Maximum Likelihood Method on The Construction of Phylogenetic Tree for Identification the Spreading of SARS Epidemic

[1]Siti Amiroch, [2]M. Syaiful Pradana
Mathematics Department of
Universitas Islam Darul ‘Ulum
Lamongan, Indonesia
siti.amiroch@unisda.ac.id, syaifulp@unisda.ac.id

[3]M. Isa Irawan, [4]Imam Mukhlash
Mathematics Department of
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
mii@its.ac.id, imamm@matematika.its.ac.id

***Abstract*—There are several phylogenetic tree construction methods that can be used to illustrate the spread of disease epidemics. One of the methods used is probability-based maximum likelihood. This paper explains how to identify the spread of the SARS epidemic in the world using the formation of phylogenetic trees using the maximum likelihood method. The output obtained was a description of the spread of SARS virus with the selection of tree candidates after performing a heuristic search of the Stepwise Addition method. The result of the research showed that the closest distance to the palm civet is GZ 02/18/03 indicating that the initial spread of the SARS epidemic started from Guangzhou.**

Keywords—maximum likelihood; phylogenetic tree; SARS epidemic.

I. Introduction

Relationships between organisms can be represented as tree diagrams. This tree diagrams become a very precise model to show the correct pattern of relationships between species, individuals, even genes. In DNA sequencing technology, it can be inferred that phylogenetic tree diagrams are closely related to DNA analysis based on the overall comparison of the genome [1]. One of the methods used for phylogenetic tree inference is the probabilistic-based method maximum likelihood [2][3][4][5].

Severe Acute Respiratory Syndrome (SARS) is one example of cases that can be identified with this phylogenetic tree inference. The pattern of the spread of SARS virus is thought to be a phylogenetic tree in which all SARS viruses are related to each other. This started from a single virus emerging in China [1] which then formed a branched relationship network as an SARS epidemic transmitted from one individual to another. The branching process of the tree diagram is to determine the beginning of the epidemic [1] [6].

There have been some previous studies conducted in Phylogenetic analysis of SARS epidemic. Several biologists [7] [8][9] conducted the analysis by using the CLUSTAL W application revealing that the relationship was most closely associated with the coronavirus in a palm civet originating from Guangdong, China. Another researcher conducted a research on the phylogenetic tree construction using a distance-based and neighbor joining algorithm that was simulated in Matlab [6]. The researcher then conducted a more detail analysis using multiple alignment method especially in stable and unstable areas, network topology and decomposition of mutation network [10] with the result indicating that civet (palm civet) as the host of the SARS epidemic. Meanwhile [11] used phylogenetic trees using Unweighted Pair Group Methods with Arithmetic Average (UPGMA) method to identify various types and the spread of Ebola virus.

With regards to maximum likelihood method, a number of studies have been conducted. Thorne [12] presented an evolutionary model for maximum likelihood alignment of DNA sequences. In [13], maximum-likelihood is used to fit a model to the scores, avoiding any costly simulation of random sequences. The method is applied in detail to the Smith-Waterman algorithm when gaps are allowed, and is shown to give results very similar to those obtained by simulation. Serdoz et all [14] introduced a maximum likelihood estimator for the inversion distance between a pair of genomes, using a grup-theoretic approach to modelling inversions introduced recently. However, there has been no study on the use of maximum likelihood method to identify the spread of SARS using phylogenetic trees. This study explains how to identify the spread of the SARS epidemic through the formation of phylogenetic trees using the maximum likelihood method.

