Trabalho apresentado no XLIII CNMAC, Centro de Convenções do Armação Resort - Porto de Galinhas - PE, 2024

Computational Assessment of Spike Protein Diversity in SARS-CoV-2 Lineages

Emanoelle La Santrer[1]
PPGMB/EP-SCMBH, Belo Horizonte, MG
Edgar L. Aguiar [2]
PPGMMC/CEFET-MG, Belo Horizonte, MG
Cláudia B. Assunção [3]
PPGMB/EP-SCMBH, Belo Horizonte, MG
Sandro R. Dias [4]
DECOM/CEFET-MG, Belo Horizonte, MG
Thiago. S. Rodrigues [5]
PPGMMC, CEFET-MG, Belo Horizonte, MG
Rachel B. Caligiorne [6]
PPGMB/EP-SCMBH, Belo Horizonte, MG

Abstract. In December 2019, a new beta-coronavirus was identified in Wuhan, China. The new Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) led to a global pandemic due to rapid human transmission. The disease causes severe respiratory illness that can manifest as a mild cold or total respiratory commitment, often leading to death in severe cases. In this study, 153 SARS-CoV-2 samples were selected from 43 countries and analyzed. We aimed to perform a phylogenetic analysis of the Spike protein of SARS-CoV-2 to identify the mutations and their occurrence. This work aimed to perform a comparative analysis of the evolution of the Spike protein and the major events that affected it. A Maximum likelihood tree was inferred using the software RAXML-NG. The confidence in the inferred tree topology was performed through bootstrap replicates and branch support was calculated with the Transfer index. Each clade was grouped by a clear set of mutations, some mutations were presented in multiple lineages. The lineage that presented the most original set of substitution was BA.1. An evident change was observed when the substitution profile of the lineages was observed. The ability to reconstruct the functional evolution of proteins and stipulate probable evolutionary paths allows for a better understanding of our universe and a greater preparation for future challenges.

Keywords. Phylogenetics, Bioinformatic, Virology

1 Introduction

The SARS-CoV-2 virus has four crucial structural proteins: membrane protein (M), envelope protein (E), nucleocapsid protein (N), and spike protein (S) [1]. Among these, the spike protein (S) stands out as a class I fusion protein consisting of a homotrimeric spike glycoprotein envelope that binds cellular receptors. Comprised of 1273 amino acids, the spike protein has been a primary

[1] manusantrer@gmail.com
[2] edgarlaguiar@gmail.com
[3] assuncaoclb@gmail.com
[4] sandrord@cefetmg.br
[5] thiagothiagodcc@gmail.com
[6] rachelbc@faculdadesantacasa.edu.br

DOI: 10.5540/03.2025.011.01.0365

focus of evolutionary mutations and comprises two subunits, S1 and S2 [2]. Figure 1 illustrates both subunits and their respective domains. The spike protein is pivotal for the virus's ability to enter host cells and initiate infection.

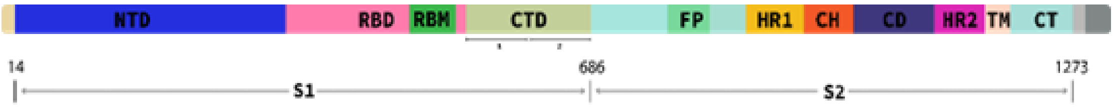


Figure 1: Domains present in the Spike protein. The first subunit (S1) includes the N-terminal domain (NTD), the Receptor Binding Domain (RBD) along with the Receptor Binding Motif (RBM), and the C-terminal domain (CTD). The second subunit (S2) contains the Fusion peptide (FP), two heptad repeat motifs (HR1 and HR2), central helix region (CH), connector domain (CD), Transmembrane anchor (TM), and a Cytoplasmic tail (CT). Font: Author figure.

