IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-5 , Issue-9 E-ISSN: 2347-2693

Cloud Computing in Bioinformatics: Solution to Big Data Challenge

Shahid Tufail*, M. Abdul Qadeer

[1*]Dept. of Computer Engineering, Z. H. College of Engineering and Technology, Aligarh Muslim University, Aligarh, India
[2]Dept. of Computer Engineering, Z. H. College of Engineering and Technology, Aligarh Muslim University, Aligarh, India

*Corresponding Author: shahid.tufail@zhcet.ac.in

Available online at: www.ijcseonline.org

Received: 03/Sep/2017, Revised: 17/Sep/2017, Accepted: 23/Sep/2017, Published: 30/Sep/2017

***Abstract*—** The piling up of vast quantity of biological data owing to the enormous exploitation of next and third generation sequencing techniques has made their management and handling an uphill task. Cloud computing offers solution to the storage, processing and analysis issues of such a gigantic amount of biological data. The abstraction layer in cloud computing empowers an incorporated access to handling, storage and virtualization. Herein, we review various types of clouds, cloud based service models in bioinformatics and cloud computing platforms with parallel application tools. Lastly, we discuss how the cloud based platforms are being exploited for big data analysis in biology.

***Keywords*—**Cloud computing, bioinformatics, big data, handling, challenge

I. INTRODUCTION

Advances in high-throughput sequencing innovations has prompted an exponential ascent in the biological sequence information. This has put a huge challenge of storage and analysis of such a big amount of data and it is becoming very difficult for small as well large institutions to keep and maintain computational infrastructures for processing such an enormous amount of data. Presently, a promising, efficient and cost effective solution to this growing challenge is cloud computing that uses computing resources to access data over internet. The National Institute for Standards and Technology (NIST) defines cloud computing 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''* [1]. The term cloud is used in terms that the data is stored far away from your device in a manner similar to a 'cloud' literally, however, it is still accessible using a computing device via internet [2, 3]. Grid computing, distributed systems and parallelised programming concepts in association with cloud computing are quite traditional but the idea of virtualization has given cloud computing a new dimension and widespread rollout. The virtualization technology enables hosting of multiple virtual machines using a single physical machine leading to maximum utilization of hardware and capital investment [1]. Operating System (OS) is exploited by applications to interact with hardware in traditional computing while in virtualized computing, the hardware assets (CPU, RAM, storage and networking) are shared by multiple OS images, which are distributed and managed by virtualisation software known as a hypervisor or virtual machine monitor (VMM). A virtual machine (VM) is a software application that copies a physical computing environment in which an OS and applications associated with it can be run with multiple virtual machines installed on a single machine. A VMM processes the requests from the VM to the hardware (CPU, memory, hard disks and network connectivity) [1].

Herein, we review various types of cloud, cloud based service models in bioinformatics, cloud computing platforms with parallel application tools and applications of cloud based platforms in biology. This review summarizes how cloud computing handles large data including their analysis, storage and acquisition in the present scenario.

II. DIFFERENT FORMS OF CLOUD

Public cloud

Cloud possessed and handled by third parties are publicly available with some constraint of security and data variance. Service providers avail public cloud applications, resources, storage and other services to general public. Some of the examples of public cloud are Amazon Web Services (AWS) and Microsoft Azure [4].

Private cloud

Cloud maintained by organizations for their personal issues and the services are accessed only upon the permission of organizations. Microsoft, Hewlett Packard (HP) and Intel boast of having their own private cloud.

Community Cloud
Cloud shared by organizations having common resource requirements (security, jurisdiction and policy), whether managed internally or by a third-party. Organizations take advantage of sharing costs using community cloud computing. For Example Google Gov (google apps for government) has a community cloud.

Hybrid cloud
Cloud composed of two or more clouds i.e. composition of either a private or a public cloud or a private cloud or a community cloud etc. Hybrid cloud guarantees the arrangement and adaptability of in house applications with the adaptation to internal failure and versatility of cloud based platforms [4].

