ISSN: 2321-7782 (Online)
Impact Factor: 6.047

Volume 4, Issue 5, May 2016

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

SUPERBLAST: An advanced gene sequencing algorithm for Hadoop platform in Bio-informatics era

Madiha Shanam[1]
Computer Science and Engineering
CMR Institute of Technology
Bangalore – India

Dr. Sanjay Chitnis[2]
Computer Science and Engineering
CMR Institute of Technology
Bangalore – India

Abstract: ***The NCBI and gene sequencing technologies generate sequence data at an exceptional rate, a distributed strategy of virtual portioning for the database and query sequences was proposed but existing state of architecture is designed based on single-processor sequence alignment algorithms which is finding difficulty in keeping pace with current growth and it suffer from memory constraint in term of storage and computing. There comes a mechanized paradigms like timely computation, assembling and analysis of nucleotide or DNA sequencing but this become bottleneck to design very large scaling sequence and are incompetent for Big data processing – a parallelized maprduce programming of Hadoop framework. To overcome this, Super Blast, an efficient gene sequencing algorithm is proposed, a global alignments in terms of leading and lagging strands which adopts a new mapping algorithm optimized parallely for aligning sequence data by adopting an efficient scheduling technique. To evaluate its performance, nextNGS data is mapped to a nr protein database genome in a variety of configurations. That shows the effectiveness of proposed model in term of computation time and accuracy for better phylogenetic analysis.***

Keywords: Bioinformatics; Sequence Alignment; Big Data; Hadoop; Mapreduce; Genomics.

I. INTRODUCTION

This Bioinformatics use system machinery for managing biological information. Sequence, a serially related thing following each other in a particular order. Sequence alignment in bioinformatics is a method wherein we discover similar regions where similarity may be with respect to structurally, evolutionary and may functionally also. Sequences we try to align is of RNA, DNA and protein. Since the biological discoveries is getting advanced, there coming an innovative approaches for high performance computation. This computative approach is mainly requisite in favor of time manner, analysis and well-organized processes. On available huge genome databases, the importance on rapidity is essential for building algorithm practically, although there may be faster subsequent algorithms. There are many computational algorithms used for aligning the sequences with some of the methods like dynamic programming, probabilistic or heuristic methods. A BLAST, a widely use bioinformatics programs in searching of sequences similarties. BLAST, a heuristic version of Smith Waterman algorithm for pairwise local alignment which is the earliest and widely used algorithm which ensures the best performance on accuracy. As it is very time consuming we moved on to the BLAST eventhough not guaranting best possible alignment of both sequences. FASTA got introduced before BLAST by Lipman and Pearson. BLAST uses an approach of virtual partitioning where database is portioned automatically depends upon both the size of hadoop cluster and input query sequences. As existing state of BLAST architecture is designed based on single-processor sequence alignment algorithms which are finding difficulties in keeping pace with current growth and also it suffer from memory constraint in term of storage and computing. To overcome this, SUPERBLAST, a new gene sequencing algorithm is proposed which adopts mapping algorithms optimized in aligning data sequence from systems to database genome. This adopts the Hadoop MapReduce to parallelize execution on multiple computing nodes by adopting an efficient scheduling technique. This advanced gene sequencing algorithm is more time-efficient compared

to the normal BLAST algorithm which assure the optimal alignment of query and database sequence. This algorithm also ensures the best performance on accuracy and gives the more precise result. This tool gives the high performance computation with diagnostic testing which helps in phylogenetic analysis. It compares the similarities between nucleotide and the protein sequences which outputs the significant matches. It is more scalable compared with earlier algorithms. This led to the enormous generation of data where it has a rise with execution speed. The significant prerequisite in molecular diagnosis is to rapidly analyze the vast amount of datasets.

II. BACKGROUND AND RELATED WORK

Bioinformatical studies, grouping of the two main parts of science, biology as well as computer science. Learning of biology11[3] part is represented in an effecient way with system tools that jointly form bio-informatics. A scope of bioinformatics arises by means of Human Genome Project, where the entire human genome is sequenced. Now a day's bioinformatics is a keen area of interest for many scientist who has concerned for this subject hence they carry everyday research. Current discoveries of bioinformatics unified along with the enormous release of characteristics like its function, structure and also its genomic data gets swiftly mounting latent scale of sequence alignments method. This type of application is used in this module.

