Available online at www.medicinescience.org

ORIGINAL ARTICLE

Medicine Science 2022;11(1):171-5

Medicine Science International Medical Journal

A bioinformatic analysis of the spike glycoprotein & evolution of COVID-19

Atiksh Chandra[1], Sathees B Chandra[2]

[1]*Cypress Bay High School, Weston, FL, USA*
[2]*Biomedical Sciences Program, College of Nursing and Health Sciences, Barry University, Miami Shores, USA*

Received 03 July 2021; Accepted 16 August 2021
Available online 17.01.2022 with doi: 10.5455/medscience.2021.07.223

Abstract

The Severe Acute Respiratory Syndrome 2 (COVID-19/SARS-CoV-2) has become the pandemic of the century due to its drastically infectious nature. SARS & MERS are the most notable of past coronaviruses infecting merely thousands compared to COVID-19's gigantic magnitude. COVID-19's global spread has been attributed to its high asymptotic transmission and explosive infectious nature, mainly due to mutational changes in the spike glycoprotein. The purpose of this research is to comprehend & evaluate the divergent evolution of the spike glycoprotein in COVID-19, and other coronaviruses, at the molecular level via bioinformatic analysis. A phylogenetic tree was constructed using spike glycoprotein sequences from viral genomes using MEGA X program. Nucleotide composition analysis and genome organization study were carried out. Dot plot comparisons were performed using EMBOSS Dot Matcher program. Phylogenetic analysis produced four distinct clades for each coronavirus genera with a common ancestral origin sometime in recent history. More importantly, COVID-19 and SARS formed their own subclade suggesting that evolution of sequence has taken place in the spike glycoprotein over the period time. Genome organization and nucleotide composition provided further evidence of mutational changes in the spike glycoprotein. The results from this study demonstrated the divergent evolution of coronaviruses. Mutational changes in the spike glycoprotein have resulted in more virulent forms of COVID-19.

Keywords: COVID-19, spike glycoprotein, coronavirus, divergent evolution, mutation, phylogenetic analysis

Introduction

The novel coronavirus Disease 2019 (COVID-19) pandemic, caused by The Severe Acute Respiratory Syndrome 2 (SARS-CoV-2), is a catastrophe impacting millions of people medically, socially, and economically worldwide. As of mid-August 2021, over 207 million cases and more than 4.36 million deaths globally have been associated with COVID-19 [1]. In the United States alone it infected more than 36 million and killed at least 621 thousand in a just over a year [2]. Coronaviruses in general are not unknown to humans. SARS (SARS-CoV) in 2002 in China infected 8000 people and caused 800 deaths [3]. MERS (MERS-CoV) in 2012 in the Middle East infected 2500 people and killed 800 of them [4]. We were able to successfully contain the spread of SARS & MERS, before they could spread to the rest of the world, because of their low infectious nature [5-7]. However, COVID-19 has spread worldwide within a short period of time due of its high asymptotic transmission and potent infectious nature, mainly due to mutational changes in the spike protein [8-10].

Most coronaviruses have four structural proteins identified as spike (S), envelope (E), membrane (M), and nucleocapsid (N) [11,12]. The virus's entry into humans is mediated specifically by spike protein through a multi-step process [13,14]. They first need to recognize host cell-surface receptor for viral attachment. In the next step, fuse viral and host membranes for entry [13]. Therefore, Receptor-Binding Domains (RBD) present in S protein play important role in viral transmission [15]. In addition, a recent D614G mutation identified in spike protein, from A to G base change, in one of the Wuhan strains is confirmed to be a rapid transmission form of COVID-19 [16,17]. Furthermore, various mutations in spike protein have recently been reported in isolates from several geographical regions of China and other parts of the world [18,19]. Therefore, the sequence analysis of the genomic region encoding the viral spike glycoprotein is essential

***Corresponding Author:** Sathees B Chandra, Biomedical Sciences Program, College of Nursing and Health Sciences, Barry University, Miami Shores, USA
E-mail: schandra@barry.edu

in understanding explosively contagious nature of this virus and to possibly design effective long lasting viral vaccines in the future.

