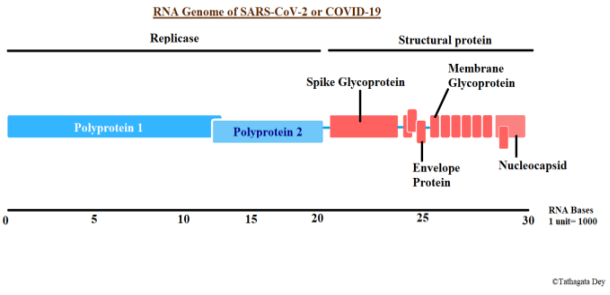
MOL2NET, International Conference Series on Multidisciplinary Sciences
USINEWS-04: US-IN-EU Worldwide Science Workshop Series, UMN, Duluth, USA, 2020

New Computational Analysis to Identify the Mutational Changes in SARS-CoV-2

Tathagata Dey [a,d], *Shreyans Chatterjee* [b,d], *Smarajit Manna* [c,d], *Ashesh Nandy* [d], *Subhash C Basak* [e,d]

[a] *Computer Science Department, Government College of Engineering and Textile Technology, Serampore-712201, India*
[b] *Microbiology Department, St. Xavier's College, Kolkata-700016, India*
[c] *Jagadis Bose National Science Talent Search, Kolkata 700107, India*
[d] *Centre for Interdisciplinary Research and Education, Kolkata-700068, India*
[e] *Department of Chemistry and Biochemistry, University of Minnesota, Duluth, MN, USA; sbasak@d.umn.edu*

Graphical Abstract



RNA Genome of SARS-CoV-2

Abstract.

The ongoing rapid spread of COVID-19 disease from its first detection in Wuhan, China in late 2019 was declared a pandemic by World Health Organization on 11th March, 2020. It is believed that to combat this deadly virus, now designated as SARS-CoV-2, designing and developing a proper vaccine is the best solution. For developing a sustainable vaccine against this virus, one should have a proper understanding of the mutational changes occurring constantly in its genome and also about the variations that may arise in different communities. Here, we report an algorithm to identify and characterize the mutational changes in the COVID-19 sequences isolated from different countries. The patterns in mutation along with the demographic analysis shown here can be very effective for community specific vaccine designing in the future.

Introduction

SARS-CoV-2 or COVID-19 is the newest member of *Coronaviridae* virus family, which has recently created a pandemic. The World Health Organization (WHO) declared this as a public health emergency of international concern (PHEIC) on 30th January, 2020 and a pandemic on 11th March. As on 28th April, 2020, according to the 99th Situation Report published by WHO, there has been a total of 2,954,222 confirmed cases of SARS-Covid-2 viral disease in 211 countries and territories across the world, amongst which 202,597 died [1].

The SARS-CoV-2 virus is the seventh type of coronavirus after 229E Alpha, HKU1 Beta, NL63 Alpha, OC43 Beta, MERS Beta and SARS Beta Coronavirus. The SARS-CoV-2 virus leading to COVID-19 disease is the fifth endemic coronavirus in human which had likely crossed the species barrier [2][3]. The symptoms of the disease include onset of fever (98%), cough (78%) and fatigue (44%) along with severe respiratory problems, lower cardiac outputs are also observed in some patients. It is a highly contagious disease and spreads via droplet infection.

Coronaviruses possess the largest genome compared to other RNA viruses. This gives them an opportunity for housing a variety of genes [4]. The novel Coronavirus or COVID-19 or SARS-CoV-2, is a spherical enveloped positive-sense single stranded RNA virus, falling under the genera Betacoronavirus. It has a genome length of around 29000 bp, coding for 7096 amino acids on an average. The genome consists of both structural and non-structural proteins.

Structural proteins are respectively: Surface. Glycoprotein (S), Membrane Glycoprotein (M), Envelope Protein (E) and Nucleocapsid (N). They occur in the 5’ to 3’ order as S, E, M, N. The non-structural proteins are: Polyproteins with nsp3, nsp6, nsp8, nsp10 and polyprotein1ab [5][6]. The arrangement of genes in the genome of Covid-19 are given in Figure 1.

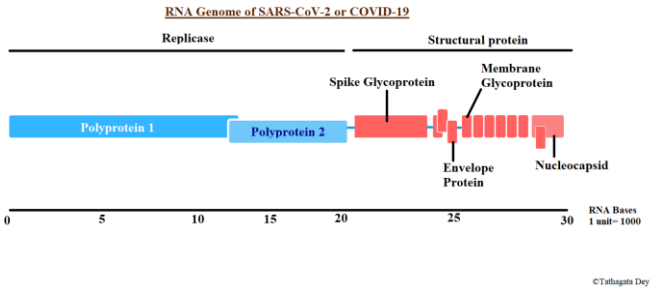


Fig. 1. Genome of SARS-CoV-2 (COVID-19)

As with all epidemics, there were no drugs or vaccines that could combat the CoVID-19 disease, although efforts are on a war footing to develop some therapeutics. Till now several vaccines are on clinical and preclinical trials. These candidate vaccines are either live-attenuated or inactivated vaccines. The mode of action of some candidate vaccines also relies on targeting a specific region in the viral genome. Here mutation becomes an important factor which determines the sustainability of the vaccine. A less mutating organism will make these vaccines more effective for a longer period of time.

