Trends in Bioinformatics

ISSN 1994-7941

science alert

ANSInet
an open access publisher
http://ansinet.com

Trends in Bioinformatics 8 (3): 75-85, 2015
ISSN 1994-7941 / DOI: 10.3923/tb.2015.75.85
© 2015 Asian Network for Scientific Information

Modified k-Tuple Method for the Construction of Phylogenetic Trees

[1]Geetika, [2]M. Hanmandlu, [1]Ashish Sani and [1]Deepti Gaur
[1]Department of Computer Science and Engineering, Institute of Technology and Management University, Gurgaon, India
[2]Department of Electrical Engineering, Indian Institute of Technology, Delhi, India

Corresponding Author: Geetika, Department of Computer Science and Engineering, Institute of Technology and Management University, Gurgaon, India

ABSTRACT

This study proposes an extension of k-tuple method which utilizes the ratio of frequency of common sub words of length k to compare two sequences. The proposed method has two stages. stage 1 extracts feature from the sequence to obtain distance matrix and stage 2 obtains clusters from similarity matrix. The proposed method is tested on four datasets and the results are compared with those of k-tuple and tree generated using clustalw. Purity of tree and symmetric distance between the tree generated from proposed method and alignment based methods have also been computed. The results of proposed method are also compared with Composition Vector and k-tuple.

Key words: Phylogenetics, nucleotides, sequence alignment

INTRODUCTION

One of the major objectives of molecular evolution study to is to construct or create phylogenetic trees. A typical method for such a tree construction is based on multiple sequence alignment where comparison is made by alignment score. It may be noted that the tree constructed is highly dependent on the alignment score and the order of the given sequence to obtain the score. When the sequences are closely related and can be aligned reliably the results obtained are incomparable, but when the sequences are divergent, a reliable alignment cannot be obtained thus affecting the quality of tree. Another limitation of the alignment-based approaches is its computational complexity and time-consumption which limit us from dealing with large-scale sequence data (Haubold, 2014; Hohl and Ragan, 2007). Therefore, various alignment free methods exist in literature which performs sequence comparison.

In the past decade, various alignment free methods are proposed; some of them are based on tuple count known as k-tuple (Haubold, 2014), i.e., frequency of k-length string present in a sequence where frequency f = f(w)/(k+n-1) with f(w) being the number of times the substring of length k present in a sequence of n length. This method requires a large memory and it cannot work on closely related genes. Lempel-Ziv (LZ) method (Louhisuo, 2004) is based on the history of complexity of sequences. The LZ Complexity is a concept by which a finite sequence S can be built using the number of steps required by a production process. It follows following steps; (i) Take an array history initially empty to store the components, (ii) Read one symbol of the sequence at a time from left to right, (iii) Scan the array history and if the current symbol is not there in the history, add it to history as one of the components, (iv) If the symbol is already present in the history, then read the next symbol till the longest component at is not present in the history is

formed, (v) Repeat step iii-iv till the last component is add to the history and (vi) Calculate the number components in the history and display it as the LZ Complexity of the sequence. In these methods extracting the history of longer sequence is very time consuming. Other methods use information theory concepts like Shannon entropy and relative entropy (Otu and Sayood, 2003) for the sequence matching statistics. Entropy is the measure of uncertainty of information content in terms of bits of outcome. In the sequence analysis, entropy is a measure of information obtained from the order or disorder in a sequence (Wei and Jiang, 2010). Shannon entropy for a random variable x with values in a finite set x is given by:

$$H(x) = -\sum p_i \log_2 p_i$$

where, p_i is the probability of occurrence of character i. The Relative entropy (Long-Hui *et al*., 2004) is a measure of uncertainty between two probability distributions. For two probability distributions p_i and q_i, the Relative Entropy is given by:

$$S(x) = -\sum p_i \log_2 \frac{p_i}{q_i}$$

In the sequence analysis, the probability distributions can be obtained from two sequences for which Relative Entropy (Long-Hui *et al*., 2004) can also be calculated. Composition vector method (Chan *et al*., 2012) is also another alignment free method used for comparing two DNA sequences. Compared with the alignment based methods, it is simple as the number of parameters is less and there are no score matrices and gap penalties. The Composition vector method uses the informative string i.e. short nucleotides. In all the above alignment free methods the feature extracted from the sequence plays a major role in the tree clustering. After obtaining features from the sequence efforts are made to cluster these features using various methods of tree generation. For tree generation generally hierarchical clustering is adopted, which proceeds successively by merging smaller clusters into a bigger one, where the clusters created are called as dendrograms. Some of these algorithms are unweighted pair group method using arithmetic mean (UPGMA), the neighbour joining and Fitch-Margoliash (Louhisuo, 2004) methods which belong to the distance based category. Fitch-Margrolish method generates an unrooted tree as it does not assume molecular clock. 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 (Yang and Zhang, 2008). 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 (Orr, 2004). Both methods consider the mostly likely tree as the one that requires the fewest number of changes to explain the data in the alignment (Louhisuo, 2004). For example as per the maximum parsimony principle among four sequences: s1 = TAGCCAA, s2 = TAGCCTT, s3 = TGCACCA, s4 = TGCAGGA 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 the long-branch attraction, which means that the long branches would be artificially connected because of accumulation of in humongous similarities, even if they are not at all phylogenetically related (Orr, 2004). Maximum likelihood method is the slowest and most computationally intensive method, but gives the best result along with the most informative tree (Louhisuo, 2004). Another method based on distance reduces the information of long sequences into evolutionary distance and is considered to be computationally efficient. The sequence pairs having the smallest number of sequence changes between them are termed as neighbours. The goal of the distance methods is to identify a tree that positions the neighbours correctly and reproduces the original data as closely as possible with its branch lengths. Distance between two sequences is expressed as the number of changes per site i.e., the ratio of the number of mutations to the number of sites (You *et al*., 2009) and it assumes that all sites can vary and in case the unvaried sites are present it will underestimate the changes occurred at the variable sites.

MATERIALS AND METHODS

K-tuple method counts the frequency of each k-tuple and builds a feature vector, but this method doesn't describe the information completely. Moreover the accuracy of ktuple method is always a concern and to overcome this problem efforts have been made to find the common substrings in the sequences and then proposed a distance formula to find the pairwise distance between the sequences based on the count of common substrings. The goal of the proposed method is to: (i) Make an efficient use of the information contained in the genomes in form of tuples which lead to better distance measurement and (ii) To cluster the sequences that belongs to the same family in the same cluster as far as possible.

The proposed method eliminates the count of redundant tuples present in two sequences by taking intersection. The large number of common subwords indicates higher similarity or matching. The Phylogeney results represent the evolutionary dendogram and the quality of these dendograms depicts diversity, evolution mechanism and disease association.

To avoid the loss of information contained in a sequence the similarity score is calculated at various values of k and the most significant results are generated for higher values of k. The whole method is explained as:

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}
Then, generating
Int (S_1, S_2) = {GA,AC,CG,CT,GA,GT,TA,TT} ; so int_{count} = 8
Union (S_1,S_2)= {AA,AC,AG,AT,CA,CC,CG,CT,GA,GC,GG,GT,TA,TC,TG,TT}
The values of union count and similarity are:

$Union_{count}$ = 16 and $\mathrm{sim}(s_1, s_2) = \frac{\mathrm{int}(s_1, s_2)}{\mathrm{union}(s1, s_2)} = 0.5$

The proposed distance method satisfies the following properties (Yang and Rannala, 1997):

- Positivity: $d(s_1, s_2) = \geq 0$ and $d(s_1, s_2) = 0$ if S1 = S2
- Symmetry: $d(s_1, s_2) = d(s_1, s_2)$
- Triangle inequality: $d(s_1, s_2)+d(s_1, s_3) \geq d(s_1, s_3)$

