Applied Microbiology: Open Access

Singh, Appli Microbio Open Access 2016, 2:2
DOI: 10.4172/2471-9315.1000113

Review Article Open Access

Big Genomic Data in Bioinformatics Cloud

Prachi Singh*

NIIT University, Neemrana, Rajasthan, India

***Corresponding author:** Prachi Singh, NIIT University, Neemrana, Rajasthan, India, Tel: +66-969172953; E-mail: prachi.singh@st.niituniversity.in

Received date: January 02, 2016; **Accepted date:** April 27, 2016; **Published date:** April 28, 2016

Abstract

The achievement of Human Genome project has led to the proliferation of genomic sequencing data. This along with the next generation sequencing has helped to reduce the cost of sequencing, which has further increased the demand of analysis of this large genomic data. This data set and its processing has aided medical researches.

Thus, we require expertise to deal with biological big data. The concept of cloud computing and big data technologies such as the Apache Hadoop project, are hereby needed to store, handle and analyse this data. Because, these technologies provide distributed and parallelized data processing and are efficient to analyse even petabyte (PB) scale data sets. However, there are some demerits too which may include need of larger time to transfer data and lesser network bandwidth, majorly.

Keywords: Big data; Bioinformatics; Cloud computing; Genomics; Hadoop; Research

Introduction

The introduction of next generation sequencing has given unrivalled levels of sequence data. So, the modern biology is incurring challenges in the field of data management and analysis. A single human's DNA comprises around 3 billion base pairs (bp) representing approximately 100 gigabytes (GB) of data. Bioinformatics is encountering difficulty in storage and analysis of such data. Moore's Law infers that computers double in speed and half in size every 18 months. And reports say that the biological data will accumulate at even faster pace [1]. Sequencing a human genome has decreased in cost from $1 million in 2007 to $1 thousand in 2012. With this falling cost of sequencing and after the completion of the Human Genome project in 2003, inundate of biological sequence data was generated. Sequencing and cataloguing genetic information has increased many folds (as can be observed from the GenBank database of NCBI). Various medical research institutes like the National Cancer Institute are continuously targeting on sequencing of a million genomes for the understanding of biological pathways and genomic variations to predict the cause of the disease. Given, the whole genome of a tumour and a matching normal tissue sample consumes 0.1 TB of compressed data, then one million genomes will require 0.1 million TB, i.e. 10^3 PB (petabyte) [2]. The explosion of Biology's data (the scale of the data exceeds a single machine) has made it more expensive to store, process and analyse compared to its generation. This has stimulated the use of cloud to avoid large capital infrastructure and maintenance costs.

In fact, it needs deviation from the common structured data (row-column organisation) to a semi-structured or unstructured data. And there is a need to develop applications that execute in parallel on distributed data sets. With the effective use of big data in the healthcare sector, a reduction of around 8% in expenditure is possible, that would account for $300 billion saving annually.

Review

Cloud computing

Cloud computing is defined as "a pay-per-use model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction" [3]. Some of the major concepts involved are grid computing, distributed systems, parallelised programming and visualization technology. A single physical machine can host multiple virtual machines through virtualisation technology. Problem with grid computing was that effort was majorly spent on maintaining the robustness and resilience of the cluster itself. Big data technologies now have identified solutions to process huge parallelised 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.

Three types of clouds are: public cloud, Private cloud and Hybrid cloud. First one refers to resources like infrastructure, applications, platforms, etc. made available to general public, accessible only through Internet on "pay as you go" basis. Second one refers to virtualised cloud infrastructure owned, housed and managed by a single organisation. Third one refer to the connection of private and public, for scalability and fault tolerance via Virtual Private Networking (VPN). A fourth model is also proposed, namely Community Cloud. Here organisations like public sector organisations, having same interest, can contribute financially towards a cloud infrastructure.

