 doi:10.20944/preprints202005.0026.v1

Sequence analysis and amino acid variations of structural proteins deduced from novel coronavirus SARS-CoV-2 strains, isolated in different countries

Tarlan Mamedov[1, 2*], Inanc Soylu, Gunay Mammadova[1] & Gulnara Hasanova[1]

1. Akdeniz University, Department of Agricultural Biotechnology, Dumlupınar Boulevard 07058 Campus, Antalya, Turkey

2. Institute of Molecular Biology and Biotechnology, Azerbaijan National Academy of Sciences, Matbuat Avenue 2a, Baku 1073, Azerbaijan

Correspondence to: tmammedov@gmail.com

ABSTRACT

SARS-CoV-2 is a novel and highly pathogenic coronavirus, which was first diagnosed in Wuhan city, China, in 2019, and spread to 185 countries and territories, and as of April 29, 2020, more than 3.11 million cases were recorded, and more than 217,000 people were killed. Despite all worldwide efforts, there is currently no vaccine, any drugs available to protect people against deadly SARS-CoV-2 coronavirus. The world urgently needs a SARS-CoV-2 coronavirus vaccine or effective antiviral drugs to relieve the human suffering associated with the pandemic that kills thousands of people every day. The SARS-CoV-2 genome encode a non-structural proteins named as ORF1a/b, and structural proteins such as spike (S) glycoprotein, nucleocapsid protein (N), small envelop protein (E) and matrix protein (M). A number of studies have been shown that CoV spike (S) glycoprotein and nucleocapsid protein (N) could be promising targets for vaccine, antibodies and therapeutic drug development to combat with deadly, pandemic SARS-CoV-2. Purposes of the present paper is the sequence analysis and amino acid variations of structural proteins deduced from novel coronavirus SARS-CoV-2 strains, isolated in different countries. Multiple sequence alignment of S, N and E proteins from four different coronavirus species, are also described. It is expected that the data from these studies will be very useful for the the designing and development of vaccines, antibodies and therapeutic agents that can be used to combat with the highly pathogenic SARS-CoV-2 coronavirus worldwide.

Key words: coronavirus , SARS-CoV-2, Spike protein, Nucleocapsid protein, MSA.

 doi:10.20944/preprints202005.0026.v1

Introduction

SARS-CoV-2 is a novel and highly pathogenic coronavirus, which has caused an outbreak in Wuhan city, China in 2019, and then soon spread nationwide and spilled over to other countries and the world. Head of the United Nations has described this as humanity's worst crisis since World War II. Despite all worldwide efforts, there is currently no vaccine, any drugs available to protect people against deadly SARS-CoV-2 coronavirus. The world urgently needs a SARS-CoV-2 coronavirus vaccine or antiviral drugs to relieve the human suffering associated with the pandemic that kills thousands people every day. SARS -Cov-2 is a considered as betacoronavirus , like MERS-CoV or SARS-CoV, with the single stranded RNA genomes. Phylogenetic analysis on the coronavirus genomes has revealed that SARS-CoV-2 is a new member of the betacoronavirus genus, which includes SARS-CoV, MERS-CoV, bat SARS-related coronaviruses (SARSr-CoV), as well as other coronaviruses identified in humans and animal species (Zhou, P. et al. 2020; Wu, F. et al.2020; Lu et al., 2019). The two-thirds of SARS -Cov-2 RNA genome (~30 kb) encodes a non-structural proteins, named as ORF1a/b (ppa1 and pp1ab). The rest part of virus genome encode mainly structural proteins such as spike (S) glycoprotein, nucleocapsid protein (N), small envelop protein (E) and matrix protein (M). A number of studies have been shown a CoV spike (S) glycoprotein as a leading target for vaccines, antibodies, and therapeutic drug development against deadly, pandemic SARS-CoV-2. It was shown that SARS-CoV-2 share about 80% sequence identity in the spike (S) gene with SARS-CoV and other SARSr-CoVs (Zhou et al., 2020). However, bat coronavirus RaTG13 appears to be the closest relative of the SARS-CoV-2 sharing over 93.1% sequence identity. The crystal structure of the SARS-CoV-2 spike receptor-binding domain (RBD) bound to the cell receptor ACE2 at 2.45 Å resolution was quite recently determined by Zhou et al. (2020). It was demonstrated that the overall ACE2-binding mode of the SARS-CoV-2 RBD is nearly identical to that of the SARS-CoV RBD, which also utilizes ACE2 as the cell receptor (Wong et al, 2004). It should be noted that since the RBD is the critical region for receptor binding, therefore RBD could be great promise for developing highly potent cross-reactive therapeutic agents towards diverse coronavirus species including SARS-CoV-2. The spike protein of SARS-CoV-2 coronavirus is cycteine-rich protein and total of nine cysteine residues are found in the RBD, eight of which forming four pairs of disulfide bonds (Lan et al., 2020). In addition, the spike protein has 22 potential N-glycosylation sites, of which two of them are in the receptor binding domain (RBD) region.