After getting the distance matrix of similarity score, clustering is done using some clustering algorithm; here cluster refers to a group of sequences that show high similarity and least variation. Although, data origin/family is clear but there is no clustering algorithm that can precisely and accurately cluster all the data into the correct clusters. The results obtained by using various clustering algorithms like k-means, DBSCAN are varying every time because of the random initial cluster centres. In order to overcome this problem, UPGMA is used which is faster and based on polynomial time algorithm. The generated clusters/trees are evaluated to know the purity of a cluster. Also the results of the proposed method are compared with those of the alignment based tree generated using clustalw. Symmetric distance between ktuple tree and proposed method tree with respect to clustalw is calculated using treedist function of PHYLIP (Felsenstein, 2004). The symmetric distance metric is a measure of dissimilarity between the two trees by comparing the partitions induced by both the trees (Felsenstein, 2004). Consider two trees T1 and T2 of Fig. 1, all the partitions of the leaves are done and then the ones that are unique are counted. Those partitions are considered unique that do not appear in other trees. It is to be noted that if two trees have different number of leaves, then all the partitions will come out to be unique. The symmetric distance is twice the number of unique partitions. This can be explained as in Fig. 1 the cut C in T1 1 gives {AB}: {CFED} and partitions in tree 2 cut C gives us {ABD}:{CEF}, similarly for cut D in tree 1 partitions are {ABC}:{FED} in tree 2 {ABD}{FEC} for all other cuts the partitions will be similar in both trees so symmetric distance between tree1 and tree 2 is 2 as shown in Fig. 1.

The proposed method is also validated using purity of a tree, the Purity of a cluster is given by the ratio of the maximum number of species to the number of species in family correctly classified. It is defined as:

$$\text{Purity} = \frac{\text{Max}\left(\text{Number of sequence correctly classified in a family}\right)}{\text{Numberof sequence in that family}} \quad (1)$$

The proposed method is tested on 4 datasets. Dataset-1 is the mt-dloop of 14 (Yang and Rannala, 1997) species having two order, viz., primates and ferungulates with an average sequence length of 400. The mitochondrial D-loop is one of the fastest mutating sequence regions in DNA. Therefore, it is useful to compare the closely related organisms. Dataset-2 (Long-Hui *et al*., 2004) is the complete genomes sequences of 20 species with an average sequence length of 16 kB, Dataset-3 is the 48. The HEV strains (Long-Hui *et al*., 2004) with 4 classes belong to the dataset with an average sequence length of 7 kB and dataset-4 (Gupta *et al*., 2013) is the beta-globin genes of 6 species with an average sequence length of 400. All the datasets are having same structure i.e., name, Accession number, length and family. As an example details of dataset-2 is shown in Table 1.

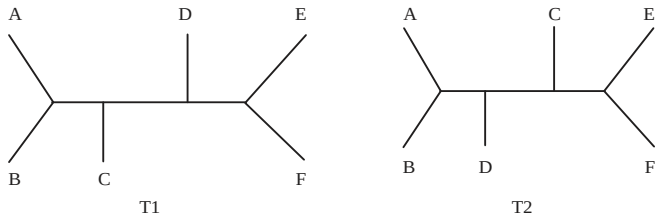

Fig. 1: Illustration to calculate symmetric distance

Table 1: Description of dataset-2

Name	Accession number	Length	Order
Human	V00662	16569	Primates
Chimpanzee	D38116	16563	Primates
Pigchimpanzee	D38113	16554	Primates
Horse	X79547	16660	Ferungulates
Rhinoceros	Y07726	16832	Ferungulates
Harbourseal	X63726	16826	Ferungulates
Bluewhale	X72204	16402	Ferungulates
Rat	X14848	16300	Rodentia
Mouse	V00711	16295	Rodentia
Greyseal	X72004	16397	Ferungulates
Finwhale	X61145	16398	Ferungulates
Opossum	AJ508398	17079	Non placental mammal
Wallaroo	Y10524	16896	Non placental mammal
Palutpus	X83427	17019	Non placental mammal
Cow	V00654	16338	Ferungulates
Cat	U20755	17009	Ferungulates
Baboon	Y18001	165521	Primates
Gibbon	X99256	16472	Primates
Orangutan	D38114	16389	Primates
Gorilla	D38115	16364	Primates

