Sequence and Structure Analysis of Surface Glycoprotein of SARS-CoV-2 from India Variants

[1]Pramodkumar P. Gupta, [2]Virupaksha A. Bastikar, [3]Alpana Bastikar, and [4]Santosh S. Chhajed

[1]School of Biotechnology and Bioinformatics, D. Y. Patil Deemed to be University, Navi -Mumbai, Maharashtra, India.,
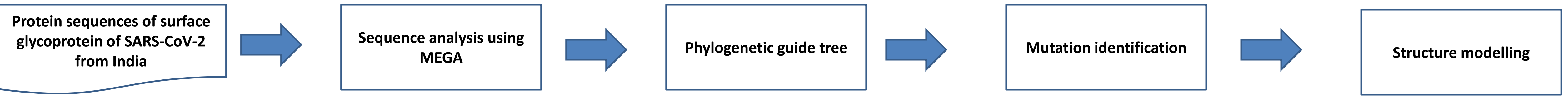[2]Amity Institute of Biotechnology (AIB), Amity University Maharashtra, Mumbai, Maharashtra India.,
[3]Department of Computer -Aided Drug Design, Navin Saxena Research and Technology Pvt. Ltd, Gandhidham, Gujarat, India.,
[4]Department of Pharmaceutical Chemistry, Bhujbal Knowledge City, MET's Institute of Pharmacy, Adgaon, Nashik, Maharashtra, India
Corresponding author: Pramodkumar P. Gupta, pramodkumar785@gmail.com

COVID19 is a pandemic across the world. It is caused by a novel coronavirus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The Surface/Spike Glycoprotein of SARS-CoV-2, which plays a key role in the receptor recognition and cell membrane fusion process, is composed of two subunits, S1 and S2. In the present work we have searched for Surface/Spike glycoprotein in the NCBI protein database and origin from "India", the search hit out 192 protein sequences as on 20 June 2020. Further, the sequences were aligned using Surface/Spike glycoprotein from Wuhan-China Origin and on the basis of the sequence length of 1273, the sequences were screened. Out of 192 input protein sequences, 177 sequences were complete in the length of 1273 amino-acids. Comparing all the sequences via sequence alignment mode in MEGA-X, exhibited a complex diversified outcome and reported 32 sequences. The protein sequence id QKI28685.1 was identified as a root and 31 protein sequences as a mutant/variant. QKI28685.1 was subjected to 3D protein structure modelling. As no full-length structural template was identified in the database. Automated homology modelling, Swiss-Model server and threading based I-Tasser were considered for the structure determination. Swiss-Model reported a partial structure from amino acid length 27 to 1146. A full-length structure is obtained from the I-Tasser server. The structures were analyzed using the ProSA and Ramachandran plot. 31 identified mutations were manually incorporated in the protein structure and a total of 31 mutants were created. Further, these proteins are in a process to study and understand the structural changes and their impact on the protein-protein interaction and protein-drug interaction.

