Spike protein modeling and single amino acid variant analysis might suggest reduced transmitability of SARS-CoV-2 in Jordan, Middle East

Walid Al-Zyoud (walid.alzyoud@gju.edu.jo)
German Jordanian University https://orcid.org/0000-0002-3772-5617
Hazem Haddad
Jordan University of Science & Technology
Ramzi Foudeh
Jordanian Society of Genetic Engineers (JSGE), Amman, Jordan

Research Article

Keywords: COVID-19, SARS, spike, variants, structure

Posted Date: June 4th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-33156/v1

Abstract

Spike protein (approx. 180 kDa) is the surface glycoprotein of the severe acute respiratory syndrome-coronavirus-2 (SARS-CoV-2) necessary for the interaction of the virus with human endothelial cell receptors on the cell membrane to be engulfed causing COVID-19 disease after binding with the angiotensin-converting enzyme 2 (ACE2) with an evident activation by type II transmembrane protease TMPRSS2 . Therefore, mutations and amino acid variants analysis are essential in characterizing the mechanism of binding of spike protein with its receptor, which totally gives insights on possibilities to design a peptide or nucleotide-based vaccine for COVID-19. Here, we employed Iterative Threading Assembly Refinement (I-TASSER) and Multiple Alignment using Fast Fourier Transform (MAFFT) to predict the three-dimensional structure and to analyze the amino acid variants for spike protein sequences of SARS-CoV-2 from GISAID database of samples collected from Jordan to try to find a justification for low number of confirmed COVID-19 in Jordan, Middle East. Our findings showed the molecules structurally close to the spike glycoprotein from the Enzyme Commission (EC) numbers and active sites included Isoleucyl-tRNA synthetase, Crystal structure of the tricorn protease (hydrolase); Crystal structure of the T. Thermophilus RNA polymerase holoenzyme (transferase); Crystal structure of the complex between pyruvate-ferredoxin oxidoreductase from Desulfovibrio africanus and pyruvate (oxidoreductase); and Reovirus core (virus). Our MAFFT findings showed that Four Amino Acid Variants (SAV) founded in 20 samples of SARS-CoV-2 were not conserved residues in spike glycoprotein. What is equal to 5% of samples showed tyrosine (polar) deletion at Y144 , 62% of samples showed aspartate (polar, acidic) substitution to glycine (nonpolar) at D614G, 5% of samples showed aspartate (polar, acidic) substitution to tyrosine (polar) at D1139Y and 5% of samples showed glycine (nonpolar) substitution to serine (polar) at G1167S respectively. By using Phyre2, our findings have shown lower sensitive mutational that cannot affect the pocket region or alpha and beta-sheet in all mutations except for D614G, which has the highest mutational sensitivity score (5 out of 9) indicating a bigger effect on the function of spike protein. This might suggest, in general, a reduced transmitability of SARS-CoV-2 in Jordan, Middle East. As the crystal structure of spike protein is not revealed yet, it was not possible to compare the predicted modes versus each other.

1. Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV–2) caused an outbreak in Wuhan city, China, at the beginning of December 2019 that rapidly spread across the country and to other nations around the world and characterized as a pandemic by the World Health Organization WHO [1]. The first case of SARS-CoV–2 was reported the ministry of health in Jordan on the 2nd of March 2020 for a Jordanian citizen who returned from Italy. To the date of this report, there are 459 confirmed cases, 364 recovered and eight deaths of COVID–19 in Jordan, according to the official web site launched by the Jordanian Ministry of health as a unified source of information about coronavirus (https://corona.moh.gov.jo/en).

SARS-CoV–2 has a positive, single-strand RNA genome that is over 29 kilobases in length, which belongs to one of the four genera of *Orthocoronaviridae,* the beta-coronavirus [2]. Moreover, SARS-CoV–2 encodes four major structural proteins, the envelope (E), membrane (M), nucleocapsid (N), and spike (S) proteins. Spike protein (approx. 180 kDa) is the surface glycoprotein of the severe acute respiratory syndrome-coronavirus–2 (SARS-CoV–2)[3]. Spike glycoprotein is necessary for the interaction of the virus with human cell receptors for a sequential combination of the viral encompass with the cell membrane to be engulfed and permit COVID–19 disease by binding with the angiotensin-converting enzyme 2 (ACE2) [4] [5] after an evident activation by type II transmembrane protease TMPRSS2 [6].

