 doi:10.20944/preprints202003.0409.v1

Structural similarity analysis of the spike protein of SARS-CoV-2 and other SARS-related coronaviruses

YoungJoon Park [1,5,#], Ju Won Ahn [1,#], Sojung Hwang [4], Kyoung Su Sung [3], Jaejoon Lim [2,*], KyuBum Kwack [1,*]

[1] Institute Department of Biomedical Science, College of Life Science, CHA University

[2] Department of Neurosurgery, Bundang CHA Medical Center, CHA University

[3] Department of Neurosurgery, Dong-A University Hospital, Dong-A University College of Medicine

[4] Global Research Supporting Center, Bundang CHA Medical Center, CHA University

[5] DERMAY Research Center, Dongtan

[#] Equally contributed in this study as first authors.

***Correspondence to** KyuBum Kwack, Department of Biomedical Science, College of Life Science, CHA University, Seongnam, Gyeonggi-do, Republic of Korea

Tel: +82-31-881-7141, e-mail: kbkwack@cha.ac.kr

Jaejoon Lim, Department of Neurosurgery, Bundang CHA Medical Center, CHA University, Yatap-dong 59, Seongnam 13496, Republic of Korea

Tel: +82-31-780-5688, Fax: 82-31-780-5269, e-mail: coolppeng@naver.com

 doi:10.20944/preprints202003.0409.v1

Abstract

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has high infectivity in humans, attributed to the strong affinity of its spike (S) protein to angiotensin-converting enzyme 2. Here, we analysed the structural similarity of the S protein between SARS-CoV-2 and other SARS-related coronaviruses. The S1 domain of the unclassified coronavirus RaTG13 was structurally very similar to that of SARS-CoV-2, implying that RaTG13 could be the origin of SARS-CoV-2.

Keywords: angiotensin-converting enzyme 2; SARS-CoV-2; spike protein

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 March 2020 doi:10.20944/preprints202003.0409.v1

In December 2019, a highly infectious novel coronavirus emerged in Wuhan, China, which was named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1]. This virus has infected more than 200,000 people worldwide in just three months since the outbreak in Wuhan. Based on phylogeny, SARS-CoV-2 has been included in the species *Severe acute respiratory syndrome-related coronavirus* (SARSr-CoV) and the genus *Betacoronavirus* [1].

Like SARS-CoV, the entry of SARS-CoV-2 into a host cell is facilitated by the binding of the spike (S) protein to angiotensin-converting enzyme 2 (ACE2) [2]. The similarity of the amino acid sequence of the S protein between SARS-CoV and SARS-CoV-2 is about 76%, with a very high degree of homology [3]. Importantly, SARS-CoV-2 is reported to have a much higher human-to-human transmission through ACE2 binding compared to SARS-CoV [4], suggesting a stronger binding affinity of the S protein of SARS-CoV-2 to ACE2.

A recent study suggested that an unclassified coronavirus, named RaTG13, from Chinese bats has the highest sequence similarity with SARS-CoV-2, with 96% identity at the whole-genome level [5]. However, it is not yet proven that RaTG13 is the origin of SARS-CoV-2. Here, we hypothesized that if RaTG13 is the origin of the highly contagious SARS-CoV-2, it will share certain structural characteristics of the S protein.

We downloaded the S protein sequence in fasta format from the NCBI database and analysed it with S protein sequences isolated from nine hosts infected with SARS-CoV-2, SARS-CoV, SARS-like CoV, or RaTG13 (Table 1). On the basis of phylogeny analysis with neighbour joining and 100 bootstrap iterations, the amino acid sequence of the S protein of RaTG13 was found to be the most similar to that of SARS-CoV-2 (Figure 1A). Several insertions were indicated in SARS-CoV, SARS-CoV-2, and RaTG13 (Figure 1B). Importantly, the amino acid sequence between the positions 331 and 583, representing the receptor-binding domain (RBD) of the S protein of SARS-CoV-2, had more similarity with RaTG13 compared with SARS-CoV (Figure 1B). This indicates that the RBD of the RaTG13 S protein might be

