F1000Research

Check for updates

BRIEF REPORT

REVISED **Analyses of spike protein from first deposited sequences of SARS-CoV2 from West Bengal, India** [version 2; peer review: 2 approved, 1 approved with reservations]

Feroza Begum[1,2]*, Debica Mukherjee[1,2]*, Dluya Thagriki[1,2]*, Sandeepan Das[1,2]*, Prem Prakash Tripathi[1,2]*, Arup Kumar Banerjee[3]*, Upasana Ray [iD][1,2]*

[1]CSIR-Indian Institute of Chemical Biology, Kolkata, West Bengal, 730032, India
[2]Academy of Scientific and Innovative Research, Ghaziabad, 201002, India
[3]Department of Biochemistry, North Bengal Medical College and Hospital, Sushrutanagar, West Bengal, 734012, India

* Equal contributors

Abstract

India has recently started sequencing SARS-CoV2 genome from clinical isolates. Currently only few sequences are available from three states in India. Kerala was the first state to deposit complete sequence from two isolates followed by one from Gujarat. On April 27, 2020, the first five sequences from the state of West Bengal (Eastern India) were deposited on GISAID, a global initiative for sharing avian flu data. In this study, we have analysed the spike protein sequences from all five isolates and also compared their similarities or differences with other sequences reported in India and with isolates of Wuhan origin. We report one unique mutation at position 723 and another at 1124 in the S2 domain of spike protein of the isolates from West Bengal only. There was one mutation downstream of the receptor binding domain at position 614 in S1 domain which was common with the sequence from Gujarat (a state of western India). Mutation in the S2 domain showed changes in the secondary structure of the spike protein at region of the mutation. We also studied molecular dynamics using normal mode analyses and found that this mutation decreases the flexibility of S2 domain. Since both S1 and S2 are important in receptor binding followed by entry in the host cells, such mutations may define the affinity or avidity of receptor binding.

Keywords

Spike, India, West Bengal, Mutations

Open Peer Review

Approval Status ✔ ? ✔

	1	2	3
version 2 (revision) 27 Sep 2023			
version 1 18 May 2020	✔ view	? view	✔ view

1. **Pawan Kumar Maurya**, Central University of Haryana, Mahendragarh, India

2. **Jalaj Gupta** [iD], Georg-Speyer-Haus, Lucknow, India

3. **Nicolas Cuburu**, National Institutes of Health, Bethesda, USA

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the Emerging Diseases and Outbreaks gateway.
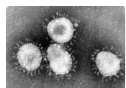
This article is included in the Coronavirus collection.

Corresponding authors: Prem Prakash Tripathi (prem.tripathi@gmail.com), Arup Kumar Banerjee (dr.arup.banerjee@gmail.com), Upasana Ray (ray.upasana@gmail.com)

Author roles: Begum F: Conceptualization, Formal Analysis, Investigation, Writing – Review & Editing; **Mukherjee D**: Conceptualization, Formal Analysis, Methodology, Writing – Review & Editing; **Thagriki D**: Formal Analysis, Writing – Review & Editing; **Das S**: Conceptualization, Formal Analysis, Writing – Review & Editing; **Tripathi PP**: Conceptualization, Writing – Review & Editing; **Banerjee AK**: Conceptualization, Writing – Review & Editing; **Ray U**: Conceptualization, Formal Analysis, Investigation, Methodology, Project Administration, Supervision, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

> **_REVISED_** **Amendments from Version 1**
>
> The affiliations 1 and 2 were typed wrong. We have corrected the same. There is no actual change in affiliations of the authors. Scientifically there is no added change.
>
> **Any further responses from the reviewers can be found at the end of the article**

Introduction

The SARS-CoV2 (a member of Coronaviruses) outbreak occurred in Wuhan, China in the year 2019, and was recently declared a pandemic, which has affected countries worldwide. To design antiviral therapeutics/vaccines it is important to understand the genetic sequence, structure and function of the viral proteins. When a virus tries to adapt to a new environment, in a new host, in a new geographical location and a new population, it may make changes to its genetic make-up which in turn bring slight modifications in viral proteins. Such variations would help the virus to utilize the host's machinery to best in favour virus survival and propagation. Since the host's immune system eventually learns to identify an infecting pathogen and starts producing protective antibodies, a virus often changes its structural proteins so that it escapes the host's immune system and continues to infect the host cells. Coronaviruses have been long known to undergo rapid mutations in its RNA genome[1]. Such mutations are reflected in changes in the amino acid sequences of its structural and non-structural proteins.

