Journal of Theoretical and Applied Information Technology
31st January 2015. Vol.71 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

CONSTRUCTION OF PHYLOGENETIC TREE USING NEIGHBOR JOINING ALGORITHMS TO IDENTIFY THE HOST AND THE SPREADING OF SARS EPIDEMIC

[1] MOHAMMAD ISA IRAWAN, [2]SITI AMIROCH

[1]Institut Teknologi Sepuluh Nopember (ITS)
Faculty of Mathematics and Natural Sciences
Department of Mathematics

[2]Darul Ulum Islamic University
Faculty of Mathematics and Natural Sciences
Department of Mathematics

E-mail: [1]mii@its.ac.id, [2]amirast_117@yahoo.com

ABSTRACT

The epidemic of SARS is a dramatic example of how quickly the virus spread throughout the world. The pattern of spread of the SARS virus can be likened to a tree called a phylogenetic tree. All of SARS virus related to one another are connected to a network that is transmitted from one individual to another. The results of this branching process to determine the beginning of the epidemic as well as the host of the corona virus.

To construct a phylogenetic tree takes inputs in the form of a distance matrix. In this study the distance matrix is derived from the alignments between sequence by using Super Pairwise Alignment (SPA). The output of the alignment is in the form of number of differences of sequences that determine in genetic distance. From the distance matrix, the genetic distance is converted to evolutionary distances using the Jukes Cantor model of which was established by phylogenetic tree by using the neighbor joining algorithm. From the results of the phylogenetic tree formation are known as host SARS Co-V is a palm civet with evolutionary distance of 0.1314. The epidemic began on December 16, 2002 in Guangzhou of South China and then spread to Zhongshan. Rallying point of the epidemic in Guangzhou Hospital and the Metropole hotel subsequently spread to Hanoi, Toronto, Singapore, Taiwan, and Hong Kong so that SARS to be international case. Furthermore, from the result of alignments of hosts and human SARS Co-V with SPA is known that the mutation is a mutation that occurs with type I protein replacement at position 77, 139, 147, 244, 344, 360, 472, 480, 487, 577, 609, 613, 665, 765, 778, and 1163.

Keywords: *SARS Coronavirus, Phylogenetic Tree, Neighbor Joining Algorithm, Super Pairwise Alignment.*

1. INTRODUCTION

Allegations of SARS (Severe Acute Respiratory Syndrome) which was originally caused by influenza virus because the symptoms are similar to the symptoms of people with influenza breakable because influenza viruses are not isolated from the patient. Finally, it was found that most of SARS patients were isolated by Corona virus when it is seen from its morphology. Virus of genome sequence is then analyzed and confirmed that the cause of SARS is Corona virus that has changed (mutation) of Corona viruses are around for this [9].

The pattern of spread of the SARS virus during the winter and spring of 2003 can be described as a tree. All of SARS virus related to one another is a branched network of relationships that are transmitted from one individual to another. The results of this process of branching of tree diagram are to determine the beginning of the epidemic and the host of the Corona virus.

Along with technological advances in the technology of DNA sequences, it was concluded that diagram of the phylogenetic tree is a task that is associated with DNA analysis and is based on a comparison of the whole genome between different SARS virus. The diagram of this tree is known to

Journal of Theoretical and Applied Information Technology
31st January 2015. Vol.71 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

be able to be analyzed of comparative DNA sequence of an organism because it is always changing and mutating [1].

Bioinformatics as a science that studies the application of computational techniques to manage and to analyze biological information includes the application of the methods of mathematics, statistics, and informatics to solve biological problems, especially with the use of DNA sequences and amino acid as well as information related to it [12]. In this case, bioinformatics plays an important role in the process of reading of the Corona virus genome. Then it searches for sequence similarity (homology alignment) with other human Corona virus that is carried by chickens, birds, dogs, pigs, rats, cats and ferrets. Further results of analysis is known that the genome of Corona virus causes SARS Corona virus is different from the other, so that the virus is called a virus SARS (SARS Co-V). Furthermore it is formed phylogenetic tree to identify the host and to determine the spread of the epidemic. This study using the neighbor joining method with the input of distance matrix which is generated from protein sequence alignments of human SARS Co-V and SARS virus sequences from several animals suspected with *Super Pairwise Alignment*. Having obtained the phylogenetic tree, the virus in the closest branching to the human SARS Co-V will be analyzed for mutations that happened to him.