 doi:10.20944/preprints202003.0409.v1

structurally similar to that of SARS-CoV-2. Therefore, the structural similarity between a S protein sequence of SARS-CoV-2 was compared with NP_828851.1 (SARS-CoV), AVP78042.1 (SARS-like CoV, bat-SL-CoVZXC21), and QHR63300.2 (Unclassified CoV, RaTG13). For this, protein structures of the four sequences were predicted by SWISS-MODEL, which involves alignment of a target sequence and template structure [7-11]. We used the template with S protein (PDB ID: 6acd.1.A) [12], which was analysed by electron microscopy. In the RBD of the S protein, an insertion event appears to have occurred in SARS-CoV and SARS-CoV-2 (Figure 2A). However, the inserted sequence of SARS-CoV was very different from that of SARS-CoV-2 or RaTG13 (Figure 2A), while the inserted sequences of SARS-CoV-2 and RaTG13 were very similar (Figure 2A). The three-dimensional (3D) structure of the RBD was similar to those of SARS-CoV, RaTG13, and SARS-CoV-2 (Figure 2B). To quantitatively evaluate the structural similarity, we calculated the template modelling (TM) score between SARS-CoV-2 and each of the three proteins using the web-based software TM-Score [6]. The TM score of RaTG13 against SARS-CoV-2 was 0.8401, while the TM scores of SARS-CoV and SARS-like CoV against SARS-CoV-2 were <0.5. A TM score >0.5 indicates the same fold between two amino acid sequences. In addition, interestingly, the distance between most residue pairs of the S1 domain (but not S2) of the S protein between SARS-CoV-2 and RaTG13 was similar (<5.0 Å) (Figure 3).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 March 2020 doi:10.20944/preprints202003.0409.v1

Table 1 Details of the coronavirus sequences compared in this analysis

Locus	Protein ID	Virus	Host	Country	Isolate	Collection date
NC_045512	YP_009724390.1	SARS-CoV-2	*Homo sapiens*	China	Wuhan-Hu-1	2019
NC_004718	NP_828851.1	SARS-CoV	*Homo sapiens*	Canada	Tor2	2003
DQ182595	ABA02260.1	SARS-CoV	*Homo sapiens*	China	ZJ0301	2003
MN996532	QHR6330.2	Unclassified CoV	*Rhinolophus affinis*	China	RaTG13	2013
MT126808	QIG55994.1	SARS-CoV-2	*Homo sapiens*	Brazil	BRA	2020
MT123291	QIE07461.1	SARS-CoV-2	*Homo sapiens*	China	IQTC02	2020
MT049951	QIA20044.1	SARS-CoV-2	*Homo sapiens*	China	Yunnan-01	2020
MG772934	AVP78042.1	SARS-like-CoV	*Rhinolophus sinicus*	China	bat-SL-CoVZXC21	2015
MG772933	AVP78031.1	SARS-like-CoV	*Rhinolophus sinicus*	China	bat-SL-CoVZC45	2017

Locus: Nucleotide accession ID in the NCBI database

 doi:10.20944/preprints202003.0409.v1

In conclusion, there are two major domains, including S1 and S2, in the S protein. The S1 domain of the S protein contains the RBD, which is reported to bind to ACE2 directly. Our results reveal that the structure of the S1 domain in RaTG13 is very similar to that of SARS-CoV-2. Functionally, RaTG13 is likely to be the origin of SARS-CoV-2 because of the close similarity of the S1 domain, which is associated with the high-infectivity characteristic of SARS-CoV-2.

