**BIOINFORMATICS PROJECT**

*Silvio Valli*

***PROJECT’S PURPOSE***

The present study aims to produce the phylogenetic tree capable of representing the evolution of the sequence responsible to produce the **Spike** protein and **Nucleocapsid** in the different viral strains derived from the SARS-CoV-2 (GenBank id: “**NC\_045512.2**”) mutations since the beginning of the COVID-19 pandemia.

***INTRODUCTION***

**SARS-CoV-2**, also known as **Severe Acute Respiratory Syndrome Coronavirus 2,** is a viral strain of the species SARSr-CoV belonging to the genus “*Betacoronavirus”*, subgenus “*Sarbecovirus*,” of coronavirus subfamily “*Orthocoronavirinae*”. *“Sarbecoviruses”* are positively coiled single-stranded RNA viruses whose genome has about 30,000 nucleotides, which is the largest known genome for RNA viruses.

The latter are capable of infecting humans and various animals such as bats and mammals. They are responsible for diseases ranging from the common cold to more serious illnesses such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). Only 6, however, are known to have the ability to infect humans: ***229E***, ***NL63***, ***OC43***, ***HKU1***, ***MERS-CoV***and ***SARS-CoV*** [1] .Therefore, SARS-CoV-2 is logically considered the seventh. At the end of 2019, SARS-CoV-2 made its first appearance in China, but unlike previous human coronaviruses, it has spread very efficiently by infecting more than 670 million individuals worldwide and causing more than 6.8 million deaths (data as of 10/03/2023) [2].

One of the peculiarities of this virus is that although it can be transmitted by air like its predecessors (i.e. SARS-CoV), transmission can occur before the onset of symptoms and even from asymptomatic individuals (about 40/45% of infected people) [3] thereby challenging public health systems of epidemic containment. Another peculiarity of this virus, which has made it one of the deadliest in the world, is the ability of interspecies transmission (*“spillover”* [4]) as in fact many studies reports [5]. Before continuing the discussion, it is necessary to specify the origin of these viruses, in fact, as is highly documented, all previous highly pathogenic human coronaviruses have had a zoonotic origin [6] including SARS-CoV-2. Although the detailed origin of the latter is not yet certain [7] viral evolution following interspecies infection is a pivotal feature in the virus' ability to infect and evolve. Indeed, SARS-CoV-2 like all **RNA viruses** is prone to accumulate mutations in its genome during each replication cycle, an ability that allows the virus to adapt nimbly to stressful situations such as may be a suboptimal immune response (e.g., of our body) to the virus thus creating a perfect environment for the selection of variants characterized by enhanced viral replication efficiency and immune evasion.

One of the targets of these frequent replications is the **Spike** protein, i.e., a glycoprotein structure found on the outside of the virus procapsid, which is primarily responsible for mediating viral entry into the host cell through docking and fusion with the membrane receptor **ACE2** (angiotensin-converting enzyme 2) [8]**.**

During its dissemination, the *“European Centre for Disease Prevention and Control”* (ECDC) monitored and classified SARS-CoV-2 variants according to 3 basic labels: **VOC**, **VOI** and **VOM** [9] based on transmissibility, severity of infection, virulence, detectability and immune evasion.

In this context, the present study aims to produce the phylogenetic tree capable of representing the evolution of the sequence responsible to produce the **Spike** protein and **Nucleocapsid** in the different viral strains derived from the SARS-CoV-2 “**NC\_045512.2”** mutations during the pandemic.

***PROJECT DEVELOPMENT***

The project has been developed using BioPython. We will analyze all the steps to produce phylogenetic trees.