Methodology

Protein sequences of surface glycoprotein of SARS-CoV-2 from India → Sequence analysis using MEGA → Phylogenetic guide tree → Mutation identification → Structure modelling
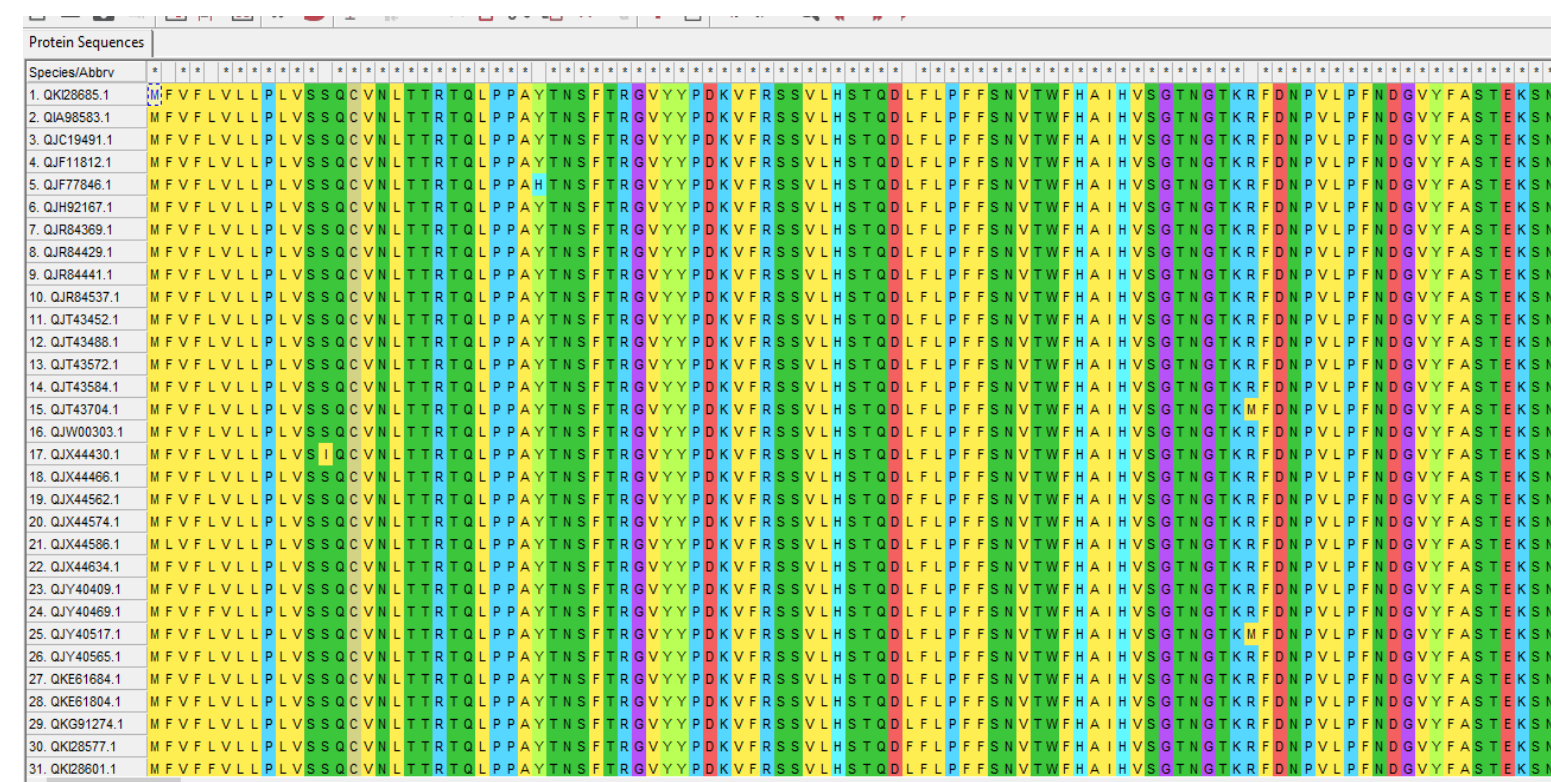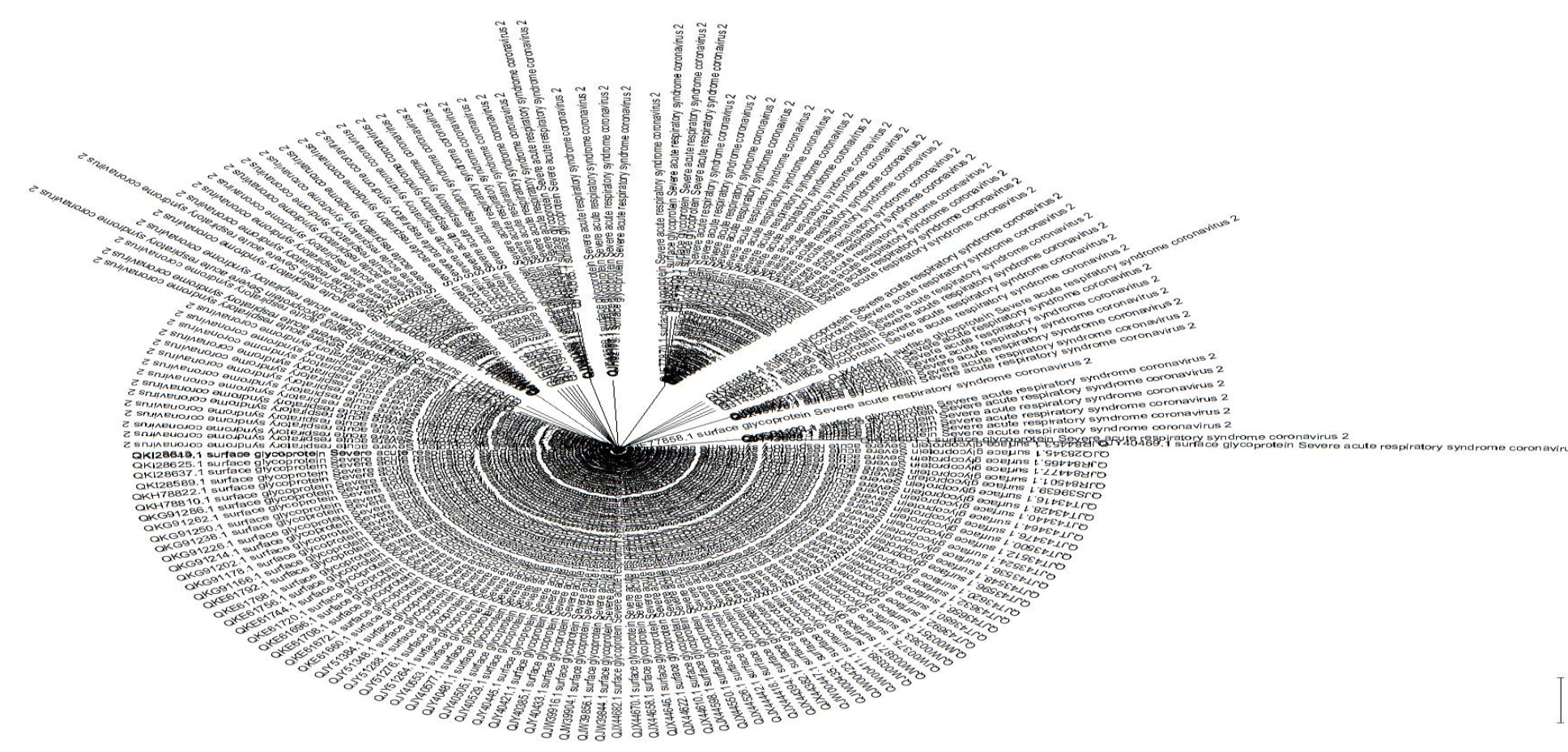
Results