II. Related Work

A. Sequence Alignment

Sequence alignment [5] is a method of position analysis and mutation type hidden within the biological sequence. The most important thing in the sequence alignment is to determine the transfer of mutations resulting in the genetic distance of each sequence. Given two sequences [5] A_1 and A_2 as shown in (1).

$$A_1 = (a_{11}, a_{12}, ..., a_{1n_a}) \text{ and } A_2 = (a_{21}, a_{22}, ..., a_{2n_a}) \quad (1)$$

978-1-5386-5280-0/18/$31.00 ©2018 IEEE

The insertion of "-" symbols into sequences A_1 and A_2 aims to form two new sequences called sequence $A_1^{'}$ and $A_2^{'}$. Furthermore, elements of sequences A_1 and A_2 have ranges from V_5={0,1,2,3,4} or *{a,c,g,t,-}*. The definition of multiple sequences is a collection of sequences expressed as (2).

$$\mathrm{A} = \{A_1, A_2, \ldots., A_m\} \qquad (2)$$

Where each A_s is a separate sequence defined in V_q, and expressed as (3).

$$A_s = (a_{s,1}, a_{s,2}, \ldots, a_{s,n_s}), \qquad s = 1,2,\ldots.,m \qquad (3)$$

with n_s as the length of the sequence A_s and *m* as the number of sequences in each group.

B. Jukes Cantor Evolutionary Model

Considering that the observed genetic distance may not pay attention to the true number of evolutionary changes, many studies have developed methods for converting that distance into actual evolutionary distances. This technique is often called the distance correction method; which purpose is to 'correct' the observed distance by estimating the number of evolutionary changes that have occurred [4]. In this model, the distance between two nucleotide sequences is given by (4).

$$d = -\frac{3}{4}\ln(1-\frac{4}{3}p) \qquad (4)$$

With *p* as a different proportion of nucleotides in two sequences.

C. Markov Chain Continuous Time

The Markov chain is defined as the stochastic process of the random variable $\{X(t), t \geq 0\}$ that forms a series that satisfies the Markov properties, i.e.

$$p_{ij}(t) = P(X(t+s) = j \mid X(s) = i) \qquad (5)$$

with $p_{ij}(t)$ as the probability of transition from state i to state *j* [15]. The transition probability matrix of a Markov chain is:

$$P(t) = \begin{bmatrix} p_{00}(t) & p_{01}(t) & \cdots & p_{0j}(t) \\ p_{10}(t) & p_{11}(t) & \cdots & p_{1j}(t) \\ \vdots & \vdots & \ddots & \vdots \\ p_{i0}(t) & p_{i1}(t) & \cdots & p_{ij}(t) \end{bmatrix}$$

with $p_{ij}(t) \geq 0$ and $\sum_j p_{ij}(t)$=1.

DNA has four nucleotides, that is *A, C, G,* and *T*. These four nucleotides are used as a set of state or state that may occur. Since, $S = \{A, C, G, T\}$, the transition probability matrix is:

$$P(t) = \begin{bmatrix} p_{AA}(t) & p_{AC}(t) & p_{AG}(t) & p_{AT}(t) \\ p_{CA}(t) & p_{CC}(t) & p_{CG}(t) & p_{CT}(t) \\ p_{GA}(t) & p_{GC}(t) & p_{GG}(t) & p_{GT}(t) \\ p_{TA}(t) & p_{TC}(t) & p_{TG}(t) & p_{GG}(t) \end{bmatrix}$$

With $t \geq 0$, $P(t) \geq 0$, and the probability value of the number of elements per line corresponds to one [16].

Nucleotide substitution from state *i* to state *j* is a Markov chain with continuous time. This means that nucleotide changes from state *i* to state *j* during that *t*+*s* time and that nucleotide changes from state *i* to state *k* over *t* time, then continues from state *k* to state *j* over *s* time. Thus, based on the Chapman Kolmogorov equation, equation (5) can be written as (6).

$$p_{ij}(t+s) = \sum_{k=0}^{\infty} p_{tk}(t) p_{kj}(s) \qquad (6)$$

Since the nucleotide substitution is a Markov chain with continuous time, it can be said that the nucleotide substitution is a Markov chain with regular continuous time.

