Implementation of Hierarchical Clustering Method in Analyzing Genetic Relationship on DNA SARS-CoV-2 Sequences

E Banjarnahor¹, A Bustamam¹, W Mangunwardoyo², D Sarwinda¹

¹Department of Mathematics (Faculty of Mathematics and Natural Science – FMIPA), Universitas Indonesia, Depok, Indonesia
²Department of Biology (Faculty of Mathematics and Natural Science – FMIPA), Universitas Indonesia, Depok, Indonesia

*Corresponding author: evanderbanjarnahor@sci.ui.ac.id

Abstract. In mid-September of 2020, WHO released data stating that more than 28 million people worldwide have contracted coronavirus. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the full name coronavirus, specifically Covid-19. This virus attacks the human respiratory system and can cause infection in the human lungs and even death. WHO noted that more than 900 thousand people worldwide have died due to being infected with the coronavirus. In Indonesia, more than 210 thousand people have been infected by the coronavirus, and more than 8.5 thousand of them have died. Based on this data, it is necessary to analyze the coronavirus's kinship to reduce the spreading. This research uses The Euclidean distance in determining the distance matrix. This research will then use the Hierarchical Clustering method for analyzing the genetic relationship on DNA SARS-CoV-2 sequences. This research will take samples of SARS-CoV-2 DNA sequences from 20 countries infected. From the simulation result, the ancestors of SARS-CoV-2 coming from China. Besides, it also found that the SARS-CoV-2 DNA sequence from Indonesia has the closest ancestor from Pakistan.

1. Introduction
Bioinformatics is a combination of two big majors in research; computer science and biology. Bioinformatics involves the computer's primary functions, such as storing, collecting, manipulating, and distributing information related to biological macromolecules such as DNA, RNA, and protein [1]. [2] also explains that Bioinformatics is a multidiscipline area of biology, chemistry, physics, mathematics, and computational sciences. Bioinformatics also learns about the evolution of genes related to the structure, function, and proteins, and entire genomes; development of methods for management and analysis of biological information arising from genomics and high achievement of experiments [3]. From this explanation, we know that Bioinformatics is a combination of several scientific fields that learn about the evolution of genes, DNA, RNA, and proteins.
Currently, the world is experiencing a virus pandemic, the COVID19 (Corona virus Disease 2019) pandemic. Wuhan, one of the districts in Hubei province, is known to be the original place of this virus that is now spreading rapidly throughout the world.

This virus originated in Wuhan, precisely in China's Hubei, province at the end of December 2019 and has now spread worldwide. Coronavirus poses a severe threat to human health as; it can cause severe acute respiratory distress. Based on WHO data, in mid-September 2020, more than 28 million people worldwide have been infected with this coronavirus (SARS-CoV-2). This virus attacks human respiratory system, which can cause lung infections and can even cause death— noted that more than 900 thousand people worldwide have died due to disease associate with the coronavirus. In Indonesia, more than 210 thousand people have been infected by the coronavirus, and more than 8,5 thousand of them have died. Therefore, to reduce the spread of this virus, it is needed to learn about this virus's kinship. The most used way to know this virus's kinship is to build a phylogenetic tree or clustering. Therefore, the researcher is interested to Implement the Hierarchical Clustering Method in Analyzing Genetic Relationship on DNA SARS-CoV-2 Sequences.

2. Literature Review

2.1. Relevant Research

In this study, researchers plans to use the SARS-Cov-2 DNA sequence data as research data; and aims to help research activities amid the world's COVID-19 pandemic situation, especially in Indonesia. Researchers used the hierarchical clustering method in analyzing kinship through the formation of phylogenetic trees. There are several studies related to DNA sequences analysis [4] and to discussing the problem of automatic classification of isolates in HIV DNA. In that study, the researcher used n-Gram-based classification and unsupervised Hierarchical Clustering. The result turn out positive and shows that the proposed technique can solve different classification problems. [5] used Bayesian, Maximum Parsimony, and UPGMA methods in analyzing rRNA sequences. His study results show that the UPGMA tree produced from 16S rRNA rows and d-loops are also identical. From the research in reference [6], the researchers used K-Means Clustering for grouping 16 DNA data from the Hepatitis B virus (HBV). In their study, they grouped HBV virus data into 2 clusters. The first cluster contains 11 HBV viruses with several base pairs (bp) of more than 3000 bp and the second group containing 5 HBV viruses less than 2,000 bp.