Fig 1: Sequence alignment

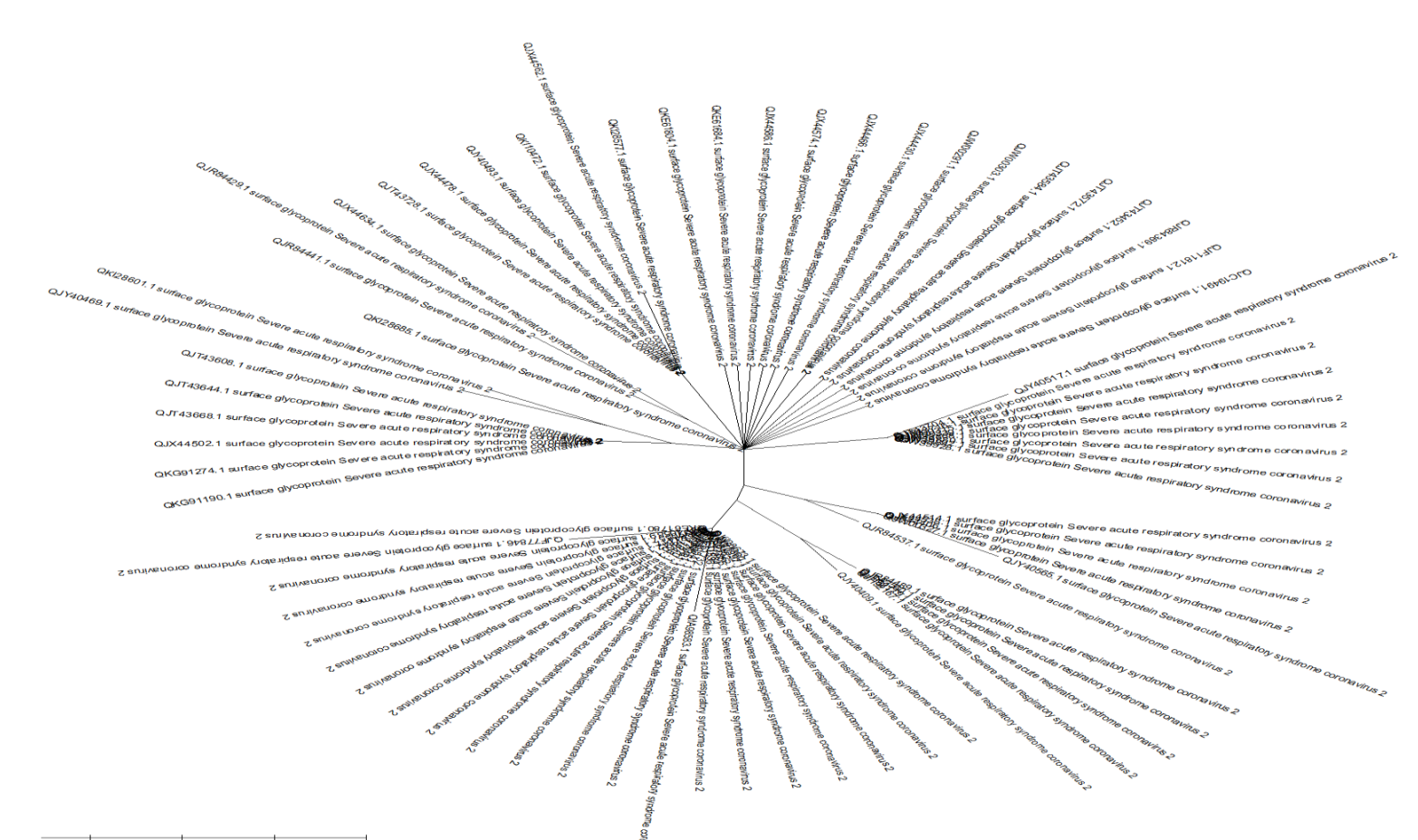Fig 2: Phylogeny of all the sequences

Fig 3: Phylogeny of all full length sequences

Table 1: Details about protein sequence, mutation and model

sr no	Protein Sequence Id	Mutation Position	Mutation Position	ProSA Score	Ramachandran Plot using Molprobity			Energy Minimization from Yasara	
					Favoured region	Allowed Region	Outlier	Initial Energy Kj/mol	Final Energy Kj/mol
1	QKI28685.1	614G		-9.91	71.52%	89.30%	10.70%	-325057.3	-503512
2	QJF77846.1	28(Y/H)	614(G/D)	-9.94	71.52%	89.20%	10.78%	-351536.1	-504211
3	QIA98583.1	614 (G/D)	930 (A/V)	-9.9	71.50%	89.20%	10.78%	-348526.6	-507212
4	QKI28673.1	614(G/D)		-9.93	71.50%	89.20%	10.78%	-351888	-504978
5	QJY40409.1	614(G/D)	677(Q/H)	-9.91	71.52%	89.20%	10.70%	1066717883	-504756
6	QJH92167.1	677(Q/H)		-9.9	71.52%	89.30%	10.70%	1066717927	-504136
7	QJR84537.1	583(E/D)	614(G/D)	-9.9	71.52%	89.20%	10.78%	-351435.5	-502811
8	QJY40565.1	583(E/D)	1083(H/Q)	-9.92	71.52%	89.30%	10.70%	-351579.1	-503256
9	QJT43488.1	583(E/D)		-9.92	71.52%	89.30%	10.70%	-351544.7	-501305
10	QJY40517.1	78(R/M)	1083(H/Q)	-9.99	71.52%	89.30%	10.70%	-350279.6	-504551
11	QJT43704.1	78(R/M)		-9.96	71.52%	89.30%	10.70%	-350268.1	-504040
12	QKG91274.1	572(T/I)		-9.9	71.52%	89.30%	10.70%	-351653.7	-504555
13	QKI28601.1	5(L/F)	572(T/I)	-9.91	71.52%	89.30%	10.70%	-351510.8	-499995
