4 (0.00%)	0 out of 3 (0.00%)	0 out of 2 (0.00%)
Membrane glycoprotein	0 out of 15 (0.00%)	0 out of 13 (0.00%)	0 out of 9 (0.00%)
Envelope protein	0 out of 8 (0.00%)	0 out of 4 (0.00%)	0 out of 4 (0.00%)

and 183, 189, 193 to 195, 217 to 220 are hydrophobic positions, while 3 to 7, 12 to 14, 16 to 20, 40 to 45, 72 and 73, 76, 103, 105 to 109, 111 to 116, 123, 125, 133, 135 and 136, 147 and 148, 152 to 167, 173 to 180, 184 to 188, 190 to 192, 196 to 216 are hydrophilic position and neutral places are 172, 181 (Figure 2E). Whereas, in case of envelop protein, hydrophobic region residues are 3 to 5, 11 to 54, 56 to 58, 60, 72 to 73 and hydrophilic region residues are 6 to 10, 55, 59, 61 to 71(Figure 2F). The GRAVY scores are 0.64, 1.45, 0.32, 0.23, 0.45, and 1.13, respectively for ORF10, ORF7b, ORF7a, ORF6, membrane glycoprotein, and envelope protein (Figure 2G).

Protein disorder

The protein disorder plot indicated that the disorder scores were higher for C-terminal half than N-terminal half for almost all selected proteins (Figure 3). Almost all proteins showed protein disorder scores indicating moderate flexible to highly flexible residues. However, membrane glycoprotein is more disordered compared to other proteins. Also, no protein revealed scores of $\leqslant 0.1$ indicating rigidity.

Amino acids composition and protein-binding propensity

The ORF10 protein consists of the highest percentage of asparagines (N), where, in case of ORF7b, ORF7a, membrane glycoprotein, and envelope protein, leucine (L) is presented in the maximum percentage. Moreover, for ORF6, it was isoleucine (Ile). The overall amino acid composition of all 6 proteins has been represented in Table 2. The binding propensity is important to influence electrostatic and aromatic interactions and also it is extremely varied with the amino acid residues. Several fluctuations have been observed in protein-binding propensity in both C-terminal half residues and N-terminal half residues (Supplemental Figure 1).

Aliphatic index and transmembrane helix

Aliphatic index values of more than 100 indicated that these proteins are highly thermo-stable over a wide temperature assortment. For ORF10, ORF7b, ORF7a, ORF6, membrane glycoprotein, and envelope protein, the aliphatic index values found were 107.63, 156.51, 100.74, 130.98, 120.86, and 144 respectively (Figure 4A). Transmembrane helices of less than one indicated these helices are less likely to interact with membrane lipids. For ORF10 and ORF7a, TH predicted spanning residues are 3 to 29 and 6 to 33, respectively. In case of ORF7b, TH predicted residues are 4 to 23 and 93 to 119. Furthermore, for ORF6 and envelope protein, the predicted TH spanning residues are 5 to 38 and 11 to 61. Finally, in case of membrane glycoprotein, the transmembrane helix residues are 18 to 59 and 61 to 105. The representation is in Figure 4B.

Protein secondary structures

In respect to ORF10, ORF7b, and envelope protein, α-helix spanning residues are 11 to 21, 4 to 35, and 4 to 64, respectively (Figure 5A, B, and F). Additionally, in ORF6, the α-helix spanning residues are 4 to 21, 26 and 27, 29 to 44, 48 to 51 (Figure 5D). In case of ORF7a, the α-helix and β-sheet spanning residues are 90 to 96, 99 and 100, and 28 to 33, 40 to 41, 53 to 66, 72 to 79, respectively (Figure 5C). The membrane glycoprotein has the α-helix and β-sheet spanning residues of 10 to 19, 22 to 36, 40 to 70, 75 to 106, 161 to 163, and 112, 118 to 123, 128 to 132, 139 to 146, 148 to 151, 154 to 159, 167 to 172, 175 to 185, 193 to 201, respectively (Figure 5E).

Protein modeling and validation

Initially, the models having low *P*-values and high-quality scores were subjected to refinement which was yielded by the IntFOLD web-based server. Then the selection was done according to the QMEAN *Z* score and Ramachandran plot score (Supplemental Table S1 and Figure 6A-F). QMEAN *Z* score and Ramachandran plot score table were added as Supplemental Table S1. In response to hydrophobic and hydrophilic properties, the majority of the proteins surfaces were found as hydrophobic (Figure 6G-L). The Ramachandran plot score details for all the proteins were found more than 90% except for membrane glycoprotein which is 87.6% (Figure 6M-R).

