 doi:10.20944/preprints202004.0281.v1

Mutation hot spots in Spike protein of COVID-19

Arup Kumar Banerjee[#3], Feroza Begum[1,2], Upasana Ray[#1,2]

[1]Infectious Biology and Immunology Division, CSIR-Indian Institute of Chemical Biology, 4, Raja S.C., Mullick Road, Jadavpur, Kolkata-700032, West Bengal, India.
[2]Academy of Scientific and Innovative Research (AcSIR), Ghaziabad- 201002, India
[3]Department of Biochemistry, North Bengal Medical College and Hospital, Sushrutanagar, Siliguri-734012, West Bengal, India

 # *Corresponding Author*

Keywords: COVID-19, SARS-CoV-2, S protein, RBD, S1, S2

Abstract

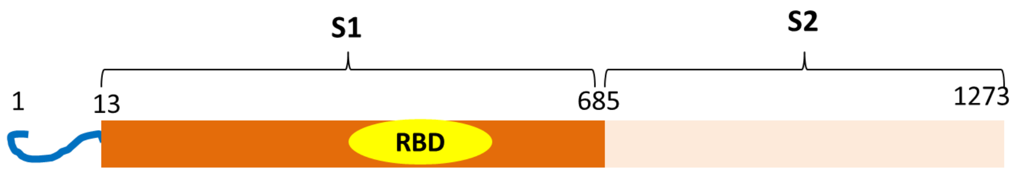
Spike protein of Coronaviruses help in receptor binding and virus entry into the host cells. While spike protein helps in receptor mediated virus entry, it is also extremely important as an immunogen as it is the most accessible part of the viral architecture. SARS-CoV2 or COVID 19 has four different structural proteins, N (nucleocapsid), M (membrane), E (envelope) and S (spike). Although all these proteins are the part of virus structure, only E, M and S are exposed towards the outer surface of the virus particle. S protein forms a knob like structure protruding outwards beyond the other structural proteins. It forms homotrimers containing an S1 and S2 as monomers and together they form the viral spikes. Mutations in structural proteins of virus play crucial role in viral virulence by determining generation of antibody escape variants and cellular tropism. In this paper we have performed in depth analyses of spike protein sequence from various parts of the world and tried to correlate the data with the current situation of virulent nature of this virus in certain parts of the world much more as compared to others. Here, we have focussed on the isolates from the North America and have pointed out three major hot spots of mutations in the S1 subunit.

Introduction

In recent times novel coronavirus 2019/ nCoV-19/ COVID 19/ SARS CoV2 infection has become a pandemic and matter of concern worldwide. As per the World Health Organization, as of 11th of April 2020, globally total confirmed COVID 19 cases have added up to 1,610,909 whereas as many as 99,690 are the number of deceased individuals. Among all the countries those comprising the Americas have seen the highest toll of affected individuals (536,664 confirmed and 19,294 deaths) after Europe. In Europe Turkey, Switzerland, Bosnia and Herzegovina, Andorra and San Marino have been declared to be facing community transmission whereas within the Americas, the entire United States of America, Canada, Brazil, Ecuador, Chile, Peru, Mexico, Panama, Dominican Republic, Combodia and Argentina have been classified to be experiencing community spread. The entire West Pacific region including the area of origin of this pandemic i.e. China has only seen sporadic spread.

In this paper we have focussed on COVID 19 isolates of the North American origin to investigate possible sequence-virulence relation of this virus in United States. We also studied the similarities and differences of North American isolates with other variants of the world.

Spike protein is one of the most important structural protein of SARS CoV2 that plays the major in virus entry. Spike protein is a 1273aa long protein with two major sub domains, S1 and S2 (Figure 1). While S1 harbours the receptor binding domain or RBD and mediates virus attachment to its ACE2 receptor, S2 carries out the function of fusion to enable successful entry. S2 contains the fusion peptide.



Schematic Representation of Spike Protein of COVID 19

Domains of Spike protein (adopted from https://covid-19.uniprot.org/uniprotkb/P0DTC2)

1-12 : Signal peptide
13-685 : S1 (contains Receptor binding domain)
686-1273 : S2
(770-778 is the fusion peptide) (Reference: UniRule: UR000875864)

Figure 1: **Schematic showing structural organization of COVID 19 spike protein**

For carrying out sequence analyses of COVID 19, we have used the protein sequence of the spike protein.

