International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 7, July - 2013

Hadoop Cloud Application In DNA Alignment And Comparison

Subhendu Bhusan Rout,
Department of CSE&A
IGIT Sarang
Dhenkanal, Odisha, India

Bhabani Sankar Prasad Mishra,
School of Computer Engineering
KIIT University
Bhubaneswar, Odisha, India

Satchidananda Dehury
Department of Systems Engineering
AJOU University
Suwon, Republic of Korea

Abstract

Now a days Bioinformatics is a good and a very upcoming technology for the recent researchers. Alignment and comparison of DNA and RNA, Gene mapping on chromosomes, Protein structure prediction, gene finding from DNA sequences are various useful tasks of bioinformatics. In recent years many techniques are being used for the DNA Alignment and Comparison. In the field of biology DNA Alignment and Comparison plays a vital role for the development a new drug. For the development of new drugs or medicines Alignment and comparison of DNA and RNA, Gene mapping on chromosomes, Protein structure prediction, gene finding from DNA sequences are very necessary parameters. Recently the Bioinformatics industry is in the fledgling condition and gaining more attention of researchers. Various Soft Computing methods like Artificial neural network, Fuzzy Logic, Genetic algorithms, Swarm optimization, etc are used for this purpose to distinguish, compare or process the various DNA, RNA Particles. It is always a big task for researchers to develop new tools and methods for the purpose of processing of data as well as development of drugs. In the field of cloud computing, Apache hadoop is a good platform which is suitable for processing huge amount of data. Till now Apache hadpoop is having the similar type of application in Facebook & Yahoo. Fault tolerance is property which is very useful for the security of data and information. In this paper we have discussed about the application of cloud technologies in DNA Alignment and Comparison and as a real time application we have discussed about the Apache hadoop that can be applied for this purpose also. We have proposed a technique, which will work for huge amount of data and for the DNA Alignment and Comparison in a large scale. It will be helpful for the recent medicine researchers to develop various drugs after study and analyzing the changes in the DNA structure.

Key Words: *Bioinformatics, Cloud computing, Gene mapping, Protein structure prediction, Apache hadoop, DNA Alignment.*

I. Introduction

Cloud computing is an internet based computing of shared resources, databases, and soft wares etc, which are generally provided over Internet with an on demand basis like an electricity grid. One can access any of the resources that live in the cloud across the Internet and don't have to worry about computing capacity, bandwidth, storage, security, and reliability. The advantages of cloud computing over traditional computing include: agility, lower entry cost, device independency, location independency, and scalability. As the popularity of Internet is growing day by day, so many applications from different sides of the world can be combined through internet to process, gather or to integrate for a particular research. To develop a drug it need a high scale research of different chromosomes, DNA, RNA etc. In order to process the huge amount of data it

needs a good platform that will serve dedicative for the simulation of data or simply for research purpose. The platform should be unbiased and error free. The cloud computing is such a good technology for this purpose.

The research upon DNA & RNA is a major project all over the world for scientist as well as researchers. In bioinformatics, a DNA & RNA sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. Aligned sequences of nucleotide or amino acid residues are typically represented as rows within a matrix. Gaps are inserted between the residues so that identical or similar characters are aligned in successive columns. There are various ways to compare and alignment these particles. Alignments are commonly represented both graphically and in text format. In almost all sequence alignment representations, sequences are presented in the forms of arrangement of rows so that aligned residues appear in successive columns. In text formats, aligned columns containing identical or similar characters are presented with a system of conservation symbols. If we represent these in an image, then an asterisk or pipe symbol can be used to show identity between two columns; other less common symbols like a colon for conservative substitutions and a period for semi conservative substitutions. Many sequence visualization programs also use color to display information about the properties of the individual sequence elements; in DNA and RNA sequences, this equates to assigning each nucleotide its own color. In Protein Structure color is often used to indicate amino acid properties to aid in judging the conservation of a given amino acid substitution[1].

