

A Mathematical model For Establishing Evolutionary Network Among DNA Sequences Based on Nucleotide Frequencies

Rinku Mathur^{*}

Department of Mathematics, School of Chemical Engineering and Physical Sciences
Lovely Professional University, Phagwara, Punjab, India.

Abstract

In this study, a novel measure dependent on frequencies of nucleotides in DNA (de-oxy ribonucleic acid) sequences has been proposed. The dissimilarities among the sequences has been evaluated by employing Euclidean distance metric. Dissimilarities obtained are used to establish the evolutionary tree between biological sequences. Moreover, a fuzzy linear programming model is proposed in this work, to establish the evolutionary network. As an application, the model has been implemented on the actual data set of mitochondrial DNA sequences of twenty species and compared with the results of method (T-Rex).

Keywords: Nucleotide frequencies; De-oxy ribonucleic acid sequences; Euclidean distance; Fuzzy linear programming model; Evolutionary network.

1. Introduction

With the advent of huge amount of DNA sequences, the scientific databases are growing very rapidly. Biological de-oxy ribonucleic sequences are generally composed of four base pairs called Guanine (G), Cytosine (C), Adenine (A), and Thymine (T). For biologists, it will be very hectic task to analyze the DNA sequences mathematically. To address the problem, good number of models are reported to obtain detailed information of sequences of different species [1-3].

Till date many models like sequence alignment, graphical techniques, and distance methods have been developed for finding similarities/dissimilarities among the DNA sequences [4-8].

A “C programming code” has been developed in this work for finding the frequencies of occurrence of four base pairs (nucleotides) in the DNA sequences. To calculate dissimilarities between each pair of DNA sequence, Euclidean distances are used on these frequencies. These Euclidean distances suggests the genetic/original evolutionary dissimilarities among organisms [9].

* Corresponding author and email: rinku.22748@lpu.co.in; rinkumathur56@gmail.com

Generally, phylogenies are suggested as basic tool for the prediction of evolutionary history of different organisms on earth. A lot of evolutionary tree construction models are cited in the literature of this field [8, 10-13].

The generalization of evolutionary trees has been visualized as evolutionary network and it predicts the possibilities of reticulate events among different organisms present in the tree [14-27]. In present study, an attempt has been made for the establishment of evolutionary network dependent on fuzzy membership distances and evolutionary tree dissimilarities among DNA sequences. A fuzzy linear programming (FLP) model has been implemented for the establishment of evolutionary network which is applied over the published dataset of mitochondrial DNA sequences of twenty organisms and validated by equating the predicted results with the results of already available model, T-Rex. [10, 18].

2. Methodology of Work: To perform the whole procedure of the methodology, three stairs have been followed in this study to establish evolutionary network.

2.1 Calculation of dissimilarities among DNA sequences:

To find out the evolutionary relationship among the DNA sequences, the dissimilarities among them has been calculated. According to literature, lot of methods are available to calculate genetic/original distances among sequences which are generally aligning the sequences and then calculating the dissimilarities among them [3, 8, 12]. In the present study, an alignment free model dependent on frequencies of presence of four base pairs in DNA sequences has been presented.

Suppose that a DNA sequence containing n nucleotides is given by $S_i = a_1, a_2, a_3, \dots, a_n$ and i represents the number of sequences considered in the study and ranges from 1 to N such that $s_1, s_2, \dots, s_N \in \{A, C, G, T\}$. The presence of each base pair in the composition of DNA sequence given by the formula [23, 27]:

$$f_{s_i} = \frac{\text{No. of } s_i}{n_j}; \quad \text{such that } s_i \in \{A, G, C, T\}$$

where n_j denotes length of the j_{th} sequence

The values of f_{s_i} obtained for every DNA sequence has been used to calculate the dissimilarities between each pair of available sequences under consideration.

Euclidean distance between two sequences m and l can be calculated by the formula[8]:

$$d_{ml} = \sqrt{\sum_{s_i} (f_{ms_i} - f_{ls_i})^2} \dots\dots\dots (1)$$

where f_{ms_i} denotes frequency of base pair s_i for the sequence m .