Methods

All available full length sequences of COVID-19 spike protein (1-1273) of different geographical origins belonging to North America (United States of America) (342), South America (2), Oceania (Australia)(1), China (63) and Europe (14) were downloaded in FASTA format from severe acute respiratory syndrome coronavirus 2 data hub of NCBI virus database of National Library of Medicine (NLM).

Multiple sequence alignments were done using alignment tool of NCBI virus server as well as CLUSTAL Omega. Sequence alignments from CLUSTAL Omega was viewed using MView tool.

Phylogenetic analyses were done using simple phylogeny tool of CLUSTAL W2 using neighbour joining method.

Results and Discussion

Multiple sequence Alignment of COVID-19 spike protein sequences from United States of America showed multiple mutations at few frequent locations whereas some of the parts of this protein was seen to be conserved. Table 1 shows a summary/ list of mutations observed in isolates of USA. Although there were many mutations dispersed at various sites in the spike protein sequence, few mutations occurred more frequently (Table 2, Figure 2). At position 614, mutation D to G occurred in 99 of the isolates which is clearly a very frequent mutation.

TABLE 1

Amino acid range	Mutations	Accession number
Upto 80	L54F	QIS30295
	A27V	QIU80973
	T29I	QIS60546
	H49Y	QHW06059
81-160	E96D	QIS60930
	D111N	QIS61338
161-240	T240I	QIU81585
	G176V	QIJ96493
241-320		
321-400	A348T	QIS30335
401-480	G476S	1. QIS30625
		2. QIQ49882
		3. QIQ50152
481-560	V483A	1. QIU81177
		2. QIU81549
		3. QIS60774
		4. QIS60882
		5. QIS60954
		6. QIS30165
	A520S	QIS60489
	H519Q	QIS61422
561-640	D614G	1. QIU81213
		2. QIU81741
		3. QIS61170
		4. QIU81609
		5. QIU81405
		6. QIS61110
		7. QIS30575
		8. QIS30615
		9. QIS30295
		10. QIS30115
		11. QIV64965
		12. QIV64989
		13. QIV15032
		14. QIV15044
		15. QIV15020
		16. QIV15188
		17. QIV14972
		18. QIV15116
		19. QIV15152
		20. QIV15128
		21. QIV15176
		22. QIV15104
		23. QIU81537
		24. QIU81681
		25. QIU81513
		26. QIU81597
		27. QIU81633
		28. QIU81693
		29. QIU81237
		30. QIU81189
		31. QIU81345
		32. QIU81717
		33. QIU81561
		34. QIU81393
		35. QIU80985
		36. QIU81141
		37. QIU81057
		38. QIU81105
		39. QIU81153
		40. QIU81117
		41. QIU81165
		42. QIU81417
		43. QIU81321
		44. QIU81465
		45. QIU81501
		46. QIU81033
		47. QIU81021
		48. QIU81129
		49. QIU80997
		50. QIT06951

Amino acid range	Mutations	Accession number
561-640	D614G	50. QIT06951
		51. QIT06963
		52. QIT06879
		53. QIT06927
		54. QIS61278
		55. QIS60834
		56. QIS60870
		57. QIS60894
		58. QIS61122
		59. QIS61182
		60. QIS60618
		61. QIS60642
		62. QIS60630
		63. QIS60666
		64. QIS61194
		65. QIS61206
		66. QIS61542
		67. QIS61554
		68. QIS61518
		69. QIS60762
		70. QIS61050
		71. QIS60810
		72. QIS60942
		73. QIS60918
		74. QIS61134
		75. QIS61242
		76. QIS61446
		77. QIS61290
		78. QIS61074
		79. QIS30665
		80. QIS30125
		81. QIS30075
		82. QIS30095
		83. QIS30535
		84. QIS30585
		85. QIS30235
		86. QIS30085
		87. QIS30245
		88. QIS30265
		89. QIS30445
		90. QIS30435
		91. QIS30195
		92. QIS30405
		93. QIS30155
		94. QIS30385
		95. QIS30305
		96. QIS30315
		97. QIS30285
		98. QIS30365
		99. QIS30415

Amino acid range	Mutations	Accession number
641-720		
721-800		
801-880		
881-960	D936Y	QIS30615
961-1040		
1041-1120	A1078V	QIS61254
1121-1200		
1201-1273	D1259H	QIS60582

TABLE 1 (CONTD.)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 April 2020 doi:10.20944/preprints202004.0281.v1

