A Study on Cloud Computing for Big Data Analysis in Bioinformatics

Manish Kumar*, Dr. (Prof.) Deva Prakash**

*(Research Scholar, Department of Mathematics & Computer Science, Magadh University, Bodhgaya, Bihar
** (Associate Professor, Head of Department, Department of Mathematics, S.M.D. College, Punpun, Bihar

Abstract **: Bioinformatics is an interdisciplinary field, advances in information technology and life sciences have had a significant impact on it. Research in bioinformatics uses a tremendous quantity of complicated data. Although still in its development, cloud computing holds the potential to assist bioinformatics experts with their enormous data storage and processing challenges. Analysis of huge data volumes is another aspect. The traditional methods employed in bioinformatics take a long time to provide results, and it is also challenging to assess the intricate structure of the data involved. With the use of cloud computing technologies, the issues experienced by bioinformatics researchers in order to conduct their study in an affordable and quick manner may be readily handled. Cloud computing is therefore beneficial for bioinformatics research. As a result, it becomes necessary to use machines with enormous processing power, which raises the cost of doing bioinformatics research. The present state of the art in bioinformatics and big data analytics is also explained, along with any possible problems that may arise in upcoming research. Bioinformatics has taken a further step ahead from internal computing infrastructure into utility-provided cloud computing offered over the Internet to manage the large volumes of biological data produced by high-throughput experimental procedures.**
In this article, we've discussed regarding how big data analytics will potentially be facilitated by cloud computing for bioinformatics researchers and we have examined the cloud-based services currently offered in the field of bioinformatics, characterize them into Data as a Service (DaaS), Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS), and also provide our perspectives on the use of cloud computing in this industry.

Keywords**: Big Data Analytics, Cloud computing, Bioinformatics, Big Data, Data storage, Data analysis**

INTRODUCTION

Bioinformatics faces challenges in the storage and interpretation of enormous volumes of biological data as a result of considerable advancements in high-throughput sequencing technology and the ensuing exponential growth of biological data. The ability of computers to process such large amounts of data is falling behind their sequencing throughput [1]. According to a February 2012 article, two nanopore sequencing technologies (GridION and MinION) can generate extremely long sequencing reads (around 100 kb) at a significantly lower cost and faster throughput [2]. This implies that the amount of biological data will increase quickly. The main problem in bioinformatics is to find the "treasure" inside the vast biological data, which places unprecedented demands on big data storage and processing. It is getting harder for small laboratories or even large institutions to set up and maintain computing infrastructures for data processing as the volume of data keeps increasing. Cloud computing [3][4][5], which fully utilizes the power of several computers and distributes computation and storage as dynamically allotted virtual resources through the Internet [6], is now a promising solution to this problem.

CLOUD COMPUTING AS A PUBLIC UTILITY

The term of "cloud computing" was inspired by the cloud symbol that is often employed to depict the Internet in flowcharts. As a matter of fact, cloud computing is not a new concept; it can date back to 1961 at the MIT Centennial when John McCarthy opined that "computation may someday be organized as a public utility" [7]. Cloud computing makes the best use of multiple computers to provide convenient and on-demand access to hosted resources (e.g., computation, storage, applications, servers, network) via Web Application Programming Interfaces (API). Due to its efficient and economical features, it is believed that cloud computing gains promise in transforming computing into a public utility [8].
Similar to extant public utilities (viz., water, electricity, gas, and telephone), computing utility packages a variety of computer resources as metered services ("*pay-as-you-go*"), which can be accessed by any person without the necessity to know where the services are hosted or how they are delivered. Whether a public utility or not, cloud computing has already become a significant technology in big data storage and analysis, exerting revolutionary influences on both academia and industry.

BACKGROUND OF BIG DATA ANALYTICS

The amount, diversity, and speed of data generation in organizations of all sizes have reached previously unheard-of levels during the past few years. Big Data is the word used to describe this quantity of data. The following are a few significant trends that have contributed to the handling of this problem and are to blame: People use the Internet to send emails, like Facebook posts, tweets, upload images to Facebook, and other things. The volume and diversity of data that are transmitted across the Internet are enormous. The Internet of Things (IoT) is a type of network of manufactured objects integrated with electronics, software, and sensors that enables objects to gather and share data.

• The IoT industry will produce more than $300 billion in sales by 2020, according to Gartner, a leading provider of information technology research and consulting services. Big data and the Internet of Things are essentially two sides of the same coin. For the majority of organizations worldwide, the main difficulty is the extraction and administration of IoT data. A business should put up a suitable analytics platform that performs well and can be expanded in the future in order to handle and retrieve IoT data appropriately.

