International Journal of Database Theory and Application
Vol.8, No.6 (2015), pp.111-124
http://dx.doi.org/10.14257/ijdta.2015.8.6.10

A New Alignment Free Method for Phylogenetic Tree Construction

Geetika Munjal*, Madasu Hanmandlu[1] and Deepti Gaur[2]

Institute of technology and management University Gurgaon, Haryana, India
India Institute of Technology New Delhi, Delhi 110016[1]
munjal.geetika@gmail.com, mhmandlu@gmail.com, deeptigaur@itmindia.edu

Abstract

In this paper various methods of sequence analysis which include the alignment based and alignment free methods of tree generation are reviewed and these find distance/similarity among the sequences of different species. Alignment free method based on tuple count and set theory is proposed and the results are compared with the guide tree obtained using alignment based method. The proposed method is tested on DNA sequence of length below 1000bp (dataset1) and Sequence of length above 16000bp (dataset2). It achieves the similar performance as that of the alignment based method but without the alignment phase.

Keywords: *sequence alignment, phylogenetic, nucleotide*

1. Introduction

There are millions of living organisms on earth today and evidence suggests that all of these organisms are genetically related. These genetic relationships can be represented by an evolutionary tree called the tree of life [20]. This tree represents the phylogeny of all organisms and it is the history of the organism's lineage as they mutate through time. Organisms have evolved over time from ancestral forms to more derived forms generating a tree of life [1, 12]. There are two types of trees: rooted and unrooted. The nodes of trees represent species and the relationship among the nodes is represented by branch. In technical terms the species are called Operational Taxanomic Units (OTUs) and the numbers of rooted and unrooted trees are 3 and 1 respectively in case of three-OTUs and 15 and 3 in case of four OTUs. As we increase the OTUs count, the number of rooted trees becomes very large. Multiple sequence alignment is considered as NP hard, which is computationally expensive. As the number of datasets in phylogeneomics increases exponentially the alignment based methods become more unaffordable. Another problem faced by the alignment based methods is that their results are dependent on the ultimate multiple sequence alignment on the initial pair-wise sequence alignments. In Multiple sequence alignment for phylogenies the very first sequences to be aligned are those most closely related to the sequence tree. If these sequences are aligned accurately, there will be a few errors in the initial alignments. However, the more distantly related are these sequences, the more errors will be committed, and these errors will be propagated to Multiple Sequence Alignment (MSA) thus generating less accurate tree. Another factor that plays a crucial role in the tree construction is a choice of suitable scoring matrices and gap penalties that apply to the set of sequences. Gaps in alignments can be thought of mutational changes in sequences, including insertions, deletions, or rearrangements of genetic material. For phylogenetic analysis the selected sequences should align with each other along their entire lengths, or else each should have a common set of patterns or

ISSN: 2005-4270 IJDTA

domains which provides a strong indication of evolutionary relatedness [14], hence the alignment quality affects the relationship created in phylogenetic tree due to the consideration of all points.

Apart from alignment there are a lot of issues to be addressed associated with tree generation. These are: i) Determination of evolutionary rate ii) The *molecular clock*, and iii) Biological validation of the trees produced. The evolutionary rate differs greatly between species and between single nucleotides; the rate is also not constant, but it may vary greatly at times. Moreover different kinds of mutations including insertions and deletions, substitutions, gene duplication pose complexity to the problem. Rough estimates of the evolutionary rate can be calculated; but their correctness is not guaranteed.

Some of the methods for tree construction are based on character data comprising the maximum parsimony and maximum likelihood. The maximum parsimony aims at finding a tree with the minimum number of substitutions. This method guarantees to find the best tree, because all possible trees relating to a group of sequences are examined [12]. But it is time consuming and not at all useful for large datasets or sequences having large variations. Maximum Likelihood method uses probability in constructing a tree that takes account of the variation in a set of sequences. It is similar to the maximum parsimony method in which analysis is performed on each column of the aligned sequences [7]. Both methods consider the mostly likely tree as the one that requires the fewest number of changes to explain the data in the alignment. For example as per the maximum parsimony principle among four sequences: s1=TAGCCAA s2=TAGCCTT, s3=TGCACCA, s4=TGCAGGA in which s1 is closer to s2 and s3 is more closely related to s4. Maximum parsimony method gives very less information about the branch lengths and suffers badly from long-branch attraction, which means that the long branches would be artificially connected because of accumulation of inhomogous similarities, even if they are not at all phylogeneticaly related[6]. Maximum Likelihood method is the slowest and most computationally intensive method, but gives the best result along with the most informative tree [17].

Another method based on distance reduces the information of long sequences into evolutionary distance and is considered to be computationally efficient [5]. The sequence pairs having the smallest number of sequence changes between them are termed as neighbors. The goal of the distance methods is to identify a tree that positions the neighbors correctly and reproduces the original data as closely as possible with its branch lengths [26]. Distance between sequence x and sequence y is expressed as the number of changes per site. This distance is based on the ratio of the number of mutations to the number of sites [3] and it assumes that all sites can vary and in case unvaried sites are present in two sequences it will underestimate the change occurred at the variable sites. Jukes cantor method assumes that every nucliotide can mutate into another with equal probability which is practically not possible The Jukes cantor distance between sequences *x* and *y* is given by Eq. 1 [9, 15]

$$D_{xy} = -\left(\frac{3}{4}\right)\ln(1-\frac{4}{3}d) \quad (1)$$

In Eq. (1) d is the observed proportion of nucleotides which differs in two sequences x and y. Another distance measure for nucleotide is by Kimura 2-parameter which estimates the evolutionary distance between two sequences in terms of the number of base substitutions per site occurred over T years of evolution [10]. This method assumes that transitions are more frequent than transversions. Using Kimura method the evolutionary distance is given by

$$k = -(1/2)\ln[\{1-2p-q)\sqrt{1-2q}\}] \quad (2)$$

where p is the probability of transitions and q is the probability of transversions [11]. Both methods find the distance among the sequences. Once the distance matrix is obtained, the tree is obtained by applying UPGMA, the neighbor joining and Fitch-Margoliash methods

[15,16] belonging to the distance based category. Fitch-Margrolish method generates an unrooted tree as it does not assume molecular clock.

