Article

A new measure of dissimilarity and fuzzy linear programming model to construct phylogenetic network among DNA sequences

Rinku Mathur[1], **Neeru Adlakha**[2]
[1]Depatment of Mathematics, School of Chemical Engineering and Physical Sciences, Lovely Professional University, Phagwara - 144411, Punjab, India
[2]Department of Applied Mathematics and Humanities, S. V. National Institute of Technology, Surat-395 007, Gujarat, India
E-mail: rinkumatur56@gmail.com, rinku.22748@lpu.co.in, neeru.adlakha21@gmail.com

Received 20 May 2019; *Accepted 30 June 2019*; *Published 1 December 2019*

Abstract
The growth of DNA databases used to store large number of biological sequence data, has stimulated the importance of alignment of sequences for phylogenetics. Most of the phylogenetic methods based on alignment of sequences consume long time to provide the results. In this regard, a new alignment free measure, based on frequency of occurrence of different nucleotides in sequences has been reported. The Euclidean distance metric has been used over these frequencies of nucleotides to obtain the dissimilarities among DNA sequences. These distances are then used to construct the phylogenetic tree among sequences. In addition, a fuzzy linear programming model has been developed here to construct the phylogenetic network which is considered as the generalized form of phylogenetic tree. As an application, the proposed method is applied over the data set of β- globin gene of nine species and is validated by comparing the obtained results with the already existing method. The results obtained are more promising over the available method and can be applied over any length of input data sequences.

Keywords Euclidean distance; DNA sequences; phylogenetic network; fuzzy linear programming.

Network Biology
ISSN 2220-8879
URL: http://www.iaees.org/publications/journals/nb/online-version.asp
RSS: http://www.iaees.org/publications/journals/nb/rss.xml
E-mail: networkbiology@iaees.org
Editor-in-Chief: WenJun Zhang
Publisher: International Academy of Ecology and Environmental Sciences

1 Introduction

The scientific databases are growing rapidly with the advent of large number of DNA sequences responsible for physiological structures of organisms. The DNA sequences are the strings of four nucleotides or characters: Adenine (A), Guanine (G), Cytosine (C) and Thymine (T). But, it is very difficult to get the information about the organisms directly from these sequences. It is one of the challenging tasks for the biologists to analyze the DNA sequences mathematically. Henceforth, a lot of methods have been developed and to get the information about the sequences of various organisms (Qi et al., 2012; Liua et al., 2006; Kimura, 1980).

Sequence alignment has one of the powerful tools to compare the two or multiple sequences based on the order of nucleotides present in the sequences (Larkin et al., 2007). The divergence of species or organisms over the time makes the non-alignment of concerning sequences. The graphical techniques are also considered as the important tools for the visualization and characterization of DNA sequences. These techniques provide the 2D or 3D representation of the DNA sequences (Liu et al., 2006; Qi and Fan, 2007; Liao and Wang, 2004; Aita et al., 2011). The distance methods are also used to obtain the similarity or dissimilarity among the number of sequences. Some of the important models for the computation of genetic distances among the sequences are as: Jukes – Cantor model, Kimura – two or three – parameter model, F84 distance model, LogDet distance model, etc (Lockhart et al., 1994).

In this work, a new alignment free measure has been reported for the calculation of dissimilarity of DNA sequences in which the frequencies of occurrence of nucleotides has been considered. Based on the frequency of occurrence of different nucleotides, the Euclidean distance has been measured among every pair of sequences. These distances provide us the genetic or original distances between the species or taxa. The genetic distances obtained are then used for the phylogenetic analysis of sequences (Grzegorzewski, 2004).

Phylogenies are generally considered as an important tool to predict the evolution of species, populations and individuals (Eslahchi et al., 2010; Bandelt et al., 1999). Phylogenetic tree is a model that gives the estimation of divergence of DNA sequences. Various methods (like NJ, UPGMA, Maximum parsimony, Fitch-Margolish method, etc) are available for the construction of these trees (Zhang, 2016, 2018). The distance and character based models are the most commonly used methods for phylogenetic analysis (Lockhart et al., 1994; Zhu et al., 2010). To see the relationships among the species, we used the simple average linkage clustering method of phylogenetic tree construction (Qi et al., 2011). But with the increase of DNA sequences, the evolutionary processes like horizontal gene transfer and recombination can occur. Such type of events could not be represented by phylogenetic tree. Meanwhile, the importance of phylogenetic networks has increased with the growth of DNA sequences.

Phylogenetic networks are generally considered as the generalization of tree which can visualize the above mentioned reticulate events. There are several methods to construct the phylogenetic networks which are based on distance matrix such as MC–Net, Neighbor NET, Q–Net, Least Squares, Node-similarity, etc (Makarenkov et al., 2004; Dress et al., 2010; Chan et al., 2005; Tang and Moret, 2005; Morrison, 2013; Zhang, 2015, 2016, 2018). Here in this work, we made an attempt to construct a phylogenetic network based on fuzzy membership matrix and tree distances among sequences. The fuzzy linear programming is used for the construction of network. As an application, the proposed method is applied over the data set of β- globin gene of nine species. The method is also validated by comparing the obtained results with the already existing method like T-Rex (Eslahchi et al., 2010; Makarenkov, 2001).

2 Basic Preliminaries

The preliminaries used in this study are discussed in subsection:

2.1 Frequency of nucleotides

Let $S = s_1, s_2, s_3, ..., s_n$ be a sequence of n nucleotides having the length n such that $s_i \in \{A, C, G, T\}$. Then composition of occurrence of every nucleotide is defined as total number of that nucleotides present in the whole sequence and is denoted by f_{s_i}.

$$f_{s_i} = \frac{s_i}{n_j} \quad \text{where } s_i \in \{A, C, G, T\} \text{ and } n_j \text{ is the length of the } j_{th} \text{ sequence.}$$

2.2 Euclidean distance matrix

A matrix M of order $n \times n$ is called a Euclidean distance matrix if it satisfies the following conditions (Grzegorzewski, 2004):

i) M is a symmetric matrix : $d_{ii} = d_{ii} \quad \forall\ i, j = 1, 2, \ldots\ldots\ldots\ldots, n$.

ii) Diagonal elements are all zero: $d_{ii} = 0 \quad \forall\ i, j = 1, 2, \ldots\ldots\ldots\ldots, n$.

iii) There exist n points: $p_1, p_2, \ldots\ldots\ldots\ldots, p_n$ on m - dimensional space such that

$$d_{ij} = \sqrt{\left(p_i - p_j\right)^2} \quad : 1 \le i, j \le n .$$

2.3 Fuzzy set and fuzzy relation set

If X is a collection of n objects generally denoted by x, then a Fuzzy set A in X is a set of ordered pairs (Zimmermann, 2001).

$$A = \{(x, \mu_A(x)) \ / \ x \in X\}$$

where $\mu_A(x)$ is the value of the membership function or grade of membership of x in A. Also $A = \{((x, y), \mu_A(x, y)) \ / \ x, y \in X\}$ is called the fuzzy relation set.

