JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

An alignment free mathematical model based on fuzzy dissimilarity metric to infer phylogenetic tree among Ebola viruses

Adamu Yusha'u[1*], Vijendra Singh Rawat[2] , Rinku Mathur [3], M S. Mahmud[4]

[1]Department of mathematical Science, Faculty of science, Bauchi State University, Gadau, Nigeria.
[2]Department of Mathematic, Faculty of Engineering, Teerthanker Mahaveer University, U.P., India.
[3]Department of Rural Development and Panchayati Raj, Jammu and Kashmir India
[4]Department of mathematics and statistics, federal university of kashere, Gombe state.

Abstract:

In this paper, we developed an alignment-free method for examining the similarity and evolutionary divergence of Ebola virus sequences taken from five African countries for faster sequence analysis. The notion of "similarity" as described here is essentially a generalization of the equivalence notion. The method utilized the use of math-lab programming to generate the number of frequencies of occurrence of all nucleotide base from each DNA sequences and then obtain a dissimilarity matrix that shows; the smaller is the element, the more similar the sequences of Ebola viruses are. The dissimilarities obtained are then converted to similarity matrix so using a mathematical concept of fuzzy transitive relation, at last the transitive closure has been used to get phylogenetic tree along with Ebola viruses.

Keywords: Ebola viruses, Phylogenetic tree, DNA sequences, Euclidean distance, Fuzzy dissimilarity matrix.

AMS Subject Classification: 90C05, 92B05, 92B10, 92D15, 92D20.

1. Introduction

Biological systems are extremely complex due to adaptability, adaptation, stability, and robustness. Much biological data is produced by not well understood biological processes. Interpretation of such data includes discovery of interconnected relationships hidden in the data. The accuracy of prediction or the knowledge extracted from a database is often unsatisfactory because of these challenges. Clearly there is ample room for further research, which requires new theoretical frameworks and computational techniques[1].These challenges serve as driving force for our research, Ebola virus (EBOV) is enveloped filo virus with 14,000 nanometers in length with a diameter of 80 nanometers and is not transcript into DNA and its single-stranded RNA genome that causes sporadic Outbreaks of lethal hemorrhagic fever in humans.

There exist a transmission between human and bats or apes, so there is need to carry sequence analysis because in West Africa is the deadliest occurrence of the disease since its discovery in 1976 [2-4]. We considered the sequences of Ebola virus in different African countries from "Gen bank" and construct a phylogenetic tree to identify the relatedness among the type of Ebola virus and it is also stated that Ebola virus is filo virus which causes sporadic outbreak[5]**.** The Ebola virus disease (EVD) is named after the Ebola River valley in Zaire in Democratic Republic of Congo where it occurs earliest in 1976 from then other outbreaks have been discovered in many parts of Central Africa, the extreme rising of the case was observed recently

around 2014 in some West African countries like Sierra Leone, Liberia, Nigeria, Guinea and Senegal. In almost all the outbreaks the initial infection is due to contacts with infected animals (hunted for food) such as fruit bats and primates (ape, chimpanzee, monkey,) this highlights the need to put into account the indirect contacts with the surrounding environment as a transmission route of the virus.

To understand the transmission dynamics of Ebola virus is advance to incorporate vaccination and change of behavior for self–protection of susceptible individual is important[6]. It was also discovered that EBOV is a human pathogen in 1976 but the high case-fatality rate and self-limited nature of EVD outbreaks suggest that EVD is a zoonosis. To get the broadest possible perspective of Ebola virus sequences evolution, we considered the Sequences of Ebola virus from T. Andriani and M. I. Irawan and construct a phylogenetic tree to identify the relatedness among the type of Ebola virus. Furthermore, the phylogenetic tree constructed by UPGMA method, sort of numerical coding method of DNA sequences, dynamic programming[6-7]. But all these methods require multiple alignment of the sequence and presume some kind of evolutionary model among them. In addition to multiple alignment problems like computational complexity and inherent uncertainty of alignment cost criteria, these methods are not providing the complete information of phylogeny. Furthermore, number of alignment-free method are reported for phylogenetic analysis in the literature[8– 13] while the literature[14-15] propose a new way of representing genetic sequences as fuzzy sets in the I12 space, expanding the approach originally introduced by Torres & Nieto and Sadegh-Zadeh & simple method to analyze the similarity of biological sequences. By taking the average contents of biological sequences and their information entropies as the variables, the fuzzy method is used to cluster them.

Here, a new method an alignment free mathematical model based on frequency dissimilarities for inferring phylogenetic tree among Ebola viruses has been proposed. An alignment-free method for examining the similarity and evolutionary divergence of Ebola virus sequences taken from five African countries for faster sequence analysis. The notion of "similarity" as described here is essentially a generalization of the equivalence notion. The method utilized the use of math-lab programming to generate the number of frequencies of occurrence of all nucleotide base from each DNA sequences and then obtain a dissimilarity matrix that shows; the smaller is the element, the more similar the sequences of Ebola viruses are. These dissimilarities obtained are then converted to similarity matrix so using a mathematical concept of fuzzy transitive relation. At last, the transitive closure has been used to establish a phylogenetic tree. Unlike other methods such as the graphical representation methods, which is usually very complex to compute some invariants of matrix derived from graphical representation, our method has less computational burden, simple algorithms and easy to implement for any DNA sequence analysis efficiently.