As mentioned above alignment is a difficult problem and the construction of a phylogenetic tree is highly dependent on alignment, i.e. the errors of alignment will result in wrong tree generation. Even the choice of scoring matrices and gap penalties plays a major role in the resultant tree [28] and to solve such problems various alignment free methods are explored. We have proposed new alignment free method here.

The paper is organized as follows. Section 2 reviews various alignment free methods, and their validation methods. Section 3 gives the proposed method which is tested on two datasets, dataset I which is mitochondrial DNA D-loop for various species, and dataset II which is mitochondrial DNA for complete sequence of various species. A comparison of the proposed method is with the alignment based method in Section 4 and Section 5 gives the conclusions.

2. Alignment Free Methods for Sequence Analysis

Entropy is defined as a measure of information; it measures the average uncertainty in terms of bits of the outcome of a genomic sequence treated as a random variable [11]. It is used to analyze the randomness of genomic sequences to measure the information. Entropy convergence increases with the increased lengths of strings. Shannon entropy H(x) is defined as

$$H(x) = -\sum_{i=1}^{n} p_i \log(p_i) \tag{3}$$

where p_i in Eq. (3) is the probability of occurrence of i^{th} symbol. High entropy gives the information that all nucleotides are uniformly distributed whereas low entropy indicates the biasness in the sequence. The entropy can be used to find information of sequence based on k-mer where k is length of subsequence and p_i is the count of the occurrences of k length sequences in a DNA sequence[10]. To compare two sequences using k-mer firstly the sequence is partitioned into k-mer, and the frequency of k-mers is used to calculate the similarity between sequence s_1 and s_2 as follows:

$$similarity = d(\text{int}(s_1, s_2)) / \sqrt{d(s_1)d(s_2)} \tag{4}$$

where $d(int(s_1,s_2))$ in Eq. (4) is number of common k-mer present in sequences s_1 and s_2 and $d(s_1)$ and $d(s_2)$ are the total number of k-mers present in sequences $s_1$and s_2 respectively.

Relative entropy is also a measure of the degree of similarity/distance among the sequences. But this cannot be called as distance as it does not satisfy the property of symmetry. So we use relative entropy to obtain the distance matrix. The method best suited to large DNA sequence is K-L (Kullback-Leibler) divergence between two sequences x and y given by:

$$p = \sum_{i=1}^{n} p_i \log(p_i(x) / p_i(y)) \tag{5}$$

The unaligned methods use information theory and statistical theory concepts to extract features from the existing sequences and then the distance/similarity among the sequences is computed using different measures such as the Euclidean distance, Kullback-Leibler divergence and Shannon's entropy. Alignment free methods involve two steps : i) Obtaining features from sequences using Eq. (3) and Eq. (5), ii) Applying the distance formula in Eq. (6) and Eq. (7) to get the distance matrix followed by tree generation using neighbor joining or UPGMA. The quality of a tree is dependent on the information extracted from the sequences.

Coming to the analysis of sequences, k-tuple sequence analysis is based on Nucliotides symbols as a DNA sequence is built from 4 bases, viz., A, T, C, G. We have a

total of 4^k substrings of k length. If we take k=2 then 4^2 =16 possible combinations can be formed from 4 bases. Similarly k=3, 4^3=64 possible combinations can be formed and so on. As we increase the value of k there will be an exponential increase in the number of the possible k-tuples and most of them may not be present in a sequence [5-6]. So the extraction of features from k-tuple based method becomes a tedious task. The time taken by the alignment free method is largely dependent on the amount and the quality of features extracted. In a k-tuple we are concerned with tuple lengths varying from 2 to 8. Based on this we obtain features *p* of sequences called as the frequency vector. Then we can apply any of the distance metrics like Euclidian, Pearson, and Manhattan distances on p to obtain distance between the sequences. Let p_i be the extracted features from the sequence i and p_j be the extracted features from the sequence *j* . Then the Euclidean distance between them is given by

$$D^{Euclidean}(i,j)= \sqrt{\sum_{i,j=1}^{m}(p_i - p_j)^2} \tag{6}$$

where m is dependent on k and the value of m is 16 if k=2 and 64 if k=3 and so on. The Manhattan distance is

$$D^{manhattan}(i,j)=\sum_{i=1}^{m}| p_i - p_j | \tag{7}$$

The above distances, viz., Euclidean, Manhattan and Pearson are examined in [2] and it is concluded that normalization methods can only improve the clustering accuracy.

Raymond [25] has used the frequency method based on k-tuple where n is the frequency of occurrence of k-string computed as f(s)=n/(N-K+1), N is the length of string and K is the length of k-string. Given a Sequence S=GATTTCAT, f(AT)=2/7, f(ATT)=1/6, the maximum entropy principle is applied to find the similarity among the sequences.

Another alignment free method for sequence analysis is based on Lempel Ziv[27] method which traces the history of a sequence and uses this history to calculate the distance, for example, for the sequence S1=TTCGGTATT , H(S1)=T.TC.G.GT.A.TT and S2= ATTCGGTTA , H(S2)=A.T.C.G.GT.TA where H(S) is the exhaustive histories of the sequence. We need to apply any distance metric to obtain the exhaustive histories of the sequences.

To validate a tree-bootstrapping that estimates the significance of branches in a tree is used. It is a shuffled representation of DNA sequence involving a time consuming task if large number of samples is taken. In bootstrap a matrix x is formed by selecting some of the columns chosen from the original matrix. The original tree forming algorithm is used to build bootstrap tree from a new matrix. The process is repeated a large number of times. The proportion of trees agreeing to the original tree is checked to know the confidence on the tree. Jacknife analysis [1, 3] is used for phylogenetic tree validation which gives internal support of phylogenetic tree. Jacknife data set is smaller than that of bootstrap so it contains less information to build the phylogenetic tree. Tree metrics can also be used to validate a tree using the concept of agreement subtrees [4] as shown in Figure 1 for trees T1 and T2 for which the maximum agreement subtree is T3 as it satisfies the leaf ordering of both trees T1 and T2.