Algorithm for the Tree Construction using proposed method

The algorithm involves two procedures (*tuple, Similarity*) described as follows:
Input: for which tree is to be costructed sequqnce
Output: tree corresponding to input sequqnces
Range of i is from 2-4^k ,
procedure tuple (x) find all possible tuple of length i present in a sequence.

```
Procedure tuple (S, i)
        For sequence S of length l
        For i = 2 to 4^k
        Tup(i)=Substring of length I in S
Procedure Similarity(S,R)// find similarity of sequence S and R of
length N and M
        For i = 1 to N
        ktup = tuple (S, i), x++ // calling procedure tuple for S
ktup is array of tuple of length I
in sequence S
        end
        for i = 1 to M
        ktup= tuple(R, i), y++ //calling procedure tuple for R
ktup is array of tuple of length I
in sequence S
        end
        for i = 1 to x
                for j = 1 to y
                                if Ktup (Sx)= Ktup (Ry)
                                intersection++
                                else
                                unique++
                end
        end
        union = intersection+unique
        similarity = union/intersection

call procedure similarity for all sequence pairs to obtain matrix D
call UPGMA clustering algorithm for matrix D to obtain tree
```

RESULTS AND DISCUSSION

The proposed method is compared with composition vector method by varying k from 4-12. Figure 2 shows the effect of k on distance between 2 sequences using proposed method. It can be seen as the value of k is increased the distance (rat, cow) and distance (rat, mouse) is increased also the gap between the distance is also increasing for example at k = 9 distance (rat, cow) 6.7 is distance (rat, mouse) is 4.8 here the gap is 1.9 (6.7-4.8) and at k = 11 distance (rat, cow) is 19.96 and distance (rat, mouse) is 9.09 gap is 10 (19.96-9.09), whereas in Fig. 3 which shows the effect of k on distance using composition vector (Chan *et al*., 2012) method value of k is increased there is no significant increase in the gap between distance of the species. At k = 9 distance (rat, cow) is 40 is distance(rat, mouse) is 39 here the gap is 1 and at k = 11 distance (rat, mouse) is 512 and distance (rat, cow) 513 which means proposed method can significantly differentiate between similar and dissimilar sequence.

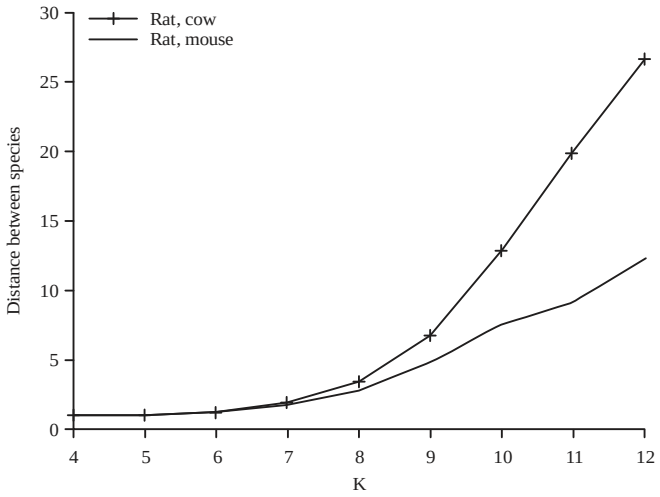

Fig. 2: Graphical Illustration of effect of k on (Rat, Cow) and (Rat, Mouse) using proposed method

