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

YoungJoon Park [a,e,#], Ju Won Ahn [a,#], Sojung Hwang [d], Kyoung Su Sung [c], Jaejoon Lim [b,*], KyuBum Kwack [a,*]

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

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

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

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

[e] 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.v2

Abstract

Objectives

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 human angiotensin-converting enzyme 2 (ACE2). Here, we analyzed the structural similarity of the S protein between SARS-CoV-2 and other SARS-related coronaviruses (CoVs).

Methods

We performed multiple alignment analysis of nine amino acid sequences of CoV S proteins from NCBI with MAFFT web-based software, followed by phylogeny analysis. Three-dimensional structure modeling was performed by SWISS-MODEL. We calculated the template modeling score between the S protein of SARS-CoV-2 and that of other SARS-related CoVs.

Results

The S1 domain of the unclassified CoV RaTG13 (the host of which is the intermediate horseshoe bat) was structurally very similar to that of SARS-CoV-2, implying that RaTG13 could be the origin of SARS-CoV-2. In addition, the folding property of the entire S protein was nearly the same between SARS-CoV-2 and RaTG13 after the PRRA amino acid insertion was removed from SARS-CoV-2.

Conclusions

RaTG13 could have a high binding affinity to ACE2, similar to SARS-CoV-2, and it is therefore highly likely to infect other animals. Therefore, massive research and monitoring of CoVs in animals is necessary to prevent future COVID-19-like disasters.

Keywords: angiotensin-converting enzyme 2; SARS-CoV-2; spike protein; COVID-19

 doi:10.20944/preprints202003.0409.v2

Introduction

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). As of 8 April, 2020, about 1.5 million people were infected and 88,000 people worldwide were died with this virus. On February 11, 2020, the World Health Organization (WHO) named the disease as coronavirus disease 2019 (COVID-19). In addition, on March 11, 2020, WHO declared COVID-19 a worldwide pandemic.

We hypothesized that if RaTG13, a unclassified coronavirus (CoV) from the intermediate horseshoe bat *Rhinolophus affinis,* is the origin of the highly contagious SARS-CoV-2, it might share certain common structural characteristics in the spike (S) protein. Therefore, we evaluated the possibility of RaTG13 being the origin of SARS-CoV-2 on the basis of the structural similarity of the S protein.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2020 doi:10.20944/preprints202003.0409.v2

Methods

Data collection

We downloaded S protein sequences isolated from nine hosts infected with SARS-CoV-2, SARS-CoV, SARS-like CoV, or RaTG13 in FASTA format from the NCBI database for analysis (Table 1).

Phylogeny and three-dimensional structure and similarity analysis

To analyze the phylogeny, we performed multiple alignment analysis of the nine amino acid sequences with MAFFT web-based software (2) by neighbor joining with 100 bootstrap iterations. Three-dimensional (3D) structure modeling was performed for four amino acid sequences, including SARS-CoV-2 (YP_009724390.1), RaTG13 (QHR63300.2), SARS-CoV (NP_828851.1), and bat-SL-CoVZXC21 (AVP78042.1) by SWISS-MODEL, which involves the alignment of a target sequence and template structure (3-7). We used the template with S protein (PDB ID: 6acd.1.A) (8), which was analyzed by electron microscopy. We calculated the template modeling (TM) score between SARS-CoV-2 and each of the three proteins using the web-based software TM-Score (9).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2020 doi:10.20944/preprints202003.0409.v2

Results

Evolutionary closeness of the S protein amino acid sequence between SARS-CoV-2 and RaTG13

On the basis of phylogeny, 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 structurally similar to that of SARS-CoV-2. 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).

Structural similarity of the S protein between SARS-CoV-2 and RaTG13

The 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 structural similarity. As a result, the 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 TM score between SARS-CoV-2 and each of the three proteins using the software TM-score (9). A TM score of >0.5 indicates the same fold between two amino acid sequences. 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. In addition, interestingly, the distance between most residue pairs of the S1 domain (but not the S2 domain) of the S protein between SARS-CoV-2 and RaTG13 was similar (<5.0 Å) (Figure 3).

 doi:10.20944/preprints202003.0409.v2

The new insertion event altered the structure of the S protein of RaTG13 and SARS-CoV-2