In this article we try to figure out the mutational changes occurring in the genome of COVID-19 and their potentiality in generating different strains across the world during person-to-person transmission. Our 2D polar coordinate plot of amino acid sequences helps us with this estimation [7]. Proper knowledge of ongoing mutations in the CoVID-19 genome, or more specifically in the spike glycoprotein and non-structural proteins (nsp), may help avoid highly variable (or mutating) zones while designing a vaccine, so as to make it long lasting and may tell us about variability in infection rates over different countries. Further studies can also reveal the relation of non-structural protein (nsp) mutations with change in mode of attack or growth of virus inside the host cells. Along with these we discussed

the mutational changes with demographics that may help one to understand the effect of mutation on various communities.

Materials and Methods

ALGORITHM TO STUDY GENOME SEQUENCES

We used 2D Polar Co-ordinate Representation of Amino Acid Sequences to study the CoVID-19 sequences. In this method we assign angles to each amino acid depending upon their relative hydrophobicity indices. While reading the sequence, we read each amino acid one-by-one and move one unit in the respective assigned direction. We assume the graph to contain unit masses at each co-ordinate found in the traversal of the graph. The distance of the centre of mass of the graph from the origin is defined as the Quotient Radius or q_R value of the graph. This value is found to be a characteristic value of a sequence [7].

We took a total of 1674 sequences of SARS-CoV-2 from NCBI database [8], consisting of full genomes, spike glycoproteins and other proteins. We compared them through graph plotting, q_R value characterization and distribution plots. Sequences having identical q_R values indicate that they have identical amino acid sequences. If sequences are having very close q_R values, then it can be concluded that one or few amino acids in those sequences have altered compared to each other. Using this principle, we try to find out the mutations in SARS-CoV-2 genome.

Another study has been done here by plotting the Mutation Distribution Curve. Here we measure the deviation of q_R values of various proteins from the q_R values of the equivalent proteins as they were in the initial Wuhan sequence. That is, we calculate the difference in their q_R values and plot them. So, in the x-axis we plot the q_R difference (*dif* q_R) and in the y-axis we plot total number of such sequences.

$$x_i = (q_R)_i - (q_R)_{wuhan} = \Delta q_R$$

$$y_i = no.of\ sequences\ having\ \Delta q_R$$

So, a point in the graph (x_1, y_1) represents, that the q_R value of the sequence is $\left(q_{R_{Wuhan}} + \Delta q_R\right)$ and there are y_1 such sequences present having that q_R value. So, reading this graph we can interpret, how different a sequence is from the Wuhan sequence and how many such sequences are present. This will give a clearer idea about the significance of the mutations in that sequence.

Results and Discussion

- **FULL GENOME**

We considered 101 full genome sequences from the database and plotted them using 2D Polar Co-ordinate Representation of Amino Acid Sequences and determined their q_R Values. The Figure 2 below represents the 2D polar plot of initial Wuhan sequence (YP_009724389) from 3’ to 5’ end. The list of the q_R values is given in additional file.

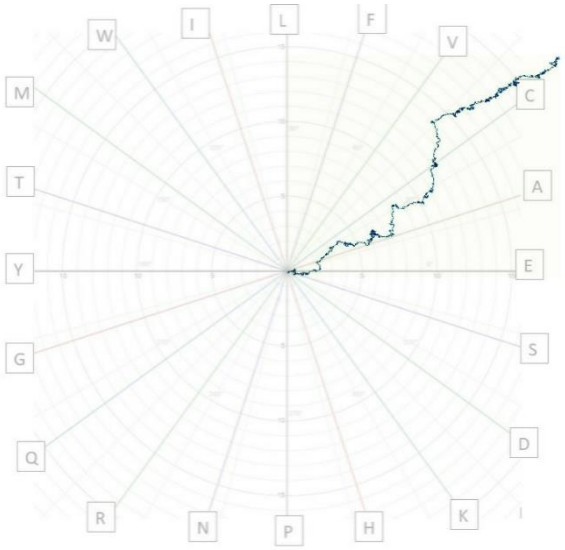


Fig. 2. 2D polar coordinate graph of Full Genome Wuhan SARS-CoV-2 Sequence

We observed that, the initial Wuhan sequence, collected in December, 2019, has a q_R value of 370.1371472. There is a total of 34 other sequences whose q_R value is equal to the initial Wuhan strain's q_R. The mutation distribution curve is shown in Figure 3 and significant clusters are marked red.