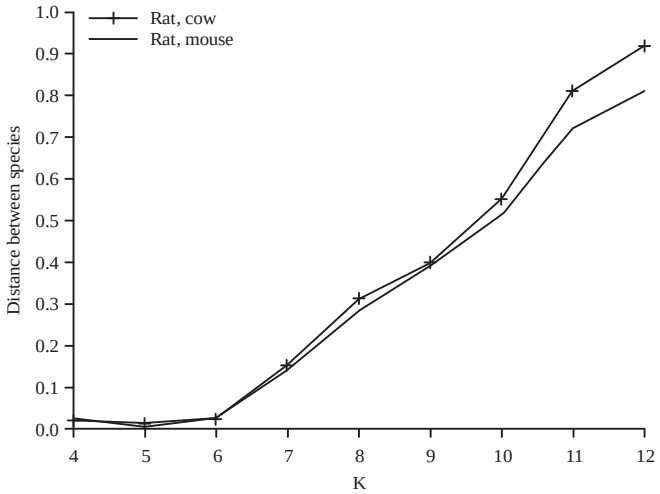

Fig. 3: Graphical Illustration of effect of k on (Rat, Cow) and (Rat, Mouse) using composition vector method

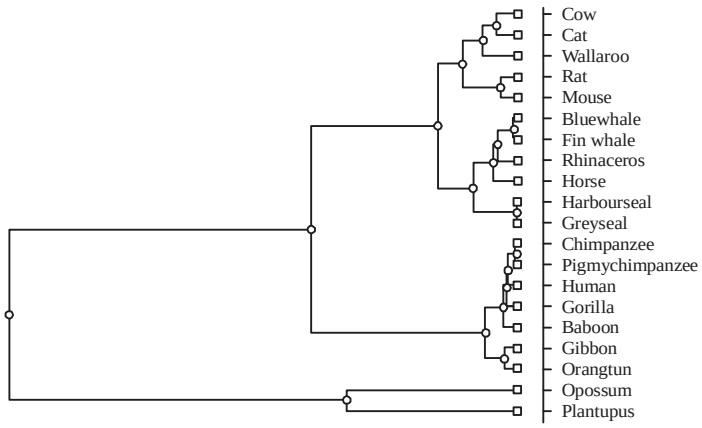

Fig. 4: Tree generated using proposed method for dataset 2 at k = 12

Table 2: Purity of the proposed method

Dataset	Purity of tree using proposed	Weighted relative entropy	Composition vector method	LZ complexity	K (for proposed method and CV)	No. of sequences	No. of families
1 (mt-DNA)	1.00	1.00	1.00	1.00	4	14	2
2 (complete genome)	0.85	1.00	1.00	0.91	8	20	4
3 (m-RNA)	0.833	0.79	0.80	1.00	6	6	3
4 (HEV strains)	1.00	0.82	0.93	1.00	9	48	4
Average purity	0.92	0.90	0.93	0.9775			

As can be seen in Fig. 4 which has a total of 20 sequences of which primates class has 7 sequences, ferungulates class has 8 sequences, rodents class has 2 sequences and the remaining 3 sequences are non placental mammals. At k = 12 for primates class all the sequences are clustered together, sequence of rodents (rat, mouse) are also clustered together. For ferungulates class (harbourseal, greyseal, horse, rhinocerous, finwhale, bluewhale) are clustered together but cow and cat have made another cluster thus affecting the purity of ferungulates. Non placental mammal (opossum and platypus) have clustered together but walaroo is misplaced again affecting the purity of cluster. For dataset 4 of HEV strains of 48 sequences belonging to 4 classes/families are clustered within their family (the prefix in specie name depicts its family) though the topology is different; hence the purity obtained for tree is 1. The purity (Eq. 1) of clusters of all dataset is given in Table 2 , as can be seen for dataset 1 and dataset 4 the purity is 1 where as for dataset 2 and 3 purity is 85 and 83, respectively. The trees generated using proposed method (Fig. 4 and 5) is compared with clustalw trees (Fig. 6 and 7) using PHYLIP (Felsenstein, 2004). And it can be seen in Table 3 the symmetric distance for all datasets is less as compared to original k-tuple except dataset 3.

