2. MATERIALS & METHODS

2.1 Prediction of orthologs

Proteome sets of SARS-CoV-2 and related genomes were retrieved from NCBI genome (https://www.ncbi.nlm.nih.gov/genome). The proteome sets were used to predict orthologous proteins of SARS-CoV-2 using reciprocal best BLAST hit (RBBH) [5]. The orthologous proteins were filtered with E-value and sequence coverage of 10e-2 and 30 percent, 10e-3 and 40 percent and 10e-5 and 50 percent respectively. The results were analysed to infer the proteins with significant evolutionary changes.

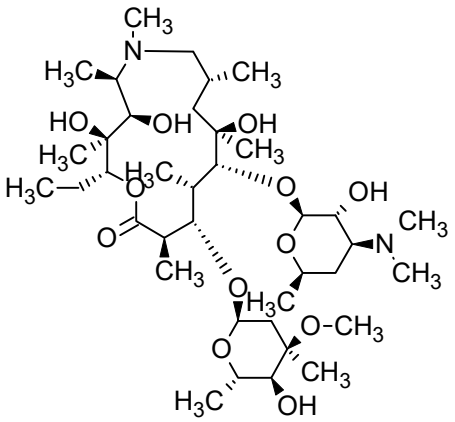
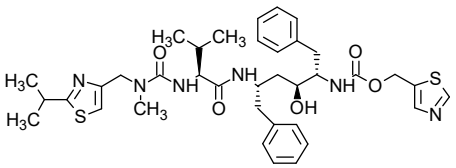
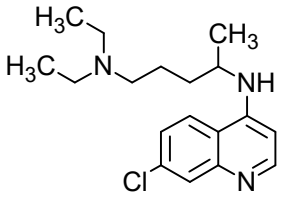
2.2 Prediction of regions of evolutionary differences

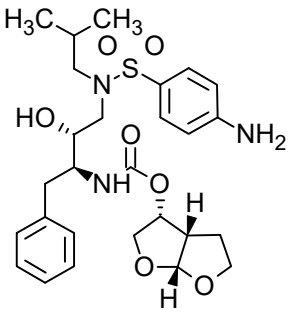
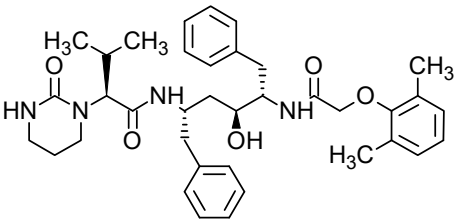
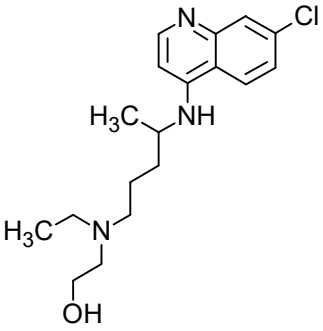
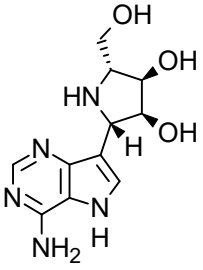
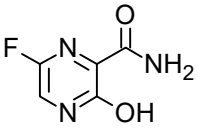
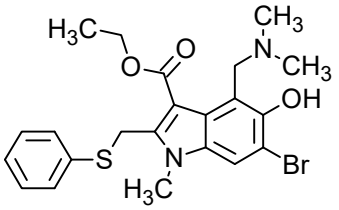
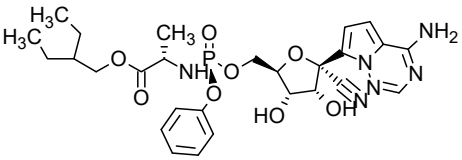
A set of spike glycoprotein including proteins from its closely related genomes was used to implement multiple sequence alignment (MSA) using Clustal Omega [6]. The resulting alignment was used to identify the regions of evolutionary differences such as homologous substitutions, insertions and deletions in the spike glycoprotein [7]. Then, the 3D protein structure of the spike glycoprotein (6LZG) [8] was extracted from RCSB Protein Data Bank. Earlier protein set along with the amino acid sequence of the 3D protein structure was used to implement again the MSA. The resulting new alignment was used to identify the homologous regions of the sites of evolutionary differences in the protein structure.