2.2 Framework for drawing evolutionary tree:

Now a days, many methods are available to draw the history of organisms called evolutionary trees like NJ, UPGMA, etc. that depends on dissimilarities of DNA sequences. In addition to it, character-based models can also be used for the establishment of evolutionary tree [21-23]. In this work, evolutionary tree among sequences has been established by using the dissimilarities generated in the above section and the concept used for that is relevant to biological evidence that “the larger the element in the dissimilarity matrix, the most-farther the species are” and vice versa. The average linkage clustering approach has been used in this study to draw the evolutionary tree [13].

Let d_{mn} and τ_{mn} denotes genetic/original distances and evolutionary tree dissimilarities between sequences m and n . In evolutionary tree, pair of organisms are clustered according to their original distances d_{mn} . Firstly, those organisms are combined together that have the smaller distance d_{mn} and so on and hence it will work till all the organisms under study should be present in the evolutionary tree. In some cases, if the tree distances $\tau_{mn} > \tau_{pq}$, corresponding to the original distance $d_{mn} < d_{pq}$, then it does not obeys the rule of evolution of organisms as per their ancestry. In such type of cases, the distance d_{mn} will be reshuffled by d_{pq} to get the tree distances τ_{mn} . At the end, evolutionary tree distances can be obtained for the organisms considered in the study.

2.3 Establishment of evolutionary network:

To establish evolutionary network, tree distances obtained in the above section has been used. Now, the pair of organisms whose $\tau_{mn} > d_{mn}$ will be pointed out to get the reticulation events in the network. The time for construction of network can also be saved by incorporating this condition. To get network, tree distances τ_{mn} for every pair of organisms have been transformed to fuzzy membership distances μ_{mn} with the function [23-24]:

$\mu: S \times S \rightarrow [0,1]$ defined by:

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

The fuzzy membership function given by equation (2) has been based on the biological fact that $\mu(m, n)$ will attain membership value 1 for the same DNA sequence, and as evolutionary dissimilarities between sequences increases, the fuzzy membership decreases.

Now, reticulation/hybridizing branches can be added to evolutionary tree by using the tree and fuzzy dissimilarities between the pair of organisms to get the network. The main aim of study is to minimize the total length of weighted evolutionary network representing reticulate/hybridizing organisms.

Here, a FLP model has been formulated for establishing the network [25-26]. The variable $e(i, j)$ denotes presence or absence of paths $(i, j) \in E(T)$ and $\tau(i, j)$ denotes length or weight of the corresponding paths in the evolutionary network or tree, whichever is smaller. Hence, to obtain the expected results for getting evolutionary network, a FLP model has been proposed and presented as:

$$\text{Minimize } Z = \sum_{i=1}^{|V|} \sum_{j=1}^{|V|} \mu(i, j) \tau(i, j) e(i, j) + \mu(s, t) \gamma(s, t)(new) e(s, t)(new) - \sum_{j=t} \mu(i, j) \tau(i, j)(old) e(i, j)(old) \quad \forall i, j \in V \quad \dots\dots\dots (3) \text{ subject to}$$

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

$$\tau(i, j) \in [0, 1] \quad \forall i, j \in V \quad \dots\dots\dots (5)$$

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

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

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

where $\gamma(s, t)(new)$ is the new weight of the edge (s, t) added in tree to get network

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

$\mu(i, j)$ is the fuzzy membership of edge (i, j) corresponding to $\tau(i, j) \forall i, j \in V$.

FLP model proposed here gives us the flexibility to evaluate or check only the vertices that are diverging in evolutionary tree, while the existing models evaluate every vertex of the tree. For establishing, the minimized evolutionary network of DNA sequences, it is important to pinpoint, how many new edges can be involved to evolutionary tree. Keeping it in mind, possible goodness-of-fit criteria making us to find out, when to stop adding new edges to evolutionary network as [21, 27]:

The formula of goodness of fit function was given and defined as [14, 21]:

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

where n is the number of leaves or number of organisms present in the evolutionary tree

Z_s presents criterion of stopping the addition of new edges to evolutionary network.

Z_1 presents updated value of Z while establishing evolutionary network.