The spike protein is a structural protein of SARS-CoV2 that forms a homotrimer on the surface of the vital lipid envelope[2]. This trimer is made up of monomers consisting of S1 and S2 subunits. S1 subunit helps in attachment to the host cell receptor, while S2 subunit helps in fusion to host cell and entry. Thus, the spike protein has been an area of interest for designing vaccine and antiviral candidates against SARS-CoV2. Since the spike protein tends to mutate, it is important to obtain a broad mutation profile of this protein from extensive genome sequencing from different geographical locations of the world. Targeting areas of the spike protein that do not undergo mutation, i.e. conserved regions, would be the key to designing effective broad spectrum antivirals or vaccine.

Methods
Sequences
We downloaded the five new SARS-CoV2 sequences from West Bengal (EPI_ISL_430468; EPI_ISL_430467; EPI_ISL_430465; EPI_ISL_430464; EPI_ISL_430466) from the GISAID database under the EpiCov icon by specifying location as India and the spike protein sequences corresponding to Kerala isolates (QIA98583 and QHS34546)[3], Gujarat isolate (QJC1949.1)[4] and Wuhan isolate (QIS29982.1) from the NCBI virus database. Wuhan isolate was used as the original SARS-CoV-2 sequence for sequence comparisons. For this study we have considered only the complete sequences (>29,000bp) with high coverage.

For the West Bengal isolates, the nucleotide sequences corresponding to the spike protein were selected and then translated on ExPASY Translate tool to obtain the protein sequences.

Sequence alignments and structure
All the spike protein sequences were aligned using the multiple sequence alignment platform of CLUSTAL Omega. The alignment file was viewed using MView and differences in the sequence or the amino acid changes were recorded.

CFSSP (Chou and Fasman secondary structure prediction) server was used to predict secondary structures of SARS-CoV2 spike protein.

To study the effect of mutation on the conformation, stability and flexibility of the spike protein, the SARS-CoV-2 structure was downloaded from RCSB PDB. We used the available SARS-CoV-2 spike ectodomain structure (open state) (PDB ID: 6VYB). The 6VYB structure was uploaded on DynaMut software (University of Melbourne, Australia)[5] and changes in vibrational entropy and the atomic fluctuations and deformation energies due to mutation were determined. For atomic fluctuation and deformation energy calculations, calculations were performed by the software over first ten non-trivial modes of the molecule.

Results and discussion
The first set of five sequencing data from clinical isolates of SARS-CoV2 from the state of West Bengal, India was submitted on 27/2/2020 by the National Institute of Biomedical Genomics (NIBMG) in collaboration with ICMR-National Institute of Cholera and Enteric Diseases (ICMR-NICED). The sequences were submitted to the GISAID database.

We downloaded all the sequences from West Bengal (Table 1) and performed a nucleotide translation to obtain respective spike protein sequences. All these spike protein sequences were first aligned in CLUSTAL Omega to check for similarities or differences. We found that at position 723, all the isolates from West Bengal except EPI_ISL_430466 had 'T' (other isolates had 'I'). Also, at position 1124, two of the isolates (EPI_ISL_430468 and EPI_ISL_430464) showed mutation from 'G' to 'V'. We used spike protein sequence translated from one of these sequences (EPI_ISL_430468) as representative of SARS-CoV2 spike from West Bengal for our further comparison with other states.

Since, we had access to sequences for SARS-CoV2 only from three states in India (Kerala, Gujarat and now West Bengal), we compared all those sequences to detect possible changes (Figure 1). We considered the original Wuhan sequence as the wild type for comparison. We found six different amino acid positions that were mutated in these isolates overall. We had recently published the details with respect to the spike protein mutations in Kerala[3] and Gujarat isolates[4].

Table 1. Accession numbers of the sequences from West Bengal used for the analyses with specific regions of sample collection.

