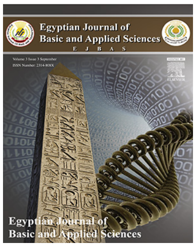
HOSTED BY

ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: http://ees.elsevier.com/ejbas/default.asp



Full Length Article

A graph theoretic model for prediction of reticulation events and phylogenetic networks for DNA sequences

Rinku Mathur *, *Neeru Adlakha*

Department of Applied Mathematics and Humanities, Sardar Vallabhbhai National Institute of Technology, Surat 395007, Gujarat, India

ARTICLE INFO

Article history:
Received 25 July 2015
Received in revised form 14 July 2016
Accepted 14 July 2016
Available online 20 July 2016

Keywords:
DNA sequences
Directed weighted graph
Distance matrix
Phylogenetic networks
Reticulation events

ABSTRACT

Phylogenies are the commonly used tools for the prediction of ancestry of present day organisms from the past decades. Several methods have been developed to construct phylogenetic trees that predict the history of species by direct linkage of edges. Very few studies have been developed for the phylogenetic networks (which is the generalization of trees). Presently, the methods used to determine phylogenetic networks are based on distance measures or character measures of the sequences of species. It is a very challenging task for computational biologists to find the exact method that can predict the accurate networks of organisms. In this study, a phylogenetic network construction model based on basic graph theory concepts is reported. This model finds the distance matrix of every sequence considered in the study. The two features (positioning and stack interactions) of every DNA sequence and their combined effect have been taken into account to calculate the distances. Results suggested that reticulate events can be observed by using the distances obtained by the proposed method and no such event is predicted by using the distances calculated by the previous method. The important results obtained in the form of distances are 1.637300, 2.000000, 0.932700, 2.331300, 2.829200 and the significance of these values is to represent the different reticulation events among the sequences for different features. Hence distances calculated by this model gives better insights to study the phylogenetic networks.

1. Introduction

The evolutionary and ancestry relationships have been investigated among the number of species since several decades in the past. The phylogenetic analysis of DNA or protein sequences is performed by most of the studies to determine the evolutionary history of organisms [1–3]. Phylogenies are the important tools to predict the evolutionary history and relationships of different organisms. The estimation of divergence of species

* Corresponding author.
E-mail address: rinkumathur56@gmail.com (R. Mathur).

http://dx.doi.org/10.1016/j.ejbas.2016.07.004

from a common ancestor can be predicted by various methods [4–6]. This estimation permits one to construct a phylogenetic tree that represents the history of species according to their ancestry. The most commonly used methods for tree construction are phenetic and cladistic [7,8]. The former is based on distances and measures the pair-wise distances between organisms to construct the phylogenetic tree. The latter is based on characters and constructs the tree by using the parsimonious methods [8,9]. Lastly, the tree that optimizes the evolution of species most is considered [1].

However, most of the available methods are unreliable when the base composition of taxa or species varies among the sequences. Various methods work on the computation of distances among the two or more sequences based on their nucleotide composition, binary composition or sequence similarity, etc. These methods generally use the Jukes Cantor (JC) distance, Kimura two parameter (K2P) distance, F84 distance, LogDet distance, etc. to find the phylogeny among the taxas [10–12]. Qi et al., 2011 reported a graph theory based method for similarity analysis. But no method based on graph theoretic concepts similar to those of Qi et al., 2011 is reported for identification of reticulation events and construction of phylogenetic networks.

In the present study, the authors modify the method developed by Qi et al., 2011 by incorporating two more features of a DNA sequence that are used to find the distance matrix of the sequence based on the graph theory and hence extend it to develop a model for identification of reticulation events and construction of phylogenetic network [13]. Every nucleotide sequence is composed of four base pairs, namely, adenine (A), guanine (G), cytosine (C) and thymine (T). Hence the adjacency matrix of the order 4×4 is obtained for both the features of every sequence. After the computation of distance matrices of all the available sequences of the organisms, the Euclidean distance among all the matrices of the same order is calculated to determine the phylogeny [10,14].

The phylogenetic tree is obtained by the average clustering or neighbor-joining method [5] and then reticulation events are predicted among the taxas to construct the phylogenetic network and to predict the interactions between the different species. These interactions may be of the Lateral Gene Transfer (LGT), Homoplasy, Hybridization, Genetic Crossover, Recombination type, etc. [14,15]. As an application, the method is applied over the data set of nine sequences and is validated by comparison with the available phylogenetic network construction method. The basic concepts used in the present study are defined in the subsequent sections.

2. Basic definitions

2.1. *Directed graph*

A directed graph G consists of a set of vertices $V = \{v_1, v_2, v_3, \ldots\ldots\ldots\}$, a set of edges $E = \{e_1, e_2, e_3, \ldots\ldots\ldots\}$ and a mapping that maps every edge onto some ordered pair of vertices (v_i, v_j) [16].