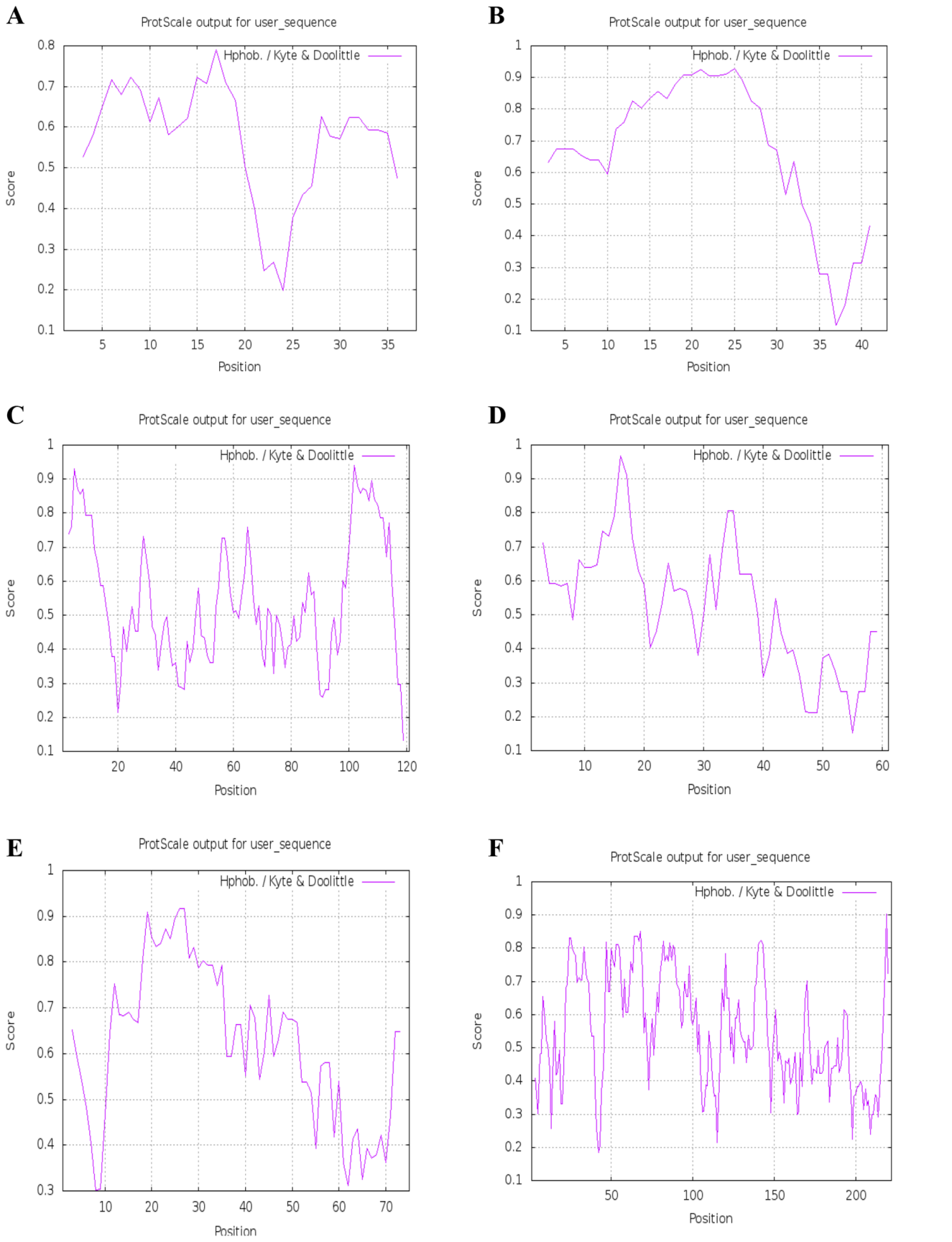

Figure 2. *(Continued)*

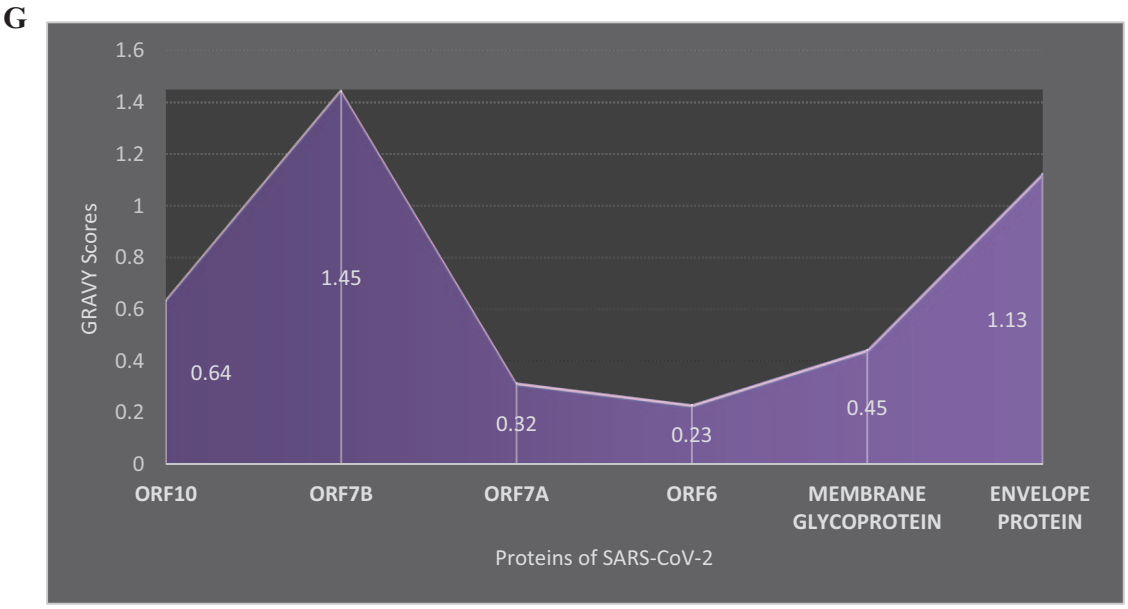

Figure 2. Hydrophobicity plot and GRAVY Scores of selected 6 proteins. (A-F) Hydrophobicity plot of ORF10, ORF7b, ORF7a, and ORF6, membrane glycoprotein, and envelope protein respectively. The hydrophobicity plots were generated according to the Kyte-Doolittle hydropathy plots. (G) GRAVY scores of ORF10, ORF7b, ORF7a, ORF6, membrane glycoprotein, and envelope protein. The numerical values for each score displayed above are their corresponding box. The proteins are recognized as mostly hydrophobic.