3. Implementation to well-known result of mt-DNA sequences of twenty organisms

3.1 DNA Sequences as Data set

To check the utility of the purposed model, we have chosen the already published data set of DNA sequences, so that comparison among the results obtained can be done. The data set is comprised of the complete mitochondrial genomes of the twenty sequences, which were also studied by Hasan and Khalid (2003) [28]. All of these sequences were retrieved from the GenBank sequence database of NCBI site [29]. The names and accession numbers of the sequences of the data set used in the study are provided in Table 1:

Table 1: ID Information for mtDNA sequences of twenty organisms

Organisms	Accession/ID	Database	Length
Human	V00662	GenBank	16569
Common-chimpanzee	D38116	GenBank	16563
Pigmy-chimpanzee	D38113	GenBank	16554
Gorilla	D38114	GenBank	16364
orangutan	D38115	GenBank	16389
Gibbon	X99256	GenBank	16472
baboon	Y18001	GenBank	16521
horse	X79547	GenBank	16660
White rhinoceros	Y07726	GenBank	16832
Harbor seal	X63726	GenBank	16826
Gray seal	X72004	GenBank	16797
cat	U20753	GenBank	17009
fin whale	X61145	GenBank	16398
blue whale	X72204	GenBank	16402
cow	V00654	GenBank	16338
rat	X14848	GenBank	16300
mouse	V00711	GenBank	16295
opossum	Z29573	GenBank	17084
wallaroo	Y10524	GenBank	16896
platypus	X83427	GenBank	17019

A program in C language has been made to find the frequencies of occurrences of nucleotide sequences whose information is given in Table 1. The results obtained from the program and Euclidean distances among the sequences are then used to calculate the dissimilarity matrix among the different pairs of DNA sequences.

3.2 Results and Discussion

The results obtained by the proposed method and program coded in C language for the frequencies of presence of four base pairs in the composition of genome sequences are presented in Table 2 as:

Table2: Frequencies of occurrence of base pairs in composition of DNA sequences of organisms

Frequency of Nucleotides Organisms	Adenine	Guanine	Cytocine	Thymine
Human	0.3092	0.1313	0.3124	0.2471
Common-chimpanzee	0.3133	0.1270	0.3069	0.1270
Pigmy-chimpanzee	0.3113	0.1289	0.3080	0.2518
Gorilla	0.3092	0.1319	0.3069	0.2519
orangutan	0.3055	0.1323	0.3244	0.2378
Gibbon	0.3059	0.1369	0.3176	0.2396
baboon	0.3144	0.1313	0.3055	0.2488
horse	0.3216	0.1342	0.2854	0.2588
White rhinoceros	0.3341	0.1289	0.2796	0.2574
Harbor seal	0.3298	0.1428	0.2743	0.2530
Gray seal	0.3296	0.1428	0.2746	0.2530
cat	0.3258	0.1415	0.2619	0.2708
fin whale	0.3268	0.1331	0.2728	0.2673
blue whale	0.3276	0.1305	0.2760	0.2659
cow	0.3342	0.1346	0.2593	0.2719
rat	0.3402	0.1240	0.2629	0.2728
mouse	0.3454	0.1234	0.2440	0.2872
opossum	0.3532	0.1204	0.2115	0.3149
wallaroo	0.3318	0.1290	0.2628	0.2764
platypus	0.3143	0.1363	0.2350	0.3144

Now these frequencies of nucleotides have been used to calculate the evolutionary distance among the organisms by using the distance measure (d_{mn}) given by equation (1). After getting these evolutionary dissimilarities among organisms, the evolutionary tree among these organisms based on the proposed model has been constructed. Tree obtained suggests evolutionary history of these organisms and will be supposed as basis of establishment of evolutionary network. Optimized value of Z obtained for the phylogenetic tree with fuzzy membership function is 0.4847 and without fuzzy function is 0.5056.

Some of pairs of the organisms presents in evolutionary tree distracted from their original/genetic evolutionary dissimilarities and hence added to evolutionary tree to get evolutionary network based on proposed FLP model. The addition of new edges tells about the information about the reticulate or hybridizing organisms/events.

Based on tree dissimilarities, fuzzy membership distances between these organisms has been evaluated by using the formula of equation (2). Evolutionary tree dissimilarities and fuzzy distances have been used to draw evolutionary network. After applying the proposed model, the four reticulation branches (shown by red lines in Fig. 1) has been added to the evolutionary tree to convert it to evolutionary network. Addition of first reticulation branch reduces the value of Z from 0.4847 to 0.4389. In accordance with it, the optimized Z values has been reduced for the second, third and fourth reticulation branches are 0.3974, 0.3649 and 0.3349. The resulting network obtained is shown in Figure 1.