The proposed method extracts common and uncommon subsequences of length k from the sequences to find similarity score. It is compared with four alignment free methods (Composition vector method (Chan *et al*., 2012), relative entropy (Bai *et al*., 2013), LZ (Otu and Sayood, 2003) and k-tuple methods (Yang and Zhang, 2008) along with alignment based tree generated using CLUSTALW on parameter including purity, symmetric distance. In Table 2 the average purity achieved from proposed method is 0.92 where as for (Bai *et al*., 2013) method the average purity is 0 90, for LZ(5) purity is 0.97 and for CV (Chan *et al*., 2012) is 0.93. In Table 3 average symmetric distance for proposed method is 8.75 whereas using k-tuple(Yang and Zhang, 2008) is 11.5. Average

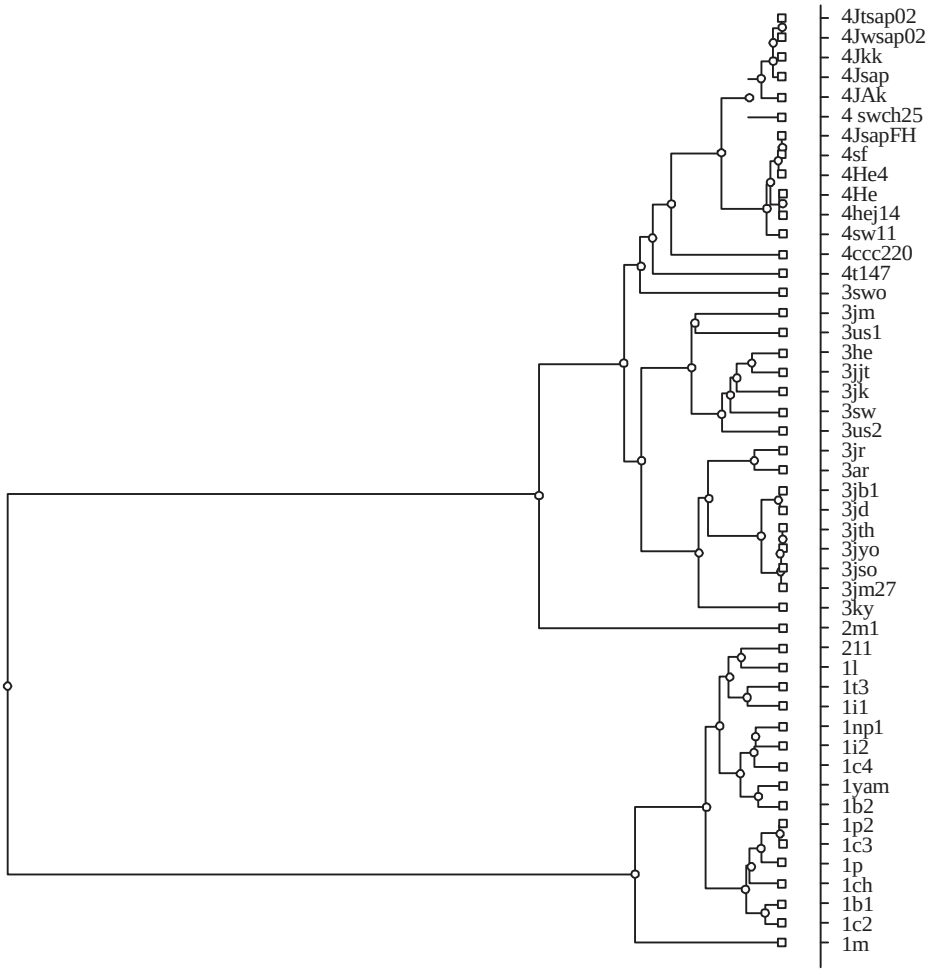

Fig. 5: Tree generated using proposed method for dataset 2 at k = 10 prefix 1, 2, 3 and 4 in specie name indicates its family 1m 1c9 belong to family 1

