 doi:10.20944/preprints202004.0450.v1

Novel mutations in the S1 domain of COVID 19 Spike protein of isolate from Gujarat origin, Western India

Arup Kumar Banerjee[#3], [$]Feroza Begum[1,2], [$]Dlyuya Thagriki[1,2], Prem Prakash Tripathi[#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
$ Joint second authors

Keywords: COVID-19, S protein, RBD, S1, S2, Mutation

Abstract

Till date there are three full length COVID 19 virus genome sequences available from India. The earlier two were reported from Kerala, Southern India and the more recent one has been reported from Gujarat, western part of India. In this paper we report two novel mutations in the Spike protein sequence of Gujarat's isolate. These mutations are based on comparison with the original Wuhan sequence. The two mutations have been found to be located just upstream and downstream of the receptor binding domain (RBD). Out of these two one has also been found to affect the secondary structure of the S1 domain. Since both of these mutations lie near the receptor binding domain, they might influence the spike receptor interactions by changing the conformation of the spike protein S1 domain. These mutations are uniquely placed and might be important in the context of vaccine engineering and therapeutic interventions.

Introduction

The outbreak of COVID 19 virus or severe acute respiratory syndrome coronavirus 2 (SARS-CoV2) started in December 2019 in Wuhan [1]. Eventually this virus spread to other parts of the world and took the shape of a pandemic recently. Today it is a global pandemic and has taken many lives while many others are suffering. Almost every continent has been touched by this novel SARS coronavirus and the risk is increasing every day. Borders have been sealed and travel restricted. In many countries community spread has also started which gets updated on regular basis by the World Health Organization.

Governments throughout the globe are emphasizing on elaborate screening, genome sequencing, drug repurposing, new drug designing and urgent vaccine development. The therapeutic interventions need in depth knowledge about the genetic sequence of this virus. Thus, sequencing and sequence analyses are important handles for therapeutic targeting against COVID 19.

 doi:10.20944/preprints202004.0450.v1

COVID 19 is an enveloped virus having four different structural proteins, N (nucleocapsid), M (membrane), E (envelope) and S (spike) [2]. S protein mediates receptor binding and virus entry. S protein has two domains S1 and S2. Each S protein forms a homotrimer of S1 and S2 and form knob like structures protruding outwards from the viral envelope. S1 is responsible for receptor attachment and S2 mediates cell fusion.

Since S protein, owing to its importance in receptor binding and virus tropism, has been the most attractive drug target. Thus, in this paper we have focussed on the S protein and studied the recently sequenced isolates from India to identify the similarities and differences in the protein sequence with respect to Wuhan isolates.

Methods

Sequences

All available full-length sequences of COVID-19 spike protein (1-1273) belonging to India (3), China (63), Pakistan (2) and Nepal (1) were downloaded in FASTA format from severe acute respiratory syndrome coronavirus 2 data hub of NCBI virus database of National Library of Medicine (NLM).

Sequence analysis tools

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

CFSSP (Chou and Fasman secondary structure prediction) server [3] was used to predict the secondary structure of S1 domain. Since this server can take up to 1000aa residues at a time, we use 1-1000aa sequence of S protein of COVID 19 for secondary structure prediction.

Results and Discussion

Currently we have three full sequences of COVID 19 genome that are available of public databases. The most recent one (collection date: April 5, 2020) is from Gujarat, a state in western part of India. Earlier we have published our studies on the other two previously announced Indian COVID 19 isolates [4]. Here, we studied the Spike protein sequence of the Gujarat isolate. Since the origin of this outbreak was from Wuhan, we have considered the sequence from Wuhan as our template of comparisons. Multiple sequence alignment revealed that all the Wuhan spike sequences were identical. So, we chose one of these for our further studies. Multiple sequence alignment of spike protein sequence (1-1273) of Wuhan isolate with the Gujarat isolate revealed that there were two mutations in the Gujarat isolate (Figure 1).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 April 2020 doi:10.20944/preprints202004.0450.v1

