International Journal of Recent Development in Computer Technology & Software Applications
Vol. 1, Issue 1- 2017

GENOMICS WITH CLOUD COMPUTING

GITA NAWALE*, RB DHANDE*

ABSTRACT

Now day's cloud computing and big data technologies are mostlyuse in bioinformatics. Genomics is study of genome which provides large amount of data for which large storage and computation power is needed. To understand genomics analysis large amount is biological information is used. These issues are solved by cloud computing that provides various cloud platforms for genomics. These platforms providesmany services to user like easy access to data, easy sharing and transfer, providing storage in hundreds of terabytes, more computational power.

Ever improving next generation sequencing technologies has led to an unprecedented proliferation of genomic sequence data. Biology is now one of the fastest growing fields of big data Science. A consequence ofis that the medical discoveries of the future will largely depend on our ability to process and analyse large genomic data sets, which continue to expand as the cost of sequencing decreases.Here authors provide an overview of genomic cloud computingand various cloud computing platforms for genomicsare discussed.

KEYWORDS*:* Genomic, Cloud Computing, Bioinformatics, Medical, Genes Etc.

INTRODUCTION

Genomic research is the most illustrative area in bioinformatics, as it is the underlying advance of a few sorts of investigations and it is additionally required in a few different bioinformatics fields. It thinks about genomic highlights-DNA arrange ments, qualities, administrative groupings, or other genomic auxiliary segments of various living beings. When all is said in done, similar genomics begins with the arrangement of genomic. It is utilized to see that how much two species are firmly identified with each other that intend to ponder comparability and contrasts of different living beings and it likewise used to recognize any infection. Some genome analysis tools are BLAST and PSI-BLAST, and FASTA etc. Genome analysis includes the detection of similarity in two or more sequences which is important in research and diagnostic work .genome Sequencing is a prominent example of a big data technology because of the massive amount of information it produces .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.

*Department of Computer Science and Engineering, Anuradha Engineering College, Chikhli.
Correspondence E-mail Id: editor@eurekajournals.com

The sensor nodes of the wireless sensor network is permits irregular sending in difficult to reach territories, this implies convention of the wireless sensor is self-sorted out, another imperative component of the wireless sensor network is agreeable exertion of the sensor nodes. Sensor nodes are gathering data about condition, in the wake of gathering it they process it and after that transmit to the base station. Base station gives an interface amongst client and web. Essential characteristic of the wireless sensor network are restricted energy, dynamic network topology, bring down power, hub disappointment and Mobility of the nodes, short-run communicate correspondence and multi-jump steering and extensive size of deployment cost. These services of hardware and computational power can be provided using user friendly web interfaces. 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. 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 protected, encoded, repetitive condition and EC2 gives an adaptable, versatile and stable computational condition. Clients can make virtual machines of various sizes like machine with 60 GB of RAM and 88 centers parallel work process Elastic Map Reduce gives a system to parallelizing employments, so assignments that may have taken days before would now be able to be performed in a matter of hours. Every one of these administrations all in all gives look into organizations the fundamental ability to store and investigate sequencing information. In recent years, Google and Amazon have struggled to tackle the problem of big data which are handled by their specific cloud applicationslike 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 machining 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.

GENOMICS

GENOMICS is investigation of the considerable number of qualities that aggregately make up a living being. All qualities on the whole are known as genome. So genomics is investigation of genome which gives data of highlights of creatures genomics can be partitioned into three classifications: 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 and genomes [2].

Figure 1.genomes

COMPUTING CLOUD

Cloud computing is defined as “a pay-per-use model” for empowering helpful, on-request organize access to a mutual pool of configurable registering assets (ex. systems, servers, stockpiling, applications and administrations) that can be quickly provisioned and discharged with negligible administration exertion or specialist co-op association [3]. Some of the significant ideas included are matrix figuring, circulated frameworks, parallelized programming and representation innovation. A solitary physical machine can have different virtual machines through virtualization innovation. Issue with network figuring was that exertion was significantly spent on keeping up the vigor and strength of the bunch itself. Big data technologies now have identified solutions to process huge parallelized data sets cost effectively. Cloud computing and big data technologies are two different things, one is facilitating the cost effective storage and the other is a Platform as a Service (Paas), respectively.

Figure 2.cloud computing

GENOMIC CLOUD COMPUTING

Genomic cloud computing can be defined as a versatile administration where hereditary succession data is put away and prepared for all intents and purposes (ie. in the 'cloud') as a rule by means of arranged, substantial scale server farms available remotely through different customers and stages over the Internet. As opposed to purchasing more servers for the nearby research site, as was done before, genomic distributed computing enables analysts to utilize advances, for example, application programming interfaces (APIs) to dispatch servers. Different cloud computing stages have risen for genomic analysts, including Galaxy, Bionimbus and DNA nexus which enable analysts to perform genomic examinations utilizing just a web program. These stages thus may keep

running on specific mists gave by cloud service providers (CSPs).