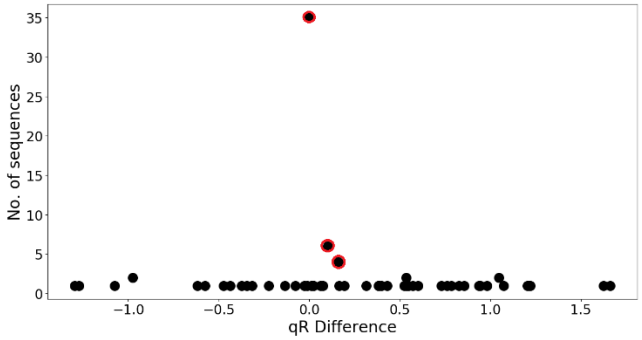


Fig. 3. Mutation Distribution curve of Full genome of SARS-CoV-2

As discussed earlier, here in this graph, the x-axis is labelled as $(q_R\ dif)$ and y-axis represents number of such sequences. So, the co-ordinate $(0,35)$ signifies that there are 35 such sequences whose q_R value difference with Wuhan q_R value is 0, i.e. they have identical sequences. Also, there are two other small clusters shown, which have slightly deviated on the +ve side of Wuhan sequence q_R. All the significant clusters are marked with red here.

By comparing the respective q_R values for each sequence we find deviations in sequences of Australia, Finland, USA from the initial Wuhan sequence. Among the rest, two other significant clusters have been noticed for two group of sequences. One of them having 6 sequences of same q_R as 370.238609 (sequences collected from - USA on 17th February and 22nd January; Brazil on 10th February and 2nd March; China on 17th January) and the other having 4 sequences with same q_R as 370.2969564 (sequences collected from China on 26th and 28th January. (All these sequences, their Accession IDs, sources, collection dates and q_R values are given in additional file. The sequences having same q_R are highlighted with same colour too.)

The q_R value of the first cluster (figure 3) is first seen in China, in a sequence collected on 17th January, 2020 and later on found in sequences collected from Brazil and USA in February, but we didn't find this value anywhere else after that. On the other hand, the second cluster is completely bound inside China and found in the sequences reported during 26th to 28th January, 2020. We expect both of these mutations to be less important as they are not reported elsewhere after the said time period. The rest q_R values cannot be clustered together which signifies small point mutations worldwide throughout the time span. These point mutation sequences are mostly from China while some are from USA and rest of the countries.

We observe that, the q_R differences are ranging from -1.3 to +1.7. Only two cases are found where ordinate is higher than usual. For the rest sequences, all the ordinates are either 1 or 2, which means that not many sequences are reported having that Δq_R value. This is the evidence of several point mutations. [9]

- **SPIKE GLYCOPROTEIN**

Our sample set contained 302 Spike glycoprotein sequences of SARS-CoV-2 or COVID-19. The mutation distribution curve of the whole set is shown below in Figure 4.

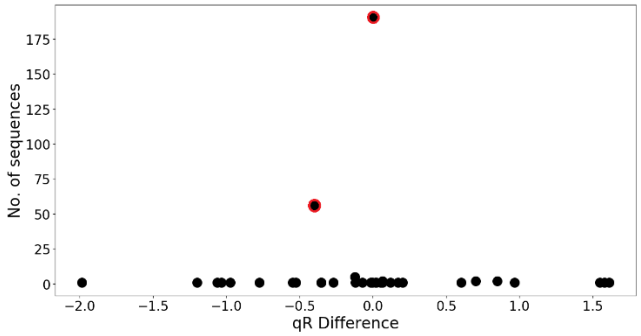


Fig. 4. Mutation Distribution curve of Spike Glycoprotein of SARS-CoV-2

This figure is also drawn following the same method of mutation distribution plot. The spike glycoprotein of initial Wuhan strain had a q_R value of 30.16105055. According to Figure 4, we see that the y value at x=0 is 190. So, out of 302, 190 sequences are of initial Wuhan sequence type.

Amongst rest only one set of significant mutation is seen having q_R value of 29.76636445, which is also marked with red in Figure 4. There is a total of 56 sequences present of the second type, making it a very significant mutation. The second type of q_R value (29.76636445) initially appeared on 11th March in a sequence reported from USA. Later on, the number of such sequences increased rapidly, making the count to be 56 as on 24th March,2020.

The sequences from the set of 56 are all from USA. 4 of them were collected from San Francisco while rest are from Washington state. This strain isn't reported anywhere else till 24th March, 2020, so it is very evident that the mutation took place in USA only. Still most of the USA sequences are of initial Wuhan type.