2. REVIEW OF RELATED LITERATURE
3.

2.1 Sequence Alignment

Sequence alignment is an important method in the analysis of the position and type of mutation that is hidden in biological sequences and allow appropriate comparison is done. The most important thing in *a sequence alignment* is to determine the displacement of mutation [12]. Given two sequences A_1 and A_2 which are defined as follows:

$A_1 = (a_{11}, a_{12}, \dots, a_{1n_a})$, $A_2 = (a_{21}, a_{22}, \dots, a_{2n_a})$

Insertion of symbol "-" to the *sequences* A_1 and A_2, and aims to establish two new *sequences*, namely, the *sequences* A_1' and A_2'. Furthermore, the elements of the *sequences* A_1 and A_2 that represent the DNA sequence having the range V_5 = (0,1,2,3,4) or (a, c, g, t,-). As for the protein sequence has the range V_{21} = {A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V, - }.

2.2 Super Pairwise Alignment

Super Pairwise Alignment or SPA predicts the existence of *insertions* or deletions and long *insertions* that rely on *sequence local similarity.*

2.3 Model Jukes Cantor

Model of Jukes Cantor is one of the simplest model of evolution of sequence that assumes that the four nucleotide has the same frequency, and that all substitutions are equally likely [5]. In this model the distance between two nucleotide sequences is given by:

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

Where p is the proportion of different nucleotides in the two sequences.

As for the protein sequence, the model used is:

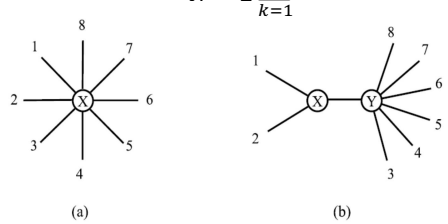
$$d = -\frac{19}{20} ln\left(1 - \frac{20}{19}p\right)$$

2.4 Neighbor-Joining Algorithm

Neighbor-joining method is a distance-based method used to construct phylogenetic tree [10]. For the set of N sequences, computational steps are given as follows:

1. Determine the distance matrix of N sequences.
2. Assume a tree with all OTU (operational taxonomic units) in the matrix as branching from a central point, then form of the *star-like* pattern as in the schematic representation in Figure 2.2a.
3. For each OTU, compute S, where S is the sum of the distance (D) between the OTU one with other OTU, divided by (N-2), where N is the total number of OTU.

$$S_i = \frac{1}{N-2} \sum_{k=1}^{N} D_{ik}$$



Picture. 2.2 a, b. Neighbor-joining. a. initial structure of star-like b. structure of tree-like after joining nodes 1 and 2.

4. Identify OTU pair with the minimum value:

$$M_{ij} = D_{ij} - S_i - S_j$$

5. Combine 2 taxa at a node in a sub-tree
6. Calculate the length of the branch:

$$D_{xi} = \frac{(D_{ij} + S_i - S_j)}{2}$$

$$D_{xj} = \frac{(D_{ij} + S_j - S_i)}{2}$$

7. Calculate a new distance matrix by connecting i and j and replace it with a node (x) which connect:

Journal of Theoretical and Applied Information Technology
31st January 2015. Vol.71 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$D_{xk} = \frac{(D_{ik} + D_{jk} - D_{ij})}{2}$$

8. Repeat the above steps until N = 3.

3. ANALYSIS AND DISCUSSION

3.1 Results of Phylogenetic Tree Formation To Identify Host

Before constructing phylogenetic tree, all of protein sequences are aligned first with the SPA method to determine the difference (genetic distance). Furthermore, genetic distances represented in the form of a distance matrix is first converted into a matrix of evolutionary distance with Jukes Cantor method. Evolutionary distance matrix is formed, to be input that is used for the process of the formation of the phylogenetic tree.

Table 3.1 Matrix Of Evolutionary Distance Cycle 1

	A	B	C	D	.....	L
A	0	0,119	2,205	1,188		0,458
B	0,119	0	1,200	1,192		0,459
C	2,205	1,200	0	0,131		1,218
D	1,188	1,192	0,131	0		1,212
E	1,361	1,444	1,401	1,392		1,387
⋮						
L	0,458	0,459	1,218	1,213		0

a. Step 1:
Obtained results of the calculation of each Si is:
S_A = 1.08618 ; S_B = 1.0002 ; S_C = 1.3991 ;
S_D = 1.2942 ; S_E = 1.42744 ; S_F = 1.4156 ;
S_G = 1.4312 ; S_H = 1.4889 ; S_I = 0.95125 ;
S_J = 0.9172 ; S_K = 0.9158 ; S_L = 0.9252.

b. Step 2:
Wanted minimum value for each pair of sequences:

$$M_{ij} = D_{ij} - S_i - S_j$$

M is the smallest pair obtained M_{CD} = -2.5620

c. Step 3:
Define a new OTU, namely, U1 replaces the smallest pair (C and D). Furthermore, these taxa are combined as U1 following formula:

$$s_{CU_1} = 0.5(D_{CD} + S_C - S_D) = 0.11815$$
$$s_{DU_1} = 0.5(D_{CD} + S_D - S_C) = 0.01325$$

d. Step 4:
Connect taxa U1 with C and U1 with D, each with an *edge* length or distance follow as the result of the calculation in step 3.