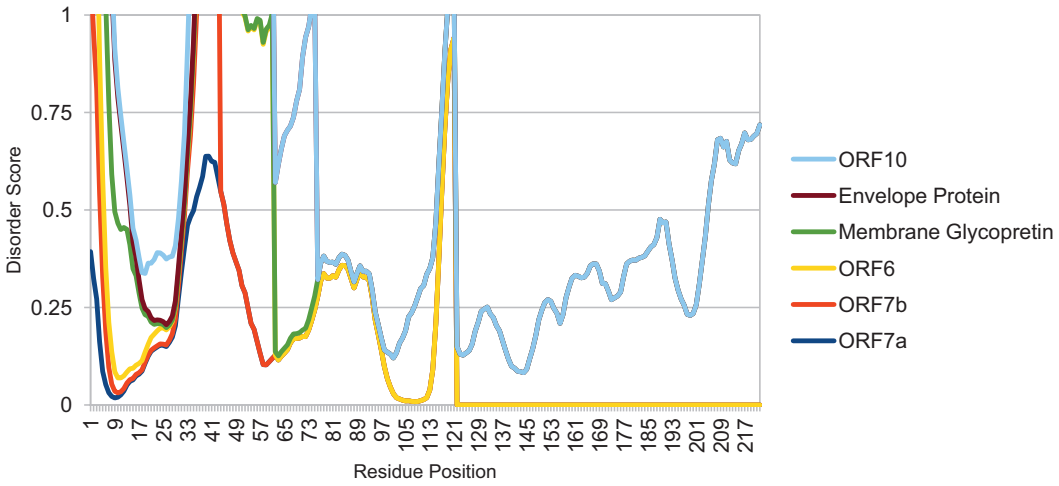

Figure 3. Per-residue disorder plot for ORF10, ORF7b, ORF7a, ORF6, membrane glycoprotein, and envelope protein of SARS-CoV2. All proteins found as highly flexible and ordered. Scores ⩾0.5 indicate disorder residues, while scores within 0.25 to 0.5 and 0.1 to 0.25 suggest highly flexible and moderate flexible residues. Scores ⩽0.1 indicate rigidity.

Discussion

The phylogenetic data in this study proposes that ORF10 protein in Spain is most distantly related to all other ORF10 proteins. Whereas, the high similarity was detected in the other remaining selected ORF10 proteins of SARS-CoV-2. For ORF7b, ORF7a, and ORF6, membrane glycoprotein, and envelope protein, there was no distant relationship among the strains selected from the other different countries. All these proteins showed 100% conserved region, thus, mutations in these regions were not detected. It also revealed that these proteins shared a strong phylogenic relationship with their common ancestors in the past. Conserved regions of ORF7b,

Table 2. Amino acid composition of ORF10, ORF7b, ORF7a, ORF6, membrane glycoprotein, and envelope protein in percentage (%).

AMINO ACID	ORF10 (%)	ORF7B (%)	ORF7A (%)	ORF6 (%)	MEMBRANE GLYCOPROTEIN (%)	ENVELOPE PROTEIN (%)
Ala (A)	5.3	4.7	7.4	1.6	8.6	5.3
Arg (R)	5.3	0.0	4.1	1.6	6.3	4.0
Asn (N)	13.2	2.3	1.7	6.6	5.0	6.7
Asp (D)	2.6	4.7	1.7	6.6	2.7	1.3
Cys (C)	2.6	4.7	5.0	0.0	1.8	4.0
Gln (Q)	2.6	2.3	4.1	4.9	1.8	0.0
Glu (E)	0.0	7.0	6.6	8.2	3.2	2.7
Gly (G)	2.6	0.0	3.3	0.0	6.3	1.3
His (H)	0.0	4.7	2.5	1.6	2.3	0.0
Ile (I)	7.9	11.6	6.6	16.4	9.0	4.0
Leu (L)	10.5	25.6	12.4	13.1	15.8	18.7
Lys (K)	0.0	0.0	5.8	6.6	3.2	2.7
Met (M)	5.3	4.7	0.8	4.9	1.8	1.3
Phe (F)	10.5	14.0	8.3	4.9	5.0	6.7
Pro (P)	2.6	0.0	5.0	1.6	2.3	2.7
Ser (S)	5.3	4.7	5.8	6.6	6.8	10.7
Thr (T)	5.3	2.3	8.3	4.9	5.9	5.3
Trp (W)	0.0	2.3	0.0	1.6	3.2	0.0
Tyr (Y)	7.9	2.3	4.1	3.3	4.1	5.3
Val (V)	10.5	2.3	6.6	4.9	5.4	17.3
Pyl (O)	0.0	0.0	0.0	0.0	0.0	0.0
Sec (U)	0.0	0.0	0.0	0.0	0.0	0.0

ORF7a, and ORF6, membrane glycoprotein, and envelope protein could be playing a fundamental role in the assembly of particular proteins, formation of protein structure, and/or demonstrating virulent functions by facilitating precise protein interactions. Although, defining relationships between specific sequences is not entirely possible when based solely on sequence data.[30] We can predict that the selected proteins of SARS-CoV-2 here are highly ordered as intrinsically disordered proteins have a tendency to be phosphorylated that leads to disorder-to-order and order-to-disorder transitions.[31]

Phosphorylation controls the function of a particular protein and cell signaling by changing conformational shape in the phosphorylated protein which maintains the catalytic property of the protein. Thus, activation or inactivation of proteins mainly depends on phosphorylation.[32] Prediction of phosphorylation site of selected 6 SARS-CoV-2 proteins conceded that phosphorylated serine, threonine, or tyrosine was mostly not present though ORF7a had a single phosphorylated serine. In every conceivable way, the phosphorylation of a distinct protein is able to modify its activities which include inflection of protein's intrinsic biological property, proper sub-cellular location, docking with other related proteins, and half-life. It also decides the level and period of a response given by a protein which acts as an input to signal integration.[33] Moreover, sites of phosphorylation are more prone to be evolutionary conserved than other interfacial residues.[34]

The purpose of the hydropathy index of amino acids is mainly to predict the function of a structurally or functionally unknown protein. The distribution of hydropathy clusters in a particular protein appears to recommend that these cluster location is principally conserved in a given group of proteins.[35] In the present study, selected 6 SARS-CoV-2 proteins