1. *Identify the SARS-CoV-2 sequences for Spike and Nucleocapsid sequences*

First, it was necessary to find on NCBI the ID of the genetic material of ***SARS-CoV-2***, namely ***NC\_045512.2*** which was subsequently used to extract the sequences coding for the proteins of our interest

1. *Identify the genomes of SARS-CoV-2 variants*

Next, it was necessary to identify the SARS-CoV-2 variants on which the phylogenetic study was carried out. The identification of these genomes was done by identifying the viral strains of greatest interest according to ECDC and catalogued on the organization's own website. Following the nomenclature ***“Pango Lineage”*** [10], a nomenclature specifically created to follow the development and evolution of SARS-CoV-2 virus around the world, the genomes of the viral strains shown in the table were identified on ***NCBI virus****.* To make the study complete, complete sequences of the genomes of interest were chosen [11]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| PANGO LINEAGE | WHO label | ECDC classification | NCBI id | FIRST DETECTION |
| BA.2.86 | Omicron | VOI | PP791391.1 | - |
| KP.3 | Omicron | VOI | PQ968032.1 | - |
| XEC | Omicron | VOM | BS016399.1 | - |
| B.1.1.7 | Alpha | Declassified | BS016085.1 | United Kingdom |
| P.2 | Zeta | Declassified | PV002179.1 | Brazil |
| B.1.351 | Beta | Declassified | PQ964503.1 | South Africa |
| P.1 | Gamma | Declassified | PV002009.1 | Brazil |
| B.1.617.2 | Delta | Declassified | PV002455.1 | India |
| C.37 | Lambda | Declassified | PV002221.1 | Perú |
| B.1.617.1 | Kappa | Declassified | PV002147.1 | India |

1. *Upload the sequences*

Once the genome ids of SARS-CoV-2 and its variants were obtained, using *"****Entrez.efecth()***”, the records in *fasta* format of the (integer) genomes of the variants and the sequence of SARS-CoV-2 responsible for encoding the Spike protein were obtained. The custom functions implemented for these purposes are ***“getGene()***” and ***“getGenomes()”***

1. *Find the gene within the genomes*

The coding sequence for the Spike protein was searched within the variant genomes through the use of a custom ***"findGene()*”** function using the ***“Aligner.score()*”** function from the **AlignIO** module in BioPython.

1. *Multiple Alignments*

After obtaining the sequences homologous to the original one responsible for encoding the Spike protein from the genomes of the viral variants, a multiple sequence alignment was initiated. This was performed with the ***“local\_msa()”*** function, which in turn uses a command that runs the **ClustalW2** program installed on the PC used on a specific user-defined file. This was achieved with the ***“subprocess”*** module, which allows programs or applications to be executed via Python code. The input file is a 'fasta' format file containing all the sequences identified in the previous step.

1. *Phylogenetic tree generation*

From the previous step we obtained an *“aln”* format file representing the result of multiple sequence alignment between the nucleotide sequences responsible for encoding the Spike protein in the different viral strains. This file was converted to *“phylip*” format to be so used by the ***“tree\_generation()”*** function, which in turn runs the ***“PhyML-3.1”*** program on the pc for generating a phylogenetic tree. In this case, a subprocess was not used but rather the ***“PhymlCommandline()”*** function imported from the ***“Bio.Phylo.Applications”*** module.

1. *Pipeline creation*

Eventually, a pipeline was created by implementing all the previous steps. Using the pipeline, the same study was carried out, on the same viral strains, but comparing the genetic sequences responsible for encoding the protein used by the virus for **Nucleocapsid** production.

***RESULTS***

***A diagram of a tree

Description automatically generated***

***A diagram of a tree

Description automatically generated***

***A graph of a graph

Description automatically generated with medium confidence***

***A graph of a number of blue lines

Description automatically generated with medium confidence***

***DISCUSSION***

Looking at graphs related to the alignment score between the sequence encoding the protein of interest and the genomes of the mutated viral strains, it can be seen that the mutation of the Spike protein is more pronounced among the affected genomes than the mutation of the protein responsible for Nucleocapsid formation.