The SARS-CoV-2 nucleocapsid (N) protein is multi functional RNA binding protein, which is responsible for viral RNA transcription and replication. Nucleocapsid protein consist three domains, i) RNA binding domain, ii) C-terminal dimerization domain (CTD) and iii) Ser/Arg (SR)-rich linker. Previous studies have shown that NTD is responsible for RNA binding, dimerization domain for oligomerization and SR for phosphorylation (Lo et al., 2013; Chen et al., 2013; Chang et al., 2013; Chang et al., 2006; Wootton et al., 2002). A number of studies

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 May 2020 doi:10.20944/preprints202005.0026.v1

have been shown that N protein is highly produced during infection, and induced protective immune response against SARC-CoV as well as SARC-CoV-2 (Ahmed et al., 2020; Liu et al., 2006; Shang et al., 2005; Lin et al., 2003).

The purpose of this paper is to generate data that would be useful for the design and development of vaccines, antibodies, and therapeutic drugs to combat with deadly pandemic SARS-Cov2, by the sequence analysis and analysis of amino acid variations of novel coronavirus SARS-CoV-2 strains, isolated worldwide.

Results

1. Sequence analysis of structural proteins deduced from novel coronavirus SARS-CoV-2 strains, isolated in different countries.

All available sequences of 2019 Novel Coronavirus (SARS-CoV-2) strains, isolated in different countries were downloaded from NCBI as of 24.04.2020. We aligned 1330 spike protein sequences of SARS-CoV-2 strains, isolated in different countries and found to be nearly 100% identical. The only difference was in 614th position of the consensus sequence (Figure 1); variation in this position of residues G and D. It may due to a SNP (e.g. G***A***T to G***G***T, Asp to Gly or viceversa) (Figure 1). Similarly, when 1334 nucleocapsid proteins sequences were aligned, there was almost 100% percent identity; the only difference was in a pair of aminoacids at 204-205 position, RG to KR (Figure 2). We also aligned membrane and envelope proteins of SARS-CoV-2 coronovirus strains, isolated in different countries. When 1326 sequences of envelope protein was aligned the following differences were found: there is one (out of 1326 sequences A to V conversion) mutation at 36th position, and two mutations (L to H and L to R conversion) at 37th position, and one mutation (S to F) at 55th position. When 1316 sequences of membrane proteins, was aligned, there are differences in the following: two sequences have mutation at 2nd and 3rd position, one sequence at at 57th position, L to V, and one sequence have V to F mutation at 70th position, one sequence A to I mutation at 73th position, two sequences A to S at 85th, at 89th one sequence G to R , at 133th one sequence L to M, at 142th one sequence A to P, at 175th four sequences T to M, at 190th one sequence D to N, at 195 and 196th one sequence AY to VH. Thus, there are more amino acid variations in membrane proteins compared to S, N, and E. In general, these results demonstrate that there are no significant amino acid variations in the S, P or E proteins of the coronavirus strains (SARS-CoV-2) isolated in different countries.

2. Comparison of S, N and E proteins of closely and distantly related coronaviruses.

Based on a previous phylogenetic analysis of the coronavirus genomes, it was demonstrated that, like SARS-CoV, MERS-CoV and bat SARS-related coronaviruses, SARS-CoV-2 is a new member of the beta-coronavirus genus (Zhou et al. 2020; Wu et al. 2020; Lu et al., 2019). SARS-CoV-2 share about 80% sequence identity in the spike (S) gene with SARS-CoV and other SARSr-CoVs (Zhou et al., 2020). However, bat coronavirus RaTG13 exhibited a high

 doi:10.20944/preprints202005.0026.v1

sequence identity to SARS-CoV-2, sharing over 93.1%. We performed a multiple sequence alignment (MSA) of SARS-CoV (GenBank: NC_004718.3), SARS-CoV2 (GenBank: NC_045512.2), Bat SARS-like Coronavirus WIV1(GenBank: KF367457.1) and Pangolin Coronavirus (GenBank: MT072864.1). Sequences have 80% average percent identity over the alignment and 68.7% identical sites. As can be seen from Figure 2 (under the alignment), amino acids after the first furin cleavage site, as previously reported, are not well conserved, however, the amino acids on the right second cleavage site are very well conserved. Another finding is that the rich N-glycosylation sites (Figure 3, top left) and the cysteine-rich region are well conserved (Figure 3, top right) in S protein from different coronavirus species.