The Apache Hadoop is a good technology for the Processing and sharing of data upon a common platform. In this approach of cloud computing project develops open-source software for reliable, scalable, distributed computing. The Apache Hadoop software library is a framework that allows for the distributed processing of large amount of data sets across clusters of computers using a simple programming model. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Not only depending upon hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures. Not only in a single cluster it may process data from multiple clusters that may situate in several area of the globe through internet. Researcher from different clusters may share distribute or communicate their research work and knowledge over the network. It can possible only for the capability of handling large amount of data.

In this paper in section II we have discussed briefly about the cloud computing with all about Service Oriented Architecture and Apache hadoop in subsection 'A' and 'B'. In section III we have provide a brief idea about DNA, RNA Alignment & Comparison. In section IV of this paper we have proposed an idea about the application of cloud computing in the field of DNA, RNA Alignment & Comparison as well as a proposed example for real time application of apache Hadoop in processing of huge amount of data. Finally the paper concludes in section V.

II. Cloud Computing: A Brief Introduction

Cloud Computing is being a suitable platform for the processing of huge amount of data in a computer fields for various applications. It has emerged as a computing infrastructure that enables rapid delivery of computing resources as a utility in a dynamically scalable, virtualized manner. There are various advantages of cloud computing over traditional computing which include: agility, lower entry cost, device independency, location independency, and scalability. There are many cloud computing initiatives from IT giants such as Microsoft, IBM, Google, Amazon as well as start-ups such as Parascale, Elastra and Appirio. Various Services can be access by users without having the necessary requirements. Basically there are three types of resources that can be shared and consumed over the Internet. They can be shared among users by leveraging economy of scale. One of the major objectives of Cloud Computing is to leverage Internet for various applications and to provision resources to users [2]. The three type of resources that can be consumed by means of cloud computing is

- Infrastructure as a service
- Platform as a service
- Software as a service

A. *Service Oriented Architecture*

SOA and cloud computing are related, specifically, SOA is an architectural pattern that guides business solutions to create, organize and reuse its computing components, while cloud

computing is a set of enabling technology that services a bigger environment. The distinction from current cloud implementations is that the cloud computing resources in SOCCA are componentized into independent services such as Storage Service, Computing Service and Communication Service, with open-standardized interfaces, so they can be combined with services from other cloud providers to build a cross-platform virtual computer on the clouds. In other words, SOA and cloud computing will coexist, complement, and support each other. There have been several initiatives at attempting bridging SOA and cloud computing. Cloud providers might not confirm to the standards rigidly; they might also have implemented extra features that are not included in the standards. Cloud Ontology Mapping Layer exists to mask the differences among the different individual cloud providers and it can help the migration of cloud application from one cloud to another[2].

Multi-tenancy architecture is also a very good technology which has the capability to handle Single application Instance and Multiple service instances. According to SOCCA, it supports Single application instance and multiple service instances. The motivation behind this pattern is that the workloads are often not distributed evenly among application components, and the performance of the single application instance is limited by the application components having lower throughput. By using multitenant architecture a user can get multiple services in a single application that services can be a single cloud or from multiple clouds. If a particular service is not available form one resource at a time, than from another source this service can be provided to the customer. Better scalability is not only the benefit from this Single Application Instance and Multiple Service Instance pattern but easy customizability is another gain of service[2]. If at a time according to the list of demand services a service is not in the service instance, it can be easily plugged into the existing service instances group from another. This may the application of fault tolerance system.

B. Apache Hadoop: A Cloud Computing Approach

Apache hadoop is an application of cloud computing in the application of data processing of huge and similar type of data. Now a day it has the applications in the major data clusters & search engines like Facebook, Google, Yahoo, etc. Hadoop Distributed File System (HDFS) is a distributed file system that provides high-throughput access to application data. Hadoop has the good technology like fault tolerance capacity. In August 2010 hadoop applied to worlds largest data cluster Facebook [3]. Now Hadoop is used in searching machines like yahoo, amazon, zvents etc. It has the application like log processing in case of Facebook, Yahoo, ContextWeb. Joost, Last.fm etc. Not only these the Hadoop technology has the applications in the field of data warehousing & processing applications as well as video and image analysis in the field of Facebook, AOL and New York Times, Eyealike respectively. In these field it can able to process a huge amount of data that comes from many users or simply many part of the globe. It has the capability to handle, retrieve or manipulate huge amount of data.