Genomics through big data technologies

With the implementation of big data technologies in storing, processing and analysing genomics data of medical research can profoundly impact mankind. Timely processing of data, and subsequent analysis are still a challenge. Solutions could be implementation of leading big data technologies like Hadoop. There

have been studies regarding the utilisation of Apache Hadoop platform in bioinformatics projects [4].

Bioinformatics tools developed

MapReduce projects [5]

Crossbow project [6]

BlastReduce project [7]

CloudBurst [8]

CrossBow [9]

Cloudera

Cloudera, being the service provider in the big data platform is the leading Apache Hadoop software. It is contributing >50% of its output into open source (Apache licensed) projects, drawing a cutting edge in the development of big data technology and the Hadoop framework. It was established by Google, Yahoo and Facebook leading engineers along with an Oracle executive, who were later joined by the founder of Apache Hadoop project. [3]

Cloudera is a pioneer of big data and cloud computing in the biomedical researches. The chief scientist and the co-founder of Cloudera, is aiming to dedicate 25% of their time towards the use of computational biology in genomics [10]. Hence, leading pioneers of big data and computational biology along with leading multinationals are now committing to aid medical discoveries through contribution towards analysis of large biological data, for the understanding, diagnosis and treatment of diseases. In fact, this is the need of the hour, because the annual growth for healthcare computing is going to be around 20.5% through 2017 [11].

Hadoop

Two key modules: i) MapReduce ii) Hadoop Distributed File System (HDFS)

1. A computational program is divided into many small sub-problems. Distributed on multiple nodes of the computer.
2. A distributed file system for storing data on these nodes.

Such softwares are designed for load balancing among different nodes and allowing distributed processing of large datasets, enabling fault-tolerant parallelized analysis. Bioinformatics cloud involve services like data storage, acquisition, analysis, etc. as the cloud platform delivers hosted services over the Internet. It could be categorized into four categories namely, Data as a Service, Software as a Service, Platform as a Service, and Infrastructure as a Service [12-16].

Data as a service (DaaS)

Bioinformatics clouds are dependent on data for downstream analyses. "It is reported that annual worldwide sequencing capacity is beyond 13 Pbp and on an increase by a factor of five every year" [17]. Due to this unrevealed explosion of data, Data as a Service (DaaS) delivery via Internet has gained importance. It provides dynamic data access on demand, along with up-to-date data access to a wide range of devices, connected over the Web.

Amazon Web Services (AWS) provide a centralized cloud of public data sets (e.g. archives of GenBank, Ensembl databases, 1000 Genomes, Model Organism Encyclopedia, Unigene, etc.) of biology, chemistry, economics, etc. as services [18].

Software as a service (SaaS)

SaaS delivers a large variety of software services online for different types of data analysis facilitating remote access of various heavy bioinformatics softwares. Thus, it eliminates the need for local installation, thereby easing software maintenance. Up-to-date cloud-based services for bioinformatic data analysis has made life easy for the users.

Efforts have been made to develop cloud-scale and cloud-based sequence mapping [19], multiple sequence alignment [20], expression analysis [21], identification of epistatic interactions of SNPs (single nucleotide polymorphisms) [22], and NGS (Next-Generation Sequencing).

Platform as a service (PaaS)

PaaS allow users to develop, test and use cloud applications in an environment where computer resources scale to match application demand automatically and dynamically. This scalability factor helps in developing applications for biological data.

Two PaaS platforms:

1. Eoulsan, cloud-based-for high-throughput sequencing analyses [23];
2. Galaxy Cloud, cloud-scale-for large-scale data analyses [24].

Infrastructure as a service (IaaS)

IaaS delivers all kinds of resources (virtualized) including CPU (hardwares), OS (softwares) etc. summing up a full computer infrastructure, reaching to the full potential of computer resources via Internet. Virtualized resources can be accessed as a public utility by users and thereby paying for the cloud resources that they utilize. Flexibility and customization give freedom to different users to access different cloud resources, as per their requirement, thus meeting the customized needs of different users.