2.4 Fuzzy membership function

A function defined on the set X to the membership space ranges from the 0 to1 is called a fuzzy membership function and is denoted as $\mu_A(x)$. i.e.,

$$\mu_A(x): X \to [0,1]$$

Also $\mu_A(x, y): X \times X \to [0,1]$ is called fuzzy relation membership function (Zimmermann, 2001; Mohaddes and Mohayidin, 2008).

2.5 Phylogenetic tree

Any connected graph with a unique path between any two distinct vertices u and v is called a tree and is denoted as $T(uv)$. A tree T has exactly $|V| - 1$ edges.

Related to the set X, the X - trees are associated by two properties:

(i) the set of leaves of T is X;

(ii) for any $v \in V - X$, $\partial(v) \ge 3$.

An X tree with n leaves has at most $n-2$ internal vertices, $2n - 2$ vertices and thus $2n - 3$ edges and any given Phylogenetic tree representing the evolutionary history of taxon or organisms can be transformed into a binary tree by adding links of length zero wherever it is necessary (Makarenkov and Legendre, 2004; Semple and Steel, 2003; Mathur and Adlakha, 2013).

2.6 Phylogenetic network

Any phylogenetic tree with loops is called a phylogenetic network or alternatively any directed acyclic graph in which every node except the root satisfies one of the following conditions (Tusserkani et al., 2011):

i) It has indegree 2 and outdegree 1. These nodes are called reticulation nodes.

ii) It has indegree 1 and outdegree 2. These nodes are called binary nodes.

iii) It has indegree 1 and outdegree 0. These nodes are called leaves.

Any phylogenetic network inferred from tree with n leaves can have at most $n(n-1)/2$ branches.

3 Methodology

The procedure to obtain the phylogenetic network in this study is divided into three steps which are discussed in subsequent subsections:

3.1 Evaluation of genetic distances among DNA sequences

Let "N" be the number of the number of DNA sequences which has to be analyzed to get the evolutionary relationship among them. As per the available methods, the genetic distances among sequences are generally

measured by firstly aligning and then calculating the distances among them by the available model like Jukes-Cantor, Kimura-two-parameter, etc (Lockhart et al., 1994; Zhu et al., 2010). In this paper, we have presented, a alignment free method based on the frequency of occurrence of nucleotides in sequences.

Let $S_i = s_1, s_2, s_3, ..., s_n$ be the length of a sequence having n nucleotides and i ranges from $1\ to\ N$ (number of sequences) and $s_1, s_2, ..., s_n \in \{A, C, G, T\}$.

For every sequence, the percentage or composition of occurrence of each nucleotide in sequence is given by: $f_{s_i} = \frac{No.of\ s_i}{n_j}\ such that\ \ s_i \in \{A, C, G, T\}$ where n_j is the length of the j_{th} sequence.

After finding the value of f_{s_i} for each sequence, we will evaluate the Eucledian distance among every pair of available sequences m and l given by the formula (Lockhart et al., 1994):

$$d_{ml} = \sqrt{\sum_{s_i} \left(f_{ms_i} - f_{ls_i} \right)^2} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ (1)$$

where f_{ms_i} is the frequency of nucleotide s_i for the sequence m.

At last, the symmetric original (genetic) distance matrix among all the pair of species is obtained by using the above distance formula. These distances are then used to construct the phylogenetic tree.

3.2 Construction of phylogenetic tree

Presently, most of the methods are available for the construction of tree like UPGMA, NJ, etc which are based on distances of sequences. Also, the character based methods are available for the tree construction. Based on the dissimilarity matrix generated in the above section, the tree is constructed. The smaller the element in the dissimilarity matrix, the most closer the species are and vice versa. In this paper, we will use the average linkage clustering approach to construct the phylogenetic tree (Qi et al., 2011).

Let d_{mn} and τ_{mn} denotes the genetic and tree distances among the sequences m and n. The pair of species in a tree are clustered according to their genetic distances d_{mn}. The pair of species are clustered firstly having the smaller d_{mn} and so on. Proceeding in the similar fashion, all the species should remain present in the tree. If for some $d_{mn} < d_{pq}$, the value of $\tau_{mn} > \tau_{pq}$, then it violates the rule of evolution of species according to their ancestry. So, in this case, we will reshuffle d_{mn} by d_{pq} to get the tree distances τ_{mn}. Finally, we will obtain the symmetric tree distance matrix for all the species considered in the study.

3.3 Find the phylogenetic network

The tree distance matrix obtained in the above section has been used to construct the phylogenetic network. In search of reticulation branches of network, we will detect the pair of species whose $\tau_{mn} > d_{mn}$ which signifies that these are diverging with time. We will consider reticulation branches as the only that pair of species that are diverging in the tree from their genetic distances. By incorporating this condition, we are saving the time for construction of network. Then, the tree distances τ_{mn} for every pair of species are converted into fuzzy relational membership matrix μ_{mn} by defining the function (Mathur and Adlakha, 2013):

$\mu : X \times X \to [0, 1]$ which is defined as:

$$\mu(m,n)=\frac{1}{1+\tau(m,n)} \qquad \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

The function (2) is based on the fact that $\mu(m,n)$ attains the highest value for the same nucleotide sequence. As the evolutionary distances among sequences are increasing, the membership is decreasing.

