International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 2, Issue 1, ISSN No. 2455-2143, Pages 22-26
Published Online November-December 2016 in IJEAST (http://www.ijeast.com)

EVALUATING THE CONSTRUCTION OF PHYLOGENETIC TREE USING HIERARCHICAL CLUSTERING AND BOOTSTRAPPING

Pushpinder Kaur
M-Tech, CSE Dept.
Amritsar college of Engineering and Technology
Amritsar, India

Navneet Bawa
Associate Professor, CSE Dept.
Amritsar college of Engineering and Technology
Amritsar, India

Abstract- **Phylogenetic trees are constructed from the sequences of the different species. These are actually needed to find the relationship between the different species and also different time gaps from the actual origin. The main reason for phylogenetic tree construction is to foresee the structure of unknown sequence of different species. Methods for construction of phylogenetic tree are distance or character data. In the present work, the methods used for constructing a distance based NJ, Jukes Cantor based phylogenetic tree and phylogenetic tree using bootstrapping. There are different data formats available out of which fasta format is used. Bootstrapping technique is applied to resample data number of times. More times the data is sampled better the analysis becomes. Ultimate results are also shown in Clustering form.**

Keywords— ***Bioinformatics; Phylogenetic tree; UPGMA; NJ; Bootstrapping.***

I. INTRODUCTION

Bioinformatics is a scientific discipline formed from the Combination of biology and computer. Bioinformatics is basically a tool for biological data. The main solicitation of bioinformatics is categorized into sequence analysis and structure analysis. Bioinformatics is an combination of mathematical, statistical and computer methods to analyse biological, biochemical and biophysical data. There are macromolecules in it such as DNA, RNA and proteins. DNA is transcribed to RNA which is futher translated to proteins. Phylogenetics is the study of evolutionary history of some species. It is used to detect genetic relationship between different species

These are actually needed to find the relation between the different species. Unweighted Pair Group Method With Arithmetic Mean is a hierarchical clustering Technique used for the construction of phylogenetic tree. Hierarchical clustering is used to retrieve the results. Bootstrapping is a technique that allows rough quantification of confidence levels attached with the overall tree and its branches.

II. PHYLOGENETIC ANALYSIS

Phylogenetic, illustrate patterns of shared history between biological replicators, such as species or genes. Two main types of trees are:

1) Rooted trees—In this all nodes are derived from single node.

2) Unrooted trees—It is not clear that where the nodes originated from.

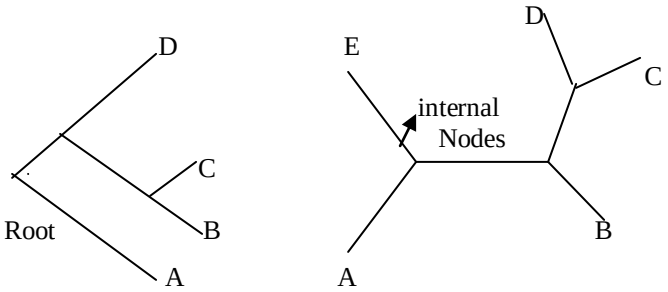

Fig. 1. Figures representing rooted and unrooted trees

III. PHYLOGENETIC TREE CONSTRUCTION

Phylogenetic trees consists of various input sequences and tree is build using various distance method

Various distance based methods include:

A. Distance Based Methods

Distance Method makes a table that describes pairwise dissimilarity .In this we calculate all the distance between species. Based on the distance, we construct a tree .This method is good for continuous character. Distance based Methods are UPGMA and Neighbour Joining.

1) *UPGMA*- Unweighted Pair Group Method With Arithmetic Mean take closely related pair of sequences to

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 2, Issue 1, ISSN No. 2455-2143, Pages 22-26
Published Online November-December 2016 in IJEAST (http://www.ijeast.com)

build a phylogenetic tree. UPGMA works by clustering the most similar taxa until all the taxa form clock-like tree.

(1) At the first each species is a cluster on its own.

(2) Closest two clusters are combine and recalculates distance of the pair by taking the average.

(3) The procedure is repeated again and again until all species are joined in single cluster.

2) NEIGHBOUR JOINING (NJ) - NJ is a clustering technique for the reconstruction of phylogenetic trees. NJ method provides resultant tree with branched lengths.

