

Mathematical approach to establish dissimilarities and phylogenetic trees among biological sequencesRinku Mathur^{1,*}, Sumeet Kumar Sharma² and Akhil Sharma¹¹Department of Mathematics, School of Chemical Engineering and Physical Sciences
Lovely Professional University, Phagwara, Punjab, India² Department of Mathematics, Govt Degree College, Nowshera, Rajouri, J&K, India.**Abstract**

Evolution is a fundamental part of life of every species on earth. Evolutionary history of species can be visualized by the phylogenetic trees which in turn can be constructed with the help of dissimilarities among the DNA (de-oxy ribonucleic acid) sequences of the species. In this work, a mathematical approach based on graph theoretic concept has been reported to find the distance matrices of DNA sequences of species. These fourth order matrices are then used to find the dissimilarities among biological sequences by using Euclidean distance metric. The proposed approach has been implemented on the 11-real-life DNA sequences collected from the NCBI database and the published literature. The approach reported works on two parameters of sequences namely positioning of nucleotide and stack interaction and their combined effects has been taken into consideration for finding dissimilarities among DNA sequences. At last, an UPGMA (unweighted pair-group method using arithmetic averages) has been used to establish phylogenetic trees among the biological sequences. Moreover, programs based on graph theoretic concept has been coded in C++ language to convert DNA sequences into distance matrices of order four in short time. The results obtained by the mathematical approach are presented in the form of phylogenetic trees which are then compared and validated with the results obtained by using existing software T-Rex for the establishment of phylogenetic trees among biological sequences.

Keywords: *De-oxy ribonucleic acid sequences, dissimilarities, directed weighted graph, Euclidean distance metric, phylogenetic tree.*

1. Introduction

A phylogenetic tree is a tool used to visualize the evolutionary relationship between nucleotides sequences, genes, chromosomes, or species. Basically, it is a tool used for prediction of evolutionary history or ancestry of the present-day organism from past decades. This network is employed when reticulation events such as hybridization that refers to combining of quality of two organisms of different species, breeds, through reproduction to

produce a new variety and moreover this process is involved in duplication of a region of DNA that contains the gene. Phylogenetic trees have been designed to represent evolutionary history of species[1-3].

Before making phylogenetic tree our first step is to find a unique characteristic of biological sequence on the basis of which we can distinguish between the different species and specify their origin. In other words, we can say that the factor that is totally unique in each organism and the DNA sequence of the gene of an organism which differs it from another one. The genetic material in a living cell is made up of de-oxy ribonucleic acid, a storage site for nucleotide information[1, 3-4].

By using DNA sequence of different organisms, we will determine the dissimilarity between the different sequences in reference to different species on the basis of distance measures of the character of sequences of different biological species[3, 5-6].

So, in this approach, a graph theoretic concept has been used to evaluate the distance measure and finding the dissimilarity between the different DNA sequences. Two parameters positioning of nucleotide, stack interaction and their joint effect have been used to judge out the dissimilarities among the biological sequences concerned[4,7-10].

2. Basic terminologies

2.1 DNA (De-oxy-ribonucleic acid)

The genetic material in a living cell is made up of DNA. It is composed of two polynucleotide chains, where the backbone is made by sugar-phosphate, and the bases present inside. It is present in all the cells including white blood cells, semen, hair roots, bone, teeth and other body tissue[3, 11]. It is composed of four nucleotides that are considered as storage site for nucleotide information which is in the form of a code that is made of four chemical components i.e. Adenine (A), Guanine (G), Cytosine (C), Thiamine (T) where (A, G) are named as purines and (C, T) are named as pyrimidines (Purines and Pyrimidines are nitrogenous bases that are composed of two different types of nucleotide bases in DNA. In contrast to DNA, RNA (ribonucleic acid) also composed four different bases in which three of these are the same as in DNA: adenine, guanine, and cytosine. It contains uracil (U) instead of thymine (T) [3,10-11].

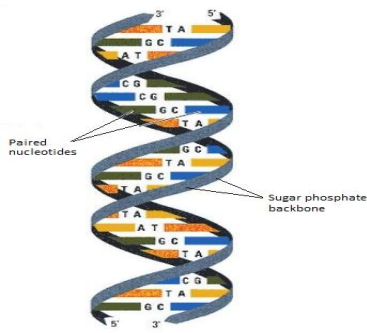


Fig. 1. Double helix structure of DNA [11].

2.2 Directed graph

A directed graph G is a graph which consists of set $V = \{v_1, v_2, v_3 \dots v_n\}$ of nodes/vertices that are connected together through set of edges $E = \{e_1, e_2, e_3 \dots e_n\}$ where all the edges are directed from one vertex to another vertex. In this type of graph each edge of graph G has a direction. In contrast, if each edge of the graph G doesn't have direction then the graph is known as undirected [5].

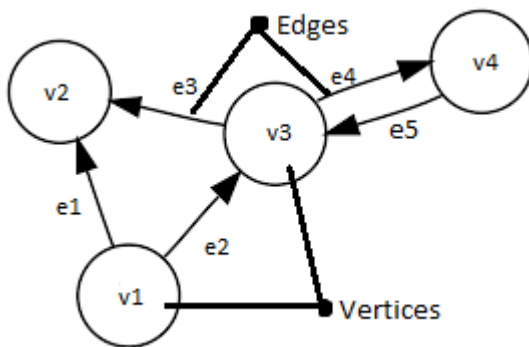


Fig.2. Directed graph

2.3 Adjacency matrix

Let us assume a graph G with set of vertices $V = \{v_1, v_2, v_3 \dots v_n\}$ and set of edges $E = \{e_1, e_2, e_3 \dots e_n\}$. Adjacency matrix of graph G is denoted by $A(G)$ is a $n \times n$ matrix in which rows and columns of $A(G)$ are indexed by set $V(G)$. So, matrix $A(G) = [x_{ij}]$ is adjacent where $x_{ij} = 1$ if there is edge directed between i^{th} and j^{th} or 0 otherwise. For a simple graph with no self-looping the adjacency matrix having 0 on the diagonal entries [7].