From the available data of tree and fuzzy distances among the pair of species, we can find the network by adding the new branches to the tree as the reticulation branches. Our aim is to minimize or optimize the total length of the weighted network which represents the reticulate events. So, an alternative fuzzy linear programming (FLP) formulation of the network based on dissimilarity and fuzzy matrices is proposed (Sridhar et al., 2007; Mathur and Adlakha, 2014). In this model, the variable $e(i,j)$ is used to denote the presence or absence of the edges $(i,j)\in E(T)$ and $\tau(i,j)$ is the length or weight of the corresponding edge in the phylogenetic tree or network, whichever is smaller. Thus, the proposed fuzzy LP model to get the desired results for the evolutionary network is described as under:

$$Minimize \quad Z = \sum_{i=1}^{|V|}\sum_{j=1}^{|V|}\mu(i,j)\tau(i,j)e(i,j) + \mu(s,t)\gamma(s,t)(new)e(s,t)(new) - \sum_{\substack{i=s,\\ j=t}}\mu(i,j)\tau(i,j)(old)e(i,j)(old) \qquad \forall \; i,\; j \in V$$

$$\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

subject to

$$\mu(s,t)\tau(s,t)(old)e(s,t)(old) - \mu(s,t)\gamma(s,t)(new)e(s,t)(new) \geq 0 \qquad \ldots\ldots (4)$$

$$\tau(i,j)\in[0,1] \qquad \forall \; i,j \in V \qquad \ldots\ldots\ldots\ldots\ldots (5)$$

$$\sum_{i=1}^{|V|}\sum_{j=1}^{|V|}\mu(i,j)\gamma(i,j)e(i,j) = \sum_{i=1}^{|V|}\sum_{j=1}^{|V|}\mu(j,i)\gamma(j,i)e(j,i) \qquad \forall \; i,j \in V \qquad \ldots\ldots (6)$$

$$\sum_{j=1}^{|V|}e(i,j) \geq 1 \qquad if\ 'i'\ is\ an\ initial\ or\ terminal\ vertex. \qquad \ldots\ldots\ldots\ldots (7)$$

$$\sum_{j=1}^{|V|}e(i,j) \geq 2 \qquad if\ 'i'\ is\ an\ internal\ node\ or\ vertex. \qquad \ldots\ldots\ldots\ldots\ldots (8)$$

where $\gamma(s,t)(new)$ is the new length of the edge (s,t) added in the tree to construct network.

$\tau(i,j)(old)$ is the length of the edge (i,j) in the tree.

$\mu(i,j)$ is the membership of edge (i,j) corresponding to $\tau(i,j)$ for all $i,\; j \in V$.

The first term on the right side of (3) represents total length of the phylogenetic tree where $\tau(i,j)e(i,j)$ is the distance between nodes i and j for edge $e(i,j)$. The second term on the right side of (3) represents the distance between nodes s and t which is the length of new edge $e(s,t)$ being added to phylogenetic network where $\gamma(s,t)(new)e(s,t)(new)$ denotes the length of new edge $e(s,t)$. The third term on the right side of (3) denotes the length of old edge $e(i,j)$ being removed from the distance of the network to obtain the optimum Z where $\tau(i,j)(old)e(i,j)(old)$ denotes distance between nodes (i,j) for the edge $e(i,j)(old)$.

The constraint (4) imposes the condition that the length of new edge added should be less than the length of the existing edge $e(s,t)(old)$ in the tree and it may be equal to the original distance among particular species. Constraint (5) shows that the length of each edge in the phylogenetic tree should be in between 0 and 1. Constraint (6) imposes the condition that the total length among all the vertices in the network remains same while going from each and every path of the network. The constraint (7) means that there is at-least one link or edge between the selected and other vertices if selected vertex is either initial or terminal. The constraint (8) means that if the selected vertex is internal, then there are at-least two links or edges between the selected and the other vertices.

The present fuzzy LP formulation gives us flexibility to evaluate only the nodes which are diverging in the tree while existing methods evaluate each node in the phylogenetic tree. In order to construct the optimized network, it is necessary to identify that how many new branches can be added to the tree. In view of above, the possible goodness-of-fit criteria allowing one to determine, when to stop adding branches to a phylogenetic network is mentioned below (Makarenkov and Legendre, 2004:

The total number of nodes in an unrooted binary phylogenetic tree with n leaves is $2n-2$. Therefore, the maximum number of branches one might place in a reticulated network, inferred from a binary phylogenetic tree with n leaves, is $(2n-2)(2n-3)/2$. However, any metric distance can be represented by a complete graph with $n(n-1)/2$ branches. Thus, any of the two limits $(2n-2)(2n-3)/2$ or $n(n-1)/2$ can be considered as the maximum possible number of branches in a reticulated network. If the latter limit is considered, the number of degrees of freedom of a phylogenetic network with N branches can be defined as $[n(n-1)/2]-\mathrm{N}$ (Makarenkov and Legendre, 2004).

Thus, the goodness-of-fit function is given by:

$$Z_s = \frac{Z - Z_1}{[n(n-1)/2]-\mathrm{N}} \qquad \text{..... (9)}$$

where Z_s is the criterion to stop the addition of new branches to the network and Z_1 is the updated value of Z while constructing the network.

The function Z_s considers the number of reticulations up to which the value of Z starts increasing and the number of minimum values indicating number of possible reticulations will lie in the range minimum one over the interval $[2n-3,\ n(n-1)/2]$ of possible values of N.

4 Application to Experimental Result of β - globin Gene of Nine Species

4.1 Data material

To check the utility of this method, we take the first exon of β - globin gene for nine different species, which were also studied by Qi et al. (2012) and Liu and Wang (2006). These sequences were retrieved from the NCBI site (http://www.ncbi.nlm.nih.gov/taxonomy). Table 1 provides the detailed information about these nine sequences, while Table 2 presents the nine coding sequences of the organisms. The sequences of Table 2 are analyzed to calculate the dissimilarity matrix based on the frequency of occurrence of nucleotides in sequences and Euclidean distances among them.

Table 1 ID Information for Exon-1of β - globin gene of nine species.

Species	ID/Accesion	Database	Exon 1 location	Length
Human (H)	AH001475	NCBI	1612-1703	92
Goat (G)	M15387	NCBI	279-364	86
Opossum (Op)	J03643	NCBI	467-558	92
Gallus (Gl)	V00409	NCBI	465-556	92
Lemur(L)	M15734	NCBI	154-245	92
Mouse (M)	V00722	NCBI	275-367	93
Rat (R)	X06701	NCBI	310-401	92
Bovine (B)	X00376	NCBI	278-363	86
Chimpanzee (Ch)	X02345	NCBI	4189-4293	105

Table 2 Coding sequences of Exon-1of β - globin gene of nine species.

