Constructing Evolutionary Tree [Phylogeny] using BIRCH Algorithm

V. Bhuvaneswari[1] R.Sindhu[2]

Abstract

As the number of sequences in the GenBank database has increased exponentially, biologists are scrambling to place the information in meaningful context. Researchers try to reconstruct the branching by looking at the similarities and differences of the DNAs of the present day individuals. Phylogeny is the study of the historical pattern of relationships among organisms which has resulted from the actions of many different evolutionary processes. Phylogenetic relationships are depicted by branching diagrams called cladograms or phylogenetic trees. Mining biological data is an emerging area of intersection between bioinformatics and data mining. The data mining techniques are widely used in biology to interpret the information underlying in the sequences. Clustering is a useful data mining technique for the discovery of data distribution and patterns in the data. In this research work an attempt is made to find the evolutionary relationship using a data mining hierarchal agglomerative clustering technique. The BIRCH algorithm is implemented to cluster DNA sequences to find the phylogenetic relationship among the organisms.

Keywords: Phylogeny, Bioinformatics, Data Mining, Hierarchical clustering, BIRCH algorithm, DNA sequences.

1. Introduction

Bioinformatics is an interdisciplinary field at the intersection of biology, computer science, and information technology[15]. Deoxyribonucleic Acid or DNA is commonly known to be responsible for encoding the information of life[4][5]. DNA is passed on as hereditary material to offspring from parents. Through many generations of reproductions, with mutations going on between every generation, different species emerge [28]. One of the most fundamental aspects of bioinformatics in understanding sequence evolution and relationships is molecular phylogenetics, in which the evolutionary histories of living organisms are represented by finite directed (weighted) trees.

A phylogenetic analysis of a family of related nucleic acid or protein sequences is a determination of how the family might have been derived during evolution. The evolutionary relationships among the sequences are depicted by placing the sequences as outer branches on a tree. The branching relationships on the inner part of the tree reflect the degree to which different sequences are related. Two sequences that are very much alike is located as neighboring outside branches and are joined to a common branch beneath them[9].

A phylogeny or evolutionary tree, represents the evolutionary relationships among a set of organisms or groups of organisms, called taxa . The tips of the treerepresent groups of descendent taxa (often species) and the nodes on the tree represent the common

[1]Lecturer , [2]MPhil Research Scholar, School of Computer Science and Engineering, Bharathiar University. Email : bhuvanes_v@yahoo.com , sindhurmohan@gmail.com

ancestors of those descendents. Two descendents that split from the same node are called sister groups.

In Figure 1, species A & B are sister groups — they are each other's closest relatives. A taxon outside the group of interest is known as an outgroup. The outgroup stems from the base of the tree. An outgroup gives a sense of where on the bigger tree of life the main group of organisms falls. It is also useful when constructing evolutionary trees[4].

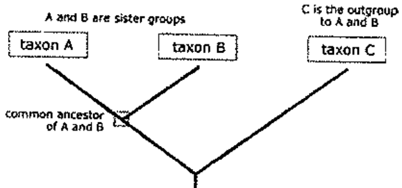

A and B are sister groups
C is the outgroup to A and B
taxon A
taxon B
taxon C
common ancestor of A and B

Figure 1: Phylogenetic Tree

The aim of the researcher is to analyze the evolutionary relationships among organisms by using a clustering technique. Clustering is a technique in which similar items are put together in a group. There are number of clustering algorithms applied successfully for various domains to infer knowledge. An attempt is made in this work to cluster the biological sequences to infer the evolutionary relationships that exist among different sequences.

The paper is organized as follows. In Section 2 literature review of phylogeny methods and literature review on data mining clustering techniques are discussed. Section 3 describes methodology of phylogeny tree construction using BIRCH algorithm and Section 5 describes the experimental results of BIRCH algorithm.

2. Overview of Phylogenetic Methods

The most popular and frequently used methods of tree building is classified into two major categories .Phenetic methods based on distances and Cladistic methods based on characters. The phenetic methods are distance-based methods that measure the pair-wise differences among sequences under study and build the tree from the resultant distance matrix. Distance methods compress sequence information into a single number and the two sequences with the shortest distance are considered as closely related taxa. In the phenogram, taxa with the shorter distances are classified as more closely related[9] . The two different algorithms in distance based method are Cluster-based and Optimality-based.