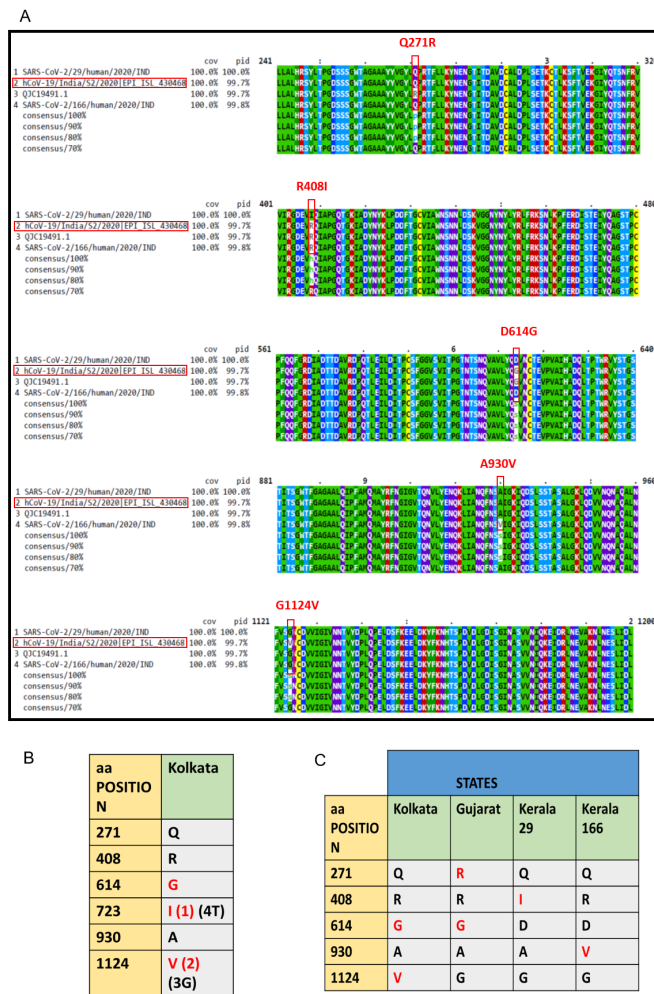| Virus name	Accession number	Region of collection within West Bengal	Gender, Age (years)
hCoV-19/India/S2/2020	EPI_ISL_430468	Kolkata	Male, 48
hCoV-19/India/S11/2020	EPI_ISL_430467	East Medinipur	Male, 25
hCoV-19/India/S5/2020	EPI_ISL_430465	Darjeeling	Female, 44
hCoV-19/India/S3/2020	EPI_ISL_430464	Kolkata	Male, 20
hCoV-19/India/S6/2020	EPI_ISL_430465	Tehatta	Male, 11

Figure 2: Mutation analysis of isolates from Kolkata, Gujarat and Kerala
A. Multiple sequence alignment of Spike protein sequence of Kolkata isolate with sequences obtained from other parts in India. Sites of mutation are showed in Red
B. Tabulation of amino acid mutations among isolates from Kolkata. Mutations are shown in red. Number/s in parenthesis show number of isolates that showed the amino acid type.
C. Tabulation of amino acid present at the points of mutation for isolates from different parts of India.

Figure 1. Mutation analysis of isolates from Kolkata, Gujarat and Kerala. (**A**) Multiple sequence alignment of Spike protein sequence of Kolkata isolate with sequences obtained from other parts of India. Sites of mutation are showed in Red. (**B**) Tabulation of amino acid mutations among isolates from kolkata. Mutations are shown in red. Number/s in parenthesis show number of osolates that showed the amino acid type. (**C**) Tabulation of amino acid present at the points of mutation for isolates from different parts of india.

Here, we report that in West Bengal isolates there were three mutations in the spike protein. One of these mutations was D614G in the S1 domain. This lies near the receptor bending domain at a downstream position. Another mutation, T723I lie further downstream in the S protein in the S2 domain. Mutation G1124V is the third mutation and lie in the S2 domain. While D614G was also found in the isolate from Gujarat but not in Kerala isolates, T723I and G1124V mutation were exclusively found in the isolates of West Bengal. None of the isolates from other parts of India had this mutation.

