International Journal of Education and Humanities
ISSN: 2770-6702 | Vol. 3, No. 3, 2022

Uncovering of the Evolutionary Relationship of SARS-COV-2 by Analyzing 13 Genomic Sequences

Xingchen Liu*

College of Biological Sciences, University of California, Davis, CA95616, USA
*Corresponding author: Xingchen Liu

Abstract: The outbreak of acute respiratory disease caused by a novel coronavirus (SARS-CoV-2) is spreading rapidly around the world. However, the variation and evolution of this virus are still largely unknown. It is urgently necessary to predict the damages caused by this virus to humans in the future. Here, I provided the clues for the SARS-CoV-2 mutated prediction, drug treatment, morbidity, or infectivity through whole-genome analysis using different types of viruses. The 13 SARS-COV-2 genomic sequences were employed to be analyzed to look for their evolutionary relationship. I found that the spike protein in the SARS-CoV-2 had the most mutations than the other proteins, which suggested that spike protein plays a key role in the processing of viral infection and the evolution of the SARS-CoV-2. Moreover, 12 possible phosphorylation sites in spike protein were predicted indicating that these amino acids may have potential significance in viral pathogenesis. Together, our study may provide several novel clues for researchers to study the morbidity or infectivity of the virus and to find the strategies to control it in the near future.

Keywords: Compulsory isolation, Drug abstainer, Meaning in life, Relapse tendency.

1. Introduction

In December 2019, multiple cases of pneumonia of unknown cause were first reported in Wuhan, Hubei Province, China. Immediately, this disease boosted worldwide in early 2020. Simultaneously, it was confirmed to be caused by a novel type of coronavirus. On February 11, 2020, the World Health Organization (WHO) announced that the disease caused by this coronavirus would be named Coronavirus disease 2019 (COVID-19). Subsequently, the WHO named this new type of coronavirus severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).

The virus, SARS-CoV-2 contains a positive-sense single-stranded RNA genome (Kadam et al., 2021), which means that the frequency of mutated occurrence is high in its genome. (Grubaugh et al., 2020). So far, SARS-CoV-2 has developed many variants. The WHO uses the Greek alphabet to classify and label these variants. This paper mainly focuses on the previously circulating Variants of concern (VOCs): alpha and beta variants, and currently circulating VOCs: delta and omicron, for research and analysis.

The Genome structure of SARS-CoV-2 (Figure 1) includes 5' and 3' untranslated regions (UTRs), open reading frames (ORFs), and structural protein regions (Kadam et al., 2021). The ORFs include ORF1ab, ORF3a, ORF6, ORF7a, ORF8, and ORF10. Structural proteins have spike protein (S), envelope protein (E), membrane protein (M), and nucleocapsid protein (N). Among these proteins in SARS-CoV-2, the S protein was found to play a critical role in the infection of host cells. The S protein is a class I virus fusion protein, and it exists extensively on the viral membrane in the form of a trimer (Herrera et al., 2020). These trimeric proteins consist of extracellular, transmembrane, and intracellular domains (Kadam et al., 2021). The extracellular region mainly includes two subdomains, S1 and S2. S1 is primarily a membrane surface antigen, while S2 mainly acts on membrane fusion (Kadam et al., 2021). Therefore, the S protein has received intense attention from scientists.

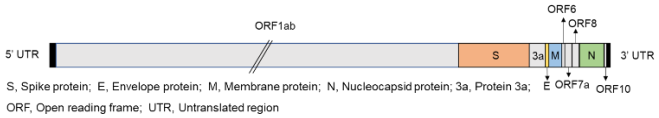

Figure 1. A diagram of SARS-CoV-2 genome and it encoded proteins

The mutation in the nuclear sequences is the driving force of viral evolution. It is necessary to analyze the mutations that existed among the different types of SARS-CoV-2, including alpha, beta, delta, and omicron. Through analysis of 13 different SARS-CoV-2 types using the programs embedded in the software or websites, I found that the sequences encoding the S protein existed the most mutations than other regions in the whole genome. The phylogenetic trees constructed using 13 S protein-coding sequences were similar to that using the 13 whole genomic sequences. Moreover, 12 potential phosphorylation residues in S protein were predicted by use of the NetPhos websites. These data may provide some novel clues for dissecting the viral pathogenesis and may be useful for making the strategies to control this disease in the near future.