Figure 4 demonstrate multi sequence alignment of nucleocapsid proteins of distantly and closely related coronaviruses such as SARS-CoV (GenBank: NC_004718.3), SARS-CoV2 (GenBank: NC_045512.2), Bat SARS-like Coronavirus WIV1(GenBank: KF367457.1), PangolinCoronavirus (GenBank: MT072864.1, BovineCoronavirus (GenBank: NC_003045.1) and Porcine Coronavirus (GenBank: NC_039208.1). As seen from the top alignment, N-protein is strongly conserved amongst closely related CoV's. Distantly related CoV's (Bovine CoV and Porcine CoV) are employed to highlight broadly conserved regions (Figure 4). By comparing two alignments four regions of high similarity emerges (Figure 4. blue squares). Sequence at top of the figure, highlights highly-conserved Serine/arginine (SR)- rich region. N-flank of this region can be used in the design and development of a therapeutic vaccine that can activate immune response against closely related coronaviruses, SARS-Cov2, SARS-Cov, Bat coronovirus, Bovine coronovirus, Pangolin coronovirus and Porcine coronovirus.

Figure 5 demonstrate multi sequence alignment of envelope proteins of distantly and closely related coronaviruses: SAR SARS-CoV (GenBank:NC_004718.3), SARS-CoV2 (GenBank:NC_045512.2), Bat SARS-like Coronavirus WIV1(GenBank: KF367457.1), Pangolin Coronavirus (GenBank:MT072864.1, Bovine Coronavirus (GenBank: NC_003045.1) and Porcine Coronavirus (GenBank: NC_039208.1). As seen from the top alignment, E-protein is strongly conserved among closely related CoV's. An interesting, but highly conserved C-XX-C region emerges from distantly aligned CoV's. The CxxC motif are employed by many redox proteins for reduction, formation and isomerization of disulfide bonds (Fomenko, Gladyshev, 2003). Thus, distantly related CoV's (Bovine CoV and Porcine CoV) are employed to highlight broadly conserved regions. Post-translational modifications such as N-linked glycosylation and phosphorylation sites detected by ProSite, which can be seen above the alignments, are well conserved in SARS-Cov2, SARS-Cov, Bat coronovirus, Bovine coronovirus, Pangolin coronovirus and Porcine coronovirus.

Discussion

 doi:10.20944/preprints202005.0026.v1

The novel conronovirus, currently designated as SARS-CoV-2, is an emerging virus, we know relatively little about it. It has observed that SARS-CoV-2 may be transmitted from infected people without symptoms, therefore, it increases the challenges of controlling a deadly pandemic without the use of a vaccine. The goal of the present paper is the sequence analysis and analysis of amino acid variations of structural proteins, deduced from novel coronavirus SARS-CoV-2 strains, isolated in different countries, which would generate bases for the designing and development of vaccines, antibodies and therapeutic agents to combat with the highly pathogenic SARS-CoV-2 coronavirus worldwide. S protein of SARS-CoV-2, a type I transmembrane glycoprotein that plays an important role in virus binding and entry and also is a major inducer of neutralizing antibodies. S protein consists of a signal peptide, and two domains, extracellular domain and transmembrane domain. Its extracellular domain consists of two S1 subunits and, and S2, the carboxy-terminal membrane fusion subunit (Wrapp et al., 2020). The furin-like cleavage site has been recently predicted in SARC-Cov-2, which lack in the other SARS-like CoVs (Coutard et al. 2020). A number of studies have been shown S and N proteins are most promising targets for vaccine and anybody development against coronovirus, including SARS-CoV-2 (Chen, et al., 2020). Our amino acid sequences analysis of structural proteins, demonstrates that, despite the higher number of amino acid variations in membrane protein (M), however, there are no significant amino acid variations observed in the structural proteins of S, P or E-proteins obtained from new strains of SARS-CoV-2 coronavirus isolated in different countries . These data are believed can provide a basis for the development of vaccines and antibodies to combat the deadly SARS-CoV-2 outbreak, isolated in different countries.

SARS-CoV-2 share about 80% sequence identity in the S gene with SARS-CoV and other SARSr-CoVs (Zhou et al., 2020). However, bat coronavirus RaTG13 exhibited a high sequence identity to SARS-CoV-2, sharing over 93.1%. Our sequence analysis showed that N protein of SARS-COV2 is more similar to pangolin (Figure 6). It should be noted that this protein is 100% identical in SARS-CoV and bat coronavirus. Given the similarity of SARS-CoV-2 to bat SARS-CoV-like coronaviruses2 in S protein and pangolin in N protein, this suggest that bats and pangolin may serve as reservoir hosts for progenitor of SARS-CoV-2**.**

By amino acid sequences analysis we also identified sequences, which are conserved in many coronaviruses, including new coronovirus SARS-CoV-2. Spike protein of NNTVYDPLQPELDSFKEELDKYFKNHTSP (Figure 3) and Nucleocapsid protein of PKGFYAEGSRGGSQASSRSSSRSR (Figure 4) was found to be particularly well conserved in many coronaviruses. Such sequences could be important for developing vaccines, antibodies, and also would be important for diagnostic purposes.