• People may now communicate with others and engage in personal computer activities like social networking, photo sharing, microblogging, etc. thanks to ubiquitous gadgets. The growth of big data has increased the need for data analytics assistance through widely available devices so that decisions may be made from any location without restriction to a particular workplace.
• Infinite processing power is now available at the lowest cost thanks to cloud computing. The cloud's scalable storage capacity, limitless computational power, and elastic resources make it the perfect location for big data.

Big Data is a vast and comprehensive phrase for data sets, hence conventional analytics are insufficient for it. A new method of analytics called "Big Data Analytics" is used to analyse large amounts of data [9].
Here is a detailed explanation of big data analytics.
Large Data Analytics is the proactive process of looking at both quantitative and qualitative big data to find valid, hidden, beneficial patterns and connections that can be utilized to improve choices for any company.

In essence, there are four primary analytical techniques, and they can be either proactive or reactive.
1. Company intelligence is a proactive method of producing standard and customized business reports in response to requests. This type of analytics produces reports based on the past's fixed data.
2. Big Data Business Intelligence [10] is a methodology similar to Business Intelligence but usable with enormous data sets.
3. According to Sandryhaila [11], big analytics is a proactive approach to analytics that does optimization, text mining, forecasting, statistical analysis, predictive analysis, and other tasks. However, as a typical storage environment, it does not apply to Big Data [12].
4. Big Data Analytics is a preventative strategy that may be used with Big Data. It is beneficial for extracting hidden, undiscovered, and practical data from terabytes, petabytes, and exabytes of data.
Understanding the language used in big data analytics is essential for analysis and interpretation since big data analytics works with big data. The next part will go into great detail on big data analytics.

CLOUD -BASED RESOURCES IN BIOINFORMATICS

Open-source applications that helped with its deployment, such Hadoop and related software, are partially responsible for the rise in popularity of cloud computing. Two essential components of Hadoop are MapReduce and Hadoop Distributed File System (HDFS). HDFS offers a distributed file system that stores data on these nodes while MapReduce breaks a computational program down into several little subproblems and distributes them over multiple computer nodes. In order to manage load balancing across several nodes and identify node failures that may be automatically carried out on any node, Hadoop and its accompanying software were created. In essence, Hadoop offers fault-tolerant parallelized analysis, supports big data scaling (HDFS, HBase), and permits distributed processing of massive datasets over numerous computer nodes (MapReduce).
Hadoop therefore satisfies the requirements of bioinformatics, and various research have effectively applied Hadoop in bioinformatics [13][14][15], resulting in cloud-based bioinformatics resources. As already established, hosted services are delivered through the Internet using cloud computing. Consequently, bioinformatics clouds encompass a wide range of services, from data storage to data analysis, which generally fall into four categories (Figure (Figure1),1), namely, Data as a Service, Software as a Service, Platform as a Service, and Infrastructure as a Service [16]. The following lists the current cloud-based resources for bioinformatics and categorizes them into these four groups (Table) (Table11).

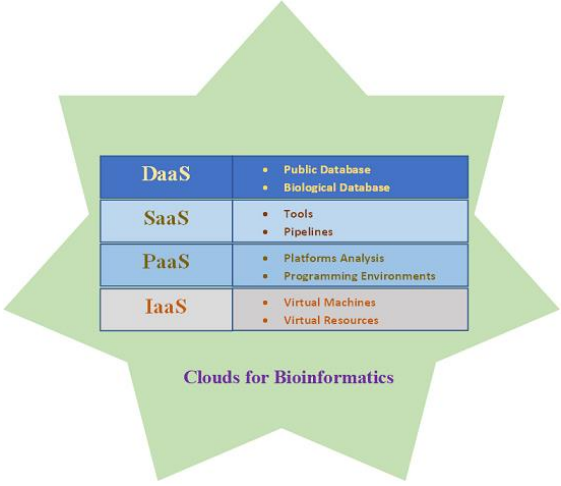

ILLUSTRATION OF BIOINFORMATICS CLOUD

Data as a Service (DaaS), Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service are the four categories under which cloud-based services in bioinformatics are categorized (IaaS).