The cluster-based method algorithms build a phylogenetic tree based on a distance matrix[9] .The algorithms of cluster-based include, Unweighted pair group method using arithmetic average (UPGMA) , Neighbor joining (NJ) method. The optimality-based method algorithms compare numerous different tree topologies and select the one which is believed to best fit between computed distances in the trees and the desired evolutionary distances which is often referred as actual evolutionary distances [10].Algorithms of optimality based include, Fitch-Margoliash , Minimum evolution method .

3. Related Works On Phylogeny Tree Construction

The literature study carried out it was found that the distance is used as a measure to construct phylogeny tree for the sequences. The aim of the researcher is to cluster the sequence based on their evolutionary distance using an unsupervised hierarchal clustering technique. The various reviews related to the work are given below Kimura M (1980) [23], has developed formulae which is used to estimate evolutionary distances in terms of the number of nucleotide substitutions. The two nucleotide substitutions types allowed are those between transitions and transversions. Saitou N, Nei M (1987) [8], has developed a new method called the neighbor-joining method for reconstructing phylogenetic trees from evolutionary distance data. John Sourdis, Masatoshi Nei (1988) [24], has studied the relative efficiencies of the maximum parsimony (MP) and distance-matrix methods in obtaining the correct tree (topology) using computer simulation. The results obtained indicates that if the number of nucleotide substitutions per site is small and a relatively small number of nucleotides are used, then the probability of obtaining the correct topology is generally lower in the maximum parsimony method than in distance-matrix methods. William J. Bruno, Nicholas D (1999) [20], has developed a distance-based phylogeny reconstruction method called 'weighted neighbor joining' or 'Weighbor' for short. The results obtained indicate that parsimony, NJ and Fitch had performed worser than Weighbor on four-taxon trees with a short internal branch and two long branches that are not joined. Stephane Guindon , Oliver Gascuel (2003) [12] has developed a fast reliable algorithm for Phylogeny tree reconstruction. Algorithm is based on the maximum-likelihood principle.The topological accuracy of this new method is high as that of the existing maximum-likelihood programs and much higher than the performance of distance-based and parsimony approaches. Gerton Lunter , István Miklos (2005) [25], has developed a fully Bayesian Markov chain Monte Carlo method for co-estimating phylogeny and sequence alignment, under the Thorne-Kishino-Felsenstein model of substitution and single nucleotide insertion-deletion (indel) events Benjamin D.Redelings, Marc A .Suchard (2005) [10], has developed a novel model and algorithm for simultaneously estimating multiple alignments for biological sequences and the phylogenetic trees that relate the sequences. Bayesian approach is used in this algorithm.

Issac Elias, Jens Lagergren (2007) [22], has developed a model for estimating the mutational distance between two sequences. P.

A., Murty, M.N., Subramanian, D.K. and Subramanian (2003) [1], developed an algorithm for effective clustering and prototype selection for pattern classification. The main objective of this unsupervised learning technique is to find a natural grouping or meaningful partition by using a distance or similarity function for finding the subgroups/subclusters within each cluster which is used to find the superfamily, family and subfamily relationships of protein sequences. E.Ramaraj, M.Punithavalli (2005) [17], outlined a method for taxonomical clustering of species of the organisms based on the genetic profile using Principal Component Analysis and Self Organizing Neural Network. In this work, Principal Component Analysis is used for reducing the dimension of the data. Kmeans algorithm is used for clustering. The initial partitioning is randomly generated and it randomly initializes the centroids to some points in the region of the space. In each iteration step, a new set of centroids is generated using the existing set of centroids Recent field of molecular phylogenetics uses nucleotide sequences encoding genes or amino acid sequences encoding proteins as the basis for classification. Many forms of molecular phylogenetics are closely related and make extensive use of sequence alignment in constructing and refining phylogenetic trees, which are used to classify the evolutionary relationships between homologous genes represented in the genomes of divergent species[6] [7].