 doi:10.20944/preprints202003.0409.v1

References

1. Coronaviridae Study Group of the International Committee on Taxonomy of V. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat Microbiol.* 2020.
2. Hoffmann M, Kleine-Weber H, Schroeder S, Kruger N, Herrler T, Erichsen S, et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell.* 2020.
3. Xu X, Chen P, Wang J, Feng J, Zhou H, Li X, et al. Evolution of the novel coronavirus from the ongoing Wuhan outbreak and modeling of its spike protein for risk of human transmission. *Sci China Life Sci.* 2020;63(3):457-60.
4. Wan Y, Shang J, Graham R, Baric RS, and Li F. Receptor recognition by novel coronavirus from Wuhan: An analysis based on decade-long structural studies of SARS. *J Virol.* 2020.
5. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature.* 2020;579(7798):270-3.
6. Zhang Y, and Skolnick J. Scoring function for automated assessment of protein structure template quality. *Proteins.* 2004;57(4):702-10.
7. Benkert P, Biasini M, and Schwede T. Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics.* 2011;27(3):343-50.
8. Bertoni M, Kiefer F, Biasini M, Bordoli L, and Schwede T. Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. *Sci Rep.* 2017;7(1):10480.
9. Bienert S, Waterhouse A, de Beer TA, Tauriello G, Studer G, Bordoli L, et al. The SWISS-MODEL Repository-new features and functionality. *Nucleic Acids Res.* 2017;45(D1):D313-D9.
10. 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;46(W1):W296-W303.
11. Guex N, Peitsch MC, and Schwede T. Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: a historical perspective. *Electrophoresis.* 2009;30 Suppl 1:S162-73.
12. Song W, Gui M, Wang X, and 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 March 2020 doi:10.20944/preprints202003.0409.v1