Notable, the spike protein of SARS-CoV-2 coronavirus is cycteine-rich protein and total of nine cysteine residues are found in the RBD, eight of which forming four pairs of disulfide bonds (Lan et al., 2020). Correct formation of disulfide bridges is essential for proper folding of cycteine-rich proteins (Mamedov et al., 2019). In addition, the spike protein has 22 potential N-glycosylation sites, of which two of them are in the receptor binding domain (RBD) region.

 doi:10.20944/preprints202005.0026.v1

Thus, the correct formation of disulfide bridges and, accordingly, the correct status of N-glycosylation will be crucial for the correct folding of the S-protein when this protein is recombinantly produced in a heterologous system.

Materials and Methods

First, all sequences are downloaded from NCBI Virus as of 24.04.2020, then these bulk files are pre-processed with *in-house built* python (also including numpy and pandas libraries) scripts. Eventually nearly 1300 sequence obtained for each given protein (spike, nucleocapsid, membrane and envelope). Then these sequences (separately) piped into local server (available at Akdeniz University, Biotechnolgy Department) to perform MSA with ClustalO (defaults used) and a MSA acquired for each protein. These MSA then fed to Geneious Prime to visualize and statistically calculate each and every mutation.

For distant and related pairwise aligments, SARS-CoV (GenBank:NC_004718.3), SARS-CoV2 (GenBank: NC_045512.2), Bat SARS-like Coronavirus WIV1(GenBank: KF367457.1) and Pangolin Coronavirus (GenBank: MT072864.1) are used. MSA generated by using in-built Geneious Prime MSA builder. Results analyzed with Geneious Prime.

References

Ahmed SF, Quadeer AA & McKay MR. Preliminary Identification of Potential Vaccine Targets for the COVID-19 Coronavirus (SARS-CoV-2) Based on SARS-CoV Immunological Studies. *Viruses* 2020; 12.

Chang CK, Chen CMM, Chiang MH, Hsu YL & Huang TH. Transient Oligomerization of the SARS-CoV N Protein - Implication for Virus Ribonucleoprotein Packaging. *PLoS ONE* 2013; **8**.

Chang CK, Sue SC, Yu TH, Hsieh CM, Tsai CK, Chiang YC et al. Modular organization of SARS coronavirus nucleocapsid protein. *Journal of Biomedical Science* 2006; 13: 59-72.

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.

Chen WH, Hotez PJ, Bottazzi ME.Potential for developing a SARS-CoV receptor-binding domain (RBD) recombinant protein as a heterologous human vaccine against coronavirus infectious disease (COVID)-19.

Chen IJ, Yuann JMP, Chang YM, Lin SY, Zhao J, Perlman S et al. Crystal structure-based exploration of the important role of Arg106 in the RNA-binding domain of human coronavirus

OC43 nucleocapsid protein. *Biochimica et Biophysica Acta - Proteins and Proteomics* 2013; 1834: 1054-1062.

Lu, R. et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* 395, 565–574 (2020).

Liu SJ, Leng CH, Lien SP, Chi HY, Huang CY, Lin CL et al. Immunological characterizations of the nucleocapsid protein based SARS vaccine candidates. *Vaccine* 2006; 24: 3100-3108.

Lin Y, Shen X, Yang RF, Li YX, Ji YY, He YY et al. Identification of an epitope of SARS coronavirus nucleocapsid protein. *Cell Research* 2003; **13**: 141-145.

Lo YS, Lin SY, Wang SM, Wang CT, Chiu YL, Huang TH et al. Oligomerization of the carboxyl terminal domain of the human coronavirus 229E nucleocapsid protein. *FEBS Letter* 2013; **587**: 120-127.

Mamedov et al., A Plant-Produced in vivo deglycosylated full-length Pfs48/45 as a Transmission-Blocking Vaccine Candidate against malaria. Sci Rep. 2019 Jul 8;9(1): 9868.

Shang B, Wang XY, Yuan JW, Vabret A, Wu XD, Yang RF et al. Characterization and application of monoclonal antibodies against N protein of SARS-coronavirus. *Biochem Biophys Res Commun* 2005; **336**: 110-117.

Zhou, P. et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 579, 270–273 (2020). https://doi.org/10.1038/s41586-020-2012-7.

Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D.Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein.Cell. 2020 Apr 16;181(2):281-292.

Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh CL, Abiona O, Graham BS, McLellan JS.Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation.Science. 2020 Mar 13;367(6483):1260-1263.

Wootton SK, Rowland RRR & Yoo D. Phosphorylation of the porcine reproductive and respiratory syndrome virus nucleocapsid protein. *Journal of Virology* 2002; **76**: 10569-10576.
Wu, F. et al. A new coronavirus associated with human respiratory disease in China. *Nature* 579, 265–269 (2020). https://doi.org/10.1038/s41586-020-2008-3.