4. Birch Algorithm

BIRCH(Balanced Iterative Reducing and Clustering using Hierarchies): BIRCH is an integrated hierarchical clustering method that introduces two concepts , Branching Factor and Clustering Feature tree (CF tree); which are used to summarize cluster representations [11] .These structures help in achieving good speed and scalability in large databases.

BIRCH is based on the principle of agglomerative clustering. At any given stage there are smaller sub clusters and the decision at the current stage is to merge the sub clusters based on some criteria[3]. Instead of maintaining all the objects of a sub cluster, BIRCH maintains a set of Cluster Features of the sub cluster.

The criteria for merging two sub clusters is taken from the information provided by the set of CF's of the respective sub clusters. The Cluster Features of the different subclusters are maintained in a tree called CF Tree [11]. The nicety of the algorithm is that it requires only one pass to construct the CF Tree, and the subsequent stages works on this tree rather than the actual database

The concepts of Clustering Feature and CF Tree are at the core of BIRCH'S incremental clustering. A Clustering Feature is a triple summarizing the information that the user maintain about a cluster[36].Cluster Feature vector(CF) is defined as, CF = (n,ls,s) where n is the number of data objects in cluster C, ls is the linear sum of data objects and ss is the square sum of data objects.BIRCH algorithm works in two phases:

4.1 Methodology for Phylogeny Tree Construction

The nucleotide sequences of organisms from the same taxa are retrieved from National center for Biotechnology Information (NCBI). The architecture given below is used to cluster sequences to determining the evolutionary relationship among organisms.

4.1.1 Input

The nucleotide sequences for human taxa and bacteria taxa are retrieved from GenBank database and saved in text format. Sequences are then converted to character array to calculate the distance matrix using Kimura Two Parameter Model.

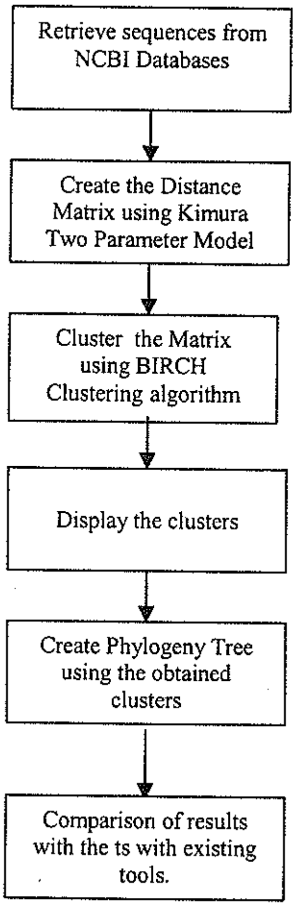

Retrieve sequences from NCBI Databases
Create the Distance Matrix using Kimura Two Parameter Model
Cluster the Matrix using BIRCH Clustering algorithm
Display the clusters
Create Phylogeny Tree using the obtained clusters
Comparison of results with the ts with existing tools.

Figure 2 : System Architecture

4.1.2 Construction of Distance Matrix

The Distance Matrix is created using Kimura Two Parameter Model. On implementation the distance matrix obtained using Kimura Two Parameter Model was found to be symmetric, the distance between same objects are zero and there are no negative values. The pair wise distance of the sequences calculated using Kimura Two Parameter Model. is given below in the equation1.

$$d = 1/2 * (\log(1/(1-(2 * p-q) + 1/4 * (\log(1/(1-(2 * q)) \quad -$$

Equation 1

Where d is the pairwise evolutionary distance, p is the transition count and q is the transversion count. The output is a symmetric matrix. The distance matrix generated is shown below in fig 2

```
x = [0      4.9633   4.0283   4.4747   3.8216   3.602    3.728
  4.9633   0        959.     2923     .2641    .272     4.1461
  4.0284   .1959    0        .282     .2766    .2857    3.6916
  4.4747   .2923    .2820    0        .2497    .2728    4.225
  3.8216   .2641    .2766    .2497    0        .1836    3.777
  3.6020   .2729    .2857    .2728    .1836    0        4.2541
  .3728    4.1461   3.6916   4.225    3.777    4.254    0   ]
```

4.1.3 Cluster Construction

The distance matrix constructed using Kimura Two Parameter-Model is given as the input for the BIRCH algorithm. A CF Tree construction is done in the first phase. For constructing the CF Tree, CF vectors, centroid, diameter and Euclidean distance of objects are calculated. The second phase forms clusters by applying a hierarchical clustering technique on the CF Tree. Different clusters are obtained for different threshold values. Correct tree is obtained by changing the threshold value. After a particular threshold value, the output remains unchanged.

