ORF10: Molecular insights into the contagious nature of pandemic novel coronavirus 2019-nCoV

Seema Mishra

Department of Biochemistry

School of Life Sciences

University of Hyderabad, India

Email: seema_uoh@yahoo.com

Summary:

2019-novel coronavirus (nCoV), the virus causing Covid19 disease, is a highly contagious virus and has created a pandemic. In efforts to combat this virus, subunit vaccine designing using immunoinformatics tools was undertaken. During these studies, new hypotheses began to emerge, the prominent among these was that the newly acquired ORF10 protein may be the key protein responsible for its highly contagious nature. ORF10 is a hitherto unknown protein with no homology to any known protein in organisms present till date. Through immunoinformatics studies, it has been observed that among all ten 2019-nCoV proteins, ORF10 presents amongst the highest number of immunogenic, promiscuous CTL epitopes. These epitopes are part of a cluster with HTL epitopes, suggesting that there is a high degree of epitope conservation in ORF10. Conservation of protein sequence across organisms is not seen, and there is no known structural template on which to model and derive a

structure to get structural and functional insights. Because there is altogether no conservation of its sequence, or structure, it may be presented as a novel protein to the immune system. Further, the human body may not have been able to utilize any memory B and T cells generated against other microorganisms to target ORF10 and fight this pathogen, contributing to its deadly, contagious nature.

Introduction:

Novel coronavirus (nCoV) is highly contagious virus which first emerged in population in December 2019, resulting in Covid-19 disease. WHO has declared it as pandemic (https://www.who.int/emergencies/diseases/novel-coronavirus-2019) and there are concerted efforts to prevent and treat it as there is no cure available for it. Based on newly available 2019-nCoV genome sequence, this study was undertaken to design potential vaccine candidates for subunit vaccine. On way to ranking and designing these epitopes on the basis of several immunological parameters, several useful insights into the deadly nature of this pathogen were found along the way and are detailed below.

Open Reading Frame 10 (ORF10) may be the reason for contagious nature of this virion

ORF10 is a predicted protein, 38 amino acids in length, and is located after the nucleocapsid region in the 2019-nCoV genome. This genome is 29882 nucleotides in length and ORF10 gene region is from 29558..29674 nucleotides, as per GenBank sequence accession number MT106054.1/RefSeq sequence NC_045512.2. The protein sequence is as follows: MGYINVFAFPFTIYSLLLCRMNSRNYIAQVDVVNFNLT. It has no known structure and function, and not much relevant information is available on it in literature. As per Swiss-Model webpage harboring the structures of all 2019-nCoV proteins (https://swissmodel.expasy.org/repository/species/2697049), HHblits identified zero significant

templates for ORF10 and hence the model could not be built. HHblits is an iterative protein sequence search tool using HMM-HMM alignment. It is the only protein among all 2019-nCoV proteins which has no homologs for building even a low quality model. BLASTp and PSI-BLAST searches against 'nr' database, too, identified no homolog. This is an interesting observation, given that this sequence is located towards the end of the 2019-nCoV genome. Had it been located in-between anywhere along the genome, it would have been called an insertion sequence or a result of possible DNA rearrangement. Hence, this may be a novel event distinct from standard DNA rearrangement events such as transposition and recombination. This is also evident from the fact that it is a 38-amino acids small predicted protein coded by 117 nucleotides in the 2019-nCoV genome, while transposons and insertion sequences are much longer sequences, characterized by sizes ranging from 700 bp to several thousands bp (1). Even there is no possibility of occurence of long terminal repeat sequences, since the ORF10 protein or nucleotide sequence is not a repetitive sequence as observed from GenBank sequence accession number MT106054.1/RefSeq sequence NC_045512.2 harbouring sequences from two different human sources, and there are no flanking repeated sequences, either. Protein family search in UniProt or in MobiDB database for disordered proteins' function identified no known protein family.

