Review on Genome Sequencing using Hadoop and MapReduce

[1]Anushree Raj, [2]Rio D'Souza

[1]Assistant Professor, [2]Professor
[1]M.Sc. Big Data Analytics Department, [2] Computer Science and Engineering Department,
[1]St Agnes College Autonomous Mangalore, Karnataka, India, [2]St Joseph Engineering College Mangalore, Karnataka, India

Abstract: **Next generation sequencing has led to the generation of billions of sequence data, making it increasingly infeasible for sequence alignment to be performed on standalone machines. The Digital data can be stored in the form of genome sequence, which requires techniques to synthesise and sequence into the DNA sequences. This paper reviews about taking a dataset of DNA sequencing as an input and split them across the cluster machine by applying MapReduce implementation of Hadoop to make the search efficient for large scale genome sequencing applications.**

IndexTerms**: Genome Sequence, DNA Sequence, Hadoop, Mapreduce.**

I. INTRODUCTION

The rapid progress in genome sequencing technologies leads to availability of high amounts of genomic data. However, one growing concern is the ability to protect the privacy of sensitive information. DNA can be encoded or digitized and stored as a sequence on a computer database. This process is, ostensibly, what the various genome sequencing efforts are all about. DNA can also be synthesized, and, in fact, it can be synthesized using the prior, digital form of DNA in databases. Storage has been a fundamental requirement for the Humans. In the modern era of computing and communication, huge amount of data is being generated and there is a pressing need for dense storage medium which is cost effective. Clustering [1] is a popular technique used for classifying data set into groups. Data points under particular group share similar features. It is widely used for pattern recognition, data mining etc. To achieve globally optimal solution, it requires iterating over all possible clustering. Hadoop Mapreduce [2] is a parallel programming technique build on the frameworks of Google app engine mapreduce. It is used for processing large data in a distributed environment. It is highly scalable and can be build using commodity hardware. Hadoop mapreduce splits the input data into particular sized chunks and processes these chunks simultaneously over the cluster.

II. HADOOP

Hadoop is an open source project, implementing MapReduce, supporting distributed computing and also providing scalability [3]. Hadoop in short provides distributed storage and analysis system [4]. Storage is provided by the Hadoop Distributed File System (HDFS) [5] and the analysis system is the MapReduce programming model or the computational model. The Hadoop framework plays an important role since it is capable of handling large data sets and also the general aspects of distributed computing is automatically handled. Hadoop is distributed and column oriented database. HBase uses HDFS for its more efficient system storage [6].

HDFS (Hadoop Distributed File System)

Hadoop provides a distributed file system called Hadoop Distributed File System (HDFS) which runs on commodity hardware machines and store very large data set. One of the assumptions of HDFS designing was that "Moving computations is much cheaper than Moving Data". The core idea is to move larger set of applications closer to where the data is located and perform computations on it. The throughput is to be increased in HDFS by moving computations near data. HDFS scales up upto ten thousands of low cost hardware machines on which data can be stored in blocks and computation can be partitioned on these machines using MapReduce. MapReduce is a programming model which is used for processing massive data sets stored on HDFS clusters.

III. MAPREDUCE

MapReduce [7] is a popular cloud computing programming model or a framework for distributed computing used especially where large data sets are involved and analysed. In this model, a map function and a reduce function is specified by the user [8]. Map function makes the traditional data analysis in distributed manner by assigning given jobs to different nodes present in hadoop environment. In general Map Reduce model executes number of problems in parallel [9]. Reducer function receives collection of inputs from Mapper function, computing the result from input sets and gives as final output [10]. Apache hadoop is the best tool for processing large amount of database.

