

ABSTRACT

In Biological analysis, 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. 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. Today’s existing freeware tools produce accurate sequence alignment but these tools performance relatively low while considering big data. This paper addresses time efficient DNA sequence alignment methodology using sahara hadoop cluster in openstack cloud improves execution speed and accuracy. Virtual cluster environment denotes resource availability highly possible. In addition, proposed system performs the sequence alignment very quickly due to parallel and distributed processing aspects of hadoop. Other than above mentioned features, the proposed methodology highly suited for large scale DNA sequence alignment. A final result indicates time efficiency of proposed system.

KEYWORDS DNA Sequence alignment, homology, Sahara Hadoop, Openstack, Virtual cluster.

INTRODUCTION

Today, big data is important term in information science. The omics fields of life science includes voluminous amount of data in many forms used to represent various levels of biological data that includes genome data, epigenome data, proteome data and transcriptomic data of different people. The biological data becomes too big and available in peta bytes and exa bytes. This data produce large amount of meaningful values. It introduces many challenges such as handling of complex information; integration of heterogeneous resources; analysis on big data. Big data analytics is one of the most booming markets. When Google search engine launched image search feature, it had indexed more than 300 million images. In every minute so many video content are uploaded in YouTube. Twitter handles millions of tweets per day. Face book users update their wall in every minute. Search engines logging 600 million queries daily. There are different data centers where people can store vast amount of data, such as IBM Server, EMC Server etc. On the other hand AWS (Amazon Web Services) provide a host of services to store, process and analyze the data at scale in a cost effective manner. Big data term refers collection of large datasets that are distributed, multi dimensional and complex that it becomes difficult to processing on hand traditional data processing applications [3].

Unstructured data includes image files, text files, audio files and video files. In general online social networks generates large amount unstructured and semi structured data. Relational Database Management System contains structured data [6]. The three dimensions of big data are volume, velocity and variety [5]. The following areas are identified as most challenging in big data: Data storage, Analytics, Security and Privacy. Data storage includes relational databases and NoSQL related processing aspects. In big data analytics machine learning plays major role [4].

RELATED WORK

Needleman-Wunsch [18] and Smith-Waterman [19] are the widely used sequence alignment algorithms. 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 [20], and T-COFFEE [21] tools are uses heuristic algorithms. It uses progressive approximation method. These tools identify similar sequence in very fast. ClustalW tool uses profile to profile alignment used to represent probability of sequence in particular position. This gives better results compare to previous one. Sequence alignment score improved using iteration based approach. Hidden Markov Model (HMM) [22], Generic algorithms [23] 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 [24] is the best example for structure based alignments using priori data. It uses baseisan alignment for its alignment purpose. Basic Local Alignment Search Tool (BLAST) [25] used for sequence to sequence alignment. PCI-BLAST [26] is used for profile to sequence alignment. Align-M [27] is example for non progressive approach that results gives better accuracy for the sequence aligned distantly. Probalign [28] calculate the probability of pair wise using posteriori information. MAFFT [29] and MUSCLE [30] tools also uses posteriori information. NRAlign [31] uses horizontal alignment information to give accurate result. Gap penalty plays major role in alignment process. DIALIGN-T produces high accuracy in terms of gaps. ISP Align [32] is combined HMM and Probcons that identify the sequence using the intermediate sequence profile to assure the accuracy. Using fast fourier transform MAFFT [33] reduces the CPU Execution time to identify similar regions. It uses iterative base approach to calculate the result. PartTree [34] used to create guide tree for constructing sequence alignment. Partial order alignment used to achieve improved result in sequence alignment. Computational complexity of progressive methods improved using grammar based sequence alignment distance calculation. Guide tree that denote collection of sequence
alignment constructed using Simultaneous Alignment and Tree Construction using Hidden Markov mOdels (SATCHMO). Local sequence alignment algorithm need high cost for computation. Now a days FGPA [35] is used for improving the computation. Relative alignments identified using COACH. Higher end machines needed to improve the performance of the system [36]. To search the alignment in homology database needs parallel architecture supported machines [36].

CEPAR [37] perform the biological related processing using parallel processor. pBLAST [38] support parallel processor, hash table and query processing to achieve computational accuracy. Sequence alignment considered as NP-Hard problem. Genetic programming gives better results compared to dynamic programming Distributed and parallel environment support achieved in ClustalW-MPI and W.ND-CLAST [39].

BAliBASE [40] 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. Hadoop [41] will replace the higher computation power with the help of available commodity machines. Hadoop environment easily integrated with dynamic programming suited for large scale computation problem. But hadoop multi cluster environment needs physical hardware to provide parallel and distributed computing environment. Proposed the dynamic programming with improved gap penalty in virtual cluster. Virtual cluster management completely achieved using analytics as a service of openstack cloud environment.