Here, to understand the early steps of COVID–19 infection, we predicted a three-dimensional structure of the spike glycoprotein of SARS-CoV–2 from positive nasopharyngeal specimens collected in Jordan and sequenced by Biolab Diagnostic Laboratories (Jordan) & Andersen lab at Scripps Research (USA) who published sequences were retrieved from GISAID, a maintained global database based in Germany. The insight in this work is helpful for scientists to understand different molecular and cytological approaches involved in vaccine development for COVID–19.

2. Materials And Methods

2.1 Genomic sequence retrieval

A total of 19 whole-genome sequences of SARS-CoV–2 collected from Jordan were retrieved from GISAID database and analyzed at the amino sequence level of the spike glycoprotein. The database showed that the nasopharyngeal specimens were collected through March 2020 only with GISAID sequential accession number from EPI_ISL_429992 to EPI_ISL_4300015.

2.2 Iterative Threading ASSEmbly Refinement (I-TASSER)

To produce a predicted three-dimensional structure for the S-protein of SARS-CoV–2 collected in Jordan as a PDB file, a hierarchical approach to protein structure and function prediction known as I-TASSER server was used. The I-TASSER pipeline consists of three steps: 1) identification of models, 2) assembly of full-length structures, and 3) annotation of structure-based functions.

2.3 Submitting sequence in FASTA format and Multiple Alignment using Fast Fourier Transform

The FASTA formats of the spike gene were aligned (Appendix A), isolated and translated into 1273 amino acids from the whole genome 20 (Jordan) sequences plus 1 reference sequence (accession number YP_009724390.1) of the SARS-CoV–2 by using an open-source functions by the The University of Alcalá, Madrid, Spain at (http://biomodel.uah.es/en/lab/cybertory/analysis/trans.htm) and the BLAST function at the NCBI, a web-based service, in addition to Multiple Alignment using Fast Fourier Transform (MAFFT)

[7] and viewed by Jalview [8] of Dundee University Scotland. Then the FASTA format of an amino acid sequence of S-protein was submitted to I-TASSER server to get protein structure and function prediction (see appendix for the submitted Sequence in FASTA format).

2.4 Single Amino Acid Variant (SAV) Phenotype, protein modeling, and mutation analysis

Four Amino Acid Variant (SAV) found from 23 spike glycoprotein submitted and retrieved by *Phyre2* server to predict mutational sensitivity [9–11].

2.5 Nomenclature sequence Amino Acid Variant (SAV) and annotation used the accession number

Surface glycoprotein [Severe acute respiratory syndrome coronavirus 2] with accession number YP_009724390.1 was used as a reference sequence to compare with, and it was downloaded from https://www.ncbi.nlm.nih.gov/protein/YP_009724390.1?report = fasta.

3. Results

3.1 Predicted Secondary Structure and Predicted Solvent Accessibility

Initially, the I-TASSER recognizes basic templates from the PDB by multiple threading approach LOMETS, with full-length atomic models produced by iterative fragment assembly simulations based on templates. Function insights of the targeted molecule are then obtained by rethreading the three-dimensional models via the BioLiP database of protein functions.

Figure 1 shows the first part of the predicted secondary structure of the SARS-CoV–2 spike glycoprotein tested in Jordan defined as (H) Helix, (S) Strand and (C) Coil, in addition to the predicted accessibility of the solvent within a value range from zero (lowest accessible) to nine (highest accessible).

3.2 Predicted normalized B-factor

Figure 2 shows the B-factor, which is a value indicating the extent of inherent residue/atomic thermal mobility in proteins. In I-TASSER, in conjunction with sequence profiles obtained from sequence databases, this value is deduced from the PDB threading template proteins. The B-factor profile described in the figure below corresponds to the target protein's normalized B-factor, as determined by B=(B'-u)/s.

3.3 Top Ten threading templates used by I-TASSER

I-TASSER modeling starts from the PDB library structure templates, which LOMETS identifies. LOMETS is a meta-server threading approach with multiple threading programs, where each threading program can create tens of thousands of template alignments. I-TASSER uses only the most important models in the threading alignments, the value of which is determined by the Z-score, i.e., the difference between the raw and the average scores in the standard deviation unit. The templates in Figure 3 are the ten best templates from the LOMETS threading programs chosen. Typically, a prototype with the highest Z-value is chosen for each threading program, where the threading programs are sorted according to the average efficiency of the large-scale tests.