The genomic analysis in our study will answer two fundamental questions. Firstly, is COVID-19 the product of the divergent evolution of coronaviruses? Secondly, are the mutational changes in spike glycoprotein the cause for evolution of more virulent form of COVID-19? Therefore, the purpose of this research is to comprehend & evaluate the divergent evolution of the spike glycoprotein in COVID-19, and other coronaviruses, at the molecular level via bioinformatic analysis.

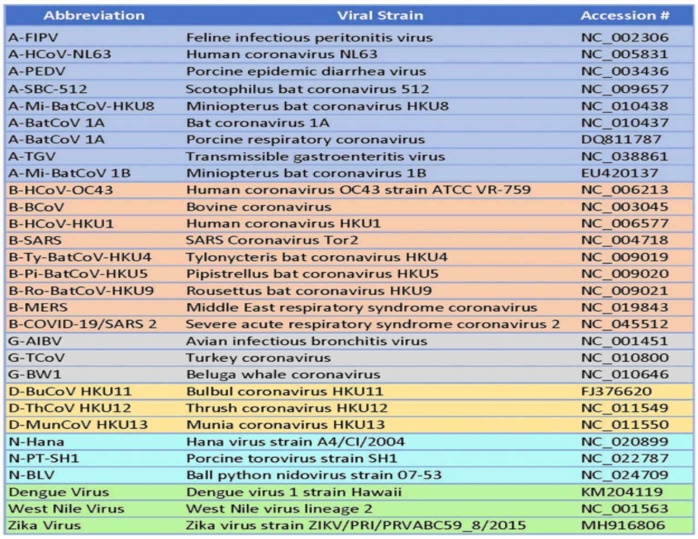
Materials and Methods

Sequence Selection

All coronaviruses fall under the Coronaviridae family within the Nidovirales order.

Coronaviruses can be further divided into 4 genera: Alphacoronaviruses, Betacoronaviruses, Gammacoronaviruses and Deltacoronaviruses. Genera are primary differentiated by their main reservoir hosts. Alpha- and Betacoronaviruses are known to infect mammalian species, Gammacoronaviruses are known to infect avian species, and Deltacoronaviruses are known to infect both. [20,21] In our analysis, 30 viral genome sequences were selected and evaluated (Table 1). 24 sequences were chosen from Coronaviridae family, 3 sequences were taken from Nidovirales Order, and other 3 represented commonly well-known human viruses. To ensure diversity in the Coronaviridae family, sequences were selected in accordance with genre: 9 Alphacoronaviruses, 9 Betacoronaviruses, 3 Deltacoronaviruses, and 3 Gammacoronaviruses. These sequences were chosen based on availability of full genome sequence data, relative genome length and current literature review from a range of mammalian and non-mammalian hosts for these viruses. All sequences were obtained from National Center for Biotechnology Information (NCBI) viral genome database (URL: http://www.ncbi.nlm.nih.gov).

Table 1. A complete list with names of the viral strains used in the analysis. Naming abbreviations for different groups: Alphacoronaviruses (Dark Blue), Betacoronaviruses (Orange), Gammacoronaviruses (Grey), Deltacoronaviruses (Yellow), Nidovirales strains (Aqua), and Commonly known viruses (Green). NCBI accession number for each strain is indicated.