This homotrimeric glycoprotein consists of two subunits, with the receptor-binding domain (RBD) responsible for recognizing and binding to the host cell receptor ACE2 [3]. Its unique features distinguish it from other betacoronaviruses, contributing to its high infectivity and pathogenicity in human cells. Notably, the RBD of the SARS-CoV-2 spike protein exhibits a significantly higher affinity for the human ACE2 receptor compared to the SARS-CoV-1 spike protein [4]. Another distinctive characteristic of the SARS-CoV-2 spike protein is the presence of a furin cleavage site, which is absent in other betacoronaviruses such as SARS-CoV-1 and Middle East respiratory syndrome coronavirus (MERS-CoV) [5]. This site enables the spike protein to be cleaved by furin protease in host cells, leading to enhanced infectivity and pathogenicity [1].

The spike protein's RBD also demonstrates a high degree of flexibility, allowing it to adopt various conformations and evade immune recognition [2]. This flexibility aids in evading neutralizing antibodies and potentially contributes to the emergence of new variants [6]. Lan et al. (2020) [7] provided insights into the RBD's structure and its interaction with the ACE2 receptor, which drives the high-affinity binding of the SARS-CoV-2 spike protein to human ACE2. The S1 subunit of the spike protein in SARS-CoV-2 comprises three primary domains: the N-terminal domain (NTD), receptor-binding domain (RBD), and subdomain 1 (SD1) [3]. The RBD, particularly the receptor-binding motif (RBM), is a crucial component of the spike protein's receptor-binding domain in SARS-CoV-2. It is a small loop structure within the RBD that directly interacts with the host cell receptor ACE2. The SD1 domain contains several amino acid residues critical for binding to ACE2. These residues form direct hydrogen bonds with ACE2, contributing to the stabilization of the RBD in the "up" conformation. Several mutations within the SD1 domain are associated with emerging variants, including Alpha, Beta, Gamma, and Delta [8].

Currently, constant global monitoring has been carried out to monitor the virus, its evolution, and its variants. This effort, however, was hampered by the pandemic. Metagenomics and bioinformatics have played an essential role in discovering, identifying, and characterizing COVID-19 cases. The data generated through bioinformatics tools and techniques led to the interpretation of the virus's architecture, and different research methodologies were essential in discovering complementary information about the new coronavirus [8] [9]. In silico studies considered the evolutionary relationship and genetic variation of the strains according to their discovery and allowed the application of clinical reasoning in terms of therapeutic drugs and population immunization [10] [11]. To effectively infect and replicate in host cells, viruses rely on genetic variability to adapt to the biological conditions of their host and environment. The emergence of new variants since the pandemic reflects viruses' natural evolution [12]. These mutations often revolve around cell entry, replication, stabilization of proteins, and evasion mechanisms against the host's immune response [13].

DOI: 10.5540/03.2025.011.01.0365

Due to the genetic variability of SARS-CoV-2 and the emergence of possible variants of clinical interest, the relative field of computational genomic analysis made it possible to verify the diversity present in the available data and extrapolate a complete characterization of the SARS-CoV-2 genome, as well as inferring the characteristics that can be studied and understanding the complexity of the studied organism. In this study, the main goal was to perform a phylogenetic comparative analysis of the Spike protein of SARS-CoV-2 to identify the mutations and their occurrence. A phylogenetic tree based on clades and mutations throughout the pandemic was created to analyze the evolution of the Spike protein and the major events that affected it. However, there are several factors to be considered when performing a comparative analysis of the Spike protein, resulting in the need for research focused on the behavior of this protein through the different causal factors of evolutionary pressure associated with the global environment. Given this dilemma, the present research aimed to perform a comparative phylogenetic analysis based on lineages post the appearance of the D614G mutation.

2 Methods

2.1 Curatorship

For this initiative, two data banks were selected. On August 2022, two sequences were selected from the non-redundant NCBI database to root the tree according to their phylogenomic association to the SARS-CoV-2 virus. SARS-CoV BJ01 (GenBank accession number AY278488.2) and the Middle East respiratory syndrome coronavirus isolate KFU-HKU 1 (GENBANK accession number KJ650297.1). Following that 153 sequences were selected randomly from the curated database of complete genomes available at GISAID to represent their respective clades, the distribution of the number of sequences per clade is dependent on the number of sequences available that fit the inclusive criteria of this study.