In Figure 3, all remaining residues are colored in black; the color is therefore given to those residues that are the same as the residue in the sequence of the request. The coloring mechanism is based on the property of amino acids, which are vividly colored by polar while dark shaded non-polar residues. The rank of templates lists the top 10 thread templates used by I-TASSER. Ident1 is the template sequence percentage identity in the area that is aligned to the query sequence of the thread. Ident2 is the sequence identity percentage for the entire query sequence template chains. Cov represents the alignment coverage and is proportional to the number of aligned residues divided by the query protein frequency. Norm. Z mark is the threading alignment's uniform Z symbol. Aligning to the standardized Z-point>1 is good alignment and vice versa. The top 10 alignments reported above (in order of their ranking) are from the following threading programs:

1: MUSTER 2: FFAS-3D 3: SPARKS-X 4: HHSEARCH2 5: HHSEARCH I 6: Neff-PPAS

7: HHSEARCH 8: pGenTHREADER 9: PROSPECT2 10: PRC [12].

3.4 Top five final models predicted by I-TASSER

For each target, an extensive collection of structural conformations is generated by I-TASSER simulations called decoys. I-TASSER uses the SPICKER to cluster all architectural structures based on the pair-sided similarity and records up to 5 models corresponding to the five largest structural clusters. The reliability of each model is evaluated quantitatively by a C-score based on the value of threaded prototype alignments and the parameters of convergence of structural mounting simulations. C-score is usually [–5,2], where a higher-value C-score means a more positive and vice versa scale.

Following the association observed between these attributes, the TM-score and RMSD are calculated using the C and the protein frequency. Since the group size classes the top 5 models, in some situations, a higher C-score is possible for the lower-ranking models. While the first model is better in most cases, lower-level models can also be better than higher-level models as seen in our research. If the I-TASSER simulations converge, less than 5 clusters can have been generated; it usually shows that because of the converged simulations, the models have good quality (Figure 4). The top five proteins structurally close to

the spike glycoprotein in the Protein Data Bank (as identified by TM-align) are listed in Table 1. In Table 2 the top five hits of closest Enzyme Commission (EC) numbers and active sites are listed.

Protein rankings are based on the structural alignment TM score in the PDB library between the query template and known structures. **RMSD[a]** the RMSD among structurally aligned residues of TM-align; **IDEN[a]** is the structurally related region's percentage sequence identity; **Cov** reflects the alignment range of the TM-alignment and is proportional to the sum by the length of query protein of structurally aligned residues. **5x58A**: Prefusion structure of SARS-CoV spike glycoprotein, conformation 1 (viral protein); **6nzkA**: Structural basis for human coronavirus attachment to sialic acid receptors (viral protein); **3aoiM**: RNA polymerase-Gfh1 complex (Crystal type 2), (transcription, transferase/DNA/RNA); **1ileA**: Isoleucyl-tRNA synthetase (aminoacyl-tRNA synthetase); **1ug9A**: Crystal Structure of Glucodextranase from Arthrobacter globiformis I42 (hydrolase).

1ileA: Isoleucyl-tRNA synthetase (aminoacyl-tRNA synthetase): **1k32A**: Crystal structure of the tricorn protease (hydrolase); **3eqlM**: Crystal structure of the *T. Thermophilus* RNA polymerase holoenzyme in complex with antibiotic myxopyronin (transferase); **2pdaA**: Crystal structure of the complex between pyruvate-ferredoxin oxidoreductase from *Desulfovibrio africanus* and pyruvate (oxidoreductase); **1ej6A**: Reovirus core (virus).

One powerful way of multiple secuence allihnemnt is the Multiple Alignment using Fast Fourier Transform (MAFFT) as shown in Figure 5 (a&b) below [8].

4. Discussion

In this study, we used the spike gene sequences from 21 (20+1) whole-genome sequences of SARS-CoV-2 collected from Jordan were retrieved from GISAID database and analyzed at the amino sequence level of the spike glycoprotein including a reference sequence of the surface glycoprotein [Severe acute respiratory syndrome coronavirus 2; (SARS-CoV-2)] own the accession number YP_009724390.1. Our findings showed that, the molecules which were structurally close to the spike glycoprotein from the Enzyme Commission (EC) numbers and active sites included Isoleucyl-tRNA synthetase, Crystal structure of the tricorn protease (hydrolase); Crystal structure of the *T. Thermophilus* RNA polymerase holoenzyme (transferase); Crystal structure of the complex between pyruvate-ferredoxin oxidoreductase from *Desulfovibrio africanus* and pyruvate (oxidoreductase); and Reovirus core (virus). All might explain the ability of SARS-CoV-2 in getting inside the human target cells.