TABLE 2

List of mutations	Occurrence of mutations (Number of isolates that showed the mutation)
L54F	1
A27V	1
T29I	1
H49Y	1
E96D	1
D111N	1
T240I	1
G176V	1
A348T	1
G476S	3
V483A	6
A520S	1
H519Q	1
D614G	116
D936Y	1
A1078V	1
D1259H	1

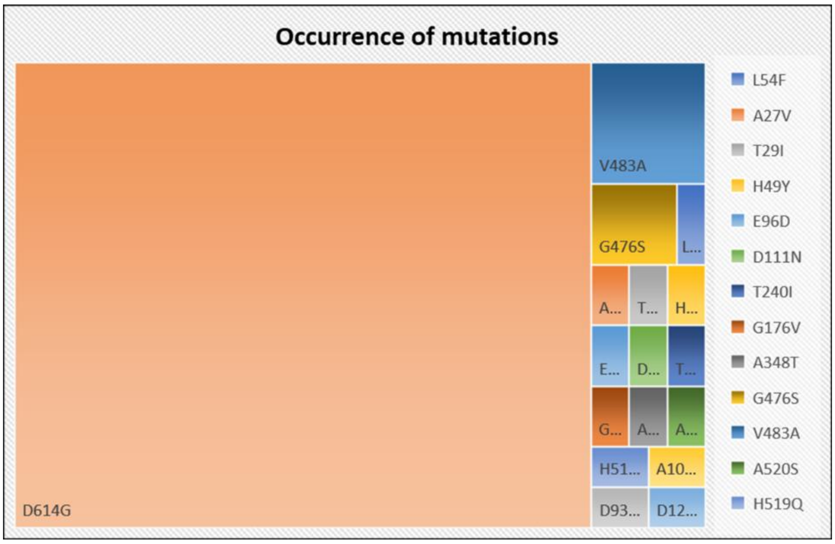


Figure 2: **Overall distribution of mutations in the analysed in isolates from North America**. Graph was plotted based on Table 2. The mutations have been shown in different colours on right side of the figure.

Receptor binding domain (RBD) of COVID 10 falls between the amino acids 331 and 524 [1]. In the receptor binding domain three different sites were seen to be mutated: A348T; G476S and V483A. Out of these, V483A repeated more frequently followed by G476S (Figure 3).

We compared the sequences of North American origin with all the available sequences from South America (Figure 4), Europe (Figure 5) and China (Figure 6). In case of South American isolates, one of the samples showed mutation of position 614 from D to G as seen in case of the isolates from North America. However, the other mutations at positions 348, 476 and 483 were not present. None of these mutations were seen

 doi:10.20944/preprints202004.0281.v1

in Australian isolates. Unlike sequences from Asia, Australia and China, four out of fourteen European isolates aligned showed the same mutation at position 614 as seen in case of isolates from USA. It is thus possible that a branch of mutants of European origin entered USA. Thus, European form of the COVID 19 seems to be closer to that of American virus type with respect to the spike protein sequence.

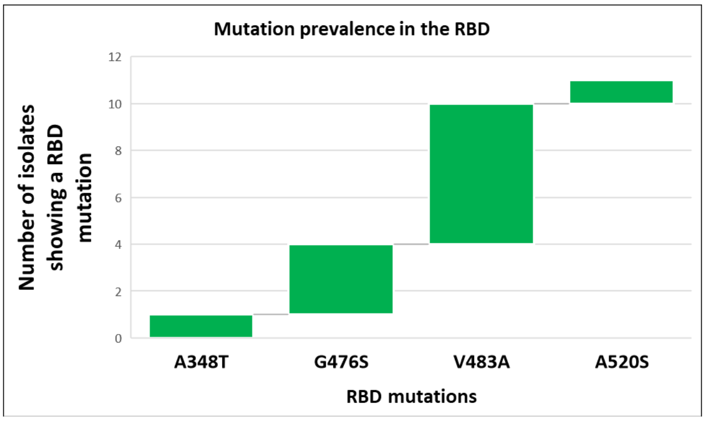


Figure 3: **Distribution of mutations in the RBD of Spike protein in isolates from North America**. X axis shows the mutations and the Y axis shows number of isolates harbouring the mutation.