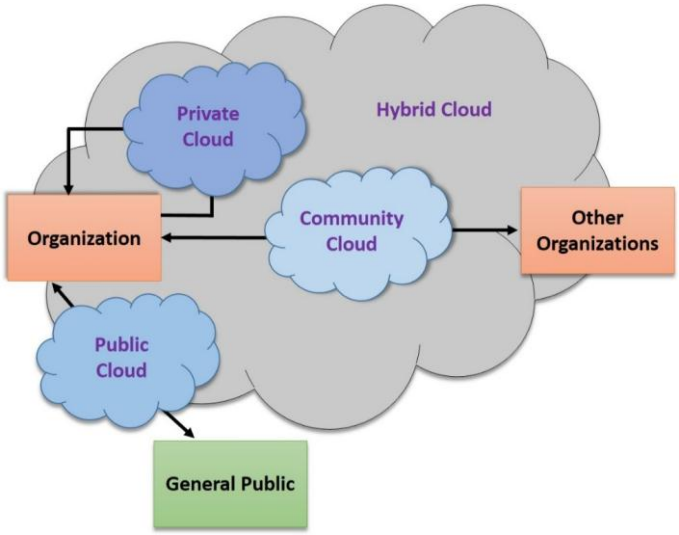

Figure 1. Illustration of various types of clouds

III. CLOUD-BASED SERVICE MODELS IN BIOINFORMATICS

Hadoop and its associated software can be credited for the popularity of cloud computing. Hadoop has two vital components—MapReduce and Hadoop Distributed File System (HDFS). MapReduce divides a computational program into multiple small sub-problems and allocates them on different computer nodes. HDFS offers a distributed file system for putting away information on these nodes. These softwares adjust stack among various nodes [5]. In other words, Hadoop takes into account distributed handling of huge datasets over multiple computer nodes, bolsters huge information scaling (HDFS, HBase), and empowers fault tolerant parallelized examination (MapReduce). Hence, Hadoop is suitable for large data handling in bioinformatics; rather there are evidences of successful exploitation of Hadoop in bioinformatics [6-8]. This paves avenues for cloud-based bioinformatics resources. As cloud computing conveys facilitated services over the Web, bioinformatics clouds include an expansive assortment of services from data stockpiling, data procurement, to data analysis, which all in all fall into four classifications.

Data as a Service (DaaS)
Bioinformatics clouds require data for downstream analyses. Owing to advancement in high throughput technologies, there is an unprecedented continuous growth of biological data. This makes DaaS via Internet one of the most crucial services. DaaS makes dynamic information available on request and furthermore gives access to latest data displayed over the Web. For instance, AWS which offers store of public data sets like chronicles of Ensembl, GenBank, Influenza Virus, Model Organism Encyclopedia of DNA Elements, 1000 Genomes, Unigene, and so forth.

Software as a Service (SaaS)
A variety of softwares are required for different types of data analyses in bioinformatics. SaaS delivers software services on demand to the clients online. Hence, SaaS reduces the need of installing softwares locally and provides latest cloud-based services for bioinformatics data analysis over the Web [5]. Efforts have been made to develop cloud-scale tools, including sequence mapping [9-11], alignment [12], expression analysis [13-15], orthology detection [16], peak caller for ChIP-seq data [17], functional annotation of variants from multiple personal genomes [18], identification of epistatic interactions of single nucleotide polymorphisms (SNPs) [19], and various cloud-based applications for NGS (Next-Generation Sequencing).

Platform as a Service (PaaS)
PaaS offers a working platform as a service. PaaS capsulizes the working environment and the required software to the provider and this platform can be used by clients. Using PaaS, users can develop, test and deploy cloud applications where computer resources scale automatically and dynamically. Programming language execution environments, databases and web servers fall among some of the environments provided by PaaS. Currently, Eoulsan and Galaxy Cloud are the two PaaS platforms in bioinformatics over the Web; Eoulsan (5, 20) uses cloud for high-throughput sequencing analyses, and Galaxy Cloud [5, 21, 22], is a cloud-scale Galaxy for large-scale data analyses.

Infrastructure as a Service (IaaS)
IaaS provides a full computer infrastructure by delivering various virtualized resources via the Internet, including hardware and software. Usage of virtualized resources are in the public domain and paid. Since different users often need different cloud resources, flexibility and customization are essential to IaaS. Examples of IaaS in bioinformatics are; Cloud BioLinux [5, 23], is a VM that is public utility for

performing high-throughput bioinformatics computing, and CloVR [5, 24], is a portable VM used to integrate several channels in order to perform automated sequence analysis.

