A theoretical analysis of the putative ORF10 protein in SARS-CoV-2

Noah A. Schuster[1*]

[1]Department of Biology, DePauw University, Greencastle, IN

***Correspondence: noahschuster_2021@depauw.edu; +1 (219)-336-5808**

Abstract

Found just upstream of the 3'-untranslated region in the SARS-CoV-2 genome is the putative ORF10 which has been proposed to encode for the hypothetical ORF10 protein. Even though current research suggests this protein is not likely to be produced, further investigations into this protein are still warranted. Herein, this study uses multiple bioinformatic programs to theoretically characterize and construct the ORF10 protein in SARS-CoV-2. Results indicate this protein is mostly ordered and hydrophobic with high protein-binding propensity, especially in the N-terminus. Although minimal, an assessment of twenty-two missense mutations for this protein suggest slight changes in protein flexibility and hydrophobicity. When compared against two other protein models, this study's model was found to possess higher quality. As such, this model suggests the ORF10 protein contains a β-α-β motif with a β-molecular recognition feature occurring as the first β-strand. Furthermore, this protein also shares a strong phylogenetic relationship with other putative ORF10 protein's in closely related coronaviruses. Despite not yielding evidence for the existence of this protein within SARS-CoV-2, this study does present theoretical examinations that can serve as platforms to drive additional experimental work that assess the biological relevance of this hypothetical protein in SARS-CoV-2.

1. Introduction

An initial outbreak of coronavirus disease 2019 (COVID-19) in Wuhan, China has resulted in a massive global pandemic causing, at the time this manuscript was prepared, over 37,000,000 confirmed cases and 1,070,000 deaths.[1] The virus responsible, SARS-CoV-2, contains a single-stranded positive-sense RNA (+ssRNA) genome that is ~29.8 kilobases in length.[2] In particular, the 3'-terminus contains several ORFs encoding four main structural proteins: the spike (S), membrane (M), envelope (E), and nucleocapsid (N) proteins.[3] In addition, many shorter length ORFs have been detected and proposed to encode for approximately nine different accessory proteins (3a, 3b, 6, 7a, 7b, 8, 9b, 9c, 10).[2,4] Although prior work has detected for the expression of RNA transcripts corresponding to majority of these proteins, transcripts belonging to ORF10 have not been observed.[4,5] Furthermore, Pancer *et al.* 2020 had identified multiple SARS-CoV-2 variants with prematurely terminated ORF10 sequences.[6] In their study, they found disease was not attenuated, transmission was not hindered, and replication proceeded similarly to strains possessing intact OFR10 sequences.[6] Altogether, current research has led to uncertainty regarding the biological relevance of ORF10 in SARS-CoV-2, with suggestions that its genome annotation should be revised. As an unintended consequence, the putatively encoded protein's structure and function has remained largely unresolved.

Found upstream of the 3'-untranslated region (3'-UTR), ORF10 (117 nt long) supposedly encodes for a protein that is 38 amino acids in length.[7,8] Although there are sequences homologous to this protein in other closely related coronaviruses (CoVs), none possess any known structure. There is also no experimentally derived crystal structure for the ORF10 protein; however, previous studies that attempted to predict the secondary structure and/or model the

protein reveal one α-helix and, depending on the study, two β-strands.[8,9,10] This protein was also found to lack significant levels of disorder; although, there is bioinformatic evidence that suggests a short molecular recognition feature (MoRF) spans residues 3-7.[11] The ORF10 protein is also hydrophobic with the α-helix identified as a possible transmembrane helix (TH).[9,12]

The greater implications are not clear, but the ORF10 protein contains high numbers of immunogenic promiscuous cytotoxic T-cell epitopes, primarily on the α-helix.[8,9] Assuming there exists a likely function, the ORF10 protein was computationally shown to closely associate with members of the $CUL2^{ZYG11B}$ complex.[10] It was suggested this protein directly interacts with the substrate adapter ZYG11B, thereby overtaking the complex and modulating certain aspects of ubiquitination to enhance viral pathogenesis.[10] Even so, the mechanisms and residues involved in this interaction have not been described in any great detail. Alternatively, ORF10 may act itself or serve as a precursor for other RNAs in regulating gene expression/replication, translation efficiency, or interfering with antiviral pathways.[5] From an evolutionary standpoint, ORF10 was found to be under the influence of strong positive selection (dN/dS = 3.82) without relaxed constraints.[12] This leads to the possibility ORF10 might encode a conserved and functional protein; however, additional sequence data is necessary to confirm this hypothesis.

Overall, there still remains an extremely few number of research efforts spent on investigating the putative ORF10 protein within SARS-CoV-2. As such, more detailed studies could further elucidate its structure and in turn provide additional insights as to a possible function for this protein. Herein, this study uses a strictly bioinformatic approach to theoretically characterize and construct the putative ORF10 protein in SARS-CoV-2.

2. Materials and Methods

2.1 Sequence Alignment and Phylogenetic Analysis

The reference sequence corresponding to the ORF10 protein in SARS-CoV-2, along with twenty-two variants, were acquired from NCBI's Protein Database; furthermore, sequences from closely related bat and pangolin CoVs were acquired as well (Table 1). Sequences were aligned using MUSCLE on the MEGA-X v10.1.7 software.[13,14] Both clustering methods were changed to neighbor joining while the other default settings were maintained. Alignment reliability was determined by overall mean distance and calculated using the *p-distance* substitution model. If the overall mean distance was found to be ≤ 0.7 then the alignment was considered reliable.[15]

A single protein tree was constructed using the neighbor-joining method and visualized on MEGA-X.[16] In the construction of this tree, uniform rates and partial deletion were used along with a site coverage cutoff of 95%. The phylogeny was tested using the bootstrap method. This tree was not rooted with any specified outgroup.