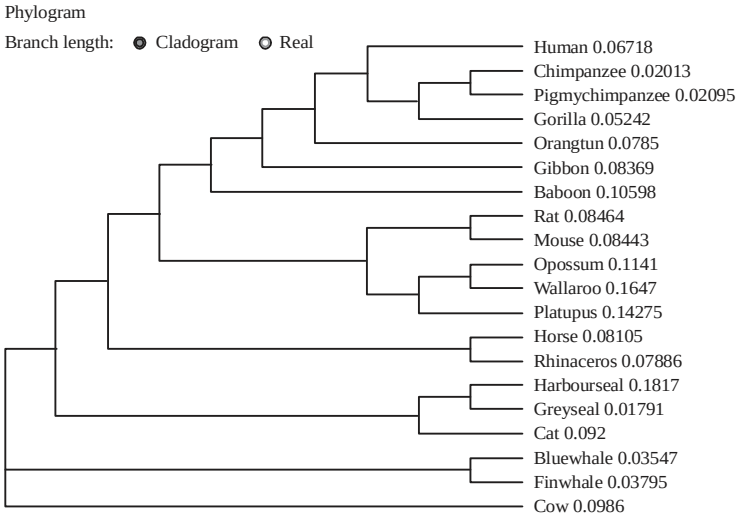

Fig. 6: Tree generated using clustalW for dataset 2

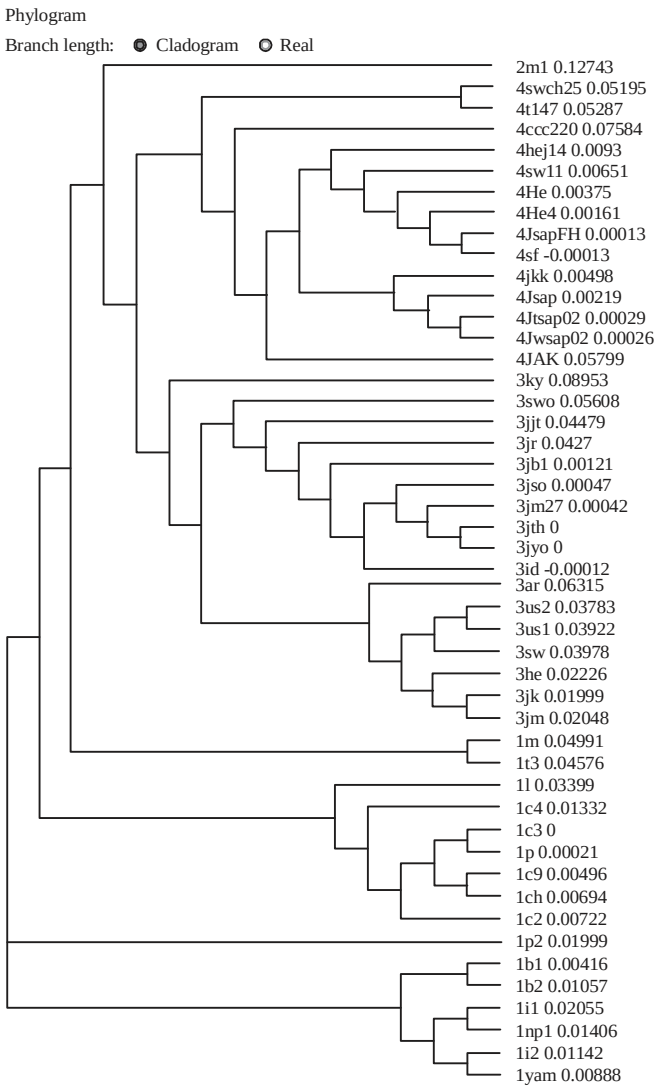

Fig. 7: Tree generates using clustalW for dataset 4

Table 3: Symmetric distance from the tree generated using clustalw

	Dataset				
Parameters	1	2	3	4	Average
Proposed	1	24	2	8	8.75
k-tuple method	2	32	0	12	11.50

purity is highest for LZ complexity but finding histories for long sequences is a very time and space consuming task. The purity and symmetric distance is calculated by taking tree generated using CLUSTALW as reference tree. Though there is no boundaries on k but if the value of k is increased i.e., the tuple length the time taken to find similarity between sequences is increases exponentially. The performance of all alignment free methods is dependent on number of species and length of sequences which is contradictory here as in dataset 4 with 48 sequences the purity is 1 where as for dataset 3 with number of sequences six the purity is 0.833.

