Abstract

Phylogenetic analysis is means of estimating evolutionary or historical relationship among group of organisms based on their genetic closeness. Phylogenetic Tree are constructed using different tree methods based on distance and character for different living organisms to evaluate the relation between the different species. In the proposed method, a sample of DNA sequences of various species are taken and a phylogenetic tree is constructed using jukes cantor method for distance calculation. It is a quite good method as it is a 1-parameter model for generation of phylogenetic tree. By this method we can calculate distance using information of Amino Acid and DNA sequences.

Keywords: Bioinformatics, Phylogenetic tree, UPGMA, NJ.

1. Introduction

Bioinformatics is the science of collecting, managing, analyzing complex biological data from biological sequences. Bioinformatics uses computers for storage and retrieval of information related to biological molecules such as Deoxyribonucleic acid (DNA), Ribonucleic acid (RNA) and Proteins. The fundamental aspect of bioinformatics is phylogenetic analysis. Phylogenetics is the study of evolutionary history of some organisms. It is used to determine genetic relationship between different species. Phylogenetic trees are constructed from the molecular sequences of the different living organisms. These are actually needed to evaluate the relation between the different species and also the different time gaps from the actual origin. Unweighted Pair Group Method With Arithmetic Mean is a hierarchical clustering Technique used for the creation of phylogenetic tree. Hierarchical clustering is used to retrieve the results. America once a time consist of large variety of felines such as sabertooth tigers, scimitar tooth tigers, Cheetah, brown bears and cats. The analysis of relationship of these cats is possible only due to availability of DNA sequence now a days.

2. Phylogenetic Tree Construction

Phylogenetic tree represent graph with no cycles that show evolutionary relationship among different biological sequences. Various distance based methods include:

2.1 Distance Based Methods

Distance methods are used in phylogeny to construct a tree using a pair-wise distances. This method compress all the separable differences among pair of sequences into a distinct number. One of the main advantage of distance based method is that they are less computationally intensive Distance based Methods are UPGMA and Neighbour Joining (NJ).

a) UPGMA – Unweighted Pair Group Method With Arithmetic Mean choose closely related pair of sequences to build a tree. It is a clustering technique where each species is a cluster of its own and the process is repeated again and again until all the species are group into a single cluster. Distance between any two cluster A & B is taken to be average of objects x & y and mean distance is given by

\[ \frac{1}{|A| \cdot |B|} \sum_{x \in A} \sum_{y \in B} d(x,y) \]
(1) Assume that initially each species is a cluster on its own.

(2) Combine the two clusters that are closest and recalculates distance of the join pair by taking the average.

(3) Repeat the process again and again until all species are grouped in a single cluster.

b) NEIGHBOUR JOINING (NJ) - NJ Methods initially sum individual distance to calculate the difference of an organism from all other organisms and then based on this sum corrected distance method is calculated. It generally gives better results than UPGMA.

3. RELATED WORK

Pushpinder Kaur et. al. (2016) presented that Phylogenetic analysis is means of estimating evolutionary or historical relationship among group of organisms. Phylogenetic Tree are constructed for different living organisms. Phylogenetic Tree can be build from different tree building methods so as to retrieve an optimal result. Data Mining is an tool to determine the hidden data and patterns from large set of data. Different Data Mining Methods have been used to study molecular phylogeny in order to determine the relationship between different organisms. Phylogenetic Tree are constructed for different living organisms. Data Mining technique is used for extraction of data.[1]

Rajbir Singh et. al. (2015) presented an approach that Data mining is an essential tool to discover knowledge from a large data set. The biological data is available in different formats. Data mining and machine learning techniques have been used to scale to the size of the problems. Data mining methods have been used to study molecular phylogeny to discover the degree of relationship within a group of organisms. Phylogenetic trees are constructed from the molecular sequences of the different living organisms. These are needed to evaluate the relationship between the different species. [2]

Shaminder Kaur et. al. (2015) presented that Bioinformatics is a branch of biology science and information technology in the arena of research and development. Phylogeny is the study of evolutionary history amongs organisms based on genetic information codes. The genetic relationships between species can be represented using phylogenetic trees. To construct a phylogenetic tree is a very challenging problem. The main purpose of phylogenetic tree is to determine the structure of unknown sequence and to predict the genetic difference between different species or organisms. There are different methods for phylogenetic tree construction from distance or character data. [3]

Joseph L. Staton et. al. (2015) suggested that Interpretation of phylogenetic trees is essential in understanding the relationships between organisms, their traits or characteristics, and even their genomic and developmental biology. As trees appear more often in basic texts, many students, and even their teachers, clearly understand little of how they are constructed and even less about what can be inferred from them about the history of the representatives analyzed. [4]

J.Jayapiya et. al. (2015) illustrates an algorithm based on one of the swarm intelligence techniques for constructing the Phylogenetic tree, which is used to study the evolutionary relationship between different species. [5]

Anand Patwardhan et. al. (2014) described that the Usages of molecular markers in the phylogenetic studies of various organisms have become increasingly important in recent times. It gives an overview of different molecular markers employed by researchers for the purpose of phylogenetic studies. In molecular phylogeny, the relationship between different organism or genes are studied by comparing homologues of DNA or protein sequences. [6]

Yang Ruan et. al. (2014) discussed that the Phylogenetic analysis is commonly used to analyze genetic sequence data while clustering techniques commonly are used to analyze sequence data. In comparison with traditional tree display methods, the correlations between the tree and the clustering can be detected. [7]

Mike Hanrahan et. al. (2014) described that Bayesian inference has been widely useful for phylogenetic analysis in recent years. High performance computing platform have been built using Open-MPI with free-cost open source Ubuntu Linux operating systems, and then apply the Bayesian inference model to construct a phylogenetic tree of various biological species based on similarities and differences in their physical and genetic characteristics. Phylogenies has been able to reconstruct the evolutionary history of the mutations. It is possible because of the better access to DNA sequences available now by utilizing the resources and massive computing power of super computers. [8]
Amardeep Singh et. al. (2013) discussed that bioinformatics represents the symbolic relationship between computational and biological sciences. "The main use of Bioinformatics therefore lies in harnessing information pertaining to these genetic functions in order to understand how human beings and other living organisms operate. Phylogenetics is the study of evolutionary relationships among groups of organisms, which are discovered through molecular sequence of data. It represents the evolutionary divergences by finite directed graphs, or directed trees, known as Phylogeny.[9]

Kenji Sorimachi et. al. (2013) evaluated the appropriateness of phylogenetic trees in biological evolution, we expanded a pre-existing baseline data set of randomly selected organisms by incorporating a collection of intentionally chosen organisms. Using two different clustering algorithms Ward’s method and neighbour-joining method to constructed phylogenetic trees.[10]

Gayatri Mahapatro et.al (2012) suggested that clustering algorithm gives the most effective clusters for biological data are k-means ,k-mediod and density-based algorithm. Phylogenetic tree for individual clusters are formed nad finally joined to create phylogenetic tree.[11]

Jasmine Kaur et. al. (2012) discussed that phylogeny is the concept of phylogenetic trees which is a graphical representation of the evolutionary relationships among three or more genes or organisms. The main existing methods for reconstructing phylogenetic trees are based on maximum likelihood, Bayesian inference, maximum parsimony or distance. 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.[12]

4. METHODOLOGY

![Flowchart diagram]

- **Start**
- Collect data from genomic bank
- Enter the sequence for different datasets
- Are sequences nucleotide?
  - Yes
  - Are seq. amino acid?
    - Yes
    - NO
    - YES
  - Distribution of nucleotide bases for cytochrome B for different data sets
  - To compute nucleotide density
    - Draw Nucleotide density A,G,T,C,G
  - Compose datasets and translated protein fragments
    - Calculate the histogram of Aminic acid count for cytochrome B of panther tiger, panther leo, wild cat.
STEP1- Data from genbank can be collected and enter it for different sequences.
STEP2- Find whether the seq are nucleotide and amino acids.
STEP3- Distribute Nucleotide base for different data sets.
STEP4- Draw Nucleotide Density for A,T,C,G and A-T,G-C.
STEP5- Calculate histogram of amino acids for different sequences

5. RESULTS AND DISCUSSION

Biological sequence for different species that include the accession no, gene index no is taken as input.

<table>
<thead>
<tr>
<th>ACCESSION NO.</th>
<th>ORGANISM</th>
<th>GENE INDEX NO.</th>
</tr>
</thead>
<tbody>
<tr>
<td>JQ937109</td>
<td>Homotherium DNA</td>
<td>&gt;gi</td>
</tr>
<tr>
<td>EF211571</td>
<td>Cheetah DNA</td>
<td>&gt;gi</td>
</tr>
<tr>
<td>NZ_KB291588</td>
<td>American Cat DNA</td>
<td>&gt;gi</td>
</tr>
<tr>
<td>AY525057</td>
<td>Chinese Desert cat DNA</td>
<td>&gt;gi</td>
</tr>
<tr>
<td>NM_001009330</td>
<td>Felis silvertris Cat A DNA</td>
<td>&gt;gi</td>
</tr>
<tr>
<td>KC538825</td>
<td>Wild Cat DNA</td>
<td>&gt;gi</td>
</tr>
<tr>
<td>HQ168692</td>
<td>Sabertooth DNA</td>
<td>&gt;gi</td>
</tr>
<tr>
<td>KC405559</td>
<td>Panthera Tigris DNA</td>
<td>&gt;gi</td>
</tr>
<tr>
<td>JX274401</td>
<td>Leopard Whipray DNA</td>
<td>&gt;gi</td>
</tr>
<tr>
<td>JX146847</td>
<td>Panthera Leo DNA</td>
<td>&gt;gi</td>
</tr>
<tr>
<td>KC414578</td>
<td>Wolf DNA</td>
<td>&gt;gi</td>
</tr>
<tr>
<td>NM_001003052</td>
<td>Dog DNA</td>
<td>&gt;gi</td>
</tr>
<tr>
<td>JX971706</td>
<td>American Black Bear DNA</td>
<td>&gt;gi</td>
</tr>
<tr>
<td>NC_011112</td>
<td>Cave Bear DNA</td>
<td>&gt;gi</td>
</tr>
<tr>
<td>JQ823257</td>
<td>Brown Bear DNA</td>
<td>&gt;gi</td>
</tr>
</tbody>
</table>
6. EXPERIMENTAL ANALYSIS

**Fig 1:** Histogram of nucleotide sequence for Felis silvetris

**Fig 2:** Histogram of nucleotide sequence of Homotherium

**Fig 3:** Histogram of nucleotide sequence of American Cat

**Fig 4:** Nucleotide Density of Felis silvetris

**Fig 5:** Nucleotide Density of Panthera Tigris

**Fig 6:** Histogram of Amino Acid for Puma
7. CONCLUSION
Phylogenetic analysis usually describes the genetic relationship between different species and their evolutionary history. In this work, in order to construct the phylogenetic tree, firstly, we need to know some statistical properties about species in order to determine the distribution of nucleotide bases for different sequences and to compute the histogram of amino acid content for sequences. Here, the GC content varies between all of the species, with some areas of similar base usage and others that are widely different. The same happens for amino acid content, with variations in the number of each amino acid used in each species. Tree constructed using different sequences shows different results.
References


AUTHOR

Pushpinder Kaur has received B-Tech degree from Punjab Technical University, Jalandhar in 2014 and currently is pursuing Mtech From PTU, Jalandhar. Her topic of interest is Bioinformatics. She has attend many conferences.