IV. DATA MINING

Data mining is the progression of take out patterns from data. Data mining is an significant tool to change the data into useful information .It is the progression of finding out information from bulks of data and this can be achieved through the examination of relations and trends within commercial databases. It is used in areas like space, Statistics, bioinformatics and medical research. Data Mining Techniques:-

1) Cluster Analysis

- Hierarchical Clustering Analysis
- Non- Hierarchical Clustering Analysis

(A) Cluster Analysis-

Cluster analysis means examine the data objects. The data objects are grouped based on the principle that maximum the resemblance of intraclass and minimize the resemblance of interclass.

In this Clusters of objects are made so that the objects that are inside a cluster have high similarity in comparison to one another but are very dissimilar to other clusters.

1)Hierarchical Clustering- This technique create a ladder of cluster from small to big .It produces a set of related cluster and can be structured as hierarchical tree.

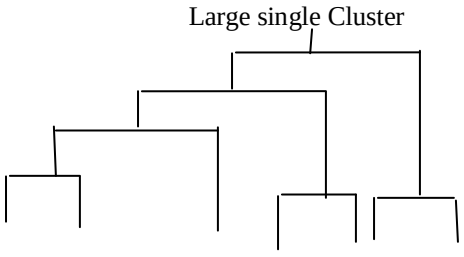

Fig. 2 shows Hierarchical Clustering

2) Non-Hierarchical Clustering-Non-hierarchical clustering do not create a hierarchy of clusters. This technique is very profligate to compute.

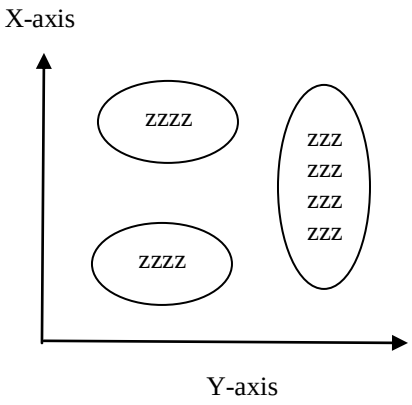

Fig.3 shows Non Hierarchical Clustering

V. METHODOLOGY

The methodology for this work can be used to explore the different phylogenetic tree construction techniques . In case of phylogenetic tree construction accuracy measure is one of the most important. Large number of samples taken more is the accurate confidence estimate bootstrap analysis becomes necessary to have an accurate estimate.Re-sampling technique is applied to re-sample the data number of times, so that it becomes possible to put reliability weights on each branch of the tree and provides the corresponding bootstrap score. Thus, high bootstrap score provides greater reliability.

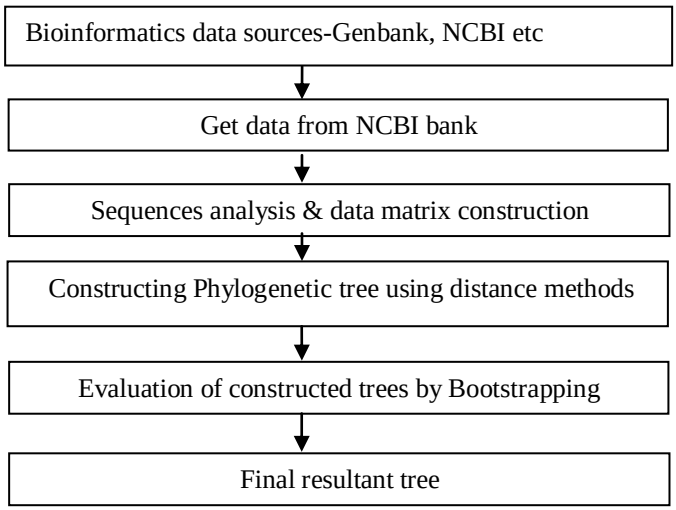

Fig:4 Methodology for Proposed work

A. Jukes Cantor Model

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 2, Issue 1, ISSN No. 2455-2143, Pages 22-26
Published Online November-December 2016 in IJEAST (http://www.ijeast.com)

Jukes Cantor model figures the probability of substitution from one state to another. In this model we can find distance between 2 sequences. We can find the distance both for nucleotide as well as amino acids. In this model probability of fluctuating from one state to another is always equal. The distance between two species is given by the following formula.

