RESEARCH ARTICLE OPEN ACCESS

Hadoop Mapreduce Based Distributed Phylogenetic Analysis

Lavanya K [1], V Nagaveni [2]
PG Student [1], Assistant Professor [2]
Department of Computer Science and Engineering
Acharya Institute of Technology, Bengaluru
Karnataka - India

ABSTRACT
Phylogenetic analysis is most important in scientific research of evolution of life, it is a measure of footprints between organisms and analysis requires multiple sequence alignment as input. Even though algorithms such as Needle-Wunsch Algorithm (NWA) and Smith-Waterman Algorithm (SWA) produce accurate alignments but they are not applicable to larger length genome sequence that increases computational complexity. The proposed approach uses complete composition vector (CCV) to represent each sequence as vector derived from K-mere by passing for multiple sequence alignment and Unweighted Pair Group Method with Arithmetic mean (UPGMA) which produces tree. The aim is to improve and optimize the performance of phylogenetic analysis for large sequence data by map reduce programming model.
***Keywords*:** Sequence alignments, Phylogenetic analysis, CCV, Hadoop map reduce, UPGMA, NWA.

I. INTRODUCTION

Phylogenetic analysis [1] is estimating relationship among genetically related group of organisms. It is a means of depicting evolutionary history between species and supposition regarding natural selection history of taxonomic groups called as Phylogeny. This phylogeny is usually represented as tree like diagrams. This branching tree diagram represents evolutionary relationships between species. It is also called as cladistics [2]. A phylogenetic analysis takes place in two steps: First one is alignment of sequences and second one is clustering to produce gene trees.

Sequence alignments are mostly done through utilities such as NWA, CLUSTAL W and SWA. This dynamic programming solution cannot be used for large length sequence and they are computationally intensive. One such idea that is proposed to handle enormous amount of data is Composition vector (CV) which has been, used to describe protein sequence as vectors, it uses sliding window to represent each sequence as a vector, where each element calculation is based on expected and actual frequency of k-mere. Distance between any two sequences can be calculated using vector representation with distance matrix. The CV method produces trees matched existing taxonomies and later it will be expanded to Complete Composition Vector (CCV) [3], which uses sliding window over wide range of length sequence and distance calculated between sequence remains same, so there is no dependency. CCV uses k-mers Generator: this module is used to generate k-mers of the genome sequences of different size. The second step of phylogenetic analysis is Clustering of genetic sequence to produce trees, which require multiple sequence alignment as an input. Distances between sequences are required to construct distance matrix which in turn depends on the distance based method to build phylogenetic tree. Distance method such as UPGMA is used, which uses hierarchical clustering.

Apache Map reduce is an open source framework which supports scalability and fault tolerance. It is a programming framework which is used to process large data sets in parallel. It consists of two phases such as map phase that is used to perform sorting as well as filtering and reduce phase that is used to combine all values from a previous map stage.

II. RELATED WORK

Next generation sequencing has made the length of sequences to rise up to trillions and billions of data, raising new challenges for interpretation and processing large genomes, where it makes impossible for processing by traditional sequence methods. Some of the traditional methods using cloud are BioCloud [4], which provides scalable, high availability, robust computing service and a combination of hadoop framework. A detailed study on overall performance is made that has many overheads in implementing cloud and even the computation time it offered is very large. Two

bioinformatics tools such as BLAST [5] (basic local alignment search tool) and another one GSEA(Gene set enrichment analysis) are used to parallelize bioinformatics applications for finding the disease that is inherited from our ancestors, which provides promising results and have a large range of bioinformatics applications maintaining good efficiency and maintenance is easy. [6] 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.[7] methodology called GATK(Genome analysis toolkit) is used for analyzing genome with a programming paradigm called map reduce programming model. It consist of wealthy data access patterns and it also defines Hadoop, which provides functionality of map reduce for HDFS (Hadoop Distributed File System) and illustrate an example word count class application called hadoop bam and explains the execution of example with GATK tool. [8] is MSA (Multiple sequence alignment) which is a Dynamic Programming for parallelizing the sequence alignment process by processing in multiple levels by improving on the computation time and maintaining accuracy.