Openstack

Openstack is open source cloud computing platform. Openstack is example for Infrastructure provisioning model. This platform contains various components to control the hardware, network and storages. Openstack support Single node setup as well as multi node setup. Multi node setup improves the horizontal scalability. Openstack consists of Nova, Keystone, Swift, Glance, Cinder, Heat, Neutron, Horizon, Sahara, and Trove. Cloud environment with openstack provides Infrastructure as a Service, VPN as a Service, Database as a Service, Firewall as a Service, Load balancing as a Service, Storage as a Service and Analytics as a Service. IaaS service provided with the help of Nova, Glance, Cinder and Keystone. Analytics as a Service provides using Sahara component. Horizon component gives best Graphical User Interface for administrator. VPN as a Service, Load balancing as a Service, Network as a Service achieved using Neutron component. Swift provides Storage as a Service. Swift provides object storage where as Cinder provides block level storage. Ceilometer is the component used to monitor the cloud environment. Heat used to provide orchestration service. Openstack include featured components such as Zaqar, Ironic, Designate, Barbican, Manila etc. Openstack have so many releases like Icehouse, Juno, etc. Recent release of openstack is Kilo.

Analytics as a Service: Sahara

Sahara project of openstack provides Analytics platform as a service on the top of cloud environment using virtual resources. The first name of this project is Savana now it renamed to Sahara due to trademark problems. Sahara on red hat platform provides elastic data processing. Analytics as a service of openstack closely similar to Elastic Map Reduce service of Amazon web service. Data processing support achieved with the active participation various data processing community such as Apache, Hortonworks and Red hat. Sahara architecture includes hadoop plugin, hortonworks plugin, apache spark plugin and simple data processing plugin. And also it gives efficient cluster management with virtual resources. In this work, time efficient DNA sequence alignment on biological data analysis performed in virtual cluster environment using sahara.

Hadoop

Sahara Component includes hadoop plugin for big data processing. Apaches first-class project hadoop is a platform that provides parallel and distributed processing aspects. The first version of hadoop framework named as Nutch, an open source search engine and crawler. After the successful implementation of Google File system (GFS), parallel processing with Map Reduce Model on Nutch gives new era Apache hadoop. This framework mainly used to analyze large amount of raw dataset. It follows distributed master-slave architecture. Hadoop framework consists of Map Reduce for parallel processing and Hadoop Distributed File System (HDFS) for storage. HDFS [1] is optimized for high throughput when reading and writing huge files and have own library to access and its logical representation have name node and data node. The high-tech companies like Facebook, Yahoo, Twitter and eBay uses hadoop. Facebook uses Hadoop, Hive and HBase for data warehousing. Twitter uses Hadoop, Pig and HBase for data analytics, visualization and social graph analysis. Yahoo uses this framework for data analytics, machine learning and search ranking. Now Microsoft is also starting to work with hortonworks to ensure that hadoop works on its platform.

Mapreduce

Map Reduce is one of the heart components of Apache Hadoop, which is used to process huge amount of data in parallel with the help of commodity machines. 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 [7]. Reducer function receives collection of inputs from Mapper function, computing the result from input sets and gives as final output [8]. Apache hadoop is the best tool for processing large amount of database. Hadoop is distributed and column oriented database. HBase uses HDFS for its more efficient system storage [9]. Map Reduce is the programming model of Apache hadoop project. Map Reduce function complete tons of unstructured data within short period. The combiner function performs the partial operation of Reducer function. The shuffle phase of programming model shuffles the output of Mapper function using Map Reduce library. The Mapper function takes key-value pairs as input values and it gives collection of key-value pairs as output values [6]. The intermediate key-value pairs are stored in local disk of data node. Reducer function is user defined function that
reduces the given input collections based on constraints specified in code [10]. The input data stored in HDFS. InputFormat class available in new mapreduce API used to divide the given input block into multiple blocks. Execution time depends on shuffle phase [11]. Recordreader class used to accept the data in Mapper function. Map Reduce execution includes following steps [6].

- Input request passed in the form of jar file which contains driver code, Mapper code and Reducer code.
- Job tracker in name node assign the Mapper tasks by tracking the logic from requested jar file to all the available task trackers of data nodes.
- Once all the task trackers are done with Mapper processes, they send the result status back to job tracker.
- After that job tracker present in data node initiates sort and shuffle phase on Mapper module output.
- After the completion of sort and shuffle, job tracker initiates reducer phase on all available task trackers.
- Once all task trackers of data nodes do with reducer phase, they update the status back to the job tracker.

