Application of new distance matrix to phylogenetic tree construction

P.V.Lakshmi pvl_7097@rediffmail.com
Computer Science & Engg Dept
GITAM Institute of Technology
GITAM University
Andhra Pradesh
India

Allam Appa Rao apparaoallam@gmail.com
Jawaharlal Nehru Technological University
Kakinada
Andhra Pradesh
India

Abstract

Phylogenies are the main tool for representing the relationship among biological entities. Phylogenetic reconstruction methods attempt to find the evolutionary history of given set of species. This history is usually described by an edge-weighted tree, where edges correspond to different branches of evolution, and the weight of an edge corresponds to the amount of evolutionary change on that particular branch. Phylogenetic tree is constructed based on multiple sequence alignment, but sometimes alignment fails if the data set is large and complex. In this paper a new distance matrix is proposed to reconstruct phylogenetic tree. The pair-wise scores of input sequences were transformed to distance matrix by Feng Doolittle formula before solved by neighbor-joining algorithm. Two data sets were tested with the algorithm: BChE sequences of mammals, BChE sequences of bacteria. We compared the performance and tree of our result with ClustalX and found to be similar.

Keywords: Phylogeny, Bioinformatics, Distance matrix, Phylogenetic tree, neighbor-joining algorithm, Clustal X.

1. INTRODUCTION

Phylogenetic reconstruction methods attempt to find the evolutionary history of a given set of species. Phylogenies are reconstructed using data of all kinds, from molecular data, metabolic data, morphological data to geographical and geological data [1]**.** Phylogenetic analysis elucidate functional relationship within living cells [2-4]. With more and more DNA and protein sequences have been obtained, [5-7] the problem of inferring the evolutionary history and constructing the phylogenetic tree has become one of the major problems in computational biology. There are three major methods for performing a phylogenetic analysis, distance method, maximum parsimony, and maximum likelihood methods. Distance-matrix methods such as neighbor-joining or UPGMA calculate genetic distance from multiple sequence alignments. Maximum parsimony method implies an implicit model of evolution, predicts the evolutionary tree that minimizes the number of steps required to generate the observed variation in the sequences from common ancestral sequence. Maximum likelihood method start with simple model, in this case a model of rates of evolutionary change in nucleic acid or protein sequences and tree models that represent a pattern of evolutionary change, and then adjust the model until there is best fit of observed data. All this method requires multiple sequence alignment. Multiple sequence alignment is extension of pairwise sequence alignment. Needleman - Wunsch[8] and Smith-waterman [9] are classical dynamic programming algorithm for pair-wise sequence alignment with time and space cost of algorithm as O(mn), where m, n are lengths of two sequences to be aligned. Proposed algorithm takes O(m+n) time to generate score matrix .

2. Phylogenetic tree construction definition and previous work

A phylogenetic tree, is a model of the evolutionary history for a set of species. The neighbor-joining method by Saitou and Nei is a widely used method for constructing phylogenetic trees. It is a distance based method for constructing phylogenetic trees. It was introduced by Saitou and Nei [6], and the running time was later improved by Studier and Keppler [10].The neighbor-joining method is a greedy algorithm which attempts to minimize the sum of all branch-lengths on the constructed phylogenetic tree. It starts out with a star-formed tree where each leaf corresponds to a species, and iteratively picks two nodes adjacent to the root and joins them by inserting a new node between the root and the two selected nodes. When joining nodes, the method selects the pair of nodes i, j that minimizes the branch-length sum of the resulting new tree. Select the pair of nodes i, j that minimizes

$Q_{ij} = (r - 2)\, d_{ij} - (R_i + R_j)$--- (1)

where dij is the distance between nodes i and j.(assumed symmetric, i.e., $d_{ij} = d_{ji}$), R_k is the row sum over row k of the distance matrix. $R_k = \Sigma_i\, d_{ik}$ (where i ranges over all nodes adjacent to the root node), and r is the remaining number of nodes adjacent to the root. When nodes i and j are joined, they are replaced with a new node ‘U’ with distance to a remaining node k given by

$d_{Uk} = (d_{ik} + d_{jk} - d_{ij})/2$------- (2).

Repeat this until single node. We implemented neighbor-joining method taking distance matrix obtained by multiple progressive alignment technique.

