Application of the Smith Waterman and Jukes Cantor Algorithm in the Arrangement of the SARS CoV-2 Virus

Tony Yulianto¹, Mohamad Tafrikan², Rica Amalia¹, Emi Yunita³, Moch. Haikal⁴, Fathorrozi Ariyanto⁵, Zuhrotul Hasanah¹
¹Department of Mathematics, Universitas Islam Madura, Indonesia
²Department of Mathematics, Universitas Islam Negeri Walisongo Semarang, Indonesia
³Department of Midwifery, Universitas Islam Madura, Indonesia
⁴Department of Biology Education, Universitas Islam Madura, Indonesia
⁵Department of Informatics Engineering, Universitas Islam Madura, Indonesia

Abstracts
In early 2020, the world was shocked by an outbreak of a new pneumonia that started in Wuhan, Hubei Province, which then spread rapidly to more than 190 countries and territories. This outbreak was named coronavirus disease 2019 (COVID-19) caused by Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2). The spread of this disease has had a wide social and economic impact. There is still a lot of controversy surrounding this disease, including in the aspects of diagnosis, treatment, and prevention. Therefore, a study was carried out on studies related to COVID-19 that have been widely published since the beginning of 2020 until the end of March 2020. So to overcome this problem, the Smith Waterman Jukes Cantor Algorithm was made to align Covid19 by taking the a pair of DNA and RNA sequences to align protein sequences. From this alignment, the percentage of identical and mutations will be known. The identical percentage in the genetic code will prove that although the symptoms caused by the disease are almost the same, the protein sequences are not necessarily the same. Based on the simulation results of the distance between sequences that produce a phylogenetic tree using the jukes cantor method, it was obtained that 4 groups of 26 sequences were divided into groups, namely, group 1 consists of 16 sequences, group 2 consists of 6 sequences, group 3 consists of 2 sequences, group 4 consists of 2 sequences. Based on these groups, it turns out that the China Wuhan sequence (sequence MT291826) is located in group 1 and other countries that are almost similar to the sequence in China Wuhan, namely the country of Timor Leste with the sequence MT641766 also located in group 1.

Corresponding author: toniyulianto65@gmail.com
Received : 10 Jan 2022
Revised : 10 May 2022
Accepted : 25 June 2022

Keywords: Covid19; DNA; RNA; Jukes Cantor; Smith Waterman Algorithm
1. Introduction

Although genes are getting more attention for research and discussion, it is actually proteins that have a major role in carrying out life functions and make up the majority of cellular structures. If a gene is disturbed, causing the protein it encodes to be unable to carry out its normal function, it will result in a genetic defect [10, 11, 18, 20, 27, 32].

To find some of the causes of a disorder in the body’s organs that are difficult to study what the cause is, DNA is used as one of the main tools in various research in the field of biology. Genetic diversity based on mitochondrial DNA is currently highly developed because mitochondrial DNA has a high number of descendants. In addition, RNA is also needed for a ribonucleic acid contained in the genetic information flow of organisms in the form of the central dogma of DNA->RNA->protein, namely DNA is transcribed into RNA, and then RNA is translated into protein. [1, 2, 8, 9, 17].

As for in this study focused on the application of alignment with the algorithm smith-Waterman. Smith Waterman’s algorithm is a type of local alignment algorithm. From this alignment, the percentage of identical and mutations will be known. The identical percentage in the genetic code will prove that although the symptoms caused by the disease are almost the same, the protein sequences are not necessarily the same. Although it looks simple to develop algorithms based on dynamic programming with appropriate local alignment, this algorithm is very important in bioinformatics [12, 13, 14, 19, 23].

Along with the development of bioinformatics as a science that applies computational techniques to manage and analyze biological information, bioinformatics also includes the application of mathematical, statistical and informatics methods to solve biological problems, especially by using information contained in a DNA or protein sequence. By aligning and analyzing the protein sequence and the level of the virus will be known. The fundamental contribution to a field of science given in this research is the contribution to the field of Bioinformatics. As it is known that bioinformatics is a combination of mathematical, statistical, and informatics methods to process biological data [21, 30, 24, 25].