For Nucleotide Sequence

$$D = -3/4 \log (1-4/3\ Nd/N)$$

Where D defines distance between the two sequences and
Nd is the number of mutations between sequences
N is the nucleotide length

For Amino Acid Sequence

Distance is calculated by D= -19/20*log(1-p*20/19)

Where D defines evolutionary distances between the sequences
p defines the sequence distance measured by the proportion of substitutions

B. Algorithm for Jukes Cantor Model

This research work includes the phylogenetic tree construction for the different sabertooth varieties.
The phylogenetic tree is constructed and distance is calculated by jukes cantor method. The resulting tree is obtained by using bootstrapping technique
The algorithm steps are written below:

- Database is made by taking the sequence from NCBI site
- Distance is calculated by jukes cantor method.
- Construct the matrix –I based on Jukes Cantor distance.
- Find the smallest distance by examine the original matrix.
- Construct the reduced matrix and distance is calculated
- Find the smallest distance by examining the reduced Matrix .Process is repeated again and again until all species are choosen.
- Construct the tree 1.
- Create another matrix-2
- Repeat process again and again
- Construct final tree
- Apply the bootstrapping technique

Then the bootstrapping process starts. The replication of data is build. The bases of data are randomly samples and then combine to make new sequences.
It is basically shuffled representation of data