Legends to Figures

Figure 1. A subset of MSA of 1330 different SARS-CoV-2 S-protein isolates. Sequences are obtained from NCBI Virus website (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/) after clean-up with *in-house built* scripts, MSA is done by ClustalO. Features are detected by ProSite.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 May 2020 doi:10.20944/preprints202005.0026.v1

Figure. 2. A subset of MSA of 1334 different SARS-CoV-2 N-protein isolates. Sequences are obtained from NCBI Virus website (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/) after clean-up with *in-house built* scripts, MSA is done by ClustalO. Features are detected by ProSite.

Figure. 3. MSA of 4 different Coronavirus species; Bat coronavirus, Pangolin coronavirus, SARS-CoV and SARS-CoV2. Cleavage sites are anotated by blue squares. Rich in N-glycosylation site sequences, top left. Cystein rich-region of the S-protein, top right. Sequences are obtained from NCBI: SARS-CoV (GenBank: NC_004718.3), SARS-CoV2 (GenBank: NC_045512.2), Bat SARS-like Coronavirus WIV1(GenBank:KF367457.1) and PangolinCoronavirus (GenBank:MT072864.1) and MSA done by Geneious Prime. Features are detected by ProSite.

Figure. 4 MSA of N-protein of distantly and closely related coronaviruses (SARS-Cov2, SARS-Cov, Bat coronovirus, Bovine coronovirus, Pangolin coronovirus, Porcine coronovirus).

Sequences are obtained distantly (Bovine CoV and Porcine CoV) and closely related coronaviruses (SARS-Cov2, SARS-Cov, Bat coronovirus and Pangolin coronovirus) are employed to highlight broadly conserved regions. By comparing two alignments 4 regions of high similarity emerges (blue squares). SR rich region, on the top. Sequences are obtained from NCBI: SARS-CoV (GenBank:NC_004718.3), SARS-CoV2 (GenBank:NC_045512.2), Bat SARS-like Coronavirus WIV1(GenBank: KF367457.1) and Pangolin Coronavirus (GenBank:MT072864.1) and MSA is done by Geneious Prime. Features are detected by ProSite.

Figure. 5 MSA of E-protein of distantly and closely related coronaviruses.
Distantly related CoV's (Bovine CoV and Porcine CoV) are employed to highlight broadly conserved regions. N-linked-glycosylation, Phosphorylation sites and C-X-X-C motif indicated by arrows . Sequences are obtainedf rom NCBI: SARS-CoV (GenBank:NC_004718.3), SARS-CoV2 (GenBank:NC_045512.2), Bat SARS-like Coronavirus WIV1(GenBank: KF367457.1) andPangolinCoronavirus (GenBank:MT072864.1) and MSA is done by Geneious Prime. Features are detected by ProSite.

Figure 6. Phylogenetic relationships of N protein of closely and distantly related Coronaviruses. Phylogenetic tree shows that SARS-CoV-2 is more similar to pangolin coronavirus suggestion pangolin coronavirus can be a reservoir for SARS-CoV-2 progenitor. Sequences are obtained from NCBI: SARS-CoV (GenBank:NC_004718.3), SARS-CoV2 (GenBank:NC_045512.2**),** Bat SARS-like Coronavirus WIV1(GenBank: KF367457.1), Bovine Coronavirus (GenBank: NC_003045.1) and Pangolin Coronavirus (GenBank:MT072864.1) and MSA is done by Geneious Prime. Phylogenetic Tree is constructed by UPGMA algorithm of Geneious Prime with default parameters.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 May 2020 doi:10.20944/preprints202005.0026.v1