Let the directed graph G is given in the figure below:

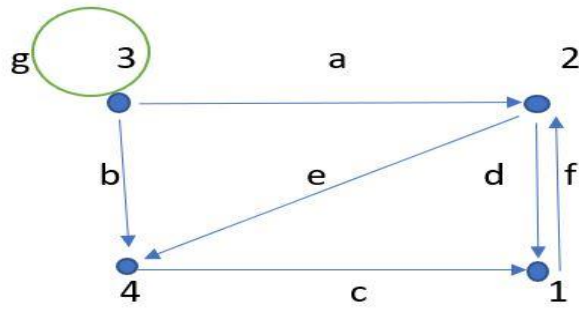


Fig. 3: Directed graph (G)

- A graph is given in the figure which represent 4 vertices $V(G)=\{1,2,3,4\}$ and 7 edges $E(G)=\{a, b, c, d, e, f, g\}$.
- Adjacency matrix corresponding to given graph G is

$$A(G) = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix} & \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \end{matrix}$$

One thing here in the above matrix is to be noticed that one of the diagonalelements is non-zero that is because of the self-looping at node 3 in the directed graph.

2.4 Distance matrix

In graph theory, a distance matrix is a square matrix containing the pairwise distance taken between the elements of the set. In graph theory applications the elements of matrix more referred to points nodes or vertices. The distance matrix of a graph is defined in a similar way as the adjacency matrix. The entry in the i^{th} row, the j^{th} column is the distance (length of the shortest path) between the i^{th} and j^{th} vertex.

2.5 Euclidean distance matrix

A Euclidean matrix is an $n \times n$ distance matrix M of order "n" if the following conditions are satisfied:

- Matrix M is symmetric if $d_{ij}=d_{ji}$ for every $i, j \in 1, 2, 3, \dots, n$.
- All the diagonal entries of matrix are zero i.e. $d_{ii}=0$ for every i, j .
- There exists $p_1, p_2, p_3, \dots, p_n$ on n-dimensional space such that $d_{ij}=\sqrt{(p_i - p_j)^2}$ where $1 \leq i, j \leq n$ [4].

2.6 Directed weighted graph

A graph G where vertices or edges or both vertices, as well as edges, have some weights or values. It is an ordered pair (V, μ) where $V = \{v_1, v_2, v_3, \dots, v_n\}$ is set of vertices of graph and μ is the weighting function which gives real non-negative value $\mu(v_i, v_j)$ to pair of nodes (v_i, v_j) . The weighted graph is called directed if the weighting function is non-symmetric i. e $\mu(v_i, v_j) \neq \mu(v_j, v_i)$. In contrast if symmetry holds for all v_i, v_j where $v_i \neq v_j$ then the graph is considered as undirected [8].

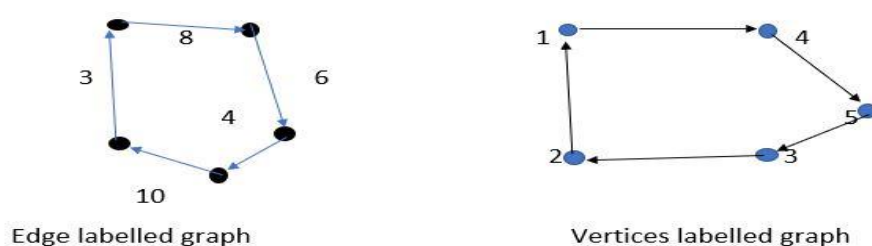


Fig. 4: Directed weighted graphs

Basically, each edge of a graph has an associated weight. In real life application, the weight may be a cost or length of road between source and destination, it may be the capacity of a pipeline that is often involved in maximal flow problems etc.

2.7 Phylogenetic tree

A connected graph with a single path between every pair of vertices u and v with no loops is called a phylogenetic tree and is denoted by $T(u, v)$. A phylogenetic tree T has exactly $|V| - 1$ edges and any vertex v is called a leaf if $\deg(v) = 1$. A tree T with n leaves has at most $n - 2$ internal vertices, $2n - 2$ vertices and thus $2n - 3$ edges. Any evolutionary tree representing the history of evolution of species or taxas can be converted into a binary tree by incorporating links of length zero wherever it will be necessary [4,6,12].

3. Methodology

3.1 Construction of directed weighted graph and distance matrix of DNA sequence with respect to nucleotides positions

As we know DNA sequences is made up of four bases A, G, C & T which will be considered as vertices.

Let us suppose a sequence $X = x_1, x_2, x_3, x_4, \dots, x_n$ such that $x_i \in \{A, G, C, T\}$

As the vertices are $V = \{A, G, C, T\}$, so 4×4 adjacency matrix is obtained no matter what the length of the sequence is. To describe the process of getting a directed weighted graph for the sequence $x_1, x_2, x_3, x_4, \dots, x_n$. For each pair of nucleotides X_m and X_n in X with $m < n$, direct an edge from X_m to X_n with weights as $(1/n - m)$ so that $(1/n - m)$ is decreasing function of $(n-m)$ which describes that nucleotides with smaller distance among them have strong interaction [4, 13].

Suppose we have a sequence $\{G, A, T, C, C\}$ we can observe that position of G is at 1, A is at 2, T is at 3, C is at 4 and 5 also.

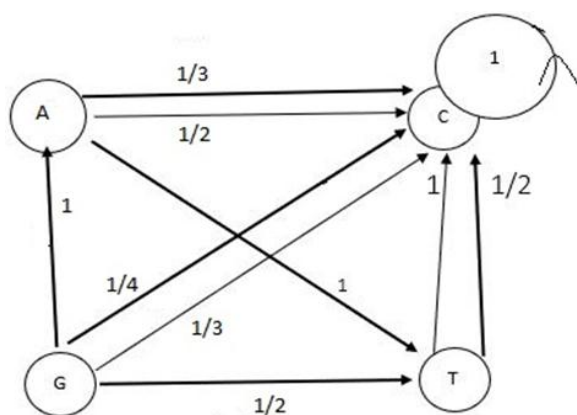


Fig.5: Directed weighted graph w.r.t position of nucleotides.

As there are two edges in the same direction among the same vertices. So, to simplify the above graph, all the weights of the edges directing in the same direction should be added, then we obtained the following distance matrix.

$$\begin{array}{c}
 A \quad G \quad C \quad T \\
 M_P = \begin{array}{c}
 A \begin{bmatrix} 0 & 0 & .83 & 1 \end{bmatrix} \\
 G \begin{bmatrix} 1 & 0 & .58 & .5 \end{bmatrix} \\
 C \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix} \\
 T \begin{bmatrix} 0 & 0 & 1.5 & 0 \end{bmatrix}
 \end{array}
 \end{array}$$

In this study, the two parameters have been taken in to account, one is based upon the nucleotides positions and other one is dependent upon stack interaction of nucleotides.