In case of distributed file system there is a single namespace for the entire cluster. As the largest data cluster like Facebook In this method data co-herency is available so that once writing many nodes can able to read it. Clients can only append to existing files. Files are broken in to typically 128-256 of block size & each block can replicated on multiple data nodes. Clients can find location of various blocks that are available & clients can access data directly from data nodes. The HDFS (Hadoop distributed file system) is having 10K nodes, 1 billion files, 100pb of data. The files are replicated to handle hardware failure and by detecting failures it recover from them. It is optimized for batch processing system i.e. Data locations exposed so that computations can move to where data resides which works with high bandwidth also. It is very user friendly and runs on heterogeneous operating system.

Now a day's more than 500 million active facebook users generate & share 30billion pieces of content every month. As a statistics 20 TB of compressed new data added per day with 3 PB of compressed data scanned per day. After all 480K compute hours is spending per day. In some cases HDFS is used for the storage of online application form. In this way hadoop has a large scale application in the field of distributed operating as well as a capability of processing huge amount of data [3].

III. DNA, RNA Alignment & comparision

Biology is the science of living things. Living organisms are characterized by both diversity and unity. The evolutionary theory is originally developed in the nineteenth century and currently undergoing a renaissance of deeper understanding that helps us to pinpoint the mechanisms, which lead to the amazing diversity we find among living beings. We are using biology science in various ways and

multitudes of causes. This is for various problems regarding living bodies or living cells.

The Genetic Theory is one of the basic principles of biology. The main concept of this theory is that traits are passed from parents to offspring through gene transmission. Genes are located on chromosomes and consist of DNA. They are passed from parent to offspring through reproduction. The principles that govern heredity were introduced by a Gregor Mendel in the 1860's. These principles are now called Mendel's law of segregation and law of independent assortment[4]. In this field Information Technology helps us in touch with this emerging scientific theory and related inventions. The researches that are from various clusters are connected by means of a network connectivity that associated with information technology.

```
AAB24882        TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT 60
AAB24881        --------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK 40
                                    ****: .***:  * *:** * :****.:* *******..

AAB24882        PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE-CNQCGKAFAQ- 116
AAB24881        HSHLQCHKRTHTGEKPYECNQCGKAFSQHGLLQRHKRTHTGEKPYMNVINMVKPLHNS 98
                 **** *:***********:***:**.: .***************    :  *.: :
```

Fig 1. A sequence alignment, produced by ClustalW, of two human zinc finger proteins, identified on the left by GenBank accession number. Key: Single letters: amino acids. Red: small, hydrophobic, aromatic, not Y. Blue: acidic. Magenta: basic. Green: hydroxyl, amine, amide, basic. Gray: others. "*": identical. ":": conserved substitutions (same colour group). ".": semi-conserved substitution (similar shapes)[1].

DNA & RNA sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. The fig 1 shows a sequence of alignment in the form of letters. When two sequences in an alignment share a common ancestor, mismatches can be interpreted as point mutations. The gaps may introduce in one or both lineages in the time since they diverged from one another. In sequence alignments of proteins, the degree of similarity between amino acids occupying a particular position in the sequence can be interpreted as a rough measure considering as how it can conserved a particular region or sequence motif is among lineages. If there is the absence of substitutions, or there is the presence of very conservative substitutions in a particular region then the sequence, suggest that in this region there is structural or functional importance. Although DNA and RNA nucleotide bases are more similar to each other than are amino acids, the conservation of base pairs can indicate a similar functional or structural role.

Pairwise sequence alignment methods are used to find the best-matching piecewise local and global alignments of two query sequences. It is the possibility of Pairwise alignments onlyd between two sequences at a time, but they are efficient to calculate and are often used for methods that do not require extreme precision like searching a database for sequences with high similarity to a query. The three basic primary methods of producing pairwise alignments are dot-matrix methods, dynamic programming, and word methods. However, multiple sequence alignment techniques can also align pairs of sequences. Although each method has its individual strengths and weaknesses, all three pairwise methods have difficulty with highly repetitive sequences of low information content, especially where the number of repetitions differ in the two sequences to be aligned. The maximum unique match' (MUM) is the way of quantifying the utility of a given pairwise alignment or the longest subsequence that occurs in both query sequence. Longer MUM sequences typically reflect closer relatedness. As these types of process or research take place with a huge amount with large number of clusters so a large number of DNA RNA sequence may come in this matter. A cost effective, efficient model is necessary for this purpose to identify the similarity from the structural, functional relationship of the living particles. So it needs to share data, information in a good platform for the development a final result. This result may

apply for the design of various drugs which will be useful for various genes in different clusters.

