ASIAN JOURNAL OF HEALTH AND APPLIED SCIENCES
2022, VOL. 1, NO. 2, 15–19
https://doi.org/10.53402/ajhas.v1i1.6

LIGHTHOUSE

OPEN ACCESS

Phylogenetic analysis of the SARS-CoV-2 virus gene-based on the chain A glycoprotein spike in Wuhan

Suyarta Efrida Pakpahan[1*]
Ira Prima Sari[2]
Dea Koesmawati[3]

[1, 2, 3] Medical Laboratory Technology, Rajawali Health of Institute, Indonesia

ABSTRACT

The spread of COVID-19 in various countries has increased the death toll due to COVID-19. Spike glycoproteins of the SARS-CoV-2 virus have an important role in binding to host cell receptors. This spike can identify the SARS-CoV-2 kinship in Wuhan and other countries by phylogenetic analysis. This study aims to determine the phylogenetic relationship of COVID-19 from Wuhan with the other countries obtained from the NCBI Gene bank based on spike chain A glycoproteins. The method uses NCBI's BLAST program to search for similar sequences, ClustalW's program to perform multiple alignment sequencing, and MegaX to create a phylogenetic tree. The research results Spike glycoprotein chain A in Wuhan has the closest kinship with the United States. This is indicated by the formation of tree branches close to each other but still in the same group as spike glycoproteins in other countries. Phylogenetic tree validated by the bootstrap test that value of 100%, which means it shows the sturdiness of the tree can be trusted. The conclusion is there is no significant difference in the characteristics of spike glycoprotein chain A, as for some countries that have spike amino acid differences in glycoprotein chain A, such as Pakistan, Poland, and Wuhan. The amino acid difference is considered normal because the virus will continue to evolve in order to adapt to the environment.

KEYWORDS

Spike glycoproteins Chain A; SARS-COV-2; Phylogenetic

Received: January 11, 2022
Accepted: March 3, 2022
Published: June 27, 2022

Introduction

SARS-CoV-2 is a virus that attacks the respiratory system and can cause mild respiratory irritation, severe lung infections, and even death (Huang et al., 2020). SARS-COV-2 is a single-stranded RNA virus. The proteins that make up the structure of the virus include the envelope gene (E), membrane gene (M), nucleocapsid gene (N), and spike gene (S) (Ansori et al., 2020).

COVID-19 infections increased rapidly from December 2019 to January 2020. In less than a month, the disease had spread to many other provinces in China, Thailand, Japan, and South Korea (Susilo et al., 2020). The rapid global spread accompanied by severe clinical symptoms made the World Health Organization set the status of the COVID-19 pandemic on March 11, 2020, until April 16, 2021. According to data from WHO, there were 137,886,311 confirmed cases of COVID-19, including 2,965,707 deaths in the world caused by COVID-19.

Phylogenetics is an evolutionary relationship and hereditary patterns in groups of organisms (Dharyamanti, 2011). Character changes can identify the evolutionary history of organisms. The same character is the basis for analysing the relationship between one species and another (Mahfut, 2020). In phylogenetic studies, the most appropriate way to link several organisms is to create a phylogenetic tree (Dharyamanti, 2020).

A phylogenetic tree is a method that can show the evolutionary relationships between organisms (Mahfut, 2020). One of the goals of a phylogenetic tree is to accurately establish relationships between organisms and determine the differences that occur from one ancestor to their offspring (Dharyamanti, 2020)

Phylogenetic studies can be carried out through molecular phylogenetic analysis (Neapolitan, 2009). Molecular phylogenetics combines molecular biology techniques with statistics to reconstruct phylogenetic relationships (Hidayat and Pancoro, 2006). The phylogenetic relationship of a species can be determined based on the nucleotide sequence of DNA or RNA and the sequence of amino acids. These proteins are used to build a phylogenetic tree (Stan, 2002). Phylogenetic tree analysis can determine the evolutionary relationship of viruses and early studies in vaccine development to identify gene similarities with other viruses.