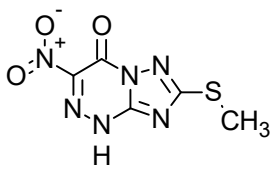
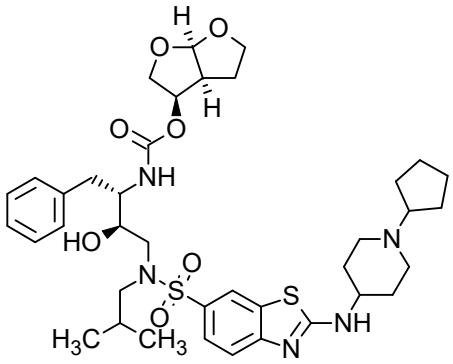
2.3 Docking analysis

The list of potential candidates for Covid-19 was obtained from DrugBank [9] white paper updated on March 26, 2020. The chemical names, structures and composition of these candidates is shown in the following **Table 1**.

Table 1. Chemical structures of potential candidates for Covid-19.

Sr. No.	DrugBank ID	Name	Composition	Structure
1	DB00207	Azithromycin	C38H72N2O12	
2	DB00503	Ritonavir	C37H48N6O5S2	
3	DB00608	Chloroquine	C18H26ClN3	

4	DB01264	Darunavir	C27H37N3O7S	
5	DB01601	Lopinavir	C37H48N4O5	
6	DB01611	Hydroxychloroquine	C18H26ClN3O	
7	DB11676	Galidesivir	C11H15N5O3	
8	DB12466	Favipiravir	C5H4FN3O2	
9	DB13609	Umifenovir	C22H25BrN2O3S	
10	DB14761	Remdesivir	C27H35N6O8P	

11	DB15622	Triazavirin	C5H4N6O3S	
12	DB15623	TMC-310911	C38H53N5O7S2	

The potential candidates in PDB format were downloaded from DrugBank database. Docking analysis was performed through AutoDock Vina [10] using spike glycoprotein as the protein molecule and the potential candidates as the ligand molecules. In the AutoDock software, the protein molecule was prepared by removing water, ions, ligands and non-amino acid atoms from the 3D structure. Polar hydrogens and Kollman charges were added to the protein structure and then saved as PDBQT file. The ligand molecules were prepared by adding Gasteiger charges to the 3D structures. The number of rotable bonds in the ligand structures was identified and then saved as PDBQT files. Size and position of the GridBox covering the regions of evolutionary differences was identified to predict the interaction and binding energy between the ligand molecules and the amino acids in the mutations sites. Similarly, size and position of the GridBox covering the whole protein was identified separately to predict the interaction and binding energy between the ligand molecules and the binding sites of the entire protein molecule. The grid coordinates, PDBQT file of the protein and PDBQT file of each of the ligands was used to implement dock analysis using AutoDock Vina. Docking analysis is implemented separately in the homologous region of the evolutionary differences and in the entire protein.

3. RESULTS & DISCUSSION

3.1 Orthologs of SARS-CoV-2 proteins

The number of SARS-CoV-2 and related genomes available at NCBI genome were 44. Implementation of RBBH produced 6,989 orthologs of total 12 proteins in SARS-CoV-2. Filtering of the orthologs with different E-values and coverage as mentioned in the methods section, different sets of orthologous proteins were obtained. However, after filtering, only 10 proteins of SARS-CoV-2 had orthologous proteins while the other two proteins viz., ORF7b and ORF10 had no protein orthologs. The comparison of number of orthologs with different E-values and coverage can be visualized from the **Fig. 1**. The figure clearly shows that the significant evolutionary changes occurred in spike glycoprotein, membrane glycoprotein, orf1a polyprotein and nucleocapsid phosphoprotein. However, the number of orthologous proteins of YP_009724390.1, the surface spike glycoprotein, was drastically reduced when both the E-value cut off reduced from 10e-2 to 10e-3 and the coverage increased from 30 percent to 40 percent. This implies that significant evolutionary changes were occurred in the spike glycoprotein.