2.2. *Adjacency matrix*

Let G be a directed graph with n vertices and E edges [16]. Then the adjacency matrix $M = [x_{ij}]$ of the diagraph G is an $n \times n$ matrix whose element

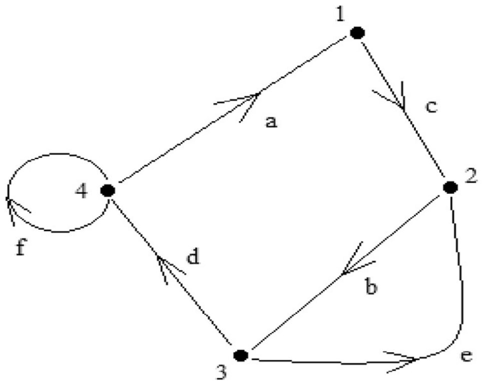
$$x_{ij} = 1, \quad \text{if there is an edge directed from } i\text{th vertex to } j\text{th vertex.}$$
$$= 0, \quad \text{otherwise.}$$

A directed graph and its adjacency matrix is shown in Fig. 1.

2.3. *Stack interactions*

The role of hydrogen bonding and stack interactions is considered as an important role in biological structures [17]. So, the DNA structure is also maintained by intra-strand base stacking interactions. Henceforth, the stacking interaction and length of sequence play vital roles in the stability of DNA sequence. The average stacking interactions of GC base pairs are two or three times stronger than AT base pairs [18,19].

In comparison to other interactions (hydrogen bonds and hydrophobic interactions) involved in holding double stranded DNA sequence together, the energies of the stacking interactions in DNA sequence are significantly large [18,19]. The probability of these interactions is considered in the study to calculate the distances among sequences.



$$M = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

Fig. 1 – Directed graph and its corresponding adjacency matrix.

2.4. *Euclidean distance matrix*

A matrix M of order $n \times n$ is called a Euclidean distance matrix if it satisfies the following conditions [20]:

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

(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

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

2.5. *Phylogenetic tree*

Any connected graph with a unique path between any two distinct vertices u and v and no cycles is called a tree and is denoted as T. A tree T has exactly $|V| - 1$ edges. The degree of a vertex v, $\deg(v)$, is defined to be the number of edges attached to v. In a tree, any vertex v with $\deg(v) = 1$ is called a leaf [21,22].

2.6. *Phylogenetic network*

Any phylogenetic tree with edge lengths and cycles is called a network. It is represented as a triplet (V, E, l), where l is a function of edge lengths assigning real nonnegative numbers to the edges. The creation of new sequence x from parent nodes Z and z' at a recombination node is called a “recombination or reticulation event” [23,24].

3. Methodology

The DNA sequences are composed of the four bases adenine, guanine, cytosine and thymine, i.e. A, G, C, T. Let us assume that any sequence of length n having the four characters can be represented as:

$$S = s_1, s_2, s_3, \ldots\ldots\ldots\ldots, s_n \quad \text{such that} \quad s_i \in \{A, C, G, T\}.$$

To describe the process of getting the distance matrix of a sequence by the concept of adjacency of graph theory, authors will construct the directed weighted graph (DWG) for $s_1, s_2, s_3, \ldots\ldots\ldots\ldots, s_n$ which is denoted by $G_w = (V(G_w), E(G_w))$. The vertex set $V(G_w) = \{A, C, G, T\}$ and the edge set $E(G_w)$ comprise pairs of two vertices. For each pair of nucleotides S_i and S_j in S with $i < j$, put an arc from S_i to S_j with the weight as $\left({}^{1}\!/_{j-i}\right)$ so that $\left({}^{1}\!/_{j-i}\right)$ is a decreasing function of $(j - i)$ which reflects the fact that the two nucleotides with smaller distance would have stronger interactive relationship than those with bigger distance.

In this work, the two features of the nucleotide sequences have been incorporated in the model; one is based on the position of the nucleotides and the other is based on the stack interactions among the nucleotides of the sequences. In case of second feature, the directed weighted graph of the sequence is constructed by adding to it the content of probability of interactions among nucleotides [14]. The rule used to construct the DWG for the feature of stack interaction is defined as:

$$\text{Stack Interaction} = \frac{\text{Probability of interaction among nucleotides}}{\text{Position among them}}$$

The three probabilities for the interactions among the nucleotides are considered which are given below:

Probability of conversion from purines (A, G) to pyrimidines (C, T) or vice-versa which is called transversions is 1/2.
Probability of conversion from purines to purines or pyrimidines to pyrimidines which is called transitions is 1/3.
Probability of conversion into the same nucleotide is 1/6.

These probabilities are assumed according to the biological effects of these conversions on the evolution of species and the total probability is considered as 1.

Let us take a sequence of length five as an illustration to construct the directed weighted graph for both the features analyzed in the present study.

Sequence = CGATA

Theorem 3.1. There exists one–one mapping between the DNA sequence and its corresponding directed weighted graph with respect to the positioning of nucleotides [13].

Proof. Let the DNA sequence for which a graph is to be constructed with respect to the position of nucleotides be CGATA. Then the graph represented by this sequence with the mapping from vertices (nucleotides) to edges (pair of nucleotides) is defined by $\left({}^{1}\!/_{j-i}\right)$. The graph obtained by using this mapping is shown in Fig. 2.