Abbreviation	Viral Strain	Accession #
A-FIPV	Feline infectious peritonitis virus	NC_002306
A-HCoV-NL63	Human coronavirus NL63	NC_005831
A-PEDV	Porcine epidemic diarrhea virus	NC_003436
A-SBC-512	Scotophilus bat coronavirus 512	NC_009657
A-Mi-BatCoV-HKU8	Miniopterus bat coronavirus HKU8	NC_010438
A-BatCoV 1A	Bat coronavirus 1A	NC_010437
A-BatCoV 1A	Porcine respiratory coronavirus	DQ811787
A-TGV	Transmissible gastroenteritis virus	NC_038861
A-Mi-BatCoV 1B	Miniopterus bat coronavirus 1B	EU420137
B-HCoV-OC43	Human coronavirus OC43 strain ATCC VR-759	NC_006213
B-BCoV	Bovine coronavirus	NC_003045
B-HCoV-HKU1	Human coronavirus HKU1	NC_006577
B-SARS	SARS Coronavirus Tor2	NC_004718
B-Ty-BatCoV-HKU4	Tylonycteris bat coronavirus HKU4	NC_009019
B-Pi-BatCoV-HKU5	Pipistrellus bat coronavirus HKU5	NC_009020
B-Ro-BatCoV-HKU9	Rousettus bat coronavirus HKU9	NC_009021
B-MERS	Middle East respiratory syndrome coronavirus	NC_019843
B-COVID-19/SARS 2	Severe acute respiratory syndrome coronavirus 2	NC_045512
G-AIBV	Avian infectious bronchitis virus	NC_001451
G-TCoV	Turkey coronavirus	NC_010800
G-BW1	Beluga whale coronavirus	NC_010646
D-BuCoV HKU11	Bulbul coronavirus HKU11	FJ376620
D-ThCoV HKU12	Thrush coronavirus HKU12	NC_011549
D-MunCoV HKU13	Munia coronavirus HKU13	NC_011550
N-Hana	Hana virus strain A4/CI/2004	NC_020899
N-PT-SH1	Porcine torovirus strain SH1	NC_022787
N-BLV	Ball python nidovirus strain 07-53	NC_024709
Dengue Virus	Dengue virus 1 strain Hawaii	KM204119
West Nile Virus	West Nile virus lineage 2	NC_001563
Zika Virus	Zika virus strain ZIKV/PRI/PRVABC59_8/2015	MH916806

Nucleotide Composition Analysis: Nucleotide composition data was summarized using Mega X program: Molecular Evolutionary Genetics Analysis (Mega Version X). All sequences were first imported into the "alignment explorer" platform from which nucleotide frequency data was produced.

Genome Organization: A genome organization of sequences was created to visually characterize structural similarities between members of Coronaviridae genres and other commonly known viruses. The information available in NCBI viral genome database was applied to map the localizations and size of the 4 main structural proteins: Spike(S), Membrane(M), Envelope(E), and Nucleocapsid(N). To discover similarity patterns in genome organization, 12 sequences (3 Betacoronaviruses, 2 Alphacoronaviruses, 2 Gammacoronaviruses, and 2 Deltacoronaviruses and 3 non-coronavirus strains) were utilized in this analysis. Graphics and imagery were developed using Microsoft PowerPoint Graphical Tools.

Spike Protein Dot Plot Comparisons: Dot plot comparisons using the EMBOSS DotMatcher program were constructed to analyze the spike glycoprotein similarity in Betacoronavirus and non-coronavirus strains relative to B-COVID-19 [22]. Similarity of sequences is assessed through the development of a straight diagonal line. A well-formed straight diagonal line represents high degree similarities; a scattered diagonal line represents low degree similarity. The sequences were assessed in the dot matcher program using a specified substitution matrix and a scoring matrix with window size of 10 and a threshold of 23 [23]. 6 sequences (3 from the Betacoronavirus genre and 3 from the Nidovirales Order) were compared with the B-COVID-19 sequence in this analysis.

Phylogenetic Analysis

For the construction of the phylogenetic tree, only S protein sequences from Coronaviridae family were utilized. ClustalW algorithm with standard settings was used in the alignment of 24 spike glycoprotein sequences of coronavirus strains. The phylogenetic analysis was carried out by using the Maximum Likelihood method and JTT matrix-based model in MEGA X Tree Constructor [24,25]. Neighbor-Join and BioNJ algorithms were used in JTT matrix-based model. A bootstrap procedure was implemented with 100 repetitions to ensure position of the clades and statistical accuracy.