The MapReduce architecture consists of one master (Jobtracker) and many workers (Tasktrackers). The JobTracker receives job from the user, breaks it down into map and reduce tasks, then assigns those tasks to Tasktrackers, after that monitors the progress of the Tasktrackers, and finally when all the tasks are completed, reports the user about the completionof jobs. Each Tasktracker has fixed number of map and reduce task slots which determines how many map and reduce tasks it can run at a single time. The

Hadoop File System HDFS supports reliability and fault tolerance of MapReduce computation by replicating i.e. generating copies of the inputs and outputs of a Hadoop jobs and then storing them uniquely.

IV. GENOME SEQUENCING

The technique that allows for researchers to read and convert the genetic information found in the DNA of any organisms is called Genome Sequencing.

DNA Sequencing

DNA sequencing is the process of determining the accurate order of nucleotides along chromosomes and genomes. The sequencing procedure consists of the following steps:

1. ***Read DNA fragments***

The raw data produced consists of pairs of short fragments of DNA called reads, each with current technology and biochemistry about 100 bases long. Each fragment is first read from one end and then from the other, producing a total of two reads. With the HiSeq 2000 the read length is typically about 100 bases, meaning that for each DNA fragment of 300–500 nucleotides we have read the first and the last 100 or so, leaving an unread section in between. From a single run the sequencing machine can produce 800 GB of data, consisting of billions of records—two for each fragment—each containing four pieces of data: a key identifying a DNA fragment; the read number (one or two); [11] the DNA sequence read; the quality score for each base in the DNA sequence, which estimates the probability of a reading error at each base.

2. ***Map fragments to reference genome***

The genomic sequence must be reconstructed by determining the original locations of the fragments [12]. When aligning read pairs, the fact that the distance between the two fragments can be estimated is used to direct the alignment process to choose a position where both reads can be aligned within a statistically reasonable distance from each other. This alignment process is known as Short Read Alignment.

3. ***Detect and remove duplicate reads***

The duplicate reads must be eliminated to avoid introducing statistical biases into the data.

4. ***Recalibrate base quality scores.***

Adjust the quality scores of the detected bases to take into account several factors in addition to the optical quality measures used by the sequencing machine.

DNA Sequence alignment

DNA Sequence alignment is a way of comparing two or more different DNA sequences by searching for a meaningful character patterns that are in the same order in the sequences [13]. Sequence alignment mainly used to identify functional and evolutionary relationship between two different biological sequence, homology study, Evolutionary linkage and Molecular structure. While aligning a different sequence, execution speed and alignment accuracy are considered as major aspects.

Needleman-Wunsch [14] performs global sequence alignment between two nucleotide or amino acid sequences and find out structural or functional similarity. Smith-Waterman [15] performs local sequence alignment; that is, for determining similar regions between two strings of nucleic acid sequences or protein sequences, Instead of looking at the entire sequence. Sequence alignment algorithms works based on dynamic programming. These are all produces accurate alignment score. These algorithms need high computation for processing the data. Computation complexity depends on sequence size. When sequence size increases, complexity increases exponentially. ClustalW [16] uses progressive alignment methods, which align the most similar sequences first and work their way down to the least similar sequences until a global alignment is created, and T-COFFEE [17] (Tree-based Consistency Objective Function for Alignment Evaluation) is a multiple sequence alignment software using a progressive approach. It generates a library of pairwise alignments to guide the multiple sequence alignment. It uses progressive approximation method. These tools identify similar sequence in very fast. Hidden Markov Model (HMM) [18], Generic algorithms [19] are used in iteration based approach. HMM is created using already aligned sequence. It tests the sequence with respect to HMM or not. Dynamic Programming methodology produces better result but it needs higher computation power. Heuristics algorithms are too fast and it needs local maxima value. Iterative based approach is relatively slow. Indonesia [20] is the best example for structure based alignments using priori data. It uses basesian alignment for its alignment purpose. Basic Local Alignment Search Tool (BLAST) [21] used for sequence to sequence alignment. It implements in a number of ways and applied in a variety of contexts including straight-forward DNA and protein sequence database searches, motif searches, gene identification searches, and in the analysis of multiple regions of similarity in long DNA sequences. PCI-BLAST [22] is used for profile to sequence alignment. It constructs a multiple alignment from BLAST output data, processes this alignment into a position-specific score matrix and uses this matrix to search the database. Align-M [23] is example for non progressive approach that gives better accuracy results for the related sequence aligned distantly. Probalign [24] calculate the probability of pair wise using posteriori information. It constructs accurate RNA alignments and produces accurate phylogenetic trees. MAFFT [25, 26] reduces the CPU Execution time to identify similar regions. It uses iterative base approach to calculate the result. and MUSCLE [27] Given an aligned pair of sequences, it compute the pairwise identity and convert to an additive distance estimate. MAFFT and MUSCLE tools uses posteriori information. NRAlign [28] uses horizontal alignment information to give accurate result. DIALIGN-T [29] produces high accuracy in terms of gaps. ISP Align [30] is combined HMM and Probcons that identifies the sequence using the intermediate sequence profile to assure the accuracy. Fast Fourier transform PartTree [31] is used to create guide tree for constructing sequence alignment. Partial order alignment is used to achieve improved result in sequence alignment. FGPA [32] is used for improving the computation. it is a system with parallel and higher density logic elements, promising to significantly improve genomic sequence searching. CUDA compatible GPU cards are