2.2 Protein Characterization and Secondary Structure Predictions

Protein disorder predictions were performed using previously described methods.[11,17] To detect phosphorylation sites, the DEPP server (http://www.pondr.com/cgi-bin/depp.cgi) was used. The grand average of hydropathicity (GRAVY) was determined by ProtParam, along with a hydrophobicity plot generated by ProtScale.[18] For structural predictions, three different online servers (PSIPRED v4.0, Jpred v4.0, and NetSurfP v2.0) were used to determine the probability scores of residues existing in either secondary structures or coils.[19,20,21] These values were then averaged to obtain individual residue scores that corresponded to each structural state (helix, strand, coil). The highest score of the three states indicated the placement of that

residue. In addition, TH predictions were performed by TMHMM and Phobius.[22,23] Predictions were then averaged and the most constant range of scores were used to designate the span of amino acids likely to be associating within the TH. Lastly, the propensity of residues involved in protein-binding interactions were evaluated using SCRIBER.[24]

2.3 Protein Modeling and Evaluation

The webserver IntFOLD was employed to make use of an *ab initio* modeling approach in constructing the ORF10 protein.[25] Models were evaluated based on IntFOLD's quality and confidence scoring. The best model was then refined using the 3Drefine webserver.[26] Among the five generated post-refinement, the model possessing the highest QMEAN Z-score was viewed as most favorable.[27] UCSF Chimera was used to visualize the model.[28] Electrostatic surfaces were generated based on the AMBER ff14SB charge model. Hydrophobicity surfaces were produced according to the default Kyte-Doolittle scale. Quality and structural comparisons against two previously constructed ORF10 protein models were also performed.[8,29]

3. Results

3.1 Sequence and Phylogenetic Analysis

The alignment had an overall mean distance of 0.059, corresponding to 94.10% identity for the entire alignment. The longest conserved region spanned residues 15-19, followed by two shorter regions spanning residues 11-12 and 33-34 (Figure 1a). Few amino acid differences occurred between SARS-CoV-2 and closely related bat and pangolin CoVs. For example, at amino acid site twenty-five SARS-CoV-2 and Pangolin-CoV MP789 possess arginine and serine, respectively. Missense mutations were identified among every SARS-CoV-2 ORF10 protein variant; furthermore, one mutation per variant was observed (Figure 1a; Table 2). Phylogenetic analysis revealed the ORF10 protein in Pangolin-CoV GX-P1E formed a lineage with the greatest evolutionary distance whereas ORF10 proteins in the remaining CoVs shared much closer relationships with significantly smaller distances (Figure 1b).

3.2 Protein Disorder, Phosphorylation, Hydropathicity, and Binding Propensity

A per-residue disorder plot for the ORF10 protein indicated that disorder scores within the C-terminal half (20-38) vary whereas the N-terminal half (1-19) held more uniform disorder scores (Figure 2). No phosphorylation sites were identified in this protein (Table S1). The hydrophobicity plot revealed two hydrophobic regions spanning residues 3-19 and 28-36, along with a single hydrophilic region spanning residues 20-27 (Figure 3a). The GRAVY score for the reference SARS-CoV-2 ORF10 protein was 0.637; however, GRAVY scores for all twenty-two variants fluctuated above and below the reference (Figure 3b). For mutations resulting in similar amino acid chemistry, polar residues that mutated into different polar residues had increased GRAVY scores. Interestingly, nonpolar residues that mutated into different nonpolar

residues had decreased GRAVY scores except in SARS-CoV-2var9 and var10. For mutations that involved shifts in polarity, all polar to nonpolar mutations had increased GRAVY scores; however, all nonpolar to polar mutations resulted in decreased GRAVY scores except for SARS-CoV-2var8 and var20. As for protein-binding propensity, residues in the N-terminal half held greater propensity scores than residues found in the C-terminal half (Figure 4).

3.3 Secondary Structure and Protein Modeling

According to secondary structure predictions, there exists a single α-helix and β-strand spanning residues 6-21 and 28-34, respectively; furthermore, one TH was predicted to span residues 7-18 (Figure 5; Table S2; Table S3). IntFOLD yielded five models with varying confidence and quality scores (Table S4). The model having a low P-value (3.33e-03) and high quality score (0.3714) was subjected to refinement. Of the five models made post-refinement, the model with a QMEAN Z-score of -0.66 was selected as the most favorable (Table S5).

Structurally, the model presented with a β-α-β motif spanning residues 3-31, along with a 3/10-helix spanning residues 34-37 (Figure 6a). An electrostatic surface map revealed two regions of positive charge and one region of negative charge (Figure 6b). Electropositive regions were influenced by R20 and R24 whereas the electronegative region was influenced by D31. As expected, a majority of the protein's surface was hydrophobic; however, hydrophilic regions did appear (Figure 6c). For example, the residues spanning 20-27 presented with surface coloring that reflected hydrophilic character, as was shown in the hydrophobicity plot as well.

3.4 Model Comparisons

Prior to this study, two ORF10 protein models were built using different methods. The first model was constructed using the QUARK web server whereas the second model was built

using the I-TASSER web server.[8,29] Both models were extensively different from the model produced within this study, mainly in terms of structure and topology (Figure 7). Furthermore, the QMEAN Z-score for both models (QUARK: -3.88 / I-TASSER: -2.63) were significantly lower indicating high levels of error and extremely poor model quality.

4. Discussion

The phylogenetic results are similar to those by Hassan *et al*. 2020 in that none of the variants resulted in the formation of distinct clades.[9] At most, the phylogenetic data in this study suggests the ORF10 protein in Pangolin-CoV GX-P1E is most distantly related to all other ORF10 proteins. With such high similarity observed in the remaining ORF10 proteins, describing relationships between specific sequences are not entirely possible when based solely on sequence data. Although unlikely, perhaps structural comparisons of each protein would assist in revealing further relationships not highlighted by sequence information. As for conserved portions in the ORF10 protein, a possible explanation exists. Assuming the ORF10 protein is synthesized, these conserved regions could be vital to protein structure and/or function. Majority of these residues possess high protein-binding propensity scores; therefore, they may be essential in facilitating specific protein interactions. Mutations in these regions were not created, thus it is unclear if these residues would be essential in maintaining protein structure.

Protein disorder predictions revealed the ORF10 protein is ordered. It is only near each terminus that residues begin expressing moderate to high flexibility. These results align well with those presented by Giri *et al.* 2020; however, results in this study do contradict those by Hassan *et al.* 2020.[9,11] Their study indicates a greater number of flexible residues, including several disordered residues found within the C-terminus, exist in the ORF10 protein.[9] This disagreement is likely attributed to the method of prediction, where Hassan *et al*. 2020 used a single prediction algorithm unlike this study which used multiple prediction algorithms to evaluate disorder.[9] In doing so, per-residue disorder scores were not over estimated by a single algorithm; therefore, more conservative but reliable predictions were achieved.[17] The data in this

study further indicates that mutations within the C-terminus instill changes in residue flexibility; however, such changes did not result in major shifts towards disorder. Conversely, mutations within the N-terminus had caused minuscule changes in residue flexibility, but as was noted in the C-terminus, these changes do not result in shifts towards disorder.

Based on GRAVY scoring, mutations in the ORF10 protein had some obvious effects on hydropathicity. For example, polar to nonpolar mutations had increased GRAVY scores which implies an increase in the protein's hydrophobic character. However, certain shifts in polarity cannot be explained as easily. For example, nonpolar to polar mutations presented decreases in GRAVY scores with the exception of SARS-CoV-2var8 and var20. If anything, nonpolar to polar mutations should lower the GRAVY score due to decreases in the protein's hydrophobic character. At this stage, the underlying cause for higher GRAVY scores in these variants remains unclear and requires further investigation. Regardless of the increases and decreases in scoring, the ORF10 protein still remained quite hydrophobic. As such, these mutations are not likely to have major effects on structure and/or function; however, Hassan *et al*. 2020 does indicate that majority of these mutations, regardless of their change in polarity, are deleterious and will decrease the protein's structural integrity.[9] Modeling of these mutations would help determine whether or not they would negatively affect the protein's structure.