Reference [7] also applied k-means clustering to gene expression microarray data, which measures human gene expression in the four stages of erythropoiesis. The study results are eight groups identified as a cluster of 450 genes (C4) that are more active towards the maturation stage, in cell division, and DNA replication processes. Other groups of 234 genes (C7) are more involved in autophagy (cell consumption/destruction), which are known to be involved in enucleation (expulsion of the nucleus from cells). Host spread of the SARS virus. The result of aligning host and SARS Co-V viruses in humans with SPA (Super Pairwise Alignment) is that mutations occur with type I protein replacement at positions 77, 139, 147, 244, 344, 360, 472, 480, 487, 577, 609, 613, 665, 765, 778, and 1163. Reference [8] used biclustering algorithms to solve the Multiple Sequence Alignment (MSA) problem. Their research developed the Local MSA computer program, Block MSA, and combined it with biclustering.

2.2. SARS-CoV-2

Coronavirus 2019 is caused by a new coronavirus, namely 2019-nCoV, also famous as Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) or human coronavirus 2019 (HCoV-2019) [9]. [10] explained that SARS-CoV-2 is also the source of this COVID-19 pandemic. In contrast to SARS-CoV and the Middle Eastern Respiratory Syndrome (MERS) Coronavirus (MERS-CoV), these two viruses are very pathogenic, which caused a global epidemic that reached Indonesia in 2003. SARS-CoV-2 can transmit rapidly from human to humans. At the present, there is no vaccine or any form of therapy to prevent and treat SARS-CoV-2 [9]. [11] also explained that COVID19 disease can spread through droplets or any direct contact with the patient. The incubation period is estimated 6,4 days, necessary reproductions of 2.24-3, 58. The main characteristic of a
person infected with this virus (or usual patience with pneumonia caused by SARS-CoV-2) is that they have a fever and which is then followed by coughing.

3. Data and Method

3.1. Sequence Method

Sequence Alignment is an effective method for analyzing the position and types of mutations hidden in biological sequences and allowing precise comparisons to be made [12]. Sequence alignment is used to see Similarities between DNA sequences. In physical, computational, and bioinformatics, studying the similarity of DNA sequences is essential. In almost all studies that explore evolutionary relationships, analysis of gene function, and prediction of protein structure and sequencing, it is necessary to calculate similarities [13].

Reference [14], explains the alignment of sequences. They assume that a linear gap model (where, \( S(-, a) = s(a, -) = -d \) for \( a \in Q \) and \( d > 0 \) resulted in the gap with the length of \( L \) is the same as \( -dL \)) and presents an algorithm, the Needleman-Wunsch [NW] algorithm, which can meet the optimal value of global alignments.

The primary purpose is to make an optimal alignment of the subsequences. Dynamic programming algorithms are algorithms that can reach optimization to improve performance for a smaller amount of data. For example, there are two sequences \( x = x_1 x_2 ... x_i ... x_n \) and \( y = y_1 y_2 ... y_i ... y_n \). We set up \((n + 1) \times (m + 1)\) matrix \( F \). The \((i,j)th\) element \(F(i,j)\) for \( i = 1, ..., n \) and \( j = 1, ..., n \) is the same as the optimal alignment score between \( x_1 ... x_i \) and \( y_1 ... y_i \). Element \( F(i,0) \) for \( i = 1, ..., n \) is the score of aligning \( x_1 ... x_i \) to a gap region of length \( i \). Equivalently, the element \( F(0,j) \) for \( j = 1, ..., m \) is the score of aligning \( y_1 ... y_j \) to a gap region of length \( j \). We can make \( F \) in \( F(0,0) = 0 \) and then fill the matrix from the top left corner to the bottom right corner. If \( F(i - 1, j - 1) \), \( F(i - 1, j) \) and \( F(i, j - 1) \) are known, calculate \( F(i,j) \) with the equation below:

\[
F(i,j) = \max \begin{cases} 
F(i-1,j-1) + s(x_i,y_j) \\
F(i-1,j) - d \\
F(i,j-1) - d
\end{cases}
\]

The three relatable ways to get the best value of \( F(i,j) \): \( x_i \) aligned to \( y_j \) (look at the 1st line), or \( x_i \) aligned to a gap (the 2nd line), or \( y_j \) aligned to a gap (the 3rd line). Knowing \( F(i,j) \) we keep a pointer to the option from which \( F(i,j) \) was produced. When it reaches \( F(n, m) \), then do restoration of optimal alignment.