A new amino acid insertion event of proline-arginine-arginine-alanine (PRRA) was detected between RaTG13 and SARS-CoV-2 (Figure 1B). To evaluate the effect of the PRRA sequence on the structural similarity between SARS-CoV-2 and RaTG13, we calculated the TM score between the amino acid sequences of the S protein of RaTG13 and SARS-CoV-2 with the PRRA sequence removed. As a result, most of the amino acid pairs between SARS-CoV-2 and RaTG13 had the same structural fold (TM score: 0.9931) (Figure 4).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2020 doi:10.20944/preprints202003.0409.v2

Discussion

COVID-19 is a high infectious disease caused by SARS-CoV-2, and it is fatal in some patients owing to lung injury or cytokine storm. In addition, at this point, 1.5 million confirmed cases have been reported globally, with 88,000 deaths. The gravity of the situation necessitates the identification of the origin of this catastrophic event. A recent paper proposed two scenarios regarding the origin of SARS-CoV-2 and its spread (10). First, SARS-CoV-2 might have been naturally selected in animals before zoonotic transfer to humans, or second, it might have been transmitted to humans from an animal and then naturally selected in humans. If the first case holds true, a disease like SARS-CoV-2 might appear again in humans in the future because the original animal host is likely to have already passed it on to other animals. In the second case, the possibility of its eradication from the animal is unlikely because natural selection occurred in humans.

CoV has an envelope protein and uses single-stranded RNA as genetic information. CoVs are divided into four genera: alpha, beta, gamma and delta. On the basis of phylogeny, SARS-CoV-2 has been included in the species *Severe acute respiratory syndrome-related coronavirus* (SARSr-CoV) within the genus *Betacoronavirus* (1). The structural proteins of CoV are divided into four types, including spike (S), envelope (E), membrane (M), and nucleocapsid (N) (11). Among the four proteins, the S protein in essential for entry into host cells via angiotensin-converting enzyme 2 (ACE2). Like SARS-CoV, the entry of SARS-CoV-2 into a host cell is facilitated by the binding of the S protein to ACE2 (12). 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 (13). Importantly, SARS-CoV-2 is reported to have a much higher human-to-human transmission through ACE2 binding compared to SARS-CoV (14), 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 intermediate

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2020 doi:10.20944/preprints202003.0409.v2

horseshoe bats (*Rhinolophus affinis*) in China, has the highest sequence similarity with SARS-CoV-2, with 96% identity at the whole-genome level (15). However, it is not yet proven that RaTG13 is the origin of SARS-CoV-2. In this study, we compared the 3D structure of the S protein between SARS-CoV-2 and other SARS-related other coronaviruses. Structurally, the S1 domain of the S protein of SARS-CoV-2 was found to be very similar with that of RaTG13 (<5.0 Å) (Figures 2 and 3).

In addition, the novel PRRA insertion in only SARS-CoV-2 directly altered S protein structure between SARS-CoV-2 and RaTG13 (Figure 4). 3D superposition of the spike proteins between SARS-CoV-2 from which the PRRA sequence was removed and QHR63300.2 (Unclassified CoV, RaTG13) were very similar (TM score: 0.9931). Most of the folding property of the entire S protein between YP_009724390.1 (SARS-CoV-2) lacking the PRRA sequence, and QHR63300.2 (Unclassified CoV, RaTG13) was the same (TM score: 0.9931), with a distance of less than 5 Å between residue pairs (Figure 4).

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. In a recent study, it was noted that due to the diversity of the RBD of the S protein between RaTG13 and SARS-CoV-2, the S protein of RaTG13 would have a lower affinity for ACE2 (10). However, our results reveal that the structure of the S1 domain in RaTG13 is very similar to that of SARS-CoV-2 (Figure 3). 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. According to a recent report, the RBD of the S protein of SARS-CoV-2 is different from that of SARS-CoV, and the binding affinity to ACE2 is stronger than that of SARS-CoV (14). As a result, the RaTG13 collected in 2013 already has a very similar receptor binding site to ACE2 for SARS-CoV-2. It is therefore likely that a virus with a similar infectivity as SARS_CoV-2 is spreading in nature. In addition, as shown in Figure 4, on removing the PRRA in SARS-CoV-2 and super-positioning it with RaTG13, the folding

 doi:10.20944/preprints202003.0409.v2

similarity of both S proteins is almost similar. This suggests that RaTG13 is the origin of SARS-CoV-2, considering the similarity of the S protein, which is most closely related to infectivity.