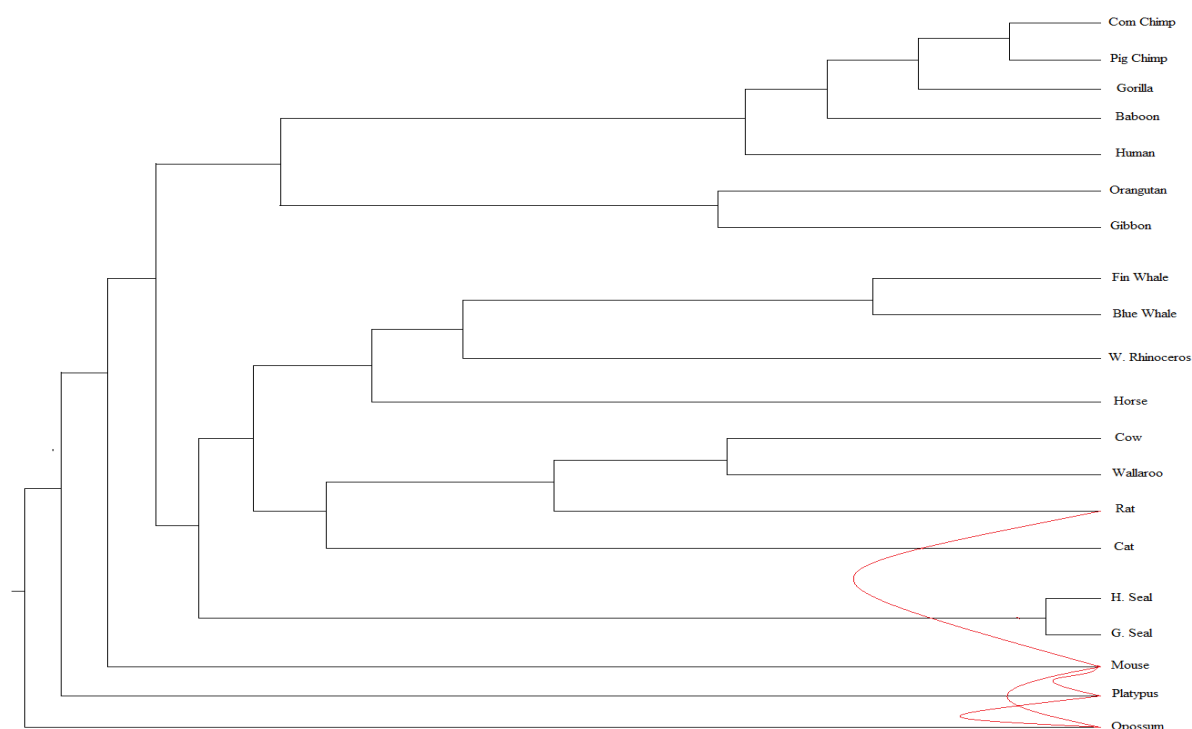


Figure1:Evolutionary network predicted by FLP modelpresenting relationships among complete mitochondrial genomes of twenty organisms.

The T-Rex online software has been used and implemented on the same data set of species to validate the results obtained and the evolutionary network obtained by the software has been presented in Figure 2.