Centroid : For a cluster C = {O1,O2...On} , the centroid is calculated using the formula,

$$O_{centroid} = \Sigma Oi / n, i = 1..n$$

Diameter : The diameter of a cluster C={O1,O2...On} is calculated using the formula,

$$D = \mathrm{Sqrt}\,(\Sigma(Oi - Oj)^2 / n(n-1)),$$
$$i = 1..n, j = 1..n$$

Euclidean Distance: For two clusters C1 and C2 with centroid O $_{centroid,1}$ and O $_{centroid,2}$, the euclidean distance is calculated using the formula,

$$D(C1,C2) = \mathrm{sqrt}(\ \Sigma\ (O^{i}_{centroid,1} - O^{i}_{centroid,2})^2)\ ,1 = 1..n$$

4.1.4 Phylogeny Representation

Phylogeny tree is constructed using the clusters obtained from BIRCH algorithm. The output clusters are converted to Newick format the phylogeny tree is generated using the matlab phylogeny tool.

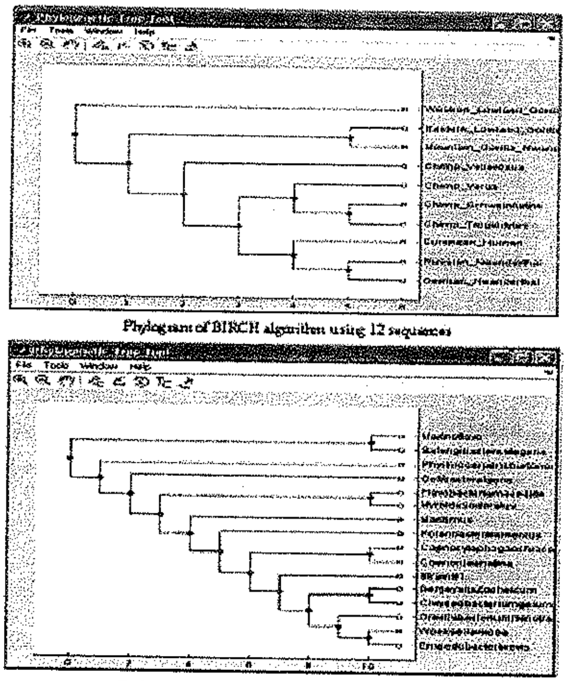
Newick format: ((Dog,(Sheep,Cow), (Mouse,rat)), (Chicken,Human));

5. Experimental Results

In order to evaluate the performance of the BIRCH Algorithm, it is tested with the nucleotide sequence dataset. The training dataset consists of sequences from human taxa and bacteria taxa. Nucleotide sequences in the genomes from mitochondria are used for evaluation of phylogeny. An example of the few sequences from the dataset taken for implementation are German_Neanderthal ,Russian_Neanderthal European_Human Mountain_Gorilla_Rwanda ,Mountain_Gorilla_Rwanda Chimp_Troglodytes Puti_Orangutan,Jari_Orangutan Western_Lowland_Gorilla ,Eastern_Lowland_Gorilla Chimp_Schweinfurthii Chimp_Vellerosus Chimp_Vellerosus

The performance of the algorithm is essentially based on the detection of correct phylogenetic tree of various species and organisms. The branching factor of the tree is fixed as 2. Threshold value is specified by the user. Clustering of objects is based on the comparison between diameter and threshold value. By changing the threshold value correct clusters are obtained. The threshold is iterated continuously to find the correct value. For threshold value of 0.5 the algorithm frames good clusters. The obtained clusters are saved in an array.

Accuracy of the algorithm is evaluated by comparing the results with proven datasets found in journals downloaded from the internet and using UPGMA Method. The comparison of the same dataset from the journal 'Neanderthal DNA sequences and the origin of modern humans' (1997) and from International Journal of Systematic and Evolutionary Microbiology (2001) using NJ Method. is done .