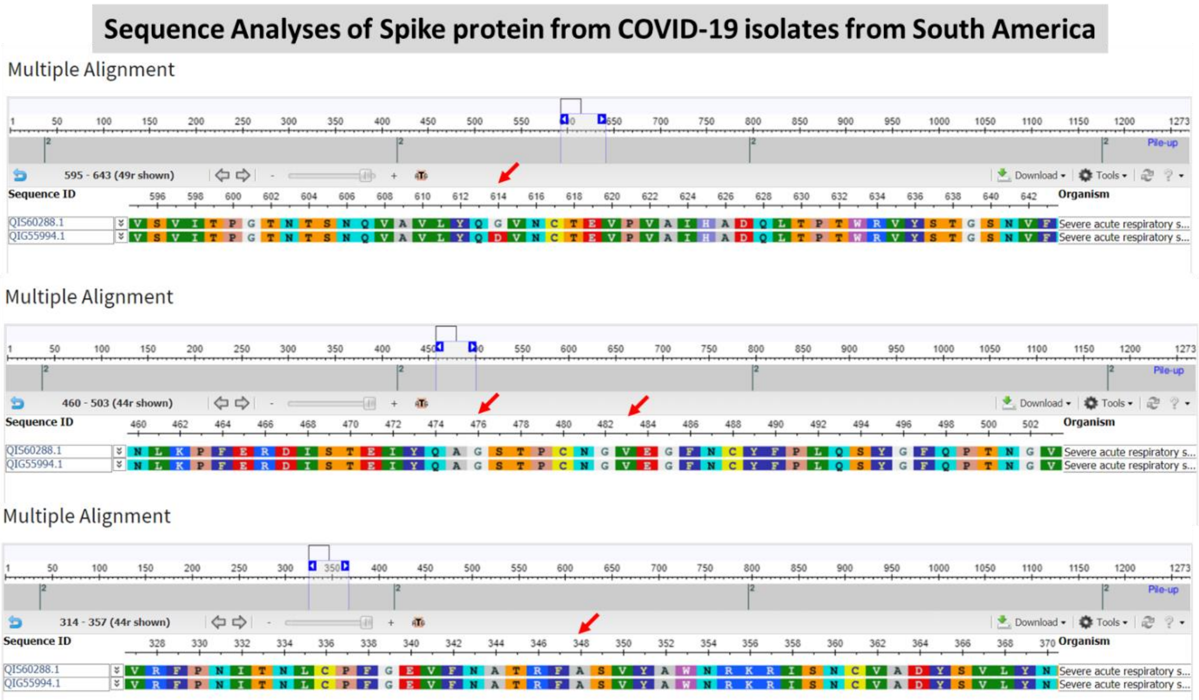


Figure 4: **Sequence alignment of spike protein of South American isolates**. Four sites of mutations that was seen in case of North American isolates have been highlighted here with red arrows: position 614 (upper panel) positions 476 and 483 (middle panel) and position 348(bottom panel)