The above observations from the different solutions that are associated to analyzing phylogenetics by using different alignment algorithm are not accurate and do not consider the dynamicity of the algorithm for improving the performance of phylogenetic analysis that does not focus on either computation or data parallelism and they are not applicable to large range of genome sequence.

III. PROPOSED ALGORITHM

The proposed solution is used to produce a gene tree (phylogenetic tree) which is accurate for a large range of sequence and it is best time efficient approach among all the other solutions because it uses Hadoop Map reduce [9], which executes program faster than traditional tools. It will be using CCV [10] to make sequence alignments, describing sequence as vectors coupled with map reduce programming prototype. Further it will be using neighbor-joining methods such as UPGMA to produce PhyloXml tree. The proposed solution considers both data and the computation parallelism by improving performance, throughput and accuracy. The parallel approach [11] is done by dividing the sequence into set of blocks and processing or making it to run parallel. High level design of proposed solution is as shown in Figure 1.

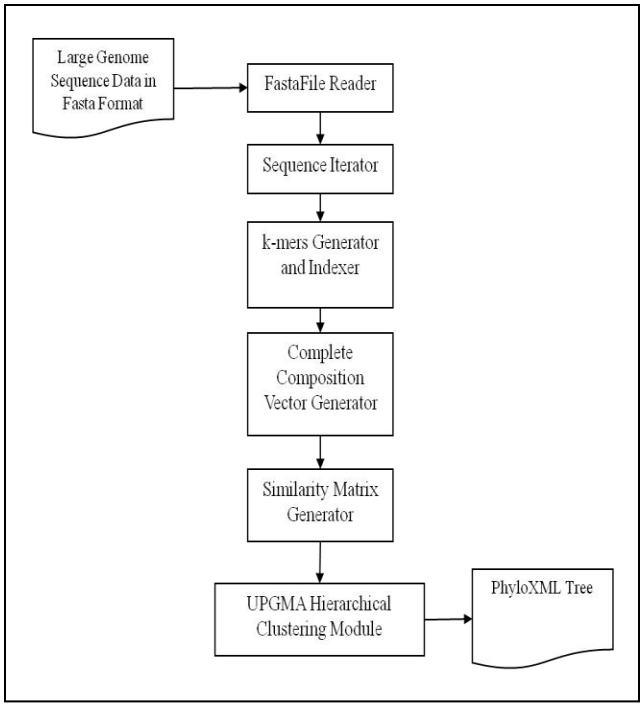

Figure 1: The High level design of proposed approach.

Here Apache hadoop map reduce applies only to complete composition vector (CCV) not for UPGMA method.
There are 3 map reduce stages, first map reduce stage reads the large set of sequence and create a custom record for each of the sequence which is identified by logical delimiter (>), next second map reduce stage, reads the records from previous map stage and forms a two identical sets of records in the same set of record what it has read from the previous stage, and make a Cartesian pair of such records, for each Cartesian pair of records it creates a custom record. Finally the third map reduce stage will align the two sequence in Cartesian record using complete composition vector and finally aligned sequence scores along with pair of sequence tags fed as input to hierarchical clustering of UPGMA[13] to produce phylogram.

The large protein sequences of plants or animal in a fasta format is represented in Figure 2, where each sequence is identified by description tag in fasta reader, as described in Figure 1,in sequence iterator the sequence is converted to ASCII format and it is provided as input to K-mere generator.

Figure 2: Example of multiple sequence data of insect in a fasta format.

Consider a sequence of length L, where K-mers [12] is a small string of length K in the sequence L, the algorithm can generate up to L-K+1 K-mers. Here sequence is converted into vectors and they are aligned accordingly. Then this K-mers is fed as input to Composition vector generators, where each sequence is considered as whole vector which in turn composed of aligned vectors of length 4,5,6,7 up to L-K+1. The length of K should be greater than two. Consider for instance the first vector in the sequence vector A is matched with the first vector of sequence vector B and then similarity between the sequence vectors is calculated using cosine similarity.

$$\text{Cosine} = \frac{\text{Dot Product}[A][B]}{\text{length}[A] * \text{length}[B]}$$

The distance matrix of this sequence is constructed based on the cosine similarity and fed as input to distance method called UPGMA [13] to produce Phylogram. UPGMA algorithm first finds the minimum distance and merges them, finds the arithmetic mean to all other sequence, this process is repeated and finally using Tree J 1.1 software it is possible to view phylogram as shown in Figure 3.

