International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-2, December 2019

Algorithm for Protein Sequence Alignment using Hadoop

Sneha Arjun Khaire

Abstract*: To develop an efficient system for matching the biological protein sequences and generating the scoring matrix using a distributed scan approach by applying Smith-Waterman(SW) algorithm. The algorithm generates fatest solution and the proposed system is comparing sequences with System, OpenMP and Hadoop. The comparison of the system leads in generating an efficient matrix of the protein sequence, beneficial for predicting the efficiency of the system.

***Keywords* : Bioinformatics; Proteinomics.**

I. INTRODUCTION

The method of sequential alignment is elementary that assists to work out with the biological associations or relationships from a huge data-sets can be achieved in parallel distributed environment. Such enormous tasks are difficult to perform using conventional strategies like string matching operation which is not effective for matching large sequences. So, In this paper we are using the proposed algorithm which is based on dynamic programming principles for alligning sequences using the local and the global information . Sequencing is the process to determine the sequence of a DNA fragment or a protein.

The Smith- waterman algorithm is far more efficient for the parallel distribution of the protein sequences using Hadoop. One of the vital applications in the field of bio-informatics includes the research based on evolutionary growth and the history of species. Using this algorithm it will be easy to predict the disease causing to the human; as it will lead to the prior precaution to the patients and also the patient who is not in the same town can be treated by the distributed approach, This is possible by using the prescription given by the doctor to the patient and also its history and its medical reports. The Smith-Waterman algorithm process local sequence alignment that is, for intuiting similar regions of protein sequences.

II. OBJECTIVE

Objective of the system is as follows:

1. To devise an unique perspective for solving problem related to alignment of biological sequences.
2. To obtain better throughput and efficient analysis.
3. To develop distributed scan approach for fast solution.

Revised Manuscript Received on December 30, 2019.
***** Correspondence Author
Sneha Arjun Khaire*, Assistant Professor, Sandip Institute of Technology and Research Centre, Nashik,India.

III. SYTEM ARCHITECTURE

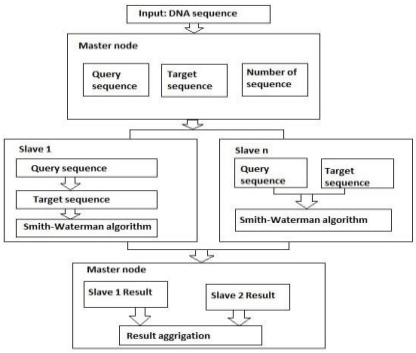

Fig: 1 Execution of Smith-Waterman

The purpose of this system is to accelerate the application of biological sequence alignment using distributed processing approach for finding optimal local and global alignment. Proposed system works on parallel distributed computing development the approach of sequential alignment is elementary that assists to work out the biological associations or relationships from a huge data-sets. To reduce computational processing time of sequence operation this system uses parallel distributed computing capabilities to get accurate and efficient implementation. Proposed system is working on HDFS framework where the processing is done using mapreduce term. In this paper we are processing the protein sequence datasets using Smith-waterman algorithm in structured format which will reduce the gap penalties. Dataset are distributed on number of nodes and query set is transformed on every system and checks whether it is matching or not. To reduce computational processing time of sequence operation this system uses parallel distributed computing capabilities to get accurate and efficient implementation. We are using structured as well as unstructured data for matching the sequences.

The query sequence are mapped on the n nodes and using threads the query is processed and as the query reaches the target sequence the sequence are reduced and result is generated accordingly. The master node maps all the sequences, the actual processing is performed on a distributed system and then a reduced result is generated of a large data as stated. To reduce computational processing time of sequence operation this system uses parallel distributed computing capabilities to get accurate and efficient implementation.

In this query processing which is the map function, the queries are processed as they are divided into the sub-sequences and if there is duplication of the query the data is aggregated.

After the entire query processing the data is reduced in the form of the final reduced result which is according to the desired outcome of the system.

A dynamic programming methodology applied to complicated downside(large downside/data) to resolve that downside and obtain optimum answer by dividing the matter into tiny sub problem then and also the answer for every sub problem. This algorithm is divides into three steps that is initialization of Dynamic programming matrix, all the matrix and optimally trace back the matrix to optimal local and global alignment. Using this algorithm it will be easy to predict the disease causing to the human; as it will lead to the prior precaution to the patients.

This is possible by using the prescription given by the doctor to the patient and also its history and its medical reports. The goal of the system is to accelerate the application of biological sequence alignment using the distributed approach using Hadoop for managing the large data. When the sequences are processed, the score matrix is generated accordingly.

These sequences are firstly matched with the CPU which takes more time in processing, than the sequences are matched on OpenMP which takes less processing time as compared to CPU. Finally the sequences are matched on Hadoop system and processed so the result is generated within very less time(ms). In very few micro seconds the result is generated on a distributed.