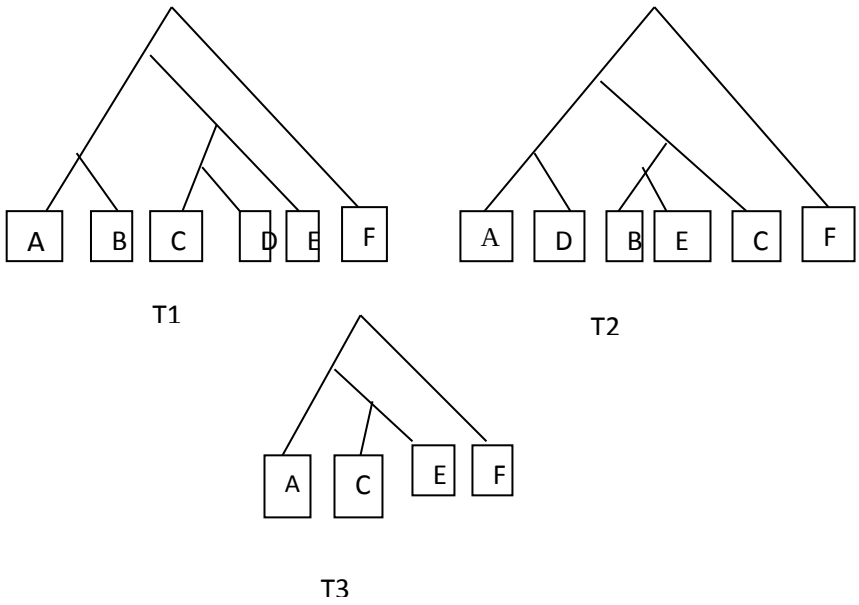

Figure 1. T3 Tree is Maximal Agreement Subtree of T1 and T2

3. The Proposed Method

The proposed 2-tuple based alignment free method finds the distance between the two sequences based on features obtained from k-tuple count. This distance between two sequences is calculated by taking the inverse of the similarity measure.

$$Sim(S1, S2) = \frac{Int(S1,S2)}{Union(S1,S2)}, \quad (8)$$

where *Int* is the intersection of k-tuple sequences and Union(S_1, S_2) is the union of k-tuple sequences in S_1 and S_2. The following steps are performed to find the similarity between the two sequences:

Algorithm:

Step1. Consider a nucleotide sequence S_1 of length L1, and S_2 of length L2. For S_1 and S_2 k-
tuples are extracted where k can vary from 2 to m.
Step 2. Find the intersection count(x) of k-tuple in S_1 and S_2.
Step 3. Find the union count of k-tuple (y) in S_1 and S_2
Step 4. Find the similarity = x/y using Eq. (7).

Let S1= ‘GATTGTGCGAGACAATGCTA’ S2= ‘CCTTACCGGTCGGAACTC’
For k=2 in S_1\{AA-1,AC-1,AG-1,AT-2,CA-1,CC-0,CG-1,CT-1,GA-3,GC-2,GG-0,GT-1,
TA-1,TC-0,TG-3,TT-1)
For k=2 in S_2(AA-1,AC-2,AG-0,AT-0,CA-0,CC-2,CG-2,CT-2,GA-1,GC-0,GG-2,GT-1,
TA-1,TC-2,TG-0,TT-1}
Int(S_1, S_2)={GA,AC,CG,CT,GA,GT,TA,TT}Int_{count}=8

Union(S_1,S_2)= {AA,AC,AG,AT,CA,CC,CG,CT,GA,GC,GG,GT,TA,TC,TG,TT}
$Union_{count}$=16
Similarity=0.5

The proposed distance metric satisfies the following properties:

1. Positivity : $d(S1,S2) \geq 0 \; and \; d(S1,S2) = 0 \; if \; S1 = S2$
2. Symmetry : $d(S1,S2) = d(S2,S1)$
3. Triangle inequality : $d(S1,S2) + d(S2,S3) \geq d(S1,S3)$

The proposed method is tested on two datasets. In dataset-I (Table1) the length of sequence is less than 1000 and in dataset-II (Table2) the length of sequence is more than 16000. For both datasets k is varied from 2 to 6. The trees are compared with the guide tree generated using the alignment based method where the distance matrix is generated using Jukes cantor method [9]. Entropy analysis is also done on sequence by varying the value of k and the results are shown in Figure 8(a) and 8(b).

Table 1. Dataset-I Sequence Length below 1000

Name of specie	Accession number	Sequence length
European_Human	X90314	379
Mountain_Gorilla	AF089820	373
Chimp_Troglodytes	AF176766	339
Puti_Orangutan	AF451972	353
Jari_Orangutan	AF451964	344
Western_Lowland_Gorilla	AY079510	376
Eastern_Lowland_Gorilla	AF050738	373
Chimp_Schweinfurth	AF176722	338
Chimp_Vellerosus	AF315498	410
Chimp_Verus	AF176731	338
finwhale	EU496282.1	686
bluewhale	EF057441.1	515

