International Journal of Latest Trends in Engineering and Technology
Vol.(8)Issue(4), pp.101-106
DOI: http://dx.doi.org/10.21172/1.84.14
e-ISSN:2278-621X

A BIG DATA APPROACH-PARALLELIZATION OF GENE DATA USING SMITH-WATERMAN ALGORITHM ON HADOOP PLATFORM

Leena I Sakri[1] and Akifa G Contractor[2]

Abstract- The broadcasting of bioinformatics has caused several recently developed determinations in living creatures. This determination would be incomplete without the innovation of genomic sequence alignment. Since the genome sequence alignment is vast and computational intense in biological sequence. Parallelization is a method where the computation of genome sequence can have less mathematical representation, which can be executed in limited cost and time. MapReduce Framework helps us to achieve the parallelization of data which works in serial manner. In this paper, Introduction of smith-waterman algorithm which works parallel on Hadoop-Mapreduce platform for the genome sequence alignment in bioinformatics

Keywords – Bioinformatics, Genome Sequence, Hadoop-MapReduce, Parallelization

I. INTRODUCTION

Big Data deals to an enormous quantity of data which is growing day by day. It has a significant job in the segments; it is as well used in emergent of applications in big data which consists forecasting of weather, cultivation, bioinformatics, learning, social sites, medicines, ecommerce websites, e-governance and so on. Due to the immense enlarge in size of the gene data and its complex structure, it is not easy to evaluate and give proper details a big sum of data.

In upcoming years, Big Data have gradually developed into a word relating datasets which are excessively huge to be tackled by conservative off-the-shelf data management systems like Traditional databases or inheritance data investigation applications [2]. An application domain, which particularly faces the bigdata test is given by bioinformatics, where genome sequencing produces several thousands of megabytes per genome. The current achievements in the developments of bioinformatics technologies are available to fabricate massive quantity of genome data in form of Gigabytes to terabyte per execution. Saving and examining such data is a difficult that exists and it needs to be handled. Bioinformatics genome sequence [6] find out the structure of DNA which will compose the reference database to identify the similarities of the regions that is, it will compare two datastream of genome sequence and match the identical section and differences between them. This process take place thousands of times per day all over the world. The output will give the maximum number of scores and the alignment of the sequence. Genome sequencing involves the bases of nucleotides such as cytosine, thymine adenine and guanine which are commonly referred as C, T, A, and G. group of three arrangement of nucleotides, determines a corrosive amino acid. For instance AGT. Usually the gene data comprises of a long sequence of this triplet nucleotides which forms amino acid.

Blast is an algorithm which is used as the search alignment tool it does very fast execution than other search sequence alignment algorithm [4] but it doesn’t guarantee the exact result or precise alignment after the execution. Smith water-man algorithm is also a search alignment algorithm which search for the sequence alignment between

[1] *Department of Information Science and Engineering SDM College of Engineering and Technology, Dharwad, Karnataka, India*
[2] *Department of Information Science and Engineering SDM College of Engineering and Technology, Dharwad, Karnataka, India*

two unlike data streams it gives you the precise arrangement of sequence alignment but it take a lot of time to execute [1].

In order to resolve those problems mentioned, A new smith-waterman algorithm is proposed in this paper .which take the input as a genome dataset uses Hadoop-MapReduce [3] to solve the inefficiency problem on large data sets.Execute in the parallel manner [7] makes the execution more efficient [5] and fast where, in this new algorithm the rapidity of Blast and accuracy of Smith-waterman algorithm is derived.

The most important contributions of this approach can be summarized as follows:

a) To conquer the instability and the sensitivity to genome data, a parallelization in Hadoop-MapReduce Method is introduced to improve the stability and accuracy.
b) The inefficiency problem in large data sets is solved by using New Smith-waterman algorithm implementing on Hadoop-MapReduce framework
c) New Smith-waterman algorithm is Faster and Accurate
d) Widespread experiments have been performed to illustrate that technique is efficient in solving the problems described above.

The remaining part of the paper is described as follows. Proposed system is explained in section II. Working of Algorithms in section III. Concluding remarks are given in section IV.