This algorithm tries to find as many similarities as possible from a pair of DNA and RNA sequences, by assigning a negative value to the base pair that is not the same (mismatch), and a positive value to the same base pair (match). So it will get the maximum positive value as the end of the alignment, and the minimum value as the beginning of the alignment. From a previous study, alignment results show similarity of 84%, with a gap of 3% [32, 34].

2. Experiments Procedure

This research will be divided into several stages as follows:

1. Literature Study
At this stage, supporting theories will be studied such as: Bioinformatics [24, 26], DNA and protein [8, 18, 30, 34], protein sequence alignment, the Smith Waterman algorithm, the application of the Smith Waterman algorithm for sequence alignment [26, 35].

2. Genbank data retrieval
The data used were taken from the National Center for Biotechnology Information [27, 36].

3. At this stage, the Smith Waterman method will be applied to the Covid data by aligning the Covid data [3, 4, 5, 6]. The Smith Waterman Algorithm is as follows [15, 16]:
   a. Alignment of protein sequences using the Smith Waterman algorithm using formula (1) [29, 30]:

   \[
   s(i,j) = \max \begin{cases}
   0 & \text{if } i = 0 \text{ or } j = 0 \\
   s(i-1,j-1) + s(x,y) & \text{if } s(i-1,j) - d, s(i,j-1) - d, s(i-1,j-1) - d
   \end{cases}
   \]

   with a series of processes and trace back algorithms.
b. Calculates the total time used to perform the alignment.
c. Calculates the identical percentage of the two sequences using Jukes Cantor's model [20, 24, 31] in equation (2).
\[ d = -\frac{3}{4} \ln \left(1 - \frac{3}{4} p \right) \]  
(2)
Where:
\[ p \]: different nucleotide proportions in two sequences 
\[ d \]: score penalty from virtual symbol
d. Indicates the mutation that occurs in the virus [7, 9, 27].
e. Forming a phylogenetic tree [34, 35, 37]

4. Program Implementation
After the Smith Waterman method was carried out, it was then applied to the Matlab software. The steps are as follows [29, 30, 31]:
a. Designing a menu interface to facilitate communication between the user and the computer (GUI)
b. Entering the FASTA code of all data sequences in a txt file.

5. Analysis and Discussion
The results of the alignment will later be tabulated and analyzed, what is the identical percentage of each alignment sequence, and the spread of the virus to Indonesia from any country [28].

6. Conclusion Drawing
This stage is the last stage in completing the research. After the research got the results from the application of the Smith Waterman method. Then the conclusions and suggestions of this research are drawn [32, 33].

3. Result and Discussion

Data Identification

During the SARS epidemic, many groups of scientists isolated and published SARS sequences. Virus that was originally recognized as the coronavirus that causes SARS, it turns out to have a different sequence with other coronaviruses that attack humans. Hence, the origin of coronavirus is thought to have come from animals [36].

In this study, we will compare some of the suspected DNA and RNA as a coronavirus host by using protein sequences that can downloaded from GenBank database [10]. GenBank is a storage site largest genetic sequence database today. GenBank is managed by the NIH (National Institute of Health) America, which is a composite of the database sequence database international nucleotides consisting of DNA DataBank of Japan (DDJB), European Molecular Biology Laboratory (EMBL) and GenBank itself in National Center of Biotechnology Information (NCBI) [26].