One of the proteins that can be used for phylogenetic analysis of COVID-19 is spike glycoprotein. Spike glycoproteins are essential in binding to host cell receptors and are the main target for neutralising antibodies (Ansori et al., 2020). Therefore, this study aimed to determine the phylogenetic relationship of COVID-19 in Wuhan and other countries obtained from NCBI Genebank data based on spike glycoprotein chain A.

CONTACT Suyarta Efrida Pakpahan suyartaefrida@gmail.com

© 2021 The Author(s). Published with license by Lighthouse Publishing.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) License (https://creativecommons.org/licenses/by-sa/4.0/), which allows others to share the work with an acknowledgement of the work's authorship and initial publication in this journal.

Methods

The method used is descriptive, using the NCBI BLAST program to search for similar sequences, the ClustalW program to perform multiple alignment sequences, and the MegaX program to construct a phylogenetic tree.

Gene preparation

SARS CoV-2 chain A was taken from NCBI. Genes are downloaded in fast format for easy analysis. Gene alignment was carried out to identify genes closely related to SARS CoV 2 Chain A***.***

Gene alignment

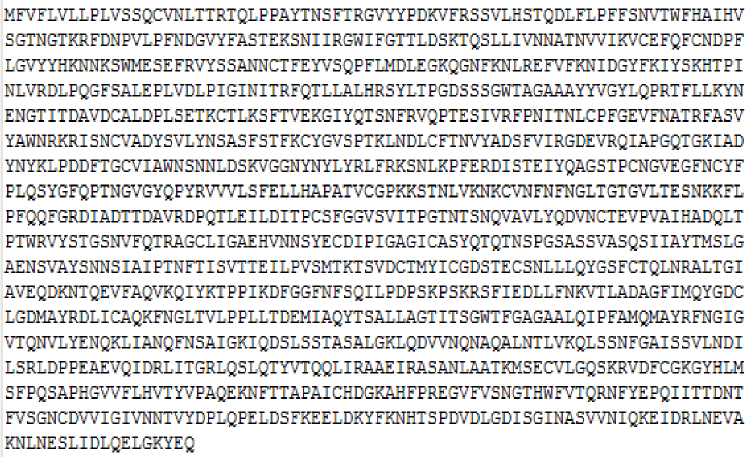
SARS-CoV 2 chain A alignment was performed using CLUTAL W. This is an early stage in creating a phylogenetic tree. This gene is a gene that is closely related to SARS CoV 2 from the NCBI database.

Phylogenetic tree

A phylogenetic tree can be created from the results of gene alignment using the Mega X tool.

Results

The Spike Glycoprotein chain-A sequence of the SARS-COV-2 virus in Wuhan obtained from NCBI (National Center for Biotechnology Information) data can be seen in Figure 1



```
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHV
SGTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPF
LGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPI
NLVRDLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYN
ENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASV
YAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIAD
YNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYF
PLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFL
PFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVPVAIHADQLT
PTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNSPGSASSVASQSIIAYTMSLG
AENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGI
AVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIMQYGDC
LGDMAYRDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIG
VTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALNTLVKQLSSNFGAISSVLNDI
LSRLDPPEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLM
SFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHFPREGVFVSNGTHWFVTQRNFYEPQIITTDNT
FVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVA
KNLNESLIDLQELGKYEQ
```

Source: www.ncbi.nih.gov/ipg/7CYP_A

Figure 1. Sample of Spike Glycoprotein chain-A virus SARS-CoV-2 sequence in Wuhan

Spike glycoprotein chain A has an Amino Acid length of 1208 aa. This spike glycoprotein sequence was then carried out by BLAST to look for similar sequences. The results of BLAST can be seen in Figure 2.