II. RELATED WORK

A genome dataset is consider to be a input which contains the four nucleotides bases in the long series that is commonly referred as A, G, T and C and these described letter gives us the information about the gene data. The genome data is fed into the HDFS [3] which turns it into distributed storage for the Hadoop-Map Reduce framework to overcome the faults of MapReduce, a parallel execution is adopted for the efficient solution and faster execution using smith-waterman algorithm for the sequence alignment

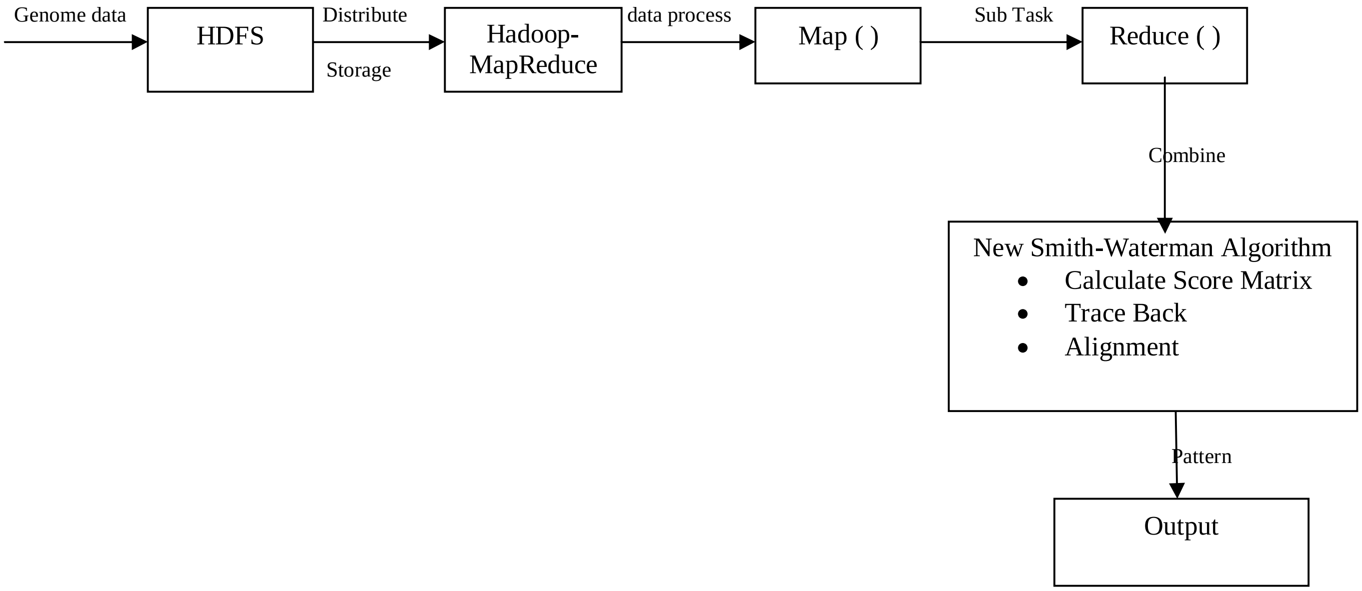

Fig 1: Data Flow of System

A. *Bioinformatic Genome Sequence Aligment*

To distinguish two proteins sequences or DNA and RNA sequences, the procedure need to find the best alignment between sequences, which has to be placed one above the other assembling apparent the communication between similar characters. In an alignment, empty spaces can be positioned in arbitrary along the DNA or RNA sequences. fundamentally, an arrangement can be inclusive, containing all lettering of the DNA or RNA sequences; local, containing substrings of the genomic sequences; or semi inclusive, collected of prefixes or suffixes of the genomic sequences, where trailing gaps are not considered.

In order to calculate the similarity between two genome sequence A1 and A2, we need to calculate the scoring matrix: provided an alignment between sequences. The scores will be assigned in each column as follows:

- If both the sequences have same character than its matching update addition of numeric one in the matrix
- If both the sequences have different characters than it is mismatching update subtraction of mumeric one in the matrix
- if both the sequence if any one of the characters is a gap update subtraction of numeric two in the matrix

The keep count is the collective result of all these values. A constant value is assigned to gaps. However, keeping gaps together Generates more significant results

B. Hadoop–Mapreduce Framework

In the MapReduce technique a double stage execution process is taken. In the beginning stage the input data which is given is divided into many chunks. Every chunk is allied with a Map worker. The Map worker gives the (key/value) pairs as outcome that is arranged on the source of the Key and values. The arranged values are given to the Reduce workers i.e. (key/sorted list (value). The Reduce workers reserve the outcome in the Hadoop distributed file System (HDFS).here for the efficiency the Hadoop-MapReduce adopts the parallel execution.