3.2. Hierarchical Clustering

Hierarchical methods and other grouping algorithms as a significant computational way to find a cluster. Generally, it is not possible to check all the group's possibilities for one set of data, especially the biggest one [15].

Distances or differences between rows of data matrices resulted in The Hierarchical clustering trees (Harrigan in [16]). The data clustering results from Hierarchical is more diverse compared to non-hierarchical because it provides more partition trees. We can obtain different grouping levels by cutting the dendrogram that has been made by the hierarchical clustering algorithm at different levels. The hierarchical clustering agglomeration has a time complexity of \( O(n^2) \). It uses nearest neighbor reciprocity and reproducibility. The same case is in the R hclust function. Hierarchical clustering starts from a single n cluster and successively combines the most similar sets to form a larger one. Repeat this step to get the desired number of groups. The algorithm is: [17].

Input:
- Input dataset \( X = x_1, x_2 \ldots x_n \);
- The linkage function \( d \);
- Desired number of clusters \( k \)
Output:
Partition cluster \( C = \{c_1, ..., c_i, ..., c_k\} \)

1. First, form the \( n \) cluster, with each having one data point \( i.e., C_i = x_i \) and \( k' = n \)
2. When \( k' > k \), run it.
3. Then, construct a similarity matrix \( S \), where the element \( S_{ij} \) is the similarity between clusters \( i-th \) and \( j-th \) (The linkage function \( d \) calculates the similarity \( S_{ij} \)).
4. Look for the two closest clusters \( C_p \) and \( C_q \) and unite them to form a new cluster \( C_r \), \( i.e., C_r = C_p \cup C_q \)
5. Let \( k' = k' - 1 \)
6. end while
7. back to \( C = \{c_1, c_2, ..., c_k\} \)

In this study, the agglomerative hierarchical clustering use is [15]:

- The Complete Linkage Method used to find the maximum distance between a point in \( A \) and \( B \).
  \[
  D(A, B) = \max\{d(y_i, y_j) \text{ for } y_i \text{ in } A \text{ and } y_j \text{ in } B\}
  \]  

- The Average Linkage Method used to find the average distance \( n_A n_B \) between a point in \( A \) \((n_A)\) and \( B \) \((n_B)\).
  \[
  D(A, B) = \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d(y_i, y_j)
  \]

where the total is overall \( y_i \) in \( A \) and all \( y_j \) in \( B \). Join the two clusters with the smallest distance in each step.

3.3. Research Flow
Below is the general flow of this research:

![Flowchart](image-url)

**Figure 1.** Flowchart of the Method.
3.4. Distance Matrix
For the continuous features matrix, The Euclidean distance might be the most popular one used. This following formula is to count the length of The Euclidean;
\[ d_e = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2} \]  
(4)
Where index i repeats all values in the vector. Euclidean distance matrix uses two components based on the magnitude to measure the geometric distance between it. In some cases, two vectors who have high correlation values might not be presented well by the Euclidean distance \( d_e \), where their length is the only one matters [18].

| Table 1. Distance Matrix of SARS-CoV-2 DNA Sequences |
|-----------------------------------------------|
| ![Distance Matrix Table](distance_matrix_table.png) |

3.5. Proposed Method
This study uses the implementation of hierarchical clustering to analyze the genetic relationship of the DNA of SARS-CoV-2 sequences through the open-source programming language R. The summary is in the flowchart in Figure.1.

3.6. Research Data
In this study, the SARS-CoV-2 DNA sequences kinship amounting to 20 DNA sequences from 20 countries in 2020. The source of these 20 SARS-CoV-2 DNA sequences is from GenBank via http://www.ncbi.nlm.nih.gov/ and https://www.gisaid.org/ (specific for Indonesia DNA data). Table 2 shows the grouping for 20 SARS-CoV-2 DNA sequences using the Hierarchical Clustering algorithm after implementing R software based on open source. The steps used in this research are finding the collection of SARS-CoV-2 DNA sequences, distance matrix, Hierarchical Clustering, and kinship grouping using phylogenetic trees.

| Table 2. Data of the SARS-CoV-2 Virus |
|----------------------------------------|
| ![Data of the SARS-CoV-2 Virus Table](data_table.png) |
4. Results