3.2 Regions of evolutionary difference in spike glycoprotein

It was observed from the orthologs analysis that the amino acid sequence of SARS-CoV-2 spike glycoprotein is closely related to spike glycoprotein of Bat coronavirus BM48-31/BGR/2008 and E2 glycoprotein precursor of severe acute respiratory syndrome-related coronavirus (SARS-CoV). This suggests the probable origin of SARS-CoV-2 from these genomes. MSA of these proteins revealed several homologous substitutions, few insertions and very few deletion regions. However, MSA of these proteins along with the amino acid sequence of the 6LZG protein PDB structure produced only an insertion region in the spike glycoprotein sequence from positions 680 to 684 with the amino acid sequence stretch of SPRRA. The region of insertion in the MSA is represented in the **Fig. 2**. The homologous substitution amino acids at the insertion regions of the protein structure were found to be GLN96, LEU97, GLN98, ALA99 and LEU100.

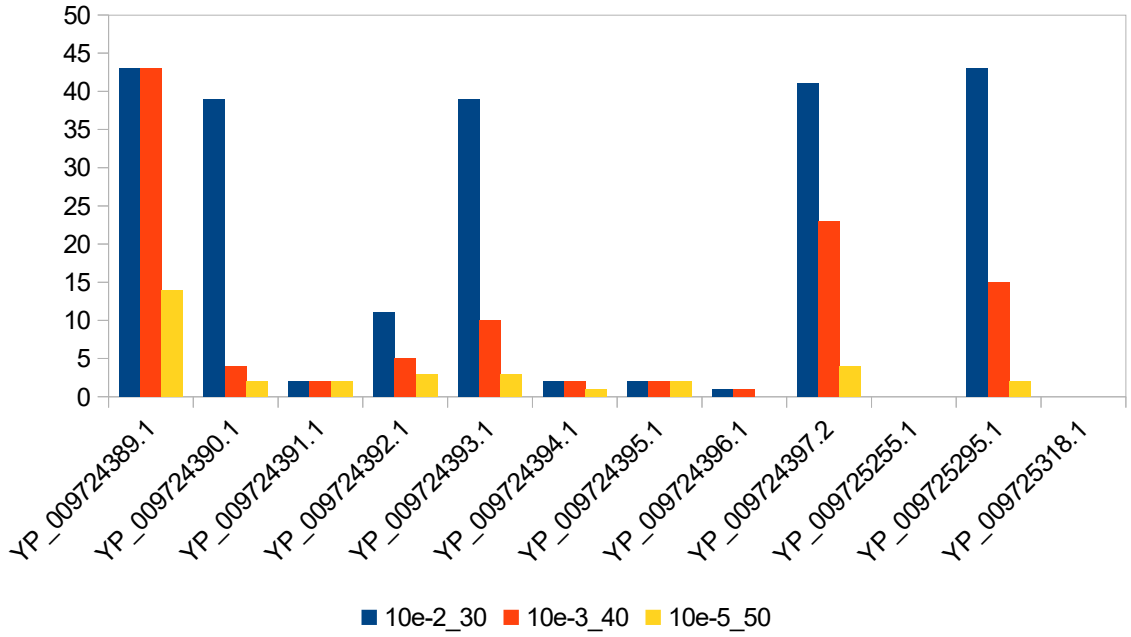


Figure 1. Bar graph of number of orthologous proteins. The legends 10e-2_30 indicate E-value cut off of 10e-2 and sequence coverage of 30 percent, 10e-3_40 indicate E-value cut off of 10e-3 and sequence of 40 percent and 10e-5_50 indicate E-value cut off of 10e-5 and sequence coverage of 50 percent. The number of orthologs of YP_009724390.1 (surface spike glycoprotein) from 10e-2_30 to 10e-30 were reduced from 39 to 4 respectively. Similar large reductions were observed in YP_009724393.1 (membrane glycoprotein) followed by YP_009725295.1 (orf1a polyprotein) and YP_009724397.2 (nucleocapsid phosphoprotein).