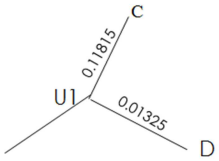


Figure 3.1 Tree on cycle 1

e. Step 5:
Combine the new distance of all of taxa to U1 The result of the new distance U1 to all taxa for subsequent inclusion in a new evolutionary distance matrix.

The next step is the same calculation as in cycle 1, the value is different for different N, the smallest Mij is different, the new distance to the end of each taxa is also different.

In the same way as in cycle 1 was obtained tree for each cycle as follows:

Figure 3.2
Tree on cycle 2

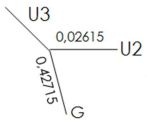


Figure 3.3
Tree on cycle 3

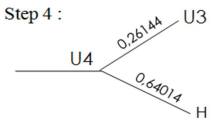


Figure 3.4
Tree on cycle 4

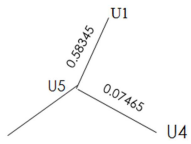


Figure 3.5
Tree on cycle 5

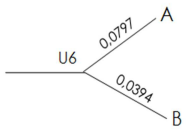


Figure 3.6
Tree on cycle 6

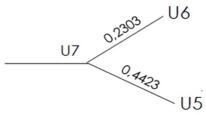


Figure 3.7
Tree on cycle 7

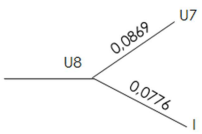


Figure 3.8
Tree on cycle 8

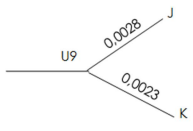


Figure 3.9
Tree on cycle 9

Journal of Theoretical and Applied Information Technology
31st January 2015. Vol.71 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

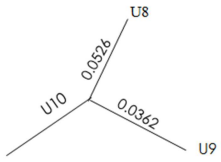


Figure 3.10 Tree on cycle 10

And the last branching is U10 to L.

After all branches combined, phylogenetic trees obtained the following:

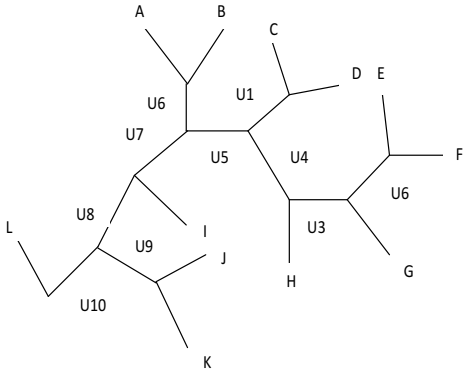


Figure 3.11 Phylogenetic Tree Of SARS Co-V Host

When depicted in the form of a Cladogram are as follows:

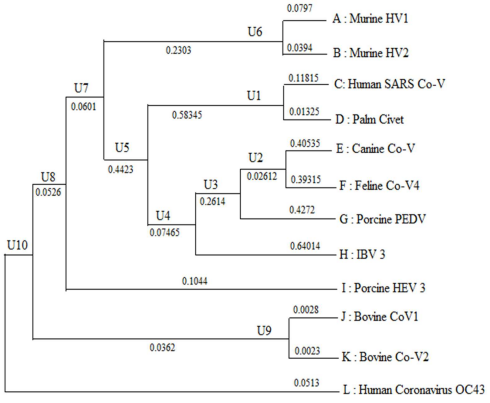


Figure 3.12: Phylogenetic Tree Of Host SARS Co-V Results Of Manual Workmanship.

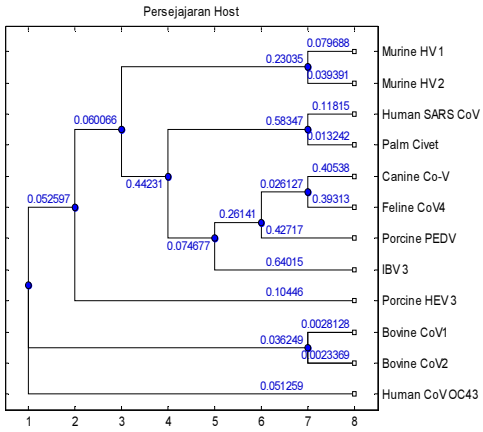


Figure 3.13: Phylogenetic Tree Of Host SARS Co-V Are Simulated In Matlab.