Species	Coding DNA Sequences
Human	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGG CAAGGTGAACGTGGATTAAGTTGGTGGTGAGGCCCTGGGCAG
Goat	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCAAGG TGAAAGTGGATGAAGTTGGTGCTGAGGCCCTGGGCAG
Opossum	ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGGT CTAAGGTGCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG
Gallus	ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTGGG GCAAGGTCAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG
Lemur	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTGGGGC AAGGTGGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG
Mouse	ATGGTTGCACCTGACTGATGCTGAGAAGTCTGCTGTCTCTTGCCTGTGGGCA AAGGTGAACCCCGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
Rat	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGGGAA AGGTGAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAG
Bovine	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGCCTTTTGGGGCAAGGTG AAAGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG
Chimpanzee	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGC AAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGG

4.2 Results and discussion

The results obtained by the proposed method for the frequency of occurrence of nucleotides in DNA sequences are listed in Table 3. Table 4 presents the dissimilarity matrix of all the species based on distance measure $\left(d_{m\,n}\right)$ given by equation (1).

Table 3 Frequency of occurrence of nucleotides in DNA sequences of species.

Species \ Freq. of Nucl.	Adenine (A)	Guanine (G)	Cytocine (C)	Thymine (T)
Human	0.18478	0.38043	0.20652	0.22826
Goat	0.19767	0.40697	0.19767	0.19767
Opossum	0.22826	0.31521	0.21739	0.23913
Gallus	0.20652	0.36956	0.26086	0.16304
Lemur	0.20652	0.38043	0.16304	0.25000
Mouse	0.18085	0.36170	0.21276	0.24468
Rat	0.21739	0.35869	0.19565	0.22826
Bovine	0.19767	0.40697	0.18604	0.20930
Chimpanzee	0.19047	0.39047	0.19047	0.22857

Table 4 Genetic (Original) distances for first exon of β - globin gene of species.

Species	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rat	Bovine	Chimpanzee
Human	----	0.04340	0.07986	0.08828	0.05331	0.02592	0.04070	0.04057	0.019793
Goat		----	0.10703	0.08166	0.06873	0.06900	0.06050	0.01647	0.03647
Opossum			----	0.10537	0.08825	0.06678	0.05091	0.10589	0.08899
Gallus				----	0.13134	0.09845	0.09347	0.09600	0.09971
Lemur					----	0.05924	0.04615	0.05446	0.03965
Mouse						----	0.04362	0.06545	0.04089
Rat							----	0.05630	0.04197
Bovine								----	0.02668
Chimpanzee									----

After getting the above evolutionary distances among the species, we constructed the phylogenetic tree among these species based on the proposed method. The tree gives us evolutionary history of these species. Table 5 presents the dissimilarity matrix of the tree distances among species. The smallest entries of Table 5

were noticed for pair (Goat, Bovine) and (Human, Chimpanzee) with $d_{GB} = 0.0164$ and $d_{HCh} = 0.0197$.

These two pairs will be clustered in the tree firstly. Analyzing all the distances of Table 5, we have three clusters for the species. Fig. 1 represents the phylogenetic tree of species corresponding to the distances of Table 5. The tree obtained by using the distances of the purposed method is in agreement with the results obtained by Liu and Wang (2006). The tree is also considered as the basis for the construction of phylogenetic network.

Table 5 Phylogenetic tree distances among the sequences and bold entries shows diverging pair of species.

Species	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rat	Bovine	Chimpanzee
Human	----	0.0197	**0.0839**	**0.0992**	**0.0535**	**0.0333**	**0.0486**	0.0197	0.0197
Goat		----	0.0839	**0.0992**	0.0535	0.0333	0.0486	0.0164	0.0197
Opossum			----	0.0992	0.0839	**0.0839**	**0.0839**	0.0839	0.0839
Gallus				----	0.0992	**0.0992**	**0.0992**	**0.0992**	0.0992
Lemur					----	0.0535	**0.0535**	0.0535	**0.0535**
Mouse						----	**0.0486**	0.0333	0.0333
Rat							----	0.0486	**0.0486**
Bovine								----	0.0197
Chimpanzee									----

Now the fifteen pair of species in the phylogenetic tree diverged from their original distances. These pairs are then added to the tree to construct the network on the basis of the proposed fuzzy linear programming. The addition of these new branches gives information about the reticulate events. These pairs of species that are diverging from the original distances are listed in Table 6 with the rate of divergence in the form of distances.

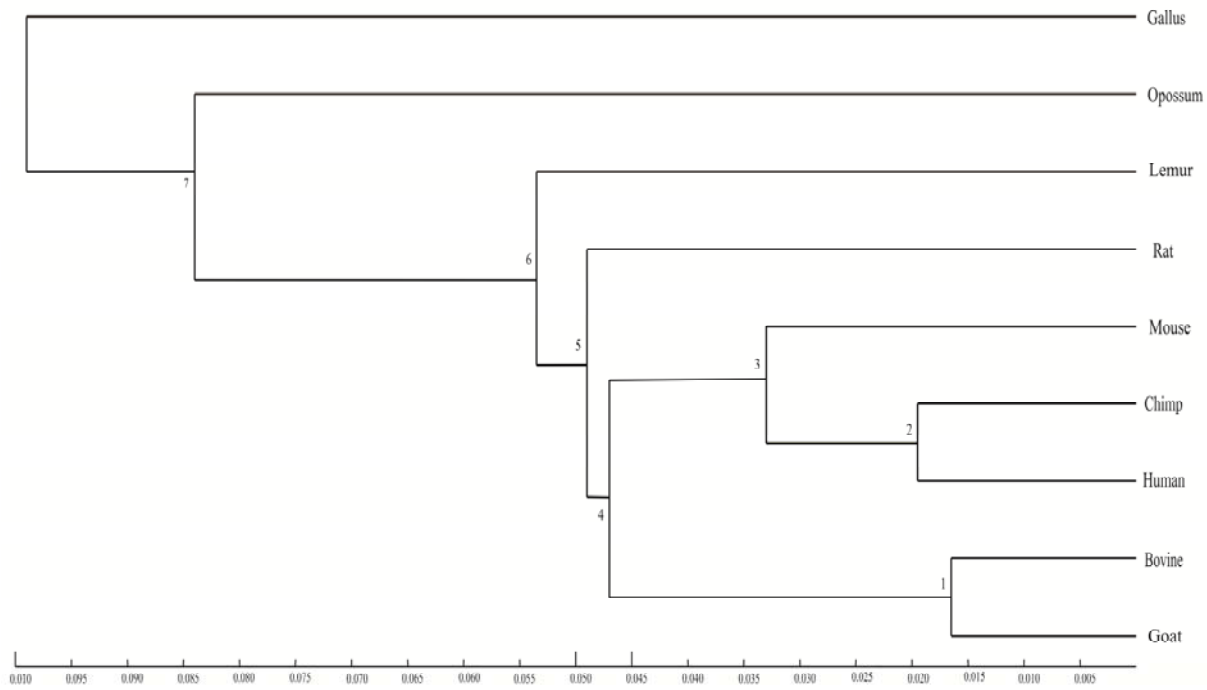

Fig. 1 Phylogenetic tree showing the evolutionary history of β - globin gene of nine speices.