A process of mapping of the sequences in order for pertaining high range of similarities is nothing but a sequence alignment. Mapping is done with database sequences and query sequences. Hence we can study the precise characteristic among these sequences and the characteristics may be of structurally, functionally or evolution arises. The generated result with alignment ensure on conveying function for different proteins, determining relationships of an organism and also forecasting the 3D protein structure. The Homology can be termed like similarities, meaning if the two sequence is similar from different, its referred Homologous.

A. *Alignments*

Alignments are adapted for discovering homology among different kinds of sequence. It can be pooled along phylogenetical study. It is used to conclude different kinds of relationship like orthologous and paralogous.

B. *Alignment Types*

- Respect to length of sequence alignment

On comparision of sequences in whole or as sections similarity:

Global alignment – Here the total length of sequence is used to scan for similarity searching. Even insertion of gaps takes place to fill unalign space.

- It spans one entire sequence with another total sequence.

- Costly in computation wise

1. Local alignment – Here it concentrates only on the high similarity regions locally. Even gaps are assigned for unaligned portions. Here, sequences subset tries for mapping of another subset sequence. cheap compared to global sequencing.

- Respect to numeral sequence

This provides two methods while calculating the amount or quantity of sequences:

1. Pairwise sequence alignment-- Here two sequences are aligned to result the best similarity region.
2. Multiple sequence alignment – It is similar to pairwise alignment but it process for greater than two sequence for a certain time. Even it is used to span all the sequences as much as possible.

C. *Parameters to produce optimum alignment is followed*

- *Maximum (max) target sequence*

  -This exhibit outcome according to the totality of sequence ie aligned.

- *Expected Threshold*

  -Its an indicator that calculate possibility where alignment results is cause with arbitrary option. Score is inversely proportional to E value. Default rate is kept 10 as 10 matches is found to be expected randomly.

- *Query matches*

  -This provide maximum match in a range of query. The result of stronger query match with weaker ones are compared.

- *Word size*

  -Here the words are matched among query as well as database sequence. This exactly searches for matching of words, when found it extends that match to the full length alignment. Three is an optimum word size for standard protein alignment and word size of two is requisite for an exact match.

- *Scoring scheme*

  -There are varities of scoring scheme algorithm which devise to attain the finest alignments. Substitution matrix help for aligning probable residual pair in addition for the generation of scores. Different PAM and BLOSUM matrices are used to ensure the excellence of pairwise sequence alignment

- *Point Accepted Mutation (PAM)*

  -The mutation can be build up to calculate amino acid substitutions through evolutions ie naturally accepted. The mutation PAM30 is applied when the length of sequence is less than 35% whereas PAM70 is used when sequences range from 30% to 50%.

- *Block amino acid substitution matrix (BLOSUM)*

  -BLOCKS are the conserved regions developed with the related protein sequence accessible inside blocks databases. Similar proteins is used to detect with the help of matrix called BLOSUM 62. The longer and weaker alignments is used by BLOSUM 45.

- *Gap cost*

  -Its a space that is established in the alignment for compensation of insertions and deletions (indels) in sequences comparative with one other. The best algorithm is the one where gaps are neglected within alignment and where penalty/score of gaps is assigned.

III. DESIGN AND IMPLEMENTATION

As computerized gene sequencing technologies is rising exponentially, there come mechanized paradigms like timely computation, assembling and analysis of nucleotide or DNA sequencing. Shorter sequences can be aligned effectively but this becomes bottleneck to design very large scaling sequence. Eventhough there are many evolutionary metaheuristic mechanisms, there arise an issue of best DNA sequencing algorithm for better outcomes. Hence project aims for super gene sequencing algorithm, reducing memory space and time for better phylogenetic analysis.

- Improved scalability over existing solutions by the use of Hadoop SUPERBLAST.
- Balance the computational work load.

- Enhanced the Blast search performance.