The Four Amino Acid Variants (SAV) founded in 20 samples of SARS-CoV-2 were not conserved residues in spike glycoprotein. What is equal to 5% of samples showed tyrosine (polar, hydrophobic) deletion at *Y144* , 62% of samples showed aspartate (polar small hydrophilic charged (-)) substitution to glycine (nonpolar hydrophobic) at *D614G*, 5% of samples showed aspartate (polar small hydrophilic charged (-)) substitution to tyrosine (polar, hydrophobic & aromatic) at *D1139Y* and 5% of samples showed glycine (nonpolar hydrophobic) substitution to serine (polar) at *G1167S* respectively. The *D614G* substitution was

previously reported as a dominant mutation in Europe [9]. Our findings by using *Phyre2* have shown lower sensitive mutational and cannot affect the pocket region or alpha and beta-sheet.

The I-TASSER predicted three-dimensional structures for the monomer of S-protein of SARS-CoV-2 had similar stability structures for all of the four Amino Acid Variant (SAV) when we aligned the reference sequence of the spike glycoprotein YP_009724390.1 (SARS-CoV-2) with FASTA sequences of spike glycoproteins from Jordanian population, no change on the three-dimensional structure.

The generated three-dimensional structure of the spike protein of SARS-CoV-2 is consistent with a perfusion conformation structure reported in the literature [3]. Like any computational bioigy study, this study is limited with the capabilities of utilized servers and algorithms is it is highly dependent on the initial templates used for calculations, so if the initial template scoring is not good enough then this might affect the final output files.

5. Conclusion

This is the first study of its kind in the Middle East to predict the three-dimensional structure of the spike glycoprotein from SARS-CoV-2 of Jordanian specimens. In addition, we reported four amino acid variants, which might explain the low number of COVID-19 cases, 459 confirmed cases, 364 recovered, and eight deaths. However, the highest frequency mutation in our study, with 62% of samples showed aspartate substitution to glycine at D614G is consistent with other reports for samples were collected in Europe at the same time of our samples collection, March 2020. In this study, we consider the mutation D614G as the dominant local mutation in Jordan. We believe that the reported fur amino acid variants have collectively reduced the spike protein affinity of SARS-CoV-2 with ACE2 receptors in the Jordanian population and, most likely, the other Middle Eastern people. It is highly recommended to keep monitory the mutation rate of SARS-CoV-2 in Jordan in monthly bases with higher number of samples to fulfil statistical power. Some of the low percentage appeared mutations e.g 5% might be increased if the population size is higher.

Declarations

Author Contributions: “Conceptualization, W.A-Z., R.F & H.H.; methodology, W.A-Z & H.H.; software, W.A-Z & H.H.; validation, W.A-Z & H.H. and R.F.; formal analysis, W.A-Z & H.H.; investigation, W.A-Z & H.H.; resources, R.F.; data curation, W.A-Z & H.H.; writing—original draft preparation, W.A-Z & H.H.; writing—review and editing, W.A-Z & H.H.; visualization, W.A-Z & H.H.; supervision, R.F.; project administration, R.F.; funding acquisition, W.A-Z & H.H. All authors have read and agreed to the published version of the manuscript.”

Funding: “This research received no external funding”

Acknowledgments: The authors acknowledge Biolab Diagnostic Laboratories (Jordan) & Andersen lab at Scripps Research (USA) who published sequences were retrieved from GISAID, a maintained global

database based in Germany.

Conflicts of Interest: “The authors declare no conflict of interest.”

