Sys Rev Pharm 2020;11(12):1282-1289
A multifaceted review journal in the field of pharmacy

The Physicochemical and Phylogenetic Properties of Twelve Spike Glycoprotein Models For SARS-CoV-2 From China, Iran, And Tunisia

Samih Abed Odhaib[a], Miaad Jassim Mohammed[b], Dina A. Jamil[c], Hayder A Al-Aubaidy[d]

[a]Faihaa Teaching Hospital, Basrah, Iraq
[b]Al-Refaee General Hospital, Thi-Qar, Iraq
[c,d]School of Life Sciences, College of Science, Health and Engineering, La Trobe University, Bundoora, VIC, Australia 3086
Corresponding Author: Hayder Al-Aubaidy, Email: h.alaubaidy@latrobe.edu.au

ABSTRACT

Background: The coronavirus spike glycoprotein is a trimeric structural surface protein that facilitates the viral adhesion through attaching receptors on the human cell surface. This study aims to analyze and compare the genomic and phylogenetic properties of these spike glycoprotiens from China, Iran, and Tunisia.

Methods: This is a descriptive cross-sectional comparative study for the different properties of S glycoprotein from 12 SARS-CoV-2 specimens from GenBank. Clustal Omega was used to study model sequences alignment, residual conservation, phylogeny, and identity matrix. SWISS-MODEL developed and validated the 3D models for three protein sequences with the highest model quality. The different physicochemical characteristics of different models were assessed by ExPASy proteomics.

Results: The Chinese and the Iranian sequences share 100% identity, although they have a different amino acids number, and 25-29.27% identity to the Tunisian sequences. The 12 models are monophyletic, with varying stages of evolutionary divergence. There are six fully, three highly, and five lowly conserved residues across the sequences. The resulting three highly reliable 3D models were of different global qualities, being the lowest for the Tunisian, and the highest for the Iranian models. All the models are highly hydrophilic. The Tunisian models were unstable in comparison to the relatively stable other models with different physicochemical characteristics.

Conclusion: The models had different N-terminal residues and side group's polarity and charge. The S glycoprotiens are neither identical nor unique in neither model structure nor the physicochemical profiles in different parts of the world. The Tunisian models are drastically biodiversity from the Chinese and Iranian models.

Keywords: Coronavirus; phylogeny; protein sequences; SWISS-MODEL

Correspondance:
Hayder Al-Aubaidy
School of Life Sciences, College of Science, Health and Engineering, La Trobe University, Bundoora, VIC, Australia 3086
Email: h.alaubaidy@latrobe.edu.au

Background

SARS coronavirus (SARS-CoV-2) is a spherical pleomorphic enveloped particle, containing single-stranded (positive-sense) RNA associated with a nucleoprotein within a capsid comprised of matrix protein. The envelope bears spike (S) glycoprotein [1]. This S-glycoprotein is trimeric and consists of three S1-S2 heterodimers that determine the adhesion and viral virulence through its ability to attach different receptors with a distinctive affinity to the angiotensin-converting enzyme 2 (ACE-2) receptors on the human cell surface [2-3]. During viral infection, the S glycoprotein is cleaved into S1 and S2 subunits at a furin-like cleavage site through a cascade of proteolytic enzymes [4]. S1 subunit determines the virus-host range and cellular tropism through the receptor-binding domain (RBD) and N-terminal domaintobind directly to the ACE-2 receptor [5-6], which serve as a functional receptor for this virus [7]. S2 subunit acts as a class I viral fusion protein and mediates virus-cell membrane fusion by two heptads repeats [8]. The viral adhesion to the host cell involves a complex pre- to post-fusion conformation transition [3], and it is ten times tighter than the corresponding S protein of other Coronaviruses to their corresponding receptors [9]. The S glycoprotein is a surface protein and acts as the main inducer of neutralizing antibodies, T-cell response, and possible protective immunity, which make this protein a target for therapeutic intervention and vaccine development [10]. This study highlighted the genomic and phylogenetic properties of SARS-CoV-2 spike glycoprotein sequences from three countries China, Iran, and Tunisia.

Materials and Methods

This is a descriptive cross-sectional comparative study analyzing the genomic and phylogenetic properties of S glycoprotein from 12 SARS-CoV-2 specimens from China (seven isolates), Iran (two isolates), and Tunisia (three isolates). The genomic coding and amino acid sequences of the SARS-CoV-2 are available in the GenBank (**https://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs/**)

[11]. The identical proteins were identified by the (Identical Protein Group) to simplify the evaluation by choosing one descriptive representative sequence from each country. The (Multiple Sequence Alignment) at Clustal Omega (**https://www.ebi.ac.uk/Tools/msa/clustalo/**) of the European Molecular Biology Laboratory- European Bioinformatics Institute (EMBL-EBI) [12]was used to study amino acid sequences alignment of the 12 sequences. Clustal Omega tools can determine any conserved residues in the 12 models. The Phylogenetic Tree and Percent Identity Matrix were used to study the nature and the percentage of the similarity between the 12 sequences. The phylogenetic tree (phylogeny) is a diagram that depicts the lines of evolutionary descent in the protein sequences. A phylogenetic tree portrays the branching history of common ancestry, and the pattern of branching using clad grams and phylogram, the branching pattern reflects different evolutionary lineages [13-14]. We chose three amino acid sequences with the highest protein identities (QHN73805.1, QIQ08768.1, and QIV64962.1) from (China, Iran, and Tunisia) respectively, for developing and validating 3D sequence models using the SWISS-MODEL(**https://swissmodel.expasy.org/interactive#structure**), along with the structural assessment tools to confirm the structured model [15-17]. The Tunisian protein sequences (QIZ14987.1 and QIZ14988.1) of Tunisia are identical. Still, we could not include any of their sequences because they contain unidentified amino acids at an abnormal position, which makes them unsuitable for modeling by the SWISS-MODEL [15]. The sequence modeling for the chosen proteins was evaluated according to their local and global quality estimates by the use of:

- Global Model Quality Estimation (GMQE) reflects modeling accuracy and reliability. The resulting score is expressed as a number between 0 and 1; the higher the number was, the more the reliability was [18].
- Qualitative Model Energy Analysis (QMEAN): The composite estimator of the different geometrical properties of the protein sequence that provides both global (for the entire structure) and local (per residue) absolute quality estimates based on one single model [18, 19].
- The QMEAN Z-score estimates the "degree of nativeness" of the structural features observed in the model on a global scale. The QMEAN Z-score compares the model QMEAN scores to the scores of the similar size expected experimental structures, through Cβ atoms only, all atoms, the solvation potential, and the torsion angle potential [18-19].
- The QMEAND is Co enhances the accuracy of the QMEAN local scores by assessing the interatomic distance of the target model against the already ensemble information of the experimentally determined proteins that share homology to the target model [19].

Prot Param tool at the ExPASy proteomics server (https://web.expasy.org/protparam/) was used to study and computes various physicochemical properties of a protein sequence [20]. It provides information about different indices and scores:

- Molecular weight.
- The number and charges of different amino acid residues.
- The chemical formula.
- Tryptophan residues availability.
- Site and type of N-terminal residues.
- Extinction coefficients indicate how much light a protein absorbs at a particular wavelength, which is useful during the purification process. Optical density (Absorbance) which is the product of dividing the extinction coefficient by the molecular weight [21].
- In vivo half-life.
- Instability index (II) which provides an estimate of the stability of your protein in a test tube. The (II) of stable proteins is <40, while for unstable proteins, the (II)>40.
- The aliphatic index is the relative volume occupied by aliphatic side chains (alanine, valine, Isoleucine, and leucine) [22].
- Grand Average of Hydropath (GRAVY) marks the protein as hydrophilic (negative values) or hydrophobic (positive values) [23].

The Protein Molecular Weight – The Sequence Manipulation Suite at Bioinformatics.org (**https://www.bioinformatics.org/sms/prot_mw.html**) was used to compare the actual molecular weight of different protein sequences, without including any unknown amino acids [24].

Results

Table 1 demonstrates the GenBank accession numbers of AA sequences of the 12 S glycoprotiens of the SARS-CoV-2 from China, Iran, and Tunisia (this represents the whole worldwide plotted sequences until April 20th, 2020). The Tunisian sequences had the most abundant amino acid number in comparison to the Iranian and Chinese sequences. The sequences (QIZ14987.1 and QIZ14988.1) are identical, and they contain an abnormal amino acid at an ambiguous position. The Identical Protein Group provides information about the identical protein sequences from each country. All the Chinese sequences are shown in Table 1, were isolated on the same day (February 11th, 2020) from the (Wuhan seafood market pneumonia virus), from the nasopharyngeal swab, serum, throat swab, and sputum. All other sequences from Iran and Tunisia were taken from the nasopharyngeal swabs. Figure 1 demonstrates the multiple sequence alignment of different sequences. There are six fully, three highly, and five lowly conserved residues. The Chinese and the Iranian sequences share 100% similarity, while the Tunisian sequences had 25 and 29.27% similarity to the Chinese and the Iranian sequences, respectively. The first Tunisian sequence described (QIV64962.1) had 99.39% similarity to the other two sequences (QIZ14987.1 and QIZ14988.1) because of the unidentified amino acid at the ambiguous position, table (2), and figure (1). The first divergence event separated the lineage that gave rise to the (China-Shenzhen 7) sequence from a lineage of the other six Chinese sequences and a lineage that gave rise to the Iranian and Tunisian sequences, i.e., the 12 sequences shared a common ancestry (monophyletic group). The (China-Shenzhen 1 and 2) share a more recent common ancestor than either share with other sequences, i.e., they are therefore more closely related to each other than either is to other sequences. The (Tunisia-Tunis 2 and 3) sequences underwent recent evolutionary divergence that occurred (Supplementary Figure 1 A and B). Because all the Chinese and Iranian sequences are of equal distance (in terms of branch arrangement) from their original ancestor, we could say that these sequences are equally related to each other. The three Tunisian sequences underwent the most recent divergence from the original evolutionary lineage. The 3D models of the S glycoprotiens from the three chosen sequences from the three countries were highly reliable models with a very high GMQE score. Still, they demonstrated that the very low global quality, very low QMEAN-scored Tunisian S glycoprotein had three chains in comparison to the Chinese sequence of medium global quality and the Iranian sequences of the highest global quality between the three countries. The Z-scores of all three sequences fall within the range between

(0.5-0.7) (Figure 2 and Supplementary Figure 2). The Tunisian sequences had the highest molecular weight, extinction index, and instability index (II) between the chosen sequences, which renders the Tunisian sequences unstable in comparison to both the Chinese and the Iranian sequences, which are relatively stable at different degrees. All the sequences are hydrophilic and share a high aliphatic index, table (3).

