MAPPING OF GENES USING CLOUD TECHNOLOGIES

Subhendu Bhusan Rou[1], Sarojananda Mishra[2], Bhabani Sankar Prasad Mishra[3]

1 Dept. of Computer Science Engineering & Application, IGIT Sarang, Odisha, India, ***subhendu.as@gmail.com***
2 Dept. of Computer Science Engineering & Application, IGIT Sarang, Odisha, India, ***sarose.mishra@gmail.com***
3 School of Computer Engineering, KIIT University, Bhubaneswar, Odisha, India, ***bspmishrafcs@kiit.ac.in***

Abstract

Bioinformatics is a much updated topic for the recent researchers. There are various tasks of bioinformatics, like Alignment and comparison of DNA and RNA, Gene mapping on chromosomes, Protein structure prediction, gene finding from DNA sequences etc. Mapping of gene is the procedure of calculating the distance between the genes in chromosomes. In the real time application the medicine researchers goes for processing huge amount of data that may comes from different clusters or from different location. The mapping results also differ from each other with huge differences. In order to process a huge amount of data we need such a good platform which will serve without fail. Cloud computing is such a good technique that can be applied for various data as well as a huge amount of data to compare, distinguish, or simply for mapping the genes in chromosomes. In the field of cloud computing, Apache hadoop is a platform which provides a good platform for processing huge amount of data. Till now Apache hadoop is having the similar type of application in Face book & Yahoo. It carries the properties like fault tolerance which can be very useful in securing the data. In this paper we have discussed about the application of cloud technologies in Gene Mapping in chromosomes and as a real time application we have discussed about the Apache hadoop that can be applied for this purpose also.

Index Terms: *Bioinformatics, Cloud computing, Gene mapping, Protein structure prediction, Apache hadoop, Chromosome.*

---***---

1. INTRODUCTION

Cloud computing is an internet based computing of shared resources, databases, soft wares etc. These services 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 genes is a major project all over the world for scientist as well as researchers. It is the process of identifying the location of the genes in chromosomes. In order to do this there has to be a way to find the specific location of genes on each individual chromosome. There are three ways in which chromosomes are mapped. One way is to map a cytogenetic map in which chromosome bands, each representing 1 million to 5 million bases, are stained and the investigator finds a correlation between people who show a particular trait and exhibit a similar staining pattern. Another way is to produce a physical map using enzymes to cut pieces of DNA into fragments containing markers along with genes whose location is to be determined. By using computers to "walk" or overlay these fragments into their proper sequence we can produce a map of a long strand of DNA. The third technique is a method that has been used for the longest time that is mapping by crossover frequency.

Genes travel as packaged trains on chromosomes. During meiosis, chromosomes can do some fairly interesting things such as losing pieces (deletion), flipping sections up-side down (inversion), and not separating from their homologous partner when they are supposed to (non- disjunction). Crossover occurs when homologous chromosomes separate towards the end of the prophase, but are still attached at a few points along their lengths. It is during this attachment that these chromosomes can exchange pieces of their genetic instructions. The frequency of this crossover is directly related to the physical distance that genes are separated from each other on the same chromosome. Genes close to one another have a lower frequency of crossover than do genes farther apart. By keeping records of genetic experiments, we can calculate the crossover frequency; this being the number of times that gene behaves and moves in the chromosomes [1]. In this genetic mapping assignment, we have proposed a technique of applying cloud technology for the research purpose through which number of drug researchers can share their knowledge in a common platform. It will give more efficiency and good result in less time.

The Apache Hadoop 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 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. Rather than rely on 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 2 we have discussed briefly about the cloud computing with all about multitenant architecture and Apache hadoop in subsection ‘2.1’ and ‘2.2’. In section 3 we have provide a brief idea about gene, gene mapping, and some gene mapping technologies. In section 4 of this paper we have proposed an idea about the application of cloud computing in gene mapping 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 5.

2. CLOUD COMPUTING: A BRIEF INTRODUCTION

Clouds have emerged as a computing infrastructure that enables rapid delivery of computing resources as a utility in a dynamically scalable, virtualized manner. The advantages of cloud computing over traditional computing 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. 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

2.1 *Multi Tenant Architecture*

Multi-tenancy architecture has the capability to handle Single application Instance and Multiple service instances. According to SOCCA [2], 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. Moreover, to enhance scalability, we want to reduce unnecessary duplications as much as possible as opposed to Multiple Application Instances pattern.

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.

2.2 *Apache Hadoop: A Cloud Computing Approach*

Apache hadoop is a real time application of cloud computing. Now a day it has the applications in 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 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 as log processing in case of Facebook, Yahoo, ContextWeb. Joost, Last.fm etc. It has the 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. 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].

3. GENOME THEORY & GENE MAPPING

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 Gene 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]. This new Science topic of Genome theory is a revolution in the way, life is understood and in the way scientific information is available online. Information Technology helps us in touch with this emerging scientific theory, the researchers involved and related news stories. The researches that are established in various clusters are connected by means of a network connectivity which is associated with information technology.