Further, it is also of particular interest to note that this ORF10 nucleotide sequence was found to be 100% aligned to all 2019-nCoV ORF10 sequences from different geographical regions, from USA (MT106054.1), India (MT012098.1), Turkey (MT327745.1), South Korea (MT304474.1), Iran (MT320891.2), Taiwan (MT066175.1), Israel (MT276597.1), Nepal (MT072688.1), Vietnam (MT192772.1), China (MT291828), Sweden (MT093571), Greece (MT328032), Italy (MT077125), France (MT320538) and Spain (MT292569).

As there was no homology at all with any protein from BLASTp search, and no domains/family were predicted, the sequence was subjected to *ab-initio* modeling. Structural modeling using *ab-initio* web server QUARK (Fig. 1) showed the structure mostly comprised of middle α-helical region with disordered regions at both ends.

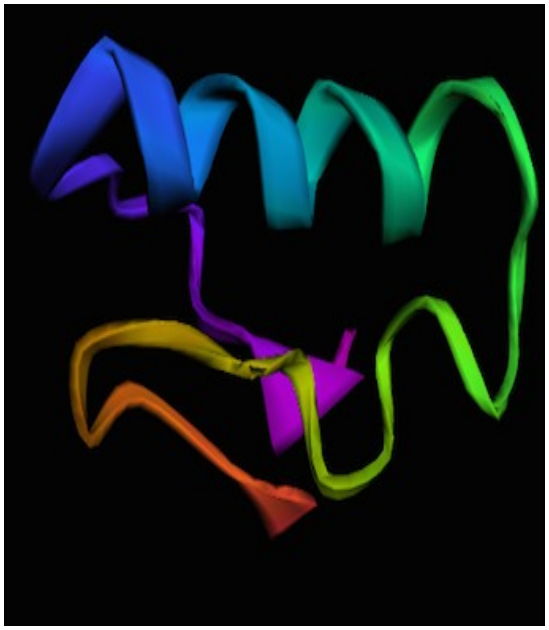


Fig. 1: ORF10 protein structure modeled using *ab-initio* modeling webserver QUARK.

All of the ten 2019-nCoV proteins, including ORF1ab replicase complex, were used to predict promiscuous cytotoxic T lymphocyte (CTL) and helper T lymphocyte (HTL) epitopes with high immunogenicity and conserved regions across SARS family of viruses (2) .

During these present studies towards designing subunit vaccines, upon identification of epitopes binding to all 12 HLA-I supertypes, a total of 9621 CTL epitopes were generated across ten 2019-nCoV proteins which include ORF1ab polyprotein. It was found that ORF10 harbored the highest number of promiscuous epitopes among all proteins, apart from nsp7 of ORF1ab polyprotein region (Fig. 2). Sequence analysis showed most of the epitopes belonged to this α-helical region (Fig. 3). It has been widely recognized that T cell epitopes are mostly found in α-helical regions of a protein (3, 4). Spike, membrane and nucleocapsid proteins had comparatively lower number of such epitopes. These 11 out of 30 CTL epitopes generated for ORF10 were predicted to have higher TAP transporter

binding, higher proteasomal cleavage as well as HLA-I binding capacity, and out of these 11 epitopes, 9 epitopes were predicted to be highly immunogenic among the top ranked candidates.