2. Result

2.1. Alignment of 13 Genomic Sequences of SARS-CoV-2

13 SARS-CoV-2 genomic sequences used in this study were downloaded from the database, gene bank and GISAID. To select the representative data, four types of different viruses including alpha, beta variants specified by the WHO previously, and delta, omicron variants specified by the WHO currently, were chosen for analysis. Therefore, I downloaded the genomic sequences of 2 alpha variants, 3 beta variants, 3 delta variants, and 4 omicron variants for further analysis. To map the variations of single nucleotide polymorphisms (SNP) or indel, the first reported strain, SARS-CoV-2 in Wuhan, Hubei Province, China (MN908947.3), was used as the reference. The information of 13 genomic sequences used in this work were summarized in Table 1.

Table 1. The information of 13 genomic sequences used in this study

WHO	Location	Virus ID	Database
	Wuhan	MN908947.3	NCBI
Alpha	USA	EPI_ISL_2758533	GISAID
Alpha	England	EPI_ISL_704783	GISAID
Beta	USA	OK238749.1	NCBI
Beta	South	EPI_ISL_7545672	GISAID
Beta	France	EPI_ISL_8376888	GISAID
Delta	USA	OL771451.2	NCBI
Delta	UK	OU398444.1	NCBI
Delta	Italy	EPI_ISL_8957550	GISAID
Omicron	USA	EPI_ISL_8950312	GISAID
Omicron	Poland	EPI_ISL_8925410	GISAID
Omicron	Mexico	EPI_ISL_8953561	GISAID
Omicron	Germany	OV669872.1	NCBI

Via alignment of 13 SARS-CoV-2 genomic sequences, I found that there were 39 deletions and 3 insertions in all sequences. Among these variations, there existed 11 deletions and 1 insertion in the ORF1ab, 22 deletions and 2 insertions in the S protein, 3 deletions in the ORF8, 3 deletions in the N protein, and 2 deletions in the 3' un-translational region (UTR) (Table 2). Form the result of aliment, the rich mutations were identified in the ORF 1ab and S proteins, indicating that these two proteins play a key role in the evolution of SARS-CoV-2.

Table 2. Summary of the variations among the 13 SARS-CoV-2 viruses

Virus ID	Variations in the position of genome (proteins)
Wuhan (MN908947.3)	Reference
Alpha-USA (EPI_ISL_2758533)	11249 (9 deletions)-ORF1ab; 21718 (6 deletions)-S; 21939 (3 deletions)-S
Alpha-England (EPI_ISL_704783)	11233 (9 deletions)-ORF1ab; 21702 (6 deletions)-S; 21923 (3 deletions)-S
Beta-USA (OK238749.1)	11249 (9 deletions)-ORF1ab; 22239 (9 deletions)-S
Beta-South Africa (EPI_ISL_7545672)	11221(9 deletions)-ORF1ab; 22119 (3 deletions)-S; 22208 (9 deletions)-S
Beta-France (EPI_ISL_8376888)	11237(9 deletions)-ORF1ab; 22227 (9 deletions)-S; 29682 (12 deletions)-3'UTR
Delta-USA (OL771451.2)	21978 (6 deletions)-S; 28191 (6 deletions)-ORF8
Delta-England (OU398444.1)	22028 (6 deletions)-S; 28241 (6 deletions)-ORF8
Delta-Italy (EPI_ISL_8957550)	479 (6 deletions)-ORF1ab; 21997 (6 deletions)-S; 28210 (6 deletions)-ORF8
Omicron-USA (EPI_ISL_8950312)	11255 (9 deletions)-ORF1ab; 12606-12641 (36 insertions)-ORF1ab; 21982 (9 deletions)-S; 22180 (3 deletions)-S; 22191-22199 (9 insertion)-S; 28355 (9 deletions)-N
Omicron-Poland (EPI_ISL_8925410)	11257 (6 deletions)-ORF1ab; 21726 (6 deletions)-S; 21941 (9 deletions)-S; 22139 (3 deletions)-S; 22175-22183 (9 insertions)-S; 28305 (9 deletions)-N; 29667 (26 deletions)-3'UTR
Omicron-Mexico (EPI_ISL_8953561)	6506 (3 deletions)-ORF1ab; 21756 (6 deletions)-S; 21971 (9 deletions)-S; 22169 (3 deletions)-S
Omicron-Germany (OV669872.1)	6510 (3 deletions)-ORF1ab; 11281 (9 deletions)-ORF1ab; 21751 (6 deletions)-S; 21966 (9 deletions)-S; 22164 (3 deletions)-S;28339 (9 deletions)-N