3.2 Construction of distance matrix with respect to stack interaction among nucleotides

In this case, the distance matrix from the DNA sequence is constructed by the probability of interaction among nucleotides.

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

The three probability for interaction among the nucleotides is considered. The probability of transformation of pyrimidines (C, T) to purines (A, G) or vice versa which is known as transversion is $\frac{1}{2}$, from pyrimidines to pyrimidines or purines to purines which is known as transitions is $\frac{1}{3}$, and conversion into the same nucleotide is $\frac{1}{6}$ [4].

Suppose we have a sequence {G, A, T, C, C} we can observe that position of G is at 1, A is at 2, T is at 3, C is at 4 and 5 also.

Distance matrix based upon stack interaction among nucleotides of the given sequence is given by

$$M_{SI} = \begin{matrix} & \begin{matrix} A & G & C & T \end{matrix} \\ \begin{matrix} A \\ G \\ C \\ T \end{matrix} & \begin{bmatrix} 0 & 0 & .291 & .5 \\ .333 & 0 & .291 & .25 \\ 0 & 0 & .166 & 0 \\ 0 & 0 & .5 & 0 \end{bmatrix} \end{matrix}$$

3.3 Euclidean distance metric to calculate dissimilarities among biological sequences

By using graph theoretic concept, distance matrices for a DNA sequence on the basis of two parameters namely nucleotides position and stack interactions are evaluated. The dissimilarity between two biological sequences said to be equal to the dissimilarities between the matrices obtained for the concerned sequences. The less dissimilarity between the two sequences represents the corresponding smaller distance among these sequences, and vice-versa. Suppose two DNA sequences P and Q and two matrices obtained corresponding to these sequences are M^P and M^Q for two parameters used. The distance measure between any two sequences has been calculated by using Euclidean distance matrices between these sequences which is generally denoted by $d(P, Q)$.

The distanced (P, Q) is given as

$$d(P, Q) = \sqrt{\sum_{i=1}^4 \sum_{j=1}^4 (M_{ij}^P - M_{ij}^Q)^2}$$

where $(M_{ij}^A - M_{ij}^B)$ determines the distance among the same positions of the matrices P and Q. The two features of the sequences are taken into consideration in the proposed approach to find out the distances among them. The total dissimilarity $d_T(P, Q)$, among the sequences P and Q is given by

$$d_T(P, Q) = d_P(P, Q) + d_{SI}(P, Q)$$

where $d_P(P, Q)$, is the dissimilarity due to positioning of nucleotide and $d_{SI}(P, Q)$, is the stack interactions dissimilarity between sequences. The comparison between the biological sequences has been determined by the distances among them. Distances obtained for both the feature and their combined effect are considered for the construction of the evolutionary tree among the given sequences [4].

3.4 Phylogenetic tree construction

It is a tool used to visualise evolutionary relationships among species. A phylogenetic tree serves a graphical relation between organisms, species, and the biological sequences, where roots represent the origin of evolution, leaves represent current organisms, and branch represents the relationship between organisms. There is number of methods by which we can construct tree but as we are having distance matrices so we should follow a distance-based method for constructing it. The main benefit of distance-based methods is the small-time complexity that makes them enable to the analysis of large data sets [6,9]. In this work, UPGMA method (Unweighted pair group method using arithmetic averages) has been used to establish the evolutionary tree among the biological sequences. UPGMA always forms an ultra-metric tree having a dendrogram like structure [12, 14]. The method forms a rooted tree because of the constant rate of evolution, though it is also possible to remove the root.

4. Application of proposed approach

To implement the proposed work on real-life data, we considered the sequences of exon-1 of the β -globin gene for 11 real life organisms. The following table represents their accession number in the NCBI (National Centre for Biotechnology Information) database [1,15].

Table 1.1: Information of Exon-1 of β -globin gene of concerned organisms.

Organisms	Accession/ID	Database	Length
HUMAN	U01317	NCBI	92
CHIMPANZEE	X02345	NCBI	105
GORILLA	X61109	NCBI	93
LEMUR	M15734	NCBI	92
RAT	XO6701	NCBI	92
MOUSE	V00722	NCBI	93
RABBIT	V00882	NCBI	92
GOAT	M15387	NCBI	86
BOVINE	X00376	NCBI	86
OPOSSUM	J03643	NCBI	92
GALLUS	V00409	NCBI	92

Table 1.2: DNA sequences of exon-1 of β -globin gene of concerned organisms[1-2, 15].

Organisms	DNA Sequences
HUMAN	ATGGTGACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAAGCCCTGGCAG
CHIMPANZEE	ATGGTGACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAAGCCCTGGCAGGTTGGTATCAAGG
GORILLA	ATGGTGACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAAGCCCTGGGCAGG
LEMUR	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTACCTCTCTGTGGGGCAAGGTGGATGTAGAGAAAGTTGGTGGCAGGCCTTG GGCAG
RAT	ATGGTGACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGGGAAAGGTGAACCTGATAATGTTGGCGCTGAGGCCCTGGCAG
MOUSE	ATGGTTGACCTGACTGATGCTGAGAAGTCTGCTGTCTTGCCTGTGGGCAAGGTGAACCCGATGAAGTTGGTGGTGAAGCCCTGGCAGG
RABBIT	ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCCTGCCCTGTGGGGCAAGGTGAATGTGGAAGAAGTTGGTGGTGAAGCCC TGGGCAG
GOAT	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGCTTCTGGGGCAAGGTGAAAGTGGATGAAGTTGGTGGTGAAGCCCTGGGCAG
BOVINE	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGCTTTTGGGGCAAGGTGAAAGTGGATGAAGTTGGTGGTGAAGCCCTGGGCAG
OPOSSUM	ATGGTGCACTTGAATTCTGAGGAGAAGAACTGCATCACTACCATCTGGTCTAAGGTGCAGGTTGACCAGACTGGTGGTGAAGCCCTGGCAG
GALLUS	ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCCGGCCTCTGGGGCAAGGTCAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG

To find distance matrix on the basis of positioning of a nucleotide of the sequence of human, we can observe from the table 1.1 and 1.2 that the length of the sequence is 92 so to calculate

it manually seems to be a difficult and time-consuming process. So, a program has been developed on the HP 245 G5 machine in C++ language that can calculate a distance matrix easily and the results are presented in the form of snapshots of output of program.