The phylogenetic tree in Figure 3.12 and 3.13 are explaining kinship between species to be estimated as the host of the Corona virus. Each *leaf* (A, B, C, D, E, F, G, H, I, J, K, L) in the *tree* is also called OTU (Operational Taxonomy Unit) which represents the aligned species. In the images appear closest branching is between OTU A with B, OTU C with D OTU, OTU E to F, and OUT J with K.

At the picture 3.12 or picture 3.13, OTU C representing Human SARS Co-V are on the same branch with the OTU D, namely, Palm Civet. Human SARS Co-V is Corona virus that attack humans, while Palm Civet is a Himalayan palm civet. Both of evolutionary is distance of 0.1314. This figure can be seen from a distance summation of branching of Palm Civet to *branch U1* (0.01325) with a distance of Human SARS Co-V to *branch U1* (0.11815). By looking at the results of phylogenetic tree analysis turns SARS Co-V relationship of evolution is closest to Palm Civet. This indicates that Palm Civet is the host to be sought in this case.

Based on the phylogenetic tree analysis above can be concluded that the evolutionary distance that is closest to the human SARS Co-V is the Palm Civet. This is because the protein sequences in Palm Civet when it is juxtaposed with the human protein sequences of SARS Co-V have been mutated as much as 0.1314 which indicates the smallest number when it is compared with the protein mutation that occurs in other host sequences. And the smallest evolutionary distance indicates the host is sought.

In Figure 3.13 the simulated phylogenetic tree in matlab gives the same results with the results of manual calculation. It looks all *leaf* and *branch* was formed in the same position. The value of the

Journal of Theoretical and Applied Information Technology
31st January 2015. Vol.71 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

evolutionary distance between species also showed the same value as the manual calculation and computation, both of them used neighbor joining algorithm with Jukes Cantor correction model.

3.2 Analysis Of Results Of Human Alignment Sars Co-V With Palm Civet

From the resulting phylogenetic tree, it is known that the host of the SARS epidemic is Palm Civet.

The sequence alignment 3 which represents Human SARS Co-V (Corona viruses that infect humans) and sequences 4 representing Palm Civet (Corona virus that attacks the weasel), both of them have the same length of sequence is 1255aa. The difference of protein of both of is 16 with length of alignment 1255aa, this indicates the genetic distance of two strains of the virus by 0.0128 which is the quotient between different genetic with length of its alignment. In the alignment occurs mutation of type I (replacement of protein) and occurs in the following positions:

a. position 77 mutation of protein G to D,
b. position 139 mutation of protein A to V,
c. position 147 mutation of protein Q into R,
d. position 244 mutation of protein I to T,
e. position 344 mutation of protein K into R,
f. position 360 mutation of protein F to S,
g. position 472 mutation of protein L to P,
h. position 480 mutation of protein D to G,
i. position 487 mutation of protein T to S,
j. position 577 mutation of protein A to S,
k. position 609 mutation of protein A to L,
l. position 613 mutation of protein D to E,
m. position 665 mutation of protein L to S,
n. position 765 mutation of protein A to V,
o. position 778 mutation of protein Y to D,
p. position 1163 mutation of protein K into E.

Alignment above when it is compared with the BLAST software will show the results as follows:

Table 3.2: Comparison of alignment of Human SARS Co-V with Palm Civet

	SPA	BLAST
▪ Alignment	Global alignment	Local Alignment
▪ The level of homologous	87,3 %	99 %
▪ Gap	0 %	0 %
▪ Genetic Distance	0.0127	0.01

3.3 Results Of Phylogenetic Tree Establishment For Spreading Sars Epidemic

Here is Phylogenetic tree with neighbor joining algorithm and Jukes Cantor model for the spread of the SARS epidemic of the simulation results from Matlab.