IV. DNA, RNA Alignment & comparision using cloud technology

Now Cloud technology is an emerging technology for various problems regarding data processing. Cloud computing is a internet based computing of shared resources, database, software etc in which one can access any of the resources that live in the globe across the Internet and don't have to worry about computing capacity, bandwidth, storage, security, and reliability of the system. Using cloud technologies a huge amount of data can be taken for a research. In a medical research it need to process data that come from different cluster with huge amount. By using cloud technologies many user can share data on a single platform so that a common result can be concluded. So in this case in a common platform huge amount of data, or a large number of chromosomes, or a huge amount of genomic data can be considered for the research.

For the development of a new drug it needs a lot of research & application that can be done upon the DNA, RNA or genes. Generally in order to develop a common result the research takes place in many places by different researcher. The researchers works on a common project with different data. For implementation of result the sharing of data, information, or the research result plays a major rule for the conclusion or simply for development of the drug. The consistent look on a DNA or RNA sequence or by applying a particular drug the researcher should study the behavior, the reaction, and the changes that happened to that particular Sequence. DNA Sequence comparison is nothing but the comparison of the structure of the DNA and to establish the relation between changes. The various developed drugs should be applied to various living cells, various chromosomes, or simply various DNA, RNA and genes before it is being commercially used for different living cells. After a high level of research upon the various particles and even after studying the reaction or changes, the drug should be designed according to that.

There are various cloud technologies like Apache Hadoop, Dryad-Linq which are used in this field for the Processing & Executing of large number of data. Till now Apache hadoop has several applications in the field of searching, log processing, data warehouse, video and image analysis etc. It has the real time application in facebook. Now a day more than thousand million active users share, upload, communicate with each other through facebook. Hadoop file system maintains data coherency that is once written it can be read by several users. In fig.2. We have discussed some features of HDFS. Hadoop distributed file system is a high fault tolerance system although in many place it may be a single point of failure. It has a query able data base which can be shared by many users or researchers. That has the open data format that is common to all users. Any people or user can share or update his data according to their own activity. It has a high quality database which never deletes any data which can be utilise for long term research.

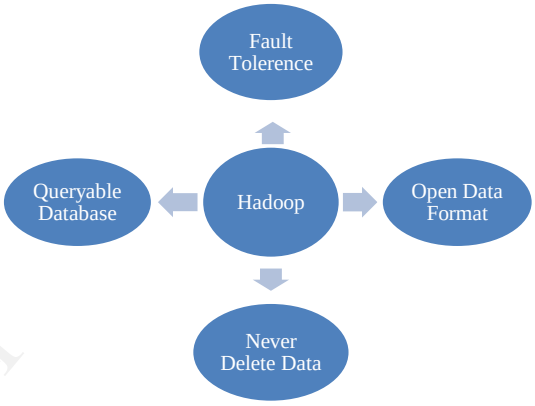

Fig. 2. Hadoop Distributed File System

cloud technologies can be used to various medical researches that take place with huge amount of data.DNA, RNA alignment and comparison is a way of arranging the sequence of DNA, RNA to findout the changes, similarity after the application of a drug. Though it is not such a complex task but the large amount of data increases the complexity of the research. Before development of a drug it is applied to various DNA, RNA sequence to study the behaviour, changes that happen to that particular drug. Not only a single cluster it is generally applied to various genes or chromosomes that comes from various area. It is not possible to keep data in a single place for the research. As in a single work many scientists or researcher works so here there is the necessity to compare, distinguish and to interpret data for the result. After a good testing and application the research conclusion may apply for the drug design. So in order to connect the entire researcher into a single platform it needs a good and error free platform so that a good result can be established. Cloud technologies are having the similar type of application in other fields. Basically Apache Hadoop is such a good technology in this filed which has the similar type of application. So for the development of a drug

many researcher can share & communicate there opinions on a common platform so that a common, efficient and cost effective result can be implemented.