Dissimilarity matrices based upon the position of nucleotides

Human

```
Enter the DNA sequence without any spaces ATGGTGCACCTGACTCCTGAGGAGAAGTCT
GCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG

Result
  A      C      G      T
A  12.09  14.09  29.55  15.40
C  12.91  18.57  27.66  20.78
G  29.10  27.50  54.83  29.66
T  13.30  16.08  41.56  14.52

-----
Process exited after 536.2 seconds with return value 0
Press any key to continue . . .
```

Chimpanzee

Result	A	C	G	T
A	14.94	15.02	35.01	18.71
C	15.87	19.06	31.29	23.39
G	34.26	29.17	65.53	39.71
T	17.91	18.18	47.60	19.14

Gallus

Result	A	C	G	T
A	15.05	20.25	29.87	14.47
C	20.91	25.44	31.14	17.89
G	29.27	40.66	49.14	18.43
T	10.41	16.50	30.83	7.34

Gorilla

Result	A	C	G	T
A	12.09	14.09	30.47	15.40
C	12.91	18.57	28.58	20.78
G	29.10	27.50	57.46	29.66
T	13.30	16.08	42.20	14.52

Lemur

Result	A	C	G	T
A	15.67	11.37	32.46	20.02
C	11.84	10.20	19.62	20.04
G	31.82	23.54	52.98	28.45
T	16.20	16.77	45.33	21.29

Rat

Result	A	C	G	T
A	17.80	17.31	30.11	20.01
C	13.35	15.09	21.93	21.31
G	28.66	29.40	46.20	25.94
T	18.91	13.42	41.36	16.80

Mouse

Result	A	C	G	T
A	13.34	14.83	27.30	16.34
C	14.33	19.14	26.53	23.39
G	26.35	29.24	46.21	29.57
T	14.08	20.52	45.44	21.20

Rabbit

Result	A	C	G	T
A	14.76	9.39	33.80	17.01
C	13.04	13.65	23.71	15.52
G	31.72	26.53	60.40	32.00
T	12.86	15.94	43.13	14.16

Goat

Result	A	C	G	T
A	13.94	11.05	31.72	13.41
C	11.54	14.38	25.76	17.16
G	29.45	29.77	52.86	25.26
T	11.48	12.93	35.38	11.13

Bovine

Result	A	C	G	T
A	13.94	10.53	31.87	13.78
C	10.97	14.90	22.39	17.35
G	29.36	27.23	53.78	25.78
T	12.15	10.53	38.39	14.28

Opossum

Result	A	C	G	T
A	17.91	22.95	29.38	19.88
C	19.48	16.60	20.74	24.46
G	26.93	25.19	36.07	24.62
T	18.97	19.86	35.91	18.65

So, these are the respective distance matrices obtained for DNA sequences of different organisms, by the program developed in the proposed approach.

Dissimilarity matrices based upon stack interaction

Human

Enter the DNA sequence without any spaces ATGGTGACCTGACTCCTGAGGAGAAGTCTGCCGTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGT
GAGGCCCTGGGCAG

Result

	A	C	G	T
A	2.015	7.046	9.851	7.702
C	6.456	3.096	13.83	6.926
G	9.699	13.75	9.138	14.83
T	6.652	5.36	20.78	2.42

Chimpanzee

Result

	A	C	G	T
A	2.491	7.508	11.67	9.355
C	7.936	3.177	15.64	7.795
G	11.42	14.58	10.92	19.86
T	8.954	6.06	23.8	3.189

Gorilla

Result

	A	C	G	T
A	2.015	7.046	10.16	7.702
C	6.456	3.096	14.29	6.926
G	9.699	13.75	9.576	14.83
T	6.652	5.36	21.1	2.42

Lemur

Result

	A	C	G	T
A	2.612	5.686	10.82	10.01
C	5.922	1.7	9.81	6.681
G	10.61	11.77	8.83	14.22
T	8.098	5.591	22.66	3.548

Rat

Result

	A	C	G	T
A	2.966	8.657	10.04	10
C	6.677	2.515	10.96	7.104
G	9.555	14.7	7.7	12.97
T	9.455	4.474	20.68	2.8

Mouse

Result

	A	C	G	T
A	2.224	7.417	9.1	8.17
C	7.163	3.191	13.27	7.796
G	8.785	14.62	7.702	14.79
T	7.04	6.84	22.72	3.533

Rabbit

Result

	A	C	G	T
A	2.46	4.693	11.27	8.505
C	6.522	2.274	11.85	5.173
G	10.57	13.26	10.07	16
T	6.428	5.313	21.57	2.36

Goat

Result	A	C	G	T
A	2.324	5.526	10.57	6.704
C	5.772	2.397	12.88	5.721
G	9.817	14.88	8.81	12.63
T	5.739	4.309	17.69	1.854

Bovine

Result	A	C	G	T
A	2.324	5.265	10.62	6.889
C	5.487	2.483	11.2	5.782
G	9.785	13.61	8.963	12.89
T	6.073	3.51	19.2	2.38

Opossum

Result	A	C	G	T
A	2.984	11.47	9.793	9.942
C	9.74	2.767	10.37	8.155
G	8.975	12.59	6.012	12.31
T	9.484	6.621	17.96	3.109

Gallus

Result	A	C	G	T
A	2.508	10.13	9.957	7.237
C	10.45	4.24	15.57	5.963
G	9.758	20.33	8.19	9.216
T	5.206	5.499	15.42	1.224

These are the 11 matrices obtained corresponding to DNA sequences on the basis of two parameters i.e. the position of nucleotides and stack interaction among nucleotides. After these results, we shall figure out the Euclidean distances among the matrices corresponding to the respective parameters and we know that the dissimilarity between two sequences is lesser, if the Euclidean distance between these sequences is smaller. For calculating Euclidean distance, a program is developed in C++ language and the dissimilarity results among 11-species are as follow.

5. Results and discussion

After obtaining the distance matrices of the concerned species, we shall find out dissimilarities among them by using the Euclidean metric distance and the result based on both of distance parameters (positioning of nucleotides and stack interaction) are given as below.

Table 1.3:Distances of dissimilarity among 11-sequences on the basis of positioning ofnucleotides.