The prediction of secondary structures for the ORF10 protein describe the presence of one α-helix and one β-strand. When compared to the structural assignment of residues in the ORF10 protein model, only 58% of residues match. For example, secondary structure predictions suggest the α-helix spans residues 6-21 whereas in the protein model the α-helix spans residues 10-22. Nonetheless, both results indicate that an α-helix and β-strand do exist in

this protein. A second β-strand, occuring before the α-helix, comprised residues 3-8 and was only observed within the model. According to Giri *et al.* 2020, a MoRF was predicted to exist in residues 3-7; furthermore, secondary structure predictions for this particular region held the lowest prediction scores.[11] A unique feature associated with MoRFs is their ability to exist in semi-disordered (i.e. flexible) states until binding, after which they can assume a designated secondary structure.[30] Altogether, it is possible residues 3-7 (or 3-8) function as a β-MoRF that upon binding to a different protein, will assume the structure of a β-strand. Residues found in this region also have high protein-binding propensity scores to support this notion.

The ORF10 protein, assuming that it's produced, would most likely be a membrane protein. Whether it is integrated within or peripheral to the membrane remains unclear, but given the extent of nonpolar residues it would seem much more likely to be integrated inside a membrane. In particular, CoVs encode multiple viroporins.[31] For example, in SARS-CoV the 8a protein oligomerizes to form ion channels inside mitochondrial membranes that induce apoptosis by depolarizing membrane potential.[12,31-33] The ORF10 in SARS-CoV-2 may oligomerize in such a way that polar and electrically charged residues coat the inside of a pore that facilitates transport of ions or small molecules for viral replication, virulence, and/or pathogenicity. A prior study found the ORF10 protein colocalized with ORF8 and ORF7b proteins in the endoplasmic reticulum (ER); however, its role in the ER was not described.[34] In this case, the ORF10 protein could associate with ER membranes as a means to assist in replication. Experimental work is necessary to evaluate this finding. Alternatively, the ORF10 protein could interact with another protein. As previously stated, the ORF10 protein was suggested to associate with the substrate

adapter ZYG11B.[10] This interaction is likely mediated by residues in the N-terminal half on account of the β-MoRF and higher residue protein-binding propensity scores.

Collectively, this study does not provide nor discredit prior evidence regarding the existence of this protein in SARS-CoV-2; however, this study does provide an interesting theoretical examination of this hypothetical protein. In summary, the putative ORF10 protein appears ordered and hydrophobic; furthermore, this protein may possess a β-MoRF followed by an α-helix and subsequent β-strand. This protein also shares strong phylogenetic relationships with ORF10 proteins in other closely related CoVs. Lastly the protein model generated in this study is of higher quality than previously generated models. Hopefully, these results can provide a foundation to drive the experimental work necessary to further assess the biological relevance of the putative ORF10 protein in SARS-CoV-2.

Acknowledgments

I would like to thank the previous researchers who had deposited the sequences used in this study onto the NCBI Protein Database. Without access to those sequences, the research described in this manuscript could have never been accomplished.

Conflict of Interest

The author declares that no conflicts of interest exist.

Tables and Figures

CoV Species and Variants*	Accession Number
SARS-CoV-2ref	YP_009725255.1
Pangolin-CoV MP789	QIG55954.1
Pangolin-CoV GX-P1E	N/A**
Bat-CoV RaTG13	N/A**
bat-SL-CoVZC45	N/A**
SARS-CoV-2var1	QIS29991.1
SARS-CoV-2var2	QKU54102.1
SARS-CoV-2var3	QKV08176.1
SARS-CoV-2var4	QKV37245.1
SARS-CoV-2var5	QLA48060.1
SARS-CoV-2var6	QLG76514.1
SARS-CoV-2var7	QLG99793.1
SARS-CoV-2var8	QLI33453.1
SARS-CoV-2var9	QLJ57416.1
SARS-CoV-2var10	QLY88596.1
SARS-CoV-2var11	QMT54534.1
SARS-CoV-2var12	QMT94417.1
SARS-CoV-2var13	QMT97141.1
SARS-CoV-2var14	QMU93213.1
SARS-CoV-2var15	QNA70543.1
SARS-CoV-2var16	QNB17780.1
SARS-CoV-2var17	QNC04532.1
SARS-CoV-2var18	QNC49349.1
SARS-CoV-2var19	QNG41574.1
SARS-CoV-2var20	QNG42985.1
SARS-CoV-2var21	QNI23218.1
SARS-CoV-2var22	QNI25281.1

Table 1. Complete list of CoV ORF10 protein sequences used in this study.

*Please note, the twenty-two individual SARS-CoV-2 ORF10 protein variants were originally collected by Hassan *et al*. 2020.[9]

**No accession number. Nucleotide sequences were acquired and then translated (http://biomodel.uah.es/en/lab/cybertory/analysis/trans.htm).

SARS-CoV-2 Variant ID	Mutation	Change in Polarity
SARS-CoV-2var1	V6I	NP → NP
SARS-CoV-2var2	Y26H	NP → P
SARS-CoV-2var3	L37P	NP → NP
SARS-CoV-2var4	T38I	P → NP
SARS-CoV-2var5	R20K	P → P
SARS-CoV-2var6	N22T	P → P
SARS-CoV-2var7	I3M	NP → NP
SARS-CoV-2var8	Y26C	NP → P
SARS-CoV-2var9	A8V	NP → NP
SARS-CoV-2var10	A28V	NP → NP
SARS-CoV-2var11	R24L	P → NP
SARS-CoV-2var12	D31Y	P → NP
SARS-CoV-2var13	S23F	P → NP
SARS-CoV-2var14	R24C	P → P
SARS-CoV-2var15	L37F	NP → NP
SARS-CoV-2var16	F35S	NP → P
SARS-CoV-2var17	R20I	P → NP
SARS-CoV-2var18	V30L	NP → NP
SARS-CoV-2var19	I4L	NP → NP
SARS-CoV-2var20	P10S	NP → P
SARS-CoV-2var21	G2D	NP → P
SARS-CoV-2var22	V30A	NP → NP

Table 2. Twenty-two SARS-CoV-2 ORF10 protein variants and their corresponding mutations, along with the changes in polarity associated with each mutation. “NP” - Nonpolar. “P” - Polar.