used to exploit the huge computational power of commonly available graphic cards, to develop high performance solutions for sequence alignment. To search the alignment in homology database needs parallel architecture supported machines [33]. CEPAR [34] perform the biological related processing using parallel processor. It performs sequence to sequence comparison, parallelizes a single query across a partitioned and distributed database and the set of queries themselves are partitioned across a set of servers with replicated or partitioned databases. pBLAST [35] support parallel processor, hash table and query processing to achieve computational accuracy. It produces the best alignment for a pair of DNA or protein sequences. Genetic programming gives better results compared to dynamic programming Distributed and parallel environment support achieved in ClustalWMPI [36] and W.ND-CLAST [37] provides an interactive tool that allows scientists to easily utilizing their available computing resources for high throughput and comprehensive sequence analyses. BAliBASE [38] tools generate numerous test cases for sequence alignment problems. Quality of the alignment improved using dynamic programming methods. It also ensures optimal alignment between the sequences. Normal computation machine needs highest computation power for alignment process.

DNA Alignment tools

Various alignment tools are used for detecting genome variations such as single nucleotide polymorphisms (SNP) and large-scale structural variations, The extensive genetic informational datasets create many serious problems and challenges for the popular alignment tools such as bowtie [39], The aligner is typically used with short reads and a large reference genome, or for whole genome analysis, RMAP [40, 41] supports paired-end reads either as read sequences or using full quality-score information, MAQ [42] particularly designed for Illumina-Solexa 1G Genetic Analyzer, and has preliminary functions to handle ABI SOLiD data., bwa [43, 44] to map low-divergent sequences against a large reference genome. There have been some initiatives towards this trend of using Hadoop such as CloudBurst [45] a Map Reduce-based read-mapping algorithm modeled after RMAP, runs in parallel on multiple machines with Hadoop, but its website and code are in-accessible, SeqMapreduce [46] parallelising sequence mapping using Hadoop. It is a performance improvement version of CloudBurst, Crossbow [47] read mapping and SNP calling software that runs in the Amazon EC2 cloud, CloudAligner [48] is designed to achieve better performance, longer reads, and extremely high scalability. It has more common functions such as bisulfite (BS) and pair-end mapping as well as a friendly user interface, and it supports more input as well as output formats. These techniques are used for detecting genome variations such as single nucleotide polymorphisms (SNP) and large-scale structural variations, which are very important in biological analyses.