D. Maximum likelihood method

Maximum Likelihood Method or maximum resemblance is the search for the maximum value for the analysis of a certain character configured between gene or protein sequences to find the greatest resemblance value in a given tree. The rationale for this method is to maximize the likelihood topology generated by a particular substitution model and the tree selected at the end has the highest likelihood value. The parameter considered is not topology but the branch length of each topology and likelihood is maximized to estimate branch length. [2]

In evolution, mutation is a change of event. The likelihood function is the conditional probability of the given data. Therefore,

$$\begin{aligned} L(\tau \mid \theta) &= P(Data \mid \tau, \theta) \qquad (7) \\ &= P(\text{sequence} \mid \text{tree, evolutionary model}). \end{aligned}$$

From equation (7), MLE of τ and θ namely $\hat{\tau}$ and $\hat{\theta}$, making Likelihood function as much as possible becomes: [2]

$$\hat{\tau}, \hat{\theta} = \arg\max_{\tau,\theta} L(\tau, \theta) \qquad (8)$$

The principle of likelihood is to prioritize the opportunities that often occur. The downside of this method is that it requires explicit models of evolution. The analysis takes a considerable amount of time and it often results in two values from one generated tree, making it difficult to conclude whether the value obtained is the maximum value. Maximum likelihood can also be used to find the rate of direct evolution of the data and to determine the best tree of kinship that can be generated from the data. This means that the maximum likelihood searches for trees and evolutionary parameters of observed data with the greatest possibility.

1) Maximum Likelihood Estimation for Two Sequences

$$L = \varphi_{x_1} p_{x_1y_1}(t)\varphi_{x_2} p_{x_2y_2}(t) \ldots \varphi_{x_n} p_{x_ny_n}(t) \qquad (9)$$

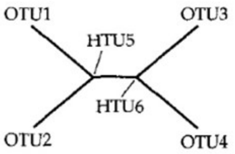
The transition probability from one sequence to another is the probability of all possible paths connecting the two sequences. Specific paths from one sequence to another can be expressed as alignment [12].

2) Maximum Likelihood Tree for Four Sequences

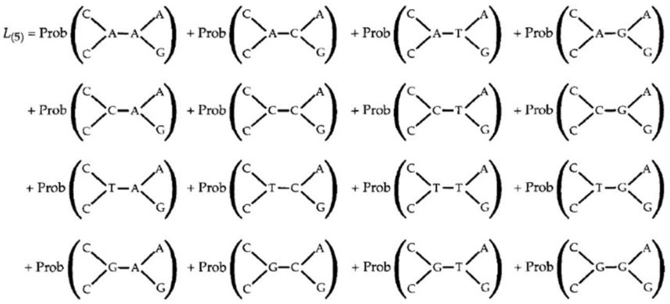
The stages in the formation of maximum likelihood trees using 4 taxa/OTU are shown as follows: [17]

OTU 1 = AACCCCTTT...N
OTU 2 = AACCCGTTA...N
OTU 3 = AACCAGTTT...N
OTU 4 = AACCGGTTT...N

On the fifth site, OTU 1, 2, 3, and 4 have nucleotides C, C, A, and G respectively. With this 5th site sample, the Maximum Likelihood method reconstructs the kinship of four OTUs by creating an unrooted tree. The probability of nucleotides of ancestors (5 & 6) is then calculated yielding nucleotide states as follows:



Since the nucleotides are 4 A, C, T, and G. Thus, there are 16 combinations of nucleotides for the two ancestral states (AA, AG, AC, AT, CA, CC, CT, CG, TA, TC, TT, TG, GA, GC, GT, GG). The probability of all combinations is calculated by using of existing sequence alignment data and then it is accumulated into likelihood values for site number 5 (L5). So for the 5th site the likelihood value is:



The Likelihood Maximum method then will calculate the likelihood value for a tree based on the entire site in alignment. Thus the likelihood value (L) of a tree is formulated by (10).

$$L = L_{(1)} \times L_{(2)} \times L_{(3)} \times .... \times L_{(n)} = \prod_{i=1}^{n} L_{(i)} \quad (10)$$

In general, the value of L is so small that it is expressed in logarithmic form (lnL). Thus, the formula becomes (11):

$$\ln L = \ln L_{(1)} + \ln L_{(2)} + \ln L_{(3)} + ... + L_{(n)} = \sum_{i=1}^{n} \ln L_{(i)} \quad (11)$$

The formula only looks for the likelihood value of a single tree.