14	QJY40469.1	5(L/F)	162(S/I)	-9.89	71.44%	89.30%	10.70%	-308769.4	-505735
15	QKE61684.1	148(N/Y)		-9.89	71.52%	89.30%	10.70%	-351390.5	-499820
16	QJX44574.1	1104(V/L)		-9.88	71.52%	89.30%	10.70%	-350126.3	-503512
17	QJX44430.1	13(S/I)		-9.92	71.52%	89.30%	10.70%	-350518.7	-508495
18	QJT43584.1	827(T/I)		-9.9	71.52%	89.30%	10.70%	-351167.8	-503572
19	QJR84369.1	1181(K/R)		-9.92	71.52%	89.30%	10.70%	-352072.2	-504759
20	QKI28577.1	54(L/F)		-9.91	71.52%	89.30%	10.70%	-351475.1	-505910
21	QJX44562.1	54(L/F)	471(E/Q)	-9.89	71.52%	89.30%	10.70%	-350705	-501060
22	QJT43572.1	892(A/V)		-9.88	71.52%	89.30%	10.70%	-351914.4	-502281
23	QJF11812.1	1250(C/F)		-9.92	71.52%	89.30%	10.70%	2.9702E+12	-504405
24	QJC19491.1	271(Q/R)		-9.93	71.52%	89.30%`	10.70%	-352094.1	-501956
25	QJT43452.1	156(E/D)		-9.92	71.52%	89.30%	10.70%	-351994.8	505783
26	QJW00303.1	177(M/I)		-9.91	71.52%	89.30%	10.70%	-351189.1	-504217
27	QJX44466.1	879(A/S)		-9.92	71.52%	89.30%	10.70%	-351994.1	-503085
28	QJX44586.1	2(F/L)		-9.89	71.52%	89.30%	10.70%	-352043	-503833
29	QKE61804.1	255(S/F)		-9.9	71.52%	89.30%	10.70%	-350978.1	-505004
30	QJR84441.1	1243(C/F)		-9.89	71.52%	89.30%	10.70%	-346291.9	-501549
31	QJX44634.1	706(A/S)		-9.92	71.52%	89.30%	10.70%	-351990.6	-503669
32	QJR84429.1	706(A/S)	1243(C/F)	-9.9	71.52%	89.30%	10.70%	-346248	-505940