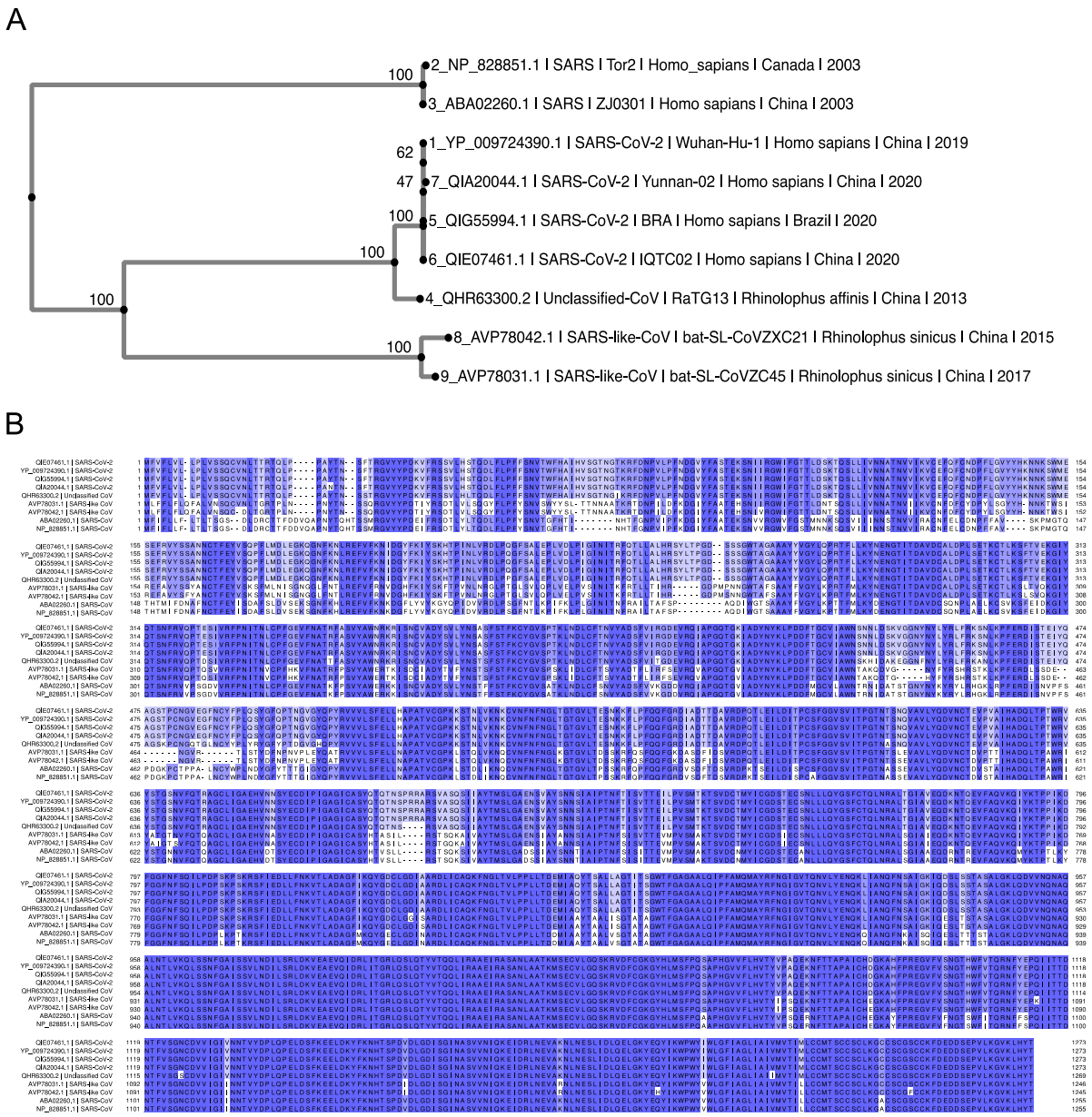
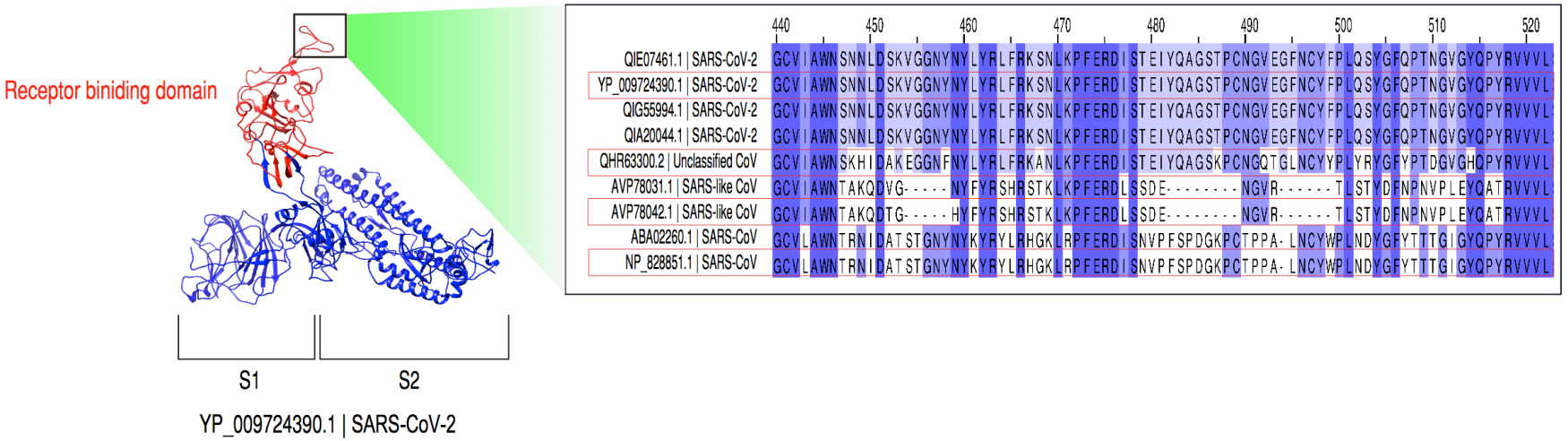