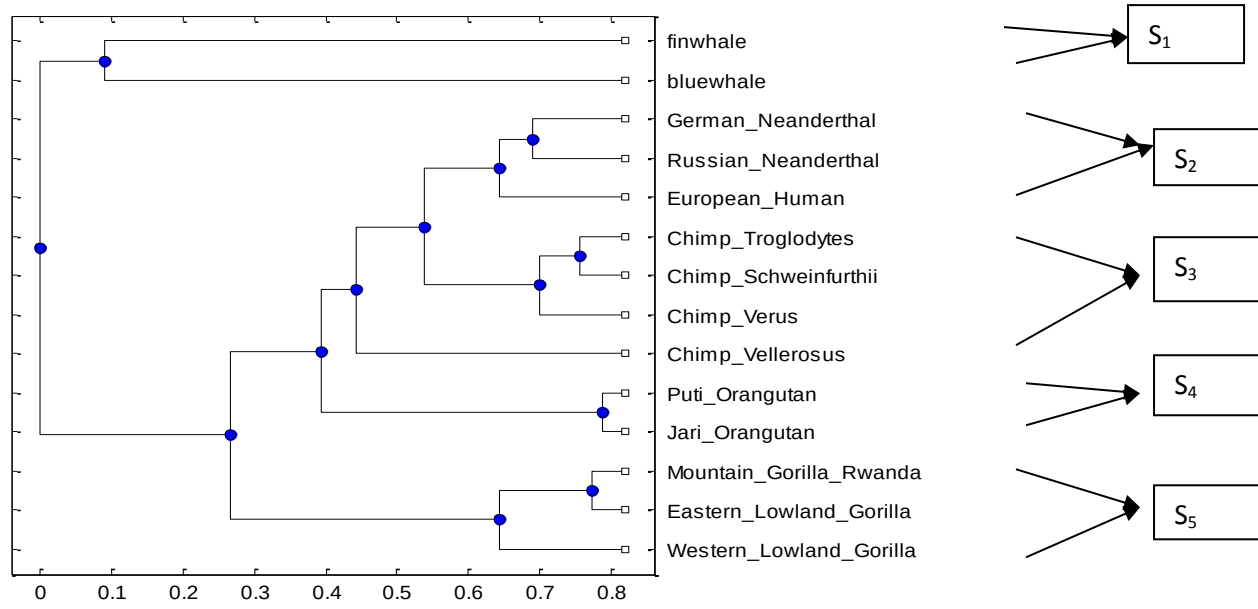

Figure 2. Guide Tree for Dataset1

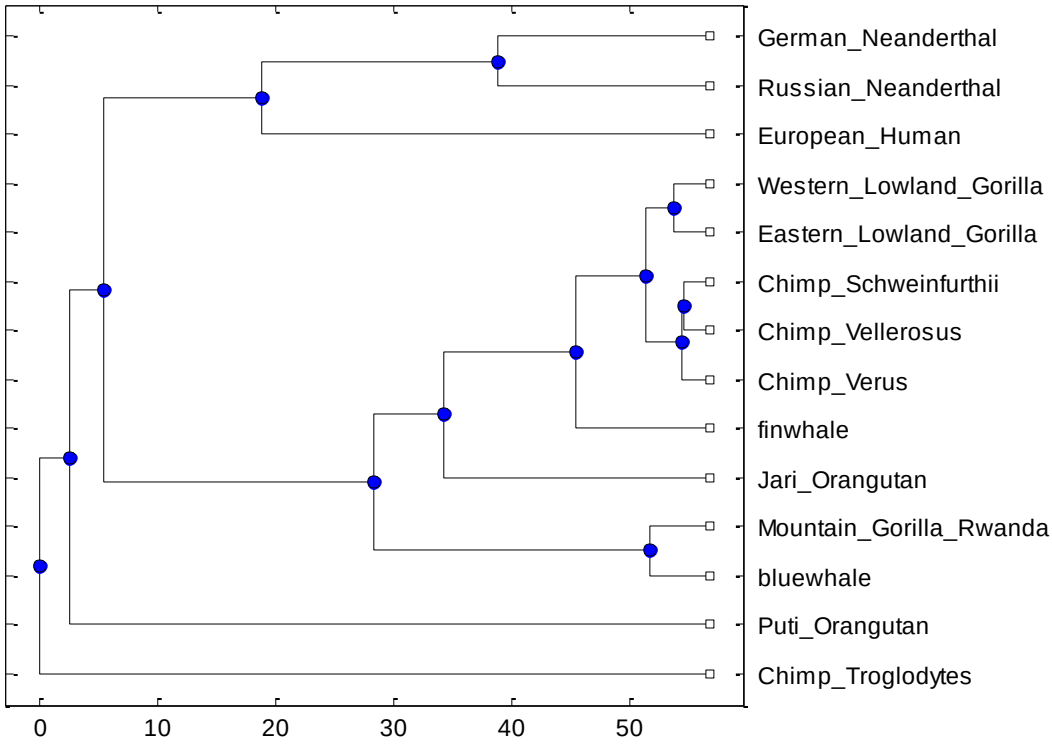

Figure 3. Tree Generated Using LZ Complexity

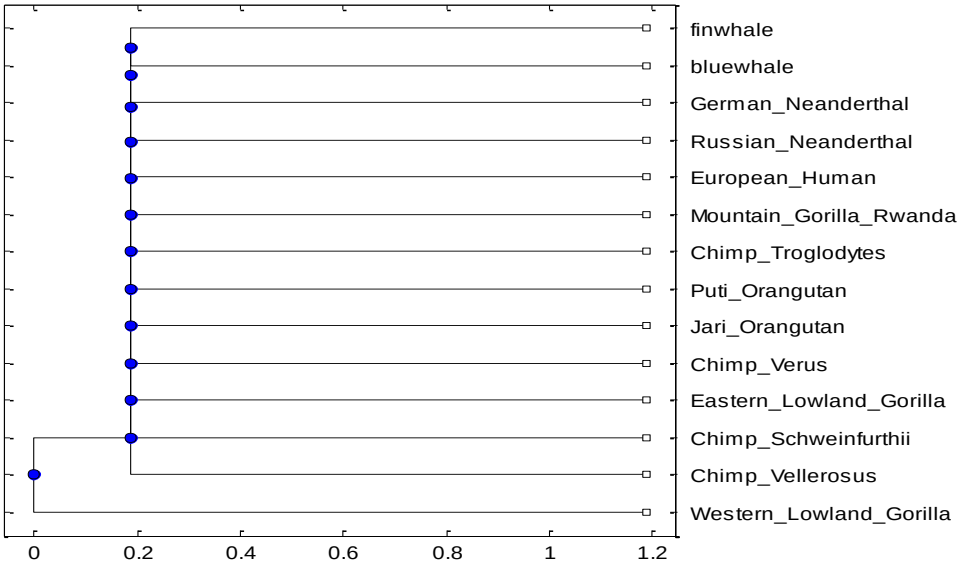

Figure 4. (a) Tree Generated for Dataset1 for k=2

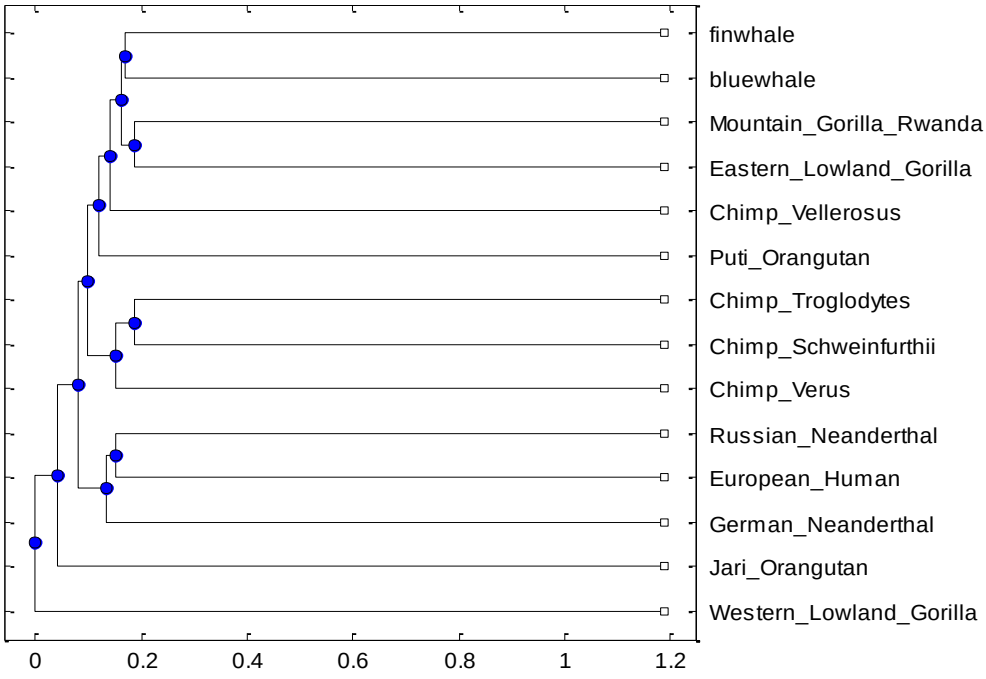