Table 6 Distances showing diverging pair of species.

S. No.	Pair of Species	Rate of divergence
1	(Human, opossum)	0.0041
2	(Human, Gallus)	0.0110
3	(Human, lemur)	0.0002
4	(Human, Mouse)	0.0074
5	(Human, Rat)	0.0079
6	(Goat, Gallus)	0.0176
7	(Opossum, Mouse)	0.0172
8	(Opossum, Rat)	0.0330
9	(Gallus, Mouse)	0.0008
10	(Gallus, Rat)	0.0058
11	(Gallus, Bovine)	0.0032
12	(Lemur, Rat)	0.0074
13	(Lemur, Chimpanzee)	0.0139
14	(Mouse, Rat)	0.0050
15	(Rat, Chimpanzee)	0.0067

On the basis of tree distances, the fuzzy relational membership matrix among these species is calculated by using the equation (2) (see Table 7). The distances of Table 5 and Table 7 are then used to construct the phylogenetic network. At last, addition of new branches was explored for the construction of network.

The value of Z for tree without fuzzy membership distances of species before adding new branches to it is as:

$$Z = \tau(G,1) + \tau(B,1) + \tau(1,4) + \tau(3,4) + \tau(2,3) + \tau(H,2) + \tau(Ch,2) + \tau(M,3) + \tau(4,5) + \tau(R,5) + \tau(5,6) + \tau(L,6) + \tau(6,7) + \tau(Op,7) + \tau(Gl,7)$$

$$= 0.4853$$

Table 7 Fuzzy relational membership matrix among the species.

Species	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rat	Bovine	Chimpanzee
Human	1.0000	0.9806	0.9225	0.9097	0.9492	0.9677	0.9536	0.9806	0.9806
Goat		1.0000	0.9225	0.9097	0.9492	0.9677	0.9536	0.9838	0.9806
Opossum			1.0000	0.9097	0.9225	0.9225	0.9225	0.9225	0.9225
Gallus				1.0000	0.9097	0.9097	0.9097	0.9097	0.9097
Lemur					1.0000	0.9492	0.9492	0.9492	0.9492
Mouse						1.0000	0.9536	0.9677	0.9677
Rat							1.0000	0.9536	0.9536
Bovine								1.0000	0.9806
Chimpanzee									1.0000

The value of Z for tree with fuzzy distances of corresponding pairs before adding new branches is as:

$$Z = \mu_f(G,1)\tau(G,1) + \mu_f(B,1)\tau(B,1) + \mu_f(1,4)\tau(1,4) + \mu_f(3,4)\tau(3,4) + \mu_f(2,3)\tau(2,3) + \mu_f(H,2)\tau(H,2) + \mu_f(Ch,2)\tau(Ch,2) + \mu_f(M,3)\tau(M,3) + \mu_f(4,5)\tau(4,5) + \mu_f(R,5)\tau(R,5) + \mu_f(5,6)\tau(5,6) + \mu_f(L,6)\tau(L,6) + \mu_f(6,7)\tau(6,7) + \mu_f(Op,7)\tau(Op,7) + \mu_f(Gl,7)\tau(Gl,7)$$

$$= 0.4597$$

Now, the value of Z has been optimized with the addition of new branches to the tree to construct the network.

Case - I: Exploring the addition of first branch.

(a) If the pair (H, OP) is added to the tree, then

$$\gamma(H,Op)(new) = 0.0798 \text{ and } \tau(H,Op)(old) = 0.0839$$

$$\therefore \quad Z = \sum \mu_f(i,j)\tau(i,j) + \mu_f(H,Op)\gamma(H,Op) - \mu_f(H,Op)\tau(H,Op)$$

$$= 0.4597 + 0.0738 - 0.0773 = 0.4562$$

(b) If the pair (H, Gl) is added to the tree, then

$$Z = 0.4597 + 0.0810 - 0.0902 = 0.4505$$

Proceeding in the similar fashion, the value of Z is calculated for all the 15 diverging pairs of species which are presented in Table 8.

It has been observed that the value of Z with the fuzzy function has been minimized for the pair (Opossum, Rat). So, the first branch between the pair Opossum and Rat can be added to tree to get the network. Now, we call it by Z_1 as the updated value of Z which is used to calculate the goodness of fit criterion for the number of reticulation branches to be added to network.

Case – II: Exploration of addition of second branch

(a) If the pair (H, OP) is added to the tree, then

$$Z = 0.4308 + 0.0738 - 0.0773 = 0.4273$$

(b) If the pair (H, Gl) is added to the tree, then

$$Z = 0.4308 + 0.0810 - 0.0902 = 0.4216$$

In order to get the optimized value of Z for second branch, all the remaining diverging pairs are calculated in the similar way. Following the same process, the values of Z for the second, third, fourth, fifth, sixth and seventh reticulation braches are 0.4160, 0.4012, 0.3885, 0.3793, 0.3721 and 0.3651.

But according to the goodness of fit criterion given by equation (9), only two reticulation branches among the pairs (Opossum, Rat) and (Goat, Gallus) can be added to the tree to construct the network. Table 9 presents the information regarding the optimized values of Z and how many branches can be added to phylogenetic network.

Table 8 Values of Z (with and without fuzzy function) corresponding to diversified pair of organisms.

S. No.	Pair of Species	Value of Z with Fuzzy function	Value of Z without fuzzy function
0	----------	0.4597	0.4853
1	(Human, opossum)	0.4562	0.4812
2	(Human, Gallus)	0.4505	0.4743
3	(Human, lemur)	0.4595	0.4851
4	(Human, Mouse)	0.4527	0.4779
5	(Human, Rat)	0.4525	0.4774
6	(Goat, Gallus)	0.4449	0.4677
7	(Opossum, Mouse)	0.4449	0.4681
8	(Opossum, Rat)	**0.4308**	0.4523

9	(Gallus, Mouse)	0.4590	0.4845
10	(Gallus, Rat)	0.4549	0.4795
11	(Gallus, Bovine)	0.4570	0.4821
12	(Lemur, Rat)	0.4530	0.4779
13	(Lemur, Chimpanzee)	0.4470	0.4714
14	(Mouse, Rat)	0.4551	0.4803
15	(Rat, Chimpanzee)	0.4536	0.4786

Table 9 gives us the notion that only two reticulation branches can be added to construct the network. These results provide the information of evolutionary relationships and reticulate events among species.