The N-terminals which signify the beginning of the S1 subunit of the S glycoprotein in SARS-CoV-2 are different in their chemical profiles:

- For the Chinese sequences, it is N (Asparagines), which is polar with a positively charged side group.
- For the Iranian sequences, it is P (Proline), which is no polar with an uncharged side group.
- For the Tunisian sequences, it is Q (Glutamine), which is polar with an uncharged side group.

These biophysical changes affect the half-life of the protein models in vitro and in vivo (Table 3).

Discussion

The coronavirus S glycoprotein is an important structural protein that is responsible for the phenotype of crown-like shape viral particles, from which the original name "coronavirus" was coined [4]. The GenBank provides unrestricted access to all types of genomic and polypeptide sequences that dealt with the novel coronavirus epidemic by SARS-CoV-2 [11], till the preparation of this paper (April 20th, 2020), only the Chinese sequences of S glycoprotein of SARS-CoV-2 were assessed and compared through the literature [25-27]. The Iranian and the Tunisian sequences were not assessed after being uploaded in the GenBank. The three Tunisian models represent the largest S glycoprotein unit, in comparison to the Chinese and Iranian S glycoprotiens, with their higher number of base pairs and a consequently higher number of amino acid residues, even though the assessment of the Percent Identity Matrix by Clustal 2.1, revealed a 100% similarity or identity between the Chinese and the Iranian sequences, although the numbers of amino acid residues are different. Two out of the three Tunisian sequences (QIZ14987.1 and QIZ14988.1) were identical 100% but share a 25% identity with the Chinese and Iranian sequences. The presence of an unknown amino acid (X) at an ambiguous position in these sequences minimally decreased the identity of the sequences (QIZ14987.1 and QIZ14988.1) to the third Tunisian sequence (QIV64962.1) by 0.61% to be 99.39%. On the other hand, the Tunisian sequence (QIV64962.1) shared a 29.27% identity with the Chinese and Iranian sequences. Shanker and colleagues had studied the amino acid sequences of the seven S glycoprotein units from China and reached the same conclusion [27]. The N-terminal identifies the start of the S1 subunit in the three modeled sequences. There was a different polarity for the N-terminal residues, being polar with a positively charged side group in the Chinese model, no polar with uncharged side group in the Iranian model and polar with uncharged side group in the Tunisian model [4-6]. The S glycoprotein remains uncleaved at the S1/S2 site during virus packaging in cells. The trimetric S-protein is processed at the S1/S2 furin-like cleavage site by host cell proteases, during infection [28]. We combined the SWISS-MODEL and the Clustal Omega modeling to find the possible furin-like S1/S2 cleavage site that was thoroughly described by Coutard and colleagues at the monobasic R residue [4], as marked in Figure 1. Following cleavage or priming, the protein is divided into an N-terminal S1-ectodomain that recognizes a cognate cell surface receptor to aid trafficking into and hijacking the host cell, and a C-terminal S2 membrane-anchored protein involved in viral entry [4, 6].The cleavage at the furin-like cleavage site occurs during virus egress for S-protein priming and may provide a gain-of-function to the virus for efficient pathogenesis compared to other betacoronaviruses lineages [4]. The SWISS-MODEL provides global and local quality assessment for the modeled sequences. The models were highly reliable, with very high GMQE scores. Still, the global and local quality scores are different even after normalization (Figure 2), that further affect the biochemical profiles of the models, table (3). The highly descriptive Prot Param tool of ExPASy provides some answers to such different quality profiles between models. All the models had no Tryptophan residues that could result in more than 10% error in the computed extinction coefficient. The extinction coefficients of the Tunisian and Iranian models were much higher compared to the Chinese models but with declining absorbance scores that are related to their molecular weight. The extinction coefficient is important when studying protein-protein and protein-ligand interactions [20]. The Iranian models had more than 20 hours' half-life in both mammalian reticulocytes in vitro and the yeast in vivo, in comparison to other models, table (3), which may be related to the no polar N-terminal (Proline) that attached the uncharged side group. The in vivo half-life predicts the elapsed time it takes for half of the protein amount in a cell to disappear after its synthesis. Prot Param relies on the (N-end rule), which relates the half-life of a protein to the N-terminal residue identity that affects the stability of the protein in vivo; the prediction is given for 3 model organisms (human, yeast and E. coli) [20]. All models share a high aliphatic index that may contribute as a positive factor to increase the relative thermo stability of the globular proteins [22]. All the models were hydrophilic, which is reflected by the highly negative GRAVY score [23], with the Tunisian model, the most hydrophilic one.

Conclusions

The S glycoproteins are not identical nor unique in model structure nor the physicochemical profiles in different parts of the world, with a special emphasis on the Tunisian S glycoprotein models, which had drastic biodiversity from the Chinese and Iranian models. Understanding spike protein models is fundamental to an understanding of viral pathogenesis, which will allow for additional protein-engineering efforts that could improve anti-genicity and protein expression for pharmaceutical development.

Abbreviations

- ACE 2: Angiotensin-converting enzyme 2.
- EMBL-EBI: European Molecular Biology Laboratory-European Bioinformatics Institute.
- GMQE: Global Model Quality Estimation.
- QMEAN: Qualitative Model Energy Analysis.
- SARS-COV-2: SARS coronavirus.