2. Methodology of the Approach

The methodology of the proposed model is presented in the following steps

Step I: Four type of Ebola virus namely Sudan Ebola virus, Zaire Ebola virus, Tai Forest Ebola virus and Bundibugyo Ebola virus also known as Cote d'Ivoire Ebola virus were considered in the studies. The eleven complete genomes of the four samples we are interested in were generated from [3].

Step 2: From the generated DNA sequences, a unique feature known as frequency of nucleotide base were identified and computed.

Step 3: Four different types of frequencies termed as frequency of Adenine, Cytosine, Guanine, and Thymine designated as "A", C, G, T respectively were identified and used in forming frequency table

Step 4: Formation of frequency table can be access in the result and discussion section of this work.

Step 5: From table 1.2, we computed the Euclidean distances among eleven (11) Ebola virus sequences.

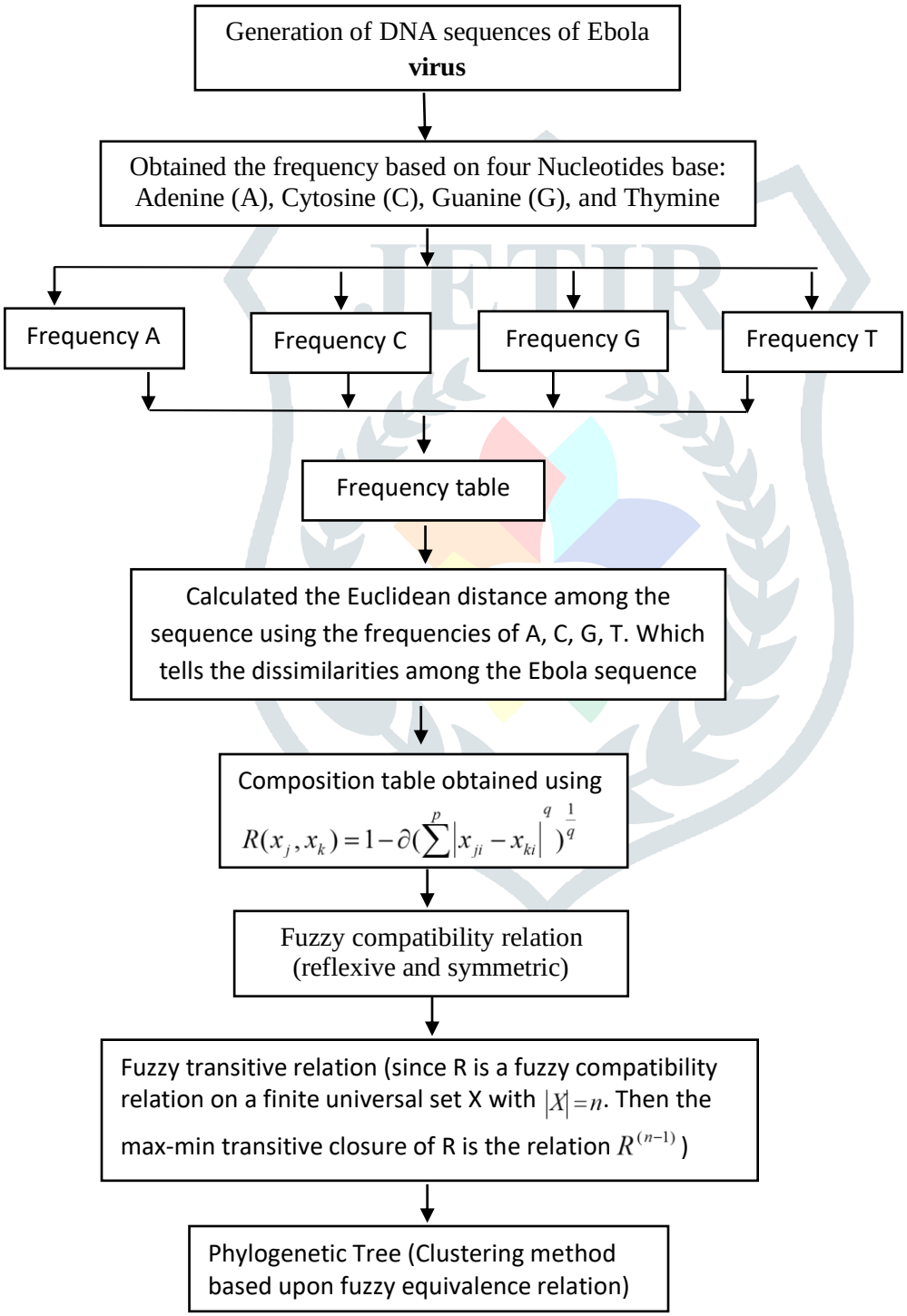
Step 6: Composition table was formed using fuzzy compatibility relationship R which is define on X in terms of the correct distance function of the Minkowski class using the relation

$R(x_j, x_k) = 1 - \partial(\sum_{i=1}^{p} |x_{ji} - x_{ki}|^q)^{\frac{1}{q}}$ where $R(x_j, x_k) \in X$, $q \in \bar{R}$ and ∂ is a constant that constrains $R(x_j, x_k) \in [0, 1]$, clearly, ∂ is the reverse value of the largest distance in X. In general, R is a fuzzy compatibility relation, but not necessarily a fuzzy equivalence relation. Hence, there is need to determine the transitive closure of R in the next step.