Examples:

1. Cloud BioLinux is a virtual machine that is publicly accessible for high-performance bioinformatics computing [25].
2. CloVR is a portable virtual machine that incorporates several pipelines for automated sequence analysis [26].

Bioinformatics cloud

Data in the cloud

Initial method of analysis involve downloading of data from NCBI, Ensembl, etc. and installation of softwares locally on in-house computers. Placing data and loading softwares in cloud, make a way to deliver them as DaaS or SaaS. Both can be seamlessly integrated into cloud. Thus, storing of biological data achieves the aim of big data analysis within the cloud. We are using conventional biological databases instead of cloud based. But, for larger sequencing projects, generating ultra-large volumes of data, would require cloud for big data analysis and sharing [27,28]. Project like Genome 10K, 1001 Genomes Project, 1KITE, TCGA etc., are similar kind of projects

requiring big data analysis, where solutions of complex biological queries involves utilization of big data tools [29].

Transferring big data

The bottleneck of cloud computing is the transfer of data into cloud. Instead of physically shiping hard drives to the cloud center, a promising solution could be the integration of innovative transferring technologies with cloud computing. One is cloud-based Easy Genomics for high speed genomic data transfer. There was a successful event of transferring genomic data across Pacific Ocean at a rate of about 10 Gigabits per second which proved technologies to be capable of dealing with big data over the Web. Apart from this, there are technologies like data compression and Peer-to-Peer (P2P) data distribution to aid big data transfer [30].

Cloud-based programming

The analysis task is implemented as pipeline through linkages between the outputs of tools with the inputs of other tools, to automate the system. Development of customized pipelines is needed for the large-scale automated and configurable data analysis on a cloud-based environment.

Similar programming paradigm is adopted through Hadoop, where a single task is distributed over multiple nodes. Computational skills are required for the development of cloud-based pipelines in Hadoop without the requirement of extensive coding, rather the setting up a system for data exchange to pave the way for programming environment [31].

Bioinformatics cloud

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 analyse data. There is more to be done with commercial clouds to provide ample 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 [32]. The availability of the cloud publicly to the scientific community is essential when data and softwares are in cloud [33]. It ensures data integration, reproducible analyses, maximum scope for sharing.

Potential Challenges

Genomics researches with enormous amounts of data has recognized the potential benefits of moving to the cloud, but at the same time cloud computing raises some concerns as well. The optimization of the genomics analysis for the cloud has provided efficient and timely services. For instance, data can be easily run from sequencing facility to analysis pipeline on the cloud, as it is generated. However, there is need to be aware of various potential challenges in adopting cloud computing technologies.

Hadoop programming requires a high level of Java expertise; it needs to be simplified to a SQL like interface to generate parallelised programs. Standardisation of reporting and summarisation of results is a problem which is not much addressed; need is to develop better analytics and visualisation technologies. Hadoop with no front end visualisation is difficult to set, use and maintain; efforts are being made towards introducing developer friendly management interfaces instead of shell/command line interfaces.

Considering the scale of the genomic data that needs to be transmitted over internet, it takes considerably large amount of time (might extend to weeks at times). Thus, the rate of transfer of data remains a bottleneck of the technology [36]. Data tenancy is another challenge. Mostly clouds provide lesser capability on data and service interoperability, making it difficult for a customer to move data and services back to an in-house IT environment or to migrate from one provider to another. Moreover, data privacy legislation, legal ownership and responsibility pertaining to data stored between international zones points at another challenge [37]. Nevertheless, genomics and proteomics research projects for sure exhibit the applications for next generation cloud based computational biology and it essentially has the potential to revolutionise the pace of research in life sciences.

Security

Privacy and confidentiality is something that is must to maintain especially when dealing with health information. Cloud computing offers the use of data encryption, password protection, secure data transfer, processes' audits, and the implementation of respective policies against data breeches and malicious use [34]. The involvement of an external entity for data storage and processing services offers added security concerns. Logging access to the data, role-based access, third party certifications, computer network security, notification alarms, change trackers, cloud usage term and associated services are made to address these concerns [35].