Although T723I was also exclusive for West Bengal, it appeared only in one isolate. Thus, we characterized mutation G1124V which appeared in two of the isolates. Both glycine (G) and valine (V) are non-polar amino acids with aliphatic R groups. Glycine has no side chain whereas valine is bulkier due to its side chain. A change from glycine to valine can thus potentially disrupt the local folding of the protein. For example, it was shown that G to V change in a P-glycoprotein changed its drug specificities[6].

Secondary structure prediction showed changes in and around the site of mutation (Figure 2). In the mutant spike there was a loss of turn structure from position 1124 and addition of four helices at positions 1123, 1124, 1125 and 1126. This change in secondary structure might lead to change in function of S2. S2 helps in fusion process of the spike protein and thus mutation in S2 may have altered receptor spike interactions and thus infectivity.

To correlate if changes in secondary structure are also reflected in the dynamics of the protein in its tertiary structure, we performed normal mode analyses and studied protein stability and flexibility. Change in vibrational entropy energy (ΔΔSVib ENCoM) between the wild type Wuhan isolate and the West Bengal isolate was -4.445 kcal.mol^{-1}.K^{-1} (Figure 3). The ΔΔG was 0.905 kcal/mol and the ΔΔG ENCoM was 4.756 kcal/mol. All these suggested a stabilizing mutation in this type of spike. The interatomic interactions have been shown in Figure 4.

Analyses of atomic fluctuations and deformation energies showed visible changes (Figure 5). Atomic fluctuations calculate the measure of absolute atomic motion whereas deformation energies detect the measure of flexibility of a protein. Figure 5 shows the visual representations of the atomic fluctuation and deformation energies where positions that could be visibly detected to be different have been marked.

This is the first report of mutations of such types in the isolates of the state of West Bengal and further sequencing followed by sequence analyses would help expanding the knowledge about variations of spike protein in human SARS-CoV. These variations might lead to virus diversification and eventual emergence of variants, antibody escape mutants, strains and serotypes. Also, mutations might help the virus to expand its tissue tropism and adjust with the host environment better. Therefore, elaborate studies on sequence variations should be done, which would in turn help in better therapeutic targeting.
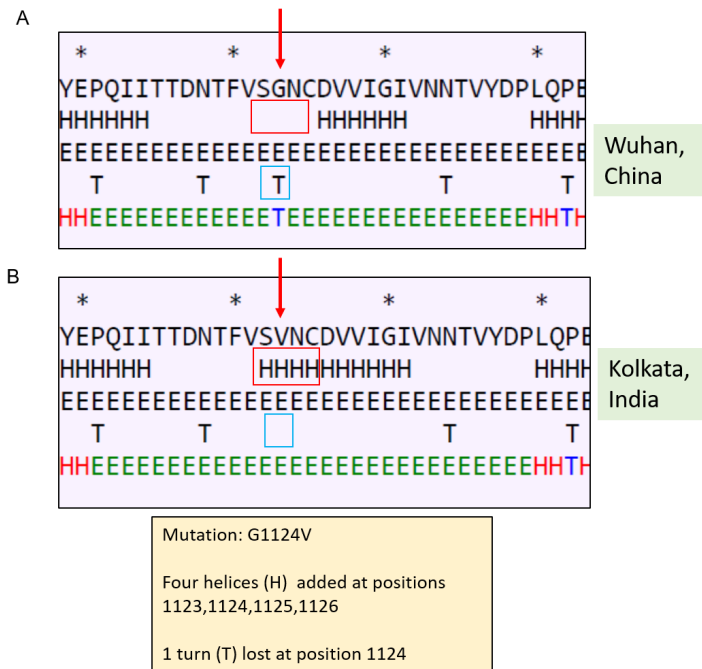