2.2. Calculation of Mutation Ratio of ORF1ab and S Proteins

Using the deduced amino acid sequences of ORF1ab and S proteins, I calculated the mutation ratio of variations in 13 ORF1ab and S proteins, respectively. By summarizing, there are 6924 amino acids in the orf1ab region, and 59 amino acids are mutated, so the mutation rate is about 0.8521%. There are 1273 nucleotides in the S protein region, and there are 53 nucleotide sites with variation, so the variation rate is about 4.163%. After comparison, the variation rate of the S protein region is higher, so the variation in S protein is relatively abundant. Therefore, the S protein existed the most mutations in the different types of SARS-CoV-2 and is more valuable for analysis furtherly.

2.3. Similarity of Phylogenetic Trees Constructed by Genomic Sequences or Deduced Amino Acid Sequences of S Proteins

To analyze the evolutionary relationship between S protein and SARS-CoV-2 viruses, I constructed the phylogenetic trees using the 13 genomic sequences of SARS-CoV-2 and the 13 deduced amino acid sequences of S proteins, respectively (Figure 2). In comparing two phylogenetic trees, (Figure 2), I found that the phylogenetic tree constructed by the S proteins better represented the taxonomic relationship of these 13 SARS-CoV-2 strains. Four clusters were obviously observed in both phylogenetic trees, including alpha, beta, delta, and omicron types. It was no surprise that the SARS-CoV-2 in Wuhan, the first reported virus strain, was found to be cluster independent others. This data implied that the variation of S protein could represent the evolutionary relationship among different types of SARS-CoV-2.

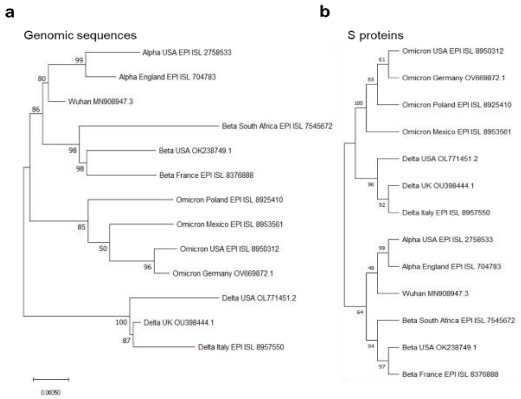

Figure 2. The phylogenetic trees constructed using the genomic sequences (a) or the deduced amino acid sequences of S proteins (b)