doi:10.20944/preprints202004.0281.v1

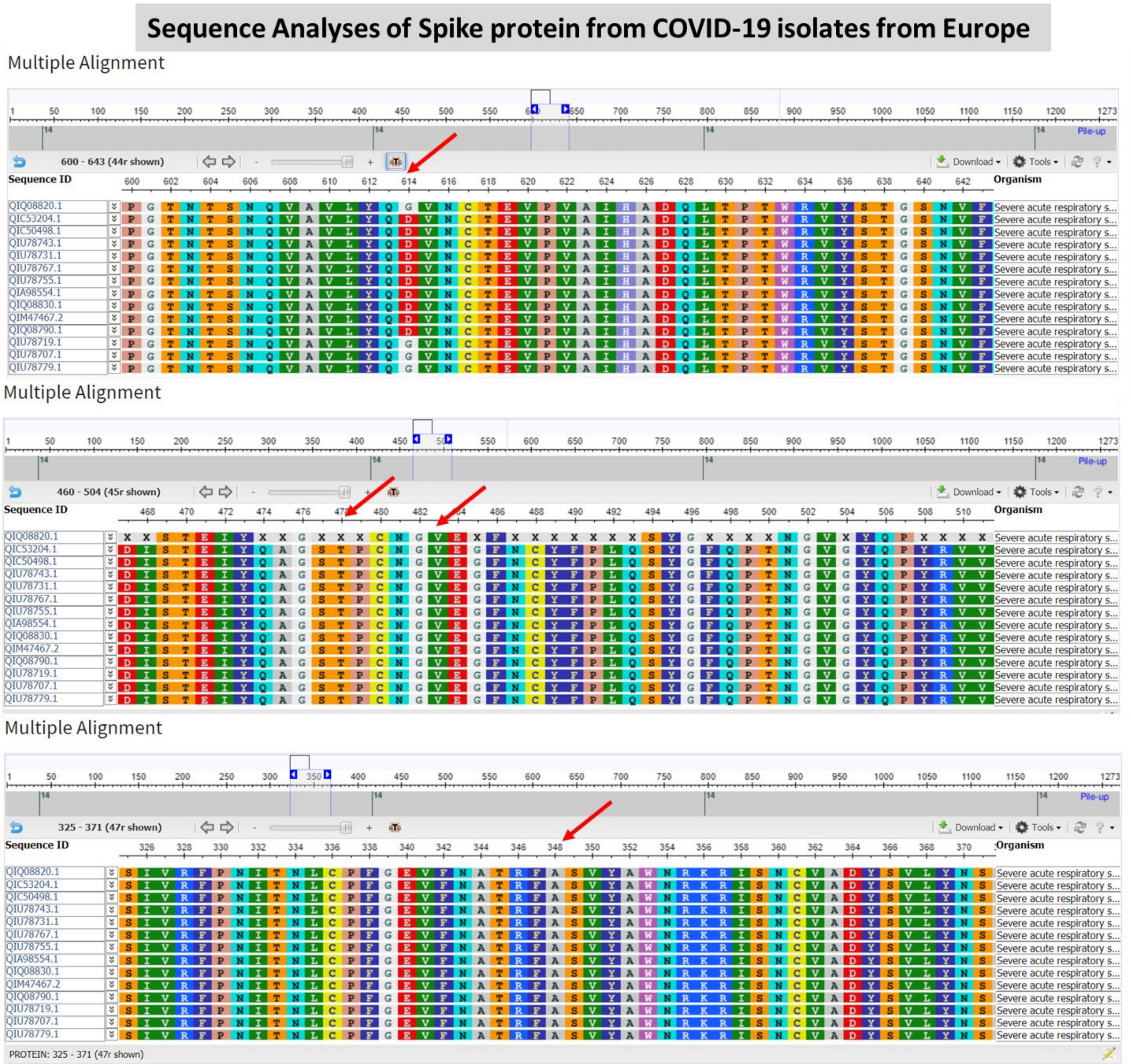


Figure 5: **Sequence alignment of spike protein of European isolates**. Four sites of mutations that was seen in case of North American isolates have been highlighted here with red arrows: position 614 (upper panel) positions 476 and 483 (middle panel) and position 348(bottom panel). Position 614 of four of the isolated showed D to G mutation similar to the USA counterparts.

We compared all available sequences from China but none of the highlighted mutations were found to exist in the Chinese sequences (Figure 6). Thus, the virus that continues to spread in the America is different based on this sequence analyses of spike protein than the original Wuhan virus.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 April 2020 doi:10.20944/preprints202004.0281.v1