V. MAPREDUCE ALGORITHM FOR DNA SEQUENCING

Typical DNA sequencing for a single data sample (about 400–900 GB in the FASTQ file format) might take 70+ hours for a very powerful single server. The goal of the MapReduce algorithm is to find the answer in a few hours and make the solution scalable. The main idea of the MapReduce is to go through some Map, Reduce and shuffle parallel computational steps. These steps can be summarized in these three points:

1. Map: reads are mapped to the reference genome in parallel on multiple machines
2. Shuffle: aggregation of the alignments to be on the same chromosome and being sorted by position
3. Scan: scanning of the sorted alignments for biological events identification in each group

The DNA sequencing involves three steps:

i) DNA Sequence alignment

In the alignment phase, MapReduce/Hadoop implementation with open source tools BWA and SAM could be used.

• Burrows-Wheeler Aligner (BWA), is an efficient program that aligns relatively short nucleotide sequences against a long reference sequence such as the human genome [49].

• Sequence Alignment/Map (SAM) tools, provide various utilities for manipulating alignments in the SAM format, including sorting, merging, indexing, and generating alignments in a per-position format [50]. The BAM format, is the binary format of a SAM file.

The map( ) function will read the input file (one single chunk) and will generate an aligned file in BAM format[51]. Here, the map( ) function uses BWA to perform the alignment process. Once the alignment is done, then it will extract all chromosomes and save them in the MapReduce filesystem (HDFS, for Hadoop).

ii) Recalibration

In the recalibration step, each map( ) function will work on a specific aligned chromosome. The mapper will perform duplicate marking, local realignment, and recalibration [52]. The goal of map( ) is to create a local recalibration table filled with covariates. local covariates will be merged by the single reducer to create the final single global file, recalibration table that will be used by the map( ) function of the third and final step of DNA sequencing, variant detection.

iii) Variant detection

The map() function will use the BAM file generated by the map() function of the recalibration step, and the final single "recalibration table" file. The map() function will use open source tools, such as GATK and SAMtools to generate partial variants, which are raw BCF—binary call format—files. The reducer will concatenate, sort and merge the raw BCF files to generate a single VCF file [53].

VI. CONCLUSION

Genome Sequencing is finding out the order of DNA nucleotides or bases in a genome, the order of A's, C's, G's and T's that make up an organism DNA. Processing of Genome Sequence is very important. This paper presents an unified MapReduce/Hadoop framework that takes huge data of genome sequence and split it across the clusters to increase the speed and efficiency. The short read mapping of genome sequencing can be easily and efficiently processed by using Hadoop and MapReduce frameworks. The MapReduce framework can speed up the processing by splitting the large data on various different clusters and have good compatibility with an open source platform Hadoop that works on commodity hardware and can result decreasing the cost of service.

