Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 7, July 2016, pg.239 – 244

A Detailed Survey on Approaches of Phylogenetic Analysis

Lavanya K

PG Student, Dept. of Computer Science and Engineering, Acharya Institute of Technology, Bengaluru, Karnataka, INDIA

lavanya.kakimallaiah@gmail.com

V Nagaveni

Assistant Professor, Dept. of Computer Science and Engineering, Acharya Institute of Technology, Bengaluru, Karnataka, INDI

nagaveni@acharya.ac.in

Abstract: All organisms have evolved from a common ancestor. The distance between these species is measured using phylogenetic analysis. It enables us to extract evolutionary relationship from sequence analysis. These relationships are depicted on phylogenetic trees. This article provides a detailed survey on different sequential approaches of sequential alignment, clustering and complete details of how a mapreduce technology improves the performance of phylogenetic analysis. A comprehensive comparison of these methods is presented in this paper.

Keywords: Phylogenetic analysis, mapreduce, approaches of phylogenetic analysis, sequence alignment, clustering.

I. INTRODUCTION

Phylogenetic analysis [1] is depicting relationship among group of organisms. It aims at uncovering the evolutionary relationships between different species in order to obtain an understanding of the evolution of life on Earth. Most of the phylogenetic analysis techniques produce phylogenetic trees. These phylogenetic trees are well suited to represent evolutionary histories in which the main events are speciation which is shown in Figure 1. They represent root as origin of evolution, leaves as current organisms (species or genomic sequence), branches as relationship between organisms, branch length as evolutionary time (in cladogram, it doesn't represent time). Phylogenetics is sometimes called cladistics because the word ‘‘clade,’’ a set of descendants from a single ancestor, is derived from the Greek word for branch. The basic tenet behind cladistics is that members of a group or clade share a common evolutionary history and are more related to each other than to members of another group.

A gene tree [1] consists of clade, taxon and node. A clade includes recent common ancestor of all members and its descendants, taxon includes a group of species but not clade and node is bifurcation point of branch. There are two types of tree in phylogenetic analysis. First one is the rooted tree i.e a common ancestor will be present to all nodes in the tree, a root node exist in the tree, the path from leaves to root node is the evolutionary time. Second one is unrooted tree i.e they are all related by descendants, no existence of root node and the path does not specify the evolutionary time. Rooted tree has cladograms, where Branch length has no meaning. In phylogram, whose branch length represents evolutionary change and finally the Ultrametric, whose branch length represent time and the length from the root to the leaves are the same. Phylogenetic analysis takes place in three steps, first one is gathering of sequences; second one is sequence alignment process and third is clustering to produce phylogenetic trees. A variety of systems and applications have been developed to infer phylogenetic trees.

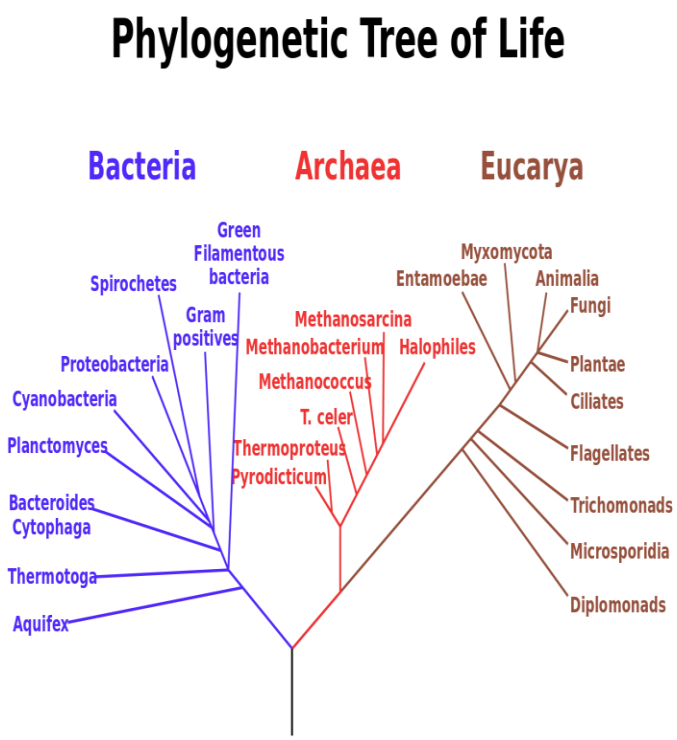

Figure 1: The Phylogenetic tree of bacteria.

