Distance based Phylogenetic Trees with Bootstrapping

Jasmine Kaur
Research Scholar
Department of C.S.E,
G.N.D.E.C., Ludhiana

Pankaj Bhambri
Assistant Prof.
Department of IT
G.N.D.E.C., Ludhiana

O.P. Gupta
Dy. Director
School of IT.
P.A.U., Ludhiana

ABSTRACT
PHYLOGENY is the concept of phylogenetic trees which is typically a graphical representation of the evolutionary relationships among three or more genes or organisms. The tree construction can be done through different tree-building methods which include methods based on distances and characters. After a phylogenetic tree is being constructed, it is important to access its accuracy which refers to the degree to which a tree approximates the true tree. The best approach to test the phylogeneies is bootstrap analysis or simply bootstrapping. Bootstrapping is not a technique to check the accuracy of a tree. Instead, it describes the robustness of the tree topology. This paper discusses the searching and analyses of different possible inputs selected on the basis of family of genes or organisms so as to obtain the most optimal result. An algorithm was developed to determine the reliability of an inferred phylogenetic tree.

Keywords
Phylogenetic Tree, Phylogeny, Bootstrapping.

1. INTRODUCTION

Bio-informatics is an interdisciplinary research area at the interface between computer science and biological science. The fundamental aspect of bioinformatics is phylogenetics. The evolutionary history of a set of taxa is usually represented by a phylogenetic tree, and this model has greatly facilitated the discussion and testing of hypotheses [1]. The construction of phylogenetic trees is done by various methods. The main existing methods for reconstructing phylogenetic trees are based on maximum likelihood, Bayesian inference, maximum parsimony or distance. There are many distance based programs for phylogenetic trees reconstruction [2]. After the tree is constructed, the accuracy measure is being done. When a tree approximately gives the value of the true tree, then accuracy is being measured. Even when the alignment is done carefully and conditions such as substitution methods or scoring matrices are carefully chosen, a meaningless tree can still be generated. It means that the consistency, efficiency and robustness of the data applied matters. Moreover, the problem with phylogenetic inference is that the data which is in a single set is of finite length. New methods for pairwise alignments are also built [3]. In order to understand, analyze, and make use of the huge amount of data, bootstrapping approach is proposed to meet the challenge. It gives a way to judge the strength on support for nodes on phylogenetic trees that are built. The objective of our paper is to describe the bootstrapping algorithm and to compute the confidence levels of the branches of an inferred phylogenetic tree.

2. PROBLEM STATEMENT AND SOLUTION APPROACH

Bioinformatics is the science of managing, analyzing and interpreting information from biological sequences and various biological structures. In this area of science, biology, computer science and information technology, all the three merges into a single discipline [4].

Tree comparisons are used for multiple purposes, from unveiling the history of species to deciphering evolutionary associations among organisms and geographical areas [5]. In the case of phylogenetic trees, the accuracy measure is most significant and is of much importance. The problem is to estimate the significance of the branches of the tree. Large number of samples had to be taken in order to have an accurate confidence estimate [6]. So it becomes necessary to do the analyses by bootstrapping technique. The more times the data are sampled, better the bootstrap analysis becomes. The importance of applying re-sampling is that by re-sampling a number of times, it is possible to put reliability weights on each internal branch of the inferred phylogenetic tree. That implies that if the data is bootstrapped a 100 times, a bootstrap score of 100 means that the corresponding branch occurs in all 100 trees made from re-sampled alignments. Thus, a high bootstrap score is a sign of greater reliability [7].

The biological sequences are taken as input listed in Table 1. It contains the accession no's, name of the organism, the source of the organism and the gene index number. This list can be processed, taking some of the sequences at a time to find out the most optimal and probable result. The results can be shown graphically as the output.