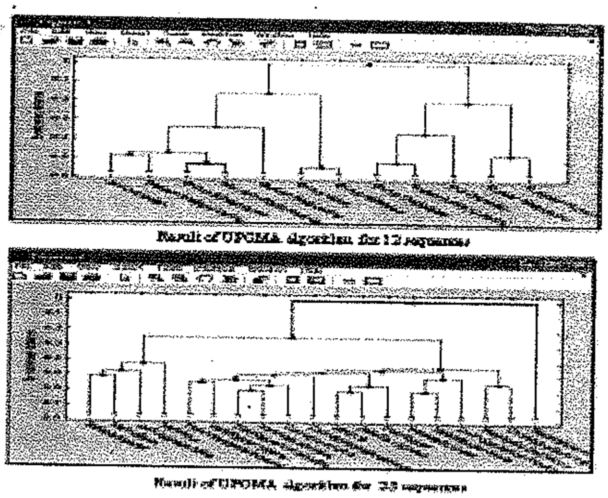

Phylogram of BIRCH algorithm using 12 sequences

Phylogram of BIRCH algorithm using 25 sequences

Figure 3 : Phylograms using BIRCH Algorithm

The following figures depict the results of using the datasets of human and bacteria taxonomy. The figure 4 depicts the screen shot of UPGMA method.

The table1 show the accuracy of the results clustered using birch algorithm with the other nearer methods. From the results it is found a BIRCH algorithm has found efficient clusters and is nearer to NJ method. The algorithm predicts good clusters when the dataset size is minimum. When the dataset size increases [above 50] the clusters found were not appropriate


Result of UPGMA algorithm for 12 sequences

Result of UPGMA algorithm for 25 sequences

Figure 4 : Result of UPGMA Algorithm

Table 1: Comparison of Clusters Obtained with Other Methods

No of sequences	BIRCH Algorithm	UPGMA Algorithm	PhyIip Tool	Journal (NJ)
5	2	2	2	2
7	4	4	4	4
10	6	6	7	7
12	7	8	5	8
15	5	9	9	9
19	13	11	9	14
25	18	18	11	20
70	19	23	21	28
100	17	30	28	34

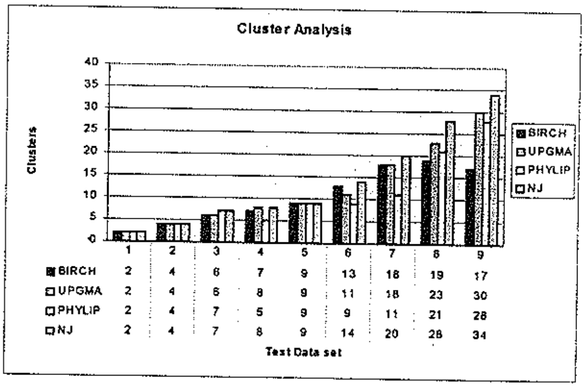


Figure 5 : Analysis of Clusters

6. Conclusion

Mining biological data is an emerging area of intersection between bioinformatics and data mining. A phylogeny or evolutionary tree, represents the evolutionary relationships among a set of organisms or groups of organisms, called taxa. Based on molecular sequences, phylogenetic trees can be built to reconstruct the evolutionary tree of species involved. Data mining is the process of discovering meaningful, new correlation patterns and trends by shifting through large amount of data store in repositories, using patterns recognition techniques as well as statistical and mathematical techniques. [3]. Clustering is the process that groups similar objects. Clustering sequential data has received a significant attention recently.

An idea is to implement Clustering DNA sequences using BIRCH algorithm to find the phylogenetic relationship among various species or organisms. The algorithm was and executed and tested with real datasets. The accuracy of the results were tested with already proven results from journal papers .In this work , it is observed that the proposed method yields good performance for small size of dataset, when the dataset size increases the clusters found does not reveal correct evolutionary relationship.The future work is to compare the same with different other data mining clustering algorithms.

Reference

1. A.Vijaya, N.Narashimhamurthy, *"An efficient incremental protein sequence clustering algorithm"*, BMC Bioinformatics, 2003.

2. Alberts Bruce , Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts and PeterLewis, *"Molecular Biology of the cell"*, ISBN, 2002.