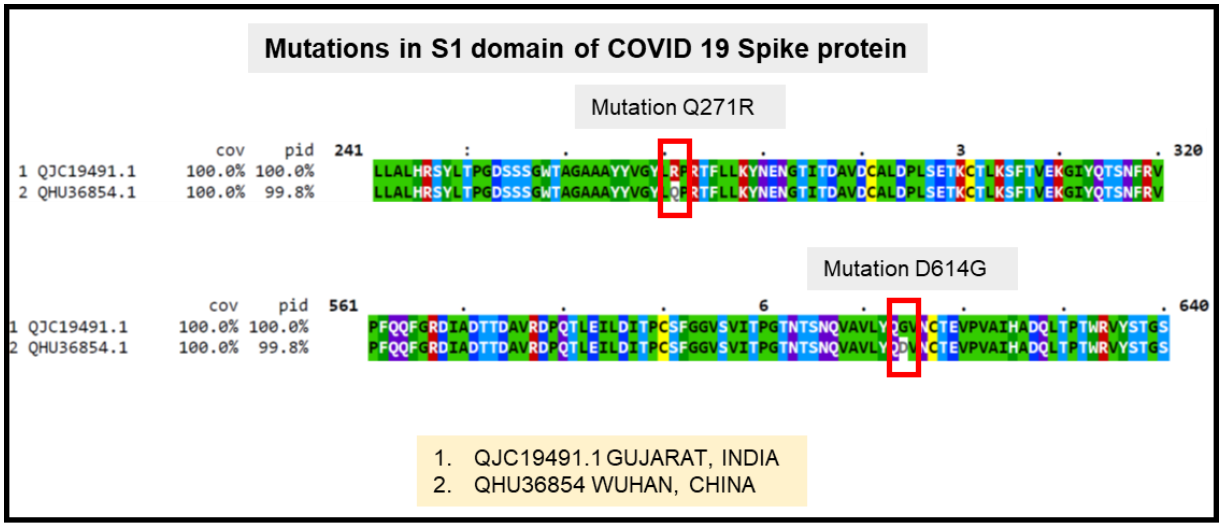
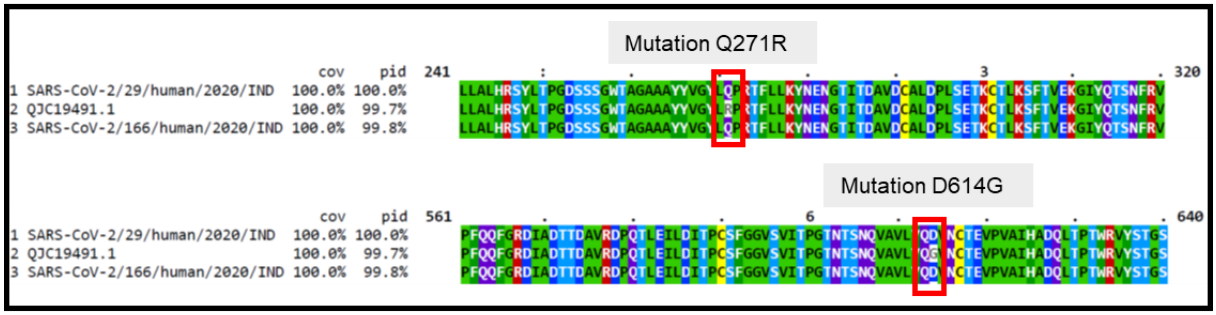


Figure 1: Mutation in S1 domain of Gujarat isolate: Multiple sequence alignment of spike protein sequence of Wuhan and Gujarat isolates showing the sites of mutation in Gujarat isolate

These mutations were at positions 271 (Q271R) and 614 (D614G). In case of Gujarat isolate at position 271, the amino acid glutamine (Q) got substituted to arginine (R) and at position 614 there was a mutation from aspartic acid (D) to glycine (G). Comparison of Gujarat isolate with that of the other two Indian isolates showed that these two mutations were novel and unique as compared to the other two and were not present in these isolates from the state of Kerala (Figure 2). Since Pakistan and Nepal are border countries for India and two full sequences from Pakistan and one from Nepal are available as of April 23, 2020 in GenBank, we also looked for the Gujarat mutations in Pakistan's and Nepal's COVID 19 spike protein sequences. However, these were absent in both the geographic origins i.e. Pakistan and Nepal (data not shown).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 April 2020 doi:10.20944/preprints202004.0450.v1

A.



B.

Isolates	Mutation Q271R	Mutation D614G
Gujarat Isolate	+	+
Kerala Isolate 29	-	-
Kerala Isolate 166	-	-

Figure 2: Mutation analysis in S1 domain of all three Indian isolates: A. Multiple sequence alignment of spike protein sequence of Indian isolates showing the sites of mutation in Gujarat isolate. B. Mutations in Gujarat isolates not found in Kerala isolates

The mutation D614G was also found in many sequences from North America and some in Europe as observed in one of our previous studies [5]. Thus, we think that a migrant from one of these parts of the world might have undergone additional mutation at position 271 to give rise to the current Gujarat variant. Aspartic acid is a bulky negatively charged, acidic amino acid and glycine neutral amino acid, quite smaller in size. Hence, a D to G mutation could lead to electrostatic alterations in the tertiary structure of the protein. Both these mutations have been found to lie near the RBD in S1 domain (Figure 3) of spike protein thus implying a possibility of modulation of the receptor binding activity of S1 domain.