3) *Find The Number of Maximum Likelihood Trees Possible*

The difficult task in phylogenetic tree reconstruction with the maximum likelihood method is to find a tree out of all possible tree structures by maximizing global possibilities. Unfortunately, there is no efficient algorithm to guarantee the best tree localization of all possible tree topologies. A number of unrooted binary topology trees increase with a number of taxa (n), which can be calculated according to (12).

$$t_n = \frac{(2n-5)!}{2^{n-3}(n-3)!} = \prod_{i=1}^{n}(2i-5) \quad (12)$$

When calculating the maximum likelihood tree, model parameters and branch lengths must be calculated for each tree, and then the tree that generates the highest likelihood value will be selected. The highest MLE value indicates that the tree can better explain the sequence alignment. Due to the large number of tree topologies, testing all possible trees is ineffective, and also computationally impossible. Thus, various heuristic methods are used to suggest the selected tree. [4]

4) *Stepwise Addition*

Stepwise Addition [4] is the first heuristic method to find the maximum likelihood of a tree. The procedure starts from an unrooted tree topology in three randomly selected taxa from a list of n. Then one of them reconstructs the appropriate maximum likelihood tree. To extend this tree, one of the remaining n - 3 taxa is randomly selected. The taxon is then inserted into each of the best tree branches. This means that branch insertion occurs in the highest likelihood. Thus, the local decision criterion selects the tree with the highest likelihood from the list of 2k - 3 trees if k taxa are already in the sub-tree. The resulting tree is then used to repeat the procedure. After step n - 3, the maximum likelihood tree is obtained. This means that the order of insertion of the taxa and the given local decision criterion indicate that there is no better tree. However, in Stepwise Addition this only calculates maximum likelihood for tree [4] . Thus, it is possible that the insertion sequence of the other taxa will provide the tree with a higher likelihood. [4]

III. Methodology

A. *Dataset*

Data used in this study were 14 DNA sequence data of SARS-infected patients as in Table I, taken from genbank, the world's largest gene database belonging to the United States government. The retrieval was performed by accessing the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov).

B. *Methodology*

Data of patient sequences infected with SARS virus were stored in FASTA file and the file was input into Molecular Biology and Evolution Toolbox (MBEToolbox) written in Matlab. From the data, an evolutionary distance matrix is formed with the Jukes Cantor model. Thus, the likelihood value is calculated from a number of possible trees.

TABLE I. GENBANK ACCESS CODE DATA OF SARS DEPLOYMENT CASE

Genbank Code	Sequence Code		
	Date	Location	Seq.
AY278489	DEC-16-2002	Guangzhou 12/16/02	A
AY394997	DEC-26-2002	ZongShan 12/26/02	B
AY395004	JAN-04-2003	ZongShan 01/04/03	C
AY394978	JAN-24-2003	Guangzhou 01/24/03	D

Genbank Code	Sequence Code		
	Date	Location	Seq.
AY394983	JAN-31-2003	Guangzhou Hospital	E
AY304495	FEB-18-2003	Guangzhou 02/18/03	F
AY278554	FEB-21-2003	Metropole 02/21/03	G
AY278741	FEB-26-2003	Hanoi 02/26/03	H
AY274119	FEB-27-2003	Toronto 02/27/03	I
AY283794	MAR-01-2003	Singapore 03/01/03	J
AY291451	MAR-08-2003	Taiwan 03/08/03	K
AY345986	MAR-19-2003	HongKong 03/19/03	L
AY394999	MAY-15-2003	HongKong 05/15/03	M
AY627048		Palm civet	N

From a number of possible likelihood trees, one tree with the highest likelihood value was analyzed. The results were obtained with the help of Molecular Biology and Evolution Toolbox (MBEToolbox) [18] written in Matlab.