A. *Sequence partitioning*

Blast provides database virtual partition on runtime where it first segregates the databases severally in lesser subdivisions where it gets upload to HDFS. Second, it forms larger ‘‘virtual partitions’’ on integrating these subdivisions. Here each partition directs to be fitted totally in a node RAM by employing the cluster nodes. SUPERBLAST ‘make-blastdb’ is an NCBI tool which translate and segregates the database in FASTA format into a set of binary files where it is use with other application. A maximum size of compile database which is divided into numerous pieces is 1.5 GB each. This each subdivision is considered as a stand-alone database as in figure below.

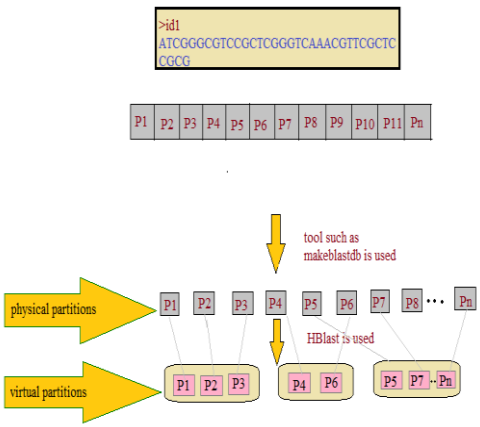

Fig 3.1 Virtual partitioning of nr dataset

By the experimentation, it was emphasized so as to if single process is being execute in single full nr database and 50 Blasts searches are being execute in 50 divisions, an acquired product of a former is as a minimum 3 three time slow than the input query size being utilized Blast MapReduce algorithm.

Hadoop is an Apache open source framework which is written in java. Hadoop framework is used for distributed storage and also for processing large amount of data sets using simple models. It includes java based file system which is used for storing enormous amount of data which is scalable and reliable. And we are using another framework for computation called map reduce. This map reduce is one of the processing technique for computations.

There are two important aspects of SUPERBLAST are taken into account while selecting the smallest partition size i.e first, a database should be fitted in a ideal way in to the RAM locally of a Hadoop nodes so that the database search process would not slow down heavy disk I/O rapidly. Second, without making the partition of the database into atleast N pieces it is impossible to make all N nodes to work. Virtual partition size is reduced from the maximum number of partition, which is necessary for covering both the cases. The two extreme cases are used in Figs. 3.2a and 3.2b.

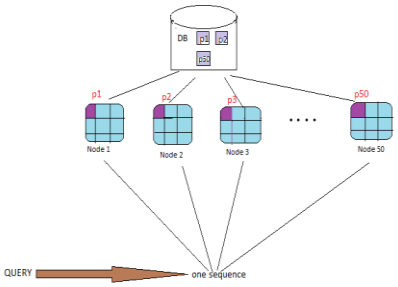

Fig 3.2a Single query sequence with number of nodes – RAM unoccupied on each node

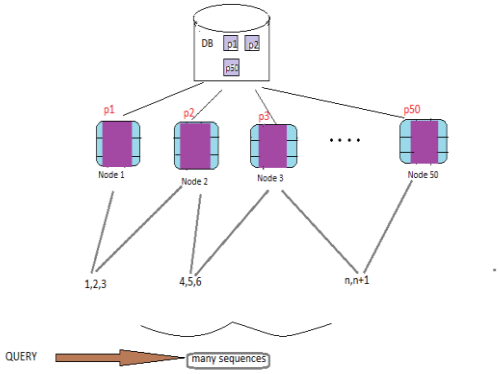

Fig 3.2b Many query sequence with number of nodes – RAM fully occupied on each node

The two extreme cases are used in Figs. 3.2a and 3.2b. In fig 3.2a cluster of 50 nodes are shown. The database can be divided into 50 parts for running of databases searching only one query . These parts of database are shared to all the nodes , the size of one database piece is so small that can be fitted in to the RAM of any node. It takes a significant attempt to accumulate the results because of which a very rapid search is being done. While selecting the best approach to search, Fig. 3.2b shows different challenge. In this case, many input queries are splitted into 50 sets which are then divided into the nodes. By anyhow the database is that so much big where some part of it can be fitted within the Hadoop RAM by a certain time. A databases is being separated within numerous partition plus to minimize the no. of I/O of the disk, processing must be done at one partition at then SUPERBLAST gets the database and query of the input sequence. It also calculates the number of tasks of the maps which is important to keep the cluster of workload in a balanced way. To assist this, the file of the input query sequence file is also being splitted. The file i.e. input record for mapping of Hadoop tasks contains a name of file for the query sequences of the input and virtual partitions i.e. the virtual partition is formed by the smaller physical database partition.