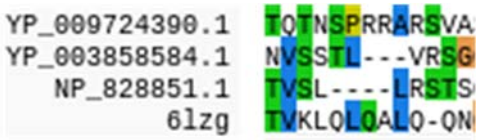


Figure 2. Region of insertion in the spike glycoprotein. YP_009724390.1 and YP_003858584.1 are the spike glycoproteins of SARS-CoV-2 and Bat coronavirus BM48-31/BGR/2008 respectively. NP_82885.1 is the E2 glycoprotein precursor of SARS-CoV.

3.3 Docking results

Docking analysis in the homologous region of insertion produced nine best conformations for each of the ligands except for Azithromycin and TMC_310911 which had only two and four best conformations respectively. Top conformation of each of the ligand was the best conformations with least binding energy and zero root-mean-square deviation (RMSD) value. Similarly, docking analysis in the whole protein region for each of the ligands produced nine best conformations and the top conformation was the best with least binding energy and zero RMSD value. The following **Table 2** shows the binding energies of the best conformation of each ligand.

Table 2. Binding energy of ligands in the protein structures.

SNo.	Ligand_Name	Binding energy at homologous insertion region (kcal/mol)	Binding energy for entire protein (kcal/mol)
1	Azithromycin	-6.9	-9.7
2	Chloroquine	-6.6	-6.7
3	Darunavir	-9.1	-9.5
4	Favipiravir	-4.9	-4.8
5	Galidesivir	-7.0	-6.6
6	Hydroxychloroquine	-6.7	-5.9

7	Lopinavir	-8.9	-9.6
8	Remdesivir	-9.5	-8.8
9	Ritonavir	-9.3	-9.9
10	TMC_310911	-9.2	-11.0
11	Triazavirin	-5.8	-5.9
12	Umifenovir	-8.0	-6.9

The table depicts that the Remdesivir has the best binding conformation with the least binding energy in the homologous region of insertion while TMC_310911 has the best binding conformation with the least binding energy when the entire protein was consider. The binding sites of Remdesivir in the spike glycoprotein were GLN98, ALA99, GLN102, ASN210 and LYS562 amino acids. Similarly, the binding sites of TMC_310911 in the spike glycoprotein were THR371, SER409 and ILE291 amino acids. It can also be observed from the table that the TMC_310911 too has a better binding conformation in the homologous region of insertion. Therefore, the present study uncovers both the Remdesivir and TMC_310911 as the highly potential candidates for the spike glycoprotein. Docking results of the Remdesivir and TMC_310911 can be visualized in the **Fig. 3** produced using PyMOL [11].