In case of point mutation, two sequences collected from Kerala, India, showed different q_R values. One of them closely matches with initial strain with q_R value 30.21560043 (Wuhan sequence q_R value: 30.16105055) while the other one has a large difference with q_R value 31.7742889. So, different strains may be present in India [10][11] right now. Most of the Chinese strains are having same value, implying that no significant mutation has taken place in spike glycoprotein of SARS-CoV-2 virus in China. All the sequences and q_R values are given in the additional file.

The Δq_R are ranging from -2 to +1.5. The coordinate y=56 is conspicuous (second type discussed here), while rest all are discrete point mutations.[12]

- **NUCLEOCAPSID**

In case of nucleocapsid, 311 sequences had been downloaded and is part of our database. The initial Wuhan sequence has q_R value of 46.77401554. 278 sequences are of the initial q_R value. So, mutation isn't very significant. The distribution graph is shown in Figure 5.

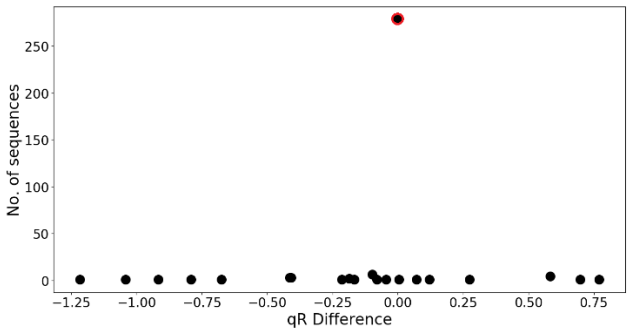


Fig. 5. Mutation Distribution curve of nucleocapsid of SARS-CoV-2

- **MEMBRANE GLYCOPROTEIN**

We collected 102 sequences of Membrane glycoprotein of SARS-CoV-2 and all of them had same q_R value as that of the initial Wuhan sequence, 31.53951. So, clearly no evidence of mutation is observed in this gene.

- **ENVELOPE PROTEN**

We took 102 sequences and 101 of them had value 16.37138. So clearly no mutations were found.

- **NSP 3a**

Our database contained 99 sequences of nsp3a. 88 of them had exactly same value of q_R , 28.88545, while 6 of them had value 29.00860427. The sources of the mutated strains are very scattered in this case. The first sequence was reported in USA in January, later on it was reported from Australia, Sweden, Brazil and Italy throughout February. But the available data of these countries in this time span being very sparse, no specific conclusions can be drawn.

- **NSP1a**

We collected 272 sequences and computed them. The distribution curve is shown in Figure 6.

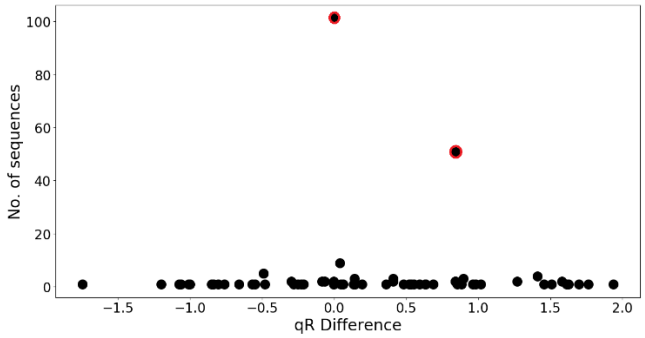


Fig. 6. Mutation Distribution curve of nsp1a of SARS-CoV-2

Out of them 101 were similar to the initial Wuhan sequence having q_R 239.3596307. Another cluster consisting of 51 sequences of q_R value 240.2023724 is also observed which indicates mutation in this gene. This mutated strain is first seen in Beijing, China on 18th January (Accession No. QIV15115). Later on, it was reported from South Korea, France and USA in March, 2020. Most of the sequences of mutated type collected from USA are either from Washington state or from Virginia. In China, large divergence in q_R values is found too. Significant clusters are marked red in the figure above. These results show that the nsp1a mutates much more than the other gene/protein sequences,

- **NSP 6**

We took 99 sequences and all had value 9.741840002. So clearly no mutations were found.

- **NSP 10**

We had 89 sequences and all had q_R value 5.260568882. So here also, no mutations have taken place.

- **NSP 8**

We took 99 sequences and all had value 11.88341642. So here also, no mutations took place.

- **NSP 7a**

We had 98 sequences which we evaluated through our methods. The mutation distribution curve is given below in Figure 7.