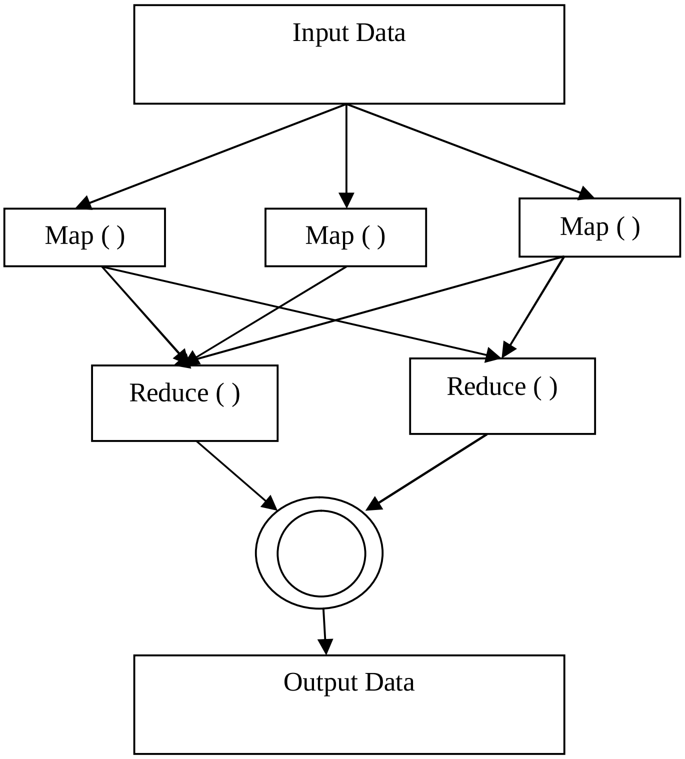

Fig 2: Working of Hadoop-MapReduce

C. Smith-Waterman Algorithm

In 1981, T. Smith and M. Waterman proposed Smith-Waterman algorithm for genome sequence alignment. The Smith-Waterman algorithm is a dynamic programming algorithm which is the search tool to find local sequence alignment that is finds out the identical section between two protein or DNA sequences. Till today it is used in many applications of bioinformatics. It guarantees the optimal alignment between query sequence and reference sequence. This algorithm is used when u cannot afford to miss any kind of information especially in the medical treatment

The algorithm consists of following two steps:

1. Compare the two dissimilar datastream and find out the identical region between them, store the highest scores of similarity in the matrix.
2. According to the dynamic programming technique, trace back the similarity matrix to search for the optimal alignment. That is from bottom to top, highest score till it reach the zero. In the algorithm, the first step will consume the largest part of the total processing time.
3. The results are precise.

D. Blast Algorithm

Blast Algorithm is Basic Local Alignment Search Tool which is also used to find the similar sequence between two unlike datastream it gives the rapid execution but the accuracy is not as precise as the Smith waterman algorithm and it doesn’t evaluate the gap penalty or gaps

III. WORKING OF SMITH WATERMAN ALGORITHM

The basic steps of algorithm are:

- Calculate a score of matrix
- Trace Back
- Alignment

To find the local sequence alignment consider two sequence
ACACACTA (sequence#1 or A)

AGCACACA (sequence #2 or B)

CALCULATE A SCORE OF MATRIX

Step 1: Initialize rows=i and column=j.

Step 2: Travel from i=0 to i=n and j=0 to j=m.

Step 3: Where n is the length of the sequence 2 and m is the length of sequence 1.

Step 4: If the residues are:
MATCHING: +2 diagonally
MISMACHING: -1 from the highest of the neighbouring boxes.

Example

$$H = \left(\begin{array}{cccccccccc}
 & - & A & C & A & C & A & C & T & A \\
- & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
A & 0 & 2 & 1 & 2 & 1 & 2 & 1 & 0 & 2 \\
G & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\
C & 0 & 0 & 3 & 2 & 3 & 2 & 3 & 2 & 1 \\
A & 0 & 2 & 2 & 5 & 4 & 5 & 4 & 3 & 4 \\
C & 0 & 1 & 4 & 4 & 7 & 6 & 7 & 6 & 5 \\
A & 0 & 2 & 3 & 6 & 6 & 9 & 8 & 7 & 8 \\
C & 0 & 1 & 4 & 5 & 8 & 8 & 11 & 10 & 9 \\
A & 0 & 2 & 3 & 6 & 7 & 10 & 10 & 10 & 12
\end{array}\right)$$

TRACE BACK

Step 5: Obtain the highest of the matrix.