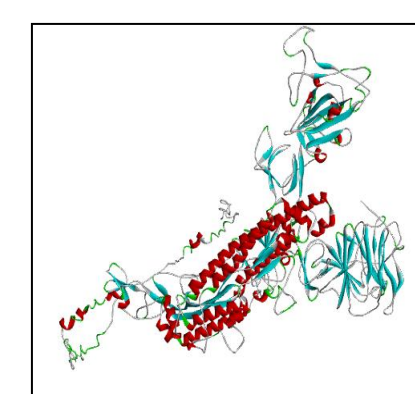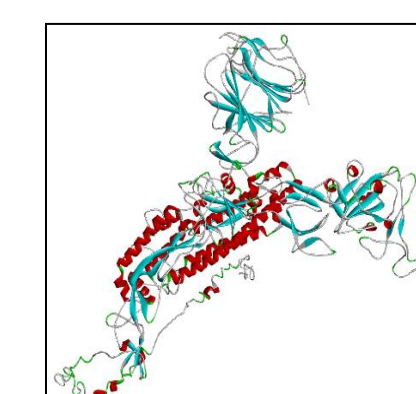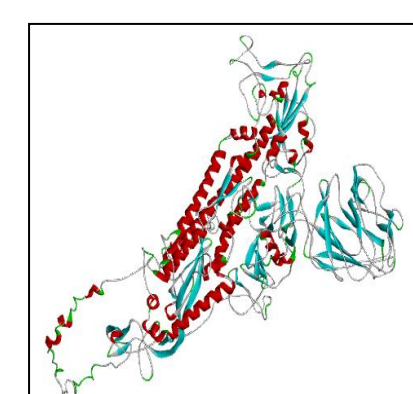
Fig 4: QKI28685.1

Fig 5: QJF77846.1

Fig 6: QIA98583.1

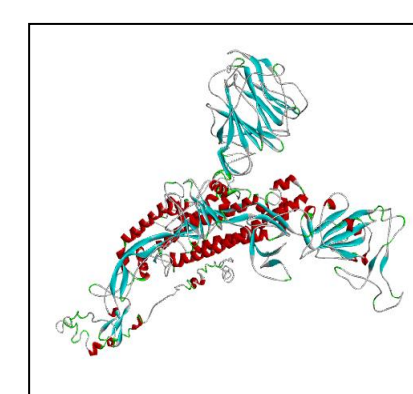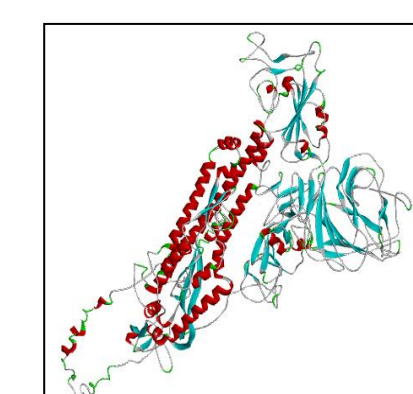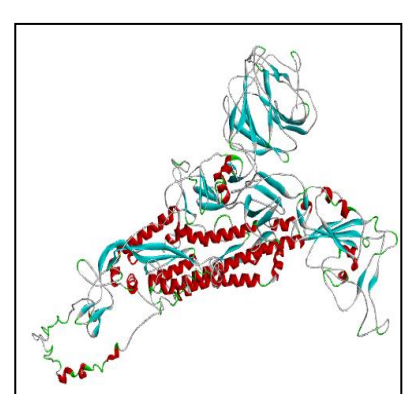
Fig 7: QKI28673.1

Fig 8: QJY40409.1

Fig 9: QJH92167.1

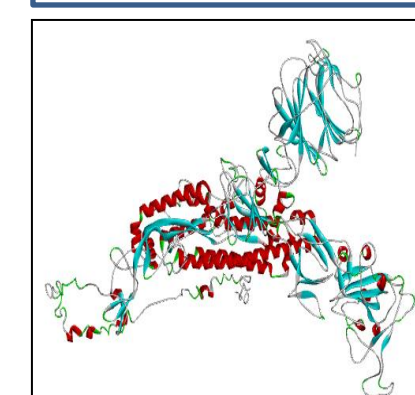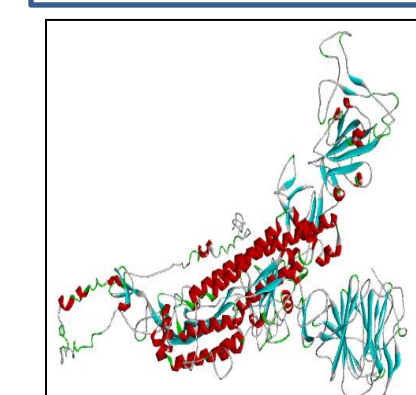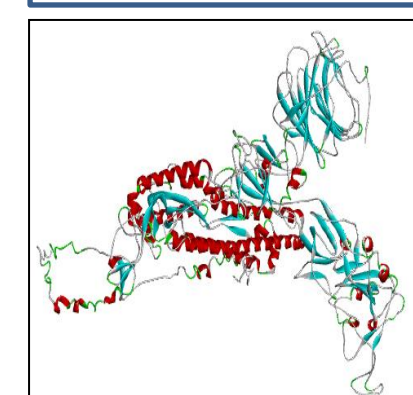
Fig 10: QJR84537.1