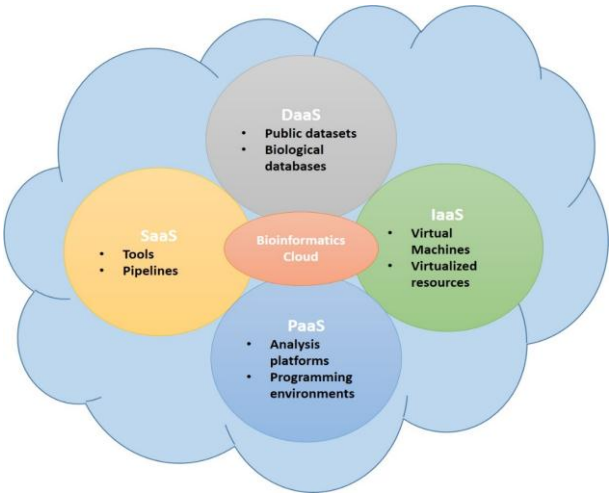

Figure 2. Cloud based service models in bioinformatics

IV. CLOUD COMPUTING PLATFORMS WITH PARALLEL PROCESSING TOOLS

As already said, large data poses huge challenge of analysis, hence, many studies have been performed with parallel applications like Message Passing Interface (MPI) and Hadoop on cloud computing in order to speed up computational time on data processing and analysis. Hadoop performs parallel processing over multiple nodes using HDFS for processing large datasets along with MapReduce [25]. Mapreduce/Hadoop is a boon to cloud computing clusters for solving embarrassingly parallel problems (EPP). Mapreduce and cloud computing programming running in parallel has been found to expedite data analysis in genomic research. CloudBLAST [26], a software combining virtual compute resources, virtual network technologies and bioinformatics platforms, uses Hadoop Vine and BLAST (Basic Local Alignment Search Tool) as application tools. CloudBurst [27], MapReduce based read map algorithm, maps single end next generation sequence (NGS) data. It also uses Amazon E2 as an application tool. Crossbrow [28] exploits applications Hadoop, bowtie, SOAPsnp and Amazon EC2 and is used for searching single nucleotide poplymorphisms (SNPs) with cloud computing. Myrna [25], a software using Hadoop, Amazon Elastic MapReduce (EMR) and HapMap performs cloud-scale RNA sequencing and differential expression analysis. SparkSeq [29] is a NGS data library with MapReduce framework and Apache Spark on the cloud. SeqPig [30] exploits scripts used by Apache Pig (a programming tool to generate MapReduce programs automatically) for analysing large sequence data sets. SeqPig scripts a way of data manipulation, analysis and access with Hadoop. AzureBlast [31], a parallel BLAST engine on Windows Azure cloud platform, is used to run BLAST on multiple instances of Azure Cloud by the query segmentation data parallel pattern. Rainbow [32], an enhancement of Crossbrow, is a tool for large scale whole genome sequencing data analysis using cloud computing. The applications used by rainbow are Crossbow, bowtie, SOAPsnp, Picard, Perl, MapReduce. BioPig [33] is a Hadoop and Apache Pig based sequence analysis tool for large acale sequencing data.

V. APPLICATIONS OF CLOUD BASED PLATFORMS IN BIOLOGY

The applications of cloud computing based platforms to store, analyse and maintain large amount of biological data has been found in various fields of biology including comparative genomics, metagenomics, genome informatics, biomedical informatics, neurosciences, RNA analysis, genome analysis and SNP detection [4].

Reciprocal Shortest Distance (RSD) algorithm has been exploited for comparative genomics studies in bioinformatics. RSD uses applications BLAST, ClustalW and Codeml for comparative studies. This algorithm is obsolete and slow but when implemented in the cloud, it calculated orthologs with the analysis of whole genome data [34, 35]. In neurosciences, cloud computing turned out to be a boon. The neurological data are not shared as the data format created by tools are in their own informal metadata format and thus, hinder the progress of neurological research. Development of an e-science cloud platform known as CARMEN [4] solved the issue. It was created to fulfil the need of sharing, analysing and integrating data online. For RNA sequence analysis, Myrna has been used. As discussed in the previous paragraph, genome analysis and SNP detection has been performed using Crossbrow, a cloud based genome sequencing platform.

VI. CONCLUSION

Herein, we reviewed various types of cloud, cloud based service models in bioinformatics, cloud computing platforms with parallel application tools and applications of cloud based platforms in biology. This review summarizes how the cloud computing is handling large data analysis, storage and acquisition in the present scenario. As with time biological data will keep accumulating owing to advanced techniques, their easy accessibility and cost effectiveness, hence, future studies should be focussed on developing bioinformatics clouds integrating both data and software tools and equipped with high-speed transfer technologies to aid big data transfer. Moreover, the technologies developed should be user-

friendly, and most importantly, be open and publicly accessible to the whole scientific community.