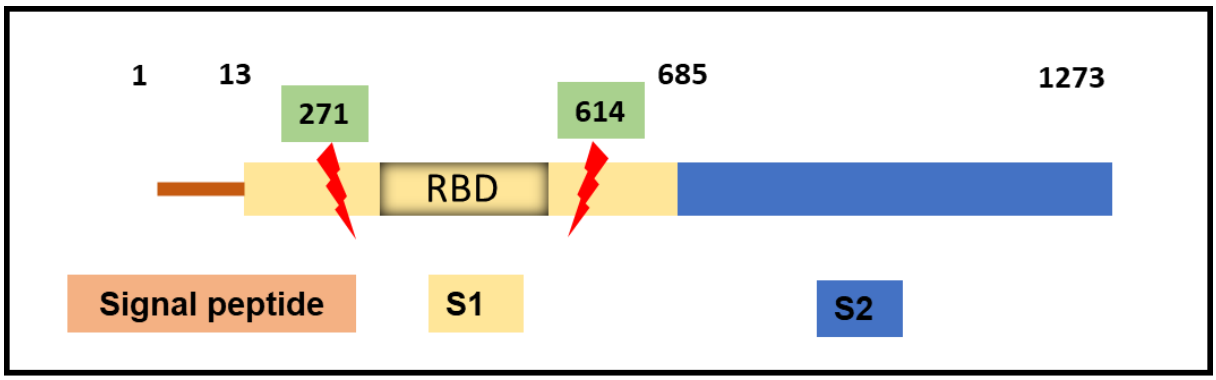


Figure 3: Schematic representation pf spike protein showing sites of mutation: 1-13aa represents the signal peptide at the N terminus region of spike protein.13-685aa belongs to S1 domain and 686-1273aa comprises the S2 domain. Two mutations of Gujarat isolate at positions 271 and 614 have been shown with red marks.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 April 2020 doi:10.20944/preprints202004.0450.v1

Secondary structure analyses revealed that in case of the mutation Q271R, there was a change in the secondary structure of Gujarat isolate as compared with that of the Wuhan isolate (Figure 4). In case of Wuhan isolates, there were seven helices from aa274-280. However, in case of Gujarat isolate where there was a mutation at 271residue, there were six helices from aa275-280. At position 274, one of the helices got changed to sheet.