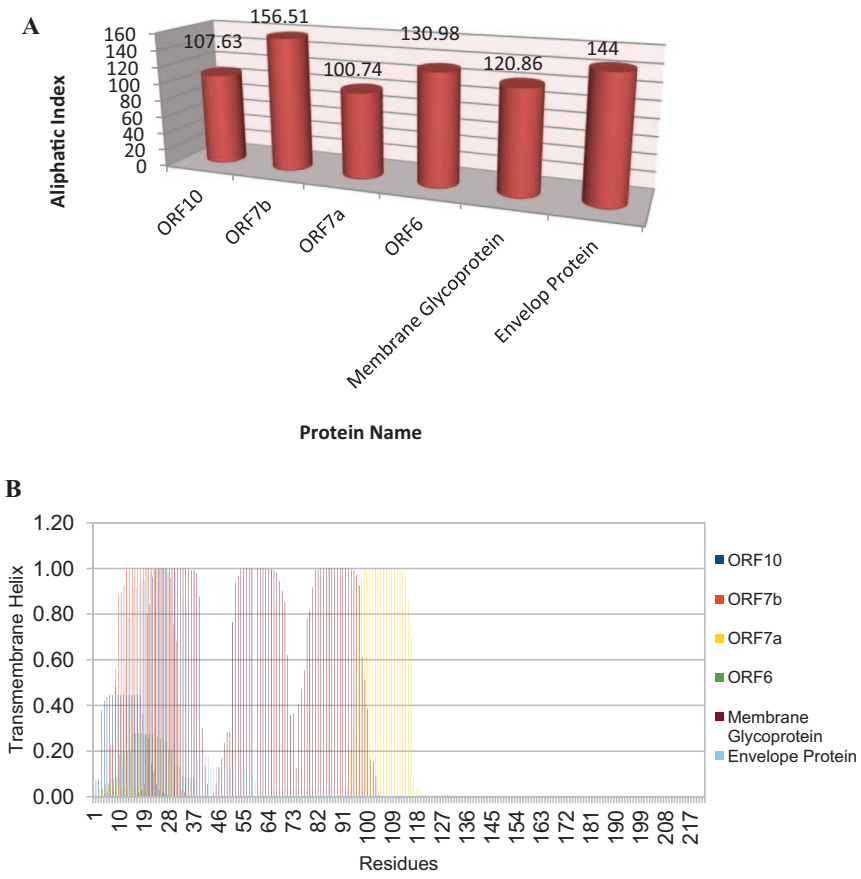

Figure 4. (A, B) Aliphatic indexes and transmembrane helix prediction scores of ORF10, ORF7b, ORF7a, ORF6, membrane glycoprotein, and envelope protein respectively. All proteins showed aliphatic index values of more than 100 indicated that they are highly thermo-stable and transmembrane helices of less than 1.

expressed hydropathy index which tended to be more hydrophobic. The literature revealed that hydrophobic proteins are more soluble and for this reason, they can function in an independent manner by avoiding undesirable interactions with watery molecules. In addition to that, these proteins are vital for protein folding which keeps it more stable and biologically active.[36]

Protein disorder predictions are an enormous challenge in structural proteomics and subsequently its function prediction including identification of those proteins that are unstructured either partially or wholly. In the current study, protein disorder predictions revealed that almost all proteins showed protein disorder scores indicating moderate flexible to highly flexible residues and no protein revealed scores which indicates rigidity. This result coincides with the interpretation presented by another research.[37] However, membrane glycoprotein was more disordered compared to other proteins in this study. Disordered regions present in specific proteins could contain short linear peptide motifs which may later play a significant role in protein function. After predicting, avoidance of prospective disordered regions in protein can augment expression, proper foldability, and stability of that expressed protein.[38]

Protein binding propensity augments the knowledge of protein-protein interactions, docking, and annotation of functional properties of that protein at the molecular level.[19] In addition, a high aliphatic index resembles to rise of the thermostability of globular proteins.[39] All 6 selected SARS-CoV-2 proteins showed aliphatic index of more than 100 which indicates these proteins are highly thermostable over a wide range of temperature. Additionally, all 6 selected proteins showed transmembrane helixes which are less than 1 and transmembrane helixes have immense importance in the study of membrane proteins.[40]

Due to the significance of structural class prediction of protein, diverse major efforts have been employed to discover a prediction model that establishes the structural class and predicts protein secondary structure depending on the sequences of specific protein.[41,42] The prediction of secondary structures for 6 selected SARS-CoV-2 proteins revealed that each ORF7a protein and membrane glycoprotein has 1 α-helix and 1 β-strand. The structural class is one of the most imperative features for its vital task in the analysis of protein function, prediction of the rate of protein folding nature, and, also, execution of a suitable approach to uncover protein tertiary structure.[43-45]