Step 7: Fuzzy Transitivity closure testing on $R(x_j, x_k)$ is done on the finite set X to ascertain equivalence relation

Step 8: Phylogenetic tree was formed based on clustering method obtained from fuzzy equivalence relation

Flowchart of the proposed model



2.1 Constructing Phylogenetic the Tree

The fuzzy compatibility relationship R obtained can be defined on X in terms of the correct distance function of the Minkowski class by the formula [16].

$$R(x_j, x_k) = 1 - \partial(\sum_{i=1}^{p} |x_{ji} - x_{ki}|^q)^{\frac{1}{q}} \dots\dots\dots\dots\dots\dots (1)$$

For all pairs, $R(x_j, x_k) \in X$ where $q \in \bar{R}$, ∂ is a constant that ensures that $R(x_j, x_k) \in [0, 1]$, clearly, ∂ is the reciprocal of the largest distance in X. In general, R defined by equation (1) is a fuzzy compatibility relation, but not necessarily a fuzzy equivalence relation. Hence, there is need to determine the transitive closure of R.

Given a relation $R(X, X)$, its transitive closure $\bar{R}(X, X)$ can be determined by simple algorithm that consists of the following three steps:

R' = R U (R o R)
If R' ≠ R, make R = R' and go to step I
Stop R' = $\bar{R}$

This algorithm applies to both smooth and fuzzy relationships, the form of composition and set union in step I must be however compatible with the definition of transitivity employed. When max-min composition and the max operator for set union are used, we called $\bar{R}$the transitive max-min closure [15].

The max–min product of the relation matrices for M and N. A similarity relation P in X is a fuzzy relation in X which is reflexive, symmetric, and transitive [17]. We can see that the concept of a similarity relation is essentially a generalization of the concept of an equivalence relation.

More specifically, if we define $\bar{R} = (\bigcup_{i=1}^{n} R^i)$ transitive closure of any fuzzy relation $\bar{R}$, then $\bar{R}$ is transitive. If R is a fuzzy relation characterized by a relation matrix of order n. Now, we can construct a transitive relation from any relation characterized by a relation matrix of order n. In fact, for any such fuzzy relation R, we compute R^ksuccessively. according to our procedure above we achieve $\bar{R} = R^K$. By applying above procedure, analyze the data for q= 1, 2. In equation (1), firstly, for q=2, which is corresponds to the Euclidean distance [18] represented by

$$E_{i,j} = \sqrt{(a_i - a_j)^2 + (c_i - c_j)^2 + (g_i - g_j)^2 + (t_i - t_j)^2}$$

In every fuzzy relation R can be uniquely represented in terms of it's $\alpha - cuts$ by the formula $R = \bigcup \alpha.^{\alpha} R\ \alpha \in [0,1]$ for any value$\alpha \in [0,1]$, create a crisp equivalence relation that represents the presence the similarity between the elements to the degreeα. Each of these equivalence relation forms a partition of X. Let $\pi(^{\alpha}R)$denote the partition corresponding to the equivalence relation $(^{\alpha}R)$. we may say two elements x and y belongs to the same block of this partition if and only if $R(x, y) \geq \alpha$.

Level set of $R\ \wedge_R = \{0.3506, 0.7148, 0.7195, 0.8887, 0.9485, 0.9511, 0.9511, 0.9675, 0.9875, 0.9919\}$
R is associated with a sequence of eleven nested partition $\pi(^{\alpha}R)$for $\alpha \in \wedge_R$ and $\alpha > 0$
The hierarchical clustering method has been used on Fuzzy equivalence relation to obtain the desired phylogenetic tree, we determine a fuzzy comparability relation in term of Euclidean distance and also hamming distance then we apply the algorithms for obtaining nested sequence for $\alpha - cut$.

3. Data Materials

3.1 Materials

Firstly**,** we have five type of Ebola virus namely Sudan Ebola virus, Zaire Ebola virus, Tai Forest Ebola virus, Bundibugyo Ebola virus also known as Cote d'Ivoire Ebola virus and Reston Ebola virus. Due to the fact that Reston Ebola virus does not infect humans, but attacks the monkey as the result of that we excluded it and consider the eleven complete genome which constitute the four other type of Ebola virus from African country were the disease was originate which later was spread to the other part of the globe.
We obtained these sequences from **[3]** which listed in the table 1below:

Table 1: **information of eleven sequences of Ebola virus**

4. Results and Discussion

The purposed method has been illustrated and validated by constructing phylogenetic tree for eleven sequences (table 1) taken into consideration. All the sequences are longer than 18000 bases presented in the table 2 that gives the frequency of each sequence of Ebola virus based on it four nucleotide bases i.e. $F_{i_{E_i a_i}}, F_{i_{E_i c_i}}, F_{i_{E_i g_i}}, F_{i_{E_i t_i}} \quad i = 1,2, \dots 11$

S. No.	Name	Accession Code	Length	Year	Location
1	Zaire Ebola virus	NC_002549.1	18959	1976	Congo
2	Zaire Ebola virus	KC242792.1	18959	1994	Gabon
3	Tai Forest Ebola virus	NC_014372.1	18935	1994	Ivory Coast
4	Sudan Ebola virus	NC_006432.1	18875	2004	Uganda
5	Zaire Ebola virus	KC242790.1	18958	2007	CongIo
6	Bundibugyo Ebola virus	NC_014373.1	18940	2007	Uganda
7	Zaire Ebola virus	KT013259.3	18958	2014	Guinea
8	Zaire Ebola virus	KP271020.1	18861	2014	Congo
9	Zaire Ebola virus	KM233042.1	18912	2014	Sierra Leone
10	Zaire Ebola virus	KP178538.1	18958	2014	Liberia
11	Sudan Ebola virus	EU338380.1	18875	2014	Sudan