Table 9 Values of Z_s corresponding to the number of reticulation branches to be added to the tree or network.

No. of branches	Degrees of freedom	Pair of Species	Values of Z	Values of Z_1	Values of Z_s
0	21	-------	0.4597	-------	0.021890
1	20	(Opossum, Rat)	0.4597	0.4308	0.001445
2	19	(Goat, Gallus)	0.4308	0.4160	0.000778
3	18	(Opossum, Mouse)	0.4160	0.4012	0.000822
4	17	(Lemur, Chimpanzee)	0.4012	0.3885	0.000747
5	16	(Human, Gallus)	0.3885	0.3793	0.000575
6	15	(Human, Rat)	0.3793	0.3721	0.00048
7	14	(Human, Mouse)	0.3721	0.3651	0.0005
8	13	(Lemur, Rat)	0.3651	0.3584	0.000515
9	12	(Rat, Chimpanzee)	0.3584	0.3523	0.000508
10	11	(Gallus, Rat)	0.3523	0.3475	0.000436
11	10	(Mouse, Rat)	0.3475	0.3429	0.00046
12	9	(Human, opossum)	0.3429	0.3397	0.000355
13	8	(Gallus, Bovine)	0.3397	0.3370	0.000375
14	7	(Gallus, Mouse)	0.3370	0.3363	0.0001
15	6	(Human, lemur)	0.3363	0.3361	0.000033

The value of Z is reduced from 0.4597 to 0.4308 when the first reticulation branch among (Opossum, Rat) is added. The addition of second reticulation branch among (Goat, Gallus) minimizes the value of Z from 0.4308 to 0.4160. The phylogenetic network obtained by this method is shown in Fig. 2.

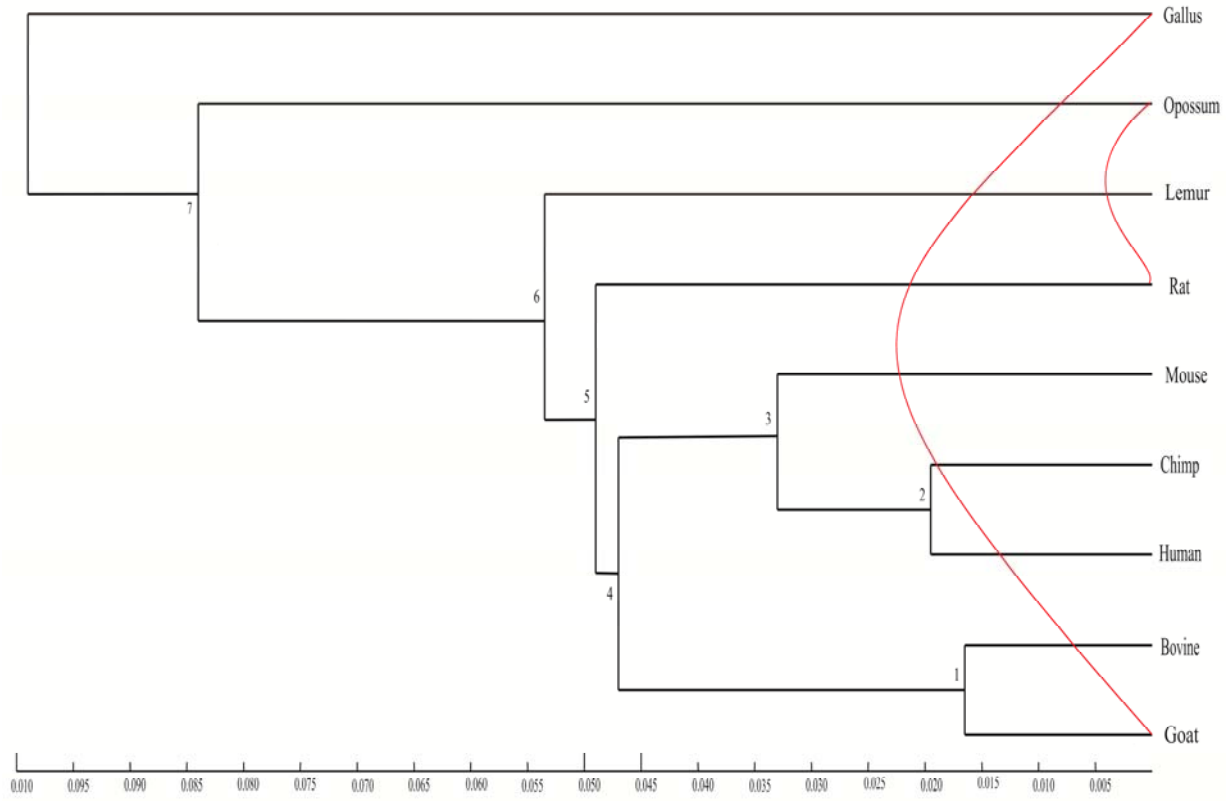

Fig. 2 Phylogenetic network obtained by FLP showing the relationship among β - globin gene of nine species.

Fig. 3 represents the phylogenetic network obtained by T-Rex software by using the sequences of Table 2. The results obtained for the phylogenetic network shown by Fig. 2 involves the reticulation events are also in agreement with that of obtained by online tool T-Rex shown by Fig. 3 (Makarenkov, 2001). The proposed method predicts the occurrence of first reticulation branch between (Opossum, Rat) while the T-Rex shows between (Opossum, Human). The second reticulation branch occurs between (Goat, Gallus) by both the methods.