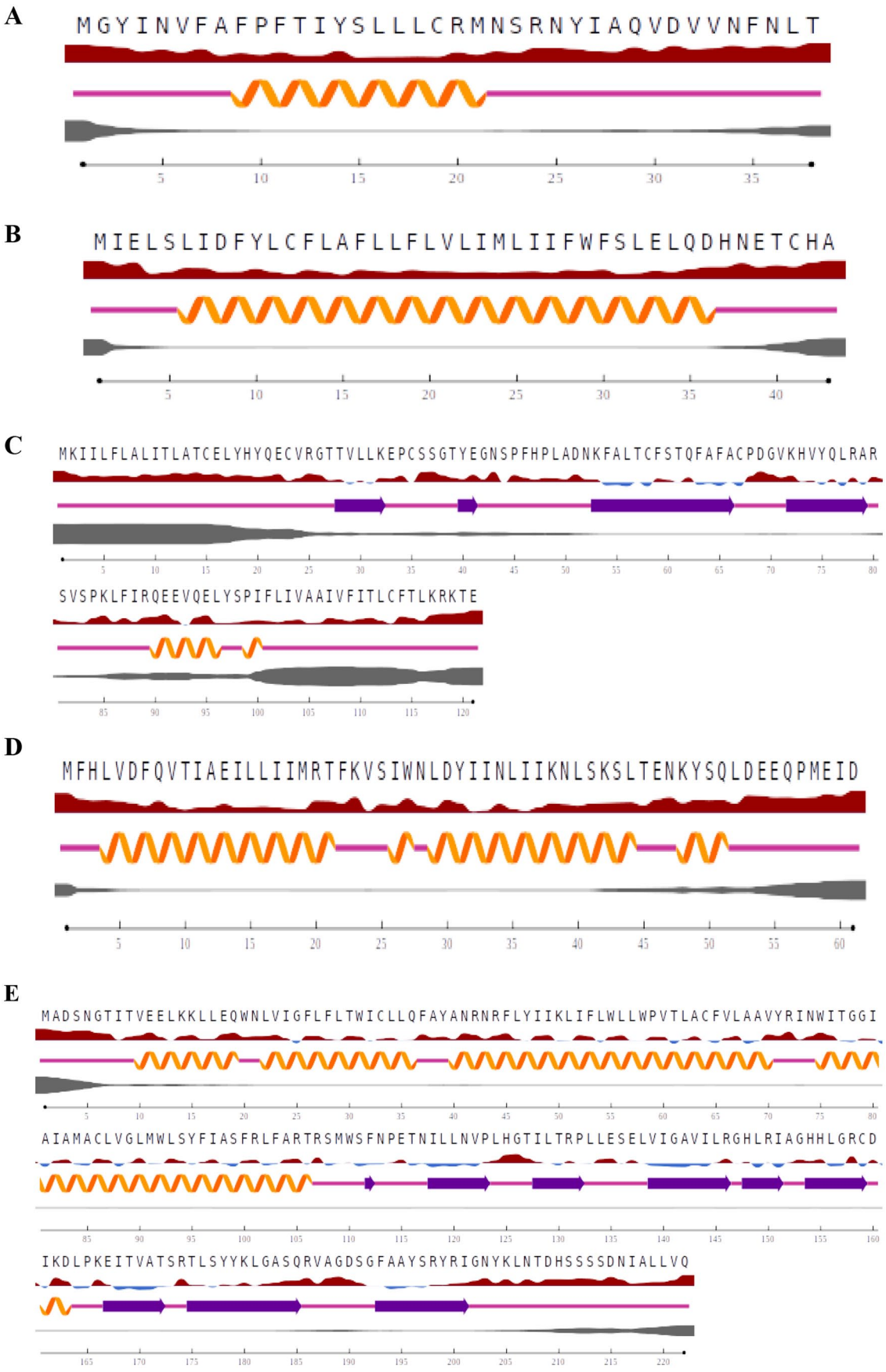

Figure 5. *(Continued)*

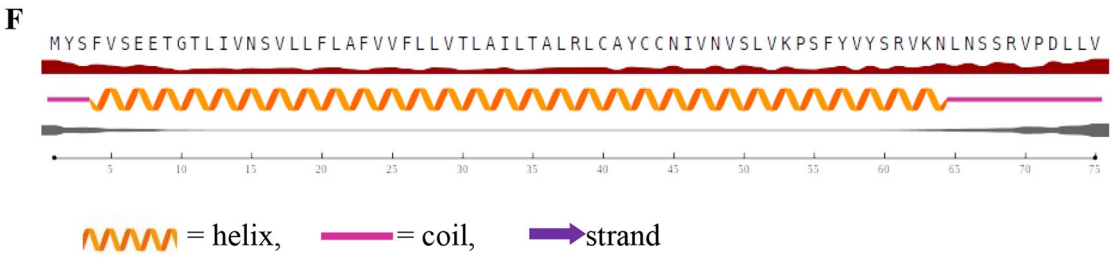

Figure 5. Secondary structure prediction of 6 selected proteins. (A-F) Secondary structure of ORF10, ORF7b, ORF7a, ORF6, membrane glycoprotein, and envelope protein respectively.

Structure based antibody against SARS-CoV-2 can be a way to suppress the infection rate caused by this particular virus. By targeting specific proteins of this virus that can invade human body by directly attaching to the host cells would be a suitable approach.[46] A recent study has demonstrated that SARS-CoV-2 attacks host cells via CD147-spike protein and this invasion of SARS-CoV-2 is mediated by a transmembrane glycoprotein from the immunoglobulin super family. An anti-CD147 humanized antibody named Meplazumab have the ability to block CD147 and subsequently prevention of SARS-CoV-2 to entrance to the host cells is occured.[47] Thus, critical characterization and function analysis of structural proteins of SARS-CoV-2 is utmost necessary issue in therapeutic perspective.[48-50]

Different proteins of SARS-CoV-2 plays significant role to express its virulence in host. ORF10 protein of SARS-CoV-2 interacts with multiple human proteins after entering the body to control over the different molecular mechanisms. Mutations in the ORF-10 present a new level of severe infection rate.[51] ORF7b protein of SARS-CoV-2 is an integral membrane protein that encoded within subgenomic RNA7. During infection, it accumulates in Golgi compartment associating with both *cis* and *trans* Golgi marker and causing Golgi compartment localization.[52] Whereas, ORF7a protein of SARS-CoV-2 hinders bone marrow stromal antigen 2 virion tethering by a new system of interference of glycosylation process.[53]

ORF6 protein of SARS-CoV-2 was able to inhibit beta interferon (IFN-β) expression by halting its synthesis and signaling.[54] Protein-protein interactions and protein-RNA interactions are significant for competent assembly of virion. Membrane glycoprotein of SARS-CoV-2 express a vital role in this purpose as formation of virus-like particle (VLP) in numerous SARS-CoV-2 involves only membrane glycoprotein, and envelope protein.[55]

Several in silico SARS-CoV-2 research presented the structure and functional perspective of the novel virus focusing its virulence transmission in human genome.[56-58] The present study explored theoretical modeling, sequence, and structure-based functional characterization of 6 accessory proteins. Phylogenetic analysis of these proteins exposed a close evolutionary relationship with the proteins of distant origins. In this present study, the stable tertiary structure of proteins was predicted which gives the primary notion about the interaction of this protein 3D structures with enzymes or host receptors. Also, in this study, hydrophobicity surface map of particular proteins was created to distinctly show the hydrophobic or hydrophilic regions of protein. Selected 6 proteins of SARS-CoV-2 Bangladeshi variant were characterized as highly flexible, structurally and electrostatically extremely stable, ordered, biologically active, hydrophobic, and closely related to the proteins of different variants. Studying these diverse proteins of the SARS-CoV-2 virus has already yielded some clues about how they connect with the human cells but much remains to be assessed. Though further comprehensive assessment with broad-scale data are required to elucidate these upshots generated in this current study.