References

1. Wu F, Zhao S, Yu B, Chen Y, Wang W, Nature ZS-, et al. A new coronavirus associated with human respiratory disease in China. nature.com.
2. Lu R, Zhao X, Li J, Niu P, Yang, B., Wu, H., et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding. Lancet. 2020 Feb 22;395(10224):565–74.
3. Daniel Wrapp, Nianshuang Wang, Kizzmekia, S.; Corbett, Jory, A. Goldsmith C-LH, Olubukola Abiona, Barney, S. Graham JSM. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. - PubMed - NCBI. Science. 2020.
4. Lan J, Ge J, Yu J, Shan S, Zhou, H., Fan, S., et al. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature. 2020 Mar 30;1–6.
5. Zhang H, Penninger JM, Li Y, Zhong, N., Slutsky AS. Angiotensin-converting enzyme 2 (ACE2) as a SARS-CoV-2 receptor: Molecular mechanisms and potential therapeutic target. Intensive Care Med. 2020 Apr 1;46(4):586–90.
6. Glowacka I, Bertram S, Muller MA, Allen P, Soilleux E, Pfefferle, S., et al. Evidence that TMPRSS2 Activates the Severe Acute Respiratory Syndrome Coronavirus Spike Protein for Membrane Fusion and Reduces Viral Control by the Humoral Immune Response. J Virol. 2011 May 1;85(9):4122–34.
7. Abio Madeira F’, Mi Park Y, Lee J, Buso N, Gur T, Madhusoodanan N, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. Web Serv issue Publ online. 2019;47.
8. Waterhouse AM, Procter JB, Martin DMA, Clamp, M., Barton GJ. Jalview Version 2-A multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009;25(9):1189–91.
9. Angyal A, Brown RL, Carrilero L, Green LR, Groves DC, Johnson KJ, et al. Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2 on behalf of the Sheffield COVID-19 Genomics Group#, LaBranche CC2, and Montefiori DC2.
10. Yates CM, Filippis I, Kelley LA, Sternberg MJE. SuSPect: Enhanced prediction of single amino acid variant (SAV) phenotype using network features. J Mol Biol. 2014 Jul 15;426(14):2692–701.
11. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 2015 Jun 30;10(6):845–58.
12. Zhang, Y. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics. 2008 Jan 23;9.

Tables

Due to technical limitations, the tables are only available as a download in the supplemental files section.

Figures

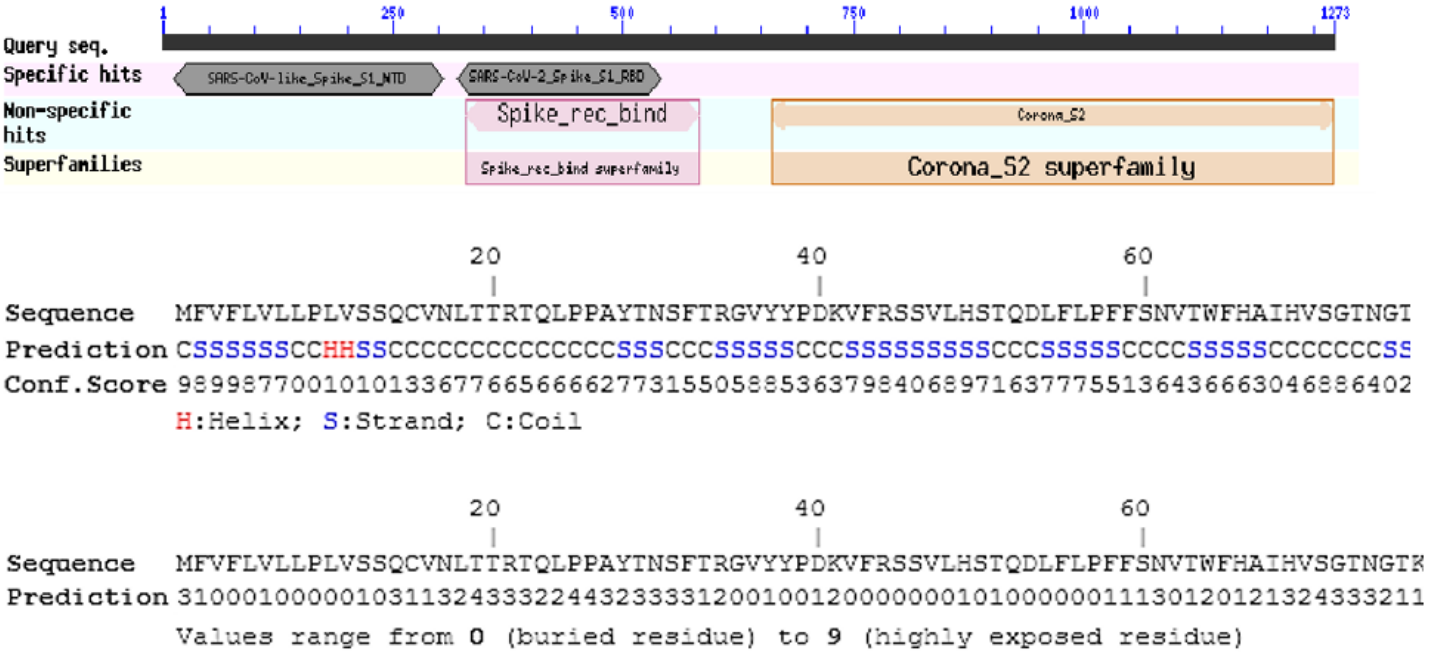

Figure 1

S-protein predicted secondary structure and predicted solvent accessibility.