Figures

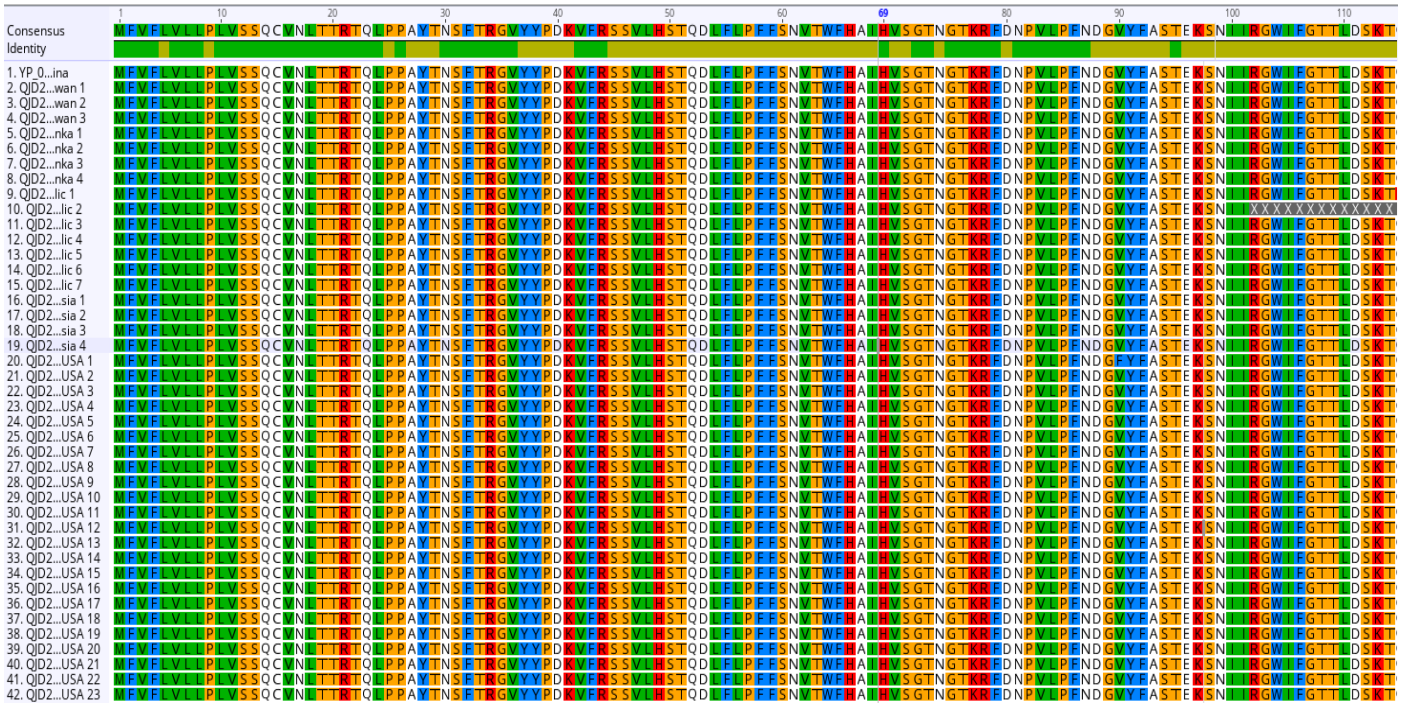

Figure 1.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 May 2020 doi:10.20944/preprints202005.0026.v1

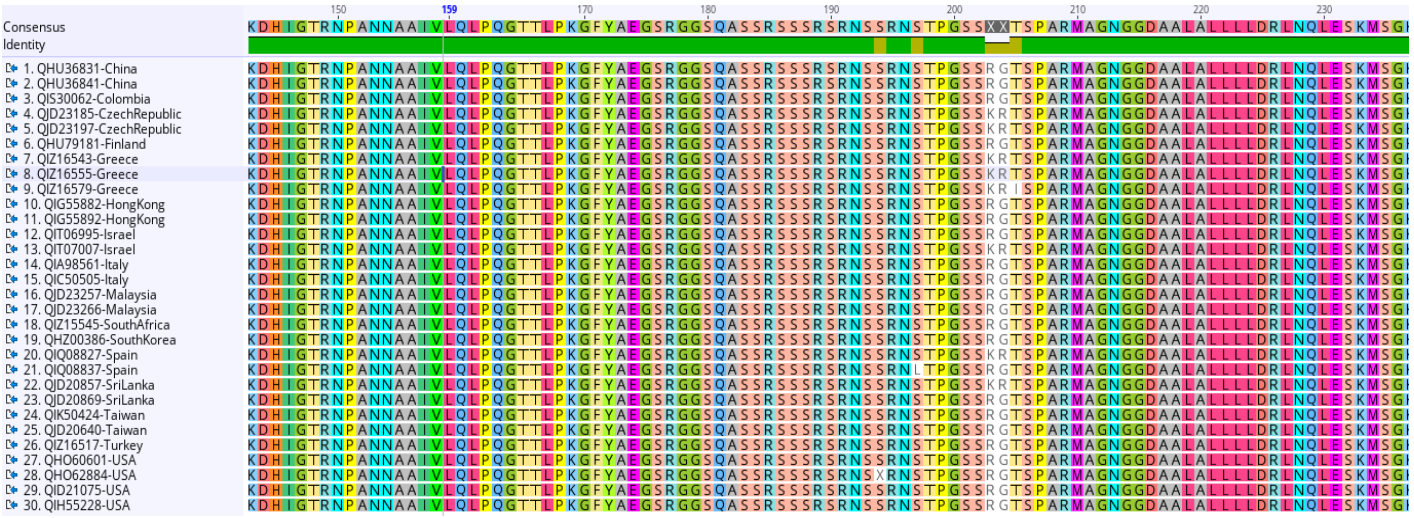

Figure 2.