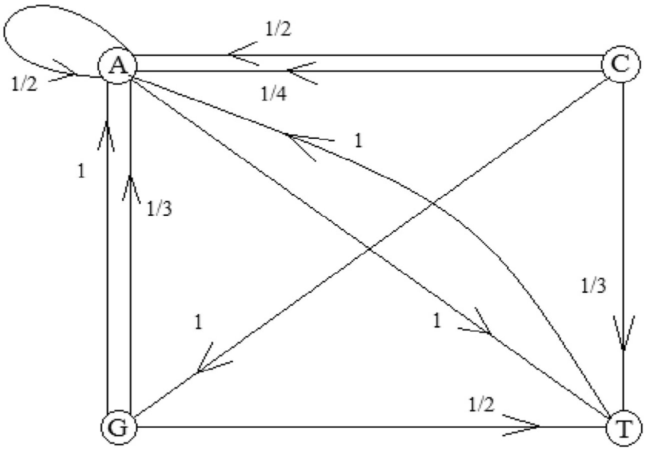


Fig. 2 – Directed weighted graph with respect to the position of the nucleotides in the sequence CGATA [5].

From the graph, one can find the first and the last nucleotides of the sequence by using the lowest weight of the arc in the graph and the direction of the arc signifies which nucleotide comes first in the sequence. Also, the loop in the graph signifies how many times the same nucleotide comes in the sequence. Further, the other positions of the nucleotides can be determined by using the weights of the particular arcs in the graph. Hence, it is easy to find the sequence from the directed weighted graph. Thus there exists one–one correspondence between DNA sequence and directed weighted graph.

3.1. *The resultant graph*

Since there are more than one arc between the same nucleotides in the same direction, to make this graph simple, authors added all the weights of the arcs occurring in the same direction. The mathematical equation for the addition of all these weights is as [13]:

$$W_R(u,v)=\sum W_w(u,v) \tag{1}$$

So, based on this rule, the resultant graph is shown in Fig. 3.

It is also observed in this study that there is no one–one correspondence between the given sequence and its weighted graph obtained for the parameter of stack interactions among the nucleotides of the sequence. But the process of construction of graph remains same and here the probabilities of conversion of nucleotides (transversion to transversion or transition to transition or vice versa) are included with the position of the nucleotides in the sequence.

Now from the resultant graph of DNA sequence obtained from the above discussion, one can obtain the 4 × 4 adjacency matrix of the sequence which is represented as below. This matrix will also show the interactions among the different nucleotides of the sequence. The higher the value of the particular entry corresponding to the pair of nucleotides in the matrix, the higher will be the interaction between them and vice-versa.

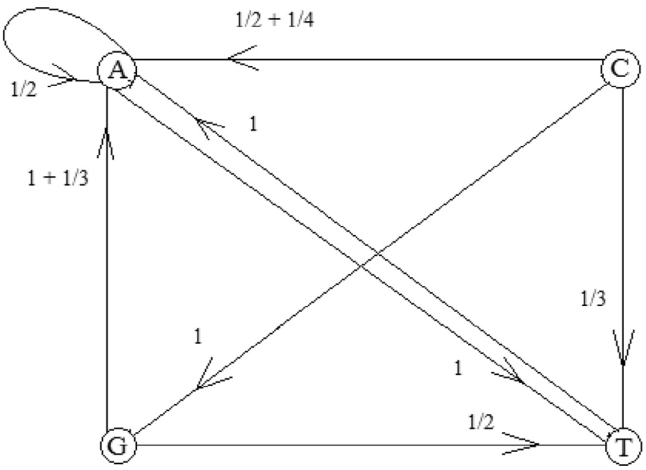


Fig. 3 – Resultant directed weighted graph with respect to the position of the nucleotides in the sequence CGATA.

$$M=\begin{array}{c|cccc} & A & C & G & T \\ \hline A & a_{11} & a_{12} & a_{13} & a_{14} \\ C & a_{21} & a_{22} & a_{23} & a_{24} \\ G & a_{31} & a_{32} & a_{33} & a_{34} \\ T & a_{41} & a_{42} & a_{43} & a_{44} \end{array}$$

The matrices obtained for the parameter of position of nucleotides of the sequence CGATA and the parameter of stack interactions among them are as below:

$$M_P=\begin{pmatrix} 0.5 & 0 & 0 & 1.0 \\ 0.75 & 0 & 1.0 & 0.3334 \\ 1.3334 & 0 & 0 & 0.5 \\ 1.0 & 0 & 0 & 0 \end{pmatrix} \quad M_{SI}=\begin{pmatrix} 0.0833 & 0 & 0 & 0.5 \\ 0.375 & 0 & 0.5 & 0.1111 \\ 0.4444 & 0 & 0 & 0.25 \\ 0.5 & 0 & 0 & 0 \end{pmatrix}$$

So, by using these matrices obtained for every sequence considered in the study, one can find the distances among the two sequences by using the appropriate distance metrics.

3.2. *Euclidean matrix distance metric to calculate dissimilarity*

In the above section, a distance matrix for a DNA sequence is derived based on two parameters, i.e. positioning and stack interactions, by using the graph-theoretic method. So, the dissimilarity between two DNA sequences becomes the same as between the matrices formed by them. The smaller the distance between two sequences, the less dissimilar they are and vice-versa. Let us consider two DNA sequences A and B and the two matrices corresponding to them are as M^A and M^B for different parameters.

The distance measurement for the two sequences is based on the Euclidean distance matrices among the sequences and is denoted by $d(A,B)$ [20]. The distance $d(A,B)$ is defined as:

$$d(A,B)=\sqrt{\sum_{i=1}^{4}\sum_{j=1}^{4}\left(M_{ij}^{A}-M_{ij}^{B}\right)^2} \tag{2}$$

where $M_{ij}^{A}-M_{ij}^{B}$ represents the difference between the same entries of the matrices A and B.