II. SEQUENCE ALIGNMENT

In bioinformatics, a sequence alignment [1,2] is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. Aligned sequences of nucleotide or amino acid residues are typically represented as rows within a matrix. Gaps are inserted between the residues so that identical or similar characters are aligned in successive columns. There are two categories of sequence alignment, local and global alignment.

Global alignment [3] is where two sequences to be aligned are assumed to be generally similar over their entire length. Alignment is carried out from beginning to the end of both the sequences to find the best possible alignment. One of the popular algorithms is needle-wunsch algorithm [4], it was proposed in 1969 by needle for global succession. It must possess all the letters from initial and the one you want to compare. Local method of alignment does not assume that the two sequences have the similarity over the entire length. It only finds local regions with the highest level of similarity between two sequences and align these regions without regard for the alignment of rest of the regions. The one which obtain enormous popularity was the algorithm produced by smith-waterman [5]. Here the two sequences are aligned locally over varying length segments, with no punishment for the unaligned bits of the succession.

i. Pair wise alignment and multiple sequence alignment

Pair wise Alignment is utilized to recognize regions of comparability that may demonstrate functional, basic or transformative connections between two organic sequences. Goal of pair wise comparison is to find conserved regions (if any) between two sequences, extrapolate information about the sequence using the known characteristics of the other sequence.

A multiple sequence alignment (MSA) is a sequence alignment of three or more biological sequences, generally protein, DNA, or RNA. In many cases, the input set of query sequences are assumed to have an evolutionary relationship by which they share a lineage and are descended from a common ancestor. From the resulting MSA, sequence homology can be inferred and phylogenetic analysis can be conducted to assess the sequences' shared evolutionary origins. Visual depictions of the alignment illustrate mutation events such as point mutations (single amino acid or nucleotide changes) that appear as differing characters in a single alignment column, and insertion or deletion mutations (indels or gaps) that appear as hyphens in one or more of the sequences in the alignment. Multiple sequence alignment is often used to assess sequence conservation of protein domains, tertiary and secondary structures, and even individual amino acids or nucleotides. For example BLAST and FASTA.

III. TREE-BUILDING METHODS

Tree building methods can be sorted into distance-based and character-based methods. Distance-based methods use the amount of dissimilarity (the distance) between two aligned sequences to derive trees. A distance method would reconstruct the true tree if all genetic divergence events were accurately recorded in the sequence. Some of the distance methods are Unweighted Pair Group Method with Arithmetic Mean (UPGMA). UPGMA [6] is a clustering algorithm—it joins tree branches based on the criterion of greatest similarity among pairs and averages of joined pairs. It is not strictly an evolutionary distance method. UPGMA is expected to generate an accurate topology with true branch lengths only when the divergence is according to a molecular clock or approximately equal to raw sequence dissimilarity.

The character-based methods have little in common with each other, besides the use of the character data at all steps in the analysis. This allows the assessment of the reliability of each base position in an alignment on the basis of all other base positions. Some of the methods are Maximum parsimony(MP) and Maximum likelihood (ML). Maximum parsimony is an optimization criterion that adheres to the principle that the best explanation of the data is the simplest, which in turn is the one requiring the fewest adhoc assumptions. In practice, ML is derived for each base position in an alignment. The likelihood is calculated in terms of the probability that the pattern of variation at a site would be produced by a particular substitution process, given a particular tree and the overall observed base frequencies. The substitution model should be optimized to fit the observed data.

The above observations from the different solutions such as Needle-Wunsch algorithm (NWA), Smith-waterman algorithm (SWA), MP and ML that are associated for analyzing phylogenetics by using different alignment algorithm which are not accurate and do not consider the dynamicity of the algorithm for improving the performance of phylogenetic analysis. These solutions are sequential and they are not executed parallel. The time taken is usually high. As data grows it is difficult to process the data, so we fasten the process of phylogenetic analysis by applying mapreduce programming model at different stages; the different techniques of hadoop are given below.

IV. SURVEY ON HADOOP BASED PHYLOGENETIC ANALYSIS

With an increase in the size of the sequence there is a need of tool that entitle efficient and fasten thee sequence alignment process. One such idea for processing these enormous quantities of data is the usage of hadoop map/reduce programming model. The computation intensive algorithms required for phylogenetic analysis can be fitted in the map/reduce model and a time efficient approach can be carved out. Big Data [7] refers to data which is too vast to be processed by traditional database systems and techniques and hence requires alternative processing methods. Such methods are needed to extract valuable information and data patterns that lie hidden in the massive datasets that constitute Big Data. A popular model and framework for handling big data is map reduce.

MapReduce is a data-parallel programming model pioneered by Google for clusters of machines. It is implemented for processing and generating large datasets to solve a variety of real-world problems and tasks. The computation takes a set of input key/value pairs and produces a set of output key/value pairs. The user of the MapReduce library expresses the computation in the form of two functions: Map () and Reduce (). Map () takes an input pair and produces a set of intermediate key/value pairs. The MapReduce library groups together all intermediate values associated with the same intermediate key K and passes them to the Reduce() function. The Reduce() function accepts an intermediate key K and a set of values for that key. It merges these values to form a (usually) smaller set of values represented in Figure 2. The intermediate values are supplied to the user's Reduce() function via an iterator. This helps us to handle lists of values that are too large to fit in memory.

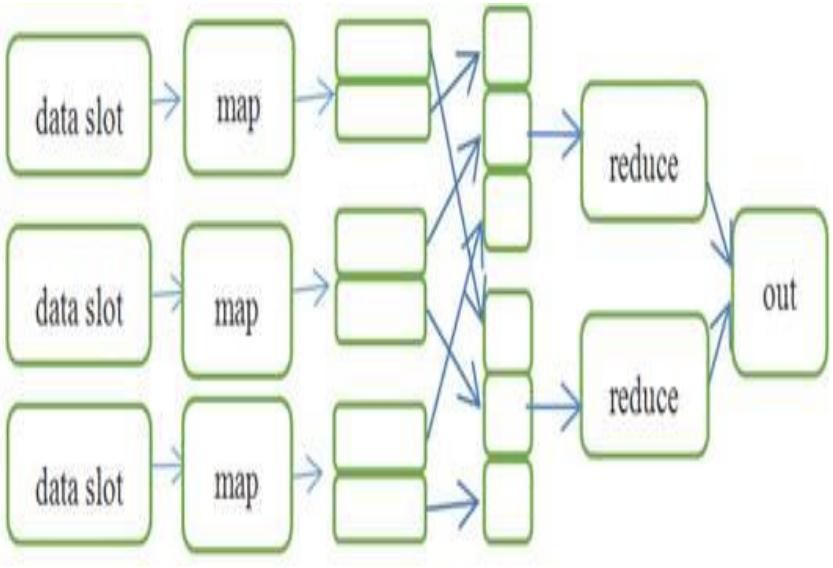

Figure 2: Mapreduce activity diagram for sequence alignment process.

[8] presented a report on application of MapReduce, using its open source implementation Hadoop, to two relevant algorithms: BLAST (Basic Local Alignment and Search Tool) and GSEA (Gene Set Enrichment Analysis). Biocloud [9] which provides scalable, high availability, robust computing service and a combination of hadoop framework. A detail study on overall performance is made. It has many overheads in implementing cloud and even the computation time it offered is very large. [10] Presented a report on phylogeny using map reduce programming model which uses NWA(Needleman-Wunsch algorithm) and UPGMA(Unweighted Pair Group Method with arithmetic mean) along with the map reduce framework to improve the performance and accuracy. [11] Presented a report on complete composition vector (CCV) employed with hadoop map reduce programming model, here sequences are converted into vectors. Using cosine similarity between the vectors, we can calculate the distance which is in turn is represented in the form of distance matrix. Using UPGMA clustering algorithm a dendogram is produced. Alignment (MSA) using Hadoop framework was presented in [12]. This methodology uses dynamic programming. This achieves parallelism in conducting MSA by using multiple levels of data processing. The proposed method of MSA improves on the computation time and also maintains the accuracy. [13] describes a methodology called GATK a genome analysis toolkit which is used for analyzing genome with a programming paradigm called map reduce programming model. It consist of wealth data access patterns and it also defines hadoop, which provides functionality of map reduce for HDFS (hadoop distribute file system) and illustrates an example word countless application called hadoop bam and explains the execution of example with GATK tool.

V. RESULTS

The experiment is conducted in one node and two node clusters in i3, i5, i7 processor. The process is executed using hadoop based approaches like CCV and then compared with sequential approaches and time is recorded. But the time taken by sequential is more when compared to parallel hadoop based approaches like CCV and even it considers the dynamicity of the algorithm, by improving the overall throughput and accuracy of the system. The results of the experiment in a one node and two nodes in i3 and i5 processor is given below

Table 1: Execution in a two node cluster of sequence in an i3 processor

Sequence length/No of sequence	25	50	75	100
250	8	9.09	10.07	11.09
500	9.45	10.51	11.56	13
750	10.3	11.49	12.54	13.58
1000	12	13.08	14.15	15.22

Table 2: Execution in a four node cluster of sequence in an i3 processor

Sequence length/No of sequence	25	50	75	100
250	14	15.1	15.52	16.30
500	15.28	16.42	17.58	18.25
750	16.17	17.23	18.35	19.42
1000	17.30	18.33	19.47	20.01

Table 3: Execution in a two node cluster of sequence in an i5 processor

Sequence length/No of sequence	25	50	75	100
250	7.15	8.21	9.10	10.27
500	8.19	9.31	10.44	11.25
750	9.4	10.28	11.23	12.31
1000	10.4	11.46	12.25	13.3s

Table 4: Execution in a four node cluster of sequence in an i5 processor

Sequence length/No of sequence	25	50	75	100
250	13	14.25	15.10	16.20
500	14.08	15.15	16.30	17.41
750	15.28	16.39	17.18	18.29
1000	16.25	17.35	18.42	19.50

The table 1-4 depicts the time recorded for different length sequence by varying the number of sequences for two nodes and four node clusters in i3 and i5 processor, the time taken by two node clusters in table 3 is less when compared to time recorded in table 1. Gradually time is reduced with increase in number of sequences, similarly the time taken by the sequences in four node cluster of an i3 processor (table 2) is more and time taken is decreased in table 4.

VI. CONCLUSTION

Phylogenetic analysis provides a transformative relationship between species. It is used to calculate distance among organism. This paper provides numerous methods for phylogenetics and various approaches for inference of huge phylogenetic trees. As the genomic data is increasing exponentially, it requires a system for handling such huge data. The results shows that map reduce programming model using Hadoop applied on different approach such as CCV is best for processing large data in less time, especially it captures the protein families in a better way.

REFERENCES

[1] Baxevanis, Andreas D., "Bioinformatics: a practical guide to the analysis of genes and proteins". Chapter 14 Vol. 43. John Wiley & Sons, 2004.

[2] Muhammad Sardaraz1, Muhammad Tahir, Ataul Aziz Ikram, Hassan Bajwa., "Applications and Algorithms for Inference of Huge Phylogenetic Trees: a Review". Department of Computing and Technology, Iqra University, Islamabad, Pakistan, American Journal of Bioinformatics Research 2012, 2(1): 21-26,DOI: 10.5923/j.bioinformatics.20120201.04

[3] Sonali Vijan .,"Biological Sequence Alignment for Bioinformatics Applications Using MATLAB". Student, Electronics Engineering, NITTTR, Chandigar.

[4] B NEEDLEMAN AND WUNCH department of biochemistry, northwestern university and nuclear medineservices,V, "a

general method applicable to the search of similarities in the amino acid sequence of two proteins" SAUL, A Research hospital chicago, Ill .60611, U.S.A 21 july 1969.
[5] Temple F Smith and Michael S Waterman, "Comparison of Biosequences" Advances in applied Mathematics, 2(4):482-489, 1981.
[6] Sukhpreet Kaur, Harwinder Singh Sohal, Rajbir Singh Cheema Dept. of CSE, LLRIET Moga, Punjab, "Implementing UPGMA and NJ Method For Phylogenetic Tree Construction Using Hierarchical Clustering" India IJCST, Vol. 4, Issue 2, April - June 2013.ISSN : 0976-8491
[7] Anurag Sarkar., "MapReduce: A Comprehensive Study on Applications, Scope and Challenges", Department of Computer Science, International Journal of Advance Research in Computer Science and Management Studies, Volume 3, Issue 7, July 2015, ISSN: 2321-7782 (Online).
[8] Gaggero, Massimo, et al. "Parallelizing bioinformatics applications with MapReduce." Cloud Computing and Its Applications (2008): 22-23.
[9] Hongfeng Zhang, Vincent Y.Liu Zhao Macau university of sciencea and technology,Taipa, "Biocloud: A systematic review and classification", Macau, may 8 2015.
[10] Siddesh G M, K G Srinivasa*, Ishank Mishra, Abhinav Anurag, Eklavya Uppal"Phylogenetic analysis using mapreduce programming model". IEEEDOI 2015:+350.
[11] Marc E Colosimo, Matthew W Peterson, Scott Mardis and Lynette Hirschman. "Nephele: genotyping via complete composition vectors and MapReduce" Colosimo et al. Source Code for Biology and Medicine 2011.
[12] Sadasivam, G. Sudha, and G. Baktavatchalam. "A novel approach to multiple sequence alignment using hadoop data grids." Proceedings of the 2010 Workshop on Massive Data Analytics on the Cloud. ACM, 2010.
[13] Maharjan, Merina. "Genome Analysis with MapReduce". June 15 (2011): 3-4.