2.4. Prediction of the Phosphorylation Residues in S Protein and Their Conservative Analysis

For analysis of the function of S protein, I predicted the phosphorylation residues in S protein using the NetPhos website. Totally, there were 139 possible phosphorylation residues in the S protein of SARS-CoV-2 (based on the S protein of virus ID MN908947.3 - QHD43416.1). Comparing these phosphorylation residues with the mutated residues in S protein, I found 9 phosphorylation residues located in the variational residues, and they were 19T, 95T, 144Y, 371S, 373S, 375S, 478T, 547T, and 716T (Table 3). According to the conservation of these 9 sites, I identified that the 19T was a unique phosphorylation site of delta variants. 144Y, 371S, 373S, 375S, and 247T were the individual phosphorylation residues of omicron variants. 478T was a mutated phosphorylation site shared by delta and omicron variants. 716T was the unique phosphorylation residue of alpha variants. However, 114Y was only related to some strains of some specific types of variants, indicating that it was not a conservative phosphorylation residue and was not a representative site. Together, the data above suggested that some possible phosphorylation residues in the S protein may play a vital role in the function of S protein because of their close association with the mutated residues in different types of the SARS-CoV-2 viruses. These 12 residues, 19T, 95T, 371S, 373S, 375S, 443S, 478T, 493S, 498Y, 547T, 716T, and 793Y, are likely to affect the structure and function of S protein and even of the SARS-CoV-2 and then change the morbidity and infectivity of SARS-CoV-2, so they have the critical value of continuing research.

3. Method

3.1. Sequence Information for SARS-CoV-2

The 13 SARS-CoV-2 sequences involved in this experiment were from the NCBI and the GISAID databases. Their specific information is shown in table 1. The nucleic acid position data used to divide the regions of SARS-CoV-2 in this experiment were obtained from the Wuhan-1

(MN908947.3) sequence in the NCBI database.

Table 3. Prediction of the phosphorylation residues conserved in the 13 SARS-CoV-2 viruses

Number of Position	X	Score	Kinase	Conservativity in the viruses
19	T	0.558	PKC	Delta USA OL771451.2 T-R Delta UK OU398444.1 T-R Delta Italy EPI ISL 8957550 T-R
95	T	0.878	Unsp	Beta South Africa EPI ISL 7545672 T-N Delta UK OU398444.1 T-I Delta Italy EPI ISL 8957550 T-I Omicron Poland EPI ISL 8925410 T-I Omicron Mexico EPI ISL 8953561 T-I Omicron Germany OV669872.1 T-I
375	S	0.707	PKC	Omicron USA EPI ISL 8950312 S-F Omicron Poland EPI ISL 8925410 S-F Omicron Mexico EPI ISL 8953561 S-F Omicron Germany OV669872.1 S-F
478	T	0.733	Unsp	Delta USA OL771451.2 T-K Delta UK OU398444.1 T-K Delta Italy EPI ISL 8957550 T-K Omicron USA EPI ISL 8950312 T-K Omicron Poland EPI ISL 8925410 T-K Omicron Mexico EPI ISL 8953561 T-K Omicron Germany OV669872.1 T-K
547	T	0.502	CKI	Omicron USA EPI ISL 8950312 T-K Omicron Mexico EPI ISL 8953561 T-K Omicron Germany OV669872.1 T-K

3.2. Sequence Alignment

The software for aligning SARS-CoV-2 genomic sequences in this experiment was Bioedit (https://www.bioedit.com/, version 7.1.11). The Clustal|W Multiple alignments program installed in the Bioeidt was used for analysis. Other operations followed the guideline of the software.

3.3. Build Phylogenetic Trees

For the construction of phylogenetic trees in this study, MEGA (version 11.0.11) software was used to construct the phylogenetic trees, and its website is https://www.megasoftware.net/. The phylogeny method was chosen as neighbor-joining. The specific parameters are as follows: Test of phylogeny->Bootstrap method; No. of bootstrap replications -> 1000; substitution method ->Poisson model; Rates among sites -> Uniform rates; Gaps/missing Data treatment -> Pairwise deletion; Number of threads -> 8. Other operations followed the guideline of the software.

3.4. Phosphorylation Prediction

The prediction of phosphorylation residues in S protein was used on the website NetPhos-3.1 (https://services.healthtech.dtu.dk/service.php?NetPhos-3.1). The specific parameters are as follows: Residues to predict -> all three; Display only the scores higher than 0; output format -> classical; generate graphics. Other operations follow the guidelines of the website.