Four deployment models of cloud computing have emerged in recent years.

- Commercial cloud infrastructure (eg, Google and Amazon) is provisioned for open use by the general public and may be owned, managed and operated by a business, academic or government organization or some combination of them. Community cloud infrastructure is provisioned for exclusive use by a specific community of consumers from organizations that have shared concerns.
- Hybrid cloud infrastructure comprises two or more distinct cloud infrastructures (private, community or commercial) that remain unique entities but are bound together by standardized technology that enables data and application portability.
- Private cloud infrastructure is provisioned for exclusive use by a single organization comprising

In addition to these categories, cloud computing can be organized into different types of service categories as follows:

INFRASTRUCTURE AS A SERVICE (IAAS)

IaaS conveys a wide range of assets (virtualized) including CPU (fittings), OS (programming projects) and so forth summing up a full PC foundation, coming to the maximum capacity of PC assets by means of Internet. Virtualized assets can be gotten to as an open utility by clients and in this way paying for the cloud assets that they use. Adaptability and customization offer opportunity to various clients to get to various cloud assets, according to their necessity, in this manner meeting the modified needs of various clients.

Examples:

1. Cloud Bio Linux is a virtual machine that is publicly accessible for high-performance bioinformatics computing.
2. Amazon Elastic Compute Cloudand Google Compute Engine

SOFTWARE AS A SERVICE(SAAS)

SaaS conveys a huge assortment of programming administrations online for various sorts of information investigation encouraging remote access of different substantial bioinformatics virtual products. Subsequently, it dispenses with the requirement for neighborhood establishment, along these lines facilitating programming support. Avant-garde cloud-based administrations for bioinformatic information investigation has made life simple for the users [14].

Example:

1. google docs

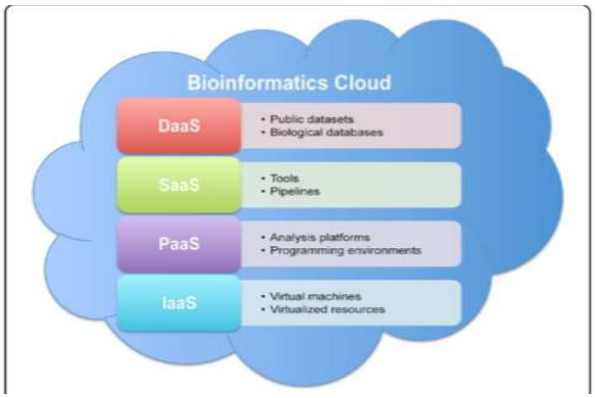

Figure 3.Bioinformatics cloud

PLATFORM AS A SERVICE (PAAS)

Paas enable clients to create, test and utilize cloud applications in a situation where PC assets scale to coordinate application request naturally and progressively. This adaptability calculate helps creating applications for natural data [14].

Two PaaS stages:

1. Eoulsan, cloud-based-for high-throughput sequencing investigations
2. System Cloud, cloud-scale-for vast scale information investigations.

DATA AS A SERVICE (DAAS)

Bioinformatics clouds are subject to data for downstream investigations. "It is accounted for that yearly overall sequencing limit is past 13 Pbp and on an expansion by a factor of five consistently". Because of this unrevealed blast of data, Data as a Service (DaaS) conveyance by means of Internet has picked up significance. It gives dynamic data access on request, alongside state-of-the-art data access to an extensive variety of gadgets, associated over the Web. Amazon Web Services (AWS) give a brought together billow of open data indexes (e.g. archives of GenBank, Ensembl databases, 1000 genomes etc[14].

PROBLEMS IN GENOMICS SOLVED BY CLOUD COMPUTING

CHALLENGE OF BIG DATA

To comprehend the living framework, huge sum is natural information is utilized. For instance, data delivered from vast tasks like 1000 Genomes will give information in Petabytes. Some different difficulties are: data exchange, get to control and administration, standardization of data formats and precise displaying of organic frameworks by coordinating data from numerous measurements. With more than 2.5 quintillion bytes made each day, information stockpiling and investigation has turned into an incredible test. Cloud computing can be the answer for Big data issue. Cloud computing is essential in BIG DATA investigation because of its application sharing and savvy properties. This innovation will help in current genomic data stockpiling and investigation.

Cloud computing postures issues for engineers and clients of cloud programming as it requires expansive data transfers over valuable low-transmission capacity. This additionally raises new protection and security issues. Be that as it may, it is an inexorably important apparatus for handling extensive datasets and it is as of now utilized by the pharmaceutical organizations, web organizations, logical labs and bioinformatics services[5].