REFERENCES

[1] O'Driscoll A, Daugelaite J, Sleator RD. (2013). 'Big data', Hadoop and cloud computing in genomics. J Biomed Inform 46(5):774-81.

[2] Mansaf Alam, Kashish Ara Shakil. (2012) Recent Developments in Cloud Based Systems: State of Art. https://arxiv.org/abs/1501.01323

[3] Ritushree Narayan. (2017) Cloud Computing In Bioinformatics: Current Status and Future Research. International Journal for Scientific Research & Development (IJSRD) 4(12):198-201.

[4] Radhe Shyam Thakur and Rajib Bandopadhyay. (2014). Role of cloud computing in bioinformatics research for handling the huge biological data. Chapter 20. In: Biology of useful plants and microbes; Edited by: Arnab Sen; Published by: Narosa Publishing House, New Delhi, India.

[5] Lin Dai, Xin Gao, Yan Guo, Jingfa Xiao, Zhang Zhang. (2012) Bioinformatics clouds for big data manipulation. Biology Direct 7:43.

[6] Dudley JT, Butte AJ: In silico research in the era of cloud computing. Nat Biotechnol 2010, 28(11):1181–1185.

[7] Stein LD: The case for cloud computing in genome informatics. Genome Biol 2010, 11(5):207.

[8] Taylor RC: An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. BMC Bioinformatics 2010, 11(Suppl 12):S1.

[9] Nguyen T, Shi W, Ruden D: CloudAligner: a fast and full-featured MapReduce based tool for sequence mapping. BMC Res Notes 2011, 4:171.

[10] Schatz MC: CloudBurst: highly sensitive read mapping with MapReduce. Bioinformatics 2009, 25(11):1363–1369.

[11] Langmead B, Schatz MC, Lin J, Pop M, Salzberg SL: Searching for SNPs with cloud computing. Genome Biol 2009, 10(11):R134.

[12] Matsunaga A, Tsugawa M, Fortes J: Combining MapReduce and Virtualization on Distributed Resources for Bioinformatics Applications. In Fourth IEEE International Conference on eScience; 2008:222–229.

[13] Hong D, Rhie A, Park SS, Lee J, Ju YS, Kim S, Yu SB, Bleazard T, Park HS, Rhee H, et al: FX: an RNA-Seq analysis tool on the cloud. Bioinformatics 2012, 28(5):721–723.

[14] Langmead B, Hansen KD, Leek JT: Cloud-scale RNA-sequencing differential expression analysis with Myrna. Genome Biol 2010, 11(8):R83.

[15] Zhang L, Gu S, Liu Y, Wang B, Azuaje F: Gene set analysis in the cloud. Bioinformatics 2012, 28(2):294–295.

[16] Wall DP, Kudtarkar P, Fusaro VA, Pivovarov R, Patil P, Tonellato PJ: Cloud computing for comparative genomics. BMC Bioinformatics 2010, 11:259.

[17] Feng X, Grossman R, Stein L: PeakRanger: a cloud-enabled peak caller for ChIP-seq data. BMC Bioinformatics 2011, 12:139.

[18] Habegger L, Balasubramanian S, Chen DZ, Khurana E, Sboner A, Harmanci A, Rozowsky J, Clarke D, Snyder M, Gerstein M: VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. Bioinformatics 2012. Epub ahead of print.

[19] Wang Z, Wang Y, Tan KL, Wong L, Agrawal D: eCEO: an efficient Cloud Epistasis cOmputing model in genome-wide association study. Bioinformatics 2011, 27(8):1045–1051.

[20] Jourdren L, Bernard M, Dillies M-A, Le Crom S: Eoulsan: a cloud computing-based framework facilitating high throughput sequencing analyses. Bioinformatics 2012. doi:2010.1093/bioinformatics/bts2165. published online April 5, 2012.

[21] Afgan E, Baker D, Coraor N, Goto H, Paul IM, Makova KD, Nekrutenko A, Taylor J: Harnessing cloud computing with Galaxy Cloud. Nat Biotechnol 2011, 29(11):972–974.

[22] Afgan E, Baker D, Coraor N, Chapman B, Nekrutenko A, Taylor J: Galaxy CloudMan: delivering cloud compute clusters. BMC Bioinformatics 2010, 11(Suppl 12):S4.