Results

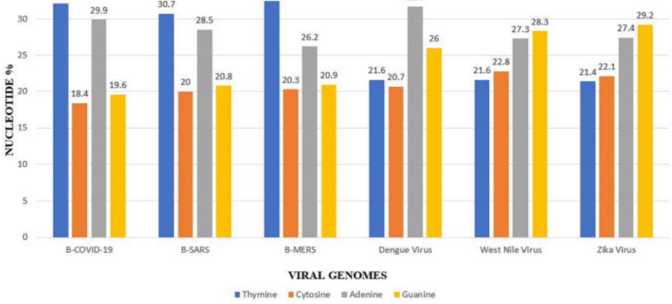


Figure 1. Multiple bar graph detailing nucleotide composition of Thymine (blue), Cytosine (Orange), Adenine (Gray), and Guanine (Yellow) for B-COVID-19, B-SARS, B-MERS, Dengue Virus, West Nile Virus, and Zika Virus.

The nucleotide composition between coronavirus and non-coronavirus strains is represented in Figure 1. The percentage of thymine is significantly higher in the coronavirus strains (B-COVID-19, B-SARS, and B-MERS) compared to the non-coronavirus strains (Dengue, West Nile and Zika). The reverse is apparent with guanine composition. Non-coronaviruses evidently have a greater average guanine composition versus coronavirus strains. However, cytosine and adenine percent composition did not show any significant variation between coronavirus and non-coronavirus strains.

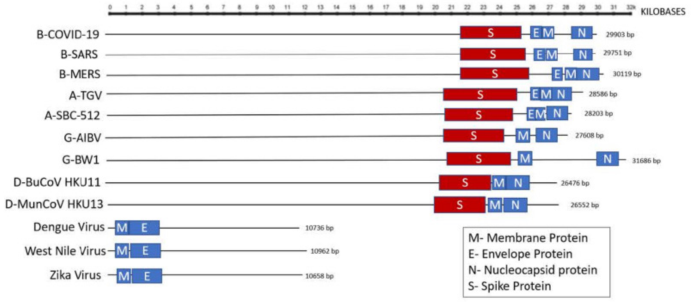


Figure 2. Genome Organization of Viral Sequences of B-COVID-19, B-SARS, B-MERS, A-TGV, A-SBC-512, G-AIBV, G-BW1, D-BuCoV HKU11, D-MunCoV HKU13, Dengue Virus, West Nile Virus and Zika Virus. Each line represents the length of each sequence along with the localizations and size of the 4 main structural proteins: Spike(S), Membrane(M), Envelope(E), and Nucleocapsid(N). The numerical length of sequence is provided at the end of the line. Data was acquired from NCBI viral genome database

Genome organizations of all 3 Betacoronaviruses indicate relatively similar sequence lengths, protein size, and localizations of all 4 structural proteins (Fig. 2). However, Alpha-, Gamma-, and Deltacoronaviruses exhibit less similarities in genome organization. The length of sequences progressively reduced from Beta- to Alpha- to Gamma- to Delta- to non-coronavirus strain. The localization of S protein appears to have a slight leftward shift. The E protein is absent in Gammacoronaviruses and Deltacoronaviruses. Dengue Virus, West Nile Virus, Zika Virus have much shorter sequence length. More significantly, this group lacked both the S and N protein.

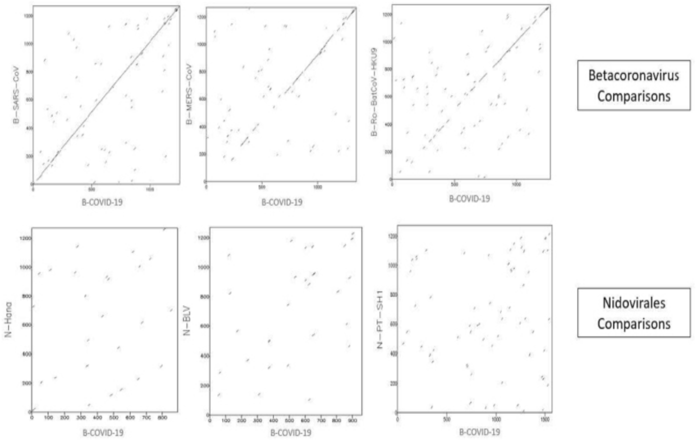


Figure 3. Spike protein sequence dot plot comparisons in EMBOSS Dot Matcher program was done with a window size of 10 and a threshold of 23. B-COVID-19 was compared to B-SARS, B-MERS, B-Ro-BatCoV-HKU9, N-Hana, N-BLV, N-PT-SH1