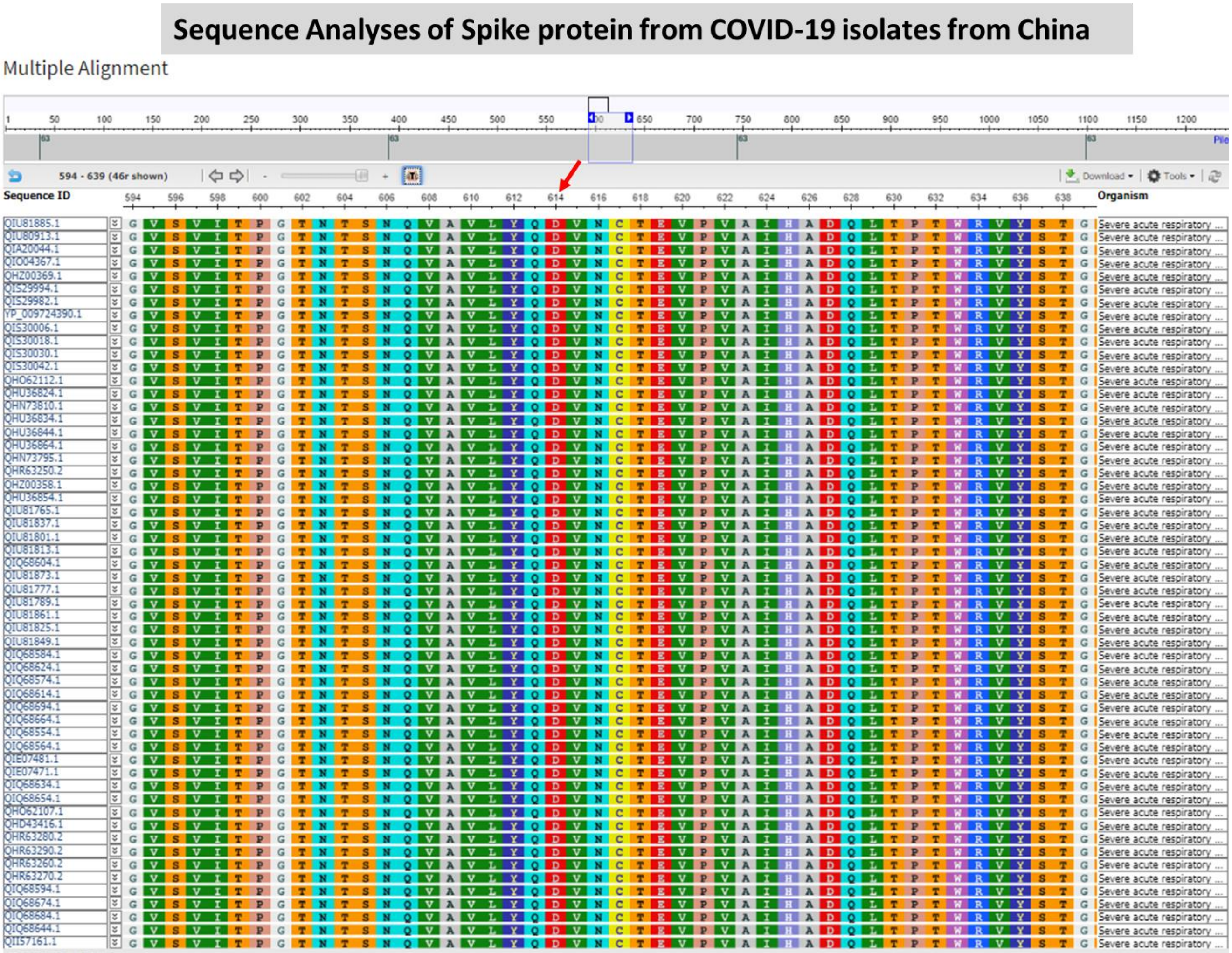


Figure 6: **Sequence alignment of spike protein of Chinese isolates**

This study revealed that the spike protein of COVID 19 virus of USA is mutating at various sites. To determine if this virus is evolving in different clusters, we performed phylogeny of all the available 342 American spike protein sequences. It was observed that this virus diverged in at least twenty clusters. Out of these three major clusters were prominent. Cluster 1 was comprised of mutation G476S (Figure 5A) and cluster 2 had mutation V483A (Figure 5B). All these mutations fall in the RBD of the spike protein. Thus, the RBD mutants might be evolving into different directions which might also get reflected in the infectivity of these isolates.

Mutation at 614 from Aspartic acid to glycine which appeared in almost 99 isolates formed a very big cluster. Mutation from Aspartic acid to glycine is a potentially crucial change in a protein sequence as Aspartic acid is a big negatively charged, acidic amino acid and on the other hand Glycine is a small neutral amino acid.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 April 2020 doi:10.20944/preprints202004.0281.v1