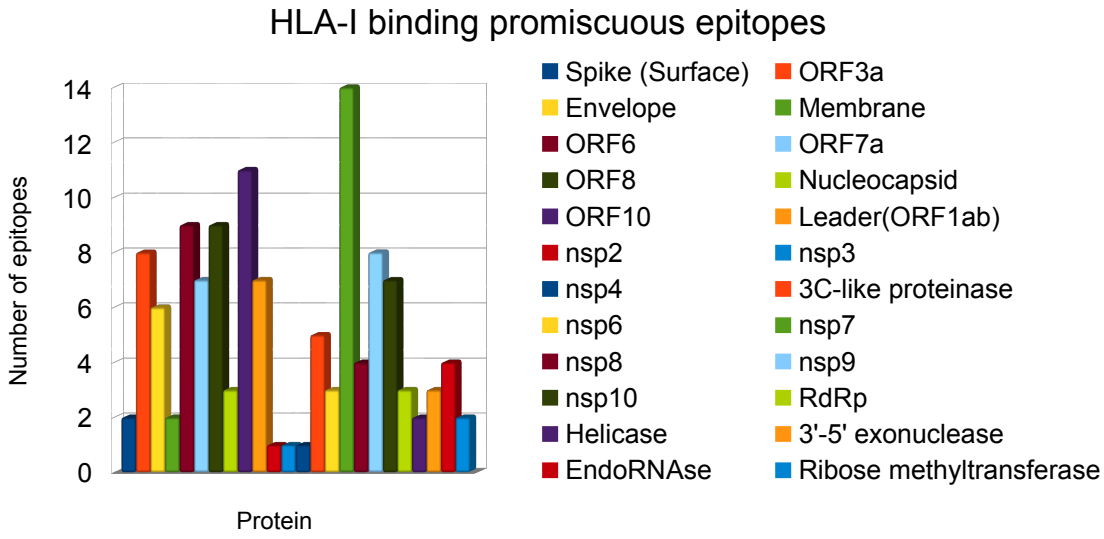


Fig. 2: Number of promiscuous HLA-I binding epitopes across 2019-nCoV proteins studied

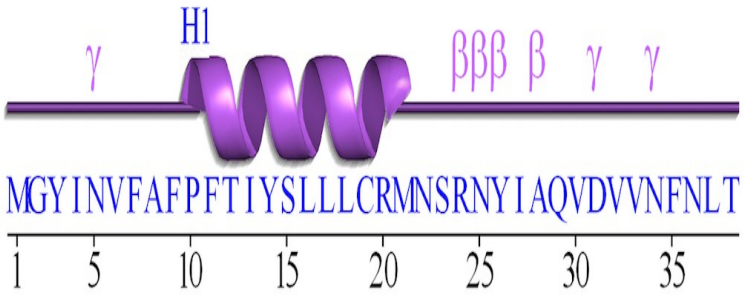


Fig. 3: Secondary structure plot of ORF10 protein, motifs β: β- turn, γ: γ- turn.

Upon further ranking of peptide epitopes, using immunogenicity screening, 9 out of 11 CTL epitopes from ORF10 were found to be highest in number among high scoring epitopes. HLA-II binding studies across all proteins showed that all the epitopes of ORF10 binding to key DRB1 alleles

were predicted to be weak binders. As HTL epitopes are required for helper T cells to boost antibody responses, it is surmised from the presence of weak binders that this may be a contributing factor towards low neutralizing antibody levels in 2019-nCoV infected patients in the initial days of infection, as reported in literature (5, 6).

Clustering analyses utilizing both HLA-I and HLA-II binding epitopes showed that all of those CTL epitopes from ORF10 that were high scoring in immunogenicity prediction had higher number of clusters with HTL epitopes. This led to high number of consensus epitopes among proteins incorporating both CTL and HTL epitopes (Fig. 4). Many epitopes belonging to other proteins were not a part of a cluster. Taken together, these analyses show that the CTL epitopes of ORF10 may be highly immunogenic. It may well function like miniprotein scaffolds used to boost the immune response by displaying binding epitopes in medical applications (7, 8). Like these miniproteins, which are generally less than 40-50 amino acids in length, ORF10 also has a well defined hydrophobic core and an alpha helical segment harboring immunogenic epitopes. Hence, together with the non-conservation of ORF10 at the sequence and structural level, this may all be the reason the human body may mount a high immune response to 2019-nCoV resulting in immunopathological conditions and subsequent consequences.