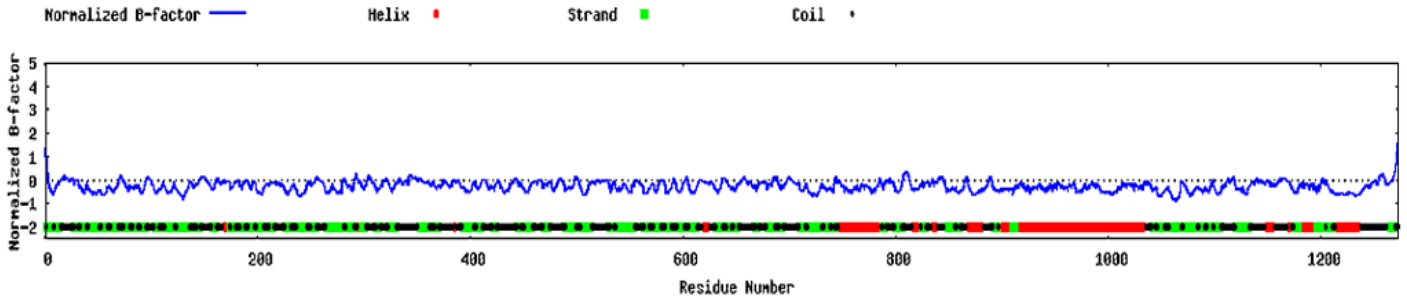

Figure 2

S-protein predicted normalized B-factor.

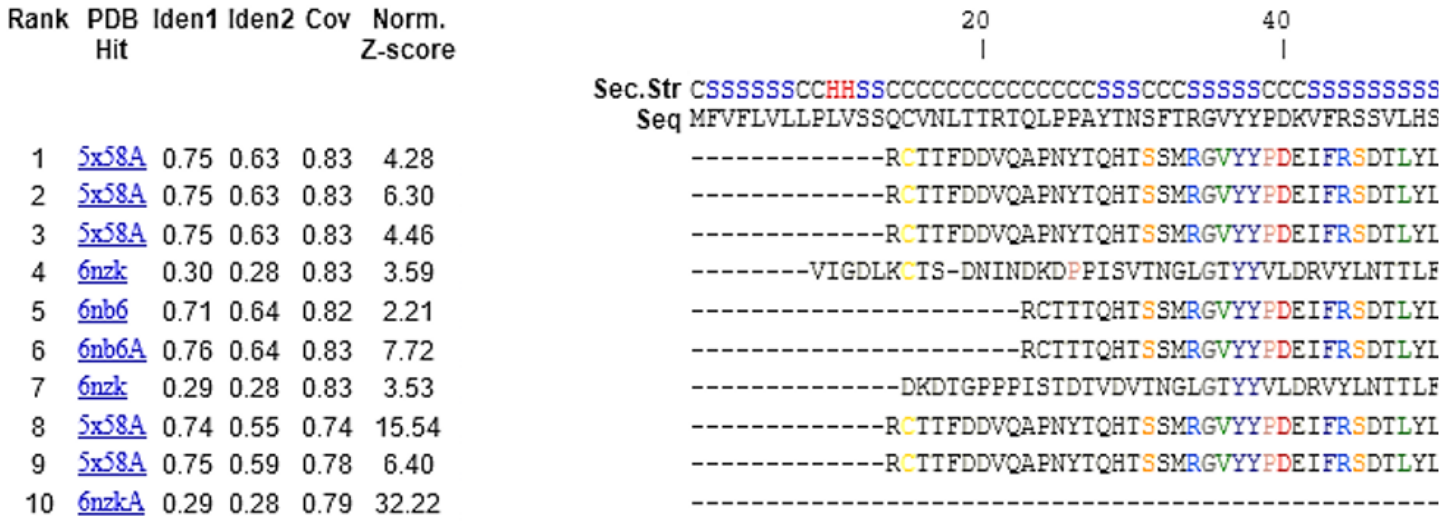

Figure 3

The top ten threading templates used by LOMETS server; 5 × 58A: Prefusion structure of SARS-CoV spike glycoprotein, conformation 1 (viral protein); 6nzkA: Structural basis for human coronavirus attachment to sialic acid receptors (viral protein); 6nb6: SARS-CoV complex with human neutralizing S230 antibody Fab fragment (state 1) (virus).