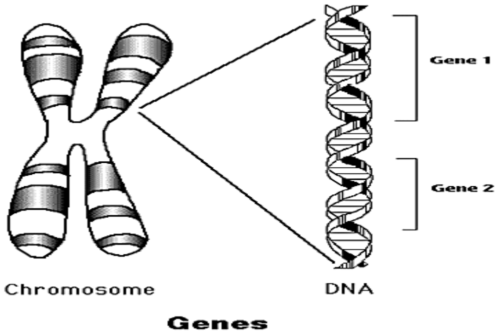

Fig 1 Gene structure in a chromosome

Gene mapping, also called genome mapping is the creation of a genetic map assigning DNA fragments to chromosomes. When a genome is first investigated, this map is nonexistent. The map improves with the scientific applications progress and becomes perfect when the genomic DNA sequencing of the species has been completed. In fig 1 it provides the gene structure in a chromosome. If we want to map between two genes we have to distinguish between the genes in the chromosomes. During this process, and for the investigation of differences in strain, the fragments are identified by small tags. These may be genetic markers or the unique sequence dependent pattern of DNA-cutting enzymes. The ordering is taken from genetic observations for these markers. Mapping is used in two different but related contexts. Two different ways of mapping are distinguished.

Genetic mapping uses classical genetic techniques to determine sequence features within a genome. Using modern molecular biology techniques for the same purpose is usually referred to as physical mapping. In physical mapping, the DNA is cut by a restriction enzyme. Once cut, the DNA fragments are separated by electrophoresis. The resulting pattern of DNA migration is used to identify what stretch of DNA is in the clone. By analysing the fingerprints, contigs are assembled by automated or manual means into overlapping DNA stretches. Macro restriction is a type of physical mapping where in the high molecular weight DNA is digested with a restriction enzyme having a low number of restriction sites. Once the map is determined, the clones can be used as a resource to efficiently contain large stretches of the genome. This type of mapping is more accurate than genetic maps. Genes can be mapped prior to the complete sequencing by independent approaches like in situ hybridization [5].

The process to identify a genetic element that signs partially or fully for a disease is also referred to as mapping. If the locus in which the search is performed is already considerably constrained, the search is called the "fine-mapping" of a gene. This information is derived from the investigation of disease-manifestations in large families (Genetic linkage) or from populations-based genetic association studies. As these types of process or research take place with a huge amount with large number of clusters so it needs to share data, information in order to develop a final result.

4. APPLICATION OF CLOUD COMPUTING IN GENE MAPPING

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 cloud across the Internet and don't have to worry about computing capacity, bandwidth, storage, security, and reliability of the system. By using cloud technologies many user can share data on a single platform so that a common result can be concluded. Using cloud technologies a huge amount of data can be taken as a research. In a medical research it need to process data that come from different cluster with huge amount. So in this case a huge amount of data, or a large number of chromosomes, or a huge amount of genomic data can be considered for the research.

In order to develop a new drug it needs a lot of research & application that can be done upon a chromosome or gene. Many times the research takes place in many places by different researcher, but the goal may same. In this case the sharing of data, information, or the research result plays a major rule for the conclusion or simply for the development of the drug. By applying a particular drug to a chromosome the researcher should study the behavior, the reaction, and the changes that happened to that particular chromosome. the development of the drug the application should be applied to various living cells, various chromosomes, or simply various genes. 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.

Apache hadoop is such a cloud technology that provides a plat form which is used to process a huge amount of data. Till now it has several application 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 500 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 the high quality databases which never delete any data. So data can be frequently used and can be access any time for further use.

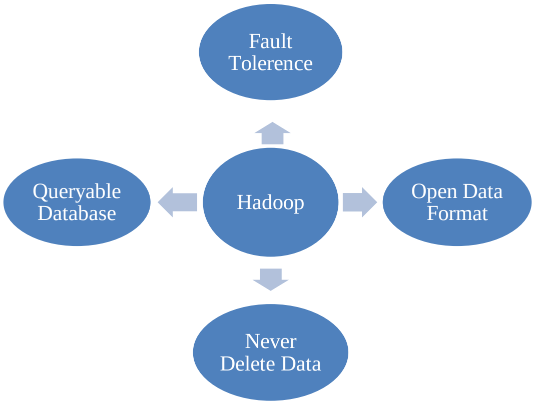

Fig. 2 Hadoop Distributed File System

So Cloud technologies can be used to various medical researches that take place with huge amount of data. Gene mapping is nothing but the distance between two genes in a chromosomes which may changes from time to time. Before development of a drug it is applied to various chromosomes 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 heavy testing or application a drug may develop. 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. By using cloud technology many researchers can communicate, share there opinions and research result in a single platform so that a common conclusion can be extract at last for the development of any medicines or drugs.

CONCLUSIONS

Cloud computing is always a good technology to develop a good networking platform. It provides such a plat form so that a dedicative communication can be take place without thinking about computing capacity, bandwidth, storage, security, and reliability of the system. Mapping of gene is nothing but to calculate the distance between genes in a chromosome, which generally comes in a huge number from different clusters. In order to develop a result it 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 towards other real time application of cloud technology upon this task.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] http://www.woodrow.org/teachers/bi/1994/chromosomes.html [Last accessed January 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] Sarkis M. , Goebel B. , Dawy Z. , Hagenauer J. , Hanus P. , Mueller J.C. , “Gene mapping of complex diseases - A comparison of methods from statistics information theory, and signal processing” Signal processing magazine, IEEE, vol-24(1)pp. 83-90, 2007.

[7] Bo Gao, Changjie Guo, Zhihu Wang, Wenhao An, Wei Sun, “Develop and Deploy Multi-Tenant Web-delivered Solutions using IBM middleware: Part 3: Resource sharing, isolation and customization in the single instance multi-tenant application”. http://www.ibm.com/developerworks/webservices/ library/wsmultitenant/index.html, 2009.

[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.