Figure 2. Effect of mutation at position 1124 on secondary structure of Spike protein. (**A**) Secondary structure of Spike protein of Wuhan isolate (the area around the residue 1124 has been shown); (**B**) secondary structure of Spike protein of Kolkata isolate showing effect of mutation on the secondary structure.
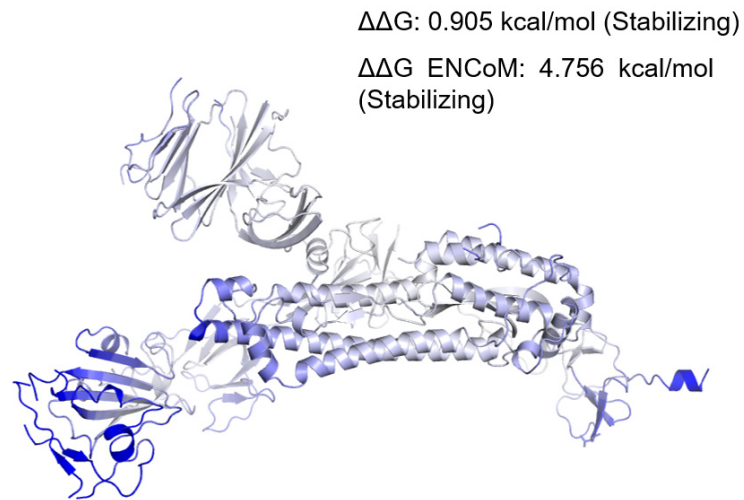
ΔΔG: 0.905 kcal/mol (Stabilizing)

ΔΔG ENCoM: 4.756 kcal/mol (Stabilizing)

Figure 3. Δ Vibrational Entropy Energy Between Wild-Type and Mutant. Δ Vibrational Entropy Energy Between Wild-Type and Mutant ΔΔSVib ENCoM: -4.445 kcal.mol-1.K-1. Amino acids were coloured as per the vibrational entropy change due to mutation. Blue represents rigidification of structure.
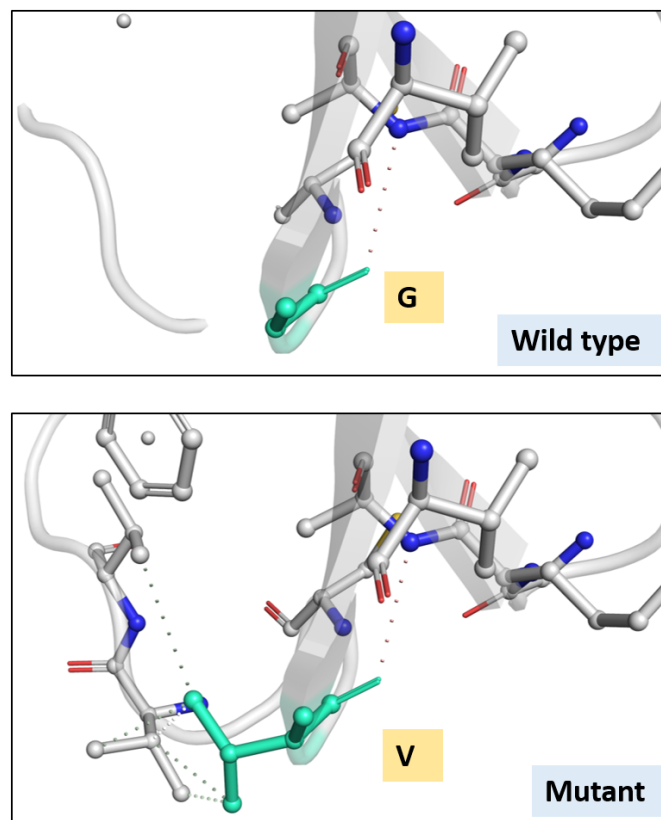