The degree of similarity between the spike protein within Betacoronavirus genre and other Nidovirales strains are expressed in Dot-plot comparison (Fig. 3). The first graph, comparing B-COVID-19 to B-SARS, has an almost perfect diagonal line. The next Betacoronavirus comparisons with B-MERS and B-Ro-BatCoV-HKU9 show a similar well-formed diagonal line. Naturally, this high level of similarity is expected since we are comparing B-COVID-19 with other closely related Betacoronaviruses. However, comparison with sequences in the Nidovirales order but not Coronaviridae family (N-Hana, N-BLV, and N-PT-SH), no diagonal lines are formed in any of these graphs indicating little to no similarities with the S protein of B-COVID-19.

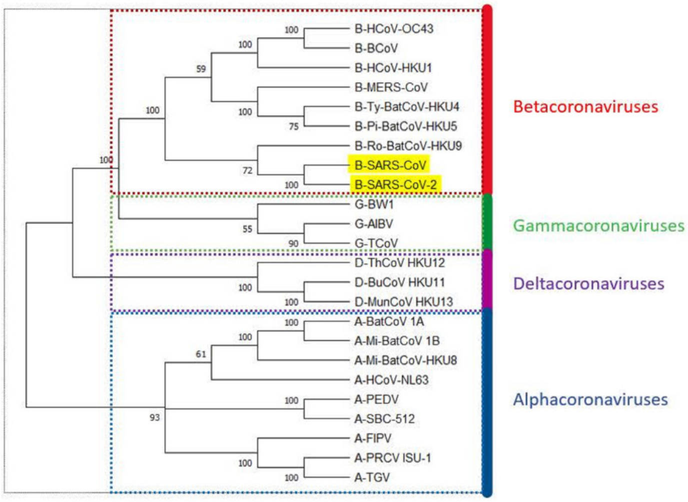


Figure 4. Phylogenetic Tree Analysis of Coronaviridae Spike Proteins. The evolutionary history was inferred by using the Maximum Likelihood method and JTT matrix-based model in MEGA X program

The phylogenetic tree analysis produced four distinct clades for each coronavirus genera (Alpha, Beta, Gamma & Delta) with common ancestral origin sometime in history (Fig. 4). This has been supported by high bootstrap values for individual viral strains and each clade. More importantly, B-COVID-19 and B-SARS formed their own subclade within the Betacoronavirus clade suggesting that evolution of sequence has taken place in the spike protein over the period time.

Discussion

We explored genome sequences of COVID-19 and other related coronaviruses by analyzing their nucleotide composition, genome organizational structure, dot-plot comparisons and by the construction of a phylogenetic tree. The evolutionary tree, produced by utilizing the spike glycoprotein sequences from our data, created 4 distinct clades (Alpha-, Beta-, Gamma- & Deltacoronaviruses) of statistically high bootstrap values with a common ancestral origin sometime in the history. Therefore, the tree supported our first hypothesis and provided evidence for divergent evolution of COVID-19 and other coronaviruses within Coronaviridae family. More importantly, COVID-19 and SARS formed their own subclade with in Betacoronavirus clade. This results further supported our second hypothesis that, mutational change in the spike protein is the cause for evolution of COVID-19, likely the virulent form strain. Our results agree with Korber et al. (2020) recent findings that variant COVID-19 strains are now spreading due to D614G mutation in spike mutation. D614G is a non-synonymous mutation where aspartic acid was replaced with glycine at position 614 of COVID-19's spike protein.