Conclusions

We revealed that the RBD of the S protein of SARS-CoV-2 was structurally different from that of SARS-CoV and similar to that of RaTG13. This indicates that RaTG13 from bats could have a high binding affinity to ACE2, similar to that of SARS-CoV-2, and it is highly likely to infect other animals. Therefore, it is necessary to monitor viruses naturally circulating in animals to prevent further spillovers that can lead to disasters such as COVID-19.

 doi:10.20944/preprints202003.0409.v2

Acknowledgments

I appreciate Chul Lee, IPBI, SNU, South Korea, comments for this study.

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 Mar 2. PubMed PMID: 32123347. Epub 2020/03/04.
2. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform. 2019 Jul 19;20(4):1160-6. PubMed PMID: 28968734. Pubmed Central PMCID: PMC6781576. Epub 2017/10/03.
3. Benkert P, Biasini M, Schwede T. Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics. 2011 Feb 1;27(3):343-50. PubMed PMID: 21134891. Pubmed Central PMCID: PMC3031035. Epub 2010/12/08.
4. Bertoni M, Kiefer F, Biasini M, Bordoli L, Schwede T. Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. Sci Rep. 2017 Sep 5;7(1):10480. PubMed PMID: 28874689. Pubmed Central PMCID: PMC5585393. Epub 2017/09/07.
5. 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 Jan 4;45(D1):D313-D9. PubMed PMID: 27899672. Pubmed Central PMCID: PMC5210589. Epub 2016/12/03.
6. 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 2;46(W1):W296-W303. PubMed PMID: 29788355. Pubmed Central PMCID: PMC6030848. Epub 2018/05/23.
7. Guex N, Peitsch MC, Schwede T. Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: a historical perspective. Electrophoresis. 2009 Jun;30 Suppl 1:S162-73. PubMed PMID: 19517507. Epub 2009/06/12.
8. 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 Aug;14(8):e1007236. PubMed PMID: 30102747. Pubmed Central PMCID: PMC6107290. Epub 2018/08/14.
9. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins. 2004 Dec 1;57(4):702-10. PubMed PMID: 15476259. Epub 2004/10/12.
10. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nature Medicine. 2020 2020/03/17.
11. Bosch BJ, van der Zee R, de Haan CA, Rottier PJ. The coronavirus spike protein is a class I virus fusion protein: structural and functional characterization of the fusion core complex. J Virol. 2003 Aug;77(16):8801-11. PubMed PMID: 12885899. Pubmed Central PMCID: PMC167208. Epub 2003/07/30.
12. 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 Mar 4. PubMed PMID: 32142651. Epub 2020/03/07.
13. 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 Mar;63(3):457-60. PubMed PMID: 32009228. Epub 2020/02/06.

 doi:10.20944/preprints202003.0409.v2

14. Wan Y, Shang J, Graham R, Baric RS, Li F. Receptor Recognition by the Novel Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural Studies of SARS Coronavirus. J Virol. 2020 Mar 17;94(7). PubMed PMID: 31996437. Pubmed Central PMCID: PMC7081895. Epub 2020/01/31.

15. 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 Mar;579(7798):270-3. PubMed PMID: 32015507. Epub 2020/02/06.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2020 doi:10.20944/preprints202003.0409.v2