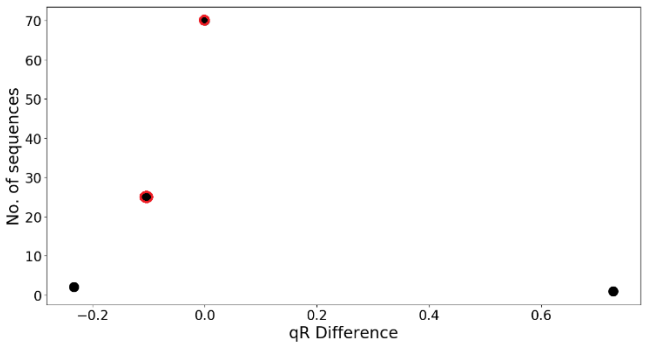


Fig. 7. Mutation Distribution curve of nsp7a of SARS-CoV-2

Two significant clusters are marked here. It is found that 70 of them had the initial Wuhan sequence q_R value of 9.244181845, implying they are the exactly same strains. Some of the sequences have mutated to gain a different q_R value of 9.139551519; as 25 such sequences are found; it can be stated as a stable mutation. The mutation is not too community dependent as we see from the data that the first appearance of this value is in a sequence collected from China around 10th January, 2020. In a sequence reported on 19th January, it is found for the first time in USA. Then, it is found successively in China and USA. Also, the Taiwan sequence has this value too. It is interesting that the two Indian sequences have different values. One of them has the mutated value and the other one has initial value. Even two such mutated strains are found in Spain too. So, observing all these phenomena we expect it to be an important mutation. The distribution graph is shown in Figure 7; It is quite different from the previous distribution graphs that no population of point mutation sequences is seen here, which raises the question whether that particular mutation is well directed.

- **COMPARISON WITH SARS**:

In another study we tried to track the direction of mutation in Covid-19. We collected 168 sequences of SARS spike glycoprotein and computed q_R values. It seemed that only 27 of them clustered together having q_R value of 37.18930424. The mutation distribution curve is shown in Figure 8. It is pertinent to say, in this case we didn't calculate the deviation of q_R from a certain sequence, rather plotted the distribution of the q_R value itself. So, in x-axis there is q_R value and in y-axis there is the number of sequences having that q_R value. Precisely, a point (x_1, y_1) represents here that there are y_1 sequences present with q_R value x_1.

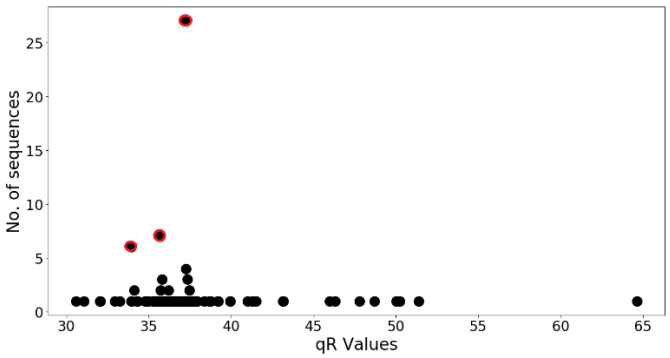


Fig. 8. Mutation distribution curve of Spike glycoprotein of SARS

In this Figure, we see most of the sequences to have a certain value of q_R (37.18930424,27), while few other small clusters can also be seen. Also, there are other random point mutations mostly having y=1 or y=2.

If we compare this with the mutational distribution graph of SARS-CoV-2 spike glycoprotein in Figure 4, we see a very similar pattern but having only 1 such significant cluster. So, if SARS-CoV-2 spike glycoprotein follows mutation pattern of SARS, then we expect to have few more such clusters in future.

- **GRAPHICAL OBSERVATION**:

To observe the mutational changes that occurred in other sequences compared to the initial Wuhan sequence, we plotted their whole protein sequences in 2D Polar Coordinate System. In these graphs the plot starts from the polyprotein1 beginning and continues through structural genes at the 5' end. We took the initial Wuhan, China Sequence (YP_009724389), USA 16th March 2020 sequence (QIK50426), India 6th March Sequence (QIA98582) and Italy 9th March Sequence (QIA98553) and plotted them.

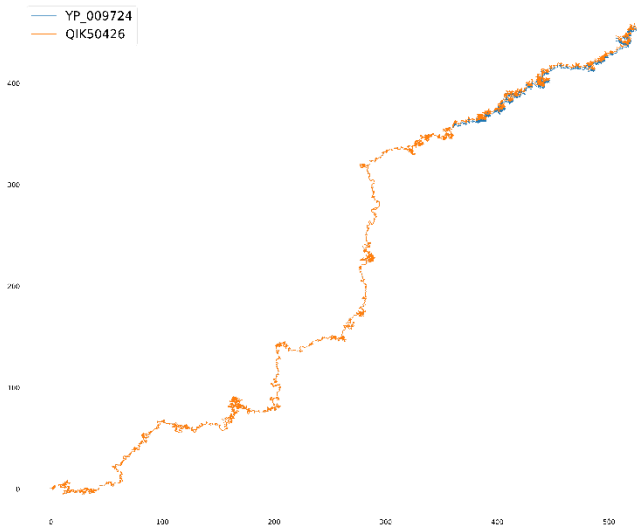


Fig. 9. Superimposition of China (YP_009724389) and USA Sequence (QIK50426) of SARS-CoV-2