Fig 11: QJY40565.1

Fig 12: QJT43488.1
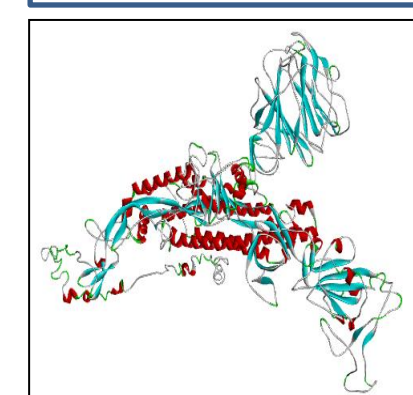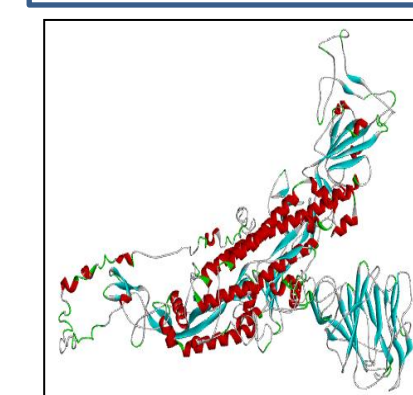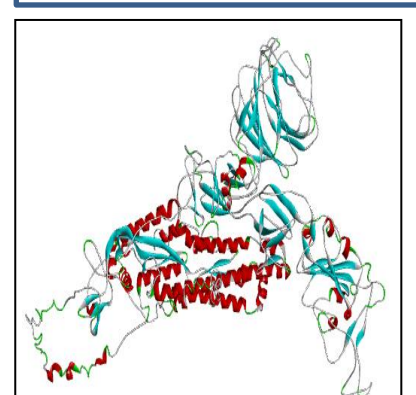