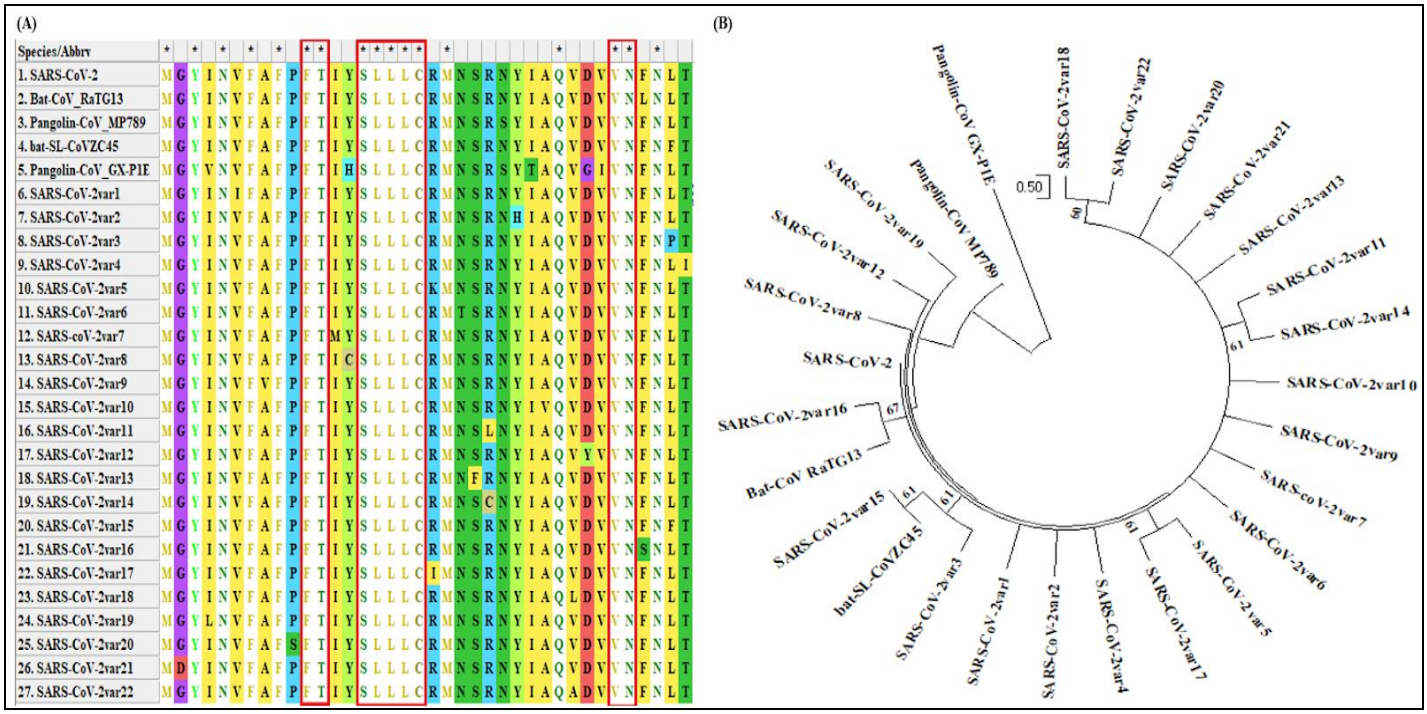

Figure 1. (A) Screen capture of the MUSCLE alignment for the ORF10 protein in SARS-CoV-2, along with closely related bat and pangolin CoVs. The red boxes indicate conserved regions with more than one residue. **(B)** Phylogenetic analysis of ORF10 protein sequences depicting evolutionary relationships within SARS-CoV-2, along with closely related bat and pangolin CoVs. Phylogeny was inferred using the number of differences model. The sum of branch lengths was 26.06 with bootstrap values calculated out of 1,000 replicates and nodes with < 60.00% support were hidden.

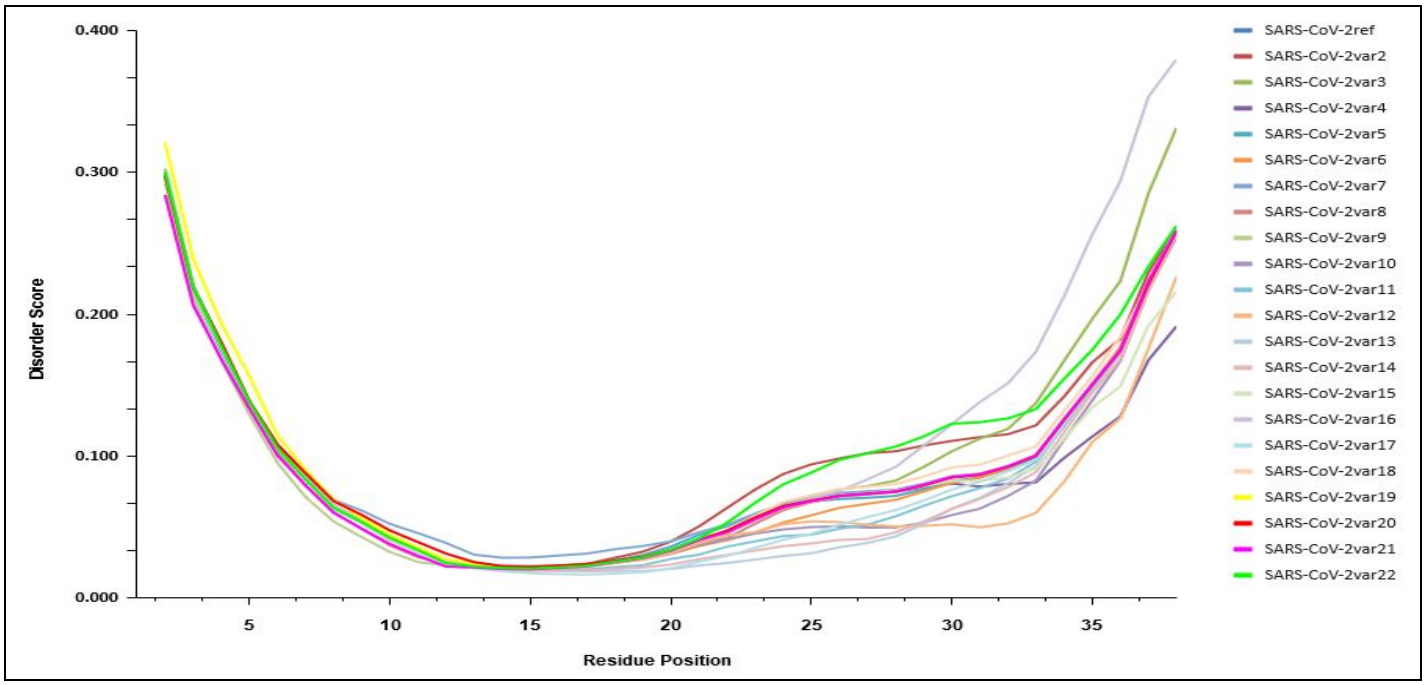

Figure 2. Per-residue disorder plot for ORF10 proteins found in SARS-CoV-2 and Pangolin-CoV MP789, along with SARS-CoV-2 protein variants. Scores ≥ 0.5 indicate disorder whereas scores between 0.25 and 0.5 indicate highly flexible residues. Any score that exists between 0.1 and 0.25 suggests moderate flexibility. Scores ≤ 0.1 indicate rigidity.