Figure 4(b). Tree Generated for Dataset 1 for k=3

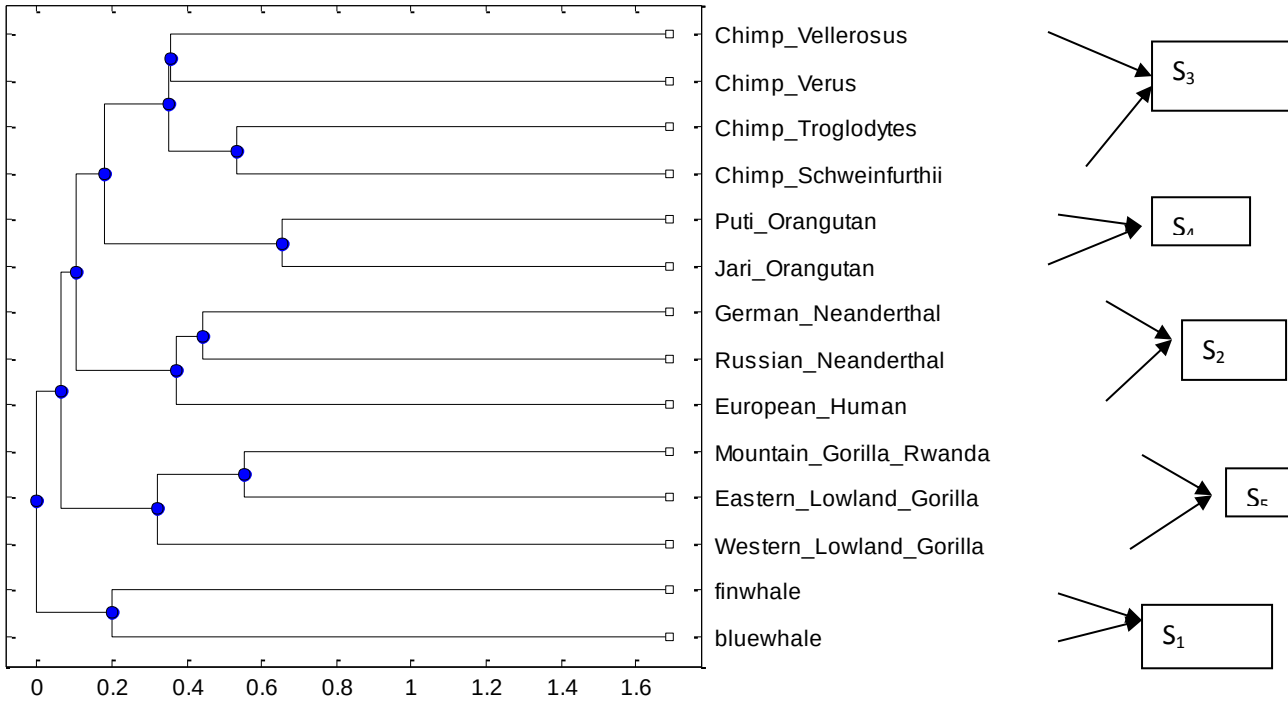

Figure 4(c). Tree Generated for Dataset 1 for k=4

Table 2

Max	min	difference	k
1.0875	1	.0875	2
1.324	1.0099	.225	3
1.9123	1.0338	.8785	4
4.8	1.0815	3.71	5
15.26	1.1049	14.16	6