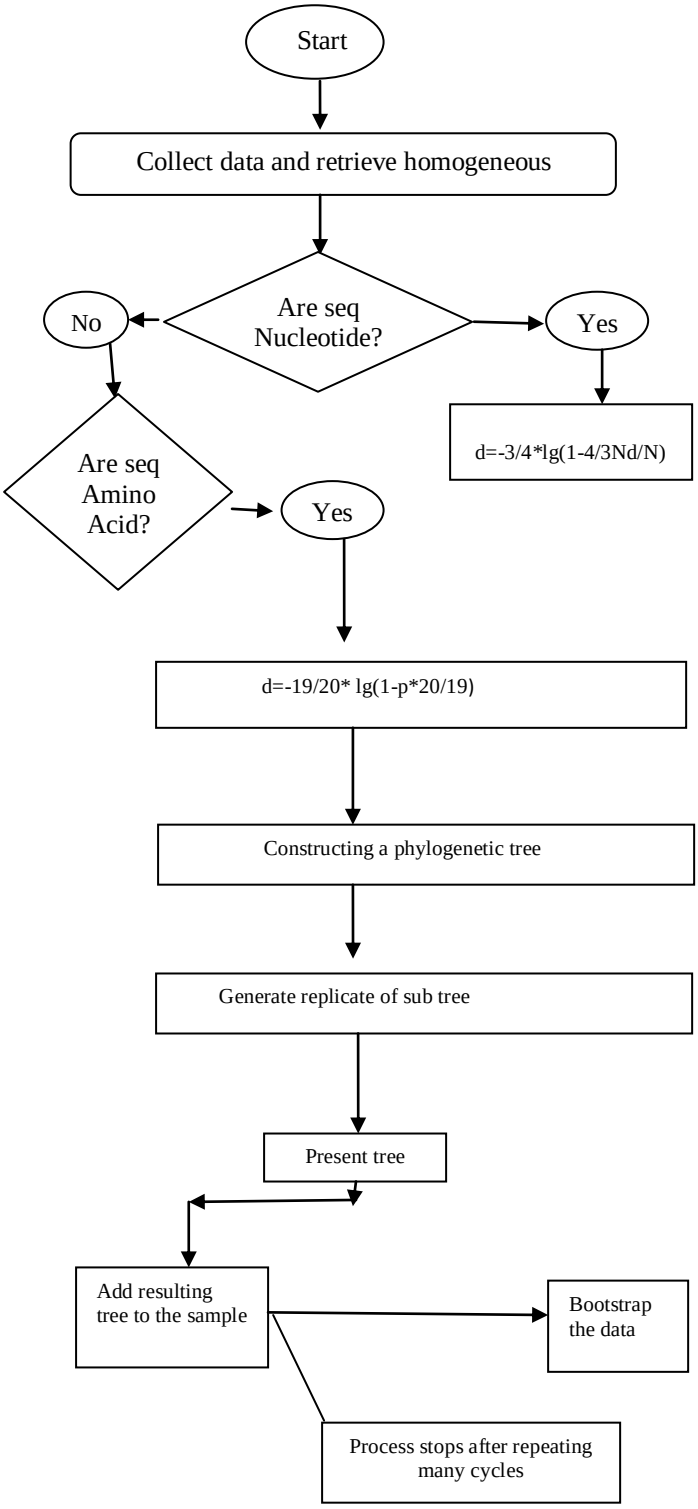

Fig 5: Flowchart of Bootstrapping process

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 2, Issue 1, ISSN No. 2455-2143, Pages 22-26
Published Online November-December 2016 in IJEAST (http://www.ijeast.com)