Table 1: Biological Sequences taken as input

ACCESSION NO.	ORGANISM	SOURCE	GENE INDEX NO.
JQ362408	Rhipicephalus Sanguineus	Mitochondrion Rhipicephalus Sanguineus (brown dog tick)	>gi|379134831
JN990983	Felis Catus	Felis Catus (domestic cat)	>gi|379067379
JN190495	Carassius Auratus	Carassius Auratus (goldfish)	>gi|379998732
NM_022414	Mus Musculus	Mus Musculus (house mouse)	>gi|31543322
NM_001206572	Homo Sapiens	Homo Sapiens (human)	>gi|330864760

3. SOLUTION METHODOLOGY

Database is created for the different sequences of the family of organisms. It contains the names, accession no.'s etc. written in FASTA format.

Comparing and computing distances between phylogenetic trees are important biological problems, especially for models where edge lengths play an important role[8]. First of all, a phylogenetic tree is constructed by computing the pairwise distances between the sequences. Sequences can also be a vector of structures or a matrix of characters as well. It will return a vector 'D' containing biological distances between each pair of sequences stored in the M elements of the cell. D is an 1-by-(M*(M-1)/2) vector, corresponding to the M*(M-1)/2 pairs of sequences. The output D is arranged in the order of ((2,1),(3,1),..., (M,1),(3,2),...(M,2),.....(M,M-1), i.e., the lower left triangle of the full M-by-M distance matrix. To get the distance between the Ith and Jth sequences (I > J), we use the formula:

D ((J-1)*(M-J/2) + I –J)

To calculate the distance we use the formula:

for NT (nucleotide sequences):

$$d = -\frac{3}{4} * log\left(1 - p * \frac{4}{3}\right)$$

for AA (amino acids):

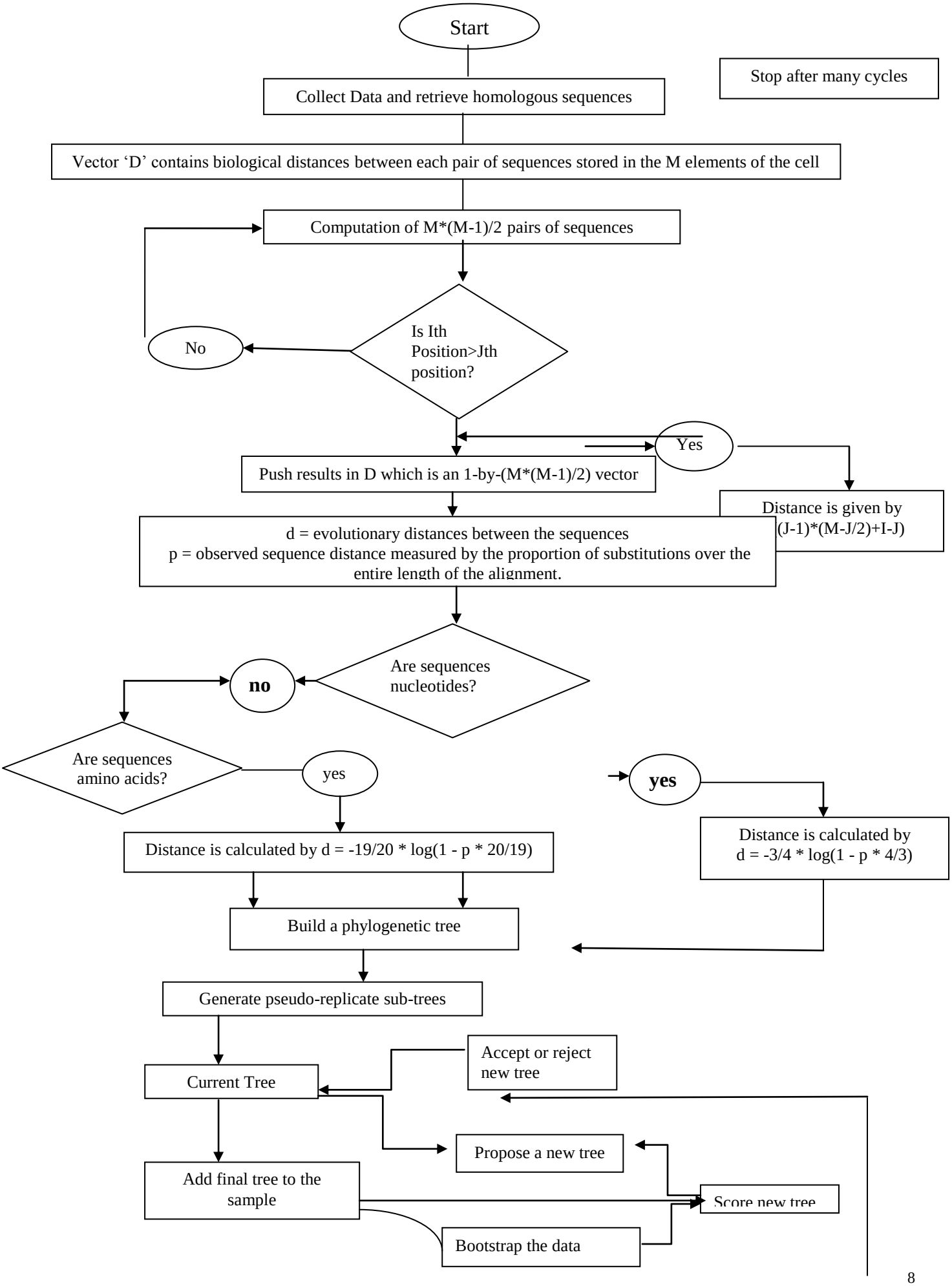
$$d = -\frac{19}{20} * log\left(1 - p * \frac{20}{19}\right)$$

