Genomics with Cloud Computing

Sukhamrit Kaur, Sandeep Kaur

Abstract: Genomics is study of genome which provides large amount of data for which large storage and computation power is needed. These issues are solved by cloud computing that provides various cloud platforms for genomics. These platforms provides many services to user like easy access to data, easy sharing and transfer, providing storage in hundreds of terabytes, more computational power. Some cloud platforms are Google genomics, DNAnexus and Globus genomics. Various features of cloud computing to genomics are like easy access and sharing of data, security of data, less cost to pay for resources but still there are some demerits like large time needed to transfer data, less network bandwidth.

Index Terms: Cloud Computing, DNA, DNAnexus, EC2, Genomics.

1 Introduction

GENOMICS is study of all the genes that collectively make up an organism. All genes collectively are known as genome. So genomics is study of genome which provides information of features of organisms genomics can be divided into three categories: Structural, Functional and Comparative Genomics. Structural genomics means determining the three dimensional structures all proteins in genome known as proteome and understanding the biological meaning of proteome. To determine structure of protein, some methods like nuclear magnetic resonance NMR are used and its main application is drug design to treat some diseases. Functional genomics is to determine functions of genes and proteins. Comparative genomics involves comparing genomes of different organisms i.e. genes, genomes and proteins. It is used to see that how much two species are closely related to each other that mean to study similarity and differences of various organisms. [1] Some genome analysis tools are BLAST and PSI-BLAST, SSEARCH, CUSHAW, FASTA, CASAVA, SAM, IMPALA, and HMMER. Genome analysis includes the detection of similarity in two or more sequences which is important in research and diagnostic work. [2**]** With genome sequencing some challenges in the biomedical field have arose like, large amounts of data produced from sequencing, data transfer, access control and management which is difficult for researchers with limited computational power and storage space. [3] Human DNA is made of approximately 3 billion base pairs which represents nearly 100 gigabytes (GB) of data, the equivalent of 102,400 photos. By the end of 2011, the global annual sequencing capacity was estimated to be 13 quadrillion bases and counting, enough data to fill a stack of DVDs two miles high. To store, process and analyse this big data is very expensive which leads to the use of cloud computing where users can hire infrastructure on a ‘‘pay as you go’’ basis, thereby avoiding large capital infrastructure and maintenance costs. These services of hardware and computational power can be provided using user friendly web interfaces. [4] A solution for this problem is cloud computing which is number of networks of computers equipped together by the Internet to work on a specific computing problem. Microsoft, Google and Amazon are providing cloud computing services for genomics which is a cost effective solution for researchers. [5] Currently, the leading cloud service provider is Amazon Web Services. They offer many resources that help to store and analyze the data produced by whole genome sequencing. Amazon S3 used to store the data in a secure, encrypted, redundant environment and EC2 provides a flexible, scalable and stable computational environment. Users can create virtual machines of different sizes like machine with 60 GB of RAM and 88 cores parallel workflow Elastic Map Reduce provides a framework for parallelizing jobs, so that tasks that may have taken days before can now be performed in a matter of hours. All these services collectively provide research institutions with the necessary capacity to store and analyze sequencing data. [3] In recent years, Google and Amazon have struggled to tackle the problem of big data which are handled by their specific cloud applications [4] like the Amazon S3 cloud computing service provides is used for storing and retrieving amount of genome data. Cloud computing is a rising technological paradigm enabling researcher to dynamically virtual machine that will be better than large computational tools in bioinformatics. It offers fast scaling, less management, pay-as-you-go pricing, code reproducibility and the potential for 100% utilization. [6]

2 Problems in genomics solved by cloud computing

2.1 Challenge of Big Data

To understand the living system, large amount is biological information is used. For example, data produced from large projects like 1000 Genomes will give information in Petabytes. Some other challenges are: data transfer, access control and management; standardization of data formats; and accurate modelling of biological systems by integrating data from multiple dimensions.

2.2 Data Transfer, Sharing, Access control and management

Analysis can increase the size of the raw data, and results of analysis can be a part of information needed and stored in one place. So, it is important to efficiently move these big data over the internet, for providing access control if the data is stored centrally to reduce storage costs and to properly manage large

- *Sukhamrit Kaur is currently pursuing masters degree program in computer science in Guru Nanak Dev University, India, PH-+919872455529.*
  E-mail: sukhamrit91@gmail.com
- *Sandeep Kaur is currently working as assistant professor in Guru Nanak DevUniversity Regional Campus Gurdaspur, India*

amount of data. Current solution is to store data on hard drives and ship it physically to customers, or upload the data onto local, temporary servers and until the hard drives arrive at lab, analysis cannot begin and data are only delivered to a single location that may be needed by many researchers at many. Therefore, data transfer, storage and sharing are time consuming and effort consuming processes.