Stage1: for calculating frequency (occurrences) of nucleotide's base present in DNA sequence

Algorithm 1:

1. For $i \in \{1,2, \dots 11\}$ do
2. $E_{i,1}$=$frequencyof_{a(file_i)}$
3. $E_{i,2}$=$frequencyof_{c(file_i)}$
4. $E_{i,3}$=$frequencyof_{g(file_i)}$
5. $E_{i,4}$=$frequencyof_{t(file_i)}$
6. end do

S. No	symbol	Name	Accession Code	Length	Frequency of 'A'	Frequency of 'C'	Frequency of 'G'	Frequency of 'T'
1	E1	Zaire Ebola virus	NC_002549.1	18959	6061	4035	3752	5111
2	E2	Zaire Ebola virus	KC242792.1	18959	6047	4052	3756	5104
3	E3	Tai Forest Ebola virus	NC_014372.1	18935	6020	4371	3630	4914
4	E4	Sudan Ebola virus	NC_006432.1	18875	5920	4071	3732	5152
5	E5	Zaire Ebola virus	KC242790.1	18958	6060	4025	3751	5122

6	E6	Bundibugyo Ebola virus	NC_014373.1	18940	5964	4324	3632	5020
7	E7	Zaire Ebola virus	KT013259.3	18958	6050	4050	3758	5100
8	E8	Zaire Ebola virus	KP271020.1	18861	6013	4051	3733	5062
9	E9	Zaire Ebola virus	KM233042.1	18912	6037	4045	3743	5087
10	E10	Zaire Ebola virus	KP178538.1	18958	6051	4052	3755	5100
11	E11	Sudan Ebola virus	EU338380.1	18875	5914	4032	3750	5179

Table 1: Frequencies obtained for 11 Ebola virus sequence based on number of 4 nucleotide bases

Stage2: Calculation for the Euclidean distance among eleven Ebola virus Sequences

Algorithm 2: **For i^{th} Euclidean distance**

1. For $i \in \{1,2,\dots 11\}$do
2. For $j \in \{1,2,\dots 11\}$do
3. For $D_{i,j} = sqrt((E_{i,1} - E_{j,1})^2 + (E_{i,2} + E_{j,2})^2 + (E_{i,3} + E_{j,3})^2 + (E_{i,4} + E_{j,4})^2)$
4. end for
5. end for

Table 2: Dissimilarities obtained by using Euclidean distance among Ebola virus sequences

$E_{i,j}$	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}	E_{11}
E_1	*0*	23.4521	410.203	152.57	14.9332	340.012	22.4277	72.9520	36.5103	22.7816	162.002
E_2	23.4521	*0*	393.026	139.11	35.3129	321.410	5.7446	58.7367	24.6374	5.7446	154.114
E_3	410.203	393.026	*0*	408.71	423.340	128.788	393.601	367.377	386.347	391.073	459.110
E_4	152.507	139.171	408.711	*0*	151.587	305.564	305.564	130.950	136.782	144.060	51.0882
E_5	14.9332	35.3129	423.340	151.57	*0*	350.979	35.4683	82.5167	47.0956	36.1939	156.897
E_6	340.012	321.410	128.788	305.54	350.979	*0*	323.640	298.156	316.191	321.069	356.321
E_7	22.4277	5.7446	393.601	305.54	35.4683	323.640	*0*	58.6430	24.2487	3.7417	158.507
E_8	72.9520	58.7367	367.377	130.90	82.5167	298.156	58.6430	*0*	36.5650	58.0775	155.375
E_9	36.5103	24.6374	386.347	136.72	47.0956	316.191	24.2487	36.5650	*0*	23.6220	154.3081
E_{10}	22.7816	5.7446	391.073	144.00	36.1939	321.069	3.7417	58.0775	23.6220	*0*	154.301
E_{11}	162.002	154.114	459.110	51.082	156.897	356.321	158.507	155.375	154.301	154.301	*0*

We implemented the above algorithms using mat-lab programming and the corresponding dissimilarity matrix has been presented in table 3 which suggest that the smaller the distance between the two sequences of Ebola virus, the more similar are the sequence of Ebola virus. Now, it will be easy to find that the two genomes, E7/Zaire Ebola virus/18958/ KT013259.3/2014/Guinea and E10/Zaire Ebola virus/18958/ KP178538.1/2014 have a close linkage.

Stage 3: For calculating fuzzy compatibility relation among sequences

Algorithm3: For evaluating j^{th},k^{th} of $\boldsymbol{R(x_j, x_k)}$

1. $rho = 1/(max(max(D)))$;
2. $R = 1 - rho * D$;
3. For $i \in \{1,2,\dots 11\}$do
4. For $j \in \{1,2,\dots 11\}$do
5. $max\ i = min(R(i,1), (R(1,j))$;
6. For $k \in \{1,2,\dots 11\}$do
7. $temp\ min = min(R(i,1), (R(1,j))$;
8. $if(temp\ min > max\ i)\ max\ i = tempmin$;
9. end for
10. $composition(i,j) = max\ i$;
11. end for
12. end for

Using the formula in equation (1) i.e. $R(x_j, x_k)$, Obtained Fuzzy compatibility relation among sequences which is shown in table 4 below and denoted by R.