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

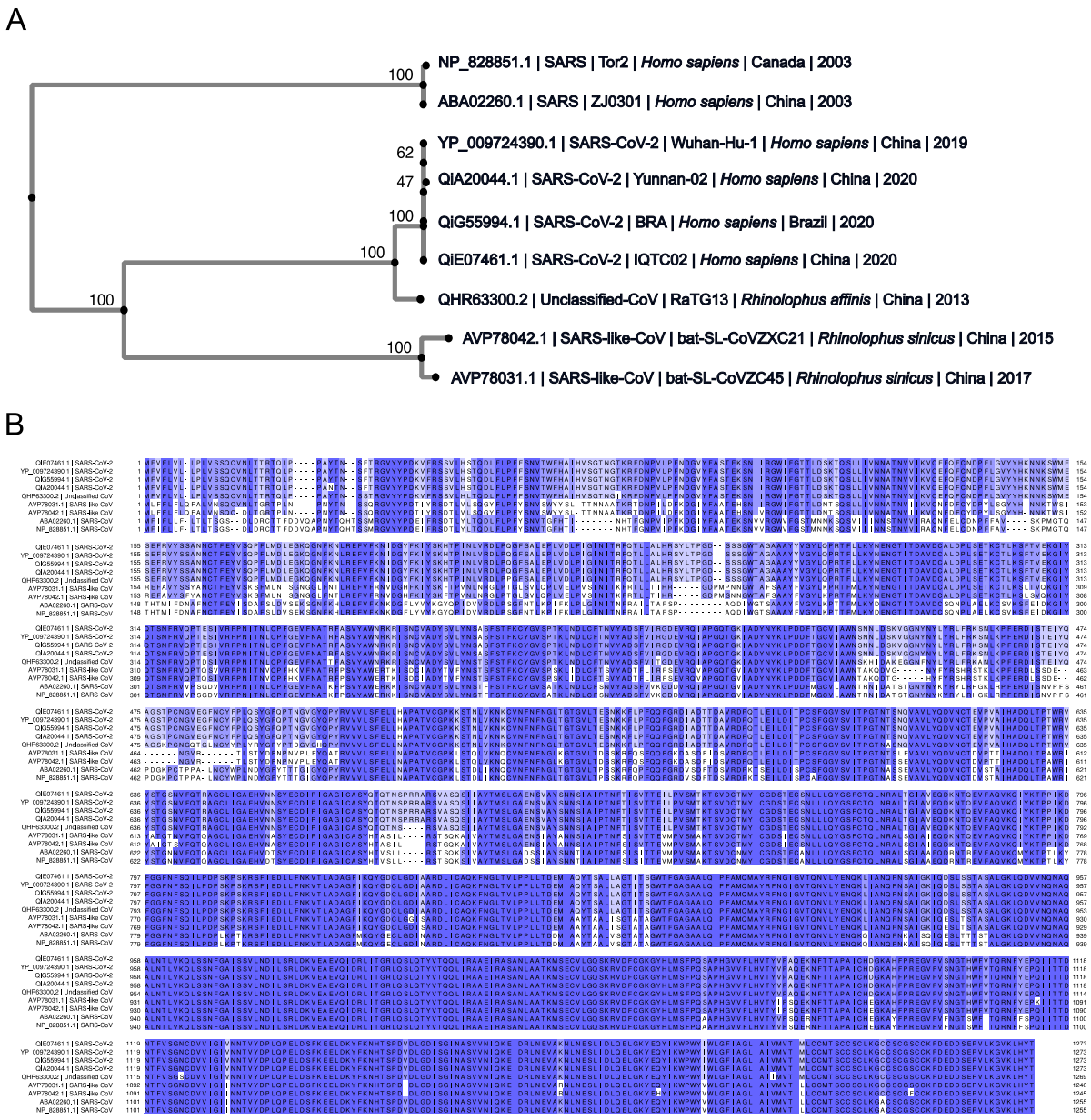
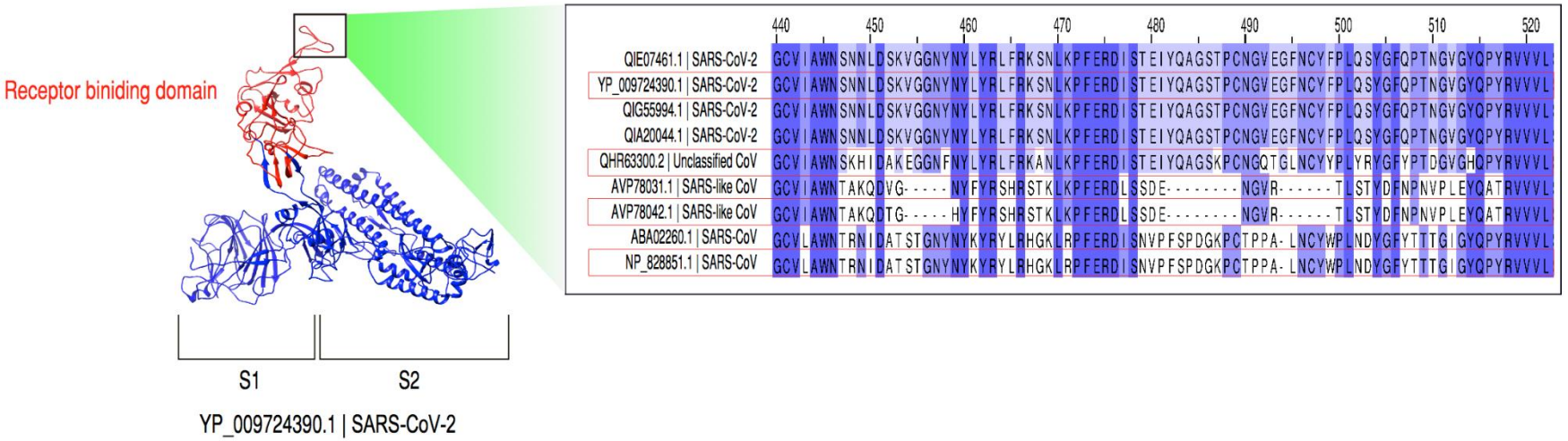