VI. EXPERMENTAL ANALYSIS

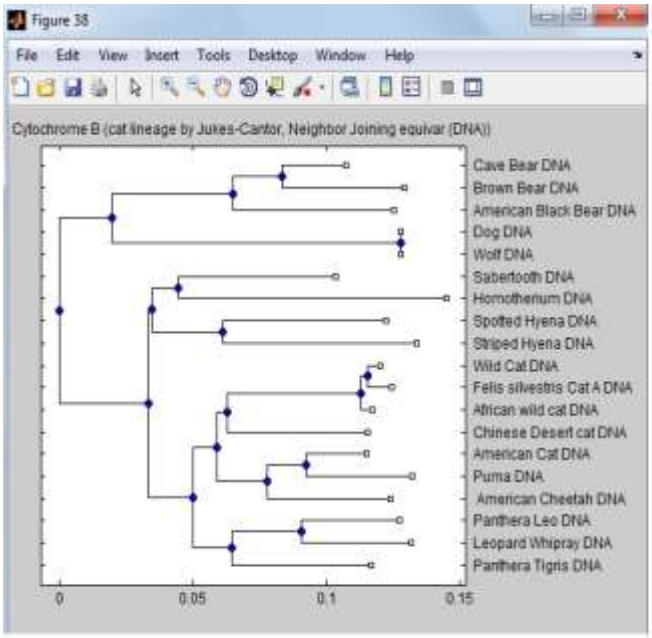

Fig.6: Phylogenetic tree construction for DNA sequence using Neighbor Joining

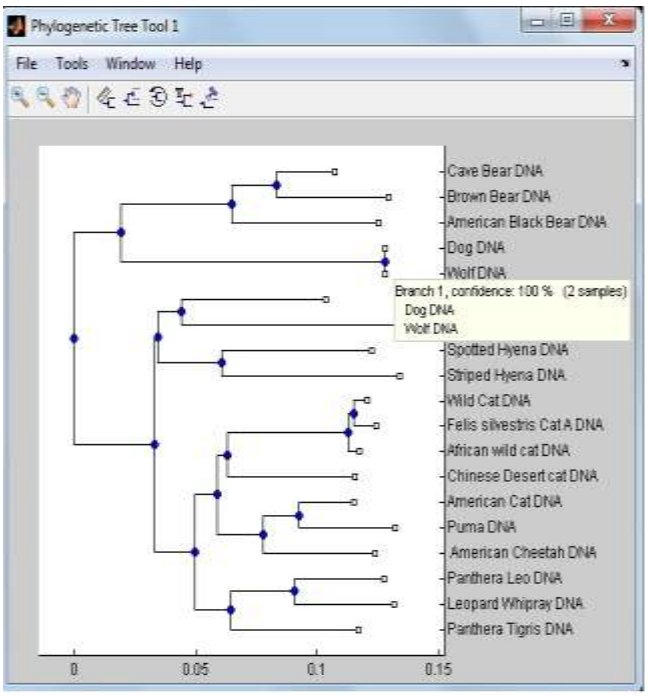

Fig.8: Phylogenetic Tree Showing Confidence Level for 2 Samples

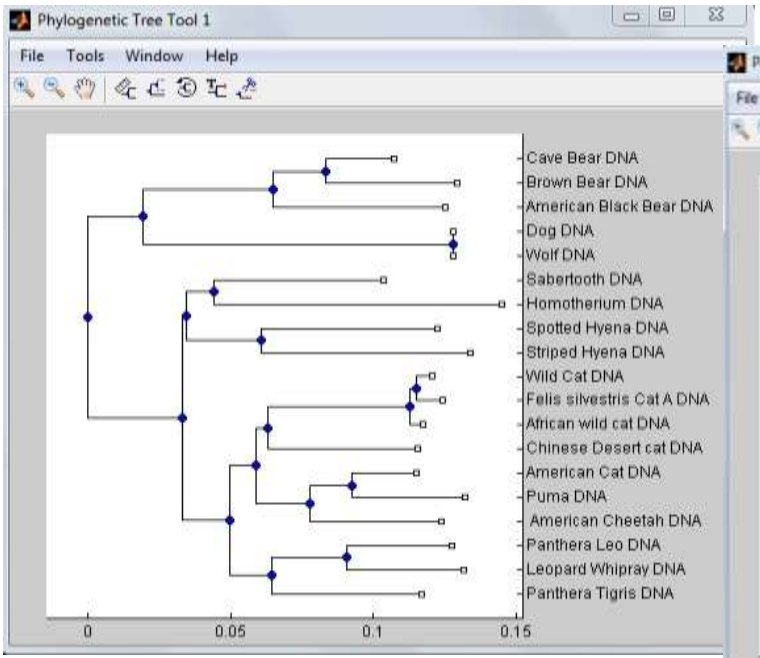

Fig.7: Phylogenetic Tree Tool 1

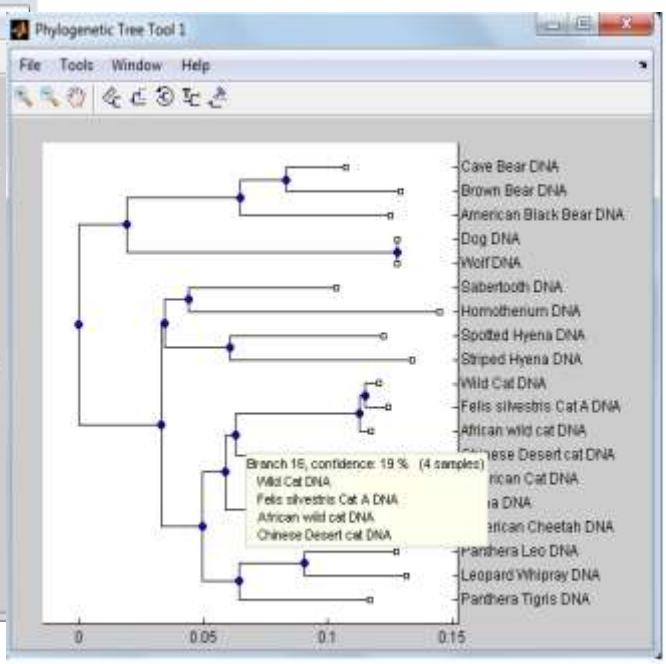

Fig.9: Phylogenetic Tree Showing Confidence Level for 4 Samples

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 2, Issue 1, ISSN No. 2455-2143, Pages 22-26
Published Online November-December 2016 in IJEAST (http://www.ijeast.com)

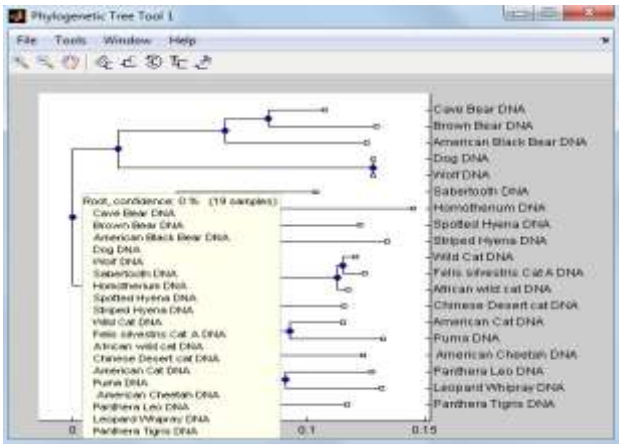

Fig.10: Phylogenetic Tree Showing Confidence Level for 19 Samples

VII. RESULTS

Following results tell us about confidence probability attached with each branch and then find the mean confidence probability:

No of samples	Branch No	Confidence	Confidence probability 1	Confidence probability 2	Confidence probability 3	Mean Confidence Probability
2	Branch 1	100%	1.0000	1.0000	1.0000	1.00
2	Branch 2	90%	0.9000	0.8900	0.8200	0.87
3	Branch 3	89%	0.8900	0.8900	0.8100	0.86
5	Branch 4	82%	0.8200	0.8500	0.7600	0.81
2	Branch 5	54%	0.5400	0.6400	0.6800	0.62
3	Branch 6	54%	0.5400	0.6400	0.6800	0.62
2	Branch 7	94%	0.9400	0.9500	0.9600	0.94
2	Branch 8	87%	0.8700	0.9000	0.9100	0.89
2	Branch 9	85%	0.8500	0.8700	0.9500	0.89
3	Branch 10	72%	0.7200	0.6800	0.8800	0.76
3	Branch 11	52%	0.5200	0.6200	0.6100	0.58
2	Branch 12	34%	0.3400	0.4400	0.4400	0.41
4	Branch 13	13%	0.1300	0.2500	0.1300	0.17
14	Branch 14	9%	0.1900	0.2200	0.2300	0.21
10	Branch 15	2%	0.0700	0.0900	0.1400	0.10
4	Branch 16	19%	0.0200	0.0400	0.0800	0.05
7	Branch 17	7%	0.0000	0.0100	0.0100	0.01
19	Branch 18	0	0.0000	0.0100	0.0100	0.01

VIII. CONCLUSIONS

The bootstrapping model is used to calculate the confidence interval between the sequences of same family of an organism. Analysis of phylogenetic tree can be done by nucleic acid and protein sequences. There are various sequence formats available out of which FASTA format is used. Bootstrapping Technique is applied across character data matrix to produce replication of data sets which is analyzed phylogenetically, with a consensus tree constructed to recapitulate the results of all replication. Bootstrapping provides a confidence interval that provides the phylogeny that would be estimated from repetition of sampling of many characters from the underlying data set of all characters.

With the help of bootstrapping, large numbers of sequences are easily handled. The overall advantage of this method is the ability to build an optimal and accurate phylogenetic tree with précised confidence time intervals. Different data sets are taken and are being studied to retrieve a consensus tree. The phylogenetic trees of all the different data are build. These trees are very much similar but have some differences.

IX. FUTURE SCOPE

- The model can be further stretched to evaluate number of nodes that requires fast and accurate algorithms.
- Bootstrapping and Jukes Cantor approach can be extended further to give an optimal result.

X. REFERENCES

[1].Pushpinder kaur and Navneet Bawa (2016) " Sequence Analysis and Phylogenetic tree construction using Jukes Cantor", International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 5, Issue 6.

[2].Pushpinder kaur and Rajbir Singh (2016) " Phylogenetic tree construction using Data Mining", International Conference on Sciences, Engineering & Technical Innovations, pp 53-56.

[3]. Rajbir Singh and Dheeraj Pal Kaur (2015) "Improved Distance based Phylogenetic Tree Construction using Bootstrapping Method", International Conference on Information Technology and Computer Science pp.11-12.

[4]. Shaminder Kaur, Baldeep Singh and Tajinder Kaur (2015) "Improved Computational Methods for Phylogenetic Tree Construction using Cluster Analysis" ,International Journal of Advanced Research in Computer Science and Software Engineering ,Vol. 5, pp. 378-385 .

[5]. Joseph L. Staton (2015) "Understanding phylogenies: Constructing and Interpreting Phylogenetic trees", Journal of the South Carolina Academy of Science, Vol. 13, pp.24-29.

[6]. J.Jayapriya and Michael Arock (2015) "Enhanced bio-inspired algorithm for constructing phylogenetic tree", ICTACT journal on soft computing, Vol.6.

[7] .Anand Patwardhan, Samit Ray and Amit Roy. "Molecular Markers in Phylogenetic Studies",Phylogenetics & Evolutionary Biology, Vol. 2, pp.2-9.

[8]. Yang Ruan, Geoffrey L. House, James D. Bever, Haixu Tang and Geoffrey Fox. "Integration of Clustering and Multidimensional Scaling to Determine Phylogenetic Trees as Spherical Phylograms Visualized in 3 Dimensions".