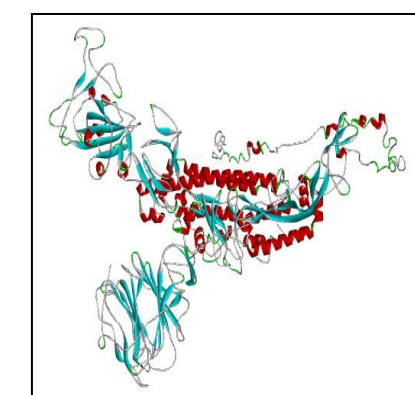Fig 13: QJY40517.1

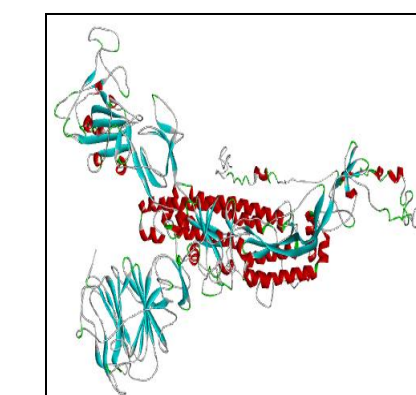Fig 14: QJT43704.1

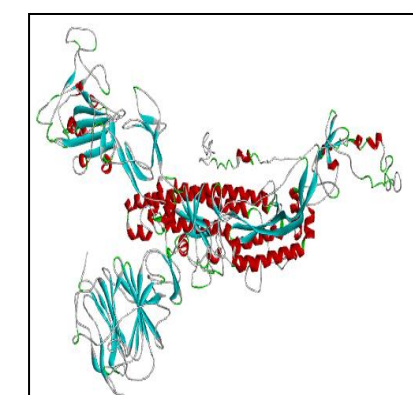Fig 15: QKG91274.1

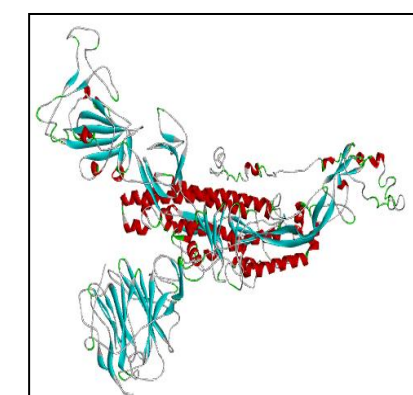Fig 16: QKI28601.1

Fig 17: QJY40469.1

Fig 18: QKE61684.1

Fig 19: QJX44574.1

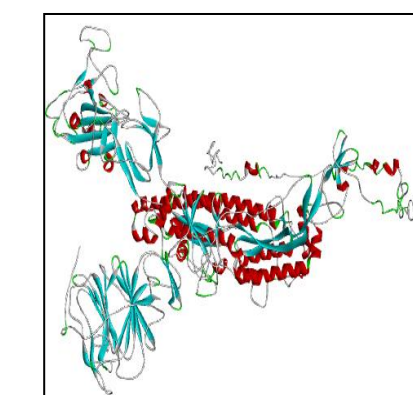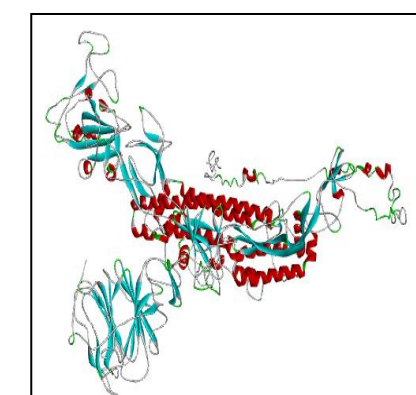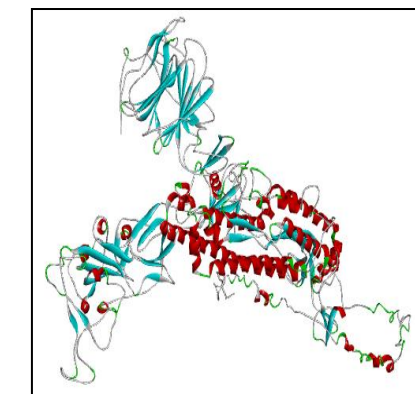
Fig 20: QJX44430.1

Fig 21: QJT43584.1

Fig 22: QJR84369.1
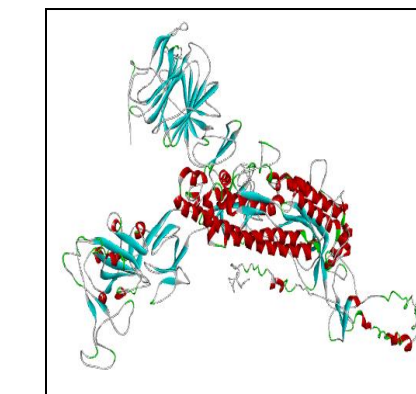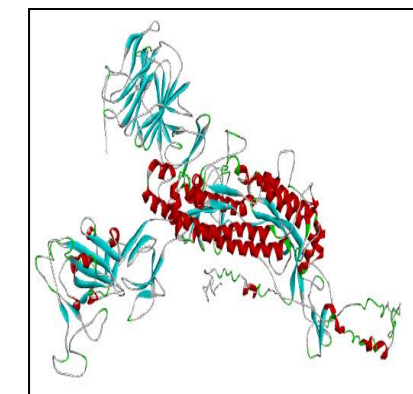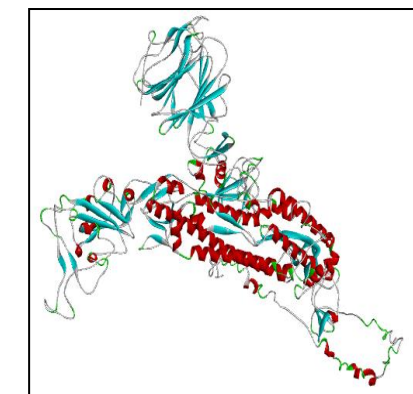