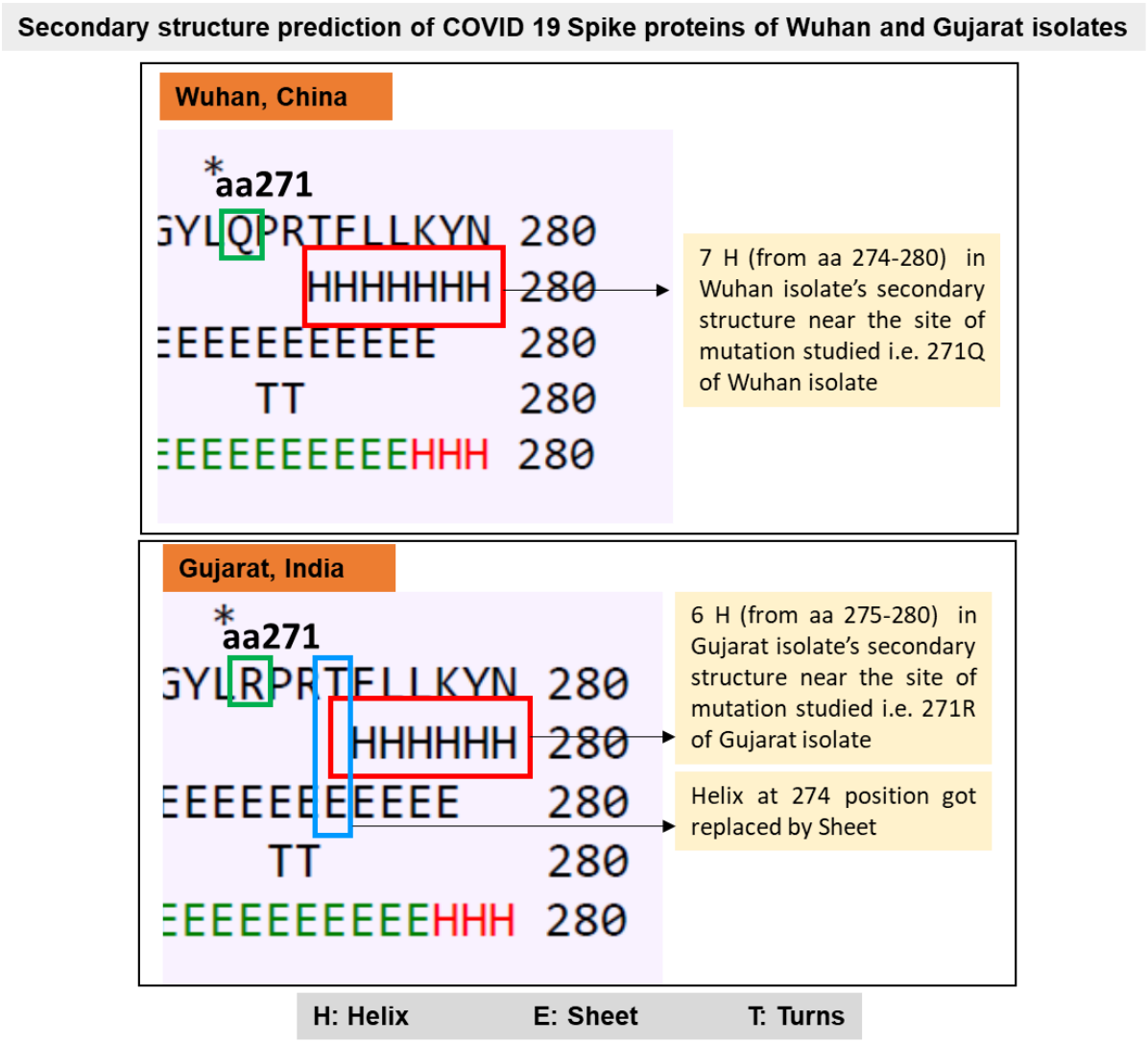


Figure 4: Secondary structure prediction of spike protein: Secondary structure prediction showing change in Gujarat isolate due to Q271R mutation

All these mutations that appeared later after the initial outbreak in Wuhan might have evolutionary advantage over the original strain. This should follow the idea of 'survival of the fittest'. Mutations in S protein might lead to (a) better binding to host cell receptor and increased virulence (b) reduced binding to host cell receptor and thus reduced virulence that might in turn help in escaping the immune system and better adaptation to the new environment (c) eventual generation of antibody escape mutants (d) generation of variants capable of binding to alternate receptors thereby expanding tissue tropism. While there are hot spots of mutation in S1, there are few conserved regions in S1 and throughout S2 domain (as seen in sequence analyses of the virus from various geographic origins). Elaborate knowledge in this area will help choosing

 doi:10.20944/preprints202004.0450.v1

better antigens/ immunogenic regions for vaccine development as well as other therapeutic interventions like antiviral designing. Poor choice of immunogenic epitopes might lead to designing of vaccine candidates that would induce production of poor antibody pool. Such antibodies instead of neutralizing the virus might potentially lead to antibody dependent enhancement of infection wherein poorly binding antibodies instead of preventing receptor mediated virus entry, would bind to Fc receptors via the Fc domain and would lead to Fc receptor mediated virus entry and increase in infection.

All the analyses and conclusions are based on available sequences. More sequencing will help validating the observations.

Conflict of Interest

Authors declare no conflict of interests.

References

1. Huang C, Wang X, Li L, Ren J, Zhao Y, et al Hu L. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. 2020; 395:497–506.
2. Tai W, He L, Zhang X, Pu J, Voronin D, Jiang S et al. Characterization of the receptor-binding domain (RBD) of 2019 novel coronavirus: implication for development of RBD protein as a viral attachment inhibitor and vaccine. 2020.
3. Feng X, Wang Z, Shi J, Deng G, Kong H et cl. Glycine at Position 622 in PB1 Contributes to the Virulence of H5N1 Avian Influenza Virus in Mice. Journal of Virology. 2016;90(4):1872-1879.
4. Saha P, Banerjee A K, Tripathi P P, Srivastava A K, Ray U. A virus that has gone viral: Amino acid mutation in S protein of Indian isolate of Coronavirus COVID-19 might impact receptor binding and thus infectivity. bioRxiv. 2020; 04.07.029132
5. Banerjee A K, Begum F, Ray U. Mutation Hot Spots in Spike Protein of COVID-19. Preprints. 2020, 2020040281