This study has concluded that this mutational change is under positive selection and has led to evolution of more virulent form of COVID-19 within a relatively short period time. Another study in Spain by Diez-Fuertes et al. (2021) also reported similar findings and summarized that an evolutionary advantage of this substitution in spike protein for COVID-19. However, a recent study by Volz et al. (2021) partially contradicted the potency of this mutational change. They concluded that spike 614G variant was more infectious but did not increase the mortality. Additionally, ongoing research work throughout the world soon will help us sort out the outcome of mutational changes in spike protein.

The nucleotide composition analysis in our study provided preliminary evidence for the evolution of among COVID-19 and other related coronaviruses. We found higher level of thymine (T) and lower level of guanine (G) in coronavirus strains. The reverse was true for non-coronavirus strains. The coronaviruses have an average thymine percent composition of 31.8% compared to 21.5% in non-coronaviruses group. The coronaviruses have an average guanine percent composition of 20.4% compared to 27.8 % in non-coronaviruses group. Pathan et al. (2020) work on prediction of COVID-19 by mutation rate analysis found that about 0.1% increment in mutation rate for mutating of nucleotides from T to G and C. This model can be handy in estimating the rate of mutation in COVID-19 variants and might be useful in managing vaccine efficacy.

Genome organization of various viral strains in the study further provided evidence of divergent evolution of sequences and viral organisms. COVID-19, SARS & MERS sequences exhibited perfect genome length and nearly similar localization of spike and other three structural proteins. Alpha-, Beta- & Deltacoronaviruses showed reduced genome size along with slightly different localization of spike protein. The non-coronavirus genomes (Dengue, West Nile & Zika virus) had drastically reduced genome length and, more importantly, lacked both spike protein and the nucleocapsid protein indicating further evidence of evolutionary progression in coronaviruses. It implies that the coronavirus spike proteins have almost completely mutated from their closest taxonomic group a long time ago. The dot-plot comparison provided further evidence of mutational change in the sequence of spike protein. Our analysis of spike protein sequences from COVID-19 vs other closely related Betacoronaviruses produced near perfect diagonal lines exemplifying the high degree of similarity. When we compared COVID-19 vs other spike protein sequences from the Nidovirales order, we found no diagonal line, indicating zero similarity, as expected if the divergent evolution was occurring over a long period of time.

Conclusion

Ultimately, we believe that our work has provided sufficient evidence for mutational changes in spike glycoprotein and its role in divergent evolution of COVID-19 and other coronaviruses. But this study is far from complete and require us to study genome sequences of new variants to better understand the evolution of these viruses. Many new variants of COVID-19 are being reported worldwide in the last few months alone, as we begin to inoculate against the original strain of this virus using both mRNA based as well as traditional vaccines. It's no surprise to see the emergence of new variants and, in fact, it is expected based on what we already knew about the coronavirus's ability to mutate. There is uncertainty about duration of protection and efficacy of vaccines against new variants of this virus. Further research involving the spike protein from many COVID-19 new variants should help us in answering these questions and pave the way for subsequent development of improved mRNA-based vaccines in the immediate future.

Conflict of interests

The authors declare that they have no competing interests.

Financial Disclosure

All authors declare no financial support.

Ethical approval

Approval of the Local Research Ethics Committee of our tertiary hospital was obtained before initiating the study (project no: İ10-622-10, date:14.01.2021)).