Table 3: Fuzzy compatibility relationships among Ebola virus sequences

R_{ij}	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}	E_{11}
E_1	1	0.9511	0.2594	0.7148	0.9675	0.3506	0.9511	0.9204	0.9485	0.9511	0.6678
E_2	0.951	1	0.2999	0.7148	0.9489	0.3506	0.9875	0.9204	0.9485	0.9875	0.6969
E_3	0.2594	0.2999	1	0.3344	0.2355	0.7195	0.2951	0.3506	0.3113	0.3007	0.2239
E_4	0.7148	0.7148	0.3344	1	0.7148	0.3506	0.7148	0.7148	0.7148	0.7148	0.8887
E_5	0.9675	0.9489	0.2355	0.7148	1	0.3506	0.9511	0.8974	0.9231	0.9504	0.6698
E_6	0.3506	0.3506	0.7195	0.3506	0.3506	1	0.3506	0.3506	0.3506	0.3506	0.3506
E_7	0.9511	0.9875	0.2951	0.7148	0.9511	0.3506	1	0.9204	0.9485	0.9919	0.6865
E_8	0.9204	0.9204	0.3506	0.7148	0.8974	0.3506	0.9204	1	0.9204	0.9204	0.7148
E_9	0.9485	0.9485	0.3113	0.7148	0.9231	0.3506	0.9485	0.9204	1	0.9485	0.7021
E_{10}	0.9511	0.9875	0.3007	0.7148	0.9504	0.3506	0.9919	0.9204	0.9485	1	0.6862
E_{11}	0.6678	0.6969	0.2239	0.8887	0.6698	0.3506	0.6865	0.7148	0.7021	0.6862	1

Now, our aims now is to get Fuzzy transitive relation, which can be obtained by performing max-min composition on Ri.e. $R^1 = R0R$ using this inequality "$2^k \geq n - 1$"where k is a positive integer value indicate the number of times to performed composition as stated in the Step-1 of the algorithms to get transitive closure, where n is the number of columns. In our study, we have n=11, therefore at most three compositions can be performed to reach to step III in the stated algorithms of transitive closure.

Now, first composition will be performed for k=1 and the resultant relationship will be obtained as $R^1 = R0R$.

Stage 4: Calculation of fuzzy transitive relationship by first composition for K=1

Algorithm 4: For evaluating j^{th},k^{th} of $R(x_j, x_k)$

1. repeat
2. $R = composition$;
3. For $i \in \{1,2,\dots 11\}$do
4. For $j \in \{1,2,\dots 11\}$do
5. $max\ i = min(R(i,1), (R(1,j))$;
6. For $k \in \{1,2,\dots 11\}$do

7. $temp\ min = min(\ R(i,1), (R(1,j))$;
8. $if\ (temp\ min > max\ i)\ max\ i = tempmin$
9. end for
10. $composition(i,j) = max\ i$;
11. end for
12. end for

Table5: Fuzzy transitive relationship among Ebola virus sequences for K=1

Table5: Fuzzy transitive relationship among Ebola virus sequences for K=1

R_{ij}	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}	E_{11}
E_1	1	0.9511	0.3506	0.7148	0.9675	0.3506	0.9511	0.9204	0.9485	0.9511	0.7148
E_2	0.9511	1	0.3506	0.7148	0.9511	0.3506	0.9875	0.9204	0.9485	0.9875	0.7148
E_3	0.3506	0.3506	1	0.3506	0.3506	0.7195	0.3506	0.3506	0.3506	0.3506	0.3506
E_4	0.7148	0.7148	0.3506	1	0.7148	0.3506	0.7148	0.7148	0.7148	0.7148	0.8887
E_5	0.9675	0.9511	0.3506	0.7148	1	0.3506	0.9511	0.9204	0.9485	0.9511	0.7148
E_6	0.3506	0.3506	0.7195	0.3506	0.3506	1	0.3506	0.3506	0.3506	0.3506	0.3506
E_7	0.9511	0.9875	0.3506	0.7148	0.9511	0.3506	1	0.9204	0.9485	0.9919	0.7148
E_8	0.9204	0.9204	0.3506	0.7148	0.9204	0.3506	0.9204	1	0.9204	0.9204	0.7148
E_9	0.9485	0.9485	0.3506	0.7148	0.9485	0.3506	0.9485	0.9204	1	0.9485	0.7148
E_{10}	0.9511	0.9875	0.3506	0.7148	0.9511	0.3506	0.9919	0.9204	0.9485	1	0.7148
E_{11}	0.7148	0.7148	0.3506	0.8887	0.7148	0.3506	0.7148	0.7148	0.7148	0.7148	1

Stage 5: Calculation of fuzzy transitive relationship by second composition for K=2

Algorithm5: for fuzzy transitive relation

1. repeat
2. $R = composition$;
3. For $i \in \{1,2, \dots 11\}$do
4. For $j \in \{1,2, \dots 11\}$do
5. $max\ i = min(\ R(i,1), (R(1,j))$;
6. For $k \in \{1,2, \dots 11\}$do
7. $temp\ min = min(\ R(i,1), (R(1,j))$;
8. $if\ (temp\ min > max\ i)\ max\ i = tempmin$
9. end for
10. $composition(i,j) = max\ i$;
11. end for
12. end for
13. until $composition <= R$

For k=2, performance of max-min composition on R^1 (i.e. $R^2 = R^1 o R^1$) has been made which realized the Step- IV of the algorithms for transitive closure (i.e. .$R^2 = \bar{R}$) and result obtained are shown in table 6.