In order to be considered for the phylogenetic analysis, the sequences had to presente high coverage with less than 1% of Ns and less than 0.05% unique amino acid mutation. No sequences that presented insertions or deletions not previously verified by the submitter were accepted. All the sequences selected needed to have entries with the complete collection date.

2.2 Phylogenetic Inference

After curation, a multiple sequence alignment (MSA) was performed using the MAFFT software v7 [14]. The software CIAlign was used to generate the image of the MSA, no editions were performed to the sequences 1.0.17 [15]. The alignment generated by MAFFT was also applied to the Guidance2 v1.1.0 software to infer a confidence score to the multiple sequence alignment and identify unreliable regions due to the high uncertainty associated with multiple sequence alignment. After aligning the sequences, a statistical selection of the model with the best fit of nucleotide substitution was carried out using the AICc bias correction criterion using the JModelTest v2.1.10 software [16].

In sequence, Maximum Likelihood trees were constructed and estimated in the RaxML-NG software with 1000 bootstrap autoMRE [17]. All annotations and images of phylogenetic trees were generated by iTol v6 software [18].

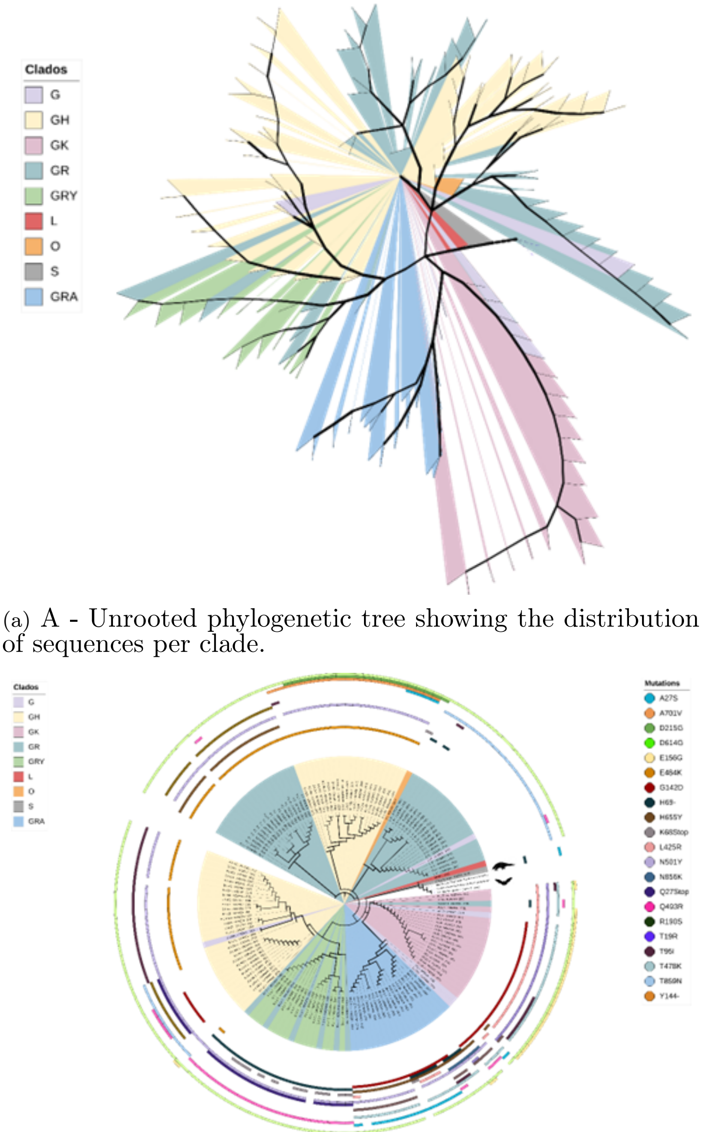
3 Results

3.1 Phylogenetic Tree

The first clade to emerge was clade O followed by clades V, S and L. Most animal genome data are found on the clades mentioned previously. With the emergence of the D614G mutation, clade L was split leading to the emergence of the G clade, which has since then become predominant. Eventually the G clade evolved into new subclades such as GR, GRY, GH, and GV reaching a climax in 2020. In 2022 with the advent of vaccination and the emergence of the lineage B.1.1.529,