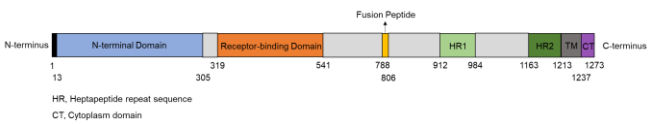

Figure 3. Exhibition of the S protein structure and its functional domains

4. Discussion

In this study, I carried out a series of data analyses with the 13 SARS-CoV-2 genomic sequences from sequences alignment and construction of phylogenetic trees to the prediction of phosphorylation residues, and it was finally found that the variation of S protein in SARS-CoV-2 could represent the evolutionary relationship of the selected 13 viruses. The S protein displayed a relatively high mutated ratio and contained multiple possible phosphorylation sites. This suggested that the S protein is likely to play an essential role in affecting the function of SARS-CoV-2. During the analysis, I can understand the evolutionary processing of SARS-CoV-2 by comparing the data and finding the critical proteins that drive the function change. The possible phosphorylation sites on vital proteins that generate variation can provide clues for future research on the pathogenicity and infection rate of SARS-CoV-2 and even drug treatment.

However, this work still has some shortcomings. This study is entirely based on data analysis, not experimental verification, which may cause the experimental results to be idealized and thus different from reality.

Compared with the other studies published before, I found that many works have proved the importance of the D614G site on the S protein of SARS-CoV-2. The mutation of the D614G site provides SARS-CoV-2 with an advantage in viral replication and increases the possibility of viral transmission (Yurkovetskiy et al., 2020), and increases the virulence of SARS-CoV-2 (Eaaswarkhanth et al., 2020). The mutation at this residue affects the morbidity and infectivity of the SARS-CoV-2. It confirms one of the results of our work that the S protein has an essential impact on the function of SARS-CoV-2. At the same time, my experiment exactly found the variation in the D614G locus in all other 12 strains compared with Wuhan-1 (MN908947.3) (Table 2).

However, the advantage of our work from the findings of others is that I summarize 12 possible phosphorylation residues. These residues may affect the structure and function of S protein and even SARS-CoV-2, which are 19T, 95T, 371S, 373S, 375S, 443S, 478T, 493S, 498Y, 547T, 716T, and 793Y. These possible phosphorylation sites may play an essential role in viral infection and transmission. S protein and these possible phosphorylation residues in it may provide a significant clue for viral pathogenesis research in the near future and even for the target sites of drug treatment in the control of COVID-19.

References

[1] SARS variants: https://www.who.int/zh/activities/tracking-SARS-CoV-2-variants.

[2] Kadam, S. B., Sukhramani, G. S., Bishnoi, P., Pable, A. A., & Barvkar, V. T. (2021). SARS-CoV-2, the pandemic coronavirus: Molecular and structural insights. Journal of Basic Microbiology, 61(3), 180-202.

[3] Grubaugh, N. D., Petrone, M. E., & Holmes, E. C. (2020). We shouldn't worry when a virus mutates during disease outbreaks. Nature microbiology, 5(4), 529-530.

[4] Herrera, N. G., Morano, N. C., Celikgil, A., Georgiev, G. I., Malonis, R. J., Lee, J. H., ... & Almo, S. C. (2020). Characterization of the SARS-CoV-2 S protein: biophysical, biochemical, structural, and antigenic analysis. ACS omega, 6(1), 85-102.

[5] Yurkovetskiy, L., Wang, X., Pascal, K. E., Tomkins-Tinch, C., Nyalile, T. P., Wang, Y., ... & Luban, J. (2020). Structural and functional analysis of the D614G SARS-CoV-2 spike protein variant. Cell, 183(3), 739-751.

[6] Eaaswarkhanth, M., Al Madhoun, A., & Al-Mulla, F. (2020). Could the D614G substitution in the SARS-CoV-2 spike (S) protein be associated with higher COVID-19 mortality?. International Journal of Infectious Diseases, 96, 459-460.