Fig.showsthe Schematic representation of a pipeline from data generated using NGS to data translationfor clinicians and researchers.Data is generated by Next-Generation DNA Sequencers (omics data such as genomes, exosomes, and other types of similar information), transferred to the “cloud” or internal servers, analyzed and visualized using different solutions that are available in the market. Finally, the data is translated as a short report to clinicians and researchers after a deep analysis for biomarkers and drug targets associated to specific disease phenotypes. Genome variants are also identified when comparing different samples, generating high-quality interpretation based on our current knowledge. This type of pipeline will ease the implementation and application of different types of omics data for the clinics and also for research purposes. Between data transfer, storage and visualization, patient data needs to be secured by encryption of the information [15].

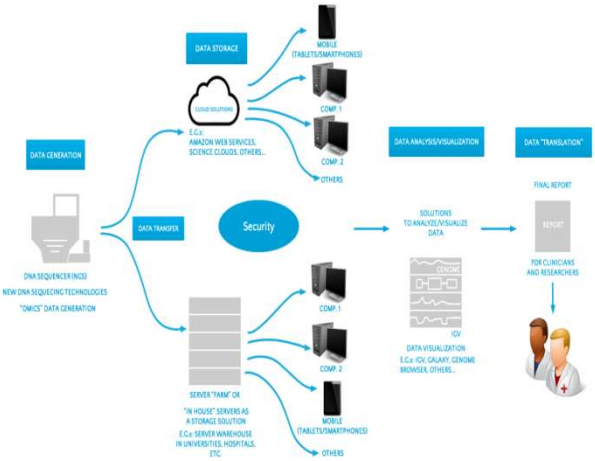

Figure 4.Data in Genomics

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 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.

DATA SHARING

Presently, the biggest cloud provider is Amazon, providing commercial clouds for big data processing. Google is another provider allowing users to develop web applications and analyses data. There is more to be done with commercial clouds to provide sample data and software, along with keeping pace of the emerging needs of researches, which require customized clouds for bioinformatics analysis. Open access and public availability of data and software are of equal significance. The availability of the cloud publicly to the scientific community is essential when data and software's are in cloud. It ensures data integration, reproducible analyses, maximum scope for sharing.

STANDARDIZING DATA FORMAT

Distinctive focuses create data in various arrangements and some investigation apparatuses expect data to be specifically organizes or require diverse sorts of data to be connected together.

Reformatting and reintegrating data can sit around idly. For instance, cutting edge sequencing organizations don't convey crude sequencing data in a configuration regular to all stages, as there is no broad standard past straightforward content records that incorporate the nucleotide arrangement and the relating quality esteems. So for grouping investigations crosswise over various stages expects instruments to be adjusted to particular stages.

The genomic data are use in clinics.The ultimate goal is to advance the use of genomic data in the clinic for improved diagnostics, treatment, and basic understanding of disease and drug responseto succeed in the Clinics it required standard data.

EXPENSIVE AND INFLEXIBLE ACCESS TO COMPUTING POWER

To get all the more processing power, cash should be spent on hardware. Tending to big data and computational difficulties requires productively focusing on restricted assets - cash, power, space and individuals to settle a use of premium. Thusly, this requires understanding and misusing the idea of the data. Variables that must be considered to take care of a specific issue most effectively include: the size and multifaceted nature of the data the simplicity with which information can be proficiently transported over the web.

CLOUD PLATFORMS FOR GENOMICS

GOOGLE GENOMICS

Google genomics is a cloud stage for putting away, handling, investigating and sharing information delivered by genomics. It enables you to store arrangements for one or numerous genomes, process information created by genomics in minutes or hours by utilizing parallel figuring like MapReduce of Amazon cloud administrations, investigate information, share genomic information between look into gatherings. Wellsprings of information incorporate dataset (gathering of genomic information and investigation), peruses (nucleotide succession delivered by sequencer with quality score and metadata), read bunch set (accumulation of peruses and metadata), variations (places of hereditary contrasts), jobs(collections of vast information). This stage gives overseeing datasets, sharing datasets, bringing in peruses, hunting down peruses and sending out of peruses, to work with genomics.

Genomic data's with petabytes rapidly growing toward exabytes. What if you could analyze massive genomic data using the same technologies that Google uses for Search, Maps and YouTube.

You can stack up petabytes of grouping peruses, variations, references, and comments, and process them all effectively. You can work with your genomics information and utilize:

Google Genomics API, our execution of the open standard from the Global Alliance for Genomics and Health, supported by Google advancements like Bigtable.