Table6: Fuzzy transitive relationship among Ebola virus sequences for K=2

$R_{i,j}$	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}	E_{11}
E_1	1	0.9511	0.3506	0.7148	0.9675	0.3506	0.9511	0.9204	0.9485	0.9511	0.7148
E_2	0.9511	1	0.3506	0.7148	0.9511	0.3506	0.9875	0.9204	0.9485	0.9875	0.7148
E_3	0.3506	0.3506	1	0.3506	0.3506	0.7195	0.3506	0.3506	0.3506	0.3506	0.3506
E_4	0.7148	0.7148	0.3506	1	0.7148	0.3506	0.7148	0.7148	0.7148	0.7148	0.8887
E_5	0.9675	0.9511	0.3506	0.7148	1	0.3506	0.9511	0.9204	0.9485	0.9511	0.7148
E_6	0.3506	0.3506	0.7195	0.3506	0.3506	1	0.3506	0.3506	0.3506	0.3506	0.3506
E_7	0.9511	0.9875	0.3506	0.7148	0.9511	0.3506	1	0.9204	0.9485	0.9919	0.7148
E_8	0.9204	0.9204	0.3506	0.7148	0.9204	0.3506	0.9204	1	0.9204	0.9204	0.7148
E_9	0.9485	0.9485	0.3506	0.7148	0.9485	0.3506	0.9485	0.9204	1	0.9485	0.7148
E_{10}	0.9511	0.9875	0.3506	0.7148	0.9511	0.3506	0.9919	0.9204	0.9485	1	0.7148
E_{11}	0.7148	0.7148	0.3506	0.8887	0.7148	0.3506	0.7148	0.7148	0.7148	0.7148	1

We also found the hamming distance Ebola virus sequences by using q=1 in equation (1) of the methodology. The same algorithms as stated when q=2, has been applied to get the dissimilarity matrix based on hamming distance and the result obtained are presented in table 7.

$R_{i,j}$	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}	E_{11}
E_1	1	1	0.76	0.16	0.99	0.81	0.99	0	0.53	0.99	0.16
E_2	1	1	0.76	0.16	0.99	0.81	0.99	0	0.53	0.99	0.16
E_3	1.24	1.24	1	0.4	1.23	1.05	1.23	0.24	0.77	1.23	0.4
E_4	1.84	1.84	1.6	1	1.83	1.65	1.83	0.84	1.37	1.83	1
E_5	1.01	1.01	0.77	0.17	1	0.82	1	0.01	0.54	1	0.17
E_6	1.19	1.19	0.95	0.35	1.18	1	1.18	0.19	0.72	1.18	0.35
E_7	1.01	1.01	0.77	0.17	1	0.82	1	0.01	0.54	1	0.17
E_8	2	2	1.76	1.16	1.99	1.81	1.99	1	1.53	1.99	1.16
E_9	1.47	1.47	1.23	0.63	1.46	1.28	1.46	0.47	1	1.46	0.63
E_{10}	1.01	1.01	0.77	0.17	1	0.82	1	0.01	0.54	1	0.17
E_{11}	1.84	1.84	1.6	1	1.83	1.65	1.83	0.84	1.37	1.83	1

Table:7 Dissimilarities among eleven Ebola virus sequences based on hamming distance

In the similar fashion, we found the transitive closure among these sequences by using hamming distance formula and the corresponding results are presented in table 8.

Table:8 Dissimilarities of transitive closure among eleven Ebola virus sequences based on hamming distance

$R_{i,j}$	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}	E_{11}
E_1	1	1	0.82	0.63	0.99	0.82	0.99	0.63	0.77	0.99	0.63
E_2	1	1	0.82	0.63	0.99	0.82	0.99	0.63	0.77	0.99	0.63
E_3	1	1	1	0.63	1	1	1	0.63	0.77	1	0.63
E_4	1	1	1	1	1	1	1	0.84	1	1	1
E_5	1	1	0.82	0.63	1	0.82	1	0.63	0.77	1	0.63
E_6	1	1	0.95	0.63	1	1	1	0.63	0.77	1	0.63
E_7	1	1	0.82	0.63	1	0.82	1	0.63	0.77	1	0.63
E_8	1	1	1	1	1	1	1	1	1	1	1
E_9	1	1	1	0.63	1	1	1	0.63	1	1	0.63
E_{10}	1	1	0.82	0.63	1	0.82	1	0.63	0.77	1	0.63
E_{11}	1	1	1	1	1	1	1	0.84	1	1	1