Competing interests

The authors declare no competing interests financial or otherwise related to this project.

Funding

This study received no specific funding.

Acknowledgment

The authors would like to thank all the international researchers and scientists who pioneered the work of sequencing the SARS-CoV-2 genome, who were greatly influential in the research field.

References

1. de Haan CA, Kuo L, Masters PS, Vennema H, Rottier PJ. Coronavirus particle assembly: primary structure requirements of the membrane protein. J Virol 1998 Aug; 72(8): 6838–50. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC109893/. PMID: 9658133; PMCID: PMC109893.
2. Tortorici MA, Veesler D. Structural insights into coronavirus entry. Adv Virus Res 2019; 105: 93–116. https://dx.doi.org/10.1016%2Fbs.aivir.2019.08.002, Epub 2019 Aug 22. PMID: 31522710; PMCID: PMC7112261.
3. Song W, Gui M, Wang X, Xiang Y. Cryo-EM structure of the SARS coronavirus spike glycoprotein in complex with its host cell receptor ACE2. PLoS Pathog 2018; 14 (8): e1007236. https://doi.org/10.1371/journal.ppat.1007236.
4. Coutard B, Valle C, de Lamballerie X, Canard B, Seidah NG, Decroly E. The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. Antiviral Res 2020 Apr; 176: 104742. https://dx.doi.org/10.1016%2Fj.antiviral.2020.104742. Epub 2020 Feb 10. PMID: 32057769; PMCID: PMC7114094.
5. Li F, Li W, Farzan M, Harrison SC. Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. Science. 2005 Sep 16; 309 (5742): 1864–8. https://doi.org/10.1126/science.1116480. PMID: 16166518.
6. Millet JK, Kien F, Cheung CY, Siu YL, Chan WL, Li H, et al. Ezrin interacts with the SARS Coronavirus spike protein and restrains infection at the entry stage. PLoS ONE 2012; 7 (11): e49566. https://doi.org/10.1371/journal.pone.0049566.
7. Li W, Moore MJ, Vasilieva N, Sui J, Wong SK, Berne MA, et al. Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. Nature 2003; 426 (6965): 450–4. https://dx.doi.org/10.1038%2Fnature02145. PMID: 14647384; PMCID: PMC7095016.
8. Xia S, Zhu Y, Liu M, Lan Q, Xu W, Wu Y, et al. Fusion mechanism of 2019-nCoV and fusion inhibitors targeting HR1 domain in spike protein. Cell Mol Immunol 2020. https://doi.org/10.1038/s41423-020-0374-2.
9. Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh CL, Abiona O, et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science 2020; 367 (6483): 1260–3. https://science.sciencemag.org/content/367/6483/1260.
10. Du L, He Y, Zhou Y, Liu S, Zheng BJ, Jiang S. The spike protein of SARS-CoV-a target for vaccine and therapeutic development. Nat Rev Microbiol 2009 Mar; 7 (3): 226–36. https://dx.doi.org/10.1038%2Fnrmicro2090 . Epub 2009 Feb 9. PMID: 19198616; PMCID: PMC2750777.
11. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. GenBank. Nucleic Acids Res 2013 Jan; 41 (Database issue): D36–42. https://dx.doi.org/10.1093%2Fnar%2Fgks1195. Epub 2012 Nov 27. PMID: 23193287; PMCID: PMC3531190.
12. Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res 2019 Jul 2; 47 (W1): W636–W641. https://dx.doi.org/10.1093%2Fnar%2Fgkz268. PMID: 30976793; PMCID: PMC6602479.
13. Baum DA, Smith SDW, Donovan SSS. The tree-thinking challenge. Science 2005; 310 (5750): 979–98. https://science.sciencemag.org/content/310/5750/979.short/.
14. Avise JC. (2006). Evolutionary Pathways in Nature: A Phylogenetic Approach. Evolutionary Pathways in Nature: A Phylogenetic Approach. Cambridge University Press. https://www.cambridge.org/iq/academic/subjects/life-sciences/genetics/evolutionary-pathways-nature-phylogenetic-approach?format=PB&isbn=9780521674171.
15. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res 2018 Jul; 46 (W1): W296–W303. https://doi.org/10.1093/nar/gky427. PMID: 29788355; PMCID: PMC6030848.
16. Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, et al. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res 2014 Jul; 42 (Web Server issue): W252–8. https://dx.doi.org/10.1093%2Fnar%2Fgku340. Epub 2014 Apr 29. PMID: 24782522; PMCID: PMC4086089.
17. Schwede T, Kopp J, Guex N, Peitsch MC. SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res 2003 Jul; 31 (13): 3381–5. https://dx.doi.org/10.1093%2Fnar%2Fgkg520. PMID: 12824332; PMCID: PMC168927.
18. Benkert P, Tosatto SCE, Schomburg D. QMEAN: A comprehensive scoring function for model quality assessment. Proteins 2008; 71. 261–77.https://doi.org/10.1002/prot.21715.
19. Studer G, Rempfer C, Waterhouse AM, Gumienny R, Haas J, Schwede T. QMEANDisCo—distance constraints applied on model quality estimation. Bioinformatics 2020; 36 (6): 1765–71. https://doi.org/10.1093/bioinformatics/btz828.
20. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, et al. Protein identification and analysis tools on the ExPASy server; (In) John M. Walker (ed): The proteomics protocols handbook, Humana Press (2005). pp. 571–607. https://doi.org/10.1385/1-59259-890-0:571.
21. Gill SC and von Hippel PH. Calculation of protein extinction coefficients from amino acid sequence data. Anal Biochem 1989; 182: 319-26. https://doi.org/10.1016/0003-2697(89)90602-7.
22. Ikai A. Thermostability and aliphatic index of globular proteins. Biochem J 1980; 88 (6): 1895–98. https://www.jstage.jst.go.jp/article/biochemistry1922/88/6/88_6_1895/_pdf/-char/en.
23. Kyte J and Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982; 157: 105–32. https://doi.org/10.1016/0022-2836(82)90515-0.
24. Stothard P. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. Biotechniques 2000; 28: 1102–4. https://doi.org/10.2144/00286ir01.
25. Chan JFW, Yuan S, Kok KH, To KKW, Chu H, Xing JYF, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. Lancet 2020; 395: 514–23. https://doi.org/10.1016/S0140-6736(20)30154-9.
26. Vankadari N and Wilce JA. Emerging WuHan (COVID-19) coronavirus: glycan shield and structure prediction of spike glycoprotein and its interaction with human CD26.