DOI: 10.5540/03.2025.011.01.0365

which initially belonged to the GR clade, new profiles of mutations unfolded. In Figure 2 it's possible to observe the distribution of sequences per clade, grouped by a clear set of mutations, some mutations were presented in multiple lineages, which is a characteristic of viruses evolving to infect more effectively their hosts.



(a) A - Unrooted phylogenetic tree showing the distribution of sequences per clade.

(b) B - Maximum Likelihood phylogenetic tree, 1000 bootstrap replicates.

Figure 2: SARS-COV-2 Spike protein phylogenetic trees. 155 Spike proteins from different sequences are represented. In Figure B, mutations are in evidence. Font: Author figure.

To accurately determine the evolutionary divergences among the Spike protein genes of SARS-CoV-2, Maximum Likelihood tree approaches were employed for the phylogenetic analysis of 153 SARS-CoV-2 Spike protein isolates belonging to different clades, lineages, and variants from all over the world. To represent outgroups in the analysis, two betacoronaviruses, namely MERS-CoV and SARS-CoV-1, as well as one Rhinolophus coronavirus and one Pangolin Spike protein gene were used to root the tree due to their phylogenetic proximity to SARS-CoV-2. As anticipated, SARS-CoV and MERS-CoV were positioned on one major branch, whereas SARS-CoV-2 formed another major branch divided mainly according to clades. The GR clade exhibited a consensus between the lineages P.1 and its descendants, within this lineage three major branches were observed, and a unique P.1.1 sequence split was identified on one of the branches. Notably, this P.1.1 sequence displayed an overall greater distance of branch length when compared to the other sequences within the same lineage. It was observed that clades GR and GRY were closely related and formed a single node, indicating their shared ancestral origin. However, these clades were considerably distinct from clade GRA and lineages BA.1 and its descendants. Within the GR clade, lineage C.37 did not display branching, with most leaves sharing a common ancestor. These findings align with the available information on this lineage and support the notion that it has a relatively homogenous evolutionary history. A single sequence from lineage B.1.617.2 was observed to be aligned with clade GK, which includes the AY lineage and its descendants, indicating a shared ancestral origin with clades G. Interestingly, clades G displayed a greater branch distance from clades GK and GR, indicating a more divergent evolutionary history. Lineage B.1.351 was found to cluster together with relatively similar branch distances observed between lineages B.1.526, B.1.621, and B.1.640. Of these lineages, B.1.640 displayed the highest branch distance within clade GH when compared to other lineages, suggesting a more distant evolutionary relationship.

3.2 Mutations and their impact on lineages

Despite the fast mutation rate, the highly deleterious mutations capable of affecting significantly the infectivity and transmissibility of the virus were rapidly purged, a selective pressure to maintain viruses' ability to infect and transmit are preserved, and any mutations that hinder replication abilities are unable to persist, a direct result of the evolutionary pressure that the virus suffered due to the variety of ethnicities, sociocultural conditions, and the response of the host organism. In Figure 3, it is possible to see the most prevalent mutations on the phylogenetic tree and their respective locations on the Spike protein.