Table 3. Dataset II

Name	Accession number	length
Human	V00662	16569
Chimpanzee	D38116	16563
pigchimpanzee	D38113	16554
horse	X79547	16660
rhinoceros	Y07726	16832
harbourseal	X63726	16826
Bluewhale	X72204	16402
Rat	X14848	16300
Mouse	V00711	16295
greyseal	X72004	16397
finwhale	X61145	16398

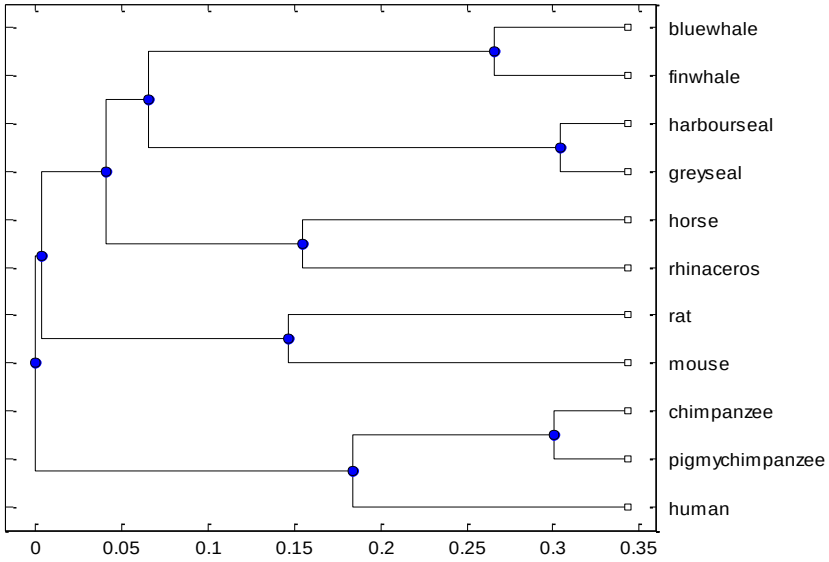

Figure 5. Guide Tree for Dataset II

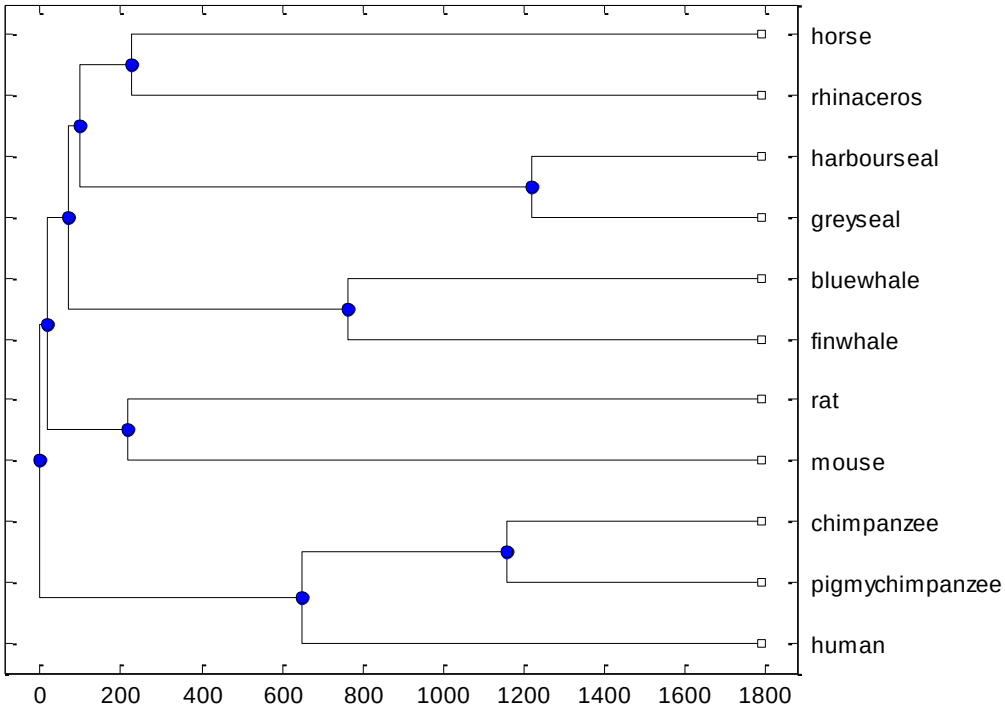

Figure 6. Tree Generated Using LZ Complexity

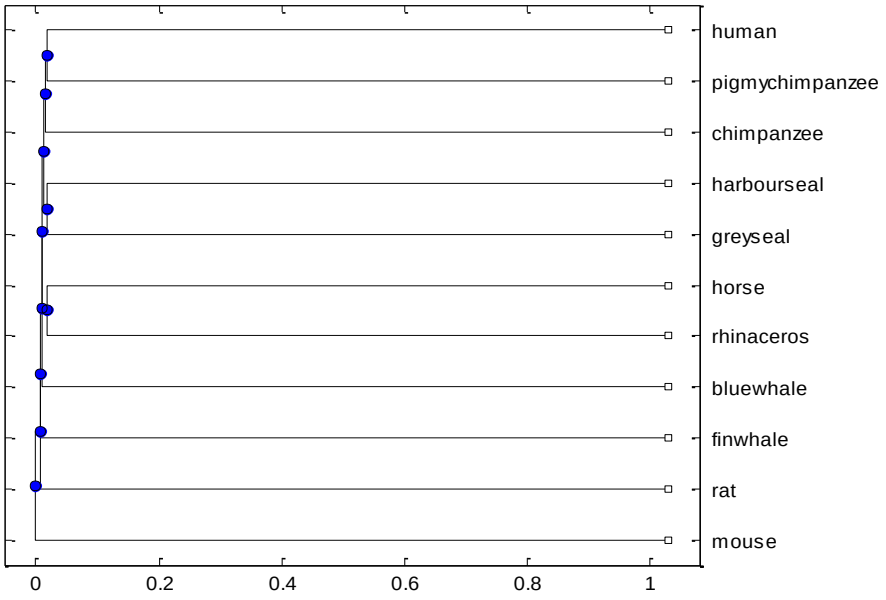

Figure 7(a). Tree Generated for Dataset II for k=5

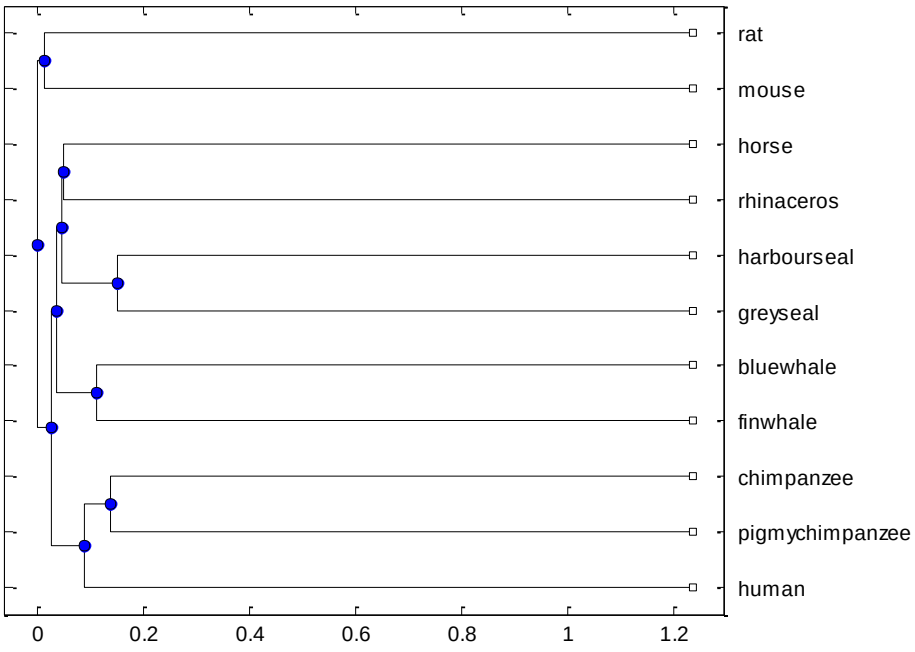

Figure 7(b). Tree Generated for Dataset II for k=6

Table 4

max	min	Difference (max – min)	k

1	1	0	2
1	1	0	3
1.0324	1.0109	.0215	5
1.2535	1.0852	.1683	6
1.9845	1.2508	.7337	7

4. Discussion of Results

The guide tree for dataset1 in which all the sequences are of length less than 1000(mtDNA D-loop) is shown in Figure 2. The 14 sequences of dataset1 are compared with those in Figure 4(c) in which the tree is generated with the proposed method for k=4. The subtrees s1, s2, s3, s4 are exactly the same in both the trees whereas the tree generated for k=2 in Figure 4(b) does not match with the guide tree. For dataset-2(complete genome) where the sequence length is greater than 16000bp the tree of Figure 5 is the guide tree. Comparing this with the tree of Figure 7 we find that the clustering of species is similar in both the trees and they agree with the properties of the maximal agreement subtrees [4]. The trees are also comparable to the tree generated using alignment free method based on LZ complexity Figure 3 and Figure 6. As can be seen in Figure 3 tree similar species are not clustered together where and even subtrees are also not similar to guide tree of Figure 2, but for dataset2 the tree generated in fig 6 using LZ complexity the subtree are similar to the subtrees of guide trees in Figure 5. Tables 2 and 4 display the maximum distance (max) and minimum (min) among the species from distance matrix for different values of k. As shown in Tables 2 and 4 the max distance for both the dataset increases with the increasing values of k just as the difference between the maximum and minimum values of distance. On comparing the trees of Figure 5 and Figure 6 it can be seen that the clustering of sequences is very clear in Figure 6 whereas in Fig. 5 the difference from the root to the species is not very clear. The same difference exists in Figure 3 and Figure 4. So it can be concluded for sequences less than 1000 the method gives good results for k=4 but for larger length sequences we have to increase the length of k to get better results. But by increasing the value of k for larger length sequences the executing time increases exponentially.

5. Conclusions

A new alignment free method of sequence analysis is presented based on k-tuple and set theory. The method is tested on mitochondrial DNA D-loop of 14 species with sequence of length <1000bp, and complete sequence of 11 species with lengths greater than 16000bp. For D-loop data the proposed method gives good results for k=4 and the tree structure remains the same till k=7. For complete genome sequences with length >16000bp improved results are attained for k=7. So it can be concluded that for sequence length less than 1000 the method gives good results for k=4 but for larger length sequences the length of k has to be increased for better results. The execution time increases exponentially for larger length sequence. The results are comparable to those of alignment based methods though alignment phase is not required thus reducing the overall complexity of the sequence comparison**.**

References

[1] Damian Bogdanowicz and Krzysztof Giaro "Matching Split Distance for Unrooted Binary Phylogenetic Trees" IEEE/ACM Trans. on Computational Biology and Bioinformatics, Volume 9 Issue 1, January 2012 pp 150-160.