Conclusions

The analysis includes detailed information regarding the characterization and structure of proteins of SARS-CoV-2 Bangladeshi variant which was performed for the first time ever to enlighten the deep mechanism behind the virulence of the particular virus. Communally, the present study provides an interesting basis for characterizing proteins of novel viruses theoretically and structurally. The selected 6 proteins characterized as stable, ordered, hydrophobic, and also share strong phylogenetic relationships with proteins of other closely related SARS-CoV-2. Finally, the tertiary models of protein constructed in this study have higher quality and stability. This analysis can offer a foundation to perform the further analysis necessary to evaluate the biological function, interaction, and relevance to viral property of the 6 proteins in SARS-CoV-2. These predicted structures would be functional for investigation of each protein interaction and their functionalities by advanced computational analysis, understanding of viral pathogenesis or to study potential vaccines and especially, to avert epidemics and pandemics.

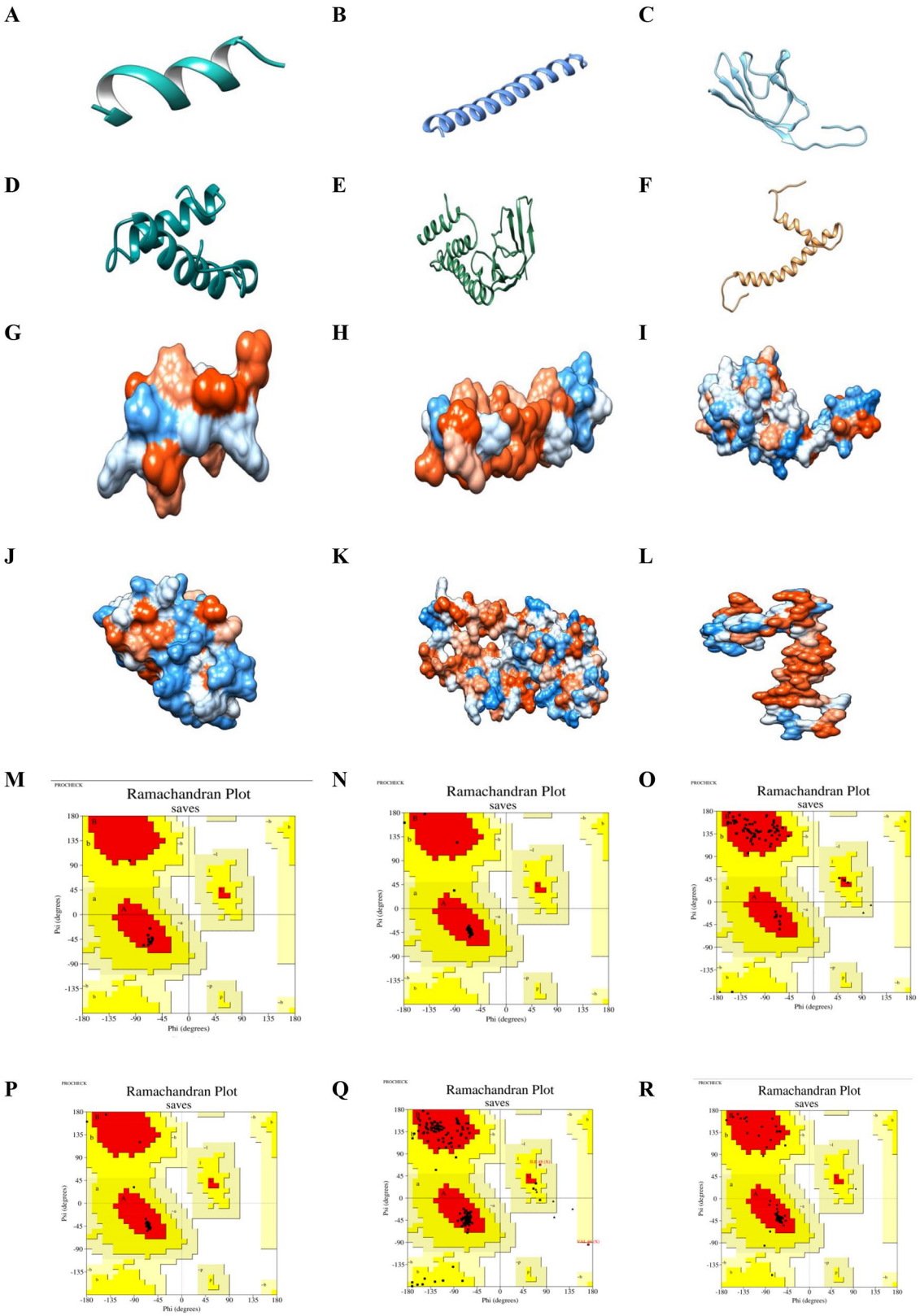

Figure 6. Protein modeling and hydrophobicity surface 3D map of the selected 6 proteins. (A-F) Ribbon diagram of ORF10, ORF7b, ORF7a, ORF6, membrane glycoprotein, and envelope protein respectively. (G-L) Hydrophobicity surface map of ORF10, ORF7b, ORF7a, ORF6, membrane glycoprotein, and envelope protein respectively. All 6 protein models are found as highly flexible and stable. The blue color represents hydrophilic regions and the orange color expresses hydrophobic regions. Where, the whitish-blue color indicates semi-hydrophobic/hydrophilic character. (M-R) Representation of Ramachandran plot of ORF10, ORF7b, ORF7a, ORF6, membrane glycoprotein, and envelope protein, respectively.

Author Contributions

Pinky Debnath: Conceived, designed, analyzed, interpreted and wrote the study.
Umama Khan: Analyzed, interpreted and wrote the study.
Md. Salauddin Khan: Critically reviewed and supervised the study.