In the present work, the two features are considered to calculate the distances among the sequences. As a result of which the total dissimilarity $d_T(A,B)$ among the sequences A and B is defined as [25]:

$$d_T(A,B)=d_P(A,B)+d_{SI}(A,B) \tag{3}$$

where $d_P(A,B)$ is the dissimilarity due to position and $d_{SI}(A,B)$ is the dissimilarity due to stack interactions among sequences.

The comparison between the DNA sequences for their dissimilarity is measured by the distances among them. The distances based on all three features are considered for the detection of the phylogenetic network among the given sequences.

Table 1 – Input data sequences with their labels.

Label	Sequence
I	ACAAG
II	ACCAG
III	GACAA
IV	AGCAA
V	AGACA
VI	AGATA
VII	CGATA
VIII	AGGTA
IX	CGGTA

4. Implication of proposed method

4.1. *Data analysis*

The data to be analyzed for this study are composed of a set of nine sequences which is provided in the Table 1.

Based on the above graph-theoretic method, the 4 × 4 matrices for all the sequences of Table 1 are computed manually. Then these matrices were used to calculate the distances among all the sequences by using the above metric. The distance tables obtained for the two parameters of the sequences and the combined result of these parameters are represented in Tables 2–4.

After getting these distances, authors were able to calculate the relationships among the DNA sequences based on the above technique.

5. Results and discussion

The dissimilarity matrices obtained by the proposed method for different features are used to analyze the sequences for their relationships. Various methods can be used to find the distances but these are based on multi-alignment of sequences where the chances of losing information about sequences increase. The proposed study finds the distance matrix of every

Table 2 – Dissimilarity distances among nine sequences based on their positions of nucleotides.

Sequence labels	I	II	III	IV	V	VI	VII	VIII	IX
I	0.0000	2.0000	2.4353	1.6373	1.8448	3.0116	3.3146	3.3146	3.9247
II	2.0000	0.0000	2.9179	2.4523	2.2142	3.4137	3.0935	3.3645	3.4621
III	2.4353	2.9179	0.0000	1.6583	1.3280	2.8161	2.6484	3.6305	3.6535
IV	1.6373	2.4523	1.6583	0.0000	1.2473	2.6247	2.7563	3.1447	3.5959
V	1.8448	2.2142	1.3280	1.2473	0.0000	2.4608	2.6484	2.9815	3.4662
VI	3.0116	3.4137	2.8161	2.6247	2.4608	0.0000	1.8296	2.1213	3.0024
VII	3.3146	3.0935	2.6484	2.7563	2.6484	1.8296	0.0000	2.5847	2.0000
VIII	3.3146	3.3645	3.6305	3.1447	2.9815	2.1213	2.5847	0.0000	2.2017
IX	3.9247	3.4621	3.6535	3.5959	3.4662	3.0024	2.0000	2.2017	0.0000

Table 3 – Dissimilarity distances among nine sequences based on the stack interactions of nucleotides.

Sequence labels	I	II	III	IV	V	VI	VII	VIII	IX
I	0.0000	0.4930	0.8421	0.6940	0.6709	1.4539	1.2898	1.3646	1.5311
II	0.4930	0.0000	0.9171	0.9104	0.7489	1.5994	1.2979	1.5096	1.4827
III	0.8421	0.9171	0.0000	0.5481	0.4668	1.3265	1.1551	1.4578	1.5080
IV	0.6940	0.9104	0.5481	0.0000	0.5740	1.3365	1.2039	1.3492	1.5162
V	0.6709	0.7489	0.4668	0.5740	0.0000	1.2304	1.2142	1.3071	1.4984
VI	1.4539	1.5994	1.3265	1.3365	1.2304	0.0000	0.7466	0.7818	1.2227
VII	1.2898	1.2979	1.1551	1.2039	1.2142	0.7466	0.0000	1.0350	0.8292
VIII	1.3646	1.5096	1.4578	1.3492	1.3071	0.7818	1.0350	0.0000	0.9327
IX	1.5311	1.4827	1.5080	1.5162	1.4984	1.2227	0.8292	0.9327	0.0000

Table 4 – Total dissimilarity distances among sequences based on position and stack interaction of their nucleotides.

Sequence labels	I	II	III	IV	V	VI	VII	VIII	IX
I	0.0000	2.4930	3.2774	2.3313	2.5157	4.4655	4.6044	4.6792	5.4558
II	2.4930	0.0000	3.8350	3.3627	2.9631	5.0131	4.3914	4.8741	4.9448
III	3.2774	3.8350	0.0000	2.2064	1.7948	4.1426	3.8035	5.0883	5.1615
IV	2.3313	3.3627	2.2064	0.0000	1.8213	3.9612	3.9602	4.4939	5.1121
V	2.5157	2.9631	1.7948	1.8213	0.0000	3.6912	3.8626	4.2886	4.9646
VI	4.4655	5.0131	4.1426	3.9612	3.6912	0.0000	2.5762	2.9031	4.2251
VII	4.6044	4.3914	3.8035	3.9602	3.8626	2.5762	0.0000	3.6197	2.8292
VIII	4.6792	4.8741	5.0883	4.4939	4.2886	2.9031	3.6197	0.0000	3.1344
IX	5.4558	4.9448	5.1615	5.1121	4.9646	4.2251	2.8292	3.1344	0.0000

Table 5 – Phylogenetic network distance matrix for the data of positions of nucleotides in the sequences.