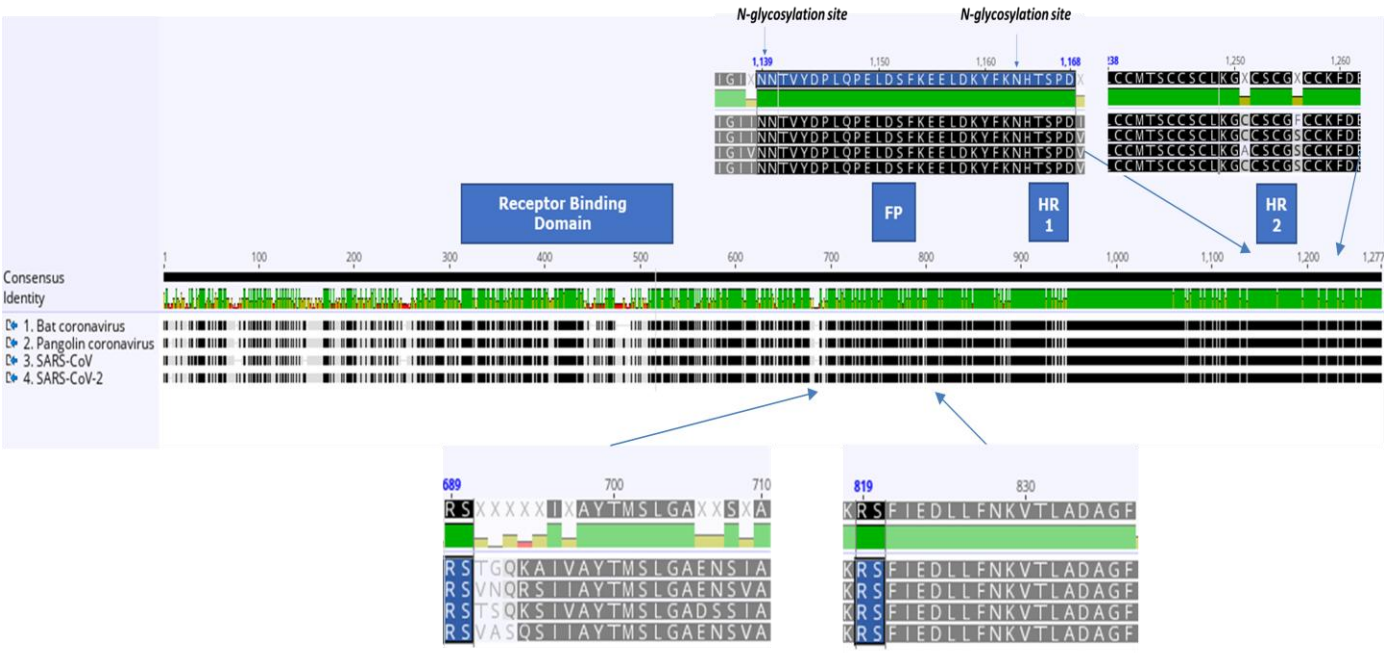

Figure 3.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 May 2020 doi:10.20944/preprints202005.0026.v1