CONCLUSION

This study focuses on finding distance between sequences to generate phylogenetic relationship. The idea is extending tuple count method to find closeness among sequences. This method does not require alignment, unequal sequence length and different positions of nucleotides do not interfere with the clustering results. The results show that the proposed method can successfully construct phylogenies using either the whole genome as shown in dataset II, or part of sequence, i.e., mt-dna dloop as in dataset I. The whole process is carried out in two stages where stage 1 finds the distance matrix and stage 2 creates the clusters from this matrix. The model is validated based on purity and symmetric distance measures. Experiment results show that the proposed method gives promising results as the value of k is increased. The results are at par with those of the alignment based methods (Fig. 6 and 7). The advantage is that it does not need sequence alignment. The performance of proposed method is better (Bai *et al*., 2013) in terms of purity and equivalent to CV (Chan *et al*., 2012) methods. In terms of symmetric distance also the method performs better that k-tuple method. Therefore it is concluded that the proposed method is capable of finding the similarity of phylogeny relationship among species.

REFERENCES

Bai, F., J. Xu and L. Liu, 2013. Weighted relative entropy for phylogenetic tree based on 2-step Markov model. Math. Biosci., 246: 8-13.

Chan, R.H., T.H. Chan, H.M. Yeung and R.W. Wang, 2012. Composition vector method based on maximum entropy principle for sequence comparison. IEEE/ACM Trans. Comput. Biol. Bioinform., 9: 79-87.

Felsenstein, J., 2004. Inferring Phylogenies. 1st Edn., Sinauer Associates, Sunderland, ISBN: 978-0-87893-177-4, Pages: 580.

Gupta, M.K., R. Niyogi and M. Misra, 2013. A framework for Alignment-free methods to perform similarity analysis of biological sequence. Proceedings of the IEEE 6th International Conference on Contemporary Computing, August 8-10, 2013, Noida, India, pp: 337-342.

Haubold, B., 2014. Alignment-free phylogenetics and population genetics. Briefings Bioinform., 15: 407-418.

Hohl, M. and M.A. Ragan, 2007. Is multiple-sequence alignment required for accurate inference of phylogeny? Syst. Biol., 56: 206-221.

Long-Hui, W., L. Juan, Z. Huai-Bei and S. Feng, 2004. A new distance metric and its application in phylogenetic tree construction. Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, October 7-8, 2004, La Jolla, CA, USA., pp: 108-115.

Louhisuo, K., 2004. Constructing phylogenetic trees with UPGMA and Fitch-Margoliash. http://www.niksula.cs.hut.fi/~klouhisu/Bioinfo/phyltree.pdf.

Orr, I., 2004. Introduction to phylogenetic analysis. Weizmann's Institute of Science, August 2004. http://bip.weizmann.ac.il/education/course/introbioinfo/04/lect10/phylogeny.pdf.

Otu, H.H. and K. Sayood, 2003. A new sequence distance measure for phylogenetic tree construction. Bioinformatics, 19: 2122-2130.

Wei, D. and Q. Jiang, 2010. A DNA sequence distance measure approach for phylogenetic tree construction. Proceedings of the IEEE 5th International Conference on Bio-Inspired Computing: Theories and Applications, September 23-26, 2010, Changsha, China, pp: 204-212.

Yang, K. and L. Zhang, 2008. Performance comparison between k-tuple distance and four model-based distances in phylogenetic tree reconstruction. Nucleic Acids Res., Vol. 36. 10.1093/nar/gkn075

Yang, Z. and B. Rannala, 1997. Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. Mol. Biol. Evol., 14: 717-724.

You, W., S. Gao, H. Deng, S. Ruan, H. Lu and Y. Zhang, 2009. Similarity analysis of DNA sequences using molecular connectivity indices method. Proceedings of the 6th International Conference on Fuzzy Systems and Knowledge Discovery, Volume 5, August 14-16, 2009, Tianjin, China, pp: 75-79.