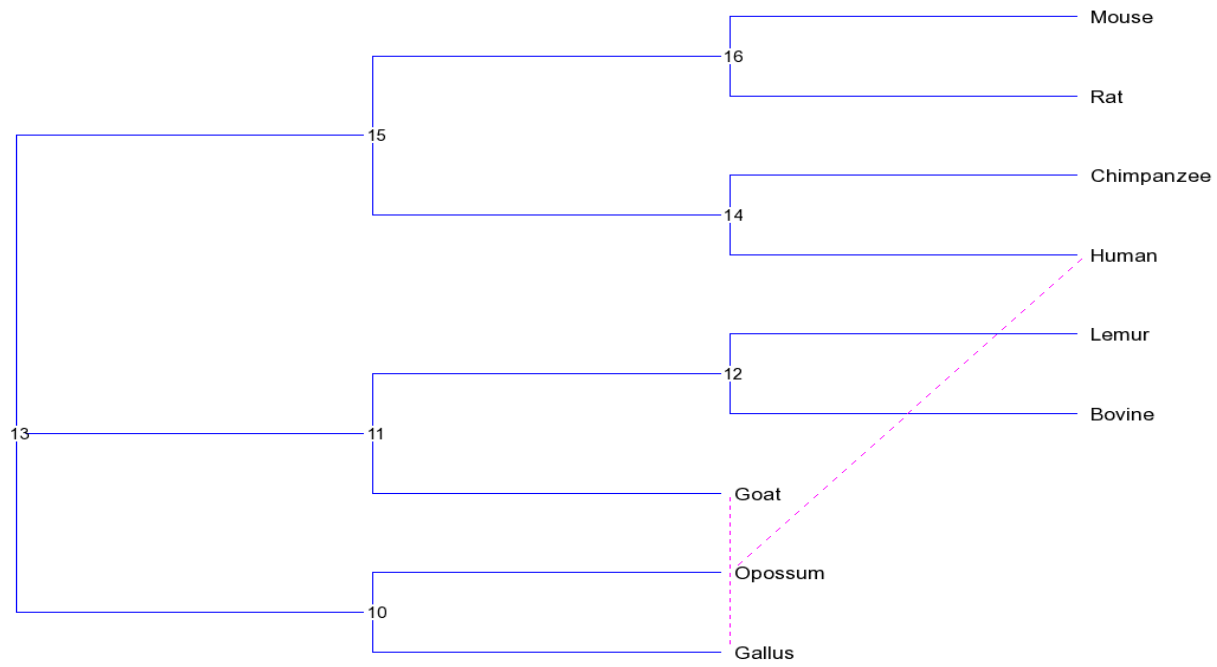

Fig. 3 Phylogenetic network obtained by T-Rex showing the relationship among β - globin gene of nine species.

The values of Z for reticulation branches obtained by the proposed fuzzy LP and without fuzzy LP are comparatively shown by Fig. 4. It has been depicted by Figure 4 that the value of $Z = 0.4853$ (without fuzzy LP) has been decreased to $Z = 0.4597$ with fuzzy LP when no reticulation branch has been added. Figure 4 also suggested that the values of Z for fuzzy LP have lower values than without fuzzy LP for all the possible reticulation branches that can be added to tree. So, fuzzy LP gives the promising results as compared to the LP for construction of phylogenetic networks.

Fig. 5 interprets the comparison between the purposed method and the existing T-Rex method for the values of Z_s that shows the addition of reticulation branches to the network. The purposed method has the higher value of $Z_s = 0.021890$ when no branch has been added to the tree. With the addition of first and second reticulation branches, the purposed method predicts the sharp decrease with the values $Z_s = 0.001445$ and $Z_s = 0.000778$ while the T-Rex gives the values $Z_s = 0.004738$ and $Z_s = 0.004584$. So, on observing the Fig. 5, it can be commented that the new method performs better.

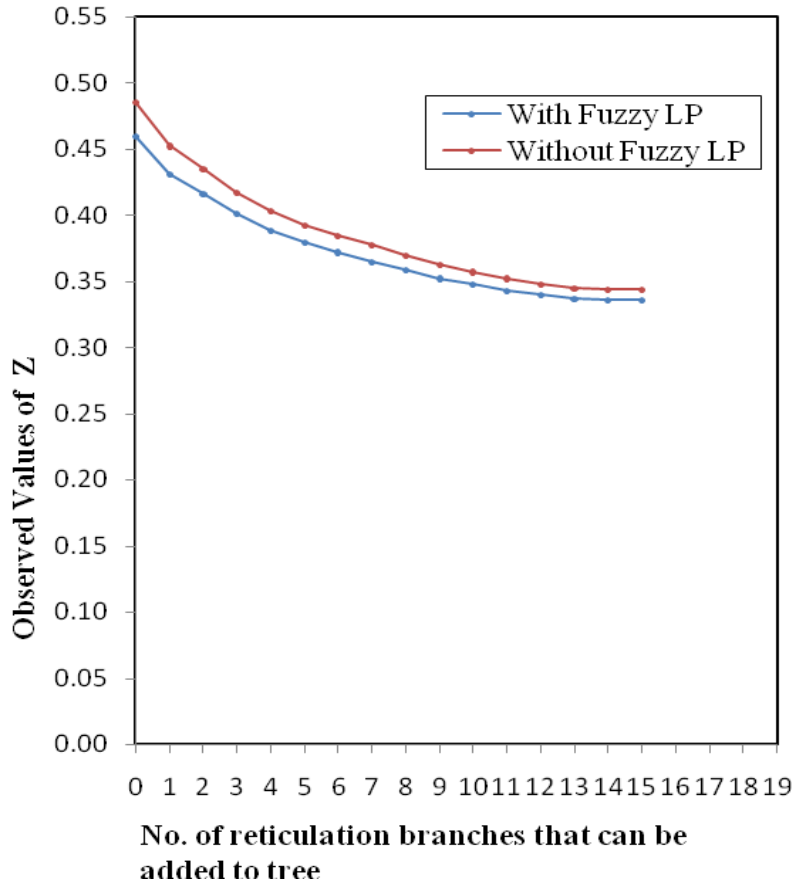

Fig. 4 The values of Z obtained for different reticulation branches in respect of FLP and without FLP.

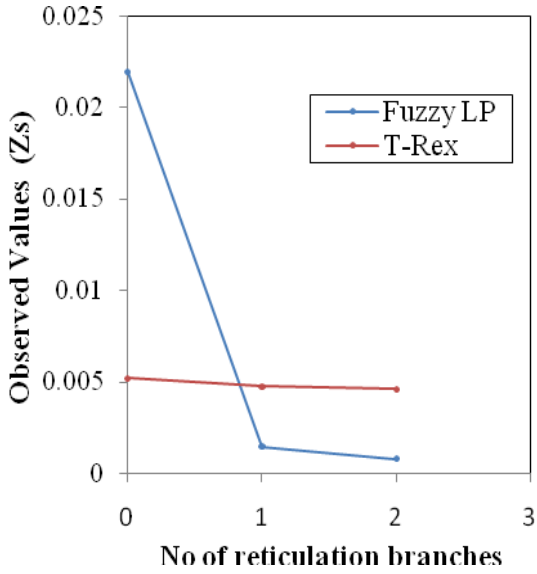

Fig. 5 The difference between the values obtained for reticulation branches by FLP and T-Rex.