3. Arndt von Haeseler, *"Introduction to Phylogeny"*, Center for Integrative Bioinformatics Vienna Gary D. Stormo, 2000.

4. Barker.D, *"Reconstructing Evolution with Parsimony and Simulated Annealing"*, University of Edinburgh, 1997.

5. Berg.j, Tymoczko.J and Stryer.L, *"Biochemistry"*, W.H. Freeman and Company,ISBN 2002.

6. *"Evolutionary Trees"*, Journal of Molecular Evolution,1996.

7. Saitu and Nei, *"Neighbor Joining Algorithm"*, ACM,1987.

8. Chaung Peng, *"DistanceBased Methods in Phylogenetic Tree construction"*, Department of Mathematics,Morehouse College, Atlanta, GA 303314, 2007.

9. Daniel A Polland, Allan M Moses, Venk N Iyer, Michael B Eisen, *"Detecting the limits of regulatory element conservation and divergence estimation using pairwise and multiplealignment"*, BMC, Bioinformatics, 2006.

10. Tian Zang, Raghu Ramakrishnan, *"BIRCH an efficient method for clustering large databases"*, ACM,1996

11. Stephene Guindon, Oliver Gascuel, *"A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood"*, System Biology ,October 2003

12. Felsenstein. J, *"Phylogenies from molecular sequences: inference and reliability"*, Annu..Rev Genet 22:521-565, 1988.

13. Felestein, *"PROTML:Maximum likelihood inference of protein phylogeny"*, Springer,2006.

14. Heiko A Schmidt, Korbiman Strimmer, Martin Vingron, *"Treepuzzle-maximum likelihood phylogenetic analysis using Quartets and parallel computing"*, BMC Bioinformatics,2002.

15. Huson. S. Nettles and T. Warnow, *"Disk-covering, a fast converging method for phylognetic tree reconstruction"*, Journal of Computational Biology, 6(3):369-386, 1999.

16. IritOrr, *"Introduction to phylogenetic Analysis"*, Weizmanx University,2004.

17. Isaac Elias and Jens Lagergren, *"Fast Neighbor Joining"*, Springer-VerlagBerlin Eidelberg,2005

18. Kimura.M, *"A simple method for evolutionary base substitution by comparative study of nucleotide sequences"*, J.Mol.Evol,1980.

19. Ling Qin,Yixin Chen,Yi Pan and Ling Chen, *"A novel approach to phylogenetic tree construction using stochastic optimization and clustering"*, BMC Bioinformatics,2006.

20. Lunter G and Miklos I, *"Bayesian coestimation of Phylogeny and sequence alignment"*, BMC Bioinformatics,2005.

21. Meng.S.W, *"Analysis of Phylogeny"*, 2004.

22. Nei.M and S. Kumar, *" Molecular Evolution and Phylogenetics"*, New York, Oxford University Press, 2000

23. Queen C, Wegman MN, Korn LJ. *" Improvements to a program for DNA analysis: a procedure to find homologies among many sequences"*, Nucleic Acids Res, 10:449-56, 1982.

24. Rob Guralnick, Allen Collins, *"Introduction to Phylogeny"*, UCMP.

Author's Biography

Ms V Bhuvaneswari received her Bachelors Degree (B.Sc.) in Computer technology from bharathiar university , India 1997 , Masters Degree (MCA) in Computer Applications from IGNOU ,India . and M.Phil in Computer Science in 2003 from Bharathiar University, India . She has qualified JRF , UGC-NET , for Lectureship in the year 2003 She is currently pursuing her doctoral research in School of Computer Science and Engineering at Bharathiar University in the area of Data mining . Her research interest include Bioinformatics, Soft computing and Databases. She is currently working as Lecturer in the School of Computer Science and Engineering, Bharathiar University, India. She has for her credit more than 15 publications in International/ National Conferences and journal publication.

Ms.Sindhu received her Master degree (MCA) , Avinashilingam University, India. She has completed her M.phil in School of Computer Science & Engineering , Bharathiar University India in 2008. Currently working as Lecturer in Hindhustan College of Arts and Science , Coimbatore. Her area of research interest includes Data Mining and Bioinformatics. She has presented a paper in National Conference.