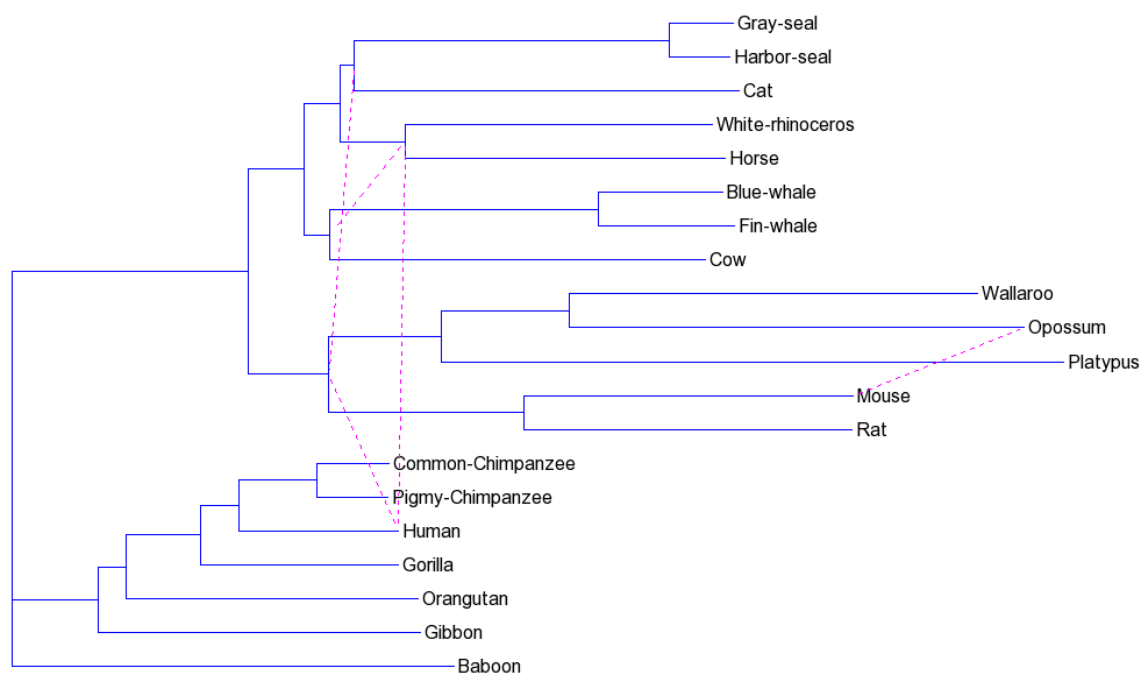


Figure2: Evolutionary network predicted by T-Rex presenting relationships among complete mitochondrial genomes of twenty organisms.

Both the figures (Fig. 1 and Fig. 2) predict almost same type of clusters for this longer dataset. The results predicted by the proposed model are in agreement with the results of existing online tool T-Rex [18]. The proposed method finds the reticulation branches among the extant species while the online T-Rex method finds only one branch between the extant species and all other branches are among the extinct species.

4. Conclusion

In this paper, a novel measure using frequencies of nucleotides in DNA sequences has been developed to evaluate dissimilarities between them. The measure does not require any kind of alignment of DNA sequences for obtaining the dissimilarities among them. Moreover, aFLP model purposed here to establish the evolutionary network which provides fast solution to the problems as it works on only extant organisms.

Henceforth, a complete model is purposed here for establishing the evolutionary network. Validation of model has been done over the data set of complete mitochondrial genomes of twenty organisms. Results predicted by the model, after the addition of four reticulation branches, value of Z has been optimized to 0.3349.

References

- [1] X. Qi, E. Fuller, Q. Wu, and C. Q. Zhang, "Numerical Characterization of DNA Sequence Based on Dinucleotides," *The Scientific World Journal*, 2012, doi:10.1100/2012/104269.
- [2] X. Q. Liua, Q. Dai, Z. Xiua, and T. Wang, "PNN-curve: a new 2D graphical representation of DNA sequences and its application," *Journal of Theoretical Biology*, vol. 243, pp. 555-61, 2006.
- [3] M. Kimura, "A Simple Method for Estimating Evolutionary Rates of Base Substitutions Through Comparative Studies of Nucleotide Sequences," *Journal of Molecular Evolution*, vol. 16, pp. 111-120, 1980.
- [4] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins, "Clustal W and Clustal X version 2.0," *Bioninformatics*, vol. 23, pp. 2947–2948, 2007.