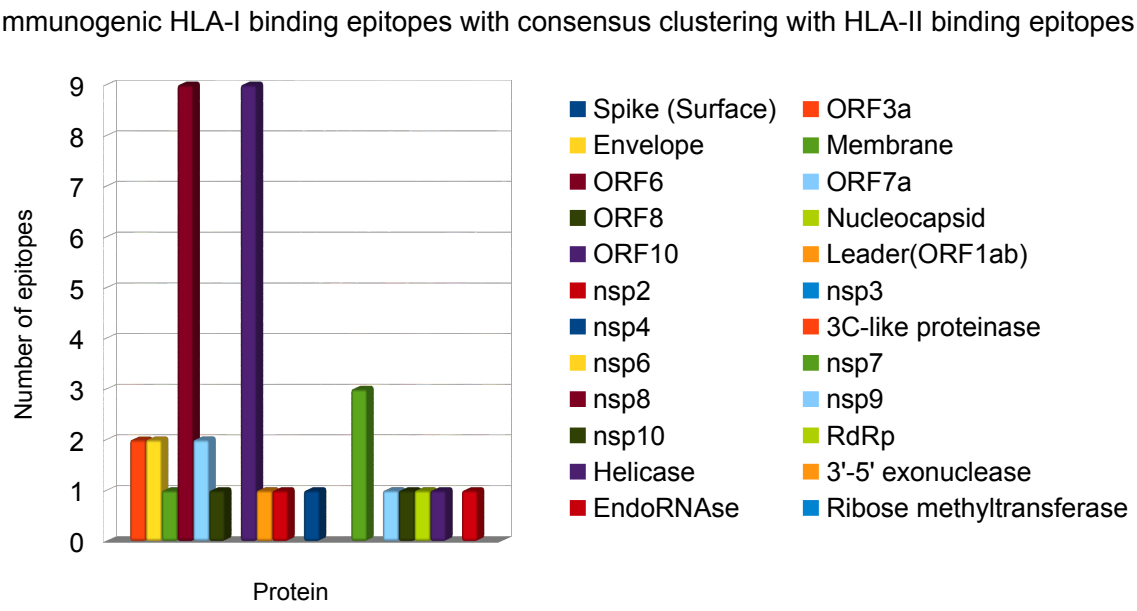
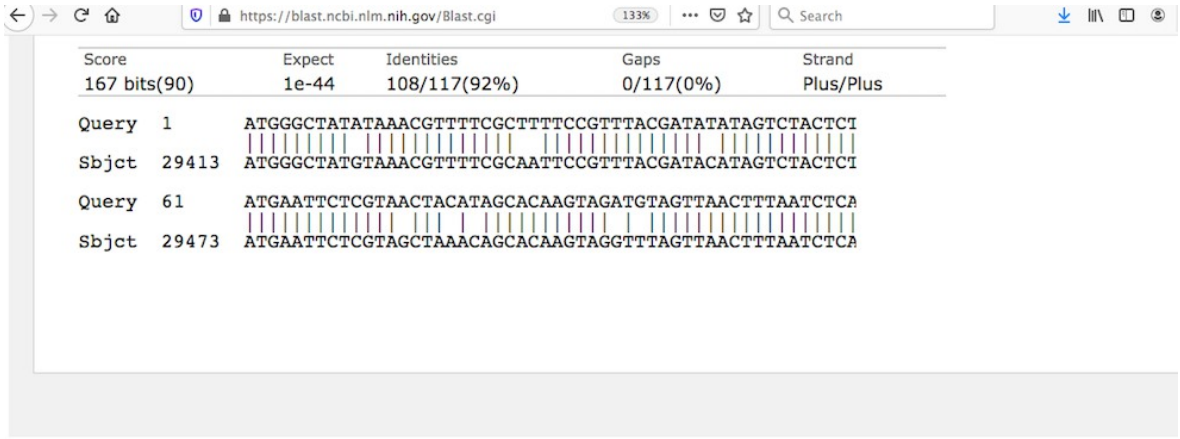


Fig. 4: Number of promiscuous immunogenic HLA-I binding epitopes across 2019-nCoV proteins studied clustered with HLA-II binding epitopes

Nucleotide Sequence Alignments:

Furthermore, there is no full sequence conservation of ORF10 in other closely related SARS and MERS viral species. It should be mentioned, however, that while the first pass in BLASTn did not tell anything, this author searched, found and aligned this region with a similar looking region in previous SARS-CoV sourced from humans with RefSeq accession number NC_004718.3. Out of 117 nucleotides in each of the two, there are 7 mutations (Fig. 5a). Further, probing of this sequence alignment with SARS-CoV sourced from bats (accession ID: KY417142) showed same mutations alongwith one more (Fig.5b)

a.



b.