Disclaimers

This paper contains original research and has not been submitted or published earlier in any journal and is not being considered for publication elsewhere.

ORCID iD

Salauddin Khan https://orcid.org/0000-0002-5001-7877

Supplemental Material

Supplemental material for this article is available online.

REFERENCES

1. Wang C, Horby PW, Hayden FG, Gao GF. A novel coronavirus outbreak of global health concern. *Lancet*. 2020;395:470-473.
2. Paules CI, Marston HD, Fauci AS. Coronavirus infections-more than just the common cold. *JAMA*. 2020;323:707-708.
3. Sahin AR, Erdogan A, Agaoglu PM, et al. 2019 novel coronavirus (COVID-19) outbreak: a review of the current literature. *Eur J Med Oncol*. 2020;4:1-7.
4. Shereen MA, Khan S, Kazmi A, Bashir N, Siddique R. COVID-19 infection: origin, transmission, and characteristics of human coronaviruses. *J Adv Res*. 2020;24:91-98.
5. Khailany RA, Safdar M, Ozaslan M. Genomic characterization of a novel SARS-CoV-2. *Gene Rep*. 2020;19:100682.
6. Schaecher SR, Mackenzie JM, Pekosz A. The ORF7b protein of severe acute respiratory syndrome coronavirus (SARS-CoV) is expressed in virus-infected cells and incorporated into SARS-CoV particles. *J Virol*. 2007;81:718-731. doi:10.1128/JVI.01691-06
7. Lu S, Wang J, Chitsaz F, et al. CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res*. 2020;48:D265-D268.
8. Geng H, Liu YM, Chan WS, et al. The putative protein 6 of the severe acute respiratory syndrome-associated coronavirus: expression and functional characterization. *FEBS Lett*. 2005;579:6763-6768.
9. Wu F, Zhao S, Yu B, et al. Author correction: a new coronavirus associated with human respiratory disease in China. *Nature*. 2020;580:E7.
10. de Haan CA, Smeets M, Vernooij F, Vennema H, Rottier PJ. Mapping of the coronavirus membrane protein domains involved in interaction with the spike protein. *J Virol*. 1999;73:7441-7452. doi:10.1128/JVI.73.9.7441-7452.1999
11. Surya W, Li Y, Torres J. Structural model of the SARS coronavirus E channel in LMPG micelles. *Biochim Biophys Acta Biomembr*. 2018;1860:1309-1317.
12. Hossain ME, Rahman MM, Alam MS, et al. Genome sequence of a SARS-CoV-2 strain from Bangladesh that is nearly identical to United Kingdom SARS-CoV-2 variant B.1.1.7. *Microbiol Resour Announc*. 2021;10:e00100-e00121.
13. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32:1792-1797. doi:10.1093/nar/gkh340
14. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol*. 2018;35:1547-1549.
15. Igic B. Phylogenetic trees made easy: a how-to manual, 2nd ed. *J Hered*. 2005;96:469-470.
16. Xue B, Dunbrack RL, Williams RW, Dunker AK, Uversky VN. PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochim Biophys Acta*. 2010;1804:996-1010.
17. Duvaud S, Gabella C, Lisacek F, Stockinger H, Ioannidis V, Durinx C. Expasy, the Swiss Bioinformatics Resource Portal, as designed by its users. *Nucleic Acids Res*. 2021. doi:10.1093/nar/gks225
18. Giri R, Bhardwaj T, Shegane M, et al. Understanding COVID-19 via comparative analysis of dark proteomes of SARS-cov-2, human SARS and bat SARS-like coronaviruses. *Cell Mol Life Sci*. 2021;78:1655-1688.
19. Zhang J, Kurgan L. SCRIBER: accurate and partner type-specific prediction of protein-binding residues from proteins sequences. *Bioinformatics*. 2019;35:i343-i353.
20. Klausen MS, Jespersen MC, Nielsen H, et al.NetSurfP-2.0: improved prediction of protein structural features by integrated deep learning. *Proteins*. 2019;87:520-527.
21. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol*. 2001;305:567-580. doi:10.1006/jmbi.2000.4315
22. Käll L, Krogh A, Sonnhammer EL. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol*. 2004;338:1027-1036. doi:10.1016/j.jmb.2004.03.016
23. McGuffin LJ, Adiyaman R, Maghrabi AHA, et al. IntFOLD: an integrated web resource for high performance protein structure and function prediction. *Nucleic Acids Res*. 2019;47:W408-W413.
24. Bhattacharya D, Cheng J. 3Drefine: consistent protein structure refinement by optimizing hydrogen bonding network and atomic-level energy minimization. *Proteins*. 2013;81:119-131.
25. Benkert P, Biasini M, Schwede T. Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics*. 2011;27:343-350.
26. Hollingsworth SA, Karplus PA. A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins. *Biomol Concepts*. 2010;1:271-283.
27. Pettersen EF, Goddard TD, Huang CC, et al. UCSF chimera—a visualization system for exploratory research and analysis. *J Comput Chem*. 2004;25:1605-1612.
28. Lineback JE, Jansma AL. PyMOL as an instructional tool to represent and manipulate the myoglobin/hemoglobin protein system. *J Chem Educ*. 2019;96:2540-2544. doi:10.1021/acs.jchemed.9b00143
29. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol*. 1982;157:105-132. doi:10.1016/0022-2836(82)90515-0
30. Schuster NA. Characterization and structural prediction of the putative ORF10 protein in SARS-CoV-2. *bioRxiv*. 2021:2020-10.
31. Collins MO, Yu L, Campuzano I, Grant SG, Choudhary JS. Phosphoproteomic analysis of the mouse brain cytosol reveals a predominance of protein phosphorylation in regions of intrinsic sequence disorder. *Mol Cell Proteomics*. 2008;7:1331-1348.
32. Stolarczyk EI, Reiling CJ, Paumi CM. Regulation of ABC transporter function via phosphorylation by protein kinases. *Curr Pharm Biotechnol*. 2011;12:621-635. doi:10.2174/138920111795164075
33. Cohen P. The regulation of protein function by multisite phosphorylation—a 25 year update. *Trends Biochem Sci*. 2000;25:596-601. doi:10.1016/s0968-0004(00)01712-6
34. Nishi H, Hashimoto K, Panchenko AR. Phosphorylation in protein-protein binding: effect on stability and function. *Structure*. 2011;19:1807-1815.
35. Damodharan L, Pattabhi V. Hydropathy analysis to correlate structure and function of proteins. *Biochem Biophys Res Commun*. 2004;323:996-1002. doi:10.1016/j.bbrc.2004.08.186
36. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R. Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. *Protein Sci*. 1997;6:53-64. doi:10.1002/pro.5560060106
37. Simm S, Einloft J, Mirus O, Schleiff E. 50 years of amino acid hydrophobicity scales: revisiting the capacity for peptide classification. *Biol Res*. 2016;49:31.
38. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. Protein disorder prediction: implications for structural proteomics. *Structure*. 2003;11:1453-1459. doi:10.1016/j.str.2003.10.002
39. Ikai A. Thermostability and aliphatic index of globular proteins. *J Biochem*. 1980;88:1895-1898. doi:10.1093/oxfordjournals.jbchem.a133168
40. Cuthbertson JM, Doyle DA, Sansom MS. Transmembrane helix prediction: a comparative evaluation and analysis. *Protein Eng Des Sel*. 2005;18:295-308. doi:10.1093/protein/gzi032
41. Chen C, Tian YX, Zou XY, Cai PX, Mo JY. Using pseudo-amino acid composition and support vector machine to predict protein structural class. *J Theor Biol*. 2006;243:444-448. doi:10.1016/j.jtbi.2006.06.025
42. Zhang TL, Ding YS, Chou KC. Prediction protein structural classes with pseudo-amino acid composition: approximate entropy and hydrophobicity pattern. *J Theor Biol*. 2008;250:186-193. doi:10.1016/j.jtbi.2007.09.014
43. Dai Q, Li Y, Liu X, Yao Y, Cao Y, He P. Comparison study on statistical features of predicted secondary structures for protein structural class prediction: from content to position. *BMC Bioinformatics*. 2013;14:152.
44. Kirtipal N, Bharadwaj S, Kang SG. From SARS to SARS-cov-2, insights on structure, pathogenicity and immunity aspects of pandemic human coronaviruses. *Infect Genet Evol*. 2020;85:104502.
45. Zhu G, Zhu C, Zhu Y, Sun F. Minireview of progress in the structural study of SARS-CoV-2 proteins. *Curr Res Microb Sci*. 2020;1:53-61.
46. Wang MY, Zhao R, Gao LJ, Gao XF, Wang DP, Cao JM. SARS-CoV-2: structure, biology, and structure-based therapeutics development. *Front Cell Infect Microbiol*. 2020;10:587269.
47. Wang K, Chen W, Zhang Z, et al. CD147-spike protein is a novel route for SARS-CoV-2 infection to host cells. *Signal Transduct Target Ther*. 2020;5:1-10.