SPECIES	HUMAN	CHIMP	GORILLA	LEMUR	RAT	MOUSE	RABBIT	GOAT	BOVINE	OPOSSUM	GALLUS
HUMAN	0	20.17	3.00	16.83	15.68	13.54	12.82	12.2	11.80	26.16	26.19
CHIMPANZEE		0	18.31	24.51	28.56	25.63	19.16	28.99	27.79	39.27	39.06
GORILLA			0	17.51	17.60	15.45	11.90	13.25	12.78	29.18	27.19
LEMUR				0	15.15	18.18	14.49	20.11	16.14	26.62	37.70
RAT					0	14.78	21.03	17.45	16.00	16.82	27.96
MOUSE						0	22.28	21.26	20.37	20.71	29.72
RABBIT							0	15.11	12.95	33.81	33.50
GOAT								0	6.72	28.66	23.62
BOVINE									0	28.73	28.05
OPOSSUM										0	31.12
GALLUS											0

Now by following similar process we shall find out the dissimilarities among the concerned species on the basis of stack interaction using Euclidean metric distances and the results are as under

Table 1.4: Distances of dissimilarity among 11-sequences on the basis of stack interactions of nucleotides

SPECIES	HUMAN	CHIMP	GORILLA	LEMUR	RAT	MOUSE	RABBIT	GOAT	BOVINE	OPOSSUM	GALLUS
HUMAN	0	7.7677	0.7762	6.2525	5.7199	3.68	4.4702	5.0274	4.9129	9.2701	11.7976
CHIMPANZEE		0	7.3907	9.647	10.1492	7.9287	7.9716	11.8872	11.6245	13.2992	16.4355
GORILLA			0	6.4668	6.0909	3.8892	4.4917	5.3432	5.275	9.7072	11.9308
LEMUR				0	5.6923	6.4963	4.7953	8.3405	6.4189	9.6464	15.9525
RAT					0	5.7332	7.2852	7.274	6.5369	6.2492	12.0463
MOUSE						0	6.3211	7.6264	7.2943	8.8321	12.5794
RABBIT							0	6.27	5.091	11.3952	14.4047
GOAT								0	2.83395	10.4816	10.1716
BOVINE									0	10.3709	12.2759
OPOSSUM										0	12.1772
GALLUS											0

Table 1.5: Total dissimilarity on the basis of positioning and stack interaction of nucleotides.

SPECIES	HUMAN	CHIMP	GORILLA	LEMUR	RAT	MOUSE	RABBIT	GOAT	BOVINE	OPOSSUM	GALLUS
HUMAN	0	27.937	3.772	23.830	21.399	17.220	17.290	17.227	16.712	35.430	37.987
CHIMPANZEE		0	25.700	33.977	38.709	33.550	27.131	40.877	39.414	52.569	55.495
GORILLA			0	23.977	23.69	19.338	16.391	18.593	18.055	38.887	39.120
LEMUR				0	20.842	24.676	19.285	28.450	22.558	36.266	53.652
RAT					0	20.513	28.315	24.724	22.536	23.069	40.006
MOUSE						0	28.601	28.886	27.664	29.542	42.299
RABBIT							0	21.380	18.041	45.205	47.904
GOAT								0	9.058	39.141	33.791
BOVINE									0	39.100	40.325
OPOSSUM										0	43.297
GALLUS											0

UPGMA methodology is implemented for evolutionary tree construction of the real-life data of 11- species obtained from NCBI database and the results of obtained trees based on the distance matrices of table 1.3, 1.4 and 1.5 are given as below:

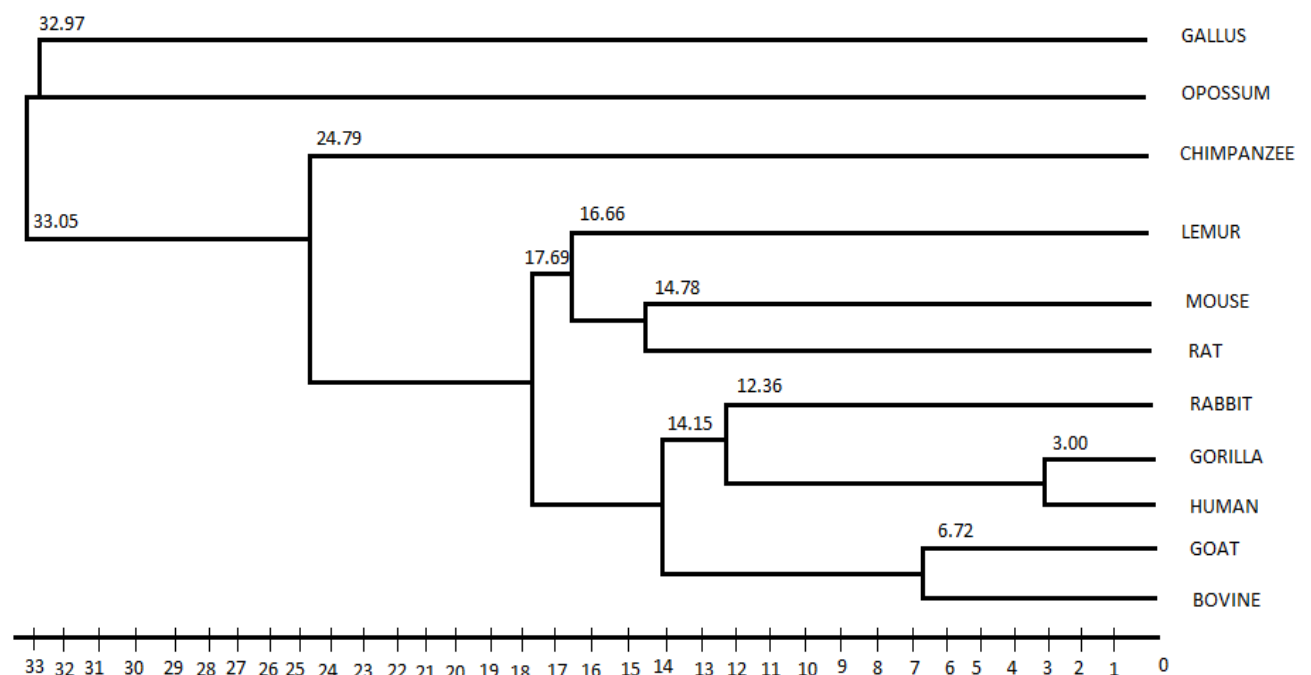


Fig.9: Evolutionary tree construction on the basis of dissimilarities calculated by positioning of nucleotides.