Fig. 5 a: Alignment of SARS-CoV region (29415 to 29531, Query) with ORF10 region in 2019-nCoV (29558 to 29674 nucleotides, Sbjct) using BLASTn. b: Same 2019-nCoV ORF10 nucleotide sequence aligned with sequence sourced from bats accession ID: KY417142.

This sequence in SARS-CoV sourced from humans does not produce correct protein structure and has one ambiguous position marked with asterisk as seen in EMBOSS Translate Tool output. The translated amino acid sequence in SARS-CoV is MGYVNVFAIPFTIHSLLLCRMNSRN*TAQVGLVNFNLT, different from ORF10 protein sequence mentioned in the beginning of this paper.

In view of these analyses, this region may be considered an inserted tail region in the 2019-nCoV genome which may be responsible for the contagious nature of 2019-nCoV. The events leading to the accumulation of such nucleotides appear to be novel/different and so, need to be studied in greater detail. These key theoretical studies await further confirmation by *in vitro* and *in vivo* experiments.

Materials and Methods:

Genome sequence:

RefSeq sequences of all ten 2019-nCoV proteins were retrieved. Specifically, the accession numbers were as follows: ORF10 protein (YP_009725255.1), nucleocapsid phosphoprotein (YP_009724397.2), ORF8 protein (GenBank: QID21074.1), ORF7a protein (YP_009724395.1), ORF6 protein (YP_009724394.1), membrane glycoprotein (YP_009724393.1), envelope protein (YP_009724392.1), ORF3a protein (YP_009724391.1), surface glycoprotein (YP_009724390.1), ORF1ab (Polyprotein accession number YP_009724389.1 and the proteins therein.

Cytotoxic T cell epitopes prediction:

Nonameric peptide epitopes were selected using NetCTLpan version 1.1 (http://www.cbs.dtu.dk/services/NetCTLpan/ (9)) and PickPocket version 1.1 (http://www.cbs.dtu.dk/services/PickPocket/(10)) with default parameters. NetCTLpan uses a combined version of three different methods to identify immunogenic epitopes: HLA-I binding, TAP transporter binding and C-terminal cleavage predictions. PickPocket works on the basis of position-specific weight matrices and NetCTL uses neural network algorithm. Epitopes from NetCTLpan were ranked according to the combined score using all three different methods, and epitopes from PickPocket algorithm were sorted by affinity (IC_{50} values in nM). In order to increase prediction accuracy, high scoring epitopes common to both these algorithms (among top 10 in PickPocket and same epitopes among high scoring ones in NetCTLpan) were fished out. 12 HLA supertypes as present in both algorithms were used (2). For ORF1ab proteins, promiscuous epitopes were selected among top 30 candidates, as not many common epitopes could be found from NetCTLpan and PickPocket.

Helper T cell epitope prediction:

NetMHCIIpan version 3.2 (http://www.cbs.dtu.dk/services/NetMHCIIpan/(11)) was used to predict and design helper T cell epitopes across several HLA-DRB1 alleles. Quantitative MHC-peptide binding affinity data obtained from the Immune Epitope Database is used in this algorithm to generate predictions. From a consensus list of 15 amino acids long epitopes, top ranked epitopes were sorted using descending order of percent rank. As per this tool paper (11), those epitopes with %rank <2% were considered strong binders, and those with %rank <10% were considered weak binders.

Immunogenicity prediction:

All the CTL epitopes were predicted for their immunogenicity using IEDB Immunogenicity tool (12). Physicochemical properties of amino acids and their positions in the predicted peptide, including amino acids with large and aromatic side chains and positions 4-6 are used to predict potential epitopes. Ranking was done from higher to lower immunogenicity score as per the authors' guidelines.

Clustering

IEDB epitope cluster analysis tool was applied to group all HLA-I and HLA-II epitopes in clusters (13). Minimum sequence identity threshold was 70% and cluster-break algorithm was applied.