IV. SYSTEM ANALYSIS

The proposed system states the large data sets divided into n query processing. To reduce computational processing time of sequence operation this system uses parallel distributed computing capabilities to get accurate and efficient implementation. The master node includes the query processing, target history of species. Using this algorithm it will be easy to predict the disease causing to the human; as it will lead to the prior precaution to the patients and also the patient who is not in the same town can be treated by the distributed approach. Than all the sequences are compared with CPU, OpenMP and Hadoop to check its performance and speed up

V.RESULTS AND DISCUSSION

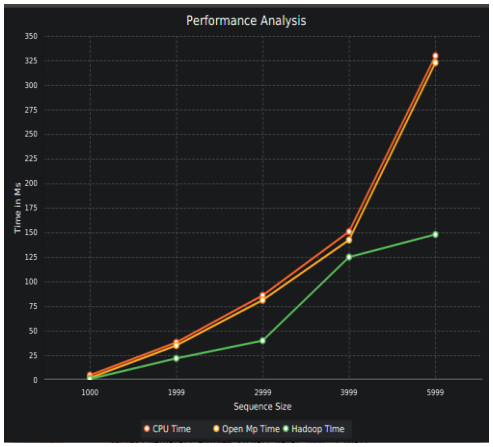

Fig. 2 Comparison of the system

Fig. 2 states the dataset details which I have calculated till 6000 sequences. The sequences are calculated as per the given no of sequence and the result is measured in (ms). In figure it is clearly stated time taken by CPU is greater as compared to OpenMP and Hadoop.

It also states the performance of the system is calculated by comparing the query with CPU, OpenMP and Hadoop system. Operation is performed on all the three system. We can see that time taken by OpenMP is bit less than CPU and Hadoop takes less time than both CPU and OpenMP to process the query sequence.

VI. CONCLUSION

The Distributed scan approach can be used for aligning long biologial sequences. This type of analysis can accelerate the alignment analysis using HADOOP for the domains related to Bioinformnatics. The Smith Waterman rule performs native sequence alignment that's, for determinative similar regions of the sequences.

REFERENCES

1. Rubio-Largo, Miguel A. Vega-Rodr iguez, David L. Gonz alez- Alvarez, ”A Hybrid Multiobjective Memetic Metaheuristic for Multiple Sequence Alignment”IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, VOL. XX,NO.XX,APRIL 2015.
2. Wael Abou El-Wafa, Hesham F. A.Hamed, Asmaa G.Seliem ”acceleration of Smith-Waterman Algorithm for short read DNA Alignment Using FPGA ”2016 IEEE 40th Annual Computer Software and Applications Conference.
3. Heba Khaled 1,H.M. Faheem 1,2, Mahmoud fayez 1,2, Iyad Katib 3 and Naif R.Aljohani, ”Performance Improvement of the Parallel Smith Water-man Algorithm Im-plementation Using Hybrid MPI Openmp Mode”, SAI Computing Conference 2016 July 13-15, 2016 London, UK.Heba Khaled 1,H.M. Faheem 1,2, Mahmoud fayez 1,2, Iyad Katib 3 and Naif R.Aljohani, ”Performance Improvement of the Parallel Smith Water-man Algorithm Im-plementation Using Hybrid MPI Openmp Mode”, SAI Computing Conference 2016 July 13-15, 2016 London, UK.
4. Huazheng Zhu, Zhongshi He, and Yuanyuan Jia, ”A Novel Approach toMultiple Sequence Alignment Using Multi-objective Evolutionary Algorithm Based on De- composition” 2168-2194 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.Technical Report CS-TR-4926, Univ. of Maryland,College Park, July 2008.
5. Giovanni Causapruno, Gianvito Urgese, Marco Vacca, Mariagrazia-Graziano, Member, IEEE, and Maurizio Zamboni, ”Protein Alignment Systolic Array Throughput Optimization”IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS 1063-8210 2014
6. M. Cottle, W. Hoover, S. Kanwal, M. Kohn, T. Strome, and N. W. Treis-ter,”Transforming Health Care Through Big Data, Institute for Health Technology Transformation” Washington DC, USA, 2013.
7. Y. Liu, B. Schmidt, and D. L. Maskell, MSAProbs, ”multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilitiesBioinformatics, vol. 26, no. 16, pp. 19581964, 2010.
8. A. R. Subramanian, M. Kaufmann, and B. Morgenstern, ”DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment Algorithms for Molecular Biology, vol. 3:6, 2008.
9. C. Notredame, ”Recent progresses in multiple sequence alignment”, a survey, Phar-macogenomics, vol. 3, no. 1, pp. 131144, 2002.
10. T. F. Smith and M. S. Waterman, ”An Improved Algorithm for Matching Biological Sequences” J. ioZ,, Biol. (1982) 162, 705-708.

AUTHORS PROFILE

Sneha Arjun Khaire, Assistant Professor of Sandip Instituteof Technology and Research Centre, Nashik.