Sequence labels	I	II	III	IV	V	VI	VII	VIII	IX
I	0.000000	1.999973	2.390675	**1.637300**	1.976846	2.958965	2.966851	3.416662	3.697089
II	1.999973	0.000000	2.595716	2.280868	2.181887	3.164006	3.171893	3.273086	3.553514
III	2.390675	2.595716	0.000000	1.659678	1.328004	2.977556	2.985442	3.435252	3.715680
IV	**1.637300**	2.280868	1.659678	0.000000	1.245849	2.662707	2.670593	3.120404	3.400831
V	1.976846	2.181887	1.328004	1.245849	0.000000	2.563727	2.571613	3.021423	3.301851
VI	2.958965	3.164006	2.977556	2.662707	2.563727	0.000000	1.924397	2.374207	2.654635
VII	2.966851	3.171893	2.985442	2.670593	2.571613	1.924397	0.000000	2.152155	**2.000000**
VIII	3.416662	3.273086	3.435252	3.120404	3.021423	2.374207	2.152155	0.000000	2.201702
IX	3.697089	3.553514	3.715680	3.400831	3.301851	2.654635	**2.000000**	2.201702	0.000000

sequence by which the complete information about the sequences can be availed. Then, the phylogenetic tree is calculated for all the distance measures by using the neighbor-joining method of T-Rex software [6]. After finding the trees, the tree distance forms the basis for the construction of the network. The phylogenetic networks are calculated to predict the relationships among different sequences of organisms. Table 5 shows the phylogenetic network distances for the input set of sequences given in Table 1, for the positioning feature of the nucleotides in the sequences.

The significant results of Table 5 are made bold and have values 1.637300 and 2.000000. These values represent branch lengths among the pairs of sequences corresponding to them and there is also the possibility of reticulation events among these sequences. The sequences corresponding to 1.637300 are sequence I and sequence IV. The sequences corresponding to 2.000000 are sequence VII and sequence IX. Phylogenetic network related to this dissimilarity data of Table 5 is also shown in Fig. 4 in which red dotted branches show the reticulated events among sequences.

In a similar fashion, the dissimilarity network distances for the stack interactions feature of nucleotides in the sequences is given in Table 6. The significant value for the feature of stack interactions is 0.932700 (in bold) among the sequences VII and IX. Corresponding to these data, the visualization of the network relationships is represented in Fig. 5.

Table 7 shows the reticulated network distances for the combined result of both the features of the nucleotides in the sequences. The values 2.331300 and 2.829200 (in bold) of Table 7 represent the occurrence of reticulation events among the sequences corresponding to them. In response to these distances, the representation of phylogenetic network is shown in Fig. 6.

Finally, the phylogenetic network distances obtained by the T-Rex package for the input data sequences are shown by Table 8 and Fig. 7 depicts the network for these sequences which is the same as Phylogenetic tree for the input sequences.

Now, it is observed that Fig. 4 represents three main clusters and the three reticulation events while Fig. 7 shows no reticulation event with two main clusters and one species as an out-group. This information tells us that positioning feature gives us the close picture of the evolutionary history of the organisms. The events marked by dotted lines in Fig. 4 show that there exist some important relationships among the non-sister species. This information could be useful for prediction of pathogens and antigens based on evolutionary relationships.

In the same way, Fig. 5 depicts some results that are different from Fig. 7 in that one reticulation event and three main clusters occur due to the parameter of stack interactions among

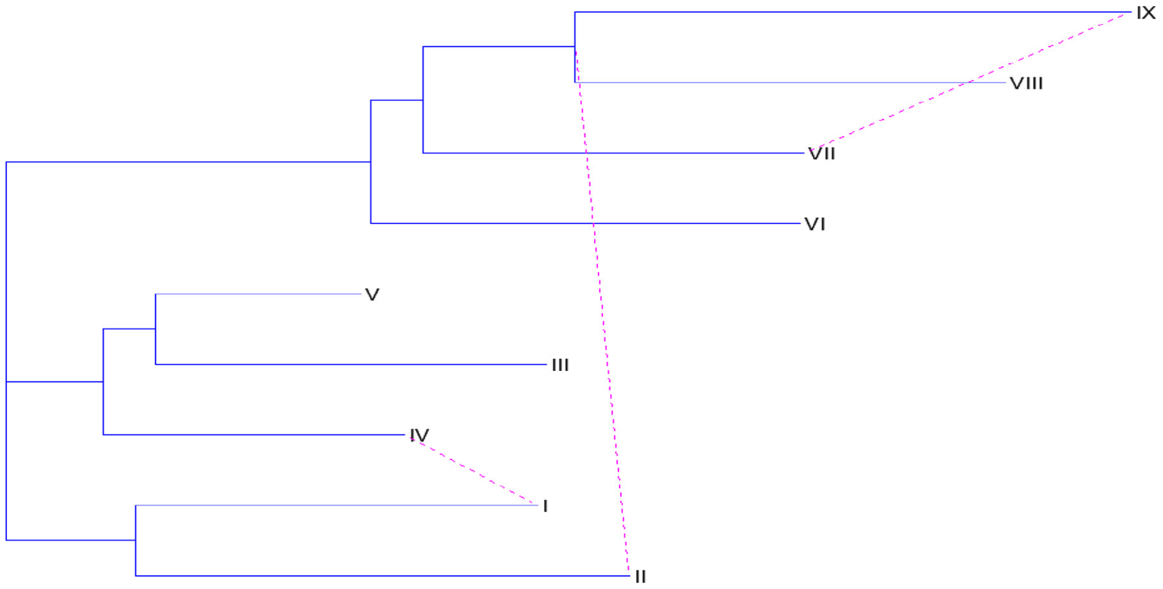


Fig. 4 – Phylogenetic network for the data of positions of nucleotides in the sequences.