Emerg Microbes Infect 2020; 9 (1): 601–4. https://doi.org/10.1080/22221751.2020.1739565.

27. Shanker AK, Divya B, Anjani A. Whole genome sequence analysis and homology modelling of a 3C like peptidase and a non-structural protein 3 of the SARS-CoV-2 shows protein ligand interaction with an Aza-peptide and a noncovalent lead inhibitor with possible antiviral properties. ChemRxiv 2020. Preprint. https://doi.org/10.26434/chemrxiv.11846943.v7.

28. Song HC, Seo MY, Stadler K, Yoo BJ, Choo QL, Coates SR, et al. Synthesis and characterization of a native, oligomeric form of recombinant severe acute respiratory syndrome coronavirus spike glycoprotein. J Virol 2004 Oct; 78(19): 10328–35. https://dx.doi.org/10.1128%2FJVI.78.19.10328-10335.2004. PMID: 15367599; PMCID: PMC516425.

Tables

Table 1: The GenBank data for the amino acid sequences of 12 S glycoprotein of SARS-CoV-2 isolates from China, Iran, and Tunisia

	GenBank Accession	**Surface Glycoprotein (S) Gene CDS Definition**	**Base Pairs[1]**	**Protein ID**	**Amino Acids**	**Collection Date[5]**	**Isolation Source**	**Locality**
1	MN938387.1	2019-nCoV_HKU-SZ-001_2020[2]	107	QHN73805.1	35	2020-01	nasopharyngeal swab	China: Shenzhen 1
2	MN938388.1	2019-nCoV_HKU-SZ-002b_2020[2]	107	QHN73806.1	35	2020-01	Serum	China: Shenzhen 2
3	MN938389.1	2019-nCoV_HKU-SZ-004_2020[2]	107	QHN73807.1	35	2020-01	nasopharyngeal swab	China: Shenzhen 3
4	MN938390.1	2019-nCoV_HKU-SZ-005_2020[2]	107	QHN73808.1	35	2020-01	Throat swab	China: Shenzhen 4
5	MN975266.1	2019-nCoV_HKU-SZ-007a_2020[2]	107	QHN73822.1	35	2020-01	nasopharyngeal swab	China: Shenzhen 5
6	MN975267.1	2019-nCoV_HKU-SZ-007b_2020[2]	107	QHN73823.1	35	2020-01	Throat swab	China: Shenzhen 6
7	MN975268.1	2019-nCoV_HKU-SZ-007c_2020[2]	107	QHN73824.1	35	2020-01	Sputum	China: Shenzhen 7
8	MT232871.1	SARS-CoV-2/human/IRN/MHKN-1/2020[3]	157	QIQ08768.1	52	2020-02-26	nasopharyngeal swab	Iran: Tehran 1
9	MT232872.1	SARS-CoV-2/human/IRN/MHKN-2/2020[3]	158	QIQ08769.1	52	2020-02-26	nasopharyngeal swab	Iran: Tehran 2
10	MT308701.1	SARS-CoV-2/human/TUN/Tunis7266/2020	493	QIV64962.1	163	2020-04-02	nasopharyngeal swab	Tunisia: Tunis 1
11	MT324679.1	SARS-CoV 2/human/TUN/Tunis_6401/2	491	QIZ14987.1	163	2020-03-29	nasopharyngeal swab	Tunisia: Tunis 2

		2020[4]						
12	MT324680.1	SARS-CoV-2/human/TUN/Tunis_7643/2020[4]	491	QIZ14988.1	163	2020-04-03	nasopharyngeal swab	Tunisia: Tunis 3

Abbreviations: CDS, coding sequence; HKU, Hong Kong University; ID, identity; IRN, Iran; kDa, kilodalton; MHKN, Mohammad Hadi Karbalaie Niya; MW, molecular weight; nCoV, novel coronavirus; SARS, severe acute respiratory syndrome; SZ, Shenzhen; TUN, Tunisia.