Resource	Description & availability
Data as a Service (DaaS):	
AWS Public Datasets	Cloud-based archives of GenBank, Ensembl, 1000 Genomes, Model Organism Encyclopedia of DNA Elements, Unigene, Influenza Virus, etc.;
Software as a Service (SaaS):	
BGI Cloud (unpublished)	Cloud-based implementations of various genomic analysis applications;
CloudAligner [20]	Fast and full-featured MapReduce-based tool for sequence mapping;
CloudBLAST [23]	A cloud-based implementation of NCBI BLAST;
CloudBurst [21]	Highly sensitive short read mapping with MapReduce;
Contrail (unpublished)	Cloud-based *de novo* assembly of large genomes;
Crossbow [22]	Read Mapping and SNP calling using cloud computing;
EasyGenomics (unpublished)	Cloud-based NGS pipelines for whole genome resequencing, exome resequencing, RNA-Seq, small RNA and de novo assembly;
eCEO [30]	Cloud-based identification of large-scale epistatic interactions in genome-wide association study (GWAS);
FX [24]	RNA-Seq analysis tool;
Gaea (unpublished)	Cloud-based genome re-sequencing assembly;
Hecate (unpublished)	Cloud-based *de novo* assembly;
Jnomics (unpublished)	Cloud-scale sequence analysis suite based on Apache Hadoop;
Myrna [25]	Differential gene expression tool for RNA-Seq;
PeakRanger [28]	Cloud-enabled peak caller for ChIP-seq data;
VAT [29]	Variant annotation tool to functionally annotate variants from multiple personal genomes at the transcript level;
RSD [27]	Reciprocal smallest distance algorithm for ortholog detection using Amazon's Elastic Computing Cloud;
YunBe [26]	Pathway-based or gene set analysis of expression data;
Platform as a Service (PaaS):	
Eoulsan [31]	Cloud-based platform for high throughput sequencing analyses;
Galaxy Cloud [32][33]	Cloud-scale Galaxy for large-scale data analysis;
Infrastructure as a Service (IaaS):	
Cloud BioLinux [34]	A publicly accessible virtual machine for high performance bioinformatics computing using cloud platforms;
CloVR [35]	A portable virtual machine for automated sequence analysis using cloud computing;

1) Data as a service

Data are a major component of bioinformatics clouds since they are essential for downstream analysis and knowledge discovery. According to reports, the yearly global sequencing capacity has reached more than 13 Pbp and is continuing growing at a factor of five each year. Delivering Data as a Service (DaaS) through the Internet is crucial due to the tremendous rise of biological data [17,18]. DaaS makes it possible to get dynamic data on demand and offers current data that is accessible by a variety of devices that are linked via the Web. One such example is Amazon Web Services (AWS), which offers a centralized database of open data sets, including archives of GenBank, Ensembl, 1000 Genomes, Model Organism Encyclopedia of DNA Elements, Unigene, Influenza Virus, etc. All public datasets in AWS are delivered as services and thus can be seamlessly integrated into cloud-based applications [19].

2) Software as a service

For various sorts of data analytics, bioinformatics needs a wide range of software tools. Software as a Service (SaaS) makes it possible to access bioinformatics software tools remotely via the Internet and provides software services online. SaaS obviates the requirement for local installation, facilitates software maintenance, and offers modern cloud-based services for bioinformatic data processing through the Internet. Sequence mapping [20][21][22], alignment [23], assembly (Contrail, Gaea, and Hecate; unpublished), expression analysis [24][25][26], sequence analysis (Jnomics; unpublished), orthology detection [27], peak caller for ChIP-seq data [28], functional annotation of variants from multiple personal genomes [29], identification of epistatic interactions of single nucleotide polymorphisms (SNPs) [30], and various cloud-based applications for NGS (Next-Generation Sequencing) data analysis (BGI Cloud and EasyGenomics; unpublished

3) Platform as a service

Platform as a Service (PaaS) provides an environment for users to develop, test, and deploy cloud applications where computer resources scale automatically and dynamically to match application demand, eliminating the need for users to estimate the number of resources needed or manually assign resources in advance. This makes the cloud programmable. The quick application development and strong scalability of PaaS make it suitable for creating specialized apps for the study of large-scale biological data. Programming language execution environments, web servers, and databases are frequently provided by PaaS. DaaS may now be seen as a development of PaaS because it offers data as a service and serves as a database. There are currently just two PaaS systems

in bioinformatics that are web servers; these are Galaxy Cloud [32][33] and Eoulsan [31], which are cloud-based platforms for large-scale data analysis and high-throughput sequencing analysis, respectively.