Table 6 – Phylogenetic network distance matrix for the data of stack interactions of nucleotides in the sequences.

Sequence labels	I	II	III	IV	V	VI	VII	VIII	IX
I	0.000000	0.492997	0.785467	0.760973	0.712495	1.405895	1.260032	1.411610	1.509931
II	0.492997	0.000000	0.873981	0.849487	0.801009	1.494409	1.348546	1.500124	1.598445
III	0.785467	0.873981	0.000000	0.597528	0.466801	1.383459	1.237596	1.389174	1.487495
IV	0.760973	0.849487	0.597528	0.000000	0.524556	1.358965	1.213102	1.364680	1.463001
V	0.712495	0.801009	0.466801	0.524556	0.000000	1.310488	1.164625	1.316202	1.414524
VI	1.405895	1.494409	1.383459	1.358965	1.310488	0.000000	0.856443	0.781800	1.106342
VII	1.260032	1.348546	1.237596	1.213102	1.164625	0.856443	0.000000	0.862157	0.829205
VIII	1.411610	1.500124	1.389174	1.364680	1.316202	0.781800	0.862157	0.000000	**0.932700**
IX	1.509931	1.598445	1.487495	1.463001	1.414524	1.106342	0.829205	**0.932700**	0.000000

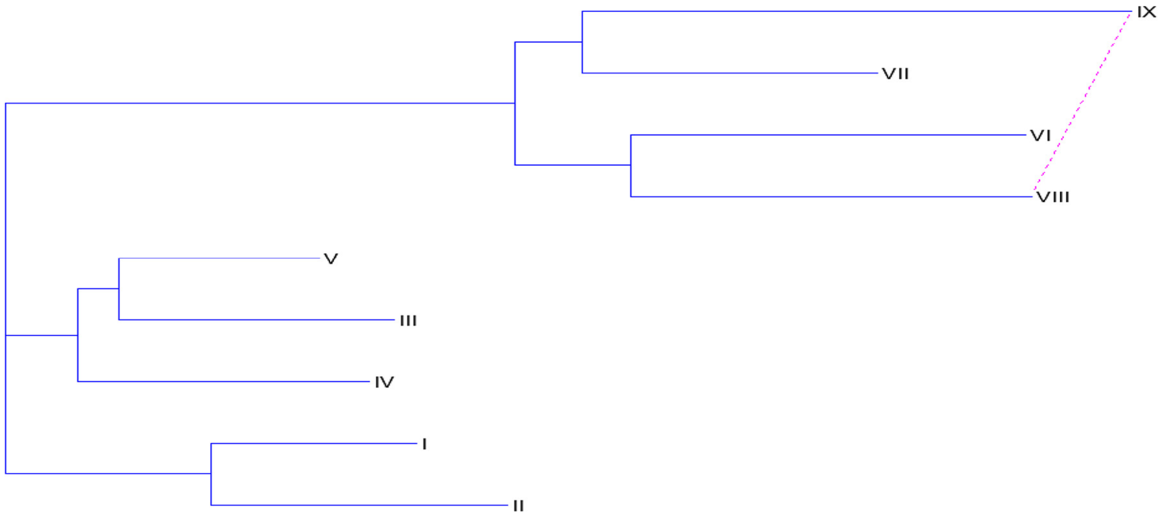


Fig. 5 – Phylogenetic network for the data of stack interactions of nucleotides in the sequences.

Table 7 – Phylogenetic network distance matrix for the combined data of positions and stack interactions of nucleotides in the sequences.

Sequence labels	I	II	III	IV	V	VI	VII	VIII	IX
I	0.000000	2.492960	3.176137	**2.331300**	2.689338	4.363562	4.210355	4.830406	5.222720
II	2.492960	0.000000	3.469693	3.130351	2.982893	4.657118	4.503911	4.713293	5.105607
III	3.176137	3.469693	0.000000	2.257206	1.794805	4.359715	4.206508	4.826559	5.218873
IV	**2.331300**	3.130351	2.257206	0.000000	1.770407	4.020373	3.867166	4.487217	4.879530
V	2.689338	2.982893	1.794805	1.770407	0.000000	3.872916	3.719709	4.339759	4.732073
VI	4.363562	4.657118	4.359715	4.020373	3.872916	0.000000	2.690632	3.310683	3.702996
VII	4.210355	4.503911	4.206508	3.867166	3.719709	2.690632	0.000000	3.028319	**2.829200**
VIII	4.830406	4.713293	4.826559	4.487217	4.339759	3.310683	3.028319	0.000000	3.134398
IX	5.222720	5.105607	5.218873	4.879530	4.732073	3.702996	**2.829200**	3.134398	0.000000

Table 8 – Phylogenetic network distance matrix for the original input sequences of nucleotides.