Figure 4. Interatomic interaction. Wild-type and mutant residues are coloured in light-green and are also represented as sticks alongside with the surrounding residues which are involved on any type of interactions.
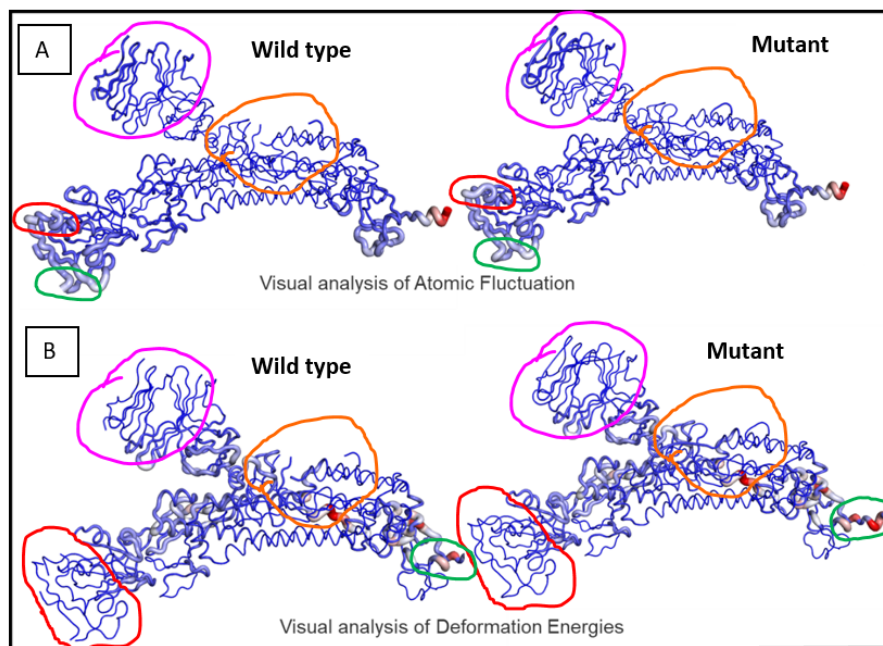
Figure 5. Visual analysis of fluctuations and deformation energies. Magnitude of (**A**) atomic fluctuation and (**B**) deformation has been shown using thin to thick lines coloured blue (low), white (moderate) and red (high).

Data availability

Underlying data

Accession numbers for sequences used in this study can be found in Table 1 and the test. GSAID is a free-to-use resource, but requires registration prior to accessing.

Extended data

Figshare: Figure F1000corrected.pdf, https://doi.org/10.6084/m9.figshare.12254033.v2[7].

This project contains the following extended data:

- Images raw images as created on DynaMut software:
  - Deformation_protein_158809623989_wild_type
  - Deformation_protein_G943V_158809623989_mutant_1124
  - Diff_A_G943V_158809623989_Kolkata_1124_mutation_6VYB23805-V1-1-fluctuation_protein_158809623989wild_type

  - Fluctuation_protein_G943V_158809623989_mutant_1124

  - Protein_cleanedcontacts_A_943_158809623989_wild_type

  - Protein_G943Vcontacts_A_943_158809623989_Mutant_Kolkata_1124

Acknowledgements

A previous version of this article is available on bioRxiv: https://biorxiv.org/cgi/content/short/2020.04.28.066985v1

References

1. Wu F, Zhao S, Yu B, *et al.*: **A new coronavirus associated with human respiratory disease in China.** *Nature.* 2020; **579**(7798): 265–269.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

2. Tai W, He L, Zhang X, *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.** *Cell Mol Immunol.* 2020; 17(6): 613–620.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

3. Saha P, Banerjee AK, Tripathi PP, *et al.*: **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.
   Publisher Full Text

4. Banerjee AK, Begum F, Thagriki D, *et al.*: **Novel Mutations in the S1 domain of COVID 19 Spike Protein of Isolate from Gujarat Origin, Western India.** *Preprints.* 2020.
   Publisher Full Text

5. Rodrigues CH, Pires DE, Ascher DB: **DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability.** *Nucleic Acids Res.* 2018; **46**(W1): W350–W355.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**

6. Choi KH, Chen CJ, Kriegler M, *et al.*: **An altered pattern of cross-resistance in multidrug-resistant human cells results from spontaneous mutations in the *mdr*1 (P-glycoprotein) gene.** *Cell.* 1988; **53**(4): 519–29.
   PubMed Abstract | **Publisher Full Text**

7. Ray U: **Figure F1000corrected.pdf.** *figshare.* Figure. 2020.
   http://www.doi.org/10.6084/m9.figshare.12254033.v2

Open Peer Review

Current Peer Review Status: ✓ ？ ✓

Version 1

Reviewer Report 18 June 2020

https://doi.org/10.5256/f1000research.26268.r63543

✓ **Nicolas Cuburu**
[1] Laboratory of Cellular Oncology, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA
[2] Laboratory of Cellular Oncology, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA

This article from Begum *et al.* entitled "Analyses of spike protein from first deposited sequences of SARS-CoV2 from West Bengal, India" reports a detailed analysis spike protein sequences deposited in GSIAD repository from clinical isolates of SARS-CoV2. The report relies on algorithms to predict the secondary structure of the SARS-CoV2 spike protein variant from Bengal and proposes that novel mutations in the SARS-CoV2 S2 domain could result in changes in the virus infectivity and escape mutants.