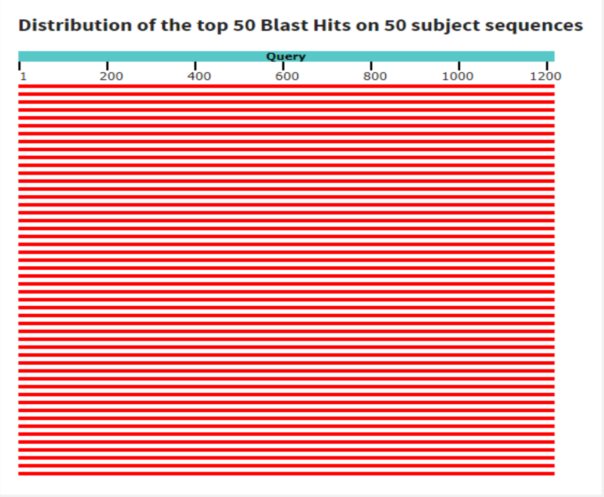
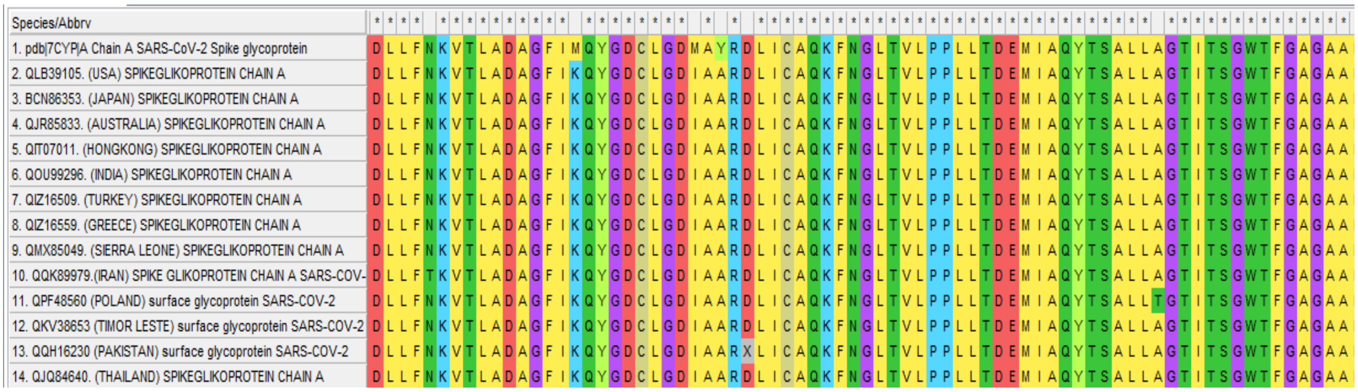