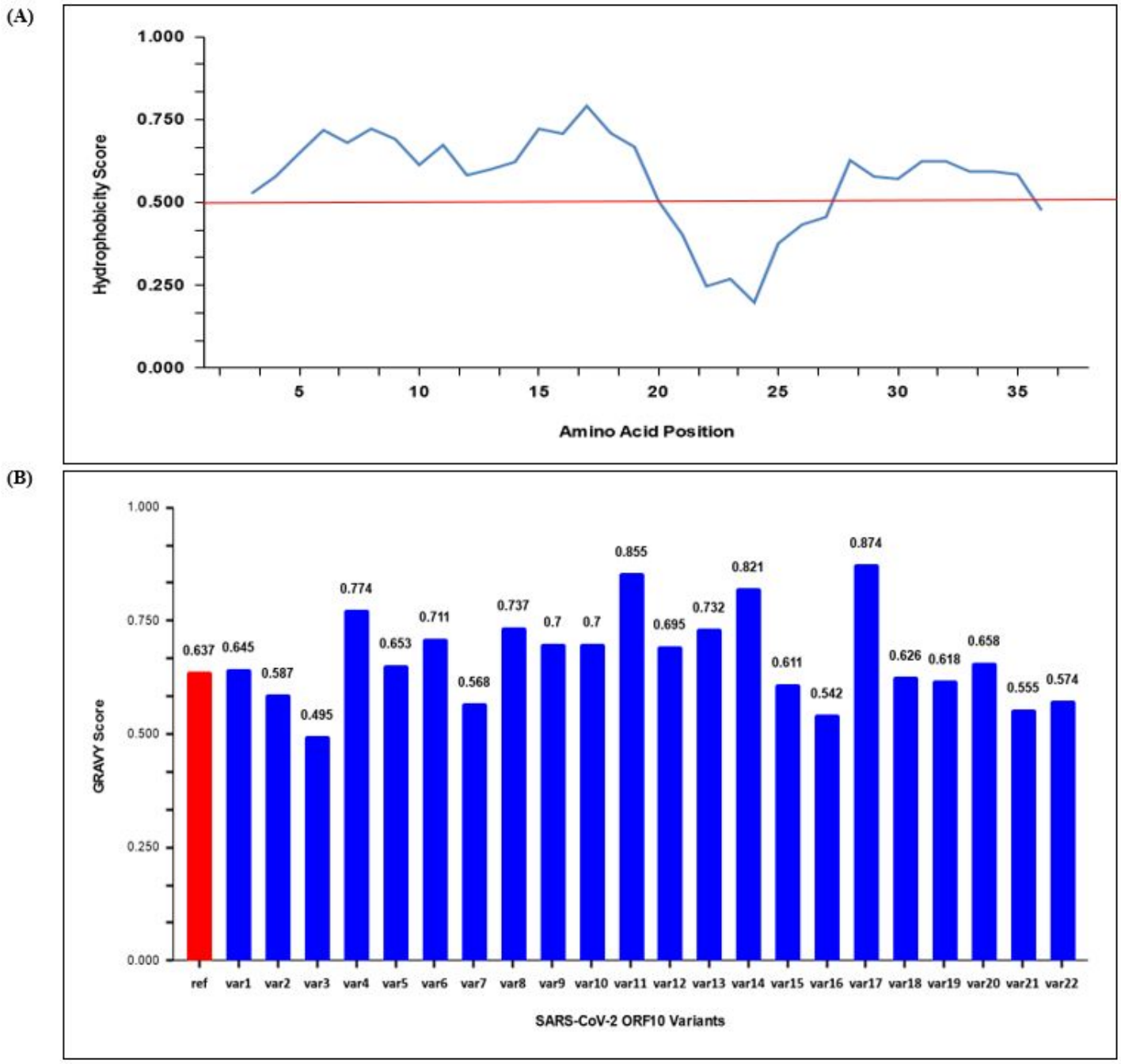

Figure 3. (A) Hydrophobicity plot for the reference SARS-CoV-2 ORF10 protein. In generating this plot, scores were normalized on a 0-1 scale and a 5-residue window was used. The red line displays the threshold where values above this line indicate hydrophobic regions and values below this line indicate hydrophilic regions. Amino acids 1-2 and 37-38 are not shown on this plot. **(B)** GRAVY scores for the SARS-CoV-2 OFR10 protein. The red box corresponds to the reference sequence while blue boxes correspond to the variants. The numerical value for each score is shown above its corresponding box. Positive scores suggest a hydrophobic protein and (despite none being found) negative scores would suggest a more hydrophilic protein.