Mapper and Reducer are called user driven phases. Mapper class output is stored in “part-m-000000” file. Actually “part-m-000000” file contains intermediate values. Similarly Reducer class output stored in “part-r-000000”.

**PROPOSED SYSTEM**

The proposed system used for DNA sequence alignment using dynamic programming concept with the help of virtual resources. From the literature survey, global sequence alignment achieved with Needleman-Wunsch alignment algorithm and affine gap penalty. Because Needleman-Wunsch alignment algorithm uses dynamic programming and it generate accurate result. Pair wise alignment achieved using Mapper and Reducer function present in sahara hadoop cluster. Here hadoop cluster framed using virtual machines. Multiple sequence alignment also performed using two different Map Reduce function. Multiple Mapper and Reducer function used in sequence alignment process perform all possible combination between the sequences.

**Needleman-Wunsch Algorithm**

Needleman-Wunsch algorithm produces optimal alignment using entire regions, therefore this kind of alignment considered as global alignment. It finds the alignment using number of matching sequence and gap penalty. This algorithm does not used to identify the similarity in local regions. Needleman-Wunsch algorithm used for DNA sequence alignment and protein sequence alignment. This is the first algorithm that uses dynamic programming concept in biological sequence comparison. In dynamic programming concept complete problem divided into independent sub problems. Sub problems are solved independently. From the smaller sub problem solution reconstruction made to obtain the complete large problem solution. This algorithm works based on dynamic programming. It is also called as optimal alignment algorithm.

In pairwise alignment algorithm, dynamic programming works from end of the sequence and perform matching based on scoring system with number of matches, mismatches and gap penalty. The result includes all possible alignment score. The highest alignment score denotes optimal alignment between the two different sequences. Dynamic programming follows recursive procedure. Needleman-Wunsch uses sore matrix to obtain a global alignment. Final score matrix S contains x number of rows and y number of columns. The variables x and y are total size of two different sequences. The score matrix formed with the help of equation 2. The actual alignment formed using matrix P which used to store the movement by backtracking method. Movement operation performed to calculate the optimal alignment score. In the proposed system for gap calculation performed using affine gap penalty. Needleman-Wunsch with affine gap function needs three more matrices T, U, and V. T is used to store the score of alignment when x_n aligned to y_n. Matrix U used to store the score of x_m aligned to a gap after the y_n. V used store y_n aligned to a gap after the x_m. The parameters m and n is the positive value used to identify the values in any matrix. S_mn contains the maximum value three different values such as T_mn, U_mn and V_mn.

In affine gap penalty calculation, two cases are considered for opening of gap and continuation of existing gap.

- Case 1: (Gap opening) U_{m+1,n} = S_{m,n} - o
- Case 2: (Gap Extension) U_{m+1,n} = S_{m,n} - e

Initialization of dynamic programming score matrix is done as follows:

\[ V_{m0} = o + (m-1) \times e \]
\[ V_{0n} = o + (n-1) \times e \]

The complete matrix formed using following equations. Maximum value of S_mn from T_mn, U_mn and V_mn denotes optimal score of alignment.

\[ S_{mn} = \max \{ T_{mn}, U_{mn}, V_{mn} \} \]
\[ T_{mn} = S_{m-1,n} + \text{score} \times (X_{mn}) \]
\[ V_{mn} = \max \{ S_{mn-1} - o, S_{mn-1} - e \} \]
\[ V_{mn} = \max \{ S_{mn-1} - o, S_{mn-1} - e \} \]

Total alignment score calculated using three moves on final score matrix. If S_mn consist of T then movement performed in diagonal. On the other hand, If S_mn consists of U_mn then movement performed in upwards i.e movement in column wise. If S_mn consist of V_mn then movement performed towards left i.e movement in row wise. For example consider two different sequences Sequence 1 = “GCTG” and Sequence2 = “GTG”.

Here gap opening penalty = -11 and gap extension penalty = -1. Final score matrix S_{0k} generated based on affine gap penalty gives alignment score as 6. This considered as optimal alignment score. The globally aligned sequences are “GCTG” and “G-TG”. Hence trace matrix moves are S_{11}→S_{12}→S_{11}.