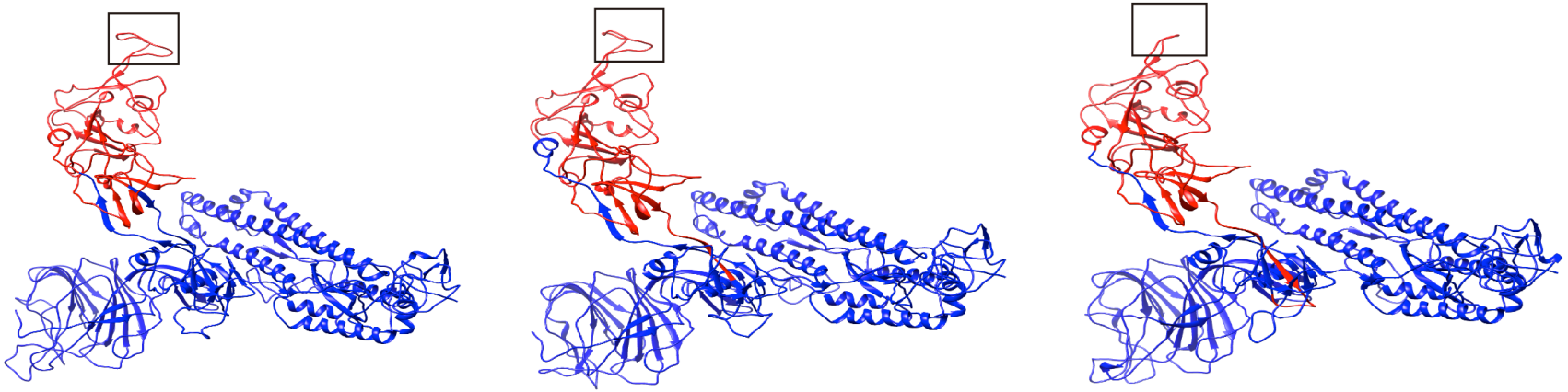
Figure 1 Comparisons of amino acid sequences of nine spike proteins from SARS-CoV-2, RaTG13, SARS-like, and SARS-CoV. Phylogenic tree generated using the neighbour-joining method and 100 bootstrap iterations (A). Multiple alignment of sequences (B).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 March 2020 doi:10.20944/preprints202003.0409.v1

A



B



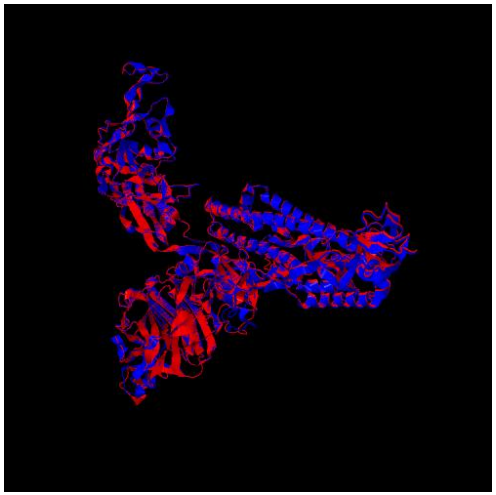
 doi:10.20944/preprints202003.0409.v1

Figure 2 Structure modelling of four spike protein sequences. Three-dimensional (3D) structure of spike protein chain A of SARS-CoV-2 and the insertion site in the receptor-binding domain (A), and 3D structure of spike proteins NP_828851.1 (SARS-CoV), AVP78042.1 (SARS-like CoV, bat-SL-CoVZXC21), and QHR63300.2 (Unclassified CoV, RaTG13) (B).

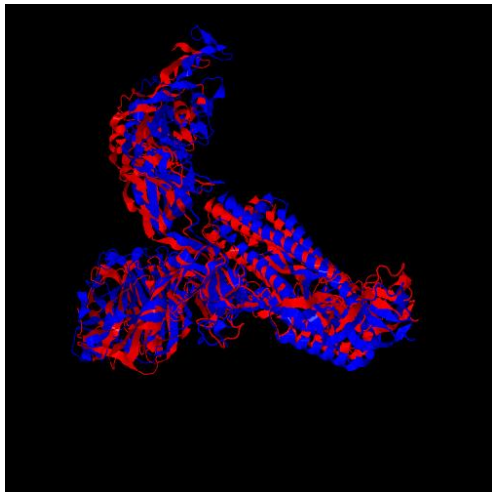
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 March 2020 doi:10.20944/preprints202003.0409.v1