3. New distance matrix algorithm

Using dynamic programming algorithm to find the score of two sequences take O(mn) time. To reduce computational time, a new method to calculate score of two sequences was proposed. Let *X* and *Y* be two sequences with lengths of *n* and *m,* respectively. The score of their alignment, namely the length of longest common subsequences, can be calculated using dynamic programming algorithm in *O*(*mn*) time. To reduce the computation time, a score estimating algorithm[11] to approximately estimate the score of a two-sequence alignment was considered. The algorithm estimates the score of the alignment of two sequences in *O*(*m*+*n*) time. The proposed algorithm consists of 4 steps each of which scans the two sequences from a different direction. Denote *X* and *Y* as the upper and lower sequence, respectively. The four steps are denoted as Left-Upper, Right-Upper, Left-Lower and Right-Lower. The step of Left-Upper starts from the first character in *X*, say *X*[0] and searches for the first matching character in *Y* from left to right. If there is no character in *Y* matching *X*[0], it restarts the scan to search for the character matching *X*[1] in *Y* . After such character, say *Y*[*j*], being found, the algorithm searches for the first character matching with *Y*[*j*] in the rest part of *X* from left to right. If there is no character in *X* matching *Y*[*j*], it restarts the scan to search for the character matching *Y*[*j*+1] in *Y* . After such character, say *X*[*i*], being found, the algorithm searches for the first character matching with *X*[*i*] in the rest of *Y* from left to right. We alternately repeat such scans until reaching the end of the sequences. As a result, the number of the matching characters can be obtained in the scan. *count* is the number of the matching characters in *X* and *Y*.

Algorithm:
Begin

```
Lt-up(X,Y,n,m,count1);
Lt-low(X,Y,n,m,count2);
Rt-up(X,Y,n,m,count3);
Rt-low(X,Y,n,m,count4);
Return (max(count1,count2,count3,count4))
End.
```

4. Method

For given set of sequences score matrix is obtained from proposed method. Distance matrix is generated using Feng –Doolittle formula.

$$d(x^{i},x^{j}) = -\log \frac{S(x^{i},x^{j}) - S_{rand}}{S_{max} - S_{rand}}$$

where S_{rand} is the mean score of two random sequences. It is taken as 4 defined constant and S_{max} is the maximum attainable score for two sequences. Phylogenetic tree was constructed using neighbor-joining algorithm.

5.Results

We tested our method on two datasets of BChE sequences, mammals and bacteria. The trees are generated using the Neighbor Joining (NJ)method [6]. And all the experiments in this paper were performed on a PC with Pentium IV CPU (ZGHZ), 512KB Cache, and 256MB RAM. We chose our group of sequences from first and second data set obtained from http://www.ncbi.nlm.nih.gov/.Mammals:Human(homosapiens,NP_000046),mouse(musmusculus, NP_033868),horse(Equuscaballus,NP_00075319),Sumatranorangutan(Pongoabeli,NP_001127509), cattle (Bos taurus, NP_001070374), domestic cat(Felis catus, NP_001009364),Norway rat(Rattus norvegicus,NP_075231),chimpanzee(Pantroglodytes,XP_516857),BengalTiger (Pantheratigristigris,AAC06262),chicken(Gallusgallus,CAC37792).Bacteria:Thiocapsaroseopersicina, Chlorobium tepidum, Rhodobacter sphaeroides, Rhodobacter capsulatus, Synechocystis sp pcc6803, Rosebacter denitrificans, Heliobacillus mobilis, Bradyrhizobium japonicum, Rhodopseudomonas palustris, Rhizobium etli cFN42, Lawsonia intracellularis, Rubrivivax gelatinosus, Candidatus Kuenenia stuttgartiensis, Bradyrhizobium sp, Chloroflexus aggregans. We applied the new distance measure to the above first data sets. Fig.1 shows the tree generated by proposed method. The tree is very close and Consistent with an earlier report published in *J Biol Chem.* 1991. The data set is applied to ClustalX[12] yielded a distance matrix, which was then analyzed by the NJ program, the result is shown in Fig.2. The results presented in Fig. 1 and Fig.2 are almost the same, but there are still some differences, for example cattle grouped with mouse and rat.