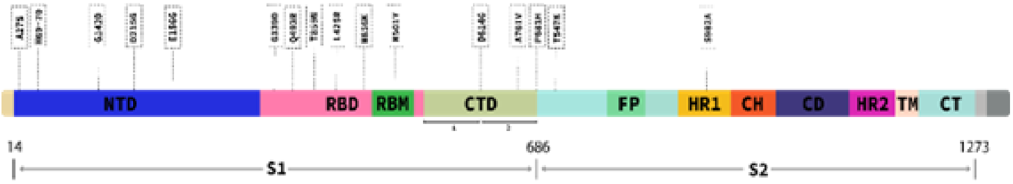


Figure 3: Spike protein subunits and the most prevalent mutations observed on the 153 sequences analyzed, each mutation is represented on their respective location on the Spike protein. Font: Author figure.

The A27S mutation is a non-synonymous substitution in the N-terminal domain (NTD) of the Spike protein of SARS-CoV-2. It is found in multiple SARS-CoV-2 lineages, including the B.1.1.7, B.1.351, and P.1 variants. Studies suggest that the A27S mutation may increase viral transmission and immune evasion by enhancing viral attachment and entry into host cells through increased binding affinity to heparan sulfate proteoglycans (HSPGs), and reducing the binding of certain neutralizing antibodies. The A27S mutation may play a significant role in the ongoing evolution of SARS-CoV-2 and impact the development of effective countermeasures against the virus. Another importante aspect of the mutations that the Spike protein underwent are deletions that occur at

DOI: 10.5540/03.2025.011.01.0365

the NTD, such as the H69, V70, and Y144 deletions. These deletions have been shown to impact the structure and function of the Spike protein and may affect the infectivity and pathogenicity of the virus. The H69 deletion, also known as the H69/V70 deletion, involves the removal of two amino acids, histidine (H) and valine (V), at positions 69 and 70, respectively. It has been identified in various lineages such as B.1.1.7, B.1.351, P.1 and B.1.617.2. It may increase the infectivity of the virus by enhancing the stability of the Spike protein in its up conformation, which facilitates binding to the host receptor [19]. Focusing further on the NTD, the D215G mutation results in the replacement of aspartic acid with glycine at position 215. It has been detected in various lineages and may enhance viral transmission by increasing the binding affinity of the Spike protein to ACE2.

Author Fang (2021) suggests that N501Y, found on multiple lineages, is the primary reason for the increase in infectivity of SARS-COV-2 due to an augmentation in binding affinity and RBD interaction surface with ACE2. There are indicators that the N501Y along with other mutations plays a vital role in evasion against antibody classes and therefore compromises immunization [19]. The L452R mutation is a non-synonymous substitution in the Spike protein of SARS-CoV-2, occurring in the RBD responsible for binding to ACE2. Its impact on the virus evolution includes increased transmissibility and immune evasion, as shown in several variants, including B.1.427/B.1.429 and B.1.617.2. The mutation affects the interaction between the Spike protein and ACE2 and alters the binding of neutralizing antibodies, reducing efficacy of vaccines and convalescent plasma therapy and increasing the risk of reinfection. Avanzato (2020) indicates that the B.1.351 lineage may have arisen from evolution between hosts through prolonged viral replication, as occurred with the N501Y mutation.

4 Conclusion

This study performs a comparative phylogenetic analysis from two public databases. It is possible to notice a clear tendency toward immune evasion when the 2021 to 2022 mutations are observed, unlike the previous mutations which were focused on more stability and higher infectivity. The pandemic has demonstrated the need to improve tools and techniques and make them available to us to analyze the enormous proportions of data generated by molecular biology and bioinformatics. The results cited above show the importance of carrying out genomic surveillance analyses. It is from the pattern of viral genomic behavior that it becomes possible to understand the evolutionary process behind the transmission and pathogenicity of a virus. This knowledge allows for a better understanding of the viral defense mechanisms of coronaviruses, their tendency towards pathogenic evolution between different species, and a global view of the cause-effect relationship behind the number of infections and deaths in the pandemic caused by COVID-19.

References

[1] T. Xiaolu et al. “On the origin and continuing evolution of SARS-CoV-2”. In: **National Science Review** 7.6 (Mar. 2020), pp. 1012–1023. ISSN: 2095-5138. DOI: 10.1093/nsr/nwaa036.