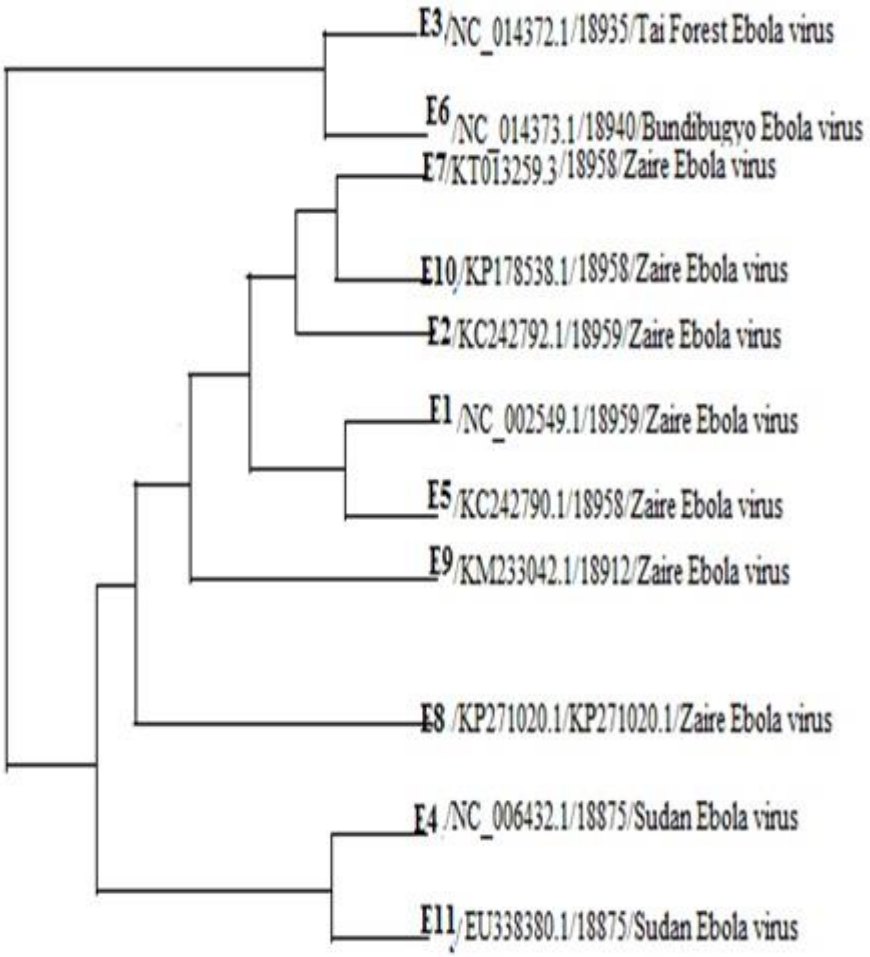
In this method we observed that when we take q=1 in equation (1) we obtained dissimilarity matrix in table 7 and subsequently transitive closure relation in table8 and due to coarseness of hamming distance, it will not be applicable in this proposed method.

From table 6, the following nested sequence among the DNA sequences of concerned Ebola virus sequences has been obtained.

Finally, the phylogenetic tree presenting the evolutionary relationship among the Ebola virus sequences has been obtained by following the purposed mathematical approach.

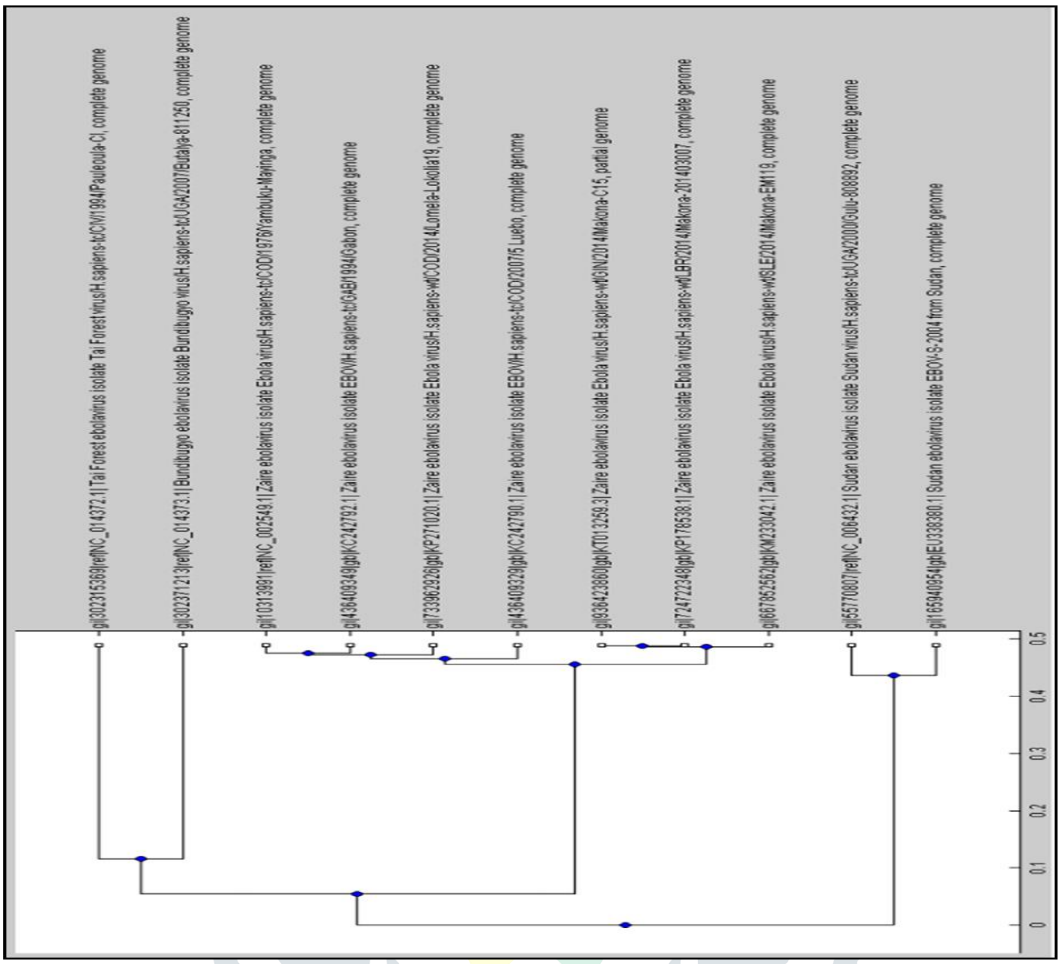
PHYLOGENETIC TREE SHOWING THE EVOLUTIONARY RELATIONSHIP AMONG THE EBOLA VIRUS BELOW

FIG 2



phylogenetic tree constructed by UPGMA method by T. Andriani[8]

Fig. 3



The evolutionary relationships (Figure 2) obtained by purposed modle are quite similar to the already established phylogenetic tree (Figure 3) by T. Andriani and M. I. Irawan[8] and both the models are implemented on the same set of sequences. So it has been pointed out that the purposed method works well and fit nicely to get the phylogenetic tree among any sequences of importance.