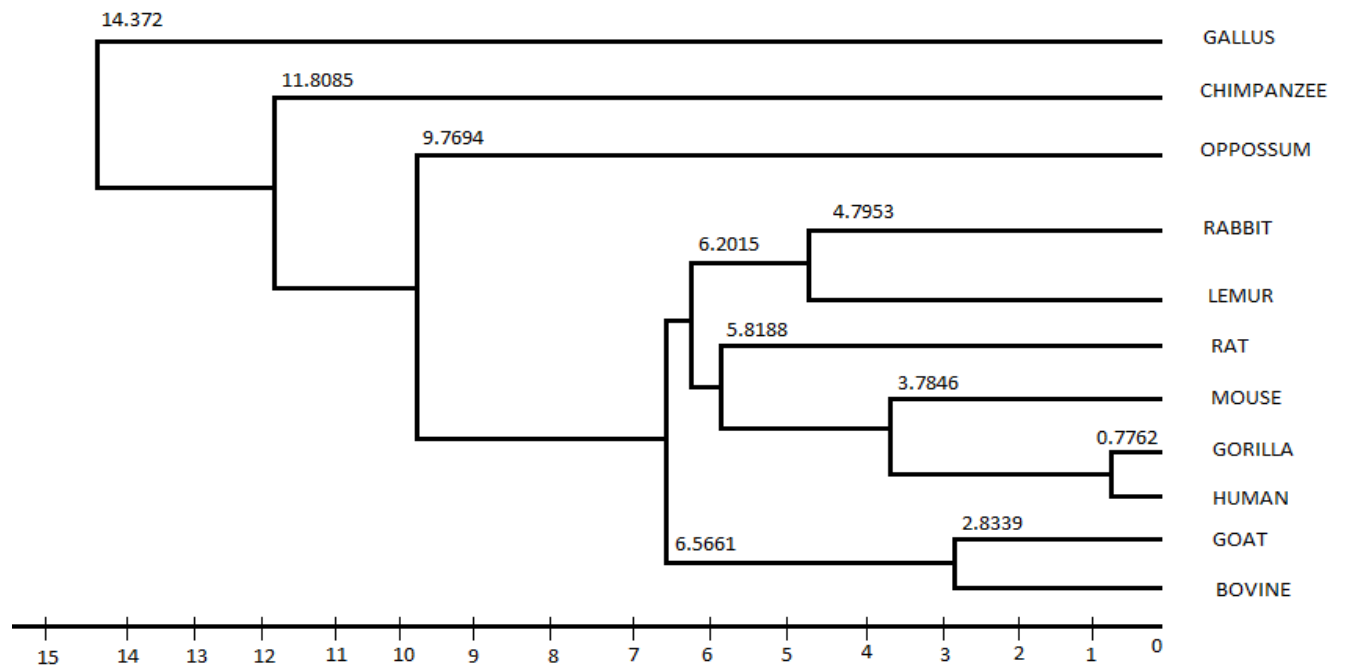


Fig.10: Evolutionary tree construction on the basis of dissimilarities calculated by stack interaction among nucleotides.

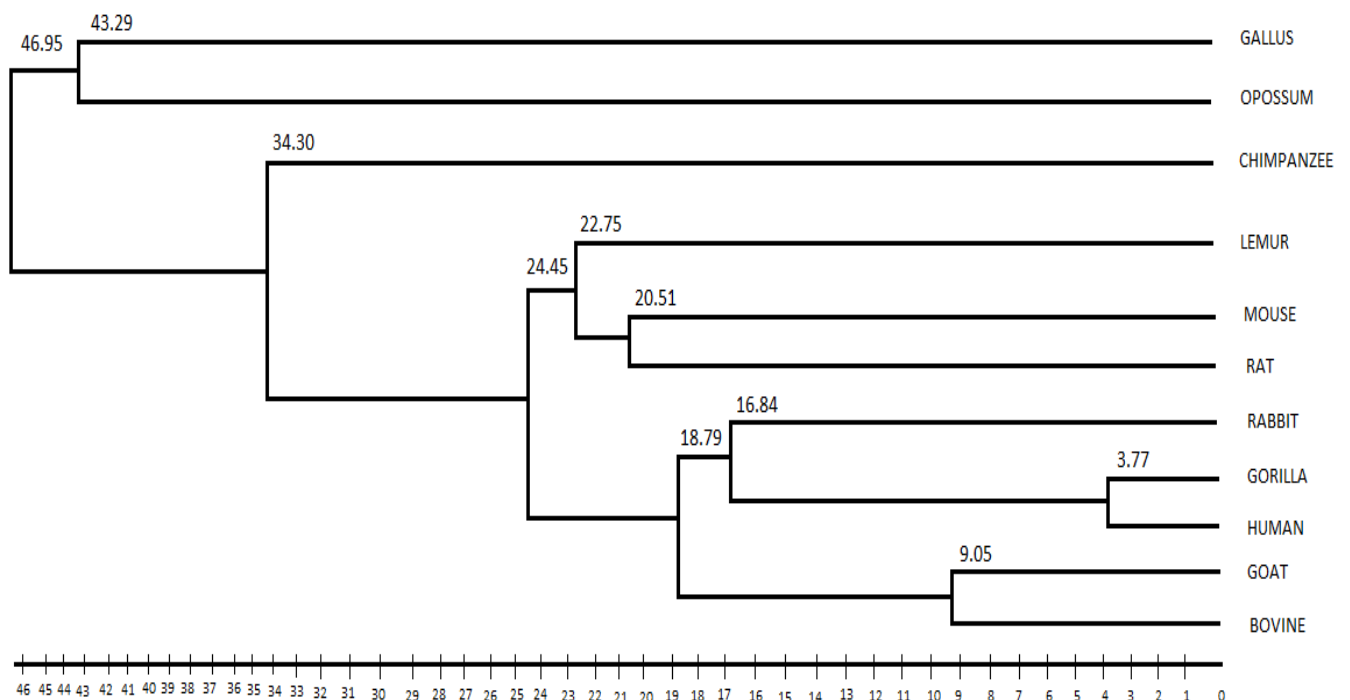


Fig.11: Evolutionary tree construction on the basis of dissimilarities calculates by totaldissimilarities.