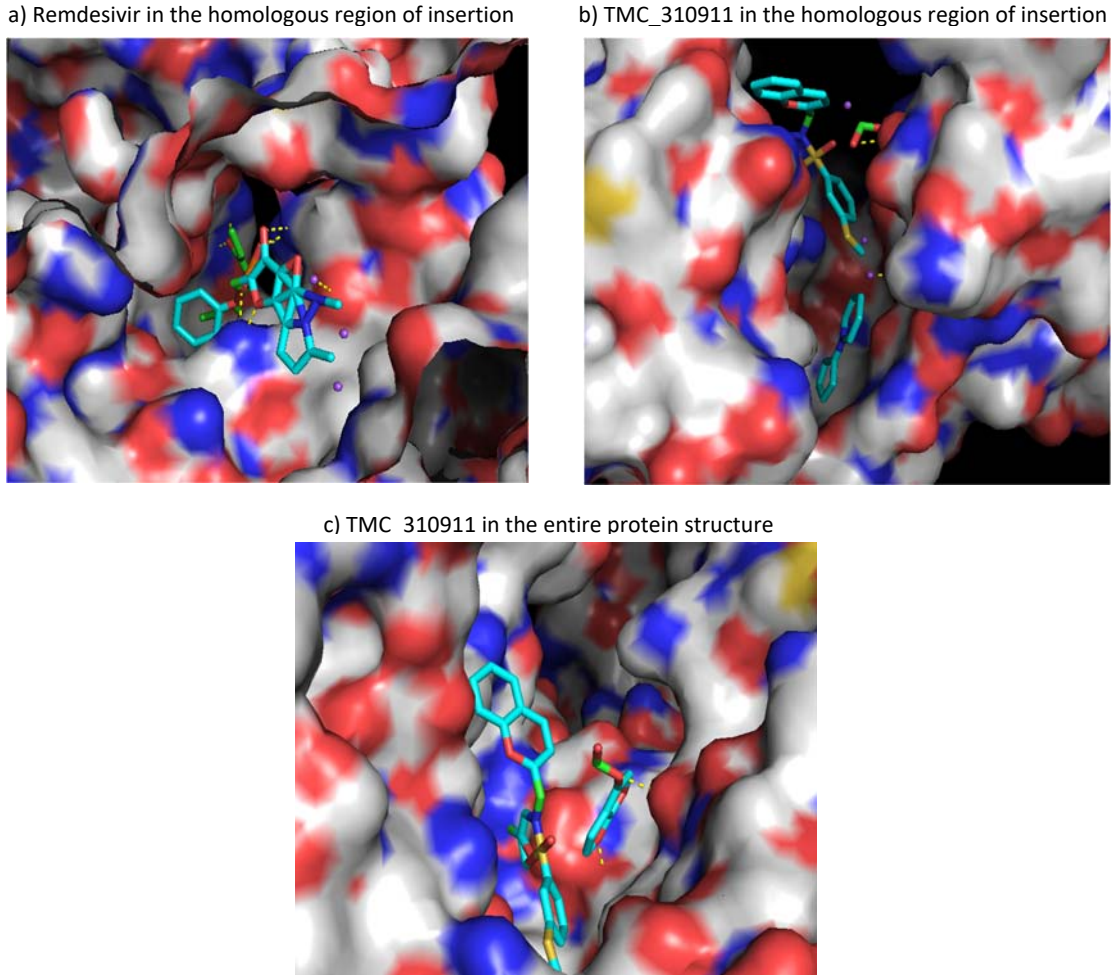


Figure 3. Docking of a) Remdesivir in the homologous region of insertion b) TMC_310911 in the homologous region of insertion and c) TMC_310911 in the entire protein structure. Binding interactions can be visualized as polar contacts.

4. CONCLUSION

The drastic evolutionary changes were observed in the spike glycoprotein indicating that it has some special function in the pathogen. The docking analysis revealed Remdesivir as the candidate with the best binding affinity in the homologous region of the insertion site while TMC_310911 as the candidate with the best binding affinity in the entire spike glycoprotein. Thus, the Remdesivir and the TMC_310911 are the highly potential candidates for SARS-CoV-2 spike glycoprotein. Therefore, the present study supports further investigation, experimentation and clinical trials of the Remdesivir and the TMC_310911 to fight against Covid-19 saving people and thus health and wealth of a country.

References

[1] D. S. Hui, E. I. Azhar, T. A. Madani, et al. The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health - The latest 2019 novel coronavirus outbreak in Wuhan, China. Int. J. Infect. Dis., 2020, 91: 264 - 266.
[2] I Astuti, Ysrafil. Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2): An overview of viral structure and host response. Diabetes Metab. Syndr. 2020, 14(4): 407-412.
[3] M. A. Tortorici, D. Veesler. Structural insights into coronavirus entry. Adv. Virus Res., 2019, 105: 93-116.
[4] A. C. Walls, Y. J. Park, M. A. Tortorici, et al. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell, 2020, 181(2): 281-292.e6.
[5] G. Moreno-Hagelsieb, K. Latimer. Choosing BLAST options for better detection of orthologs as reciprocal best hits. Bioinformatics, 2008, 24(3): 319-324.
[6] F. Sievers, A. Wilm, D. Dineen, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol., 2011, 7: 539.
[7] M. S. Rosenberg. Multiple sequence alignment accuracy and evolutionary distance estimation. BMC Bioinformatics, 2005, 6: 278.
[8] Q. Wang, Y. Zhang, L. Wu, et al. Structural and functional basis of SARS-CoV-2 entry by using Human ACE2. Cell, 2020, 181: 1-11.
[9] D. S. Wishart, Y. D. Feunang, A. C. Guo, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res., 2018, 46(D1): D1074-D1082.
[10] O Trott, A. J. Olson. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. J. Comput. Chem., 2010, 31(2): 455-461.
[11] The PyMOL Molecular Graphics System, Version 2.0 Schrodinger, LLC.