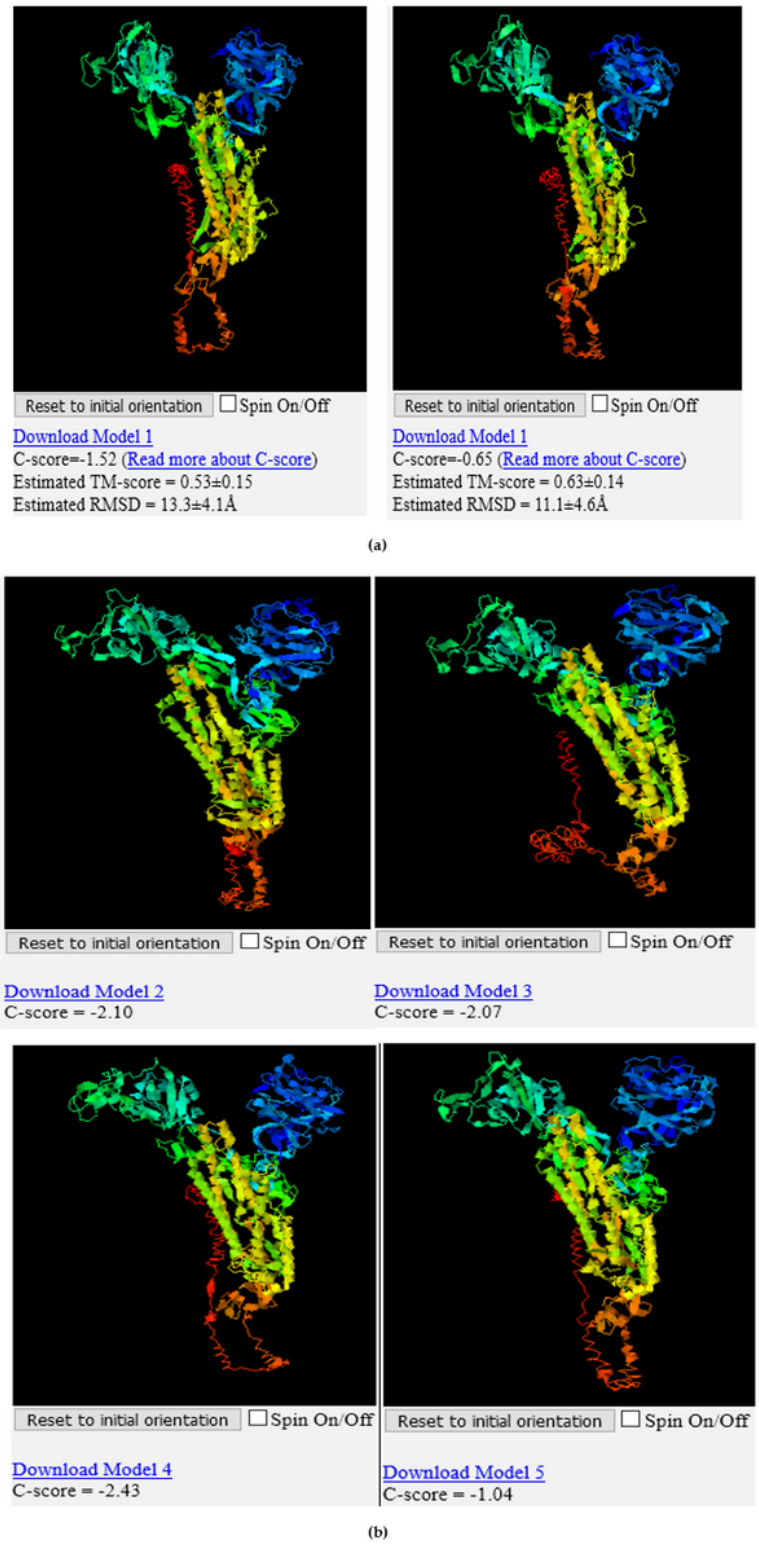

Figure 4

(a) : The final monomer models for (Wuhan) on the left side versus (Jordan) version on the right side of spike glycoproteins predicted by I-TASSER. (b): The top five final monomer models for (Jordanian) spike glycoproteins predicted by I-TASSER (the first model is in Figure 4 a).