4. Conclusion

In this work, a new method which is free from alignment of sequence and is based on fuzzy dissimilarity metric has been purposed. Our method algorithms provide a simple way to construct phylogenetic tree with less mathematical computation. It needs to convert dissimilarity matrix to a similarity matrix to create a fuzzy transitive relationship that will be used to construct phylogenetic tree using other mathematical techniques such as clustering of sequences according to their distance. It is certainly possible to obtain a matrix of dissimilarity without alignment of sequences, but most known phylogeny construction methods required multiple alignments of sequences. The result obtained by this method is in full agreement with the existing methods which are already published, and it also validates the authenticity of our model. Thus, it has been expected that this method will be fruitful for the computational scientists and biological community to find out the complex relationship among the organisms without going directly to the met lab.

Reference

[1] M. P. R. B. Dong Xu, James M Keller, *Application of Fuzzy Logic in Bioinfomatics*. RICHMOND, TX, U.S.A: by Imperial College Press (2008), 2008.

[2] H. Fausther-Bovendo, S. Mulangu, and N. J. Sullivan, "Ebolavirus vaccines for humans and apes," *Curr. Opin. Virol.*, vol. 2, no. 3, pp. 324–329, 2012, doi: 10.1016/j.coviro.2012.04.003.

[3] M. C. Georges-Courbot *et al.*, "Isolation and Phylogenetic Characterization of Ebola Viruses Causing Different Outbreaks in Gabon," *Emerg. Infect. Dis.*, vol. 3, no. 1, pp. 59–62, 1997, doi: 10.3201/eid0301.970107.

[4] V. Madelain *et al.*, "Ebola viral dynamics in nonhuman primates provides insights into virus immuno-pathogenesis and antiviral strategies," *Nat. Commun.*, vol. 9, no. 1, pp. 1–11, 2018, doi: 10.1038/s41467-018-06215-z.

[5] W. E. Diehl *et al.*, "Ebola Virus Glycoprotein with Increased Infectivity Dominated the 2013–2016 Epidemic," *Cell*, vol. 167, no. 4, pp. 1088-1098.e6, 2016, doi: 10.1016/j.cell.2016.10.014.

[6] T. Berge, M. Chapwanya, J. M. S. Lubuma, and Y. A. Terefe, "A MATHEMATICAL MODEL for EBOLA EPIDEMIC with SELF-PROTECTION MEASURES," *J. Biol. Syst.*, vol. 26, no. 1, pp. 107–131, 2018, doi: 10.1142/S0218339018500067.

[7] D. Sun, C. Xu, and Y. Zhang, "A Novel Method of 2D Graphical Representation for Proteins and Its Application," vol. 75, pp. 431–446, 2016.

[8] 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, doi: 10.1063/1.4994467.

[9] N. Gill and S. Singh, "Biological sequence matching using fuzzy logic," *Int. J. Sci. Eng. Res.*, vol. 2, no. 7, 2011.

[10] Y. Kobori and S. Mizuta, "Similarity Estimation Between DNA Sequences Based on Local Pattern Histograms of Binary Images," pp. 1–17, 2015.

[11] R. Mathur and N. Adlakha, "Binary sequences-based approach for construction of evolutionary network," vol. 7, no. 2, pp. 1–14, 2014, doi: 10.1142/S1793524514500120.

[12] R. Mathur and N. Adlakha, "A graph theoretic model for prediction of reticulation events and phylogenetic networks for DNA sequences," *Egypt. J. Basic Appl. Sci.*, vol. 3, no. 3, pp. 263–271, 2016, doi: 10.1016/j.ejbas.2016.07.004.

[13] B. Liao, Y. Zhang, K. Ding, and T. M. Wang, "Analysis of similarity/dissimilarity of DNA sequences based on a condensed curve representation," *J. Mol. Struct. THEOCHEM*, vol. 717, no. 1–3, pp. 199–203, 2005, doi: 10.1016/j.theochem.2004.12.015.

[14] W. Zhu, B. Liao, and R. Li, "A Method for Constructing Phylogenetic," vol. 63, pp. 483–492, 2010.

[15] D. N. Georgiou, T. E. Karakasidis, A. C. Megaritis, J. J. Nieto, and A. Torres, "An extension of fuzzy topological approach for comparison of genetic sequences," *J. Intell. Fuzzy Syst.*, vol. 29, no. 5, pp. 2259–2269, 2015, doi: 10.3233/IFS-151701.

[16] W. Huang, J. Zhang, Y. Wang, and D. Huang, "A simple method to analyze the similarity of biological sequences based on the fuzzy theory," *J. Theor. Biol.*, vol. 265, no. 3, pp. 323–328, 2010, doi: 10.1016/j.jtbi.2010.05.008.

[17] L. A. Zadeh, "Similarity relations and fuzzy orderings," *Inf. Sci. (Ny).*, vol. 3, no. 2, pp. 177–200, 1971, doi: 10.1016/S0020-0255(71)80005-1.

[18] D. K. Wedding, *Fuzzy sets and fuzzy logic: Theory and applications*, vol. 14, no. 3. 1997.