6. Conclusion

In this paper, we proposed a new sequence distance measure and used it to generate distance matrix for constructing phylogenetic tree. Unlike most existing phylogeny construction methods, the proposed method does not require multiple alignments and is fully automatic. We tested our method on two datasets and applied it to analyse the evolutionary relationship among BChE sequences. It is to be noted that we use no approximations and assumptions in calculating the distances between sequences, and our distance measure does not make use of any evolutionary model. It's one of the alignment free methods for phylogenetic tree construction of given sequences and is fully automatic.

Fig1 Tree constructed by ClustalX tool

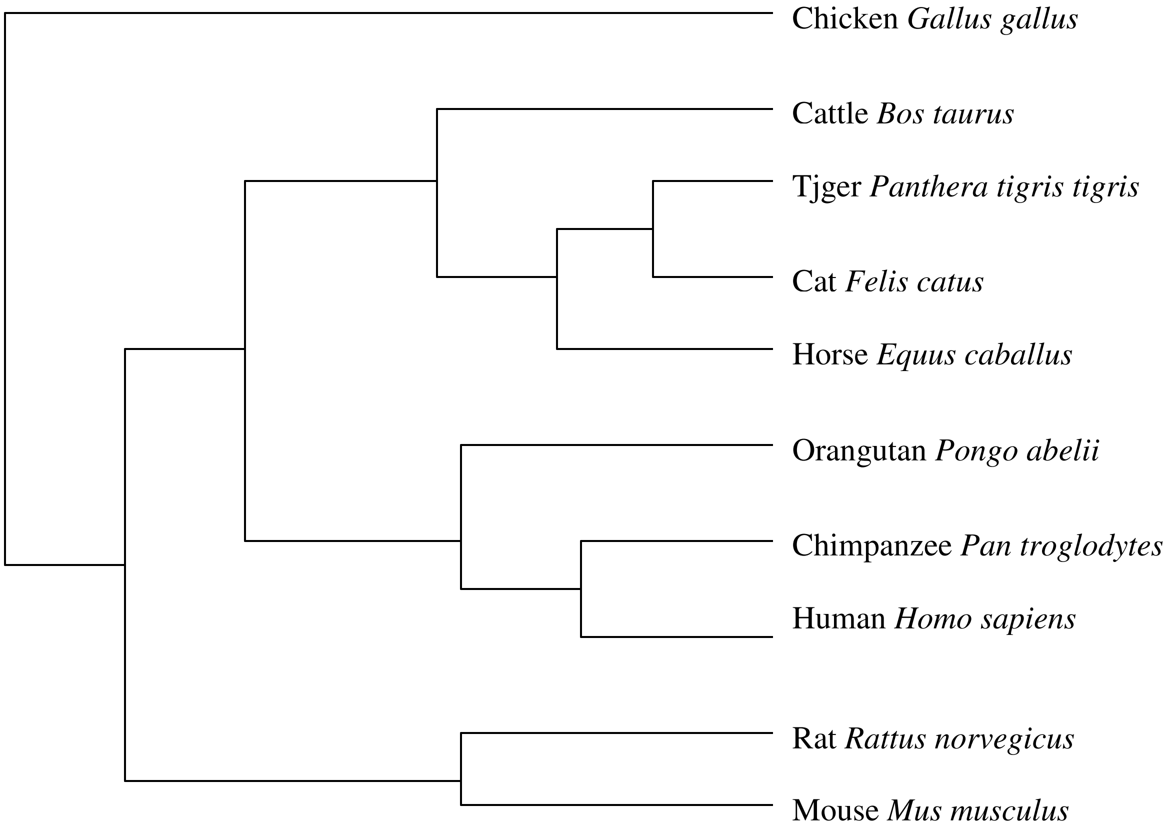
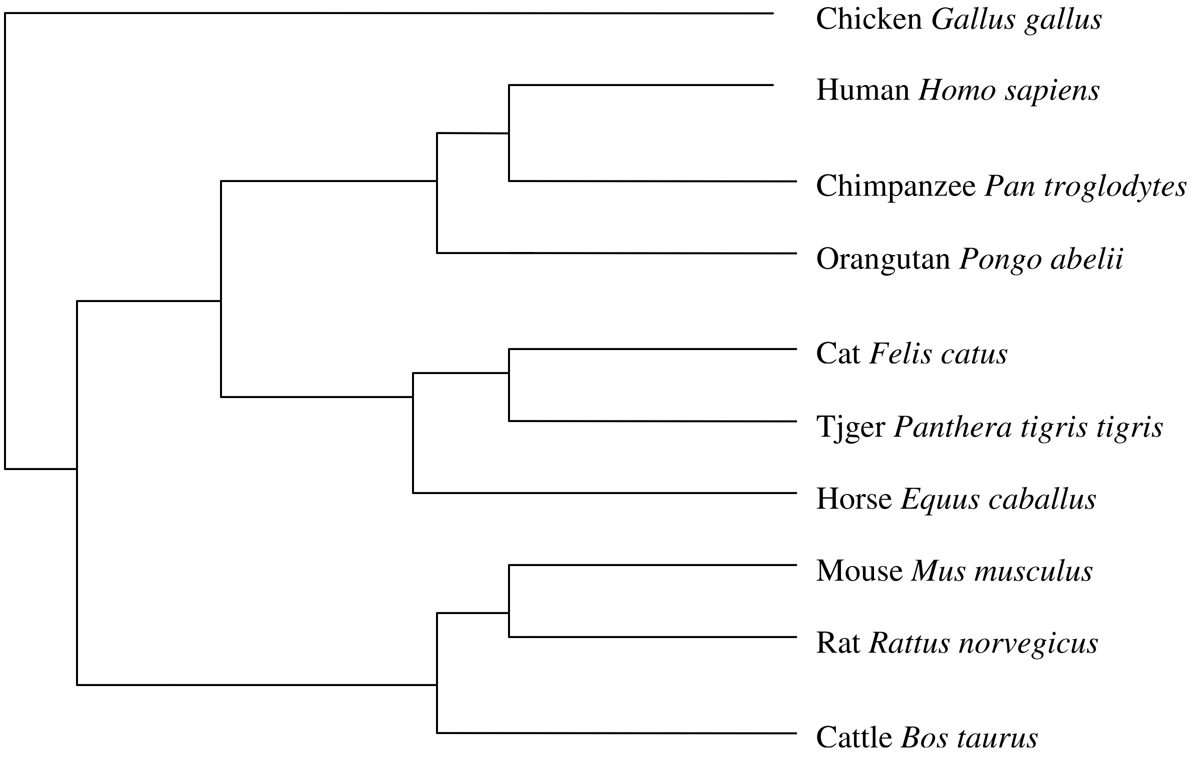