As with the program package in Phylip, there is a DNAml package that estimates the phylogenetic tree of the nucleotide sequence with maximum likelihood. The model used allows unexpected frequencies of four nucleotides, unequal transition and transversion levels, and the average rate of change that varies across different categories of sites, as well as the use of the Hidden Markov Model. While MBEToolbox_DNMAL is a modified version of DNAml on Phylip for MBEToolbox written in Matlab. [18]

IV. Result And Discussion

The following is the result of construction of a phylogenetic tree with the maximum likelihood method.

A. Acquired genetic distance with evolutionary model Jukes Cantor

Table II is the result of genetic distance, A represents sequence 1 (Guangzhou 12/16/02), B represents sequence 2 (Zongshan 12/26/02), C represents sequence 3 (Zongshan 01/04/03), D represents sequence 4 (Guangzhou 01/04/03), E represents sequence 5 (Guangzhou Hospital), F represents sequence 6 (Guangzhou 02 /18/03), G represents sequence 7 (Metropole), H represents sequence 8 (Hanoi), I represents sequence 9 (Toronto), J represents sequence 10 (Singapore), K represents sequence 11 (Taiwan), L represents sequence 12 (HongKong 03), M represents sequence 13 (HongKong05), and N represents sequence 14 (the Palm Civet).

TABLE II. THE RESULTS OF GENETIC DISTANCE

	Genetic distance with JC evolutionary model					
	A	B	C	D	E	F
A						
B	4.683					
C	4.023	3.809				
D	3.961	2.588	4.071			
E	2.715	4.815	531.297	5.035		
F	5.301	0.049	4.208	2.608	4.880	
G	5.263	0.049	4.217	2.611	4.837	0.000
H	4.952	2.670	3.334	2.500	3.839	2.682
I	4.720	2.671	3.340	2.495	3.856	2.683
J	0.049	531.297	4.875	3.923	2.556	4.248
K	4.965	2.672	3.343	2.497	3.856	2.684
L	5.263	0.049	4.198	2.610	4.880	0.000
M	5.494	0.045	3.912	2.589	4.802	0.043
N	531.297	2.537	3.677	2.818	3.881	2.552

	Genetic distance with JC evolutionary model						
	G	H	I	J	K	L	M
A							
B							
C							
D							
E							
F							
G							
H	2.677						
I	2.670	0.000					
J	4.248	5.516	5.465				
K	2.678	0.000	0.000	5.516			
L	0.000	2.682	2.675	4.248	2.682		
M	0.043	2.684	2.685	4.961	2.686	0.042	
N	2.552	3.252	3.252	531.297	3.232	2.552	2.552

B. The empirical basic frequencies obtained for the sequence nucleotide are:

Nucleotide	A	C	G	T
frequencies	0.28476	0.19982	0.20794	0.30748

Fig. 1 shows phylogenetic tree with maximum likelihood method and the value of *ln* Likelihood which is –269135.56466. This shows that the maximum value of likelihood is 269135.56466 indicating the maximum estimated value for branch length. Furthermore, this value is used to determine which tree can explain a better sequence alignment. Although the *ln* of likelihood value is negative, it only means that the corresponding probability is less than 1 since the one that matters is not the positive or negative sign but the logarithm value.

Fig. 2 shows the output of branch length and the approximation confidence limits. Fig. 2 also shows the branch length between two branches. For example, the branch length between 12 with palm civet is 5.39457 which is at a certain approximation confidence limits. This means that the distance is within confidence interval 3.63253 – 7.15668. A positive value at the confidence interval means that the branching does not need to be arranged since the interval value is still below estimation in the sense of a narrow interval indicating the branch length is more accurate. Accuracy is also reinforced with a positive significance with a value of p <0.01. However, in some branches, the confidence limit approximation has a positive significance with p <0.05. For example, at a distance/length of a branch between 12 and GZ02/18/03 of 0.00012, where this distance is within the confidential interval (0.00002 - 0.00030) with the interval length of 0.00028.