[1] All the original sequencing uses Sanger Dideoxy Sequencing Technique

[2] Identical proteins from (Wuhan seafood market pneumonia virus) from China

[3] Identical proteins from (Severe acute respiratory syndrome coronavirus 2) from Iran

[4] Identical proteins from (Severe acute respiratory syndrome coronavirus 2) from Tunisia

[5] The order of arrangement in the collection date is the order of appearance in the GenBank

Table 2: Percent Identity Matrix created by Clustal 2.1 describes the percentage of similarity between the 12 amino acid sequences

		China 1	**China 2**	**China 3**	**China 4**	**China 5**	**China 6**	**China 7**	**Iran 1**	**Iran 2**	**Tunisia 1**	**Tunisia 2**	**Tunisia 3**
	Protein ID	**QHN73805.1**	**QHN73806.1**	**QHN73807.1**	**QHN73808.1**	**QHN73822.1**	**QHN73823.1**	**QHN73824.1**	**QIQ08768.1**	**QIQ08769.1**	**QIV64962.1**	**QIZ14987.1**	**QIZ14988.1**
1	QHN73805.1	100	100	100	100	100	100	100	100	100	25	25	25
2	QHN73806.1	100	100	100	100	100	100	100	100	100	25	25	25
3	QHN73807.1	100	100	100	100	100	100	100	100	100	25	25	25
4	QHN73808.1	100	100	100	100	100	100	100	100	100	25	25	25
5	QHN73822.1	100	100	100	100	100	100	100	100	100	25	25	25
6	QHN73823.1	100	100	100	100	100	100	100	100	100	25	25	25
7	QHN73824.1	100	100	100	100	100	100	100	100	100	25	25	25
8	QIQ08768.1	100	100	100	100	100	100	100	100	100	29.27	29.27	29.27
9	QIQ08769.1	100	100	100	100	100	100	100	100	100	29.27	29.27	29.27
10	QIV64962.1	25	25	25	25	25	25	25	29.27	29.27	100	99.39	99.39
11	QIZ14987.1	25	25	25	25	25	25	25	29.27	29.27	99.39	100	100
12	QIZ14988.1	25	25	25	25	25	25	25	29.27	29.27	99.39	100	100

Table 3: The comparison of the biochemical profiles of the amino acid sequences of 12 S glycoprotiens of the SARS-CoV-2 by the use of the Prot Param tool of ExPASy

Parameters		**China 1-7[1]**	**Iran 1-2[1]**	**Tunisia 1[1]**	**Tunisia 2-3[1]**
Molecular weight (kDa)		3.93	5.76	17.67	17.67[2]
Number of negatively charged amino acids (Asp + Glu)		6	7	13	12
Number of positively charged amino acids (Arg + Lys)		4	5	15	15
Chemical formula		$C_{174}H_{267}N_{47}O_{57}$	$C_{257}H_{396}N_{66}O_{80}S_2$	$C_{777}H_{1252}N_{222}O_{238}S_5$	NA
Total number of atoms		545	801	2494	NA
Tryptophan residue availability		No	No	No	No
Extinction coefficients (M^{-1} cm^{-1})		4470	4595[3]	4595[3]	4595[3]
Absorbance Abs 0.1% (=1 g/l)		1.138	0.799	0.26	0.26
N-terminal[4]		N (Asn)	P (Pro)	Q (Gln)	Q (Gln)
The estimated half-life	Mammalian reticulocytes, in vitro	1.4 h	>20 h	0.8 h	0.8 h
	Yeast, in vivo	3 min	>20 h	10 min	10 min
	Escherichia coli, in vivo	>10 h	NA	10 h	10 h
The instability index (II)		17.20 (Stable)	5.79 (Stable)	49.67 (Unstable)	47.19 (Unstable)
Aliphatic index		78.00	82.50	94.54	94.54
GRAVY score		-0.671	-0.188	-0.050	-0.029

Abbreviations: h, hours; GRAVY, Grand average of hydropathicity; kDa, kilodaltons; min, minutes; NA, not available.