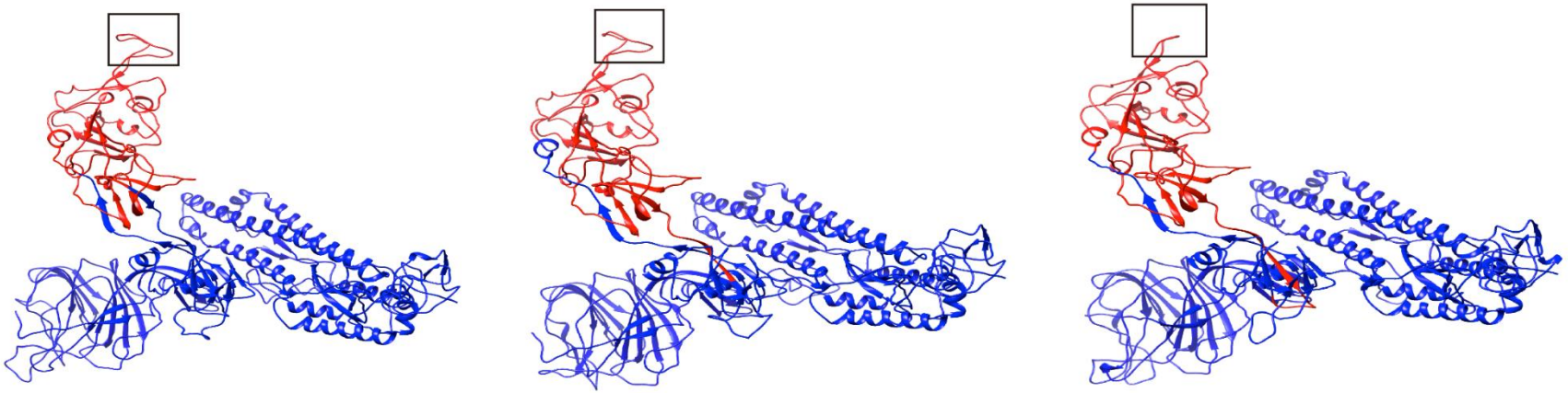
Figure 1 Comparisons of amino acid sequences of nine spike proteins from SARS-CoV-2, RaTG13, SARS-like, and SARS-CoV. Phylogenic tree generated using the neighbor-joining method and 100 bootstrap iterations (A). Multiple alignment of sequences (B).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2020 doi:10.20944/preprints202003.0409.v2

A



B



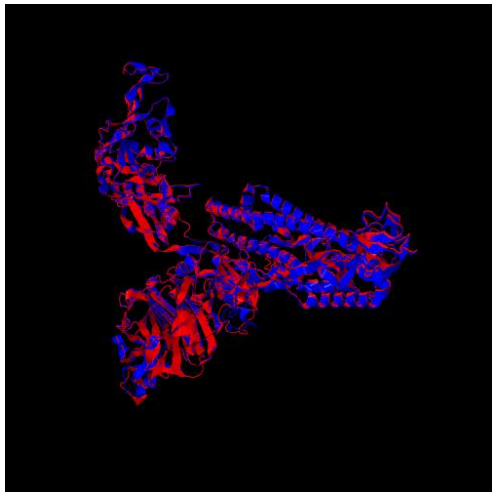
 doi:10.20944/preprints202003.0409.v2

Figure 2 Structure modeling 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 structures of spike proteins NP_828851.1 (SARS-CoV), AVP78042.1 (SARS-like CoV, bat-SL-CoVZXC21), and QHR63300.2 (Unclassified CoV, RaTG13) (B).