where d = evolutionary distances between the sequences and p = observed sequence distance measured by the proportion of substitutions over the entire length of the alignment. Then the sequences are linked with the distances by taking the average of the distances and a phylogenetic tree is build.

Ahead of that, bootstrapping process starts. The bootstrap replicates are made from the data i.e the tree data on which the tree is being build. It is a shuffled representation of the DNA sequence data. The bases of the data are sampled randomly from the sequences with replacement and get concatenated to make new sequences. The gaps between the pairs of sequences are removed to force a new pairwise alignment. After a new pairwise alignment has taken place i.e the re-sampled alignment, it represents an estimate of how a different set of sequences from the same genes and the same species may have evolved on the same tree. After it is completed, the computation of the new alignments and the phylogenetic tree distances is being done again by the same method.

When the computation is being performed, different bootstrap i.e the replicate trees are generated from the original tree. The topology of every bootstrap tree is compared with that of the original tree. Any interior branch that gives the same partition of species is counted. Since branches may be ordered differently among different trees but still represent the same partition of species it is necessary to get the canonical form for each subtree before comparison[9]. The first step is to get the canonical subtrees of the original tree. The canonical trees have the similar topology as of the original tree. To be considered as similar they must have the same topology and span the same species. After counting the species, the process is repeated. It builds up a new tree with the same branches. But this new tree contains the confidence information of all the branches of the phylogenetic tree build earlier [10]. Now this comes to a conclusion that if a new tree reconstruction on the re-sampled alignment results in a tree similar to the original one, the confidence level in the original tree is being increased. If, on the other hand, the new tree looks very different, it means that the inferred tree is unreliable [11].

MATLAB is utilized to build the phylogenetic trees in the Graphical User Interface (GUI). The original tree is shown first then after replacement, the newer tree with confidence values is being shown. The methodology is being explained in the flow chart shown.