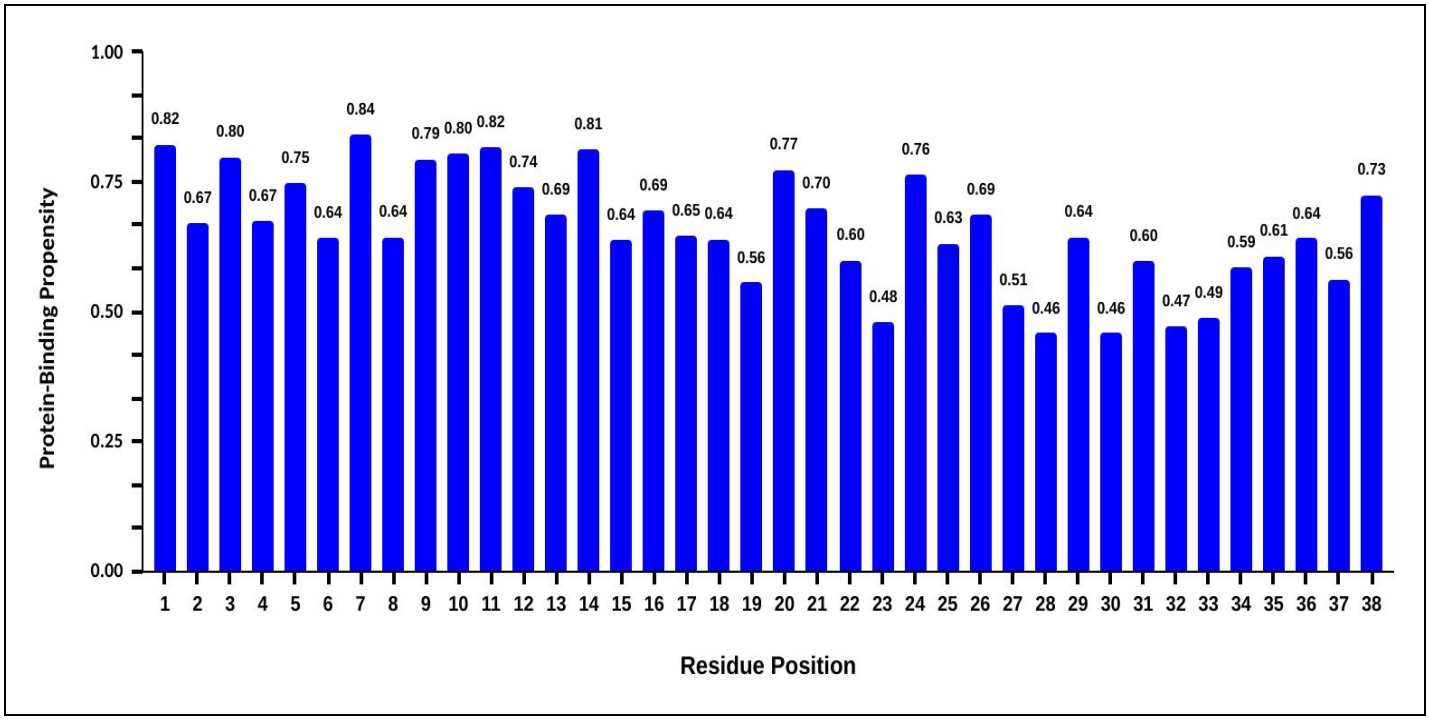

Figure 4. Protein-binding propensity scores for amino acids within the reference SARS-CoV-2 ORF10 protein. The value of each score is shown above its corresponding box. As a general rule, a higher score reflects an increased likelihood that a specific residue is to engage in protein-binding.

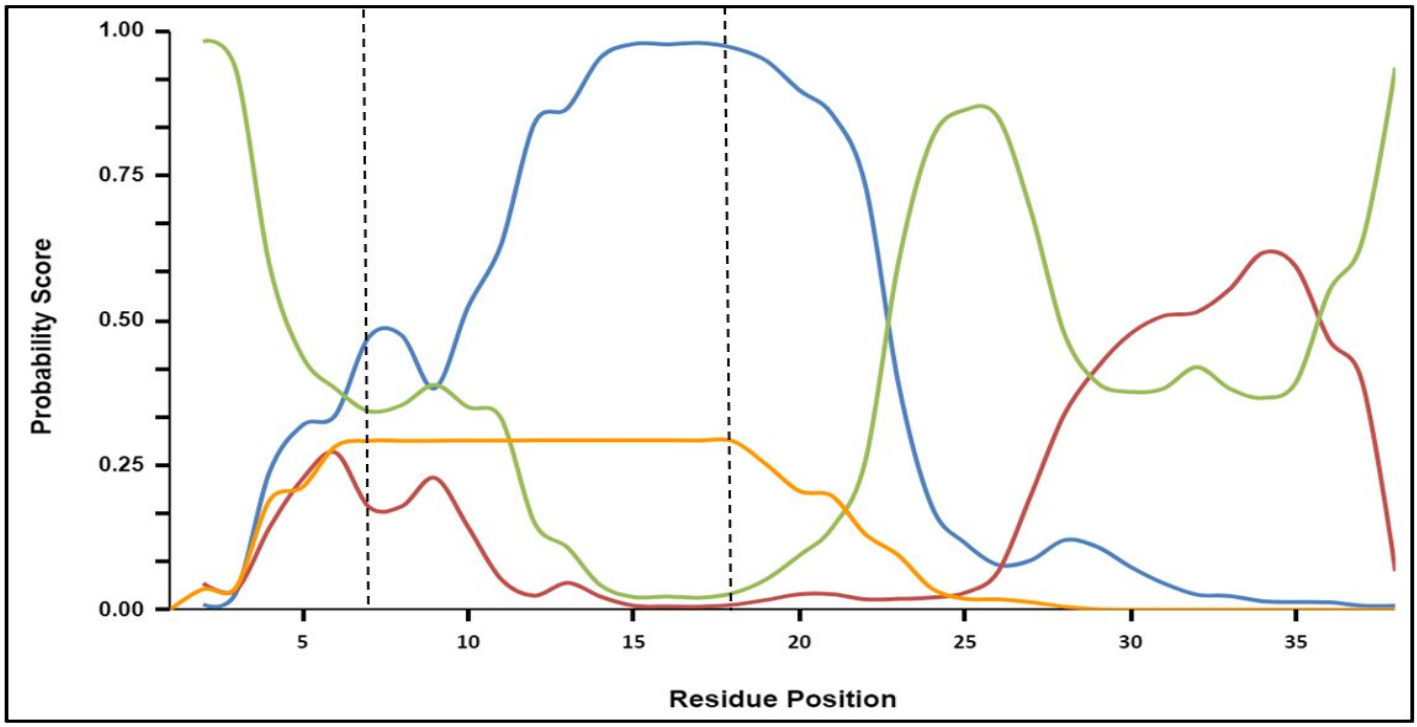

Figure 5. Secondary structure and transmembrane helix prediction scores for the ORF10 protein. Dotted lines encompass the span of residues having the most consistent stretch of transmembrane helix prediction scores. Colors for each line are defined as follows: helix - blue; strand - red; coil - green; transmembrane helix - orange.