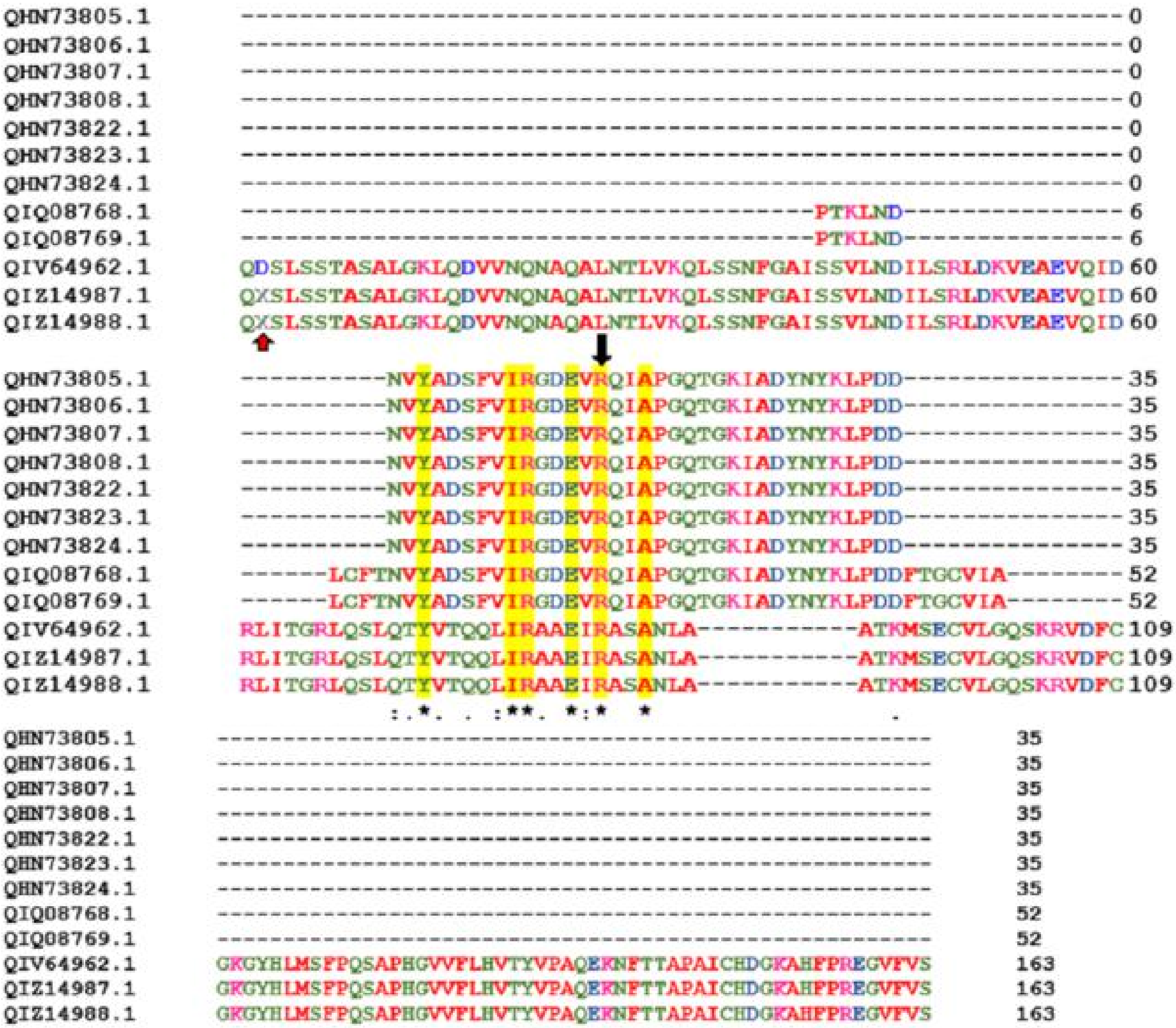
[1] China 1-7 includes identical sequences (QHN73805.1, QHN73806.1, QHN73807.1, QHN73808.1, QHN73822.1, QHN73823.1, QHN73824.1). Iran 1-2 includes identical sequences (QIQ08768.1 and QIQ08769.1). Tunisia 1 includes the sequence (QIV64962.1), which shares no identity. Tunisia 2-3 includes identical sequences (QIZ14987.1 and QIZ14988.1).

[2] The molecular weight is calculated for 163 amino acids using ExPASy is 17.67 kDa. Protein Molecular Weight - Bioinformatics.org tool calculates the molecular weight for 162 amino acids only to be (17.56 kDa), excluding the unknown amino acid.

[3] Assuming all pairs of Cys residues form cystines. While if assuming all Cys residues are reduced, the extinction coefficient will be 4470 M^{-1} cm^{-1}, and the absorbance will be 0.777 for the Iranian models (1-2), and 0.253 for all the Tunisian models.

[4] The N-terminal of the Chinese models is polar with a positively charged side group. The Iranian models had nonpolar N-terminal with an uncharged side group. The N-terminal residues of the Tunisian models are polar with an uncharged side group.

Figures



```
QHN73805.1      ------------------------------------------------------------ 0
QHN73806.1      ------------------------------------------------------------ 0
QHN73807.1      ------------------------------------------------------------ 0
QHN73808.1      ------------------------------------------------------------ 0
QHN73822.1      ------------------------------------------------------------ 0
QHN73823.1      ------------------------------------------------------------ 0
QHN73824.1      ------------------------------------------------------------ 0
QIQ08768.1      ---------------------------------------PTKLND--------------- 6
QIQ08769.1      ---------------------------------------PTKLND--------------- 6
QIV64962.1      QDSLSSTASALGKLQDVVNQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQID 60
QIZ14987.1      QXSLSSTASALGKLQDVVNQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQID 60
QIZ14988.1      QXSLSSTASALGKLQDVVNQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQID 60

QHN73805.1      ----------NVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDD--------------- 35
QHN73806.1      ----------NVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDD--------------- 35
QHN73807.1      ----------NVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDD--------------- 35
QHN73808.1      ----------NVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDD--------------- 35
QHN73822.1      ----------NVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDD--------------- 35
QHN73823.1      ----------NVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDD--------------- 35
QHN73824.1      ----------NVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDD--------------- 35
QIQ08768.1      ------LCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIA-------- 52
QIQ08769.1      ------LCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIA-------- 52
QIV64962.1      RLITGRLQSLQTYVTQQLIRAAEIRASANLA-----------ATKMSECVLGQSKRVDFC 109
QIZ14987.1      RLITGRLQSLQTYVTQQLIRAAEIRASANLA-----------ATKMSECVLGQSKRVDFC 109
QIZ14988.1      RLITGRLQSLQTYVTQQLIRAAEIRASANLA-----------ATKMSECVLGQSKRVDFC 109
                          :.*. . :**. *:*  *                .

QHN73805.1      ------------------------------------------------------     35
QHN73806.1      ------------------------------------------------------     35
QHN73807.1      ------------------------------------------------------     35
QHN73808.1      ------------------------------------------------------     35
QHN73822.1      ------------------------------------------------------     35
QHN73823.1      ------------------------------------------------------     35
QHN73824.1      ------------------------------------------------------     35
QIQ08768.1      ------------------------------------------------------     52
QIQ08769.1      ------------------------------------------------------     52
QIV64962.1      GKGYHLMSFPQSAPHGVVFLHVTYVPAQERNFTTAPAICHDGKAHFPREGVFVS     163
QIZ14987.1      GKGYHLMSFPQSAPHGVVFLHVTYVPAQERNFTTAPAICHDGKAHFPREGVFVS     163
QIZ14988.1      GKGYHLMSFPQSAPHGVVFLHVTYVPAQERNFTTAPAICHDGKAHFPREGVFVS     163
```