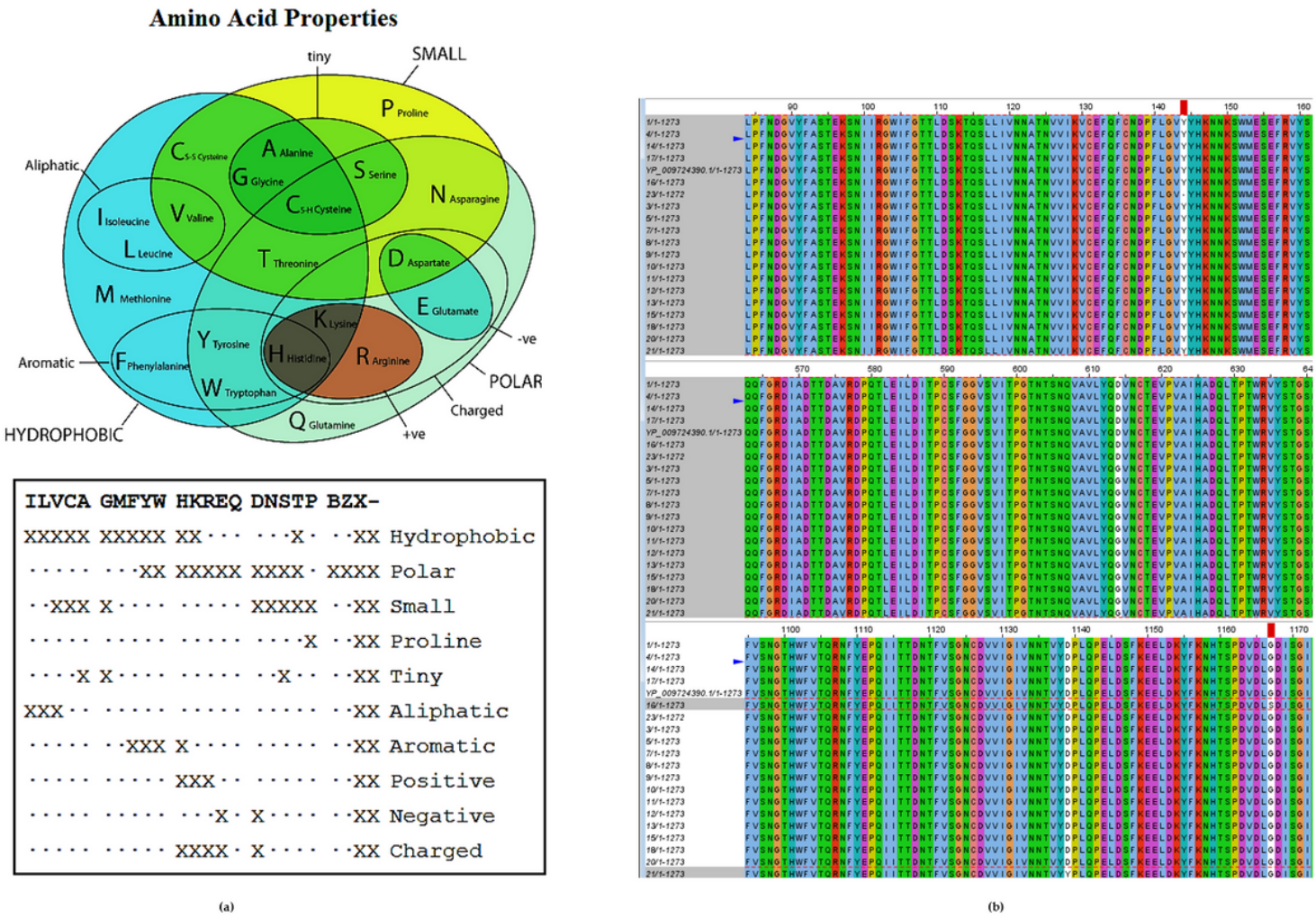

Figure 5

(a): General amino acid properties. (b): Multiple Sequence Alignment showing Amino Acid Variant (SAV) viewed by Jalview.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Tables.docx
- AppendixA.docx