Start
Stop after many cycles
Collect Data and retrieve homologous sequences
Vector 'D' contains biological distances between each pair of sequences stored in the M elements of the cell
Computation of M*(M-1)/2 pairs of sequences
Is Ith Position>Jth position?
No
Yes
Push results in D which is an 1-by-(M*(M-1)/2) vector
Distance is given by (J-1)*(M-J/2)+I-J)
d = evolutionary distances between the sequences
p = observed sequence distance measured by the proportion of substitutions over the entire length of the alignment.
Are sequences nucleotides?
no
Are sequences amino acids?
yes
yes
Distance is calculated by d = -19/20 * log(1 - p * 20/19)
Distance is calculated by d = -3/4 * log(1 - p * 4/3)
Build a phylogenetic tree
Generate pseudo-replicate sub-trees
Accept or reject new tree
Current Tree
Propose a new tree
Add final tree to the sample
Score new tree
Bootstrap the data

4. CONCLUSION

The present model computes the accuracy of the phylogenetic tree and provides the result in the form of percentage count that can be compared for different sets of input as well. The results are also shown pictorially which further provides a detailed and complete description of the confidence intervals at their respective stage level. In comparison, Toress, M. developed a tool Digafu to exploit the best characteristics of phylogenetic programs. He stated the validation of the solution and was presented using real data. In comparison, the proposed work ensures the correctness and reliability of data stated. The screen shots are represented here in the paper. Such a model will help to determine the accuracy of the branches and there reliability as well.

Screenshots

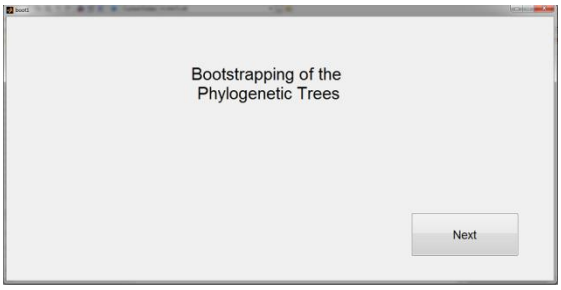


Fig 1: Home page

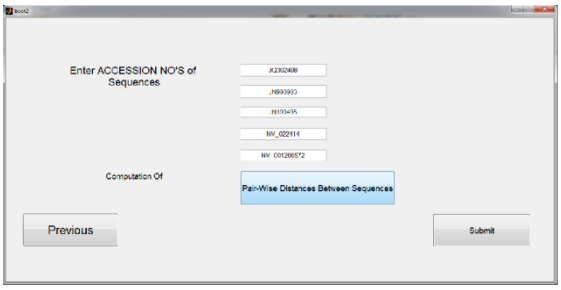


Fig 2: Entering accession no. of the sequences and calculating the paiwise distance among them