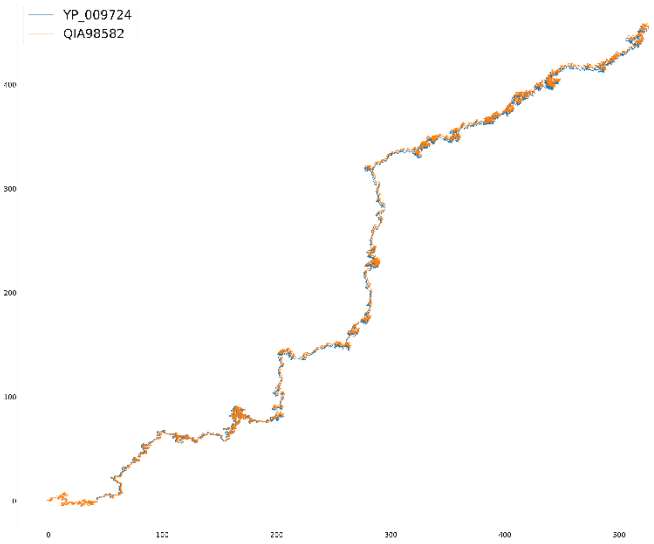


Fig. 10. Superimposition of China (YP_009724389) and India Sequence (QIA98582) of SARS-CoV-2

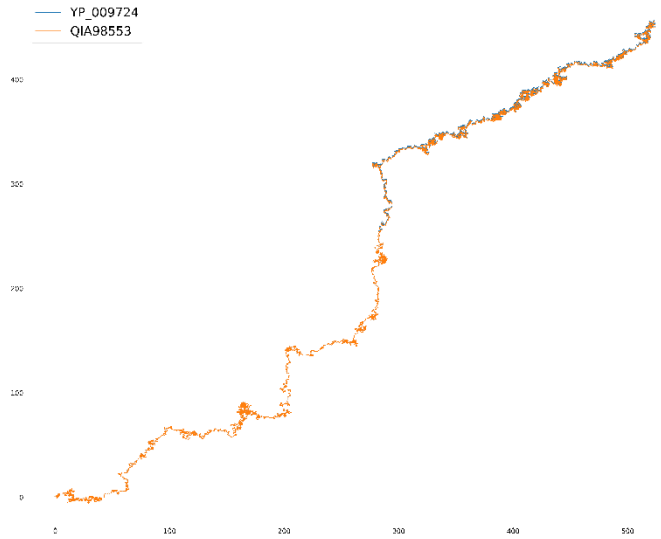


Fig. 11. Superimposition of China (YP_009724389) and Italy Sequence (QIA98553) of SARS-CoV-2

From the above three graphs in Figures 9-11 we can clearly see the position of mutations. In the China-USA comparison (Fig.9), the non-structural part seems identical and variations due to mutations begin to appear in the structural part. In the China-Italy comparison (Fig.10), the variations start a little earlier, while in the China-India comparative plot the entire Indian sequence seems to have mutational changes compared to the Wuhan sequence.

It is clearly coherent with the results obtained from q_R value analysis of spike glycoprotein. In other graphs, separation in non-structural protein also signifies our results for nsp7a and nsp1a.

SEQUENCE ACCESSION ID	COLLECTED FROM (COUNTRY)	FULL GENOME	SPIKE GLYCOPROTEIN	POLYPROTEIN 1	POLYPROTEIN 2
QIS60616	USA	99.97%	99.92%	99.97%	99.96%
QIO04366	China	99.6%	99.92%	99.47%	99.96%
QIJ96512	USA	99.98%	100%	99.97%	100%
QIH45032	China	99.98%	100%	99.97%	100%
QIG55993	Brazil	99.98%	100%	99.97%	100%
QID21047	USA	100%	100%	100%	100%
QIC53203	Sweden	99.97%	99.92%	99.95%	100%
QIA98582	India	99.95%	99.92%	99.95%	99.96%
QIA98553	Italy	100%	100%	100%	100%
QHZ00378	South Korea	99.97%	99.92%	99.97%	99.96%
QHW06058	USA	99.98%	99.92%	99.97%	100%
QHS34545	India	99.95%	99.92%	99.95%	99.96%
BCB97900	Japan	100%	100%	100%	100%

Table 1: Percentage Similarity with first COVID-19 Sequence collected from Wuhan (YP_009724389)

A summary of the main observations country wise in comparison to the original Wuhan sequence (YP_009724389) is given in Table 1. The table shows which parts of the sequences differ by how much in different countries thus providing a base for tailoring the eventual vaccine to the community where possible.

In the following graph, we plotted all the sequences of Full Genome SARS-CoV-2. It formed three unspecified clusters determining the type of mutation we found in full genome q_R analysis. We found three clusters with 35, 6 and 4 sequences. Here also we see the same pattern as for CoVID-19 being followed. All the clusters differed from each other in the beginnings of the graph, so here also it is signifying that the changes occurred in non-structural protein part.