A

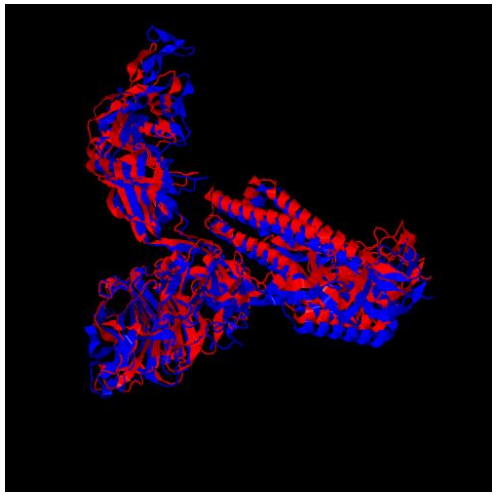
Red: QHR6330.2 | Unclassified CoV



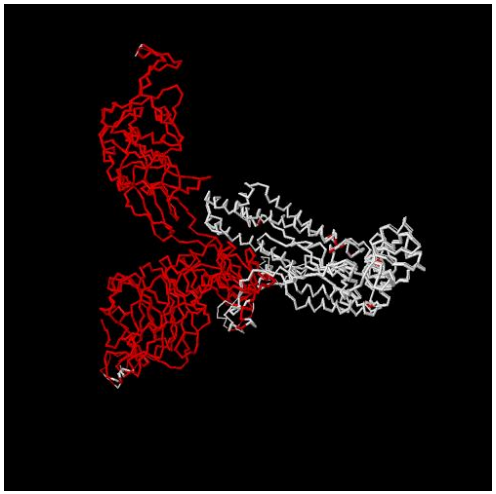
Red: NP_828851.1 | SARS-CoV

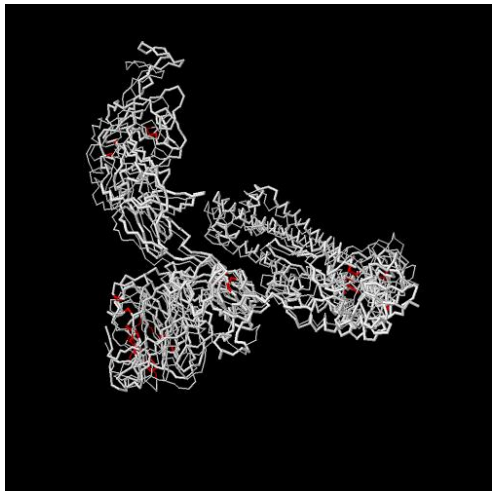


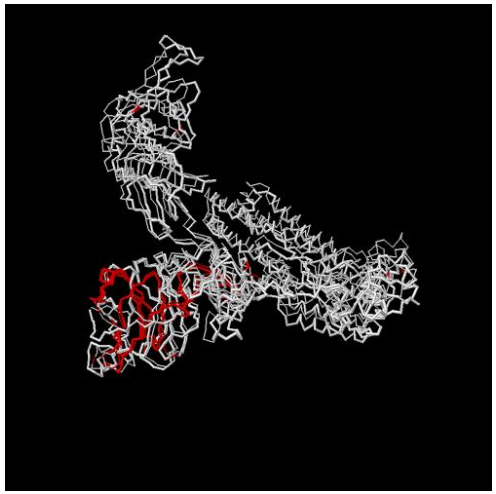
Red: AVP78042.1 | SARS-like-CoV



B







 doi:10.20944/preprints202003.0409.v1

Figure 3 3D superposition of the spike proteins of SARS-CoV-2 (YP_009724390.1) with NP_828851.1 (SARS-CoV), AVP78042.1 (SARS-like CoV, bat-SL-CoVZXC21), and QHR63300.2 (Unclassified CoV, RaTG13). The blue structure represents the spike protein of SARS-CoV-2 (A), and red colour represents a distance of less than 5 Å between residue pairs (B).