Future in microbiology research

Petabytes of raw information can revolutionize microbiology research if we are successful to figure out how to use this gold mine. Winston Hide says "In the last five years, more scientific data has been generated than in the entire history of mankind". Today the data generation is light-years faster that it was just a few years ago and thus we can't imagine the amount of digital information available to us now. Like to study respiratory disease we require capturing huge quantities of data for air quality and then match it with equivalently large datasets, are studies which involve big data. We need to engage lots of eyes in this process.

Conclusion

Cloud computing has seen a lot of hype and excitement recently but in the biotech industry it is gradually getting a recognition as a serious alternative to the hardware infrastructures already existing. Parallel DNA sequencing generates massive amount of data, and its interdisciplinary nature employ cloud computing and big data technologies in life sciences. It facilitates high throughput analytics allowing users to interrogate vast data in no time. Metagenomics, systems biology and protein structure prediction require extensive use of big data technology [12]. Metagenomics as a result of genomics revolution gave way to the sequence based analysis of the microbiome (i.e. microbial genomes), which is going to be several orders of magnitude bigger [13]. Try counting total no. of bacterial cells on earth; must be in the range of 1030, most of them still unidentified. The discovery of novel genes encode new proteins whose structure and function needs to be characterised [14].

Next generation cloud based computational biology has the potential to revolutionise life sciences. Cloud-based resources classified

as DaaS, SaaS, PaaS and IaaS bears great promises in addressing big data analysis, developing variety of services for data storage, acquisition and analysis by integration of data and softwares, as efficient high-speed transfer technologies to aid the transfer of big data. It provides a light programming environment along with develop customized pipelines publicly accessible to the whole scientific community. Despite existing challenges yet to overcome, the potential advantages that these technologies can bring to the genomic research far outweigh the disadvantages.