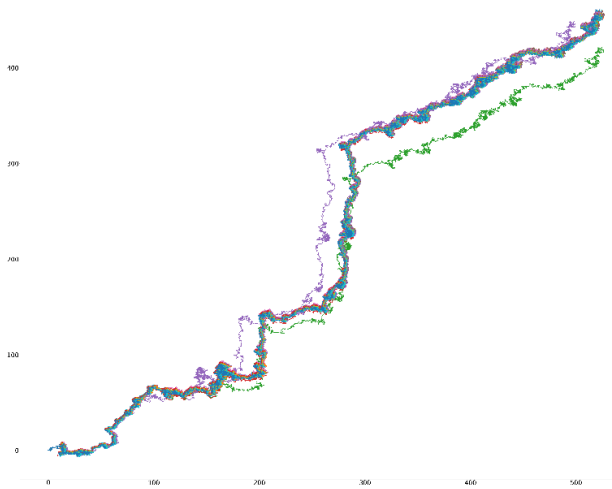


Fig. 12. Superimposition of All Full Genome of Graphs of SARS-COV-2 (101 sequences) in 2D polar coordinate plot.

- **BIOLOGICAL INFERENCES**

By studying the q_R values for the several structural and non-structural proteins we can draw these inferences:

A variety of mutations are observed in the sequences of SARS-CoV-2 isolated from USA, especially in spike glycoprotein.

Among the non-structural proteins, nsp1a and nsp7a showed mutations. nsp1a generated from orf1a assemble with orf1ab to facilitate viral transcription and replication that take place inside the host cell [13][14]. So, mutations in this region may have a significant role in pathogenicity and mode of attack.

We have not observed any significant point mutations in nsp7a region but large clustering of sequences is seen, which may indicate to a stable and well directed mutation.

For SARS-CoV-2, more point mutations are seen in spike glycoprotein. So, spike glycoprotein genes are more prone to mutations than other structural and non-structural genes of this virus. The entry of SARS-CoV-2 or COVID-19 into human cells depends on ACE2 and TMPRSS2 [15][16][17]. Some point mutations in spike glycoprotein may alter the active site for binding with host receptors changing its virulence [18]. Random point mutations may alter the 3-D structure of the folded protein, i.e., new regions of the virus spike glycoprotein may get exposed changing the potentiality of its attachment with human receptors.

SARS-CoV-2, a member of family *Coronaviridae* being an RNA virus with the largest RNA-virus genome, mutates at a moderately faster rate than DNA viruses but at a slower rate than other RNA viruses [19]. So classical vaccines which mainly consist of attenuated viral strains or inactivated viruses may become incapable of generating acquired immunity in the host body against that specific virus after a time due to mutations in the virus.

On the other hand, peptide vaccines can be a more viable tool by which we can combat this viral disease as peptide vaccines are designed to target the most conserved parts of the amino acid sequence (less mutated zones) of the spike glycoprotein. We have previously proposed a peptide vaccine for COVID-19 [20]. As mutations are occurring, one needs to make sure about choosing the conserved domains as mutations occur randomly and at the same time well directed in some cases [21]. Mostly USA sequences show such evidences, so community specific vaccines may be needed in the near future.

Conclusions

From our method of q_R analysis of SARS-CoV-2, we can easily identify the regions in its genome where mutation occurs and also quantify the extent of mutation. Our tool indicates that, higher is the Δq_R, more is the mutation. So, a complete analysis of these mutations occurring in COVID-19 can help us prepare effective and sustainable vaccines against it tailored to be community specific also where possible.

References

1. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/
2. Yong-Zhen-Zhang & Holmes, Edward C. (2020). A Genomic Perspective on the Origin and Emergence of SARS-CoV-2. 10.1016/j.cell.2020.03.035.