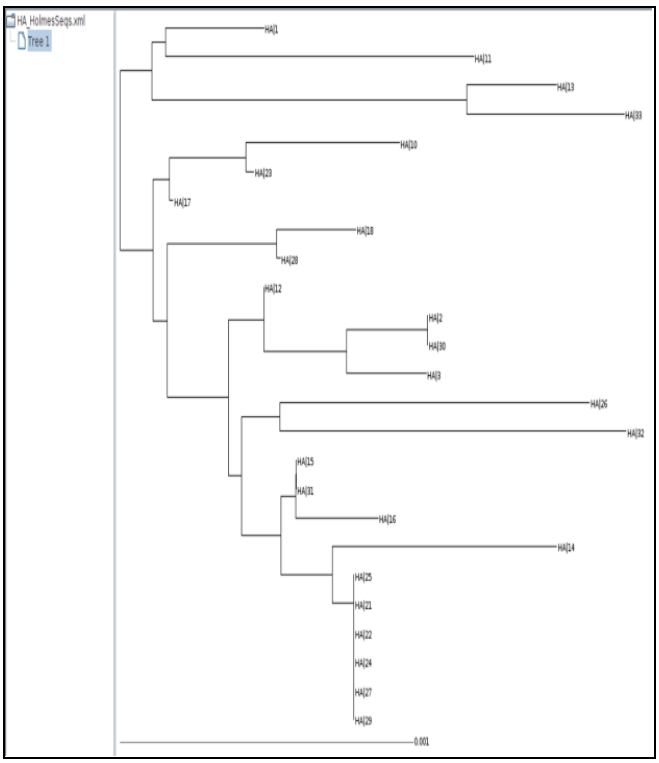

Figure 3: Phylogram representing UPGMA results.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The program is first executed for protein sequences of insect HA [14] in a Pseudo mode of one node and two node and execution time is calculated..

A. Execution time for varying sequence

A five sequence of length 500 and 1000 were given to sequence alignment. The first sequence comprises of 25 sequences, similarly second sequence compromises of 50 sequences, third of 75 sequences and 100 sequences, these sequences are run in a pseudo mode of one node and two node in a i3 and i5 processors, run time is noted for these sequences. The time taken by two nodes is less when compared to one node which is depicted in this graph as shown in figure 4 and 5. The time taken by this comparatively low when compared to the traditional prototyping model. So by using the hadoop mapreduce the time is decreased. Here time is measured in terms of minutes. The execution time for 500 length sequence is depicted in Figure 4.

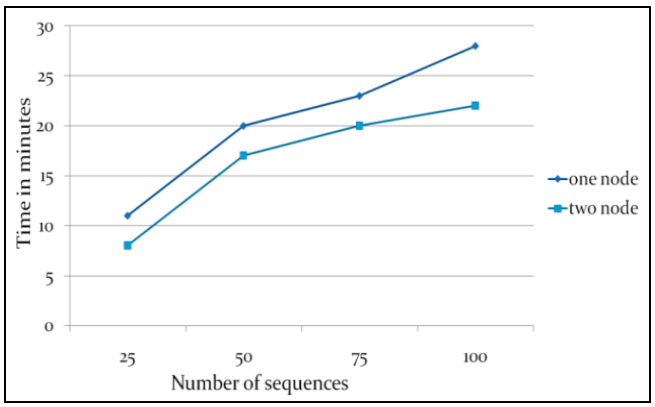

Figure 4: Execution time for 500 length sequence.

This execution time is calculated for length 1000 sequences which is depicted in Figure 5.