Figure 2. BLAST analysis graph Sequencing results

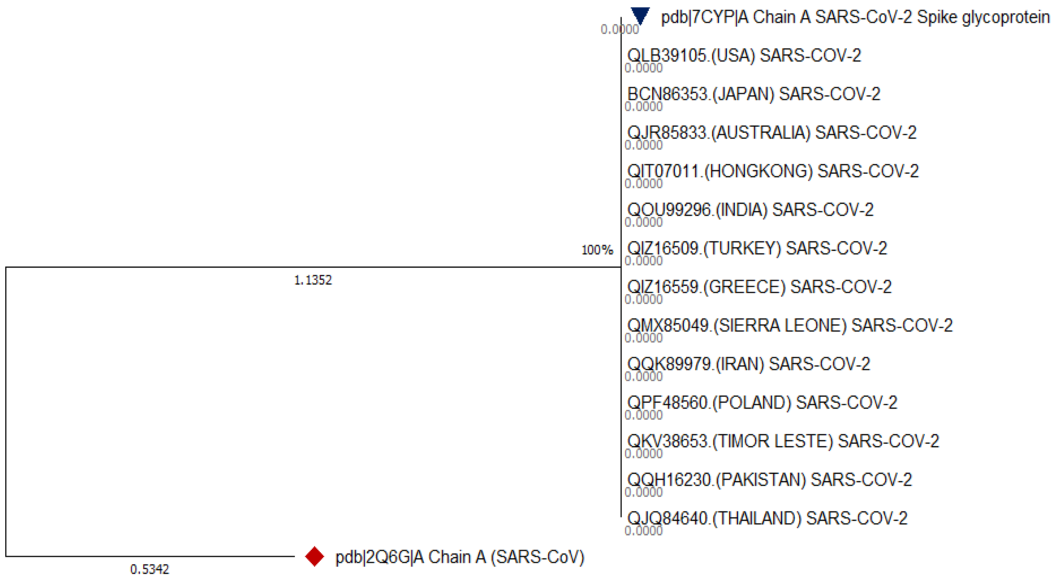
The BLAST graph shows a red colour indicating a high sequence homology level. Of the 50 BLAST data taken, only 13 sequences were taken because the location of the other 38 sequences was unknown. The 13 sequences come from the USA, Japan, Australia, Hong Kong, Turkey, Thailand, India, Greece, Sierra Leone, Poland, Pakistan, Timor Leste, and Iran. The 13 BLAST data obtained and one sample sequence from Wuhan was then performed with Multiple Alignment Sequencing using the ClustalW program.



```
Species/Abbrv                                               **** ********** ******** * * ****************************** **************
1. pdb|7CYP|A Chain A SARS-CoV-2 Spike glycoprotein         DLLFNKVTLADAGFIMQYGDCLGDMAYRDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAA
2. QLB39105. (USA) SPIKEGLIKOPROTEIN CHAIN A                DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAA
3. BCN86353. (JAPAN) SPIKEGLIKOPROTEIN CHAIN A              DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAA
4. QJR85833. (AUSTRALIA) SPIKEGLIKOPROTEIN CHAIN A          DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAA
5. QIT07011. (HONGKONG) SPIKEGLIKOPROTEIN CHAIN A           DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAA
6. QOU99296. (INDIA) SPIKEGLIKOPROTEIN CHAIN A              DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAA
7. QIZ16509. (TURKEY) SPIKEGLIKOPROTEIN CHAIN A             DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAA
8. QIZ16559. (GREECE) SPIKEGLIKOPROTEIN CHAIN A             DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAA
9. QMX85049. (SIERRA LEONE) SPIKEGLIKOPROTEIN CHAIN A       DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAA
10. QQK89979.(IRAN) SPIKE GLIKOPROTEIN CHAIN A SARS-COV-    DLLFTKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAA
11. QPF48560 (POLAND) surface glycoprotein SARS-COV-2       DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLTGTITSGWTFGAGAA
12. QKV38653 (TIMOR LESTE) surface glycoprotein SARS-COV-2  DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAA
13. QQH16230 (PAKISTAN) surface glycoprotein SARS-COV-2     DLLFNKVTLADAGFIKQYGDCLGDIAARXLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAA
14. QJQ84640. (THAILAND) SPIKEGLIKOPROTEIN CHAIN A          DLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAA
```

Figure 3. The results of the multiple alignments of 14 sequences in various countries in the ClustalW software

Based on the results of multiple sequencing alignments from 14 sequences in various countries, it can be seen that there is one different amino acid column in which Pakistan has K amino acid (Lysine). In contrast, other countries have D amino acid (aspartic acid). And differences in one amino acid column also occur. In Poland, different country sequences have an amino acid A (alanine) while Poland has a Y amino acid (tyrosine). The results of these multiple alignments are then made into a phylogenetic tree.



0.0000
pdb|7CYP|A Chain A SARS-CoV-2 Spike glycoprotein
QLB39105.(USA) SARS-COV-2
0.0000
BCN86353.(JAPAN) SARS-COV-2
0.0000
QJR85833.(AUSTRALIA) SARS-COV-2
0.0000
QIT07011.(HONGKONG) SARS-COV-2
0.0000
QOU99296.(INDIA) SARS-COV-2
0.0000
100%
QIZ16509.(TURKEY) SARS-COV-2
0.0000
1.1352
QIZ16559.(GREECE) SARS-COV-2
0.0000
QMX85049.(SIERRA LEONE) SARS-COV-2
0.0000
QQK89979.(IRAN) SARS-COV-2
0.0000
QPF48560.(POLAND) SARS-COV-2
0.0000
QKV38653.(TIMOR LESTE) SARS-COV-2
0.0000
QQH16230.(PAKISTAN) SARS-COV-2
0.0000
QJQ84640.(THAILAND) SARS-COV-2
0.0000
pdb|2Q6G|A Chain A (SARS-CoV)
0.5342

Figure 4. Phylogenetic tree of chain A glycoprotein sequences using the Neighbour-joining method and bootstrap analysis (1000 replicates) in MegaX software

Based on Figure 4, spike glycoprotein chain A SARS-COV-2 in Wuhan is in the same group as spike glycoprotein chain A in other countries, while Chain A in SARS-CoV has a different branch from the spike glycoprotein Chain A SARS-COV-2.