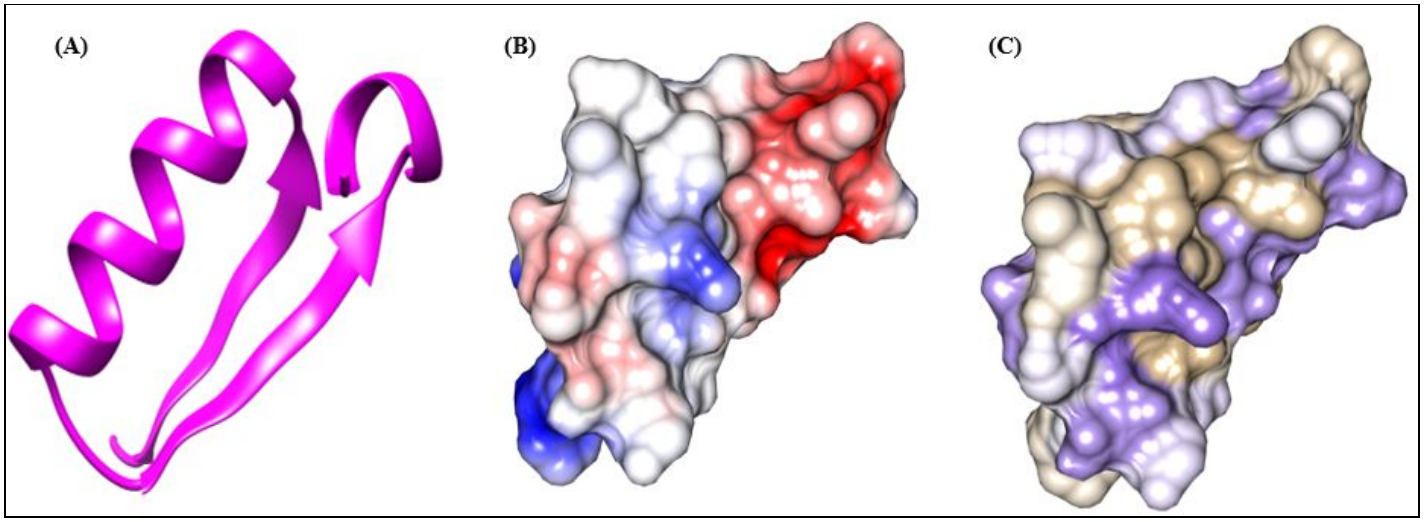

Figure 6. (A) Ribbon diagram of the ORF10 protein in SARS-CoV-2. **(B)** Electrostatic surface map of the ORF10 protein. Red indicates regions having negative charge whereas blue indicates regions having positive charge. White indicates regions having neutral charge. **(C)** Hydrophobicity surface map of the ORF10 protein. Purple indicates hydrophilic regions and brown indicates hydrophobic regions. White indicates regions presenting with semi-hydrophobic/hydrophilic character.

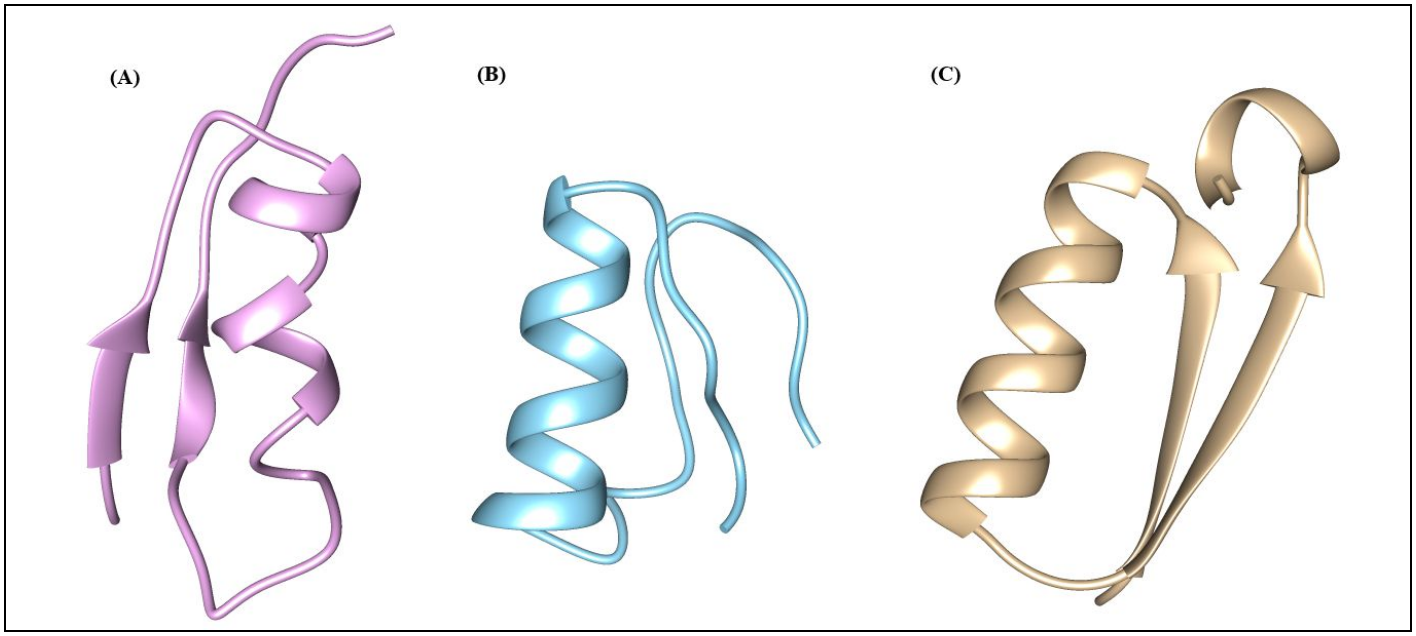

Figure 7. Three ribbon diagrams generated by different methods to represent the ORF10 protein. The **(A)** I-TASSER, **(B)** QUARK, and (**C**) current study's model of the protein. Colors are not relevant other than helping to distinguish between each model.

References

[1]WHO: World Health Organization. (2020). "Weekly Epidemiological Update Coronavirus disease 2019 (COVID-19) 12 October 2020." *Coronavirus disease (COVID-19) Weekly Epidemiological Update and Weekly Operational Update*. World Health Organization. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports

[2]Wu A, Peng Y, Huang B, et al. (2020) "Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China." *Cell Host & Microbe*. 27.3. Pg 325–328. https://doi.org/10.1016/j.chom.2020.02.001

[3]Chen Y, Liu Q, and Guo D. (2020) "Emerging coronaviruses: Genome structure, replication, and pathogenesis." *Journal of Medical Virology*. 92. Pg 418–423. https://doi.org/10.1002/jmv.25681

[4]Kim D, Lee JY, Yang JS, *et al*. (2020). "The Architecture of SARS-CoV-2 Transcriptome." *Cell*. 181. Pg 914–921. https://doi.org/10.1016/j.cell.2020.04.011

[5]George T, Rawlinson D, Featherstone L, *et al*. (2020). "Direct RNA sequencing and early evolution of SARS-CoV-2." *bioRxiv*. Preprint. https://doi.org/10.1101/2020.03.05.976167

[6]Pancer K, Milewska A, Owczarek K, *et al*. (2020). "The SARS-CoV-2 ORF10 is not essential *in vitro* or *in vivo* in humans." *bioRxiv*. Preprint. https://doi.org/10.1101/2020.08.29.257360

[7]Khailany RA, Safdar M, and Ozaslan M. (2020). "Genomic characterization of a novel SARS-CoV-2." *Gene Reports*. 19. 100682. https://doi.org/10.1016/j.genrep.2020.100682