[23] Krampis K, Booth T, Chapman B, Tiwari B, Bicak M, Field D, Nelson K: Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community. BMC Bioinformatics 2012, 13(1):42.

[24] Angiuoli SV, Matalka M, Gussman A, Galens K, Vangala M, Riley DR, Arze C, White JR, White O, Fricke WF: CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing. BMC Bioinformatics 2011, 12:356.

[25] Hyungro Lee. (2014) Using Bioinformatics Applications on the Cloud. dsc.soic.indiana.edu/publications/bioinformatics.pdf

[26] A. Matsunaga, M. Tsugawa, and J. Fortes. Cloudblast: Combining mapreduce and virtualization on distributed resources for bioinformatics applications. In eScience, 2008. eScience'08. IEEE Fourth International Conference on, pages 222{229. IEEE, 2008.

[27] M. C. Schatz. Cloudburst: highly sensitive read mapping with mapreduce. Bioinformatics,
25(11):1363{1369, 2009.

[28] B. Langmead, M. C. Schatz, J. Lin, M. Pop, and S. L. Salzberg. Searching for snps with cloud computing. Genome Biol, 10(11):R134, 2009.

[29] M. S. Wiewi_orka, A. Messina, A. Pacholewska, S. Ma_oletti, P. Gawrysiak, and M. J. Okoniewski. Sparkseq: fast, scalable, cloud-ready tool for the interactive genomic data analysis with nucleotide precision. Bioinformatics, page btu343, 2014.

[30] A. Schumacher, L. Pireddu, M. Niemenmaa, A. Kallio, E. Korpelainen, G. Zanetti, and K. Heljanko. Seqpig: simple and scalable scripting for large sequencing data sets in hadoop. Bioinformatics, 30(1):119{120, 2014.

[31] W. Lu, J. Jackson, and R. Barga. Azureblast: a case study of developing science applications on the cloud. In Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing, pages 413{420. ACM, 2010.

[32] S. Zhao, K. Prenger, L. Smith, T. Messina, H. Fan, E. Jaeger, and S. Stephens. Rainbow: a tool for large-scale whole-genome sequencing data analysis using cloud computing. BMC genomics, 14(1):425, 2013.

[33] H. Nordberg, K. Bhatia, K. Wang, and Z. Wang. Biopig: a hadoop-based analytic toolkit for large-scale sequence data. Bioinformatics, 29(23):3014{3019, 2013.

[34] Wall, D. P., Kudtarkar, P., Fusaro, V. A., Pivovarov, R., Patil, P., & Tonellato, P. J. Cloud computing for comparative genomics RSD algorithm summary. *BMC Bioinformatics*. 11, (2010). 259-270.

[35] Kudtarkar, P., Deluca, T. F., Fusaro, V. A., Tonellato, J., & Wall, D. Evolutionary Bioinformatics cost-effective cloud computing : A case study Using the comparative Genomics Tool, Roundup. *Evol Bioinfo*. 6, (2010).197-203.

Authors' Profile

Mr. S. Tufail obtained Bachelor of Technology and Master of Technology degrees from Aligarh Muslim University, Aligarh, India. His areas of interest and research are Data Science, Cloud Computing, Big Data Analysis and similar fields.

Mr A. Qadeer earned his Bachelor of Technology and Master of Technology degrees from Aligarh Muslim University, Aligarh, India. Presently, he is an Asst. Professor with the Department of Computer Engineering, Aligarh Muslim University, Aligarh, India. Earlier, he was working with Cisco Systems Inc. as a Network Consulting Engineer with the Advanced Services division in the APAC region. He has an experience of 15~ years in the area of computer networks and systems. He served as a Technical Co-Chair for IEEE WOCN 2012, 2011, 2010, Technical Co-Chair IEEE AH-ICI 2012, 2011, International Steering Committee for ICACT 2012, 2011, 2010 and as TPC member for CCNC 2013, 2012, 2011, 2010, INMIC 2009, AH-ICI 2009, WIA 2009 and MMA 2009. He has been session chair and TPC reviewer for many IEEE/ ACM conferences. He is the editor of Journal of Digital Broadcasting and Multimedia, Hindawi and is a reviewer for IET Communications Journal as well. Established global and nationwide setups of Internet Service Providers (ISP), Internet Exchange Points (IXP), Internet Data Centre (IDC) and Content Delivery Networks (CDN) both from a Networks and Systems perspective. His areas of research are computer networks, wireless networks, mobile computing, next generation networks, IMS, LTE, WiMAX, 4G, WiBro etc.