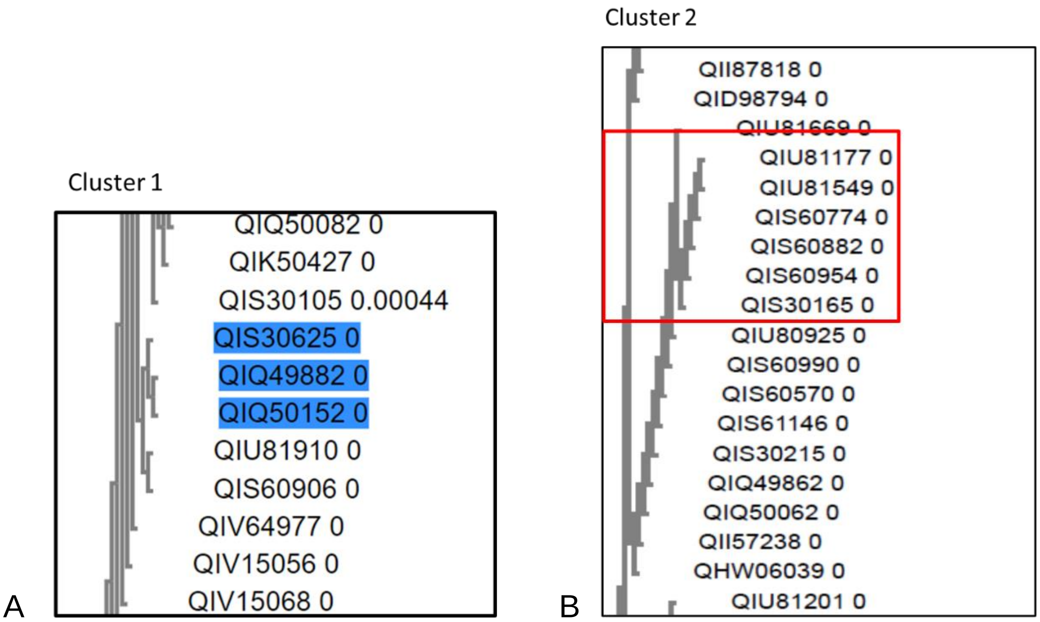


Figure 7: **Phylogeny of Spike protein sequences of North America.** Panel A shows cluster with mutation G476S and panel B shows cluster with mutation V483A

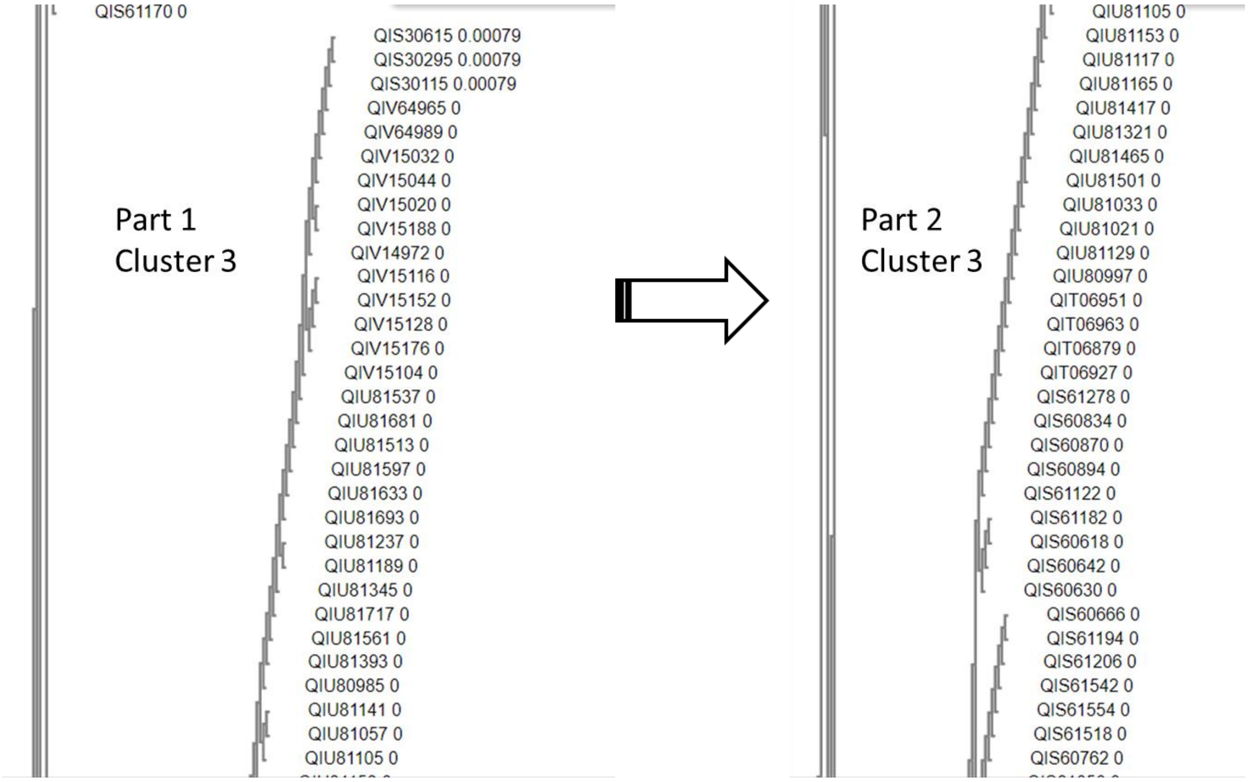


Figure 8A: **Phylogeny of Spike protein sequences of North America.** Cluster with mutation D614G. This is a very big cluster and thus it has been shown in four parts compiled in figures 8A and 8B