Discussion

This study was conducted to determine the phylogenetic relationship of COVID-19 in Wuhan and other countries obtained from NCBI data based on spike glycoprotein chain-A. In a series of research processes, starting with downloading a sample of the Wuhan spike glycoprotein chain A at NCBI and performing BLAST at NCBI, all sequences in various countries were sequenced with multiple alignments using the clustalW program, and the results of the alignment were created using a phylogenetic tree using the megaX program.

The BLAST (Basic Local Alignment Search Tools) program is used as a search tool to adjust and search for sequences that are similar to the obtained sample sequences. In this BLAST result, 50 similar sample sequences were obtained. Still, it is known that the source sequences contained 13 sequences, namely USA, Japan, Australia, Hong Kong, Turkey, Thailand, India, Greece, Sierra Leone, Poland, Pakistan, Timor Leste, and Iran.

Based on the BLAST results in Figure 2, the level of sequence homology can be shown by the values listed on the colour chart of the BLAST results. Values below 50 indicate a low homology level marked in black and blue, while green, pink and red colours indicate a homology level, which is getting higher with a value > 200. The results in the graph show a red colour which indicates that the level of sequence homology is high, so it can be concluded that all the sequences are the same and have an evolutionary relationship. This means that SARS CoV 2 from almost every country has similar genes. Therefore, these identical gene sequences can identify further studies on vaccine development from SARS CoV-2 chain A.

Multiple sequencing alignment was performed using the ClustalW program. This alignment is a necessary first step in bioinformatics for further analysis. Based on Figure 3, in the alignment process of the top row, there is an area marked with an asterisk which is an area whose sequences have similarities among other sequences or are commonly called conserved areas. Areas that are not marked with an asterisk are polymorphic regions or areas that have different sequences among other sequences. In Figure 3, there is a difference in one amino acid column in Pakistan, where isolates from other countries have D amino acids (aspartic acid) while Pakistan has K amino acids (Lysine). The difference in one amino acid occurs in sequence area 201 (attachment of Figure 3). Besides Pakistan, other polymorphic regions are Poland, Timor Leste, Iran, Turkey and Wuhan. Spike glycoprotein Chain A in Wuhan has the most polymorphic regions compared to other countries, namely the difference in 3 amino acid columns (attachment of figure 3). Polymorphic regions can be caused by gaps (marked by dotted lines) in the alignment results. The gap indicates the occurrence of a mutation process. Mutations in the virus itself are natural and normal. The virus will evolve to adapt to its environment.

The results of the multiple sequencing alignments are then made into a phylogenetic tree using the MegaX program with the neighbour-joining method. According to Dharmayanti (2011), the neighbour-joining method selects sequences that provide the best estimate of the closest branch length when combined.

Based on the phylogenetic results in Figure 4 shows that spike glycoprotein chain A in Wuhan has the closest relationship with the USA. The formation can be seen in tree branches that are close together but still in the same group as spike glycoproteins in other countries, namely Japan, Australia, and Hong Kong. , Turkey, Thailand, India, Greece, Sierra Leone, Poland, Pakistan, Timor Leste and Iran. Figure 4 also shows that the spike glycoprotein chain A SARS-CoV-2 is not in the branch of the comparison sequence, Chain-A in SAR-COV. This could be due to the spike glycoprotein Chain A SARS-COV-2 is an area with rapid evolution characterised by many amino acid differences, where SARS-COV-2 has a longer sequence length of 1208 amino acids while SARS-CoV has a longer sequence length of 1208 amino acids. 308 amino acid sequences.

The phylogenetic tree image also shows that spike glycoprotein chain A in Wuhan is in a monophyletic group with spike glycoprotein chain A in other countries and is included in the in-group category while Chain A SARS-CoV can be categorised as an outgroup. Hidayat & Adi (2008) stated that outgroup groups are needed to provide character polarisation or characteristics. Based on Figure 4, the ingroup group has a branch length that is longer (1.13) than the branch length of the outgroup group (0.53). According to Hall (2001), the greater the value of the branch length, the more sequence changes occur.