- [5] Z. H. Qi and T. R. Fan, "PN-curve: A 3D graphical representation of DNA sequences and their numerical characterization," *Chemical Physics Letters*, vol. 442, pp. 434-440, 2007.
- [6] B. Liao and T. M. Wang, "Analysis of similarity/dissimilarity of DNA sequences based on 3-D graphical representation," *Chemical Physics Letters*, vol. 388, pp. 195-200, 2004.
- [7] T. Aita, Y. Husimi, and K. Nishigaki, "A mathematical consideration of the word-composition vector method in comparison of biological sequences," *Biosystems*, vol. 106, pp. 67-75, 2011.
- [8] P. J. Lockhart, M. A. Steel, M. D. Hendy, and D. Penny, "Recovering Evolutionary Trees under a More Realistic Model of Sequence Evolution," *Molecular Biology and Evolution*, vol. 11, pp. 605-612, 1994.
- [9] P. Grzegorzewski, "Distances between intuitionistic fuzzy sets and/or interval-valued fuzzy sets based on the Hausdorff metric," *Fuzzy Sets and Systems*, vol. 148, pp. 319-328, 2004.
- [10] C. Eslahchi, M. Habibi, R. Hassanzadeh, and E. Mottaghi, "MC-Net: a method for the construction of phylogenetic networks based on the Monte-Carlo method," *BMC Evolutionary Biology*, vol. 10:254, 2010.
- [11] H. J. Bandelt, P. Forster, and A. Rohl, "Median-Joining Networks for Inferring Intraspecific Phylogenies," *Molecular Biology and Evolution*, vol. 16, pp. 37-48, 1999.
- [12] W. Zhu, B. Liao, and R. Li, "A Method for Constructing Phylogenetic Tree Based on a Dissimilarity Matrix," *MATCH Communications in Mathematical and in Computer Chemistry*, vol. 63, pp. 483-492, 2010.
- [13] X. Qi, Q. Wu, Y. Zhang, E. Fuller, and C-Q Zhang, "A novel model for DNA sequence similarity analysis based on graph theory," *Evolutionary Bioinformatics*, vol. 7, pp. 149-58, 2011.
- [14] V. Makarenkov, P. Legendre, and Y. Desdevises, "Modelling phylogenetic relationships using reticulated networks," *Zoologica Scripta*, vol. 33, pp. 89-96, 2004.
- [15] A. Dress, V. Moulton, M. Steel, and T. Wu, "Species, clusters and the 'Tree of life': a graph-theoretic perspective," *Journal of Theoretical Biology*, vol. 265, pp. 535-42, 2010.
- [16] H. L. Chan, J. Jansson, T. W. Lam, and S.-M. Yiu, *Reconstructing an Ultrametric Galled Phylogenetic Network from a Distance Matrix*: Springer-Verlag Berlin Heidelberg, 2005.
- [17] J. Tang and B. M. E. Moret, *Linear Programming for Phylogenetic Reconstruction Based on Gene Rearrangements*: Springer-Verlag Berlin Heidelberg, 2005.
- [18] V. Makarenkov, "T-Rex: reconstructing and visualizing phylogenetic trees and reticulation networks," *Bioinformatics*, vol. 18, pp. 664-68, 2001.
- [19] H. J. Zimmermann, *Fuzzy Set Theory and Its Applications*, 4th ed.: Springer, 2001.
- [20] S. A. Mohaddes and M. G. Mohayidin, "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 & Environmental Sciences*, vol. 3, pp. 636-648, 2008.
- [21] V. Makarenkov and P. Legendre, "From a Phylogenetic Tree to a Reticulated Network," *Journal of Computational Biology*, vol. 11, pp. 195-212, 2004.
- [22] C. Semple and M. Steel, *Phylogenetics*: Oxford University Press, 2003.
- [23] R. Mathur and N. Adlakha, "A fuzzy weighted least squares approach to construct phylogenetic network among subfamilies of grass species," *Journal of Applied Mathematics and Bioinformatics*, vol. 3, pp. 137-158, 2013.
- [24] R. Tusserkani, C. Eslahchi, H. Poormohammadi, and A. Azadi, "TripNet: A Method for Constructing Phylogenetic Networks from Triplets," *Computing Research Repository*, vol. abs/1104.4, 2011.
- [25] S. Sridhar, F. Lam, G. E. Blelloch, R. Ravi, and R. Schwartz, Efficiently Finding the Most Parsimonious Phylogenetic Tree Via Linear Programming: *Springer-Verlag Berlin Heidelberg*, 2007.
- [26] L. Xiaozhong, D. yuyue, and Z. Fengchao, "Fuzzy linear programming problems with fuzzy variables and fuzzy coefficients," ed, pp. 86-90.
- [27] 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.
- [28] H. H. Otu and K. Sayood, "A new sequence distance measure for phylogenetic tree construction," *Bioinformatics*, vol. 19, pp. 2122 - 2130, 2003.
- [29] <http://www.ncbi.nlm.nih.gov/genbank>.