V. Conclusion

Computers are always a very effective way of processing of huge amount of data. In order to establish a cost effective network between computers from different area cloud computing is always a good technology. It provides such a plat form so that a dedicative communication can be take place without worrying about computing capacity, bandwidth, storage, security, and reliability of the system. DNA, RNA alignment and comparison is the process of arranging the sequence of DNA, RNA to identify the similarity and differences between the sequences. This generally comes in a huge number from different clusters. In order to develop a effective result there should be a communication and sharing of data between these researchers. Cloud technologies like Apache hadoop will be helpful in this concept to develop a dedicated, error free, quarry able database for the research. Our future work will focus to wards other real time application of cloud technology upon this task.

ACKWNOLDGEMENT

We are very much thankful to wards the Faculty and staffs of IGIT Sarang for their cooperation and providing various facilities regarding this research work. We are also thankful to wards the anonymous reviewers for their valuable suggestions regarding this research work which improved the quality of the research paper.

REFERENCES

[1] http://en.wikipedia.org/wiki/Sequence_alignment [Last accessed February 2013]

[2] Wei-Tek Tsai, Xin Sun, Janaka Balasooriya, "Service-Oriented Cloud Computing Architecture", Seventh International conference on Information Technology, IEEE, pp. 684-689, 2010.

[3] Dhruba Borthakur, http://cloud.berkeley.edu/data/hdfs.pdf, "Apache Hadoop File System and its Usage in Facebook" Presented at UC Berkeley, 2011.

[4] http://www.biology.about.com/od/geneticsglossary/g/genetheory.htm, [Last Accessed January, 2013].

[5] http://www.en.wikipedia.org/wiki/Gene_mapping [Last accessed January 2013].

[6] Bauer M, Klau G, Reinert K, "Fast and Accurate Structural RNA Alignment by Progressive Lagrangian Optimization", Lecture Notes in Computer Science. Computational Life Sciences. Volume 395, 2005:217-228.

[7] Tim Wiegels, Stefan Bienert, Andrew E. Torda' "Fast alignment and comparison of RNA structures" Journal on Bioinformatics, OXFORD, 2012.

[8] Rajkumar Buyya, Chee Shin Yeo, "Cloud Computing and Emerging IT Platforms: Vision, Hype and Reality for Delivering Computing as the 5th Utility," Future Generation Computer Systems, pp. 599-616, 2009.

[9] Tinos, R., Yang, S. A self-organizing random immigrants genetic algorithm for dynamic optimization problems. Genetic Programming and Evolvable Machines, 8(3), pp. 255-286, 2007.

[10] Sarkis M., Diepold K., Westad F., "A new algorithm for gene mapping: Application of partial least squares regression with cross model validation", IEEE International Workshop on Genomic Signal Processing and Statistics, 2006.

[11] Dawy Z., Goebel B., Hagenauer J., Andreoli C., Meitinger T., Mueller J.C., "Gene mapping and marker clustering using Shannon's mutual information", Transactions on Computational Biology and Bioinformatics, IEEE/ACM , Vol-3(1) pp. 47-56, 2006.

[12] S. Mitra, R. Das, Y. Hayashi, "Genetic Networks and Soft Computing", IEEE/ACM Transactions on Computational Biology & Bioinformatics,Vol-8(1) pp. 616-635, 2011.

[13] Gong W, et al. Structure of pvu II DNA-(cytosine N4) methyltransferase, an example of domain permutation and protein fold assignment. Nucleic Acids Res 1997;25:2702–2715.

[14] S. Brunak, J. Engelbrecht, and S. Knudsen, "Prediction of human mRNA donor and acceptor sites from the DNAsequence," *J.Mol. Biol.*, vol. 220, pp. 49–65, 1991.

[15] Mathews DH, "Predicting a set of minimal free energy RNA secondary structures common to two sequences" Journal on Bioinformatics, Vol 21(1), pp.2246-2253,2005.