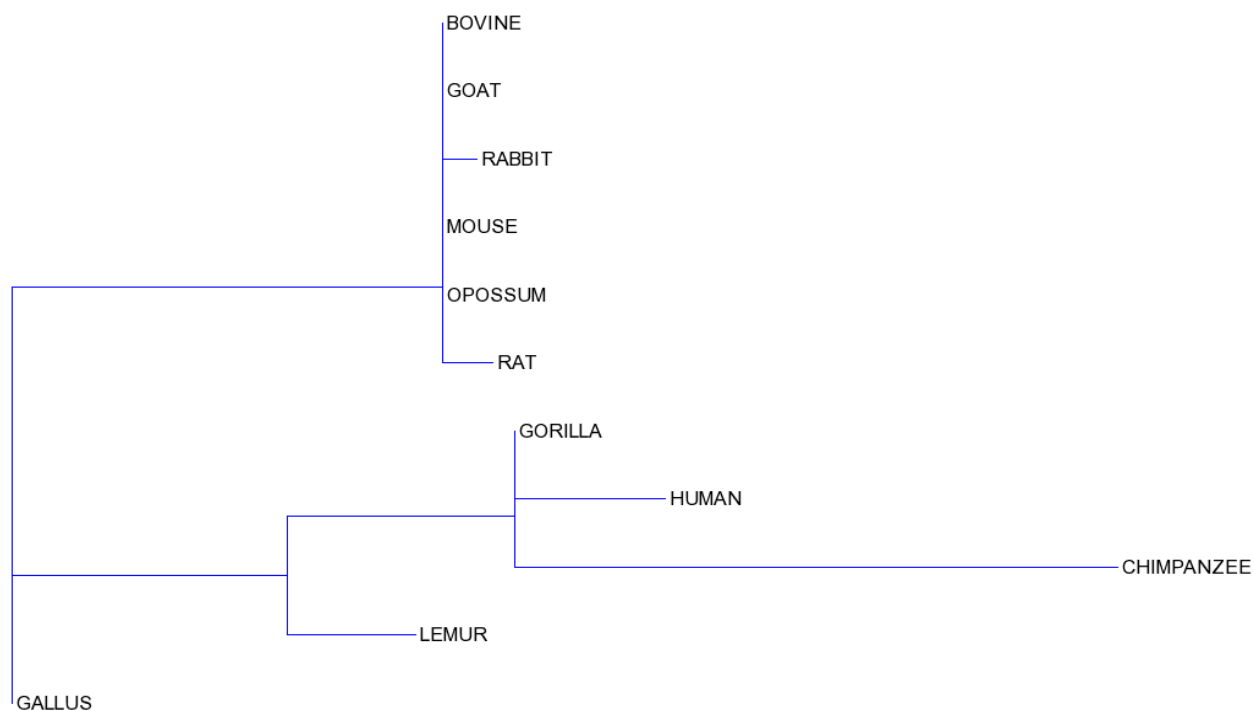


Fig.12: Evolutionary tree obtained by T-Rex software on the basis of dissimilarities calculated using positioning of nucleotides [12].

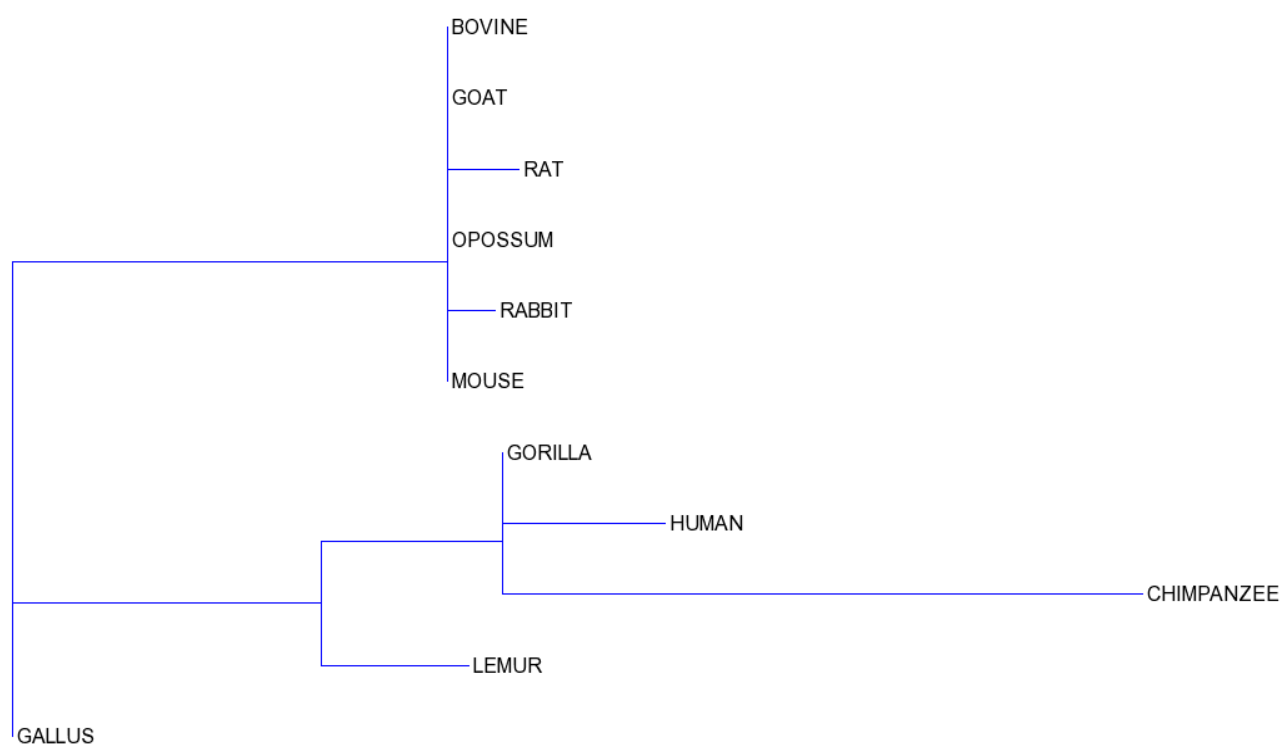


Fig.13: Evolutionary tree obtained by T-Rex software on the basis of dissimilarities calculated using stack interaction among nucleotides.