REFERENCES

[1] Jain, Anil K., M. Narasimha Murty, and Patrick J. Flynn. "Data clustering: a review." ACM computing surveys (CSUR) 31, no. 3 (1999): 264-323.
[2] White, Tom. Hadoop: the definitive guide: the definitive guide. " O'Reilly Media, Inc.", 2009.
[3] Jason Venner, Pro Hadoop, Springer, 2009.
[4] Tom White, Hadoop – The definitive Guide, 2009, O'Reilly Yahoo Press
[5] Dhruba Borthakur, "The Hadoop Distributed File System: Architecture and Design", Apache Software Foundation, 2007, Retrieved from http://hadoop.apache.org/hdfs/docs/current/hdfs_design.html.
[6] Li and Nath, "Scalable data summarization on Big Data", Distributed and Parallel Databases, vol 32,pp,313-314 ,2014
[7] Apache Software Foundation, Hadoop, 2007, http://hadoop.apache.org/
[8] J.Dean and S.Ghemawat, MapReduce: Simplified data processing on large clusters, in USENIX Symposium on Operating Systems Design and Implementation, San Francisco, CA, Dec 2004, pp.137-150.
[9] Du et al, "Review on the Application and the Handling Techniques of Bigdata in Chinese Realty Enterprise", Annals of Data Science , vol 1, no 3, pp. 339-357, 2015
[10] Lee, Hsiao and Hsieha, "A Dynamic data placement strategy for Hadoop in Heterogeneous Environment", Research on Big Data, vol1, pp.14-22, 2014.
[11] An accurate algorithm for the detection of DNA fragments from dilution pool sequencing experiments Vikas Bansal, *Bioinformatics*, Volume 34, Issue 1, 1 January 2018, Pages 155–162.
[12] A survey of sequence alignment algorithms for next-generation sequencing. Li H[1], Homer N. Brief Bioinform. 2010 Sep;11(5):473-83. doi: 10.1093/bib/bbq015. Epub 2010 May 11.
[13] A Secure Alignment Algorithm for Mapping Short Reads to Human Genome.Zhao Y, Wang X, Tang H.J Comput Biol. 2018 Jun; 25(6):529-540. Epub 2018 May 9.
[14] Sagl et al, "A general method applicable to the search for similarities in the amino acid sequence of two proteins", Journal of molecular biology, vol 48, no 3, pp.443-453,1970.
[15] Smith et al, "Identification of common molecular subsequences", Journal of molecular biology, vol 147, pp.195-197,1981.
[16] Li at al, "ClustalW-MPI: ClustalW analysis using distributed and parallel computing", Bioinformatics, vol19, no 12,pp.1585-1586, 2003.
[17] Higgins et al, "T-Coffee: A Novel method for fast and accurate multiple sequence", Journal of Molecular Biology, vol 302, no 1, pp.205- 217,2000.
[18] Holmes et al, "evolutionary HMMs: A bayesian approach to multiple alignment", Bioinformatics, vol 17, no 9, pp.803-821, 2001.
[19] Zhang et al, "A genetic algorithm for multiple molecular sequence alignment", Bioinformatics, vol13, no 6,pp.565-581, 1997.
[20] Madsen et al, "Indonesia: An integrated sequence analysis system – Manual', 2002.
[21] Stephen, Warren, Webb, Eugene and David, "BLAST", Journal of Molecular Biology, vol 215, no 3, pp.403-410, 1990.
[22] Stephen, Thomas, Alejandro, Jinghui, Zheng and David, "Gapped BLAST and PSI BLAST: a new generation of protein database search programs', Journal of nucleic acid research, vol 25, no 17,pp.3389-3402,1997.
[23] Van et al, "Align-m – a new algorithm for multiple alignment of highly divergent sequences', Bioinformatics, vol 20, no 8, pp.1428-1435, 2004.
[24] Roshanet al, "Probalign: multiple sequence alignment using partition function posterior probabilities', Bioinformatics, vol 22, no 22, pp.27152722,2006.
[25] Katoh, Misawa, Kuma and Miyata, "'MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform', Nucleic Acids Research, vol 30, no 14,pp.3059- 3066, 2002.
[26] Edgar et al, "MUSCLE: multiple sequence alignment with high accuracy and high throughput', Nucleic Acids Research, 2004.
[27] Improving accuracy of multiple sequence alignment algorithms based on alignment of neighboring residues. Yue Lu Sing-Hoi Sze, *Nucleic Acids Research*, Volume 37, Issue 2, 1 February 2009, Pages 463–472
[28] DIALIGN-T: An improved algorithm for segment-based multiple sequence alignment Amarendran R Subramanian1, Jan Weyer-Menkhoff, Michael Kaufmann1 and Burkhard Morgenstern. Published: 22 March 2005 BMC Bioinformatics 2005, 6:66 doi:10.1186/1471-2105-6-66
[29] Lu et al, "Multiple sequence alignment based on profile alignment of intermediate sequences", Journal of Computional Biology, vol15, no 7,pp. 767777,2004.
[30] Kazutakaet al, "PartTree: an algorithm to build an approximate tree from a large number of unaligned sequences", Bioinformatics, vol 23, no 3, pp.372-373, 2007.