4) Infrastructure as a service

Infrastructure as a Service (IaaS) provides a complete computer infrastructure by distributing all types of virtualized resources through the Internet, including hardware (such as CPUs) and software, in order to maximize the potential of computer resources (e.g., operating systems). Users can pay for the cloud resources they use and use virtualized resources as a public service. Flexibility and customisation are crucial to IaaS since different customers frequently require different cloud resources. Running applications inside of virtual machines is becoming more and more efficient with the continuous and fast progress of IT (VMs). Virtualization technology separates users from the underlying infrastructure and offers flexibility to accommodate the many users' individual demands. There are currently just two IaaS applications in the field of bioinformatics: Cloud BioLinux [34], a publicly available virtual machine for high-performance bioinformatics computing, and CloVR [35], a portable virtual machine with several pipelines for automated sequence analysis.

TOWARD BIOINFORMATIS CLOUD

Although still in its infancy, cloud computing has a lot of potential for solving bioinformatics' huge data storage and processing issues. We outline our opinions on the application of cloud computing

PLACING DATA AND SOFTWARE INTO THE CLOUD

The conventional approach to bioinformatics analysis frequently entails obtaining data from open sources (such as NCBI and Ensembl), installing local software tools, and doing studies using internal computer resources. Data and software may be smoothly and readily incorporated into the cloud for big data storage and analysis by storing them there and providing them as services (specifically, DaaS and SaaS). Big biological data should thus be stored and analyzed on the cloud.

In the age of big data, however, the great majority of data are still stored on traditional biological databases, and only a very small quantity of biological data is now available in the cloud (only AWS, including GenBank, Ensembl, 1000 Genomes, etc.). Long - term goals, an increasing number of sequencing initiatives, including the Genome 10K Project (a collection of DNA sequences representing the genomes of 10,000 vertebrate species), the 1001 Genomes Project (a catalog of genetic variation in 1001 strains of Arabidopsis thaliana), the 1KITE Project (1K Insect Transcriptome Evolution), the Cancer Genome Atlas, etc., would produce extremely large volumes of biological data and thus call for bioinformatics clouds for big data storage, sharing and analysis[36]. Additionally, it frequently takes the use of several instruments to solve the most significant and challenging biological puzzles [37]. However, current initiatives have only made a little impact on the majority of cloud-based solutions. It is unable to carry out extensive bioinformatics operations since the majority of software tools are built for desktop (rather than cloud) [38] and are not offered as cloud-based web services accessible over the Web. A vast quantity of biological data as well as a wide range of bioinformatics tools must fundamentally be made publically accessible in the cloud and made available as services through the Internet in order to satisfy big data storage, sharing, and analysis with reduced cost and improved efficiency. Although a vast amount of biological data is now accessible, it should be kept in mind that biology is still in its infancy (in comparison to other fields, such as physics) and that many theoretical issues in the field still need to be resolved. Hypothesis formulation, experiment design, data generation/collection, tool development (for data analysis), and knowledge formalization are frequently required to address theoretical issues and uncover fundamental biological theories. It can be useful to identify which data and tools should be stored in the cloud.

BIG DATA TRANSFER

A major bottleneck in cloud computing is the transfer of enormous volumes of biological data; at the moment, hard drives are frequently physically sent to the cloud center. Integrating cutting-edge transferring technologies with cloud computing is now a potential approach. Aspera's faspTM high-speed file transfer technology, which significantly accelerates file transfers over the Web and outperforms traditional technologies like FTP and HTTP, is used in BGI's (Beijing Genomics Institute) cloud-based EasyGenomics, as one example, to transmit high-speed genomic data. BGI demonstrated in June 2012 that high-speed transfer technologies (like Aspera's fasp) are capable of handling large amounts of data by successfully sending genetic information across the Pacific Ocean at a sustained pace of over 10 Gigabits per second with big data transfers over the Web. In addition to high-speed transfer technologies, other technologies, such as data compression [39][40] and peer-to-peer (P2P) data distribution [41][42], can also help with huge data transmission.