Structural Modeling

Ab-initio modeling conditions were applied to the ORF10 RefSeq sequence in view of no homology. The web server QUARK (https://zhanglab.ccmb.med.umich.edu/QUARK/) which predicts a structure, as per their paper (14), "replica-exchange Monte Carlo simulation under the guide of an atomic-level knowledge-based force field" is applied.

Secondary structure plot was generated by PDBsum (15) and the first structural model from QUARK was used as an input.

Acknowledgments:

This author acknowledges help in the form of nucleotide/protein sequences deposited in GenBank by several groups.

Conflict of interest: This author declares that there is no conflict of interest.

References:

1. Makałowski W., Gotea V., Pande A., Makałowska I. (2019) Transposable Elements: Classification, Identification, and Their Use As a Tool For Comparative Genomics. In: Anisimova M. (eds) Evolutionary Genomics. Methods in Molecular Biology, vol 1910. Humana, New York, NY

2. Mishra, S. T Cell Epitope-Based Vaccine Design for Pandemic Novel Coronavirus 2019-nCoV. *ChemRxiv.* Preprint. https://doi.org/10.26434/chemrxiv.12029523.v2 (2020).

3. I Berkower, G K Buckenmeyer, J A Berzofsky. Molecular mapping of a histocompatibility-restricted immunodominant T cell epitope with synthetic and natural peptides: implications for T cell antigenic structure. *The Journal of Immunology* April 1, 136 (7) 2498-2503 (1986).

4 Yusim, K., Kesmir, C et al. Clustering patterns of cytotoxic t-lymphocyte epitopes in human immunodeficiency virus type 1 (HIV-1) proteins reveal imprints of immune evasion on HIV-1 global variation. *Journal of Virology*, **76:** 8757–8768 (2002).

5. Fan Wu et al. Neutralizing antibody responses to SARS-CoV-2 in a COVID-19 recovered 1 patient cohort and their implications, *MedRxiv.* https://doi.org/10.1101/2020.03.30.20047365 (2020).

6. Haveri, A. et al. Serological and molecular findings during SARS-CoV-2 infection: the first case study in Finland, January to February 2020 *Euro Surveill.* **25**(11): 2000266 (2020).

7. Zoller, F., Haberkorn, U and Mier, W. Miniproteins as Phage Display-Scaffolds for Clinical

Applications *Molecules* ***16*****(3)**, 2467-2485 (2011).

8. Skerra, A. Engineered protein scaffolds for molecular recognition. *J. Mol. Recognit.* **13,** 167–187 (2000)

9.Stranzl, T., Larsen, M. V., Lundegaard, C., & Nielsen, M. . NetCTLpan: pan-specific MHC class I pathway epitope predictions. *Immunogenetics*, ***62***(6), 357–368 (2010). https://doi.org/10.1007/s00251-010-0441-4

10. Zhang, H., Lund, O., & Nielsen, M. The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding. *Bioinformatics (Oxford, England)*, **25**(10), 1293–1299. https://doi.org/10.1093/bioinformatics/btp137 (2009).

11. Karosiene, E., Rasmussen, M., Blicher, T., Lund, O., Buus, S., & Nielsen, M. NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. *Immunogenetics*, **65**(10), 711–724. https://doi.org/10.1007/s00251-013-0720-y (2013).

12. Calis JJA, Maybeno M, Greenbaum JA, Weiskopf D, De Silva AD, Sette A, Kesmir C, Peters B. Properties of MHC class I presented peptides that enhance immunogenicity. *PloS Comp. Biol.* **8(1)**:361. (2013).

13. Dhanda,S.K., Vaughan, K., Schulten, V., et al. Development of a novel clustering tool for linear peptide sequences. *Immunology* doi:https://doi.org/10.1111/imm.12984 (2018).

14. D. Xu and Y. Zhang. Toward optimal fragment generations for ab initio protein structure assembly. *Proteins*, **81**: 229-239 (2013).

15. R. A. Laskowski, V. V. Chistyakov, and J. M. Thornton. PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids
Nucleic Acids Res **33**, D266-D268 (2005).