[2] S. Zheng-Li C. Jie L. Fang. “Origin and evolution of pathogenic coronaviruses”. In: **Nature Reviews Microbiology** 17.3 (Mar. 2019), pp. 181–192. ISSN: 1740-1534. DOI: 10.1038/s41579-018-0118-9.

[3] W. Chen et al. “A novel coronavirus outbreak of global health concern”. In: **The Lancet** 395 (Jan. 2020). DOI: 10.1016/S0140-6736(20)30185-9.

[4] P. Leo et al. “Identification of a Novel Coronavirus in Bats”. In: **Journal of virology** 79 (Feb. 2005), pp. 2001–9. DOI: 10.1128/JVI.79.4.2001-2009.2005.

[5] Z. Na et al. “A Novel Coronavirus from Patients with Pneumonia in China, 2019”. In: **New England Journal of Medicine** 382 (Jan. 2020). DOI: 10.1056/NEJMoa2001017.

[6] N. Vinod. “Exploring COVID-19: Relating the Spike Protein to Infectivity, Pathogenicity, and Immunogenicity”. In: (Jan. 2021). DOI: 10.13140/RG.2.2.23346.22725.

[7] L. Jun et al. “Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor”. In: **Nature** 581.7807 (May 2020), pp. 215–220. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2180-5.

[8] S. Shuo et al. “Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses”. In: **Trends in Microbiology** 24 (Mar. 2016). DOI: 10.1016/j.tim.2016.03.003.

[9] L. Xiang et al. “Bat origin of a new human coronavirus: there and back again”. In: **Science China Life Sciences** 63 (Feb. 2020). DOI: 10.1007/s11427-020-1645-7.

[10] W. Fan et al. “A new coronavirus associated with human respiratory disease in China”. In: **Nature** 579 (Mar. 2020), pp. 1–8. DOI: 10.1038/s41586-020-2008-3.

[11] Z. Ali et al. “Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia”. In: **The New England journal of medicine** 367.19 (Nov. 2012), pp. 1814–1820. ISSN: 0028-4793. DOI: 10.1056/nejmoa1211721.

[12] S. Sairaj and N. Madhavan. “Structural Proteins in Severe Acute Respiratory Syndrome Coronavirus-2”. In: **Archives of Medical Research** 51 (May 2020). DOI: 10.1016/j.arcmed.2020.05.012.

[13] P. S. Masters. **The Molecular Biology of Coronaviruses**. Vol. 66. Advances in Virus Research. Academic Press, 2006, pp. 193–292. DOI: https://doi.org/10.1016/S0065-3527(06)66005-3.

[14] R. John et al. “MAFFT-DASH: integrated protein sequence and structural alignment”. In: **Nucleic Acids Research** 47.W1 (May 2019), W5–W10. ISSN: 0305-1048. DOI: 10.1093/nar/gkz342.

[15] E. B. Katherine T. Charlotte F. Andrew. “CIAlign: A highly customisable command line tool to clean, interpret and visualise multiple sequence alignments”. en. In: **PeerJ** 10 (May 2022).

[16] P. David. “jModelTest: Phylogenetic Model Averaging”. In: **Molecular Biology and Evolution** 25.7 (Apr. 2008), pp. 1253–1256. ISSN: 0737-4038. DOI: 10.1093/molbev/msn083.

[17] K. Alexey et al. “RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference”. In: **Bioinformatics** 35.21 (May 2019), pp. 4453–4455. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz305.

[18] B. Peer L. Ivica. “Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation”. In: **Nucleic Acids Research** 49.W1 (Apr. 2021), W293–W296. ISSN: 0305-1048. DOI: 10.1093/nar/gkab301.

[19] S. W. Tuck et al. “The N-terminal domain of spike glycoprotein mediates SARS-CoV-2 infection by associating with L-SIGN and DC-SIGN”. In: (Nov. 2020). DOI: 10.1101/2020.11.05.369264.