2.3 Standardizing Data Format

Different centres generate data in different formats, and some analysis tools require data to be in particular formats or require different types of data to be linked together. Reformatting and re-integrating data can waste time. So for sequence analyses across different platforms requires tools to be adapted to specific platforms.

2.4 Expensive and Inflexible Access to Computing Power

To get more computing power, money needs to be spent on hardware.

3 CLOUD PLATFORMS FOR GENOMICS

3.1 Google Genomics

Google genomics is a cloud platform for storing, processing, exploring and sharing data produced by genomics. It allows you to store alignments for one or many genomes, process data produced by genomics in minutes or hours by using parallel computing like MapReduce of Amazon cloud services, explore data, share genomic data between research groups. Sources of data include dataset (grouping of genomic data and analysis), reads (nucleotide sequence produced by sequencer with quality score and metadata), read group set(collection of reads and metadata), variants(positions of genetic differences), jobs(collections of large data). This platform provides managing datasets, sharing datasets, importing reads, searching for reads and exporting of reads, to work with genomics. [9]

3.2 DNAnexus

DNAnexus provides for solution for dna sequencing and researchers who work on data produced by sequencing. This platform solves the problem of analysis and management of this produced data. It runs on Amazon Web Services. Thousands of CPUs and 100 of Petabytes is available through DNAnexus, by accessing only those resources that are required according to our need. Its various data management features are given below**:**

- Infrastructure is provided on demand that means only that amount needed of computing power and storage is provided to user.
- It ensures the quality of sequence data before providing it to user.
- It gives access to various bioinformatics tools for sequencing.
- It is designed for the large growing sequence data.

Some features that benefits researchers are given below:

- Infrastructure is provided on demand that means only that amount needed of computing power and storage is provided to user.
- It debugs sequencing problems and provides high quality data.
- Long term storage of sequence data and provides fast or instant access to that data.
- Mapping of sequence data to genome of your choice.
- Provides fast visualization of genomic data.
- RNA-seq(method to measure gene expression) and ChIP-seq(study interaction of DNA with other molecules) applications can be accessed though this platform with one click.

DNAnexus provide security to sequencing data

- To manage and monitor security of data, it uses ISO 27002 international security standard.
- By allowing customer control i.e. user decides that how data should be processed, who should access it, and what data should be deleted
- Data centres and offices are protected by firewalls

[10]

3.3 Globus Genomics

It is combination of algorithms, data management tools, graphical workflow environment, and good computing infrastructure. It provides us with features of fast, reliable and easy to use data movement capabilities. Globus Genomics also uses Amazon Web Services for computational analysis for adjusting the resources according to research requirements, provide almost infinite pool of available resources and charges only for the resources used. [8]

7 MERITS AND DEMERITS OF CLOUD COMPUTING IN GENOMICS

7.1 Merits

- Cloud computing provides appropriate platforms for computational problems in today’s genomics research that is very difficult with previous or available methods.
- Pay only according to your usage i.e. of only those resources that you need and use. Like if you use 10 GB of storage then pay only of that but not of full hard drive price. Backup and recovery are included in this cost.
- It reduces computation time like from one day to one hour with very less cost.
- It also solves the problem of transferring and sharing data with other genomic researchers by placing data on cloud that can be accessed by many researchers from any place. [7]

7.2 Demerits

- Reduced control over distribution of computation and resources.
- Large time needed to transfer large data to and from cloud.
- It provides easy access to resources but data transfer problems still remains.
- Problem of network bandwidth makes transfer of large data difficult [7]

4 CONCLUSION

Genomics produces large amount of sequence data known as big data that is not easy to process and manage on normally using machines. So, cloud computing becomes the solution for these problem by providing appropriate platform more computational power for processing and more storage for data, easy and less costly access to resources required for processing and storing data.

Acknowledgment

Nothing concrete can be achieved without an optimum combination of inspection and perspiration. Our sincere appreciation and gratitude goes to GOD who HAS always helped us in good and bad.

REFERENCES

[1] Amanda Knowles et.al. Available: wiki.brwon.edu.

[2] M. Femminella et.al. (2014). "The ARES project: cloud services for medical genomics", pp. 15-22.

[3] (2015). Cloud Computing. Available: icbi.georgetown.edu.

[4] A. D. Driscoll et.al. (2013). 'Big data', hadoop and cloud computing in genomics. Vol. 46, pp 774-781.

[5] (2013). Answers to Genome analysis may be in the clouds. Available: www.genome.gov.

[6] P. Kudtarkar (2010). Cost Effective cloud computing: A case study using the comparative genomic tool, Roundup. pp 197-203.

[7] E.E.Schadt et.al. (2010) Computational solutions to large scale data management and analysis. 11(9), pp 647-657.

[8] Pediatric brain research laboratory uses Globus genomics to overcome its hurdles. Available at ww.globus.org.

[9] Dr. G. Mirzaa (2013). Globus genomics a solution as cutting edge as our research. Available: www. Globus.org.

[10] (2010). Available: classic.dnanexus.com.