Fig2 Tree constructed by proposed method



7. References

1. Swofford, D.L., Olsen, G.J., Waddell, P.J. and Hillis, D.M. (1996).Phylogenetic inference. In *Molecular Systematics* (ed. D.M. Hillis, B.K.Mable, and C. Moritz), pp. 407.514. Sinauer Assoc.,Sunderland, MA.
2. M.Y. Galperin and E.V. Koonin. Comparative genome analysis. Methods Biochem. Anal.,43:359–392, 2001.
3 . X. Gu. Maximum-likelihood approach for gene family evolution under functional divergence.Mol. Biol. Evol., 18(4):453–464, 2001.
4. H. Zhu and J.F. Klemic et al. Analysis of yeast protein kinases using protein chips. Nature Genetics,26(3):283–289, 2000.
5. T. Hodge, M. J. T. V. Cope, “A myosin family tree,”*Journal of Cell Science*, Vol. 113, 2000, pp.3353-3354.
6. N. Saitou and M. Nei, “The neighbor-joining method: A new method for reconstructing phylogenetic trees,” *Molecular Biology and Evolution*, Vol. 4, 1987, pp. 406-425.
7 T. H. Reijmers et al., “Using genetic algorithms for the construction of phylogenetic trees: application to G-protein coupled receptor sequences,” *Biosystems*, Vol. 49, 1999, pp31- 43.
8. S.B.Needleman, and C.D.Wunsch, “A General method applicable to the search for similarities in amino acid sequence of two proteins”, *Journal of Molecular Biology* 48pp.443-453,1970.
9 T. Smith and M. Waterman, “Identification of common molecular subsequences,” *Journal of Molecular Biology*, vol. 147, pp. 195–197, 1981. Mol.Biol.Evol.,4,406-425(1987) .
10 Studier JA, Keppler KJ: A Note on the Neighbor-Joining Method of Saitou and Nei. *Mol Biol Evol* 1988, 5(6):729-731.
11 Partitioned optimization algorithms for multiple sequence alignment. Yixin Chen[1] Yi Pan[2] Ling Chen[3] Juan Chen[3]
12 ClustalX program. Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research, 25:4876-4882