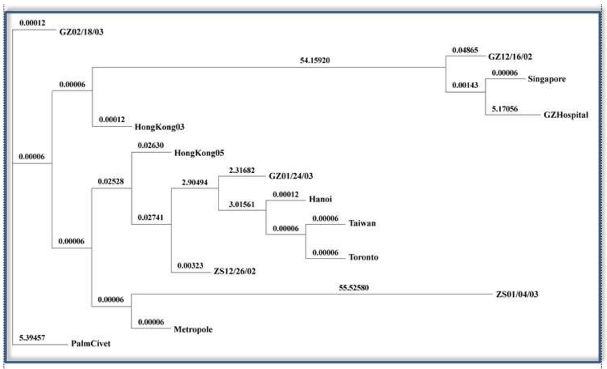
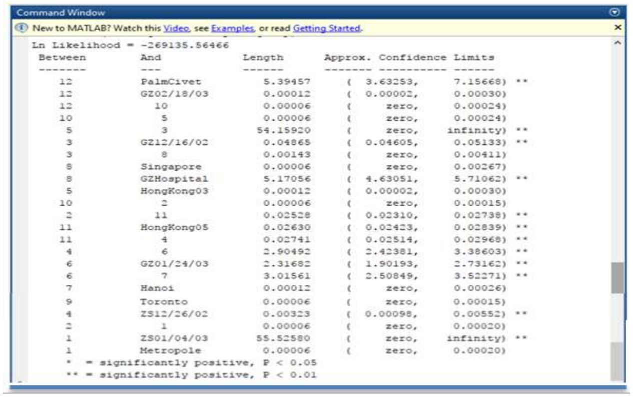


Fig. 1. Phylogenetic tree with maximum likelihood method


Command Window
New to MATLAB? Watch this Video, see Examples, or read Getting Started.

```
 Ln Likelihood = -269135.56466
  Between        And              Length      Approx. Confidence Limits
  -------        ---              ------      ------- ---------- ------
     12          PalmCivet         5.39457      (  3.63253,      7.15668) **
     12          GZ02/18/03        0.00012      (  0.00002,      0.00030)
     12            10              0.00006      (     zero,      0.00024)
     10             5              0.00006      (     zero,      0.00024)
      5             3             54.15920      (     zero,     infinity) **
      3          GZ12/16/02        0.04865      (  0.04605,      0.05133) **
      3             8              0.00143      (     zero,      0.00411)
      8          Singapore         0.00006      (     zero,      0.00267)
      8          GZHospital        5.17056      (  4.63051,      5.71062) **
      5          HongKong03        0.00012      (  0.00002,      0.00030)
     10             2              0.00006      (     zero,      0.00015)
      2            11              0.02528      (  0.02310,      0.02738) **
     11          HongKong05        0.02630      (  0.02423,      0.02839) **
     11             4              0.02741      (  0.02514,      0.02968) **
      4             6              2.90492      (  2.42381,      3.38603) **
      6          GZ01/24/03        2.31682      (  1.90193,      2.73162) **
      6             7              3.01561      (  2.50849,      3.52271) **
      7          Hanoi             0.00012      (     zero,      0.00026)
      9          Toronto           0.00006      (     zero,      0.00015)
      4          ZS12/26/02        0.00323      (  0.00098,      0.00552) **
      2             1              0.00006      (     zero,      0.00020)
      1          ZS01/04/03       55.52580      (     zero,     infinity) **
      1          Metropole         0.00006      (     zero,      0.00020)
      *  = significantly positive, P < 0.05
      ** = significantly positive, P < 0.01
```

Fig. 2. Output of branch length and approximation confidence limits

Fig. 1 points out that the closest distance to the palm civet is GZ 02/18/03 indicating that the beginning of the SARS epidemic spread is in Guangzhou. This is consistent with the results of previous studies using distance method [6][10]. However, there is a little difference in the study of the spread of SARS epidemic with neighbor joining algorithm which showed that the closest distance to palm civet is GZ16/12/2002.