References

1. JHU John Hopkins University of Medicine COVID-19 Dashboard. https://coronavirus.jhu.edu/map.html. access date 1.7.2021
2. Worldometer COVID-19 Coronavirus Pandemic. https://www.worldometers.info/coronavirus/. access date 1.7.2021
3. Ksiazek T, Erdman D, Goldsmith C, et al. A Novel Coronavirus Associated with Severe Acute Respiratory Syndrome. N Engl J Med. 2003;348:1953-66.
4. Cascella M, Rajnik M, Cuomo A, Dulebohn SC, Napoli RD. Features, Evaluation and Treatment Coronavirus (COVID-19). StatPearls Publishing. 2021.
5. Anderson RM, Fraser C, Ghani A, Donnelly C, et al. Epidemiology, transmission dynamics and control of SARS: the 2002–2003 epidemic. Philos Trans R Soc Lond B Biol Sci. 2004;359:1091–105.
6. Chowell G, Abdirizak F, Lee S, Jung E, et al.Transmission characteristics of MERS and SARS in the healthcare setting: a comparative study. BMC Med. 2015;13:210.
7. Wit ED, Doremalen NV, Falzarano D, Munster VJ. SARS and MERS: recent insights into emerging coronaviruses. Nat Rev Microbiol. 2016;14: 52334.
8. Chandra A, Chandra S. A comparative Analysis of SARS, MERS and Covid-19. J Contemp Med. 2020;10:464-70.
9. Wrapp D, Wang N, Corbett KS, Goldsmith JA, et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science. 2020:367:1260–3.
10. Pal D. Spike protein fusion loop controls SARS-CoV-2 fusogenicity and infectivity. J Struct Biol. 2021;213:107713.
11. Kim D, Lee J.-Y, Yang J-S, Kim JW, Kim VN, Chang H. The architecture of SARS-CoV-2 transcriptome. Cell. 2020;181:914–21.
12. Wu A, Peng Y, Huang B, Ding X, Wang X, et al. Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. Cell Host Microbe. 2020;27:325-8.
13. Shang J, Wan Y, Luo C, et. al. Cell entry mechanisms of SARS-CoV-2. Proc Natl Acad Sci. 2020;117:11727–34.
14. Lee S, Lee MK, Na H, et al. Comparative analysis of mutational hotspots in the spike protein of SARS-CoV-2 isolates from different geographic origins. Gene Rep. 2021;23:101100.
15. Li F. Structure, Function, and Evolution of Coronavirus Spike Proteins. Annu Rev Virol. 2016;3:237-61.
16. Korber B, Fischer WM, Gnanakaran S, Yoon H, et. al. Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell. 2020;182:812–27.
17. Zhang L, Jackson CB, Mou H, et al. SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. Nat Commun. 2020;11:6013.

18. Chand GB, Banerjee A, Azad GK. Identification of twenty-five mutations in surface glycoprotein (Spike) of SARS-CoV-2 among Indian isolates and their impact on protein dynamics. Gene Rep. 2020;21:100891.

19. Shah A, Rashid F, Aziz A, Jan A., Suleman M. Genetic characterization of structural and open reading Fram-8 proteins of SARS-CoV-2 isolates from different countries. Gene Rep 2020;21:100886.

20. Fang Li. Structure, Function, and Evolution of Coronavirus Spike Proteins. Ann Rev Virol 2016; 3:1,237-61.

21. Fehr A.R., Perlman S. Coronaviruses: An Overview of Their Replication and Pathogenesis. In: Maier H., Bickerton E., Britton P. (eds) Coronaviruses. Mol Biol 2015, vol 1282. Humana Press, New York, NY

22. Rice P, Longden I. Emboss: the European Molecular Open Software Suite. Trends Genet 2000;16:276-7.

23. Landes C, Henaut A, Risler J. Dot-Plot comparison by multivariate analysis (DOCMA): A tool for classifying protein sequences. Bioinformatics. 1998;9:191-6.

24. Jones DT, Taylor WR, and Thornton JM. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci. 1992;8: 275-82.

25. Kumar S, Stecher G, Li M, Knyaz C, and Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. Mol Biol Evol. 2018;35:1547-1549.

26. Díez-Fuertes F, Iglesias-Caballero M, García-Pérez J, et al. A Founder Effect Led Early SARS-CoV-2 Transmission in Spain. J Virol. 2021;95(3): e01583-20.

27. Volz E, Hill V, McCrone JT, Price A, Jorgensen D, et al. Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity. Cell. 2021;184:64-75.e11.

28. Pathan RK, Biswas M, Khandaker MU. Time series prediction of COVID-19 by mutation rate analysis using recurrent neural network-based LSTM model. Chaos Solution Fract. 2021;138:110018.