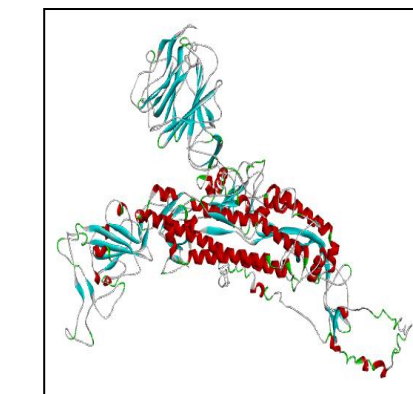Fig 23: QKI28577.1

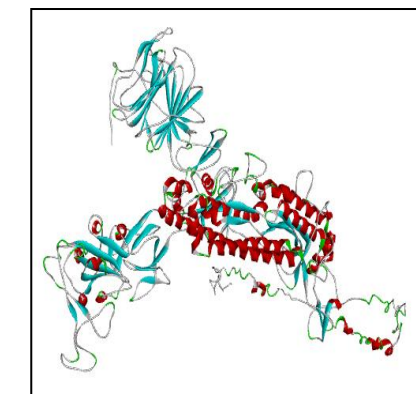Fig 24: QJX44562.1

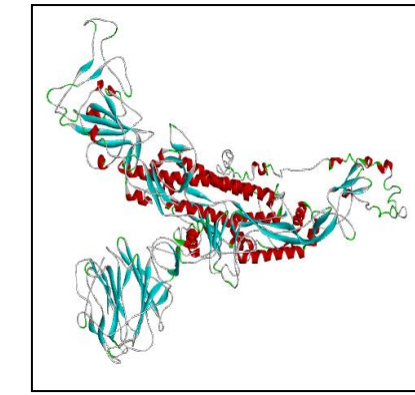Fig 25: QJT43572.1

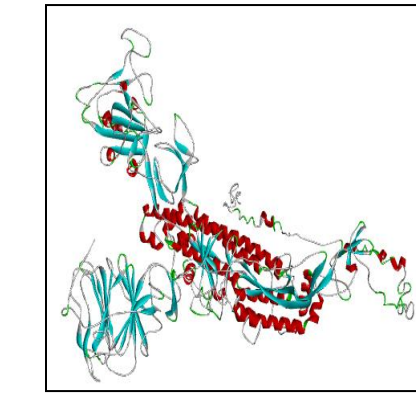Fig 26: QJF11812.1

Fig 27: QJC19491.1

Fig 28: QJT43452.1

Fig 29: QJW00303.1

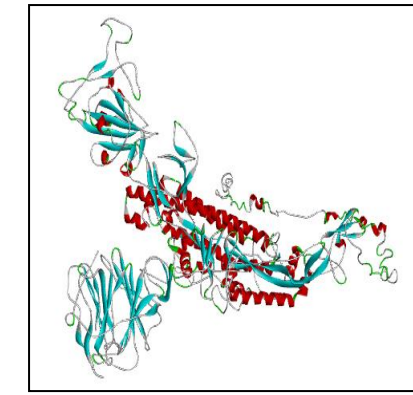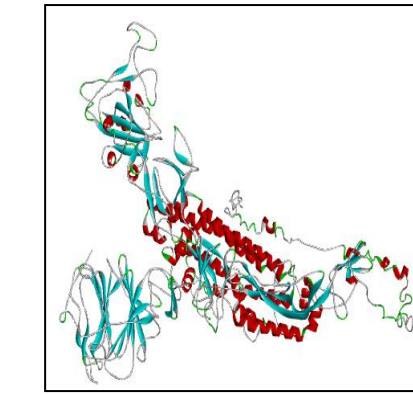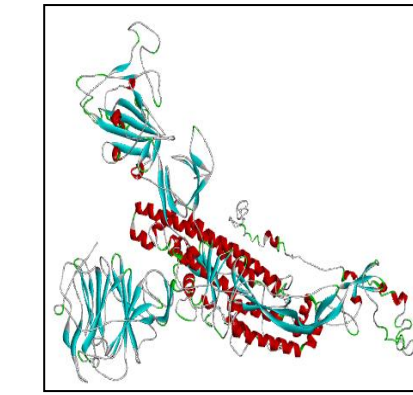
Fig 30: QJX44466.1

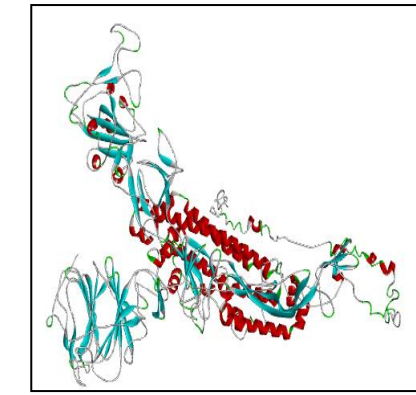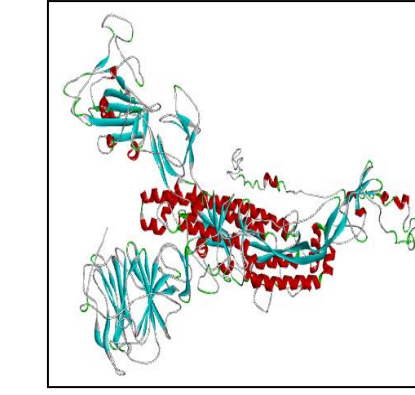Fig 31: QJX44586.1

Fig 32: QKE61804.1
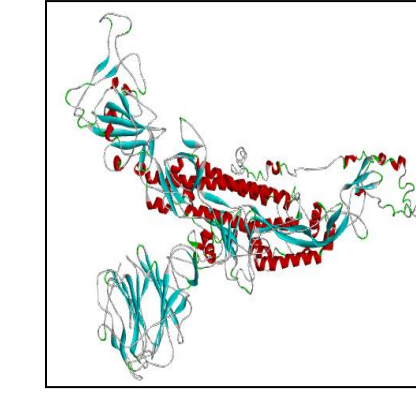

Fig 33: QJR84441.1

Fig 33: QJX44634.1

Fig 35: QJR84429.1

Conclusion

All the modelled structures were analyzed using ProSA and Ramchandran plot. Energy minimization is carried out using Yasara online server. Further these proteins are in a process to study and understand the structural changes and its impact on the protein-protein interaction and protein-drug interaction. All the modelled protein could be an initiating point to understand the impact of mutations on existing and future therapies.

Reference:

Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8. Krieger E, Joo K, Lee J, Lee J, Raman S, Thompson J, Tyka M, Baker D, Karplus K (2009) Proteins 77 Suppl 9, 114-122 PMID19768677'