This observation coincides with the increased interest that SARS-CoV-2 virus has in mutating the Spike protein, the main player in viral infection and the main target of most countermeasures devised by humans. This is especially true for the variants named under the name “Omicron” characterized by several important mutations in the Spike protein including A67V, T95I, L212I, Y145D, G339D, S373P, S371L, K417N, S375F, N440K, etc. that are responsible for the higher ability of the viral strain to escape the neutralization efficiency induced by prior vaccination or infections .[12]

Although the Omicron variant has a lower mortality rate than its predecessors, it is important to keep its evolution monitored for further mutations. The combination of high mortality rate and high spreading ability (already inherent in the viral strain) would pose a very dangerous challenge to the entire world population, the outcome of which could be more ominous than the COVID 19 pandemic.

Eventually, it is curious to observe the evolutionary proximity of the “Kappa” and “Lambda” variants in the evolution of the Spike protein although they are strains identified in opposite geographical areas (India and Peru). Also, observing the evolutionary proximity, for both proteins analyzed, of the variants’ named “Omicron” is a symptom of the correct execution of the phylogenetic study.

To verify the veracity of the trees obtained, the evolutionary relationships obtained could be compared with those defined by the “**Global Initiative on Sharing All Influenza Data” (GISAID)** at the following link: *“*[*https://gisaid.org/phylodynamics/global/fiocruz/*](https://gisaid.org/phylodynamics/global/fiocruz/)

# References

|  |  |
| --- | --- |
| [1] | S. R. Weiss, "Forty years with coronaviruses.," *The Journal of Experimental Medicine,* p. 217(5), 2020. |
| [2] | C. f. S. S. a. E. (. a. J. H. U. (JHU), "www.arcgis.com," [Online]. Available: https://www.arcgis.com/apps/dashboards/bda7594740fd40299423467b48e9ecf6. |
| [3] | D. P. &. T. E. J. Oran, "Prevalence of asymptomatic SARS-COV-2 infection.," *Annals of Internal Medicine,* pp. 173(5), 362–367, 2020. |
| [4] | S. Ryding, "NEWS Medical LifeScience - "What is a spillover event?"," [Online]. Available: https://www.news-medical.net/health/What-is-a-Spillover-Event.aspx. |
| [5] | B. B. O. S. R. S. N. D. F. M. R. J. M. E. M. R. V. D. S. A. T. P. R. A. B. M. B.-V. N. H. F. H. R. H. D. W.-B. M. C. A. B. R. J. Munnink, "Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans.," *Science,* pp. 371(6525), 172–177, 2021. |
| [6] | B. G. H. Z. P. &. S. Z. Hu, "Characteristics of SARS-COV-2 and COVID-19.," *Nature Reviews Microbiology,* pp. 19(3), 141–154, 2020. |
| [7] | A. L. J. I. P. J. E. G. S. A. S. R. H. Z. G. K. R. M. B. M. N. G. R. F. H. E. C. K. M. P. G. L. P. P. T. P. P. S. R. A. R. D. L. S. Crits-Christoph, "Genetic tracing of market wildlife and viruses at the epicenter of the COVID-19 pandemic.," *Cell,* pp. 187(19), 5468-5482, 2024. |
| [8] | A. G. Wrobel, "Mechanism and evolution of human ACE2 binding by SARS-CoV-2 spike," *Current Opinion in Structural Biology,* pp. 81, 102619, 2023. |
| [9] | E. C. f. D. a. P. a. Control, "https://www.ecdc.europa.eu/," 29 June 2023. [Online]. Available: https://www.ecdc.europa.eu/sites/default/files/documents/ECDC%20SARS-CoV-2%20variant%20classification%20criteria%20and%20recommended%20Member%20State%20actions\_1.pdf. |
| [10] | "Pango Lineages," [Online]. Available: https://cov-lineages.org/. |
| [11] | N. L. o. M. (NCBI), "NCBI Virus Help Documentation," July 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/labs/virus/vssi/docs/help/#filter-nucl-completeness. |
| [12] | S. B. M. N. S. D. K. &. C. C. Chatterjee, "A detailed overview of SARS-COV-2 Omicron: its Sub-Variants, mutations and pathophysiology, clinical characteristics, immunological landscape, immune escape, and therapies.," *Viruses,* pp. 15(1), 167, 2023. |