References

1. Eisenstein M (2012) Oxford Nanopore announcement sets sequencing sector abuzz. Nat Biotechnol 30: 295-296.
2. Managing and Analysing 1,000,000 Genomes.
3. O'Driscoll A, Daugelaite J, Sleator RD (2013) 'Big data', Hadoop and cloud computing in genomics. J Biomed Inform 46: 774-781.
4. Zou Q, Li XB, Jiang WR, Lin ZY, Li GL, et al. (2014) Survey of MapReduce frame operation in bioinformatics. Brief Bioinform 15: 637-647.
5. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297-1303.
6. Gurtowski J, Schatz MC, Langmead B (2012) Genotyping in the cloud with Crossbow. Curr Protoc Bioinformatics Chapter 15: Unit15.
7. Blastreduce: high performance short read mapping with mapreduce.
8. Schatz MC (2009) CloudBurst: highly sensitive read mapping with MapReduce. Bioinformatics 25: 1363-1369.
9. Langmead B, Schatz MC, Lin J, Pop M, Salzberg SL (2009) Searching for SNPs with cloud computing. Genome Biol 10: R134.
10. Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP (2010) Computational solutions to large-scale data management and analysis. Nat Rev Genet 11: 647-657.
11. Healthcare Cloud Computing (Clinical, EMR, SaaS, Private, Public, Hybrid) Market – Global Trends, Challenges, Opportunities & Forecasts (2012–2017).
12. Sleator RD (2010) An overview of the processes shaping protein evolution. Sci Prog 93: 1-6.
13. Sleator RD, Shortall C, Hill C (2008) Metagenomics. Lett Appl Microbiol 47: 361-366.
14. Sleator RD (2012) Prediction of protein functions. Methods Mol Biol 815: 15-24.
15. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, et al. (2011) Big data: the next frontier for innovation, competition, and productivity. McKinsey Global Institute, USA.
16. Stanoevska-Slabeva K, Wozniak T (2010) Cloud Basics - An Introduction to Cloud Computing. In: Grid and Cloud Computing: Business Perspective on Technology and Applications. Stanoevska K, Wozniak T, Ristol S (Edn.) Springer publications pp. 47-61.
17. Truong HL, Dustdar S (2009) On Analyzing and Specifying Concerns for Data as a Service. 2009 IEEE Asia-Pacific Services Computing Conference (Apscc 2009) pp.83-90.
18. Fusaro VA, Patil P, Gafni E, Wall DP, Tonellato PJ (2011) Biomedical cloud computing with Amazon Web Services. PLoS Comput Biol 7: e1002147.
19. Nguyen T, Shi W, Ruden D (2011) CloudAligner: A fast and full-featured MapReduce based tool for sequence mapping. BMC Res Notes 4: 171.
20. Matsunaga A, Tsugawa M, Fortes J (2008) Combining MapReduce and Virtualization on Distributed Resources for Bioinformatics Applications. In Fourth IEEE International Conference on eScience pp. 222-229.
21. Langmead B, Hansen KD, Leek JT (2010) Cloud-scale RNA-sequencing differential expression analysis with Myrna. Genome Biol 11: R83.
22. Wang Z, Wang Y, Tan KL, Wong L, Agrawal D (2011) eCEO: an efficient Cloud Epistasis cOmputing model in genome-wide association study. Bioinformatics 27: 1045-1051.
23. Jourdren L, Bernard M, Dillies MA, Le Crom S (2012) Eoulsan: a cloud computing-based framework facilitating high throughput sequencing analyses. Bioinformatics 28: 1542-1543.
24. Afgan E, Baker D, Coraor N, Goto H, Paul IM, et al. (2011) Harnessing cloud computing with Galaxy Cloud. Nat Biotechnol 29: 972-974.
25. Krampis K, Booth T, Chapman B, Tiwari B, Bicak M, et al. (2012) Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community. BMC Bioinformatics 13: 42.
26. Angiuoli SV, Matalka M, Gussman A, Galens K, Vangala M, et al. (2011) CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing. BMC Bioinformatics 12: 356.
27. Dai L, Gao X, Guo Y, Xiao J, Zhang Z (2012) Bioinformatics clouds for big data manipulation. Biol Direct 7: 43.
28. Dudley JT, Pouliot Y, Chen R, Morgan AA, Butte AJ (2010) Translational bioinformatics in the cloud: an affordable alternative. Genome Med 2: 51.
29. Zhang Z, Bajic VB, Yu J, Cheung KH, Townsend JP (2011) Data Integration in Bioinformatics: Current Efforts and Challenges. In: Bioinformatics - Trends and Methodologies. Mahdavi MA, Rijeka (Edn.), InTech - Open Access Publisher, Croatia.
30. Deorowicz S, Grabowski S (2011) Compression of DNA sequence reads in FASTQ format. Bioinformatics 27: 860-862.
31. Zhang Z1, Cheung KH, Townsend JP (2009) Bringing Web 2.0 to bioinformatics. Brief Bioinform 10: 1-10.
32. Marx V (2012) My data are your data. Nat Biotechnol 30: 509-511.
33. Rosenthal A, Mork P, Li MH, Stanford J, Koester D, et al. (2010) Cloud computing: a new business paradigm for biomedical information sharing. J Biomed Inform 43: 342-353.
34. Hazin R, Brothers KB, Malin BA et al. (2013) Ethical, legal, and social implications of incorporating genomic information into electronic health records. Genet Med 15: 810-816.
35. Rodrigues JJ, de La Torre I, Fernández G, López-Coronado M (2013) Analysis of the security and privacy requirements of cloud-based electronic health records systems. J Med Internet Res 15: e186.
36. Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP (2010) Computational solutions to large-scale data management and analysis. Nat Rev Genet 11: 647-657.
37. Klein CA (2011) Cloudy confidentiality: clinical and legal implications of cloud computing in health care. J Am Acad Psychiatry Law 39: 571-578.