B. *Architectural Design*

A general structure of project be logically alienated to a modules which gets processing and where data structures is referred to be Architecturally designed.

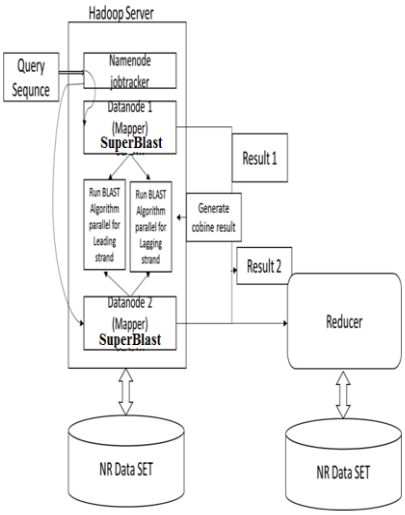

Figure 3.3 SUPERBLAST Architecture.

Query sequence is subjected to the hadoop server which includes namenode, data node, job tracker, task tracker and the algorithms like blast and SUPERBLAST. Namenode is the node which stores metadata of the file system. The actual data resides in the data anode. The job tracker functions the availability and management of resources. DNA double helix has both the leading and lagging strand. SUPERBLAST is an algorithm for both leading and lagging strands where as blast was used for only one leading strand. For each strand we are doing blast and then combine aggregates is found using SUPERBLAST with the help of 6 frame translations. This intermediary results is subjected to the Reduces where it sorts out the desired output.

***C.** Algorithm Design*

Bioinformatical studies says that BLAST for Basic Alignment hunt Tool, algorithm which compares primary genetic sequence information, like different proteins of the same or different amino-acid sequences or nucleotides of DNA sequences.

The main four components of BLAST are: queries, database, program, as well as purpose to seek out. Every molecule of DNA is comprised with 2 complementary strand one is leading and other is lagging, i.e anti-parallely arranged with each another. For each DNA strand there are 3 potentially read frames and totally 6 potentially reading frame intended for protein translation in DNA molecular region (3 frames per strand).

SUPERBLAST: Nucleotide query sequence, Leading and Lagging DNA (both strands) newly determined nucleotide against a protein sequence in database.

Step 1: For Leading DNA strand find score using BLAST algorithm

a. Quickly locate ungapped similarity region between the sequences. Instead of comparing each word of the query with each word of the database: it creates a list of "similar" words.

Now we can understand like this is for query sequence positions referred p , fetch list of length w which is scoring greater than T while pairing with word started at p it give listing of words of w length, scores greater than T in respect of word p. Identify all the exact matches with respect of Database sequences for each list of words .

b. Each match is then extended. The expansion gets stopped at the moment S scores decrease more than X when compared with the highest value obtained during the extension process.

Step 2: Find score for the lagging DNA strands by using BLAST

Follow the same procedure of above mention in step 1

Step 3: Give combine result of both step 1 and 2 using SUPERBLAST.

The DNA replication for one helix result in the outcome of 2 similar helices. Earlier DNA helix is referred to be the "parental" DNA, and the two resulting helices are referred as "daughter" helices. Any daughter helices will be similar to parental helix. The DNA parental strands can be used as a guide or templates for creation of “daughters”. A daughter strand which is newly synthesized composed by the addition of nucleotide which is complementary to the DNA parent strand. Hence replication of DNA is semi-conservative where daughter strand is always conceded by one parent strand.