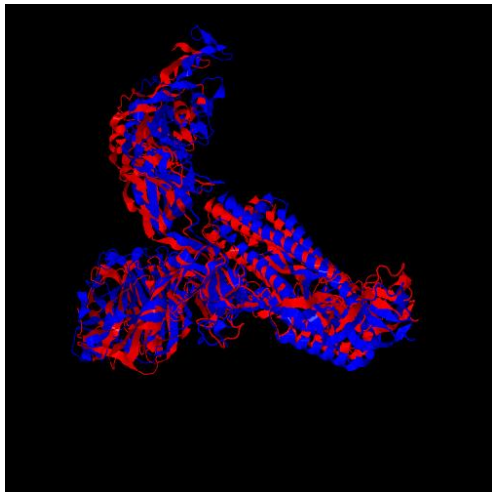
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2020 doi:10.20944/preprints202003.0409.v2

A

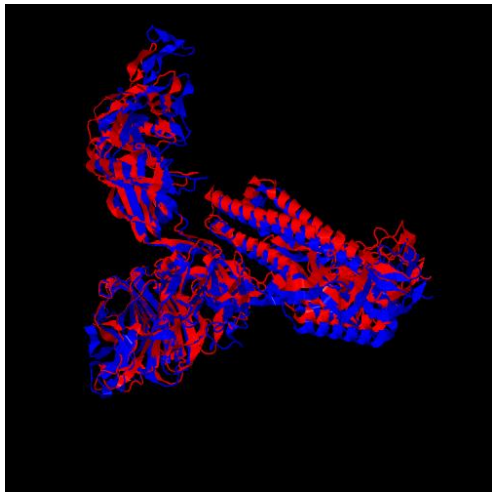
Red: QHR6330.2 | Unclassified CoV



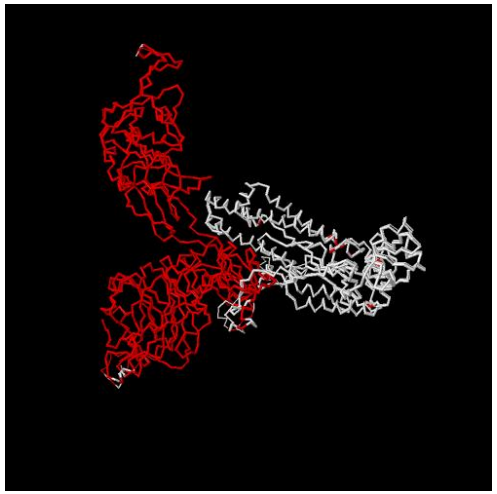
Red: NP_828851.1 | SARS-CoV

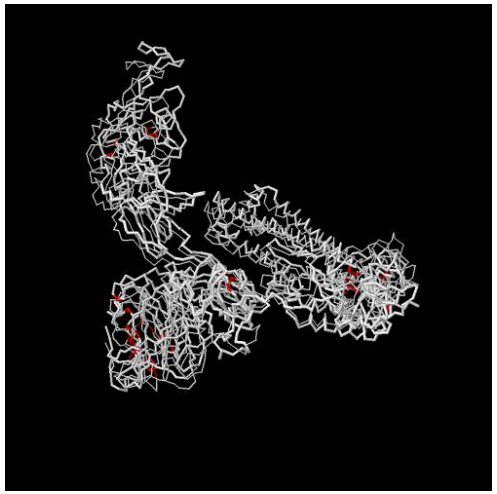


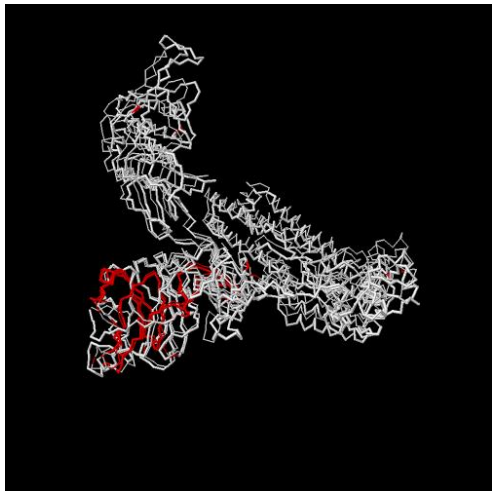
Red: AVP78042.1 | SARS-like-CoV



B







 doi:10.20944/preprints202003.0409.v2

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 color represents a distance of less than 5 Å between residue pairs (B)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 April 2020 doi:10.20944/preprints202003.0409.v2