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 April 2020 doi:10.20944/preprints202004.0281.v1

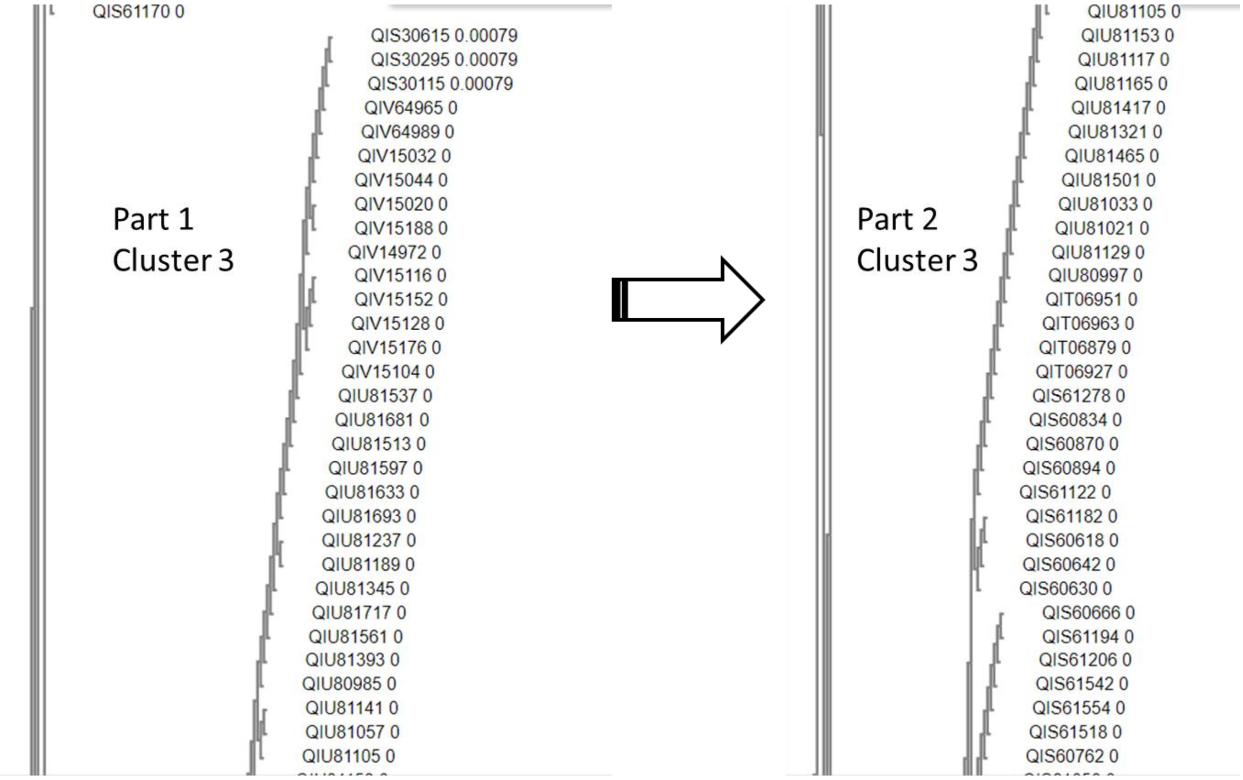


Figure 8B: **Phylogeny of Spike protein sequences of North America.** Cluster with mutation D614G. This is a very big cluster and thus it has been shown in four parts compiled in figures 8A and 8B

While many mutations have been identified in the S protein, there are portions of S protein that remained unchanged and might be conserved. These are the portions lying between 241-320; 641-880, 961-1040 and 1121-1200. Out of these major part of the sequence between 641-880, 961-1040 and 1121-1200 fall in the S2 domain of the spike protein. This indicates that majority of the conserved or non-changing zones fall in the S2 domain and thus could be used for designing therapeutic candidates or as antiviral targets. These features should also be taken care of while designing vaccine candidates for this virus where S protein is used as target.

The data presented here is based on the currently available sequences. Further sequencing from other parts of the North America and other countries would shed more light on the nature of this virus.

Conflict of Interest

Authors declare no conflict of interests.

References

1. Tai, W., He, L., Zhang, X., Pu, J., Voronin, D., Jiang, S., Zhou, Y. and Du, L., 2020. *Characterization Of The Receptor-Binding Domain (RBD) Of 2019 Novel Coronavirus: Implication For Development Of RBD Protein As A Viral Attachment Inhibitor And Vaccine*.