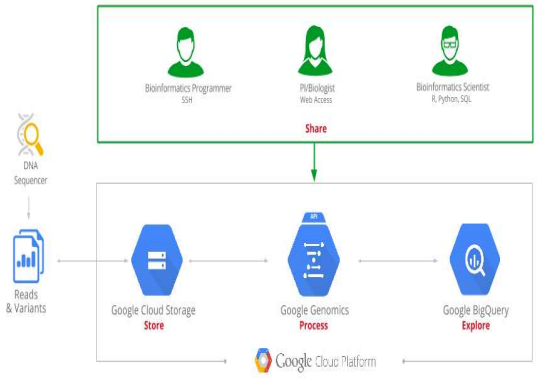

Figure 5.Google Genomics

ADVANTAGES

1 Cloud figuring gives fitting stages to computational issues in today's genomics explore that is exceptionally troublesome with past or accessible techniques.

2 Pay just as indicated by your utilization i.e. of just those assets that you need and utilize. Like in the event that you utilize 10 GB of capacity at that point pay just of that however not of full hard drive cost.

Reinforcement and recuperation are incorporated into this cost. that is, the client asks for a PC framework compose on request to fit their needs and pays for the time in which they utilized a case of that framework.

3 It lessens calculation time like from one day to one hour with less cost.

4 It likewise takes care of the issue of exchanging and offering information to other genomic scientists by putting information on cloud that can be gotten to by numerous specialists from wherever.

DISADVANTAGES

1 Reduced control over circulation of calculation and assets and the time and cost that are required to exchange expansive volumes of information to and from the cloud.

2 Large time expected to exchange extensive information to and from cloud.

CONCLUSION

Presently days singular research lab can create terabytes of information which is no suprises to new sequencing advances in genomic inquire about. Superior calculation conditions continue enhancing handling vast scale information effortlessly.

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.

In this paper we illustrate the current cloud platforms and services.by using this platform genomic data can be store and access from anywhere.

REFERENCES

[1]. DNAbarcoding-URL: http://www.dnabarcoding101.org/bioinformatics.html.

[2]. J.M. Butler, Forensic DNA Typing. Biology, Technology and Genetics of STR Markers, 2nd ed., Elsevier Academic Press, Burlington, 2005.

[3]. M. Lynch, God's signature: DNA profiling, the new gold standard in forensic science, Endeavour 27 (2003) 93–97.

[4]. Available at: www.dnabarcodes.org.

[5]. Identifying species with DNA barcoding. Available: www.barcodeoflife.org(2010).

[6]. Abayomi-Alli A., Omidiora E. O., Olabiyisi E.O., and Ojo J. A. (2012) Enhanced E-Banking System with Match-On-Card Fingerprint Authentication and Multi-Account ATM Card. The Journal of Computer Science and Its Applications, An International Journal of the Computer Society of Nigeria (NCS), V0l. 19, No. 2 December, 2012.

[7]. Kerr KCR, Stoeckle MY, Dove CJ, Weigt LA, Frances CM, Herbert PDN. (2007). Comprehensive DNA barcode coverage of North American birds.Molecular Ecology Notes, 7, 535-543.

[8]. Hollingsworth PM, Graham SW, Little DP. (2011). Choosing and using a plant DNA barcode. Plos One, 6 (5), e19254.

[9]. A.J. Jeffreys, V. Wilson, S.L. Thein, Individual-specific 'fingerprints' of human DNA,Nature 316 (1985) 76–79.

[10]. Seifert KA, Samson RA, Dewaard JR, Houbraken J, Levesque CA, Moncalvo JM, LouisSeize G, Herbert PD. (2007). Prospects for fungus identification using COIDNA barcodes, with Penicillium as a test case study. PNAS, 104(10),3901-3906.

[11]. Stoeckle M, Waggoner PE, Ausubel JS. (2004). Barcoding Life: Ten Reasons. Consortium for the Barcode of Life, v3.0. Retrieved from: http://www.barcode

oflife.org/content/barcoding-life-ten-reasons-pamphlet.

[12]. Lopez I and Erickson DL. (2012) DNA Barcodes: Methods and Protocols. Humana Press.

[13]. Stoeckle M, Waggoner PE, Ausubel JS. (2004). Barcoding Life: Ten Reasons. Consortium for the Barcode of Life, v3.0. Retrieved from: http://www.barcode oflife.org/content/barcoding-life-ten-reasons-pamphlet.

[14]. Keele JA, Carmon J, Hosler D. (2014). DNA barcoding for genetic identification of organisms, DNA barcoding standard operating procedure. Technical Memorandum No. 86-68220-14-08.

[15]. http://www.barcodeoflife.org/sites/all/themes/cbol/pdf/barcode_pipeline.pdf.