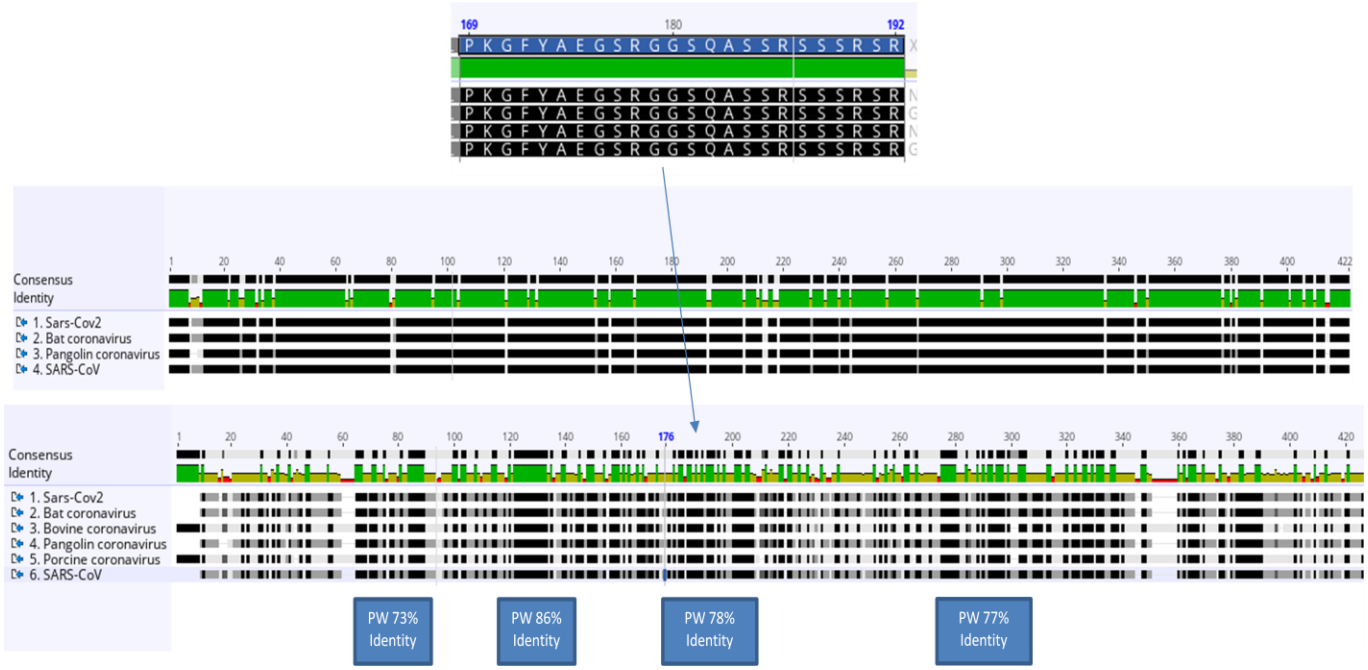

Figure 4.

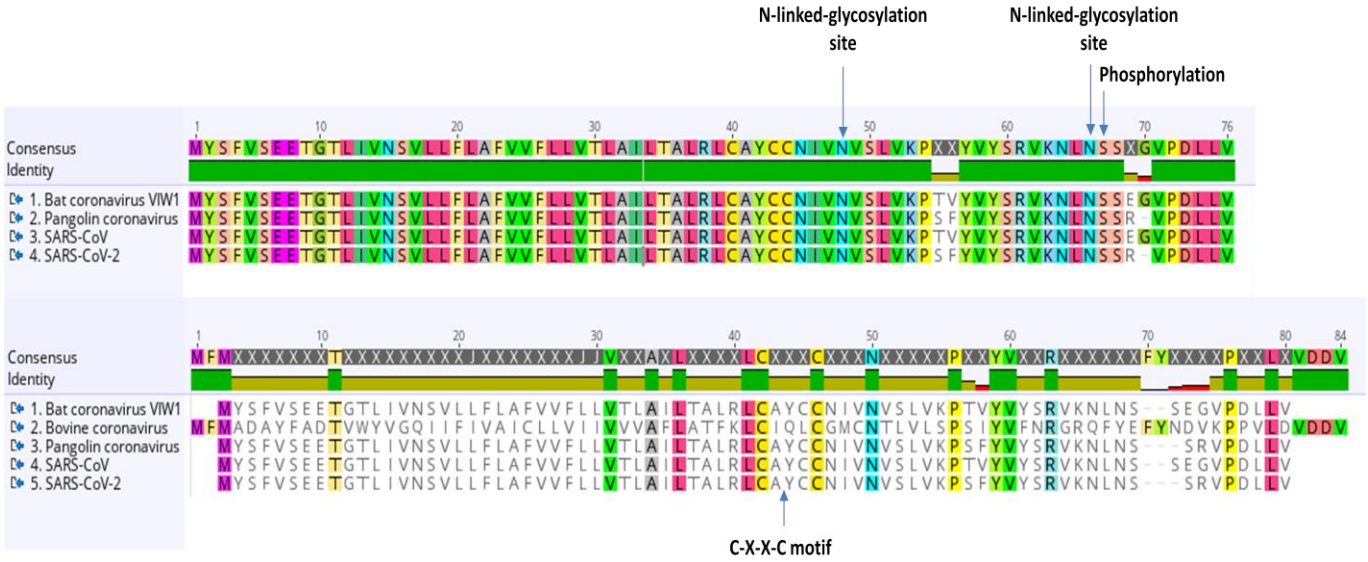

Figure 5.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 May 2020 doi:10.20944/preprints202005.0026.v1

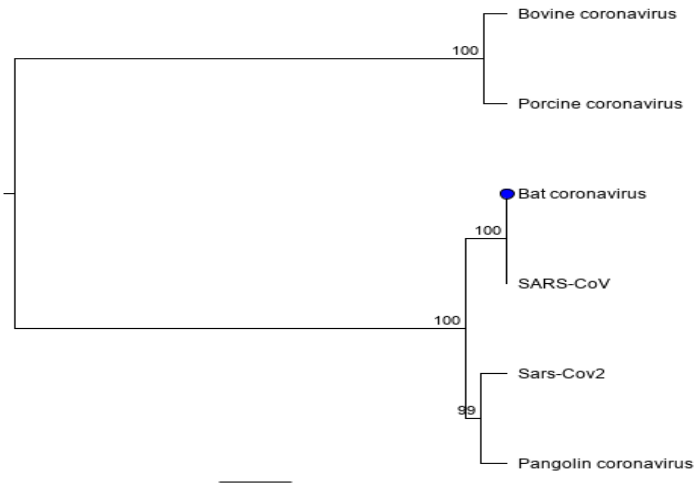

Figure 6.