DNA is made of bunch of nucleotides and it is made of phosphate group, 55 carbon sugar molecule and 4 nitrogen bases. DNA is antiparallel which means that the two strands of DNA will be moving in opposite directions. One strand is moving in 5prime to 3 prime direction and other by 3 prime to 5 prime and base pairs are attached to the two strands which makes DNA helix.

Imagine a DNA helix where base pairs are attached. On synthesizing the DNA helix opens up at a clamp like structure called helicase which is a helix braker to unzip the DNA helix. On the individual strand there resides a single stranded binding protein which protects the DNA from cleavaging and stabilize the opened up parental DNA. Parent DNA and daughter DNA both are antiparallel. DNAs are only built from 5’ to 3’ direction means that nucleotides are added in 3’ direction with the help of DNA polymerase and is called as leading strand and more number of nucleotides are added discontinuously which forms the segments with certain gaps in the opposite strand called as lagging strand. This segments of DNA is called okazaki fragments.

IV. EXCREMENTAL RESULTS

The result or output generated with step by step execution of every projected protocol for unusual values of time and speed is given by the following graph.

It describes the particular hadoop job where map and reduce functions gets processed. It gives the percentage which says how much percent map or reduce is completed. It also gives the graph with respect to map inputs along with time of computation.

This shows the final outcome of the job where both completion of 100 % is shown. Along with this it also gives number of tasks, running jobs, pending jobs, completed jobs and also the bytes which are processed. For single task CPU time spent is 5930 ms and total time for execution is 166.8 sec and for two task CPU times spent is 8750 ms and execution time is 182.848 sec and so on. So we can say that we have very efficient computation in terms of seconds. Hence we will get the correct phylogenetic outcome.

V. CONCLUSION

As the NCBI and gene sequencing technologies generate sequence data at an exceptional rate, a distributed strategy of virtual portioning for the database and query sequences was proposed. Additionally SUPERBLAST algorithm is executed parallely based on hadoop map reduce programming framework. This existing state is based on single-processor sequence alignment which finds complexity in rapidity of current growth and suffer from memory constraints in terms of storage and

computation. To overcome this here we propose an efficient gene sequencing algorithm called Super Blast, is global alignments in terms of leading and lagging strands which adopts a new mapping algorithm optimized parallely for aligning sequence data by adopting an efficient scheduling technique. Super blast is advantageous over existing SUPERBLAST approaches due to faster execution, ease of deployment and improved scalability. For its performance evaluation, proposed method is used to span next generation sequence (NGS) data to a nr protein database genomes in a varieties of configuration. The resulting outcome shows the effectiveness of proposed model in term of computation time. The future work we consider to use multi cluster hadoop environment and the introduction of complex key for further sorting in the reducer.

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany a successful completion of any task would be incomplete without the mention of people who made it possible, success is the epitome of hard work and perseverance, but steadfast of all is encouraging guidance.

So with gratitude I acknowledge all those whose guidance and encouragement served as beacon of light and crowned our effort with success.

I consider it a privilege and honour to express my sincere gratitude to *Dr. Sanjay Chitnis,* Principal and my internal guide, CMRIT, Bangalore for his valuable guidance throughout the tenure of this project work.

I would also like to thank all the members who have always been very Co-operative and generous. and all others who have done immense help directly or indirectly during my project.

References

1. *"Okazaki Fragment Metabolism"* - Lata Balakrishnan1, Robert A. Bambara, April 19, 2016.
2. *"Upcoming challenges for multiple sequence alignment methods in the high-throughput era"* - Kemena C, Notredame C, July 30, 2009.
3. *"Big Data', Hadoop and cloud computing in genomics"* Daugelaite J, O'Driscoll A, Sleator RD, 2013.
4. *"Basic local alignment search tool"*, Altschul S, Gish W, Miller W, Myers E, Lipman D, 1990
5. *"A high-performance, portable implementation of the MPI message passing interface standard. Parallel Computing"*, Gropp W, Lusk E-L, Doss N-E, Skjellum A.

AUTHOR(S) PROFILE

Madiha Shanam, received the M.Tech degree in Computer Science and Engineering from CMR Institute of Technology in 2014-16.