Mutation Result From Alignment Process

The mutation results obtained from the alignment between sequence 1 and sequence 5, and obtained 7 different sequences with the following results:

1. At the 684th position mutation from C to T
2. At position 1858 mutation from T to G
3. At the 11046th position mutation from T to G
4. At position 21674 mutation from T to C
5. At position 24288 mutation from A to G
6. At position 26107 mutation from T to G
7. At position 29658 mutation from G to A

As for the others, the mutation results obtained from the alignment between sequence 1 and sequence 10, and obtained 10 different sequences with the following results:

1. At position 701 mutation from C to T
2. At position 1875 mutation from T to G
3. At position 4382 mutation from T to C
4. At position 5042 mutation from G to T
5. At position 8762 mutation from C to T
6. At position 11063 mutation from T to G
7. At position 21691 mutation from T to C
8. At position 26124 mutation from T to G
9. At position 28124 mutation from T to C
10. At position 29675 the mutation from G to A.
Results of Phylogenetic Tree Formation

Cycle 1

a. Input: Cycle 1 evolutionary distance matrix [32, 34, 38, 40].

Step 1:

Calculate following the formula:

\[ S_i = \frac{1}{N - 2} \sum_{k=3}^{N} D_{ik} \]

Where \( N \) is the number of OTUs, \( D_{ik} \) is the distance from \( i \) to \( k \) on its evolutionary matrix. While \( N \) in this Cycle is 26.

Step 2:

Find the minimum value for each pair of sequences:

\[ M_{ij} = D_{ij} - S_i - S_j \]

When written in full, the matrix \( M_{ij} \) is as follows: The smallest \( M \) pair is obtained = - 0.0139\( M_{RX} \)

Step 3:

Define a new OTU i.e. \( U_1 \) which replaces the smallest pair \( (R \text{ and } X) \). Furthermore, these taxa are combined as \( U_1 \) follow the formula:

\[ S_{RU1} = 0.5 (D_{RX} + S_R - S_X) = -0.03421 \]
\[ S_{XU1} = 0.5 (D_{RX} + S_X - S_R) = 0.03421 \]

Step 4:

Connect taxa \( U_1 \) with \( R \) and \( U_1 \) with \( X \), respectively by following the edge length or distance as the result calculation in step 3.

Figure 1. Tree in Cycle 1

In the tree Figure 1 the length of the branch describes the evolutionary distance.

Step 5:

Merge new distances from all taxa to \( U_1 \)

\[ D_{AU1} = 0.5 (D_{RA} + D_{XA} - D_{RX}) = 0.00675 \]
\[ D_{BU1} = 0.5 (D_{RB} + D_{XB} - D_{RX}) = 0.0018 \]

Result of new distance \( U_1 \) to all taxa for further inclusion in the new evolutionary distance matrix. The next calculation step is the same as in cycle 1, the value is different because \( N \) is different, the smallest \( M_{RX} \) is different, which finally the new distance to each taxa is also different.

After all the branches are combined, the following phylogenetic tree is obtained [22, 27, 28, 38, 39, 40]:

Figure 2. Group and Distance

4. Conclusion

From the discussion that has been carried out, the conclusions obtained are as follows: the results of mutations between sequences produce differences, sometimes mutations occur and there are no mutations at all. As for the mutation, the susceptibility is very small because only the location is different and the virus is on average the same, namely the type of SAR-COV-2 virus.

Based on the simulation results of the distance between sequences that produce a phylogenetic tree using the Jukes Cantor method, 26 sequences were obtained consisting of 4 groups of them, group 1 consisted of 16 sequences, group 2 consisted of 6 sequences, group 3 consisted of 2 sequences, group 4
consisted of 2 sequences. Based on these groups, it turns out that the China Wuhan sequence (sequence MT291826) is located in group 1 and other countries that are almost similar to the sequence in China Wuhan, namely the country of Timoe Leste with the sequence MT641766 also located in group 1. The accuracy obtained is 1 hour 28 minutes 10,080 seconds

Acknowledgement

The research is supported by mathematics department, mathematics and science faculty and LPPM Universitas Islam Madura who has provided support in the form of assistance to conduct a research.

References


[40] H. Prasetya, Performance Comparison Between Kimura 2-Parameters and Jukes-Cantor Model in Constructing Phylogenetic Tree of Neighbour Joining (NJ), Bogor: IPB (Bogor Agricultural University), 2011.