[2] Chien-Hao Su, Wang, Hsu etal "The Impact of Normalization and Phylogenetic Information on Estimating the Distance for Metagenomes " IEEE/ACM Trans. On Computational Biology and Bioinformatics" Vol. 9, issue 2, March- 2012 pp. 619-628.
[3] Kuan Yang and Liqing ZhangPerformance "Comparison between k-tuple distance and four model-based distances in phylogenetic tree reconstruction" Nucleic Acids Research, Vol. 36 Feb 2008, pp. 1-9.
[4] Daniel Martin nad Bhalchandra Thatte, "The maximum agreement subtree problem", Discrete Applied Mathematics (161), Elsevier, pp-1805-1817 March 2013.
[5] Wei You, Shui Gao etal. "Similarity analysis of DNA sequences using Molecular connectivity indices method" IEEE Int. Conf. of Fuzzy systems and Knowledge Discovery, pp-75-79 2009.
[6] Wang Long-Hui etal. "A new Distance metric and its application in Phylogenetic Tree construction" IEEE Conf. on Computational Intelligence in Bioinformatics and Computational Biology , pp-108-115, 2004.
[7] C. Li and J. Wang, "Relative entropy of DNA and its application," Physica A: Statistical Mechanics and its Applications, vol- 347, pp. 465-471, Mar 2005.
[8] Dan Wei and Jiang "DNA sequence Distance measure Approach for Phylogentic Tree construction", IEEE bioinspired computing conference pp.204-212, 2010.
[9] Kirsi Louhisou, "Constructing Phylogentic Tree with UPGMA and Fitch Magrolish" Project report Aalto university of science, May 2004.
[10] Bohdan Kozarzewski "A Method for Nucleotide Sequence Analysis" computational Methods in Science and Technology Vol-18, pp-5-10 June 2012.
[11] Muhammad Sardaraz, Muhammad Tahir, Ataul Aziz Ikram, Hassan Bajwa "Applications and algorithms for inference of huge Phylogenetic Trees: a Review" American Journal of Bioinformatics Research, Vol-2, pp-21-26, Feb 2012.
[12] R.Potter, Constructing Phylogenetic Trees using Multiple Sequence Alignment, Master thesis, University of Washington 2008.
[13] Mohammed J. Zaki "Efficiently Mining FrequentTrees in a Forest: Algorithms and Applications" IEEE Trans. on Knowledge and Data Engineering, vol-17(8), pp-1021-1035, Aug-2005.
[14] Michael H. and Mark R. "Is Multiple Sequence Alignment required for accurate Inference of Phylogeny?" Oxford Journal of system Biology, vol-56, pp-206-219, 2007.
[15] Kirsi Louhisou, "Constructing Phylogentic Tree with UPGMA and Fitch Magrolish" Aalto university of science, May 2004.
[16] Jeffrey Rizzo and Eric C. Rouchka, "Review of Phylogenetic Tree Construction" University of Louisville, Bioinformatics review(report) 2007.
[17] Irrit Orr "Introduction to Phylogenetic" lecture notes Weizmann's institute of Science,.Aug-2004.
[18] Hasan H. and Khalid S. "A new sequence distance measure for Phylogenetic tree construction" OxfordJournal for Bioinformatics vol 19, pp 2122-2130, 2003.
[19] Felsenstein. J, PHYLIP - Phylogeny Inference Package (Version 3.6), Cladistics 5, pp-164–166,1989.
[20] David Mount "Bioinformatics sequence and genome analsysis", cold spring harbor laboratory press, pp.238-260.
[21] Dave Thomas "Example calculation of phylogenies The UPGMA method" notes, New Mexican for science and reason (nmsr.org), Oct 2002.
[22] Bernard M. E. Moret and Tandy Warnow, "Reconstructing Optimal Phylogenetic Trees :A challenge in Experimental Algorithmics" LNCS springer 2547, pp. 163–180, 2002.
[23] . Bichindaritz, I., Potter, S., and S.F.S, "Knowledge-Based Phylogenetic Classification Mining", Industrial Data Mining Conference, Perner,, Leipzig, Springer Verlang, Lectures Notes in Artificial Intelligence, pp. 163-172, 2004.
[24] Manoj Kumar Gupta, Rajdeep Niyogi, and Manoj Misra "A framework for Alignment-free methods to perform similarity analysis of biological sequence", IEEE sixth international conference on Contemporary Computing, Noida India, Aug -2013.
[25] Raymond H. Chan, Tony H. Chan, Hau Man Yeung, and Roger Wei Wang "Composition Vector Method Based on Maximum Entropy Principle for Sequence Comparison" IEEE/ACM Trans. on Computational Biology and Bioinformatics, Vol. 9, Issue-1, Feb-2012.
[26] Yasin Bakış, Hasan H Otu, Nivart Tasci, Cem Meydan, Neşe Bilgin, Sirri Yuzbasioglu "Testing robustness of relative complexity measure method constructing robust phylogenetic trees for Galanthus L. using relative complexity" BMC Bioinformatics,Vol-14, Jan-2013 pp.1-12.

Authors

Madasu Hanmandlu, he received the B.E. degree in Electrical Engineering from Osmania University, Hyderabad, India, in 1973, the M.Tech. Degree in power systems from R.E.C. Warangal, Jawaharlal Nehru Technological University (JNTU), India, in 1976, and the Ph.D. degree in control systems from Indian Institute of Technology, Delhi, India, in 1981. He is a Senior member of IEEE and is listed in Reference Asia; Asia's who's who of Men and Women of achievement; 5000 Personalities of the World (1998), He is presently an Associate Editor of both Pattern Recognition Journal and IEEE Transactions on Fuzzy Systems and a reviewer to other journals such as Pattern Recognition Letters, IEEE Transactions on Image Processing and Systems, Man and Cybernetics, He is associated with IIT Delhi from 1982.

Geetika Munjal, she received her B.Tech degree in computer Science and Engineering from Kurukshetra University, Haryana in 2002 and M.Tech degree in Computer Science and Engineering from Guru Nanak Dev Engineering College Ludhian, Punjab, India in 2007. And currently she is pursuing her PhD. Degree in Data mining from Institute of Technology and management Gurgaon, Haryana India.

Deepti Gaur, she has done her M.Tech, Birla Institute of Technology Mesra Ranchi, India and PhD from Banasthali University, Rajasthan, India, she is associated with Institute in Technology and Management University, Gurgaon from last 15 years.