[8]Mirsha S. (2020). "ORF10: Molecular Insights into the Contagious Nature of Pandemic Novel Coronavirus 2019-nCoV." *ChemRxiv*. Preprint. https://doi.org/10.26434/chemrxiv.12118839.v3

[9]Hassan S, Attrish D, Ghosh S, et al. (2020). "Notable sequence homology of the ORF10 protein introspects the architecture of SARS-COV-2." *bioRxiv*. Preprint. https://doi.org/10.1101/2020.09.06.284976

[10]Gordon DE, Jang GM, Bouhaddou M, *et al.* (2020). "A SARS-CoV-2 protein interaction map reveals targets for drug repurposing." *Nature*. 583. Pg 459–468. https://doi.org/10.1038/s41586-020-2286-9

[11]Giri R, Bhardwaj T, Shegane M, *et al*. (2020). "When Darkness Becomes a Ray of Light in the Dark Times: Understanding the COVID-19 via the Comparative Analysis of the Dark Proteomes

of SARS-CoV-2, Human SARS and Bat SARS-Like Coronaviruses." *bioRxiv*. Preprint. https://doi.org/10.1101/2020.03.13.990598

[12]Cagliani R, Forni D, Clerici M, *et al*. (2020). "Coding potential and sequence conservation of SARS-CoV-2 and related animal viruses." *Infection, Genetics, and Evolution*. 83. 104353. https://doi.org/10.1016/j.meegid.2020.104353

[13]Edgar RC. (2004) "MUSCLE: multiple sequence alignment with high accuracy and high throughput." *Nucleic Acids Research*. 32.5. Pg. 1792–1797. https://doi.org/10.1093/nar/gkh340

[14]Kumar S, Stecher G, Li M, *et al*. (2018) "MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms." *Molecular Biology and Evolution*. 35. Pg 1547–1549.https://doi.org/10.1093/molbev

[15]Hall BG. (2018). "Phylogenetic Trees Made Easy: A How-To Manual." Oxford University Press, New York. 5. Pg 60-61.

[16]Saitou N and Nei M. (1987). "The neighbor-joining method: A new method for reconstructing phylogenetic trees." *Molecular Biology and Evolution*. 4. Pg 406-425. https://doi.org/10.1093/oxfordjournals.molbev.a040454

[17]Cleveland SB, Davies J, and McClure MA. (2011) "A Bioinformatics Approach to the Structure, Function, and Evolution of the Nucleoprotein of the Order Mononegavirales." *PLoS ONE*. 6.5. Pg 1–13. https://doi.org/10.1371/journal.pone.0019275

[18]Gasteiger E, Hoogland C, Gattiker A, *et al*. (2005). "The Proteomics Protocols Handbook." Humana Press. Pg 571-607.

[19]Jones DT. (1999). "Protein secondary structure prediction based on position-specific scoring matrices." *Journal of Molecular Biology*. 292. Pg 195-202. https://doi.org/10.1006/jmbi.1999.3091

[20]Drozdetskiy A, Cole C, Procter J, *et al*. (2015). "JPred4: a protein secondary structure prediction server." *Nucleic Acids Research*. 43.1. Pg 389-394. https://doi.org/10.1093/nar/gkv332

[21]Klausen MS, Jespersen MC, Nielsen H, *et al*. (2019). "NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning." *Proteins: Structure, Function, and Bioinformatics*. 87.6. Pg 520-527. https://doi.org/10.1002/prot.25674

[22]Krogh A, Larsson B, Heijne GV, et al. (2001). "Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes." *Journal of Molecular Biology*. 305.3. Pg 567-580. https://doi.org/10.1006/jmbi.2000.4315

[23]Lukas K, Krogh A, & Sonnhammer ELL. (2004). "A Combined Transmembrane Topology and Signal Peptide Prediction Method." *Journal of Molecular Biology*. 3385. Pg 1027-1036. https://doi.org/10.1016/j.jmb.2004.03.016

[24]Zhang J & Kurgan L. (2019). "SCRIBER: accurate and partner type-specific prediction of protein-binding residues from proteins sequences." *Bioinformatics*. 35.14. Pg 343-353. https://doi.org/10.1093/bioinformatics/btz324

[25]McGuffin LJ, Adiyaman R, Maghrabi AHA, *et al*. (2019). "IntFOLD: an integrated web resource for high performance protein structure and function prediction." *Nucleic Acids Research*. 47. Pg 408-413. https://doi.org/10.1093/nar/gkz322

[26]Bhattacharya D & Cheng J. (2012). "3Drefine: Consistent Protein Structure Refinement by Optimizing Hydrogen Bonding Network and Atomic Level Energy Minimization." *Proteins: Structure, Function, and Bioinformatics*. 81.1. Pg 119-131. https://doi.org/10.1002/prot.24167

[27]Benkert P, Biasini M, & Schwede T. (2012). "Toward the estimation of the absolute quality of individual protein structure models." *Bioinformatics*. 27. Pg 343-350. https://doi.org/10.1093/bioinformatics/btq662

[28]Pettersen EF, Goddard TD, Huang CC, *et al*. (2004) "UCSF Chimera-a visualization system for exploratory research and analysis." *Journal of Computational Chemistry*. 25.13. Pg 1605–1612. https://doi.org/10.1002/jcc.20084

[29]Zhang Lab. University of Michigan. zhanglab.ccmb.med.umich.edu/COVID-19/. Accessed 15 Sept. 2020.

[30]Mohan A, Oldfield CJ, Radivojac P, *et al*. (2006). "Analysis of molecular recognition features (MoRFs)." *Journal of Molecular Biology*. 362.5. Pg 1043–1059. https://doi.org/10.1016/j.jmb.2006.07.087

[31]Forni D, Cagliani R, Clerici M, *et al*. (2017). "Molecular Evolution of Human Coronavirus Genomes." *Trends in Microbiology*. 25.1. Pg 35-48. https://doi.org/10.1016/j.tim.2016.09.001

[32]Chen CC, Krüger J, Sramala I, *et al*. (2011). "ORF8a of SARS-CoV forms an ion channel: Experiments and molecular dynamics simulations." *Biochimica et Biophysica Acta (BBA) - Biomembranes*. 1808.2. Pg 572-579. https://doi.org/10.1016/j.bbamem.2010.08.004

[33]Chen CY, Ping YH, Lee HC, *et al*. (2007). "Open reading frame 8a of the human severe acute respiratory syndrome coronavirus not only promotes viral replication but also induces apoptosis." *The Journal of Infectious Diseases*. 196.3. Pg 405–415. https://doi.org/10.1086/519166

[34]Zhang J, Cruz-cosme R, Zhuang MW, *et al*. (2020). "A Systemic and Molecular Study of Subcellular Localization of SARS-CoV-2 Proteins." *bioRxiv*. Preprint. https://doi.org/10.1101/2020.08.02.233023