Needleman-Wunsch global sequence alignment algorithm with affine gap penalty takes O (kMN) time to complete the pairwise sequence alignment process. The variable k represents number of matrix used for calculation. For example above computation done using four matrices therefore k=4. While
considering multiple sequence alignment, time takes to complete the n sequences is \( O(L^3) \) where as L is average length of multiple sequences chosen for computation. The above algorithm suited for parallel processing environment. Parallel implementation of above algorithm takes \( O(n^2L) \). Biovall is the best example for Needleman – Wunsch algorithm in parallel implementation. The main drawback of Biovall is large computation environment. Alchemi Grid engine provide fast computation for Needle-Wunsch algorithm. Other than above mentioned algorithms there are two parallel computation models available to analyze the biological sequences. First method used to provide homology study that give clues for structure. Second method used to predict structure of sequence that aligns homogenous sequences. For biological sequence analysis process fine grained multithreading considered preferred type while using dynamic programming.

**Affine Gap Penalty**

Scoring system for aligning a genetic code in bioinformatics is called as gap penalty. In DNA sequence Gaps created while insertion or deletion process. This process occurs due to mutation, meiosis, mispairing and translocation. Gaps represented using dashed lines in the sequence. Alignment score used to identify best possible alignment. Affine gap penalty is widely used method in scoring. It is the combined model of constant and linear scoring methods. The following equation represents mathematical form of affine gap penalty.

\[
\text{Gap} = o + (1 - 1) \cdot e
\]

In the above equation \( l \) is length of gap, \( o \) is openness i.e. gap opening penalty. It must be positive constant. \( e \) is called gap extension positive constant. Gap opening penalty assigned for first gap. Extension penalty allocated for subsequent gap in the sequence. Gap size depends on parameters openness and extension. Logarithmic gaps handled with the help of convex gap penalty method. Time complexity of affine gap penalty model is \( O(MN) \) where \( M \) and \( N \) denotes sequence length. Convex gap penalty model takes \( O(MN \ln(M+N)) \) to calculate the alignment score. The proposed system uses affine gap penalty to calculate alignment score.

**Sequence Alignment Using Sahara**

In this proposed system complete sequence alignment performed on the top of openstack cloud platform. Complete sequence alignment achieved using apache hadoop 2.6.0 plugin included in the sahara data processing architecture. Apache hadoop 2.6.0 supports YARN framework on the top of virtual cluster resources. Sequence alignment mapper function and reducer function deployed on hadoop VM. Multiple hadoop VMs are available in sahara environment and implemented mapper and reducer function for pairwise sequence alignment. Separate mapper and reducer function implemented for multiple sequence alignment. In multiple sequence alignment all possible combinations analyzed using two different functions. Input and output sequences are stored in hadoop distributed file system. Multiple mapper function with single reducer used to compare the sequences from multiple data source. Map reduce framework of hadoop provide efficient parallel processing for biological sequence analysis. Mapper function contains code for Needleman – Wunsch with affine gap penalty function. Reducer function compare all possible alignment score based on that final optimal alignment score identified. And implemented permutation function for sequence alignment in parallel. Number of permutation used for the alignment process is \( n! \). Time complexity of individual Mapper and Reducer function for two different sequences with \( M \) and \( N \) length is \( O(4^M*N) \). The number 4 denotes number of matrix (**M** × **N**) that does not change with sequence length. Block size indirectly proportional to sequence size. Therefore alignment time decreases while block size increases. Sequence alignment

**RESULT AND DISCUSSION**

Openstack environment established on IBM X3400 M3 server with 8GB RAM and Intel Xeon Processor. Time efficiency tested in terms of sequence size and block size. Various data analyzed and output readings were considered for calculation.
performed in parallel. In addition pairwise alignment of multiple sequence alignment performed in multiple blocks. The chart represented in Figure 2 plotted with block size as X-Axis and Time to complete the alignment in seconds as Y-Axis.

**CONCLUSION AND FUTURE WORK**

This paper addresses novel environment to analyze voluminous biological data. The proposed system achieved with the help of openstack and hadoop. Openstack sahara support virtual cluster management. Analyzing data with Apache hadoop is easy and effective manner. Apache hadoop is designed to run in a distributed environment and it manages the collection of various nodes running map and reduce function. In this system time efficient data analysis performed on variety of biological data. The proposed system executed with dynamic parallelism to prove accuracy. Time complexity of proposed method is \( O \left(4^M\binom{N}{Q}\right) \). Due to scalability nature of hadoop VM proposed system highly supported for large scale problems. In future, same Mapper and Reducer class implemented with openstack sahara using apache spark plugin to improve the parallelism.

**REFERENCES**

10. Highland et al, “Fitting the problem to the paradigm : Algorithm characteristics required for effective use of Map Reduce”,In the proceedings of Missouri University of Science, 2012.
of the multiple sequence alignment benchmark”,
41. http://hadoop.apache.org/core/docs/r0.17.2 Apache
Hadoop official documentation, 2002.