Because the structures were all based on the CFSSP algorithm it would be important to show how these structures relate to the known crystal structure of SARS-CoV2 spike from Wuhan. Also, the authors compared the sequences to Wuhan's original virus, but more mutations in S2 notably have been reported in clinical isolates from other parts of the world. An important question is whether the mutations from Bengal are new and unique or whether they are found also in the circulating virus in North America and Europe for instance.

Overall the observation and the hypothesis are interesting but at this point, it is purely speculative and these mutations could be passengers with no consequence on virus natural infection and transmission. These questions could only be solved by infection experiments *in vitro* or in animal models.

The authors have used the CFSSP method, but many other algorithms are available for secondary structure prediction. What was the rationale for choosing this specific algorithm? Would other algorithms yield similar predictions?

In the introduction, the authors mention that the CoV undergo rapid mutations but compared to other RNA viruses CoV tends to mutate more slowly due to the proofreading exonuclease.

Figure 1: it is not clear what consensus 100% 90% etc... represent.

Minor comments: few typos (osolate instead of isolate etc...)

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: I have limited expertise on computational approaches for protein structure prediction

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 02 June 2020

https://doi.org/10.5256/f1000research.26268.r63544

? Jalaj Gupta (ID)
1 Stem Cell Research Center, Department of Hematology, Georg-Speyer-Haus, Lucknow, Uttar Pradesh, India
2 Stem Cell Research Center, Department of Hematology, Georg-Speyer-Haus, Lucknow, Uttar Pradesh, India

Begum *et al.* submitted the manuscript 'Analyses of spike protein from first deposited sequences

of SARS-CoV2 from West Bengal, India' where COVID19 virus sequences retrieved from GISAID database have been studied for possible mutations. Authors report three mutations out of which two are unique (T723I and G1124V) and one shared (D614G) with other geographical locations. This is a timely study, unrevealing two unique mutations that can be useful for further studies on the behavior of COVID19. However, the authors did not clarify why they characterized only one mutant but not the other.

1. 1) The two mutations (T723I and G1124V) discussed in the manuscript should be similarly presented. Authors have described detailed analysis of G1124V mutation while T723I mutation is not discussed in detail. Therefore entropy calculations, interatomic interactions, deformation energy, and atomic fluctuation studies should also be done for mutation T723I.

2. What is the effect of T723I mutation on the secondary structure? This analysis should also be done and included in Figure 3. This will further improve the quality of the manuscript.

3. Authors claim that the two mutations (T723I and G1124V) are unique in the West Bengal region of India. Are they prevalent elsewhere in the World as well? Authors should expand their analysis and include more COVID19 isolates from different continents/countries.

Minor comments:
1. The accession number of Gujarat isolate in method and figure 1 do not match. Correction needed.

2. Accession number EPI_ISL_430465 in Table 1 is duplicated. One should be EPI_ISL_430466.

3. Parameters for calculation of vibrational entropy, atomic fluctuations, and deformation energies should be described in more detail.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Host-pathogen interaction, Biochemical characterization of mutant versions of proteins, Cancer biology using organoid and mouse models

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 26 May 2020

✔️ **Pawan Kumar Maurya**

[1] Department of Biochemistry, Central University of Haryana, Mahendragarh, Haryana, India
[2] Department of Biochemistry, Central University of Haryana, Mahendragarh, Haryana, India

The manuscript submitted reports mutations in newly sequenced SARS-CoV-2 isolates of West Bengal origin. The information presented here is a useful addition to the ongoing COVID 19 research. While the manuscript is well written and sufficiently documented, the authors should consider adding the following additional information.

1. Mutation 614 D to G has been in discussion in recent times extensively. The authors could add few lines in the discussion about the possible importance of this mutation as they also report the same in West Bengal's sequences.

2. DynaMut analysis for only G1124V has been done. Since mutation at the 723 position is also unique, a similar analysis for the same should also be included.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Clinical Biochemistry

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias

- You can publish traditional articles, null/negative results, case reports, data notes and more

- The peer review process is transparent and collaborative

- Your article is indexed in PubMed after passing peer review

- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research