A LIGHTWEIGHT PROGRAMMING ENVIRONMENT HOSTED IN THE CLOUD

Bioinformatic tasks are frequently implemented as pipelines by connecting the output of one tool with the input of another in order to automate data processing. A cloud-based lightweight programming environment is required to carry out large-scale data analysis and assist in the creation of corresponding bioinformatic pipelines. This environment enables quick creation of tailored pipelines from a large pool of tools and enables automated and configurable analysis on the cloud. At the moment, Hadoop [15][20][21][24][25] is the cloud-based programming paradigm that the bioinformatics community has accepted. With Hadoop, computation-intensive and data-intensive studies are generally resolved by splitting work across numerous nodes. Hadoop's programming environment is not lightweight for most biologists or persons with little or minimal programming experience, and significant computational abilities are still needed for constructing cloud-based pipelines in Hadoop. This would ideally be a simple programming environment without the need for complex keyboard coding. Instead, it would be simple to "drag drop" using a mouse. Such a setting offers remote access to various resources based on cloud computing provided by a utility, fitting nicely with the

projected e-Science movement [43]. (That is, scientific research in many disciplines is carried out via the internet). Additionally, while creating such an environment, consideration should be given to establishing a system of standards for data transfers across various software tools [44], which can in turn open the door for maximizing the potential of a lightweight programming environment.

OPEN BIOINFORMATICS CLOUDS

There are many cloud providers who are motivated by the potential large profits to be earned on a pay-as-you-go basis, and it is anticipated that there will be more providers in the years to come who will develop industrial or academic, private or public clouds. Currently, Amazon is the biggest provider of big data processing commercial clouds. Google offers a cloud platform as well that enables customers to create and host web apps as well as store and analyze data. Commercial clouds can't yet offer enough data and software for bioinformatics analysis, though. Additionally, it is highly challenging for commercial clouds to keep up with the new demands coming from academic research, necessitating the use of customized clouds for bioinformatics research. It should go without saying that open access and free software and data are very important to science [45]. Keeping the cloud open and openly accessible to the scientific community is crucial for bioinformatics research when data and software are both in the cloud [46]. Therefore, it is probable that future efforts should focus on creating open bioinformatics clouds and granting the scientific community access to them. Such bioinformatics clouds may facilitate large-scale data integration, enable repeatable and reproducible studies, maximize the sharing potential, and tap into collective intelligence for knowledge discovery. Interoperability and standards amongst clouds will become crucial challenges when there are more and more bioinformatics clouds [47][48].

CONCLUSION:

This study examined current cloud-based bioinformatics resources and categorized them into DaaS, SaaS, PaaS, and IaaS categories. Future efforts to build bioinformatics clouds involve developing a wide range of services from data storage, data acquisition, to data analysis, making utility-supplied cloud computing delivered over the Internet. This is because cloud computing shows great promise in effectively addressing big data storage and analysis. In the big data era, bioinformatics clouds should integrate both software and data tools, be outfitted with high-speed transfer technologies and other related technologies to aid in big data transfer, offer a lightweight programming environment to help people create customized pipelines for data analysis, and most importantly, be open and publicly accessible to the entire scientific community.

REFERENCES:

[1] Schatz MC, Langmead B, Salzberg SL. Cloud computing and the DNA data race. Nat Biotechnol. 2010;28(7):691–693. doi: 10.1038/nbt0710-691.

[2] Eisenstein M. Oxford Nanopore announcement sets sequencing sector abuzz. Nat Biotechnol. 2012;30(4):295–296. doi: 10.1038/nbt0412-295.

[3] Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP. Cloud and heterogeneous computing solutions exist today for the emerging big data problems in biology. Nat Rev Genet. 2011;12(3):224.

[4] Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP. Computational solutions to large-scale data management and analysis. Nat Rev Genet. 2010;11(9):647–657. doi: 10.1038/nrg2857.

[5] Grossman RL, White KP. A vision for a biomedical cloud. J Intern Med. 2012;271(2):122–130. doi: 10.1111/j.1365-2796.2011.02491.x.

[6] Armbrust M, Fox A, Griffith R, Joseph AD, Katz RH, Konwinski A, Lee G, Patterson DA, Rabkin A, Stoica I, Above the Clouds: A Berkeley View of Cloud Computing. Berkeley: EECS Department, University of California; 2009.

[7] Garfinkel SL. Architects of the Information Society: Thirty-Five Years of the Laboratory for Computer Science at MIT. Cambridge, MA: The MIT Press; 1999.

[8] Buyya R, Yeo CS, Venugopal S, Broberg J, Brandic I. Cloud computing and emerging IT platforms: vision, hype, and reality for delivering computing as the 5th utility. Future Gener Comp Sy. 2009;25(6):599–616. doi: 10.1016/j.future.2008.12.001.

[9] Berisha, B., Mëziu, E. & Shabani, I. Big data analytics in Cloud computing: an overview. *J Cloud Comp* **11**, 24 (2022). https://doi.org/10.1186/s13677-022