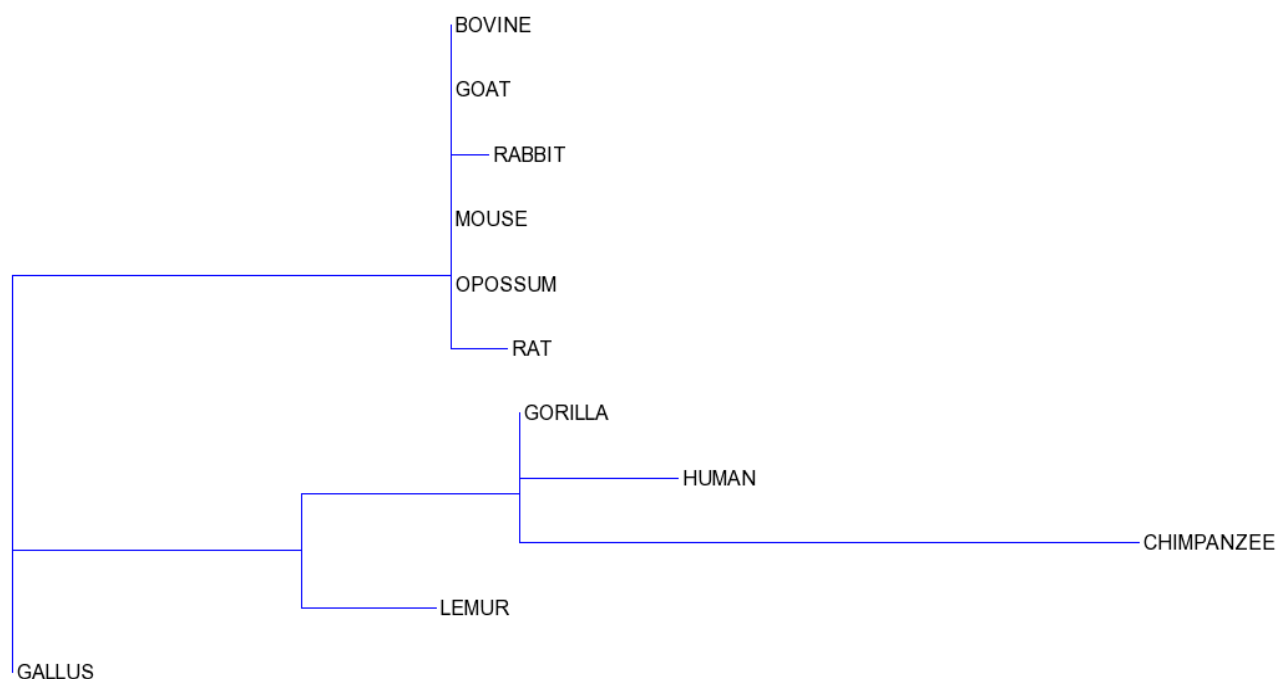


Fig.14. Evolutionary tree obtained by T-Rex software on the basis of total dissimilarities.

6. Conclusion

This study has proposed a new measure of dissimilarity based on graph theory concept to calculate distances among different species. The results obtained in the form of evolutionary trees (Fig 9, Fig 10 and Fig 11) based on the dissimilarities of three different features of the DNA sequences are validated by using existing software T-Rex for construction of evolutionary trees (Fig.12, Fig.13, Fig.14) on the same data set. By making comparison between (Fig.9, Fig.10, Fig.11) with (Fig.12, Fig.13, Fig.14) respectively we can observe that the clusters formed between concerned species are much similar. So, we can conclude the methodology proposed in this study, worked well and gave out appropriate results. The proposed method is very easy, fast and alignment free, and it can be used to analyse both short and long DNA sequences with desired accuracies.

References

- [1]. X. Qi, E. Fuller, Q. Wu, and C-Q Zhang, "Numerical characteristics of DNA sequences based on dinucleotides", *The scientific world journal*, 2012, doi:10.1100/2012/104269.
- [2]. R. Mathur and N. Adlakha, "A new measure of dissimilarity and fuzzy linear programming model to construct phylogenetic network among DNA sequences", *Network biology*, vol. 9(4), pp. 78-95, 2019.
- [3]. M. He and S. Petoukhov, "Mathematics of bioinformatics" A John Wiley & Sons, Inc. publication: 2011.
- [4]. R. Mathur, N. Adlakha, "A graph-theoretic model for prediction of reticulation events and phylogenetic networks for DNA sequences", *Egyptian journal of basic and applied sciences*, vol. 3, pp. 263-271, 2016.
- [5]. C. Vasudeva, "Graph theory with applications" New Age International Limited, Publishers: 2006.
- [6]. V. Makarenkov, D. Kevorkov and P. Legendre, "Phylogenetic network construction approaches", An article in *Applied Mycology and Biotechnology*, An international series, Bioinformatics, Elsevier Science, Vol 6, 2006.
- [7]. R.P. Bapat, "Graphs and matrices", Springer London Dordrecht Heidelberg New York. A co-publication with the Hindustan Book Agency, New Delhi (India). Doi 10.1007/978-1-84882-981-7.
- [8]. S. Umeyama, "An eigen decomposition approach to weighted graphs matching problem", *IEEE Transaction on pattern analysis and machine intelligence*, vol. 10(5), 1988
- [9]. F. Gago, "Stacking interactions and intercalative DNA binding methods", vol. 14, pp. 277-292, 1998.
- [10]. N. Deo, "Graph theory with application to engineering and computer science", New Delhi: PHI Learning Private Ltd, 2010.
- [11]. D.L. Hart, E.W. Jones, "Genetics principle and analysis", Jones and Bartlett Publishers International Barb House, Barb Mews London: 1998.
- [12]. V. Markenov, "Trex: reconstructing and visualising phylogenetic tree and reticulation networks", *Bioinformatics*, vol. 17, pp. 664-668, 2001.
- [13]. X. Qi, Q. Wu Y. Zhang, E. Fuller, C.Q. Zhang, "A novel model for DNA sequence similarity analysis based on graph theory". *Evol. Bioinform Online*, vol. 7, pp. 149-158, 2011.
- [14]. R. R. Sokal, and P. H. A. Sneath, "Principles of numerical taxonomy", W.H. Freeman, San Francisco, 1963.
- [15]. www.ncbi.in/ accessed on dated 12-10-18.