Sequence labels	I	II	III	IV	V	VI	VII	VIII	IX
I	0.000000	0.450416	1.229796	1.131206	1.257645	1.108086	1.112828	1.178000	1.135730
II	0.450416	0.000000	0.779400	0.680810	0.807249	0.657690	0.662432	0.727604	0.685335
III	1.229796	0.779400	0.000000	0.785452	0.911891	0.762332	0.767074	0.832246	0.789977
IV	1.131206	0.680810	0.785452	0.000000	0.748414	0.598855	0.603597	0.668769	0.626500
V	1.257645	0.807249	0.911891	0.748414	0.000000	0.255186	0.487507	0.631705	0.589435
VI	1.108086	0.657690	0.762332	0.598855	0.255186	0.000000	0.337948	0.482146	0.439877
VII	1.112828	0.662432	0.767074	0.603597	0.487507	0.337948	0.000000	0.486888	0.444618
VIII	1.178000	0.727604	0.832246	0.668769	0.631705	0.482146	0.486888	0.000000	0.239156
IX	1.135730	0.685335	0.789977	0.626500	0.589435	0.439877	0.444618	0.239156	0.000000

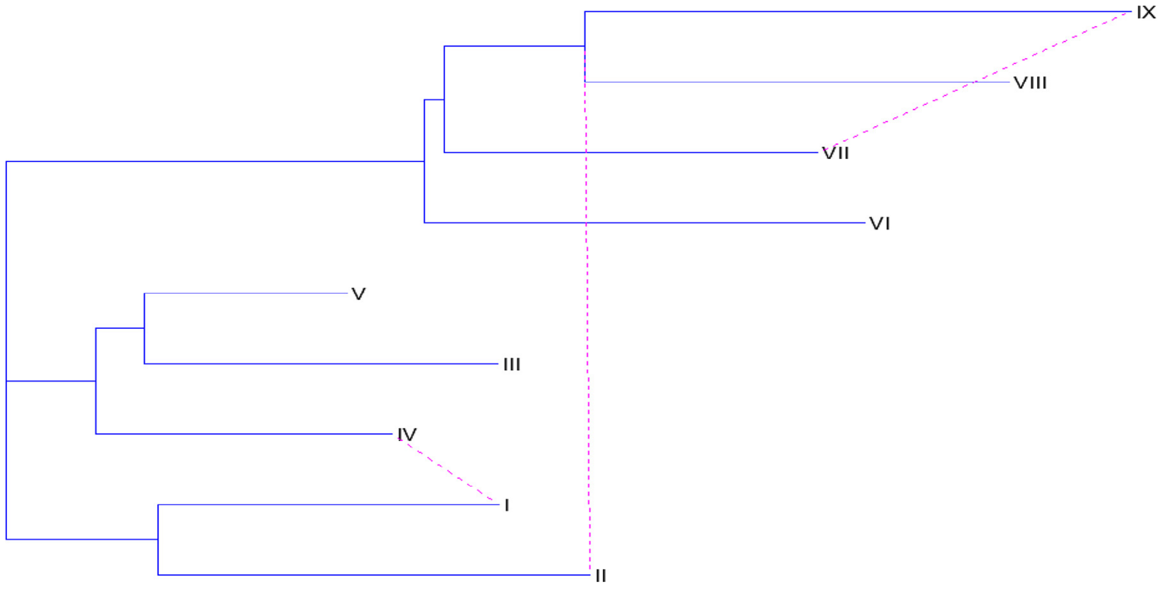


Fig. 6 – Phylogenetic network for the data of positions and stack interactions of nucleotides in the sequences.

the nucleotides. Irrespective of the same number of clusters, the species change their place of interactions with the other species in the main cluster of four species.

Likewise, the combined result of both the features (positions and stack interactions of nucleotides) highlighted by Fig. 6 shows three reticulation events among the clusters while Fig. 7 has no such type of events.

At last, authors can comment that the loss of reticulation events in Fig. 7 when data sequences are inputted is due to the alignment of sequences and then converting them into distances. But the proposed graph theory method converts each sequence into the form of matrix distances without aligning them. Hence, the possibility of losing information is decreased and the clear picture for the evolution of species can be predicted.

6. Conclusion

The proposed method gives better results and better insights about the history of organisms. It also provides the individual information of every sequence in the form of distances and thus saves the time of multi-alignment of the sequences. The results obtained by this study were compared with the existing method of T-Rex Package [6] which shows the loss of reticulation events in network. The results obtained are different for different features of nucleotide sequences. The reticulation events predicted for positioning, stack interactions and combined effect of both of them are three, one and three which are shown in Figs. 4–6. Henceforth, the proposed