3. Mousavizadeh, Leila and Ghasemi, Sorayya. (2020). Genotype and phenotype of COVID-19: Their roles in pathogenesis. 10.1016/j.jmii.2020.03.022.
4. Patrick, C.Y. Woo & Yi Huang & Sussana, K.P. Lau & Kwok-Yung Yuen. (2010) Coronavirus Genomics and Bioinformatics Analysis. 10.3390/v2081803.
5. Angeletti, Silvia & Benvenuto, Domenico & Bianchi, Martina & Giovanetti, Marta & Pascarella, Stefano & Ciccozzi, Massimo. (2020). COVID-2019: The role of the nsp2 and nsp3 in its pathogenesis. Journal of Medical Virology. 10.1002/jmv.25719.
6. Dawood, Ali. (2020). Mutated COVID-19, May Foretells Mankind in a Great Risk in the Future. New Microbes and New Infections. 100673. 10.1016/j.nmni.2020.100673.
7. Dey, Tathagata & Biswas, Shubhamoy & Chatterjee, Shreyans & Manna, Smarajit & Nandy, Ashesh & Basak, Subhash C. (2020). 2D Polar Co-ordinate Representation of Amino Acid Sequences With some applications to Ebola virus, SARS and SARS-CoV-2 (COVID-19). 10.3390/mol2net-06-06790.
8. https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=Wuhan%20seafood%20market%20pneumonia%20virus,%20taxid:2697049
9. Joob, Beuy & Wiwanitkit, Viroj. (2020). Variation of 2019 novel coronavirus complete genomes recorded in the 1 st month of outbreak: Implication for mutation. Journal of Research in Medical Sciences. 25. 33. 10.4103/jrms.JRMS_147_20.
10. Saha, Priyanka & Banerjee, Arup & Tripathi, Prem & Srivastava, Amit & Ray, Upasana. (2020). 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. 10.1101/2020.04.07.029132.
11. Joshi, Aditi & Paul, Sushmita. (2020). Phylogenetic Analysis of the Novel Coronavirus Reveals Important Variants in Indian Strains. 10.1101/2020.04.14.041301.
12. Banerjee, Arup & Begum, Feroza & Ray, Upasana. (2020). Mutation Hot Spots in Spike Protein of COVID-19. 10.20944/preprints202004.0281.v1. Issa, Elio & Merhi, Georgi & Panossian, Balig & Salloum, Tamara & Tokajian, Sima. (2020). SARS-CoV-2 and ORF3a: Non-Synonymous Mutations and Polyproline Regions. 10.1101/2020.03.27.012013.
13. Gao, Yan & Liming, Yan & Huang, Yucen & Liu, Fengjiang & Zhao, Yao & Cao, Lin & Wang, Tao & Sun, Qianqian & Ming, Zhenhua & Zhang, Lianqi & Ge, Ji & Zheng, Litao & Zhang, Ying & Wang, Haofeng & Zhu, Yan & Zhu, Chen & Hu, Tianyu & Hua, Tian & Zhang, Bing & Rao, Zihe. (2020). Structure of the RNA-dependent RNA polymerase from COVID-19 virus. Science. eabb7498. 10.1126/science.abb7498.
14. Pachetti, Maria & Marini, Bruna & Benedetti, Francesca & Giudici, Fabiola & Mauro, Elisabetta & Storici, Paola & Masciovecchio, Claudio & Angeletti, Silvia & Ciccozzi, Massimo & Gallo, Robert & Zella, Davide & Ippodrino, Rudy. (2020). Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. 10.21203/rs.3.rs-20304/v1.
15. Hoffmann, Markus & Kleine-Weber, Hannah & Schroeder, Simon & Muller, Marcel A &Drosten, Christian & Pohlmann, Stefan. (2020). SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor 10.1016/j.cell.2020.02.052.
16. Jia, Yong & Shen, Gangxu & Zhang, Yujuan & Huang, Keng-Shiang & Ho, Hsing-Ying & Hor, Wei-Shio & Yang, Chih-Hui & Li, Chengdao & Wang, Wei-Lung. (2020). Analysis of the mutation dynamics of SARS-CoV-2 reveals the spread history and emergence of 1 RBD mutant with lower ACE2 binding affinity. 10.1101/2020.04.09.034942.
17. Liu, Haiguang & Severin Lupala, Cecylia & Li, Xuanxuan & Lei, Jian & Chen, Hong & Qi, Jianxun & Su, Xiaodong. (2020). Computational simulations reveal the binding dynamics between human ACE2 and the receptor binding domain of SARS-CoV-2 spike protein. 10.1101/2020.03.24.005561.
18. Shah, Masaud & Rather, Bilal & Choi, Sangdun & Woo, Hyun Goo. (2020). Sequence variation of SARS-CoV-2 spike protein may facilitate stronger interaction with ACE2 promoting high infectivity. 10.21203/rs.3.rs-16932/v1.
19. Sanjuán, Rafael & Domingo-Calap, Pilar (2020) Mechanisms of Viral Mutation. 10.1007/s00018-016-2299-6.
20. Biswas, Shubhamoy & Chatterjee, Shreyans & Dey, Tathagata & Dey, Sumanta & Manna, Smarajit & Nandy, Ashesh & Basak, Shubhash C. (2020). In Silico Approach for Peptide Vaccine Design for CoVID 19. 10.3390/mol2net-06-06787.
21. Matyasek, Roman & Kovarik, Ales. (2020). Mutation patterns of human SARS-COV-2 and bat RaTG13 coronaviruses genomes are strongly biased towards C>U indicating rapid evolution in their hosts. 10.21203/rs.3.rs-21377/v1.

Additional Files

q_R value of the sequences we computed, by 2D polar plot are given in additional file. Out of 1674 sample sequences, we added the computational data with significant mutation evidences such as Full Genome, Spike Glycoprotein, Nsp1a, Nsp7a.