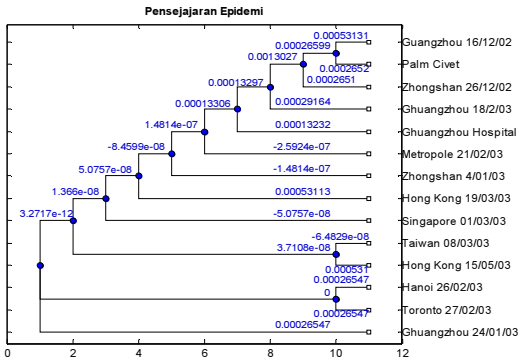


Figure 3.16: Phylogenetic tree for the spread of SARS epidemic

Phylogenetic trees above depict the story of an epidemic. Palm Civet is included in the alignment because it has been known as the host. Infection begins in Guangzhou (16/12/2002) and Zhongshan (12/26/2002), respectively it is abbreviated GZ and ZS. The epidemic began to spread in Guangzhou hospital (31/01/2003) as indeed all coronavirus-infected patients being treated there. Further it spreads to Zhongshan (01/04/2003), Guangzhou (18/02/2003), and the hotel Metropole (02/21/2003). Doctor Carlo Urbani is assigned to the Guangzhou hospital to treat SARS patients who stay at the hotel Metropole eventually contracting as well, and the other guests staying at the hotel is going home to their countries with the disease and Corona viruses on them. From there, the virus eventually spreads to Hanoi (02/26/2003), Toronto (02/27/2003), Singapore (03/01/2003), Taiwan (03/08/2003), Hong Kong (19/03/2003) and (15/05/2003) that the case of SARS to be an international outbreak.

Phylogenetic tree was simulated in matlab and results of running software clustal W2 was different because in the neighbor joining algorithm used was without distance correction. The results shown are also different from the results of running matlab program due to some residual value has the smallest *Mij* more than one. Selection of different *Mij* values will also result the different phylogenetic tree. The value of the distances between sequences on the results of running software clustal w in some strains showed the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

number 0, but the results of the matlab indicate a higher level of accuracy.

4. CONCLUSION

From the discussions that have been done, the conclusions obtained are as follows :

a. Based on the data processing of the protein sequences of various animals were suspected of hosts of SARS Corona virus and the DNA sequence of data of SARS to be infected patients were taken according to the location and date of the sample, can be proven that the phylogenetic tree constructed by using the neighbor joining algorithm with Jukes Cantor models used successfully to identify hosts and spread of the SARS epidemic.
b. The application of alignment SPA (Super Pairwise Alignment) is successfully applied to align the sequences of human SARS Corona virus with another carried by animals. From the results of alignment of hosts and human SARS Co-V with SPA is known that the mutation that occurs is a mutation of type I with the protein replacement at positions of 77, 139, 147, 244, 344, 360, 472, 480, 487, 577, 609, 613, 665, 765, 778, and 1163.
c. To construct the phylogenetic tree of the data sequence which is very high similarity is not recommended to use neighbor joining algorithm as it will be obtained several different tree (tree is not stable).

REFRENCES:

[1] Christianini, N., Hahn, M.W, 2006, *Introduction to Computational Genomics A Case studies Approach,* Cambridge University Press, New York.

[2] Durbin R., Eddy S.R., Krogh A., Mitchigan G., 1998, *Biological Sequence Probabilistic Models of Proteins and Nucleic Acids,* Cambridge University Press.

[3] Guan, et al., 2003, *Isolation and Characterization of Viruses Related to the SARS Coronavirus from Animals in Shouthern China.* Science 302, 276 (2003), The American Association for the Advancement of science, Washington, (www.sciencemag.org), *accessed on 29 January 2014*.

[4] Isaev, A., 2006, *Introduction to Mathematical Methods in Bioinformatics*, Springer.

[5] Lemey P, Salemi M, Vandamme M, 2009, *The Phylogenetic Handbook: a Practical Approach to Phylogenetic Analysis and Hypothesis Testing.* Cambridge University Press.

[6] Marra, et al., 2003, The Genome Sequence of the SARS-Associated *Coronavirus.* Science 300, 1399 (2003), The American Association for the Advancement of science, Washington, (www.sciencemag.org), *accessed on 29 January 2014* .

[7] Narshingh Deo, 1989, *Graph Theory with Application to Engineering and Computer Science.* Prentice Hallof India Private Limited, New Delhi.

[8] Page and Holmes, 1998. *Molecular Evolution*: *A Phylogenetic Approach,* Blackwell Publishing Ltd.

[9] Rota, et al., 2003, *Characterization of a Novel Coronavirus Associated with Severe Acute Respiratory Syndrome*. Science 300, 1394 (2003), The American Association for the Advancement of science, Washington, (www.sciencemag.org), *accessed on 29 January 2014* .

[10] Saitou N, Nei M, *The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees.* Mol. Biol. Evol. 4(4), 1987, PP. 406-425.

[11] *Severe Acute Respiratory Syndrome (SARS)*, Pubmed Health, (http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0004460), *accessed on 29 January 2014.*

[12] Shen, S.N., (2007), *Theory and Mathematical Methods for Bioinformatics,* Biological and Medical Physics, biomedical Engineering, Springer.