Step 6: First search the biggest value of the box and then find the source of that value and keep backtracking it until we reach zero.

$$T = \left(\begin{array}{cccccccccc}
 & - & A & C & A & C & A & C & T & A \\
- & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
A & 0 & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \leftarrow & \nwarrow \\
G & 0 & \uparrow & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \nwarrow & \uparrow \\
C & 0 & \uparrow & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \leftarrow \\
A & 0 & \nwarrow & \uparrow & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \leftarrow & \nwarrow \\
C & 0 & \uparrow & \nwarrow & \uparrow & \nwarrow & \leftarrow & \nwarrow & \leftarrow & \leftarrow \\
A & 0 & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \leftarrow & \leftarrow & \nwarrow \\
C & 0 & \uparrow & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \leftarrow & \leftarrow \\
A & 0 & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \nwarrow
\end{array} \right)$$

ALIGNMENT

Step 7- When a box value form a diagonal, take both values.
Step 8- When it comes from beside box, take only vertical axis value.
Step 9- When it comes from bottom box, take only horizontal axis value.

RESULT
Alignment will be:
Sequence 1 = A-CACACTA
Sequence 2 = AGCACAC-A

IV.CONCLUSION

The Smith-Waterman Algorithm does well when the accuracy part is considered. But when the time efficiency is considered questions are raised over the Smith-Waterman Algorithm. There is no particular efficiency of the Smith-Waterman algorithm. The efficiency depends upon the length of the sequences we choose. As there are many operations to be carried out, if they choose a big alignment this algorithm takes a lot of time to provide the result.

Thus this algorithm is not feasible for large datasets as it will take a lot of time to compute the above. So we have found the method so that it can be efficient and fast by adopting parallel execution in Hadoop-Mapreduce. And changing the direction which makes the system more efficient and fast

REFERENCES

[1] Xiaowen Feng, Hai Jin, Ran Zheng, Zhiyuan Shao, Lei Zhu,” Implementing Smith-Waterman Algorithm with Two-dimensional Cache on GPUs ” Second International Conference on Cloud and Green Computing, 2012.

[2] Novan Zulkarnain and Muhammad Anshari, "Big Data: Concept, Applications, & Challenges”, International Conference on Information Management and Technology (ICIMTech),2016.

[3] J. Ramsingh and V.Bhuvaneswari, “Data Analytic on Diabetic awareness with Hadoop Streaming using Map Reduce in Python”, IEEE International Conference on Advances in Computer Applications (ICACA), 2016.

[4] Ming Meng, Jing Gao*, Jun-jie Chen,"Blast-Parallel: The parallelization implementation of sequence alignment algorithm based on Hadoop platform ", 6th International Conference on Biomedical Engineering and Informatics (BMEI 2013) , 2013.

[5] Saad Khan Zahid, Laiq Hasan , Asif Ali Khan, Salim Ullah, "A Novel Structire of Smith-Waterman Algorithm for Efficient Sequence alignment", ISBN: 978-1-4799-6376-8/15/$31.00 ©2015 IEEE.

[6] Miss. Anju Ramesh Ekre," Genome Sequence Alignment tools: a Review", 978-1-4673-9745-2 ©2016 IEEE.

[7] Rohith K. Menon, Goutham P. Bhat and Michael C. Schatz,” Rapid Parallel Genome Indexing with MapReduce”, http://www.genome10k.

[8] Merina Maharjan, “Genome Analysis with MapReduce”, ttp://hadoop.apache.org/

[9] Miss. Anju Ramesh Ekre and Prof. Ravi. V. Mante , "Hadoop Based Clustering System for Genome Sequencing", Second International Conference on Science Technology Engineering and Management (ICONSTEM), 2016.

[10] Muhammad Shafiq, Jord`a Polo, Branimir Dickov and Tassadaq Hussain,"Modelling and Performance Evaluation of Smith-Waterman Algorithm”,13th International Bhurban Conference on Applied Science & Technology(IBCAST), 2016.