After inputting the SARS-CoV DNA data in the simulation using the R program, the sequence alignment is carried out, the results from the distance matrix can be seen in (Table 1). The values will be used to form a Phylogenetic Tree. The phylogenetic tree results obtained from the analysis of the kinship of the SARS-CoV-2 virus are as follows.

| No. | Accession Number | Length (bp) | Country    | Year |
|-----|------------------|-------------|------------|------|
| 12  | MT372481.1       | 29898       | Malaysia   | 2020 |
| 13  | LC542976.1       | 29903       | Japan      | 2020 |
| 14  | MT359865.1       | 29890       | Spain      | 2020 |
| 15  | MT569470.1       | 29782       | Serbia     | 2020 |
| 16  | MT350282.1       | 29903       | Brazil     | 2020 |
| 17  | MT371569.1       | 29798       | Czech Republic | 2020 |
| 18  | MT531537.2       | 29901       | Italy      | 2020 |
| 19  | MT499218.1       | 29786       | Tunisia    | 2020 |
| 20  | EPI_ISL_467376   | 29842       | Indonesia  | 2020 |

Afterward, the distance matrix also simulates the making of a phylogenetic tree using a complete linked method that can result in a dendrogram, as in Figure 3, and the average coupled process that produces a phylogenetic tree dendrogram as in Figure 4.

![Phylogenetic Tree in SARS-CoV-2](image1)

**Figure 2.** Phylogenetic Tree in SARS-CoV-2.

![Cluster Dendrogram](image2)

**Figure 3.** Dendrogram with Complete Linkage Method of SARS-CoV-2 DNA.
Figure 4. Dendrogram with Average Linkage Method of SARS-CoV-2 DNA.

From the dendrogram in Figure 3 and Figure 4, the conclusion is:

1. Overall, from the two methods used, it is concluded that the SARS-CoV-2 ancestor came from code 1. Code 1 is a virus with the accession number MT079846.1, which originates from China.

2. The differences between the two methods are as follows.
   - Overall, the phylogenetic tree in the complete linked method can be divided into 3 clusters, as follows
     Cluster 1: code 1
     Cluster 2: code 19, code 2, and code 12
     Cluster 3: code 3, code 7, code 14, code 15, code 4, code 17, code 18, code 11, code 8, code 5, code 6, code 20, code 16, code 9, code 10 and code 13
   - Unlike the case with the average linked method, the phylogenetic tree is divided into 4 clusters, as follows
     Cluster 1: code 1
     Cluster 2: code 19
     Cluster 3: code 2 and code 12
     Cluster 4: code 3, code 15, code 4, code 17, code 18, code 11, code 16, code 7, code 14, code 8, code 5, code 6, code 20, code 9, code 10, and code 13

3. In the complete linked method in cluster 2, code 19 is the ancestor of code 2 and code 12. Cluster 3 changed into 2 sub-clusters. From sub-cluster 3.1, code 3 becomes the ancestor of code 7, code 14, and code 15. Code 15 will be the ancestor of code 4. Code 4 becomes the ancestor of code 17 and code 18. In sub-cluster 3.2, Code 11 becomes the ancestor of code 8 and Code 16. Code 8 is the ancestor of Code 5. Code 5 is the ancestor of Code 6 and Code 20. Code 16 is the ancestor of Code 9. Code 9 is the ancestor of Code 10 and 13.

4. In the average linked method, cluster 4 changed into 2 sub-clusters. From sub-cluster 4.1, Code 3 becomes the ancestor of code 15. Code 15 becomes the ancestor of code 4. Code 4 becomes the ancestor of code 17 and 18. In sub-cluster 4.2, Code 11 becomes the ancestor of code 16. Code 16 becomes ancestor code 7, code 14, code 8, and code 9. Code 8 becomes the ancestors of code 5. Code 5 is the ancestor of code 6 and code 20. Code 9 is the ancestor of code 10 and code 13.

5. The two methods used are the complete linkage method and the average linkage method. We can conclude that code 20 with the accession number EPI_ISL_467376 originating from Indonesia has the closest ancestor to code 5 with the accession number MT500122.1 originating from Pakistan.
6. Observed from its geographic location, countries close to each other are not necessarily in the same cluster such as Malaysia (code 12) and Indonesia (code 20) even though they are both in the Southeast Asia region.