The phylogenetic tree was tested statistically using the bootstrap method with 1000 replications. Hall (2001) stated that bootstrap values of 100 to 1000 replicates were used to estimate the confidence level of a phylogenetic tree. In Figure 4, the phylogenetic tree has a 100% bootstrap value. The bootstrap value is considered high because, according to Hall (2001), a clade can be trusted with a bootstrap value of 90% and is not trusted with a bootstrap value of 25%. Ubaidillah & Sutrisno (2009) also stated that the greater the bootstrap value used, the higher the trustworthiness of the reconstructed tree topology.

Conclusion

Based on the results of the phylogenetic analysis of the SARS-CoV-2 virus gene based on the spike glycoprotein chain A in Wuhan, it can be concluded that Spike glycoprotein chain A in Wuhan has the closest relationship with the USA but is still in the same group as spike glycoprotein A in other countries. Spike Glycoprotein Chain A in each country does not significantly differ in characteristics. Based on the results of the MSA spike, glycoprotein Chain A in Wuhan has almost the same amino acid characteristics as other countries. The countries of Pakistan, Poland, Timor Leste, Iran, and Turkey have a difference of 1 amino acid. The difference in amino acids is considered normal because the virus will evolve to adapt to its environment.

Acknowledgements

Praise and gratitude, the author prays to Allah SWT because of His mercy and grace, the author was able to complete the thesis entitled "Phylogenetic Analysis of SARS-CoV-2 Virus Gene-Based on Spike Glycoprotein Chain A in Wuhan". On this occasion, the author would like to thank parents, supervisors, and friends who have provided the author's assistance, support and motivation.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

Ansori, A.N.M. *et al.* (2020). Immunobioinformatics analysis and phylogenetic tree construction of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in Indonesia: spike glycoprotein gene. Jurnal Teknologi Laboratorium, 9(1), 13–20. https://doi.org/10.29238/teknolabjournal.v9i1.221

Dharmayanti, N. I. (2020). Ulasan tentang Coronavirus : Sebagai Agen Penyakit pada Hewan dan Manusia. *Wartazoa*, 30(1), 1–14.

Huang, Chaolin, et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*, *395*(10223), 497-506. https://doi.org/10.1016/S0140-6736(20)30183-5

Ikawaty, R. (2020). Dinamika Interaksi Reseptor ACE2 dan SARS-CoV-2 Terhadap Manifestasi Klinis COVID-19. *Keluwih: Jurnal Kesehatan dan Kedokteran*, *1*(2), 70–76. https://doi.org/10.24123/kesdok.V1i2.2869

Li, F. (2016). Structure, function, and evolution of coronavirus spike proteins. *Annual review of virology*, *3*(1), 237–261. https://doi.org/10.1146/annurev-virology-110615-042301

Mahfut. (2020). Aplikasi Filogenetik Di Dunia Biologi Kesehatan: Melacak Pandemik Pathogen. *Teknosains*, *14*(2), 226–230. http://dx.doi.org/10.24252/teknosains.v14i2.15406

National Centre for Biotechnology Information (2021). *BLAST*. Available from; http://blast.ncbi.nlm.nih.gov/Blast.cgi

Qu, X. *et al.* (2019). Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. *Nature Communications*, *11*(1620). https://doi.org/10.1038/s41467-020-15562-9

Shereen, M.A., et al. (2020). COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses. *Journal of Advanced Research*, 24, 91-98. https://doi.org/10.1016/j.jare.2020.03.005

Tabibzadeh, A., et al. (2020). SARS-CoV-2 Molecular and Phylogenetic analysis in COVID-19 patients: A preliminary report from Iran. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, *84*, 104387. https://doi.org/10.1016/j.meegid.2020.104387

Walls, A. C., *et al.* (2020). Structure, Function, and Antigenicity of the SARS- Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein', *Cell*, *181*(2), 281-292. https://doi.org/10.1016/j.cell.2020.02.058

World Health Organization (2020). *WHO coronavirus disease (COVID–19) dashboard*. Available from: https://covid19.who.int/