5 Conclusion

This study has developed a new measure based on the frequency of nucleotides to calculate the dissimilarity between DNA sequences. The proposed measure has no need of sequence alignment to get the dissimilarity matrix. The importance of the method is that the chances of losing information were reduced that can occur during the alignment of sequences and it can handle the large dataset of sequences.

A fuzzy linear programming has also been purposed in this study to construct the phylogenetic network. The FLP problem gives the fast solution as it analyzes only extant species for the phylogenetic network.

Thus, a complete model based on new dissimilarity measure and FLP has been developed for the construction of network in this study. The model is validated over the β- globin gene of nine species. The value of Z is optimized to 0.4160 after the addition of two new branches to the tree. The results obtained by this method are in full agreement with the existing methods. In the end, it has been pointed out that the method is very simple, fast and can be used for the analysis of short and long DNA sequences with high efficiency. Thus, it has been expected that this method will be fruitful for the biological community to find out the complex relationships among the organisms without going directly to the wet lab.

References

Aita T, Husimi Y, Nishigaki K. 2011. A mathematical consideration of the word-composition vector method in comparison of biological sequences. Biosystems, 106: 67-75

Bandelt HJ, Forster P, Rohl A. 1999. Median-joining networks for inferring intraspecific phylogenies. Molecular Biology and Evolution, 16: 37-48

Chan HL, Jansson J, Lam T-W, Yiu SM. 2005. Reconstructing an ultrametric galled phylogenetic network from a distance matrix. In: MFCS 2005, LNCS 3618 (Jedrzejowicz J, Szepietowski A, eds). 224-235, Springer-Verlag, Berlin, Heidelberg, Germany

Dress A, Moulton V, Steel M, Wu T. 2010. Species, clusters and the 'Tree of life': a graph-theoretic perspective. Journal of Theoretical Biology, 265: 535-542

Eslahchi C, Habibi M, Hassanzadeh R, Mottaghi E. 2010. MC-Net: a method for the construction of phylogenetic networks based on the Monte-Carlo method. BMC Evolutionary Biology, 10: 254

Grzegorzewski P. 2004. Distances between intuitionistic fuzzy sets and/or interval-valued fuzzy sets based on the hausdorff metric. Fuzzy Sets and Systems, 148: 319-328

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution, 16: 111-120

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Liao B, Wang TM. 2004. Analysis of similarity/dissimilarity of DNA sequences based on 3-D graphical representation. Chemical Physics Letters, 388: 195-200

Liu N, Wang T. 2006. A weighted measure for the similarity analysis of DNA sequences. Journal of Molecular Modeling, 12: 897-903

Liu XQ, Dai Q, Xiu Z, Wang T. 2006. PNN-curve: a new 2D graphical representation of DNA sequences and its application. Journal of Theoretical Biology, 243: 555-561

Lockhart PJ, Steel MA, Hendy MD, Penny D. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Molecular Biology and Evolution, 11: 605-612

Makarenkov V, Legendre P, Desdevises Y. 2004. Modelling phylogenetic relationships using reticulated networks. Zoologica Scripta, 33: 89-96

Makarenkov V, Legendre P. 2004. From a phylogenetic tree to a reticulated network. Journal of Computational Biology, 11: 195-212

Makarenkov V. 2001. T-REX: reconstructing and visualizing phylogenetic trees and reticulation networks. Bioinformatics, 17: 664-68

Mathur R, Adlakha N. 2013. A fuzzy weighted least squares approach to construct phylogenetic network among subfamilies of grass species. Journal of Applied Mathematics and Bioinformatics, 3: 137-158

Mathur R, Adlakha N. 2014. Linear programming model to construct phylogenetic network for 16S Rrna sequences of photosynthetic organisms and influenza viruses. Interdisciplinary Sciences: Computational Life Sciences, 6: 100-107

Mohaddes SA, Mohayidin MG. 2008. Application of the fuzzy approach for agricultural production planning in a watershed, a case study of the atrak watershed, Iran. American-Eurasian Journal of Agricultural and Environmental Sciences, 3: 636-648

Morrison DA. 2013. Phylogenetic Networks are Fundamentally Different From Other Kinds of Biological Network. In: Network Biology: Theories, Methods and Applications (WenJun Zhang, ed). 23-68, Nova Science Publishers, New York, USA

Qi X, Fuller E, Wu Q, Zhang CQ. 2012. Numerical characterization of DNA sequence based on dinucleotides. The Scientific World Journal, 104269

Qi X, Wu Q, Zhang Y, Fuller E, Zhang CQ. 2011. A novel model for DNA sequence similarity analysis based on graph theory. Evolutionary Bioinformatics, 7: 149-158

Qi ZH, Fan TR. 2007. PN-curve: A 3D graphical representation of DNA sequences and their numerical characterization. Chemical Physics Letters, 442: 434-440

Semple C, Steel M. 2003. Phylogenetics. Oxford University Press, UK

Sridhar S, Lam F, Blelloch GE, Ravi R, Schwartz R. 2007. Efficiently finding the most parsimonious phylogenetic tree via linear programming. In: ISBRA 2007, LNBI 4463(Mandoiu I, Zelikovsky A, eds). 37-48, Springer-Verlag, Berlin, Heidelberg, Germany

Tang J, Moret BME. 2005. Linear Programming for Phylogenetic Reconstruction Based On Gene Rearrangements. Springer-Verlag, Berlin, Heidelberg, Germany

Tusserkani R, Eslahchi C, Pourmohammadi H, Azadi A. 2011. TripNet: A method for constructing phylogenetic networks from triplets. arXiv: 1104.4720v1 [cs.CE]

Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. 2007. Clustal W and Clustal X version 2.0. Bioniformatics, 23: 2947-2948

Zhang WJ. 2015. Prediction of missing connections in the network: A node-similarity based algorithm. Selforganizology, 2(4): 91-101

Zhang WJ. 2016. A node-similarity based algorithm for tree generation and evolution. Network Biology, 6(3): 55-64

Zhang WJ. 2018. Fundamentals of Network Biology. World Scientific Europe, London, UK

Zhu W, Liao B, Li R. 2010. A method for constructing phylogenetic tree based on a dissimilarity matrix. MATCH Communications in Mathematical and in Computer Chemistry, 63: 483-492

Zimmermann HJ. 2001. Fuzzy Set Theory and Its Applications (4th ed). Springer-Verlag, Netherlands