48. Duan L, Zheng Q, Zhang H, Niu Y, Lou Y, Wang H. The SARS-CoV-2 spike glycoprotein biosynthesis, structure, function, and antigenicity: implications for the design of spike-based vaccine immunogens. *Front Immunol*. 2020;11:576622.
49. Sternberg A, Naujokat C. Structural features of coronavirus SARS-CoV-2 spike protein: targets for vaccination. *Life Sci*. 2020;257:118056.
50. Ita K. Coronavirus disease (COVID-19): current status and prospects for drug and vaccine development. *Arch Med Res*. 2021;52:15-24.
51. Hassan SS, Attrish D, Ghosh S, et al. Notable sequence homology of the ORF10 protein introspects the architecture of SARS-CoV-2. *Int J Biol Macromol*. 2021;181:801-809.
52. Schaecher SR, Diamond MS, Pekosz A. The transmembrane domain of the severe acute respiratory syndrome coronavirus ORF7b protein is necessary and sufficient for its retention in the Golgi complex. *Virol J*. 2008;82:9477-9491.
53. Taylor JK, Coleman CM, Postel S, et al. Severe acute respiratory syndrome coronavirus ORF7a inhibits bone marrow stromal antigen 2 virion tethering through a novel mechanism of glycosylation interference. *Virol J*. 2015;89:11820-11833.
54. Kopecky-Bromberg SA, Martínez-Sobrido L, Frieman M, Baric RA, Palese P. Severe acute respiratory syndrome coronavirus open reading frame (ORF) 3b, ORF 6, and nucleocapsid proteins function as interferon antagonists. *Virol J*. 2007;81:548-557.
55. Ujike M, Taguchi F. Incorporation of spike and membrane glycoproteins into coronavirus virions. *Viruses*. 2015;7:1700-1725.
56. Shishir TA, Naser IB, Faruque SM. In silico comparative genomics of SARS-CoV-2 to determine the source and diversity of the pathogen in Bangladesh. *PLoS One*. 2021;16:e0245584.
57. Afrin SZ, Islam MT, Paul SK, Kobayashi N, Parvin R. Dynamics of SARS-CoV-2 variants of concern (VOC) in Bangladesh during the first half of 2021. *Virology*. 2022;565:29-37.
58. Hossain MS, Tonmoy MIQ, Fariha A, et al. Prediction of the effects of variants and differential expression of key host genes ACE2, TMPRSS2, and FURIN in SARS-CoV-2 pathogenesis: an in silico approach. *Bioinform Biol Insights*. 2021;15:11779322211054684.
59. Debnath P, Khan U, Khan MS. Characterization and structural prediction of ORF10, ORF7b, ORF7a, ORF6, membrane glycoprotein, and envelope protein in SARS-CoV-2 Bangladeshi variant through bioinformatics approach. *bioRxiv*. 2021.