5. Conclusion
The conclusion based on the implementation of the hierarchical clustering method in analyzing the genetic relationship through the formation of phylogenetic trees on DNA SARS-CoV-2 sequences are:

1. Based on data on the SARS-CoV-2 DNA sequence from 20 different countries in the world in 2020, the origin of the SARS-CoV-2 DNA comes from code 1 with the accession number MT079846.1, which originates from China.

2. Based on the study's simulation results, the SARS-CoV-2 ancestor closest to Indonesia (code 20) originates from Pakistan (code 5).

Further research in this direction is strongly encouraged by the recent success of hierarchical clustering in analyzing genetic relationship on DNA. We encourage readers to use other methods in determining the distance matrix. Moreover, readers could use other protein interaction data of SARS-CoV-2 or microarray data to provide more insights for discovering medicines and vaccines to cure and prevent this deadly virus.

6. References
[1] Xiong J 2006 Essential bioinformatics (New York: Cambridge University Press)
[2] Shen S 2008 Theory and mathematical methods for bioinformatics (Berlin: Springer)
[3] Higgs P G and Attwood T K 2004 Bioinformatics and Molecular Evolution (Blackwell Publishing Ltd.)
[4] Tomović A, Janičić P and Kešelj V 2006 n-Gram-based classification and unsupervised hierarchical clustering of genome sequences Comput. Methods Programs Biomed. 81 137–53
[5] Khan H A, Arif I A, Bahkali A H, Farhan A H Al and Homaidan A A Al 2008 Bayesian, Maximum Parsimony and (UPGMA) Models for Inferring the Phylogenies of Antelopes Using Mitochondrial Markers Evol. Bioinforma. 4 EBO.S934
[6] Bustamam A, Tasman H, Yuniarti N, Frisca and Mursidah I 2017 Application of k-means clustering algorithm in grouping the {DNA} sequences of hepatitis B virus ({{HBV}}) (Author(s))
[7] Saadeh H, Al Fayez R Q and Elshqeirat B 2020 Application of K-Means Clustering to Identify Similar Gene Expression Patterns during Erythroid Development Int. J. Mach. Learn. Comput. 10
[8] Wang S, Gutell R and Miranker D 2008 Biclustering As A Method For RNA Local Multiple Sequence Alignment Bioinformatics 23 3289–96
[9] Tai W, Zhang X, He Y, Jiang S and Du L 2020 Identification of {{SARS}}-{{CoV}}-2 {{RBD}}-targeting monoclonal antibodies with cross-reactive or neutralizing activity against {{SARS}}-{{CoV}}-2 Antiviral Res. 179 104820
[10] Ma C, Su S, Wang J, Wei L, Du L and Jiang S 2020 From {{SARS}}-{{CoV}} to {{SARS}}-{{CoV}}-2: safety and broad-spectrum are important for coronavirus vaccine development Microbes Infect.
[11] Lai C-C, Shih T-P, Ko W-C, Tang H-J and Hsueh P-R 2020 Severe acute respiratory syndrome coronavirus 2 ({{SARS}}-{{CoV}}-2) and coronavirus disease-2019 ({{COVID}})-19: The epidemic and the challenges Int. J. Antimicrob. Agents 55 105924
[12] Irawan M and Amiroch S 2015 Construction of phylogenetic tree using neighbor joining algorithms to identify the host and the spreading of SARS epidemic J. Applied Theor. Inf. Technol. 3171 613–65
[13] Delibašev E and Arslan A 2020 {DNA} sequence similarity analysis using image texture analysis based on first-order statistics J. Mol. Graph. Model. 99 107603
[14] Isaev A 2007 *Introduction to Mathematical Methods in Bioinformatics (Universitext)* (Berlin, Heidelberg: Springer-Verlag)

[15] Rencher A 2002 *Methods of multivariate analysis* (New York: J. Wiley)

[16] Chakerian J and Holmes S 2012 Computational tools for evaluating phylogenetic and hierarchical clustering trees *J. Comput. Graph. Stat.* 21 581–99

[17] Xu Q, Zhang Q, Liu J and Luo B 2020 Efficient synthetical clustering validity indexes for hierarchical clustering *Expert Syst. Appl.* 151 113367

[18] Matthiesen R 2010 *Bioinformatics Methods in Clinical Research* (Humana Press)

**Acknowledgments**

PDUPT 2020 grants Nos.NKB-2827/UN2RST/HKP.05.00/2020 funds for this research. We also would like to thank BRIN/DIKTI, who provided insights and expertise to improve this research in an innumerable way.