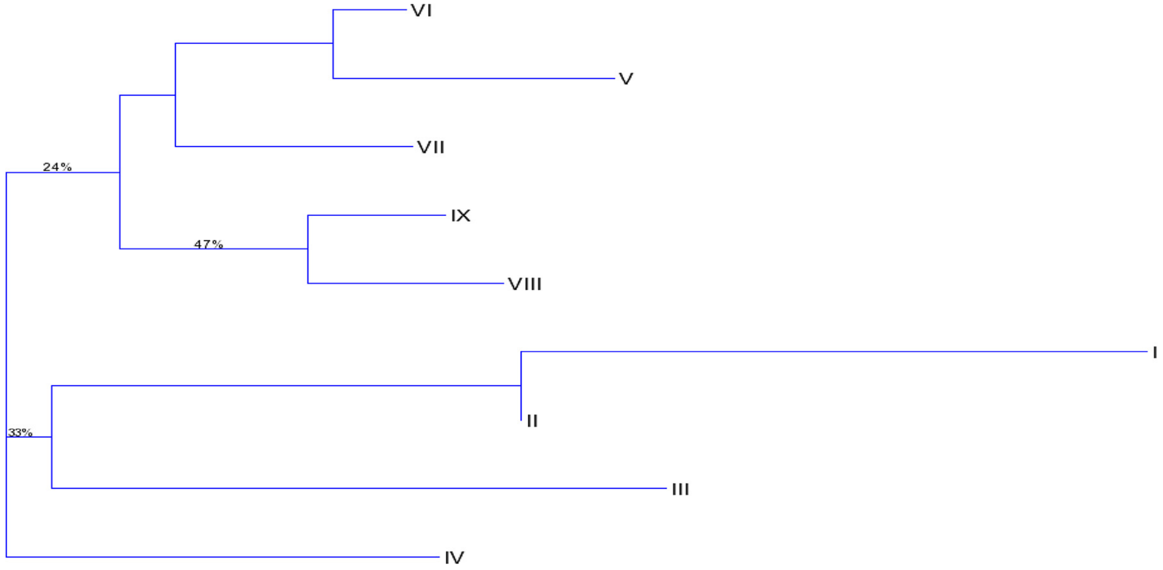


Fig. 7 – Phylogenetic network for the input sequences of nucleotides.

model performed well in determining the reticulation events and gives us better results as well.

REFERENCES

[1] Heymans M, Singh AK. Deriving phylogenetic trees from the similarity analysis of metabolic pathways. Bioinformatics 2003;19:i138–46.

[2] Bordewich M, Huber KT, Semple C. Identifying phylogenetic trees. Discrete Math 2005;300:30–43.

[3] Brandes U, Cornelson S. Phylogenetic graph models beyond trees. Discrete Appl Math 2009;157:2361–9.

[4] Dress A, Huson D, Moultan V. Analysing and visualizing sequence and distance data using SplitsTree. Discrete Appl Math 1996;71:95–109.

[5] Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 1987;4:406–25.

[6] Makarenkov V. T-Rex: reconstructing and visualizing phylogenetic trees and reticulation networks. Bioinformatics 2001;17:664–8.

[7] Bryant D, Moulton V. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol Biol Evol 2004;21:255–65.

[8] Nakhleh L, Jin G, Zhao F, Crummey JM. Reconstructing phylogenetic networks using maximum parsimony. In: Proceeding of the IEEE Computational Systems Bioinformatics Conference, August 8–11, 2005. Stanford, CA., USA: 2005. p. 93–102.

[9] Chouhan U, Pardasani KR. A maximum parsimony model to reconstruct phylogenetic network in honey bee evolution. Int J Biol Life Sci 2007;3:220–4.

[10] Lockhart PJ, Steel MA, Hendy MD, Penny D. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol Biol Evol 1994;11:605–12.

[11] Jukes TH, Cantor CR. Evolution of protein molecules. In: Munro HN, editor. Mammalian protein metabolism. New York: Academic Press; 1969. p. 21–132.

[12] Kimura M. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 1980;16:111–20.

[13] Qi X, Wu Q, Zhang Y, Fuller E, Zhang CQ. A novel model for DNA sequence similarity analysis based on graph theory. Evol Bioinform Online 2011;7:149–58.

[14] Tamura K. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. Mol Biol Evol 1992;9:678–87.

[15] Amabile–Cuevas CF, Chicurel ME. Horizontal gene transfer. Am Sci 1993;81:332–41.

[16] Deo N. Graph theory with applications to engineering and computer science. New Delhi, India: PHI Learning Private Ltd; 2010.

[17] Callahan MP, Gengeliczki Z, Svadlenak N, Valdes H, Hobza P, De Vries MS. Non-standard base pairing and stacked structures in methyl xanthine clusters. Phys Chem Chem Phys 2008;10:2819–26.

[18] Guckian KM, Schweitzer BA, Ren RXF, Sheils CJ, Paris PL, Tahmassebi DC, et al. Experimental measurement of aromatic stacking affinities in the context of duplex DNA. J Am Chem Soc 1996;118:8182–3.

[19] Gago F. Stacking interactions and intercalative DNA binding. Methods 1998;14:277–92.

[20] Grzegorzewski P. Distances between intuitionistic fuzzy sets and/or interval-valued fuzzy sets based on the Hausdorff metric. Fuzzy Set Syst 2004;148:319–28.

[21] Mathur R, Adlakha N. A least squares method to determine reticulation in eight grass Plastomes. Online Journal of Bioinformatics (OJB) 2011;12:230–42.

[22] Gascuel O. Mathematics of evolution and phylogeny. Oxford, UK: Oxford University Press; 2005 ISBN-13: 9780198566106.

[23] Gusfield D, Bansal V, Bafna V, Song YS. A decomposition theory for phylogenetic networks and incompatible characters. J Comput Biol 2007;14:1247–72.

[24] Mathur R, Adlakha N. Linear programming model to construct phylogenetic network for 16S rRNA sequences of photosynthetic organisms and influenza viruses. Interdiscip Sci 2014;6:100–7.

[25] Gowda KC, Ravi TV. Divisive clustering of symbolic objects using the concepts of both similarity and dissimilarity. Pattern Recognit 1995;28:1277–82.