Figure 1: Multiple sequence alignment of the amino acid sequences for 12 S glycoprotiens of SARS-CoV-2 from China (seven sequences), Iran (two sequences), and Tunisia (three sequences) using Clustal Omega

The red arrow marks the ambiguous sites of the unidentified amino acid of the Tunisian sequences. Asterisks represent fully conserved residues, colons represent highly conserved residues, and periods represent lowly conserved residues. The black arrow marks the possible S1/S2 furin-like cleavage site.

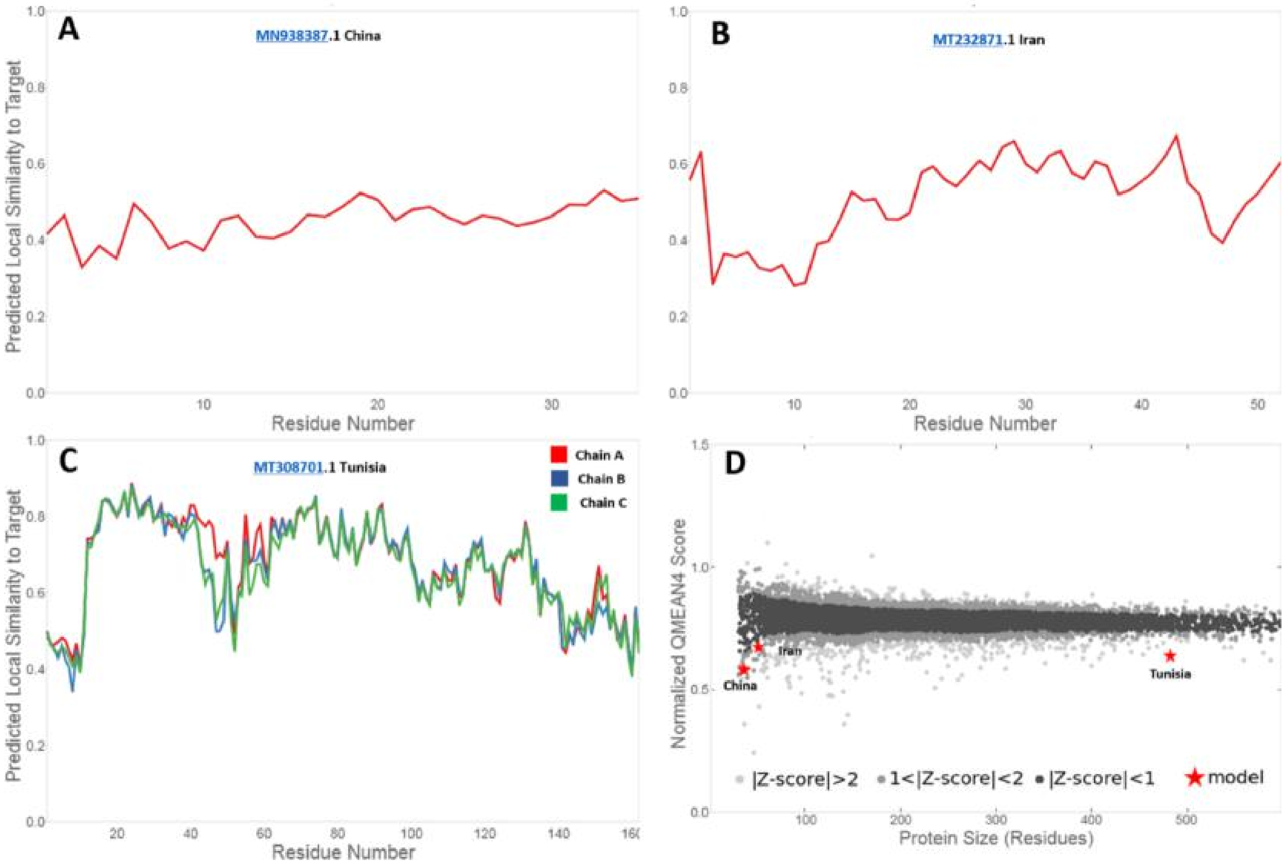


Figure 2: Local quality estimates comparison of the amino acid sequences of the three S glycoprotiens in China, Iran, and Tunisia

For panels (A, B, and C), the x-axis shows protein length (number of residues), for panel (D), the y-axis is the normalized QMEAN score. Every dot in panel (D) represents one experimental protein structure. The red stars represent the actual models normalized QMEAN score.