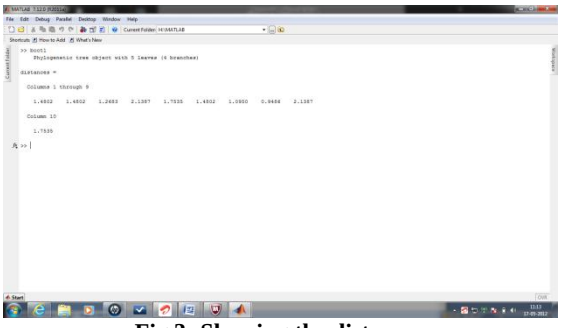


Fig 3: Showing the distances

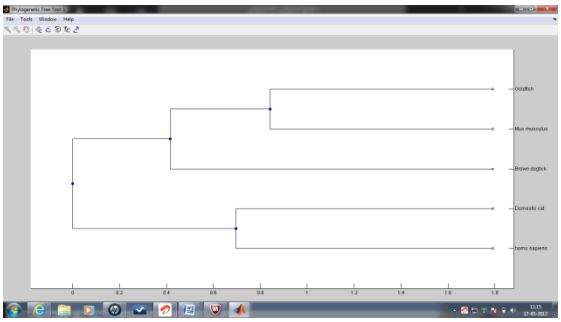


Fig 4: Phylogenetic Tree being created

Fig 5: Calculating the bootstrap replicates

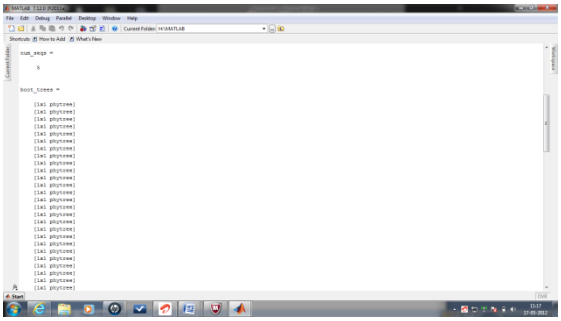


Fig 6: Showing bootstap replicate trees

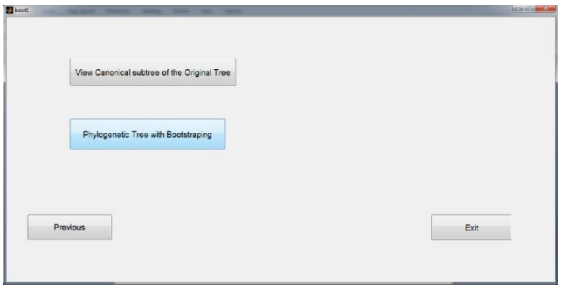


Fig 7: To build a phylogenetic tree with bootstrapping

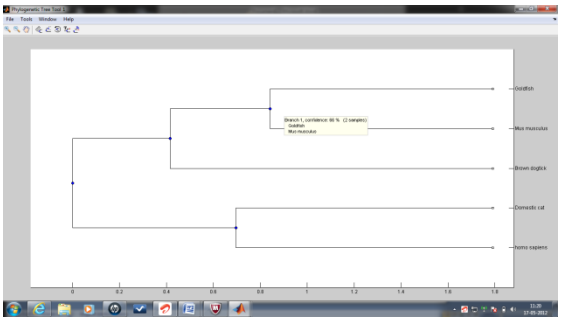


Fig 8: Displaying distance-based phylogenetic tree with confidence intervals

5. REFERENCES

[1] Huson, D. H. and Bryan, D. 2006 "Applications of phylogenetic networks in phylogenetic studies", Center for Bioinformatics (ZBIT), Tu¨bingen University, Tu¨bingen, Germany; and _Department of Mathematics, Auckland University, Auckland, New Zealand, pp 1-14.

[2] Torres, M. et al. 2011 "Tool that integrates distance based programs for reconstructing phylogenetic trees", Univ. Estaudal de Santa Cruz, Ilheus, Brazil, pp 895-901.

[3] Lin, Yu. et al. 2012 "A Metric for Phylogenetic Trees Based on Matching", EPFL, Lausanne, pp 1014-1022.

[4] Xiong, J. 2006 *Essential Bioinformatics*, Cambridge University Press.

[5] Bogdanowicz, D. and Giaro, K. 2011 "Matching Split Distance for Unrooted Binary Phylogenetic Trees", Gdansk University of Technology, Gdansk.

[6] Pevsner, J. 2009 *Bioinformatics and Functional Genomics*, 2nd ed., Wiley- BlackWell Publication.

[7] Weids, G. 2005 *Bioinformatics explained: Phylogenetics*, CLC bio, Denmark.

[8] Owen, M. and Provan, J.S. 2011, "A Fast Algorithm for Computing Geodesic Distances in Tree Space", Dept. of Math., Univ. of California, Berkeley, CA, USA, pp 2-13.

[9] Krane, D., E. and Raymer, M., L. 2006 *Fundamental concepts of bioinformatics*, 2nd ed., Pearson Education Publishers.

[10] Rastogi, S., C., Mendiratta, N. and Rastogi, P. 2007 *Genomics, Proteomics and Drug Discovery*, Prentice-Hall of India Private Limited, New Delhi.

[11] Mount, D., W. 2001 Bioinformatics*: Sequence and Genome Analysis*, 2nd ed., Cold Spring Harbor Laboratory Press.