V. Conclusion

Phylogenetic construction using the maximum likelihood method can be applied to a case of the spreading of SARS epidemic. This probability-based method allows for the acquisition of many tree topologies and thus requires a heuristic search method to shorten the time to obtain phylogenetic trees with the highest likelihood value. The result of the research showed that the maximum likelihood value was 269135.56466. This value was used to determine which tree can explain a better sequence alignment. The Fig. 1 indicates that the initial spread of the SARS epidemic started from Guangzhou, because branch length GZ02/18/02 to Palm Civet was 0.00012 that shows closest distance with Palm Civet as host SARS virus.

Acknowledgments

This research was the outcome of a cooperative research between universities in 2018 financed by Directorate of Research and Community Service, Directorate General for Research and Development, Ministry of Research, Technology and Higher Education (Kemenristekdikti) number 035/U/BU.2/A.4/IV/2018, April 2nd 2018.

References

[1] N. Cristianini and M. Hahn, *Introduction to Computational Genomics*. New York: Cambridge University Press, 2006.

[2] A Isaev, *Introduction to mathematical methods in bioinformatics*. 2004.

[3] R. Durbin, S. Eddy, a Krogh, and G. Mitchison, “Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids,” *Analysis*, p. 356, 1998.

[4] P. Lemey, M. Salemi, and A.-M. Vandamme, *The Phylogenetic Handbook; A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*, Second Edi. New York: Cambridge University Press, 2009.

[5] S. Shen and J. A. Tuszynski, *Theory and Mathematical Methods for Bioinformatics*. 2007.

[6] M. Isa Irawan and S. Amiroch, “Construction of phylogenetic tree using neighbor joining algorithms to identify the host and the spreading of SARS epidemic,” *J. Theor. Appl. Inf. Technol.*, vol. 71, no. 3, pp. 424–429, 2015.

[7] M. A. Marra *et al.*, “The genome sequence of the SARS-associated coronavirus,” *Science (80-.).*, vol. 300, no. 5624, pp. 1399–1404, 2003.

[8] P. A. Rota *et al.*, “Characterization of a novel coronavirus associated with severe acute respiratory syndrome,” *Science (80-.).*, vol. 300, no. 5624, pp. 1394–1399, 2003.

[9] Y. Guan *et al.*, “Isolation and characterization of viruses related to the SARS coronavirus from animals in Southern China,” *Science (80-.).*, vol. 302, no. 5643, pp. 276–278, 2003.

[10] S. Amiroch, M. S. Pradana, M. I. Irawan, and I. Mukhlash, “Multiple alignment analysis on phylogenetic tree of the spread of SARS epidemic using distance method,” *J. Phys. Conf. Ser.*, vol. 890, no. 1, 2017.

[11] T. Andriani and M. I. Irawan, “Application of unweighted pair group methods with arithmetic average (UPGMA) for identification of kinship types and spreading of ebola virus through establishment of phylogenetic tree,” *AIP Conf. Proc.*, vol. 1867, 2017.

[12] Thorne, “An Evolutionary Model For Maximum Likelihood Alignment of DNA Sequences,” *J. Mol. Evol.*, vol. 33, pp. 114–124, 1991.

[13] R. Mott, “Maximum-Likelihood Estimation of the Statistical Distribution of Smith-Waterman Sequence Similarity Scores,” *Bull. Math. Biol.*, vol. 54, no. 1, pp. 59–75, 1992.

[14] S. Serdoz *et al.*, “Maximum likelihood estimates of pairwise rearrangement distances,” *J. Theor. Biol.*, vol. 423, pp. 31–40, 2017.

[15] Y. Viniotis, *Probability and Random Process*. Singapore: Mc. Graw-Hill, 1998.

[16] S. M. Ross, *Introduction to Probability Models: Tenth Edition*. 2009.

[17] Swofford, Olsen, Waddell, and Hillis, *Molecular Systematics*. 1996.

[18] J. J. Cai, D. K. Smith, X. Xia, and K. Y. Yuen, “MBEToolbox: A Matlab toolbox for sequence data analysis in molecular biology and evolution,” *BMC Bioinformatics*, vol. 6, 2005.