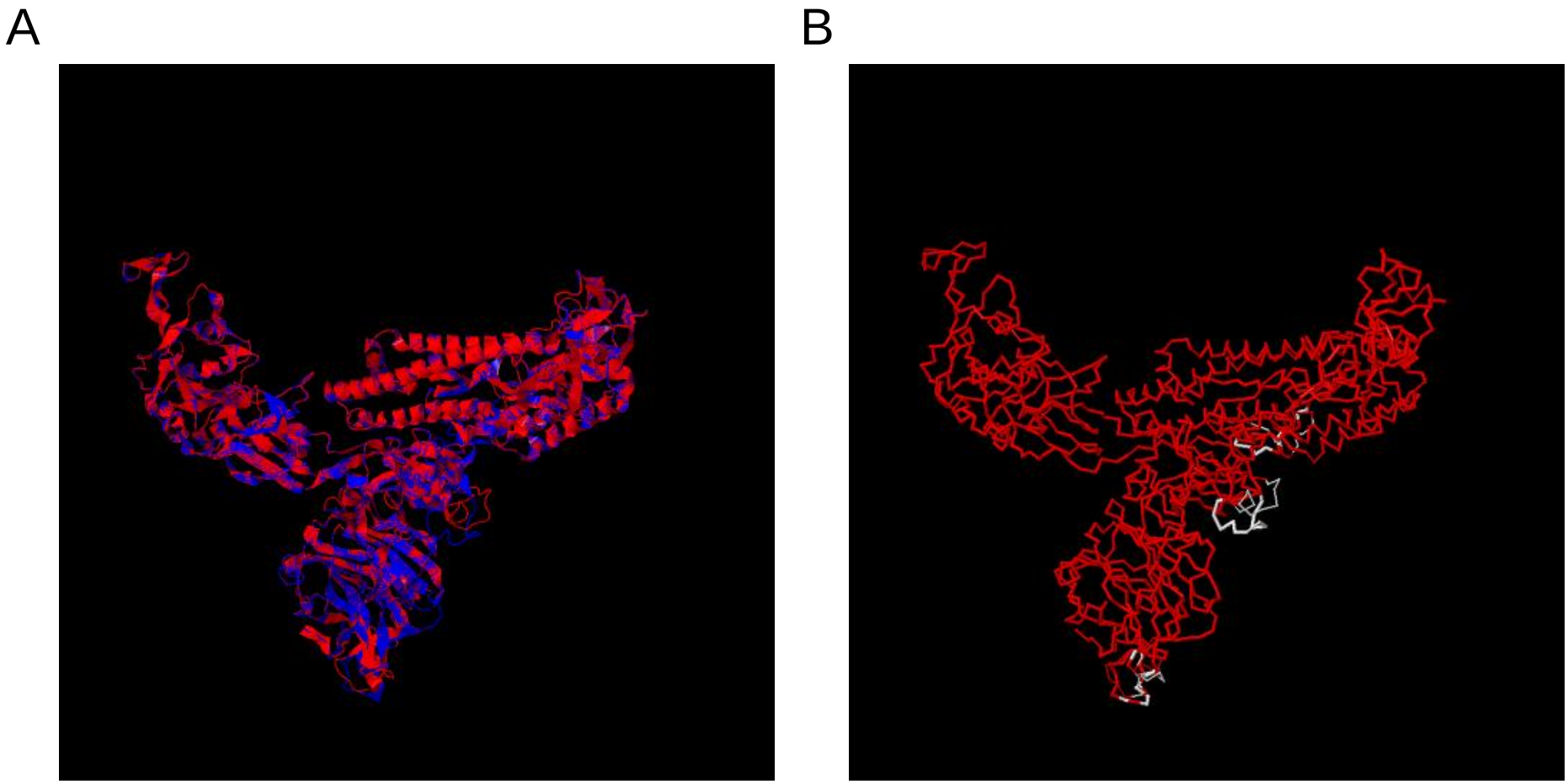


Figure 4 3D superposition of the spike proteins between SARS-CoV-2 (YP_009724390.1) from which the PRRA sequence was removed and QHR63300.2 (Unclassified CoV, RaTG13). The blue structure represents the spike protein of SARS-CoV-2 (A), and red color represents a distance of less than 5 Å between residue pairs (B).