[31] Grassoet al, "Combining partial order alignment and progressive multiple sequence alignment increases alignment speed and scalability to very large alignment problems", Bioinformatics,vol 20, pp.1546-1556,2004.
[32] Li, Shum, and Truong, "160-fold acceleration of the Smith-Waterman algorithm using a field programmable gate array (FPGA)", BMC Bioinformatics, vol 8,pp185, 2007.
[33] Manavski et al, "CUDA compatible GPU cards as efficient hardware accelerators for Smith- Waterman sequence alignment", BMC Bioinformatics,vol 9, suppl 2, 2008.
[34] Pedrettiet al, "Three Complementary Approaches to Parallelization of Local BLAST Service on Workstation Clusters", Springer Berlin, 1997
[35] Rajasekaran et al, "Randomized and parallel algorithms for distance matrix calculations in multiple sequence alignment", Journal of clinical monitoring and computing, vol 19, pp.351- 358,2005.
[36] ClustalW-MPI: ClustalW analysis using distributed and parallel computing, Kuo-Bin Li, *Bioinformatics*, Volume 19, Issue 12, 12 August 2003, Pages 1585–1586,
[37] Dowd, Zaragoza, Rodriguez, Oliver and Payton, "Windows .NET Network Distributed Basic Local Alignment Search Toolkit (W.ND- BLAST)", BMC Bioinformatics, vol 6(93),2005.
[38] Thompson et al, "BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark", PubMed, vol 61, no 1, pp. 127-136,2005.
[39] Langmead B, Trapnell C, Pop M, Salzberg S: Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 2009, 10(3):R25..
[40] Smith A, Xuan Z, Zhang M: Using quality scores and longer reads improves accuracy of Solexa read mapping. BMC Bioinformatics 2008, :128[http://www.biomedcentral.com/1471-2105/9/128].
[41] Smith A, Chung W, Hodges E, Kendall J, Hannon G, Hicks J, Xuan Z, Zhang M: Updates to the RMAP short-read mapping software. Bioinformatics 2009, 25(21):2841..
[42] Li H: Maq: Mapping and Assembly with Qualities. Version 0.6.3 2008 [http://maq.sourceforge.net/index.shtml].
[43] Li H, Durbin R: Fast and accurate short read alignment with BurrowsWheeler transform. Bioinformatics2009,25(14):1754-1760 [http://bioinformatics.oxfordjournals.org/ content/ 25/14/1754.abstract]
[44] Li H, Durbin R: Fast and accurate long-read alignment with BurrowsWheeler transform. Bioinformatics 2010, 26(5):589..
[45] Schatz M: CloudBurst: highly sensitive read mapping with MapReduce. Bioinformatics 2009, 25(11):1363.
[46] Li Y, Zhong S: SeqMapReduce: software and web service for accelerating sequence mapping. Critical Assessment of Massive Data Anaysis (CAMDA) 2009 2009.
[47] Langmead B, Schatz M, Lin J, Pop M, Salzberg S: Searching for SNPs with cloud computing. Genome Biol 2009, 10(11):R134.
[48] CloudAligner: A fast and full-featured MapReduce based tool for sequence mapping. Nguyen T[1], Shi W, Ruden D. BMC Res Notes. 2011 Jun 6;4:171. doi: 10.1186/1756-0500-4-171.
[49] http://bio-bwa.sourceforge.net/ last visited: June 20, 2016.
[50] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth,Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup
[51] http://dnasequencing.com, last visited: June 20, 2016. https://doi.org/10.1155/2012/251364
[52] RIG: Recalibration and Interrelation of Genomic Sequence Data with the GATK Ryan F. McCormick, Sandra K. Truong, and John E. Mullet
[53] Mapping short DNA sequencing reads and calling variants using mapping quality scores Heng Li, Jue Ruan, and Richard Durbin