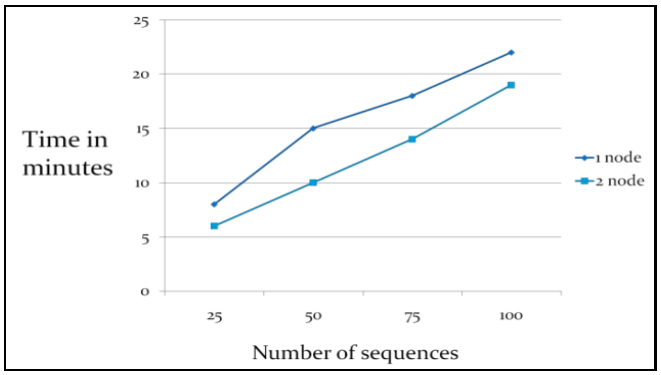

Figure 5: Execution time for 1000 length sequence.

B. Execution time for sequential and parallel phylogenetic analysis

The time taken by sequential traditional programming is more and the time taken by parallel approach is less since it employs hadoop map reduce programming model. The bar graph of sequential and parallel is shown in Figure 6.

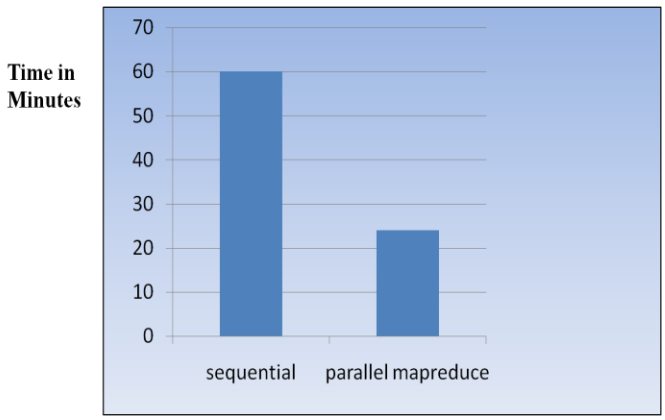

Figure 6: Bar graph for run time in sequential and parallel process.

V. CONCLUSION

The results proves that the proposed solution improves the performance of phylogenetic analysis by using apache hadoop map reduce programming model along with Complete composition vector to produce accurate alignment for large protein sequence and it uses UPGMA to produce PhyloXml tree. The proposed method is time efficient approach and it captures the protein families better when compared to other algorithms. A parallel approach of phylogenetic analysis is constructed using map reduce with CCV, that requires high end machines.

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] DeSantis TZ Jr, Hugenholtz P, Keller K, Brodie EL, Larsen N, Piceno YM, R, Andersen GL: NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. Nucleic Acids Res 2006, 34: W394-399.

[3] Wu X, Wan X-F, Wu G, Xu D, Lin G: "Whole Genome Phyogeny via Complete Composition Vectors." Technical Report TR05-06 Department of Computing Science, University of Alberta; 2005.

[4] Hongfeng Zhang1, Vincent Y. Liu2, Yu Zhao: "BioCloud: A Systemic Review and Classification" June 7, 2014; accepted May 8, 2015. doi: 10.17706/jsw.10.6.695-712.

[5] Gaggero, Massimo, et al. "Parallelizing bioinformatics applications with MapReduce." Cloud Computing and Its Applications (2008): 22-23.

[6] Siddesh G M, K G Srinivasa*, Ishank Mishra, Abhinav Anurag, Eklavya Uppal

[7] Maharjan, Merina. "Genome Analysis with MapReduce". June 15 (2011): 3-4.

[8] 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.

[9] White, Tom. Hadoop: The definitive guide. " O'Reilly Media, Inc.", 2012.

[10] Marc E Colosim, 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, 6:13.

[11] Apache Hadoop Mapreduce Programming Model, "http://hadoop.apache.org/Mapreduce"

[12] Páll Melsted autho and Jonathan K Pritchard "Efficient counting of **k**-mers in DNA sequences using a bloom filter" BMC Bioinformatics2011**12**:333

[13] [13] Sokal R and Michener C (1958). "A statistical method for evaluating systematic relationships". University of Kansas Science Bulletin **38**: 1409–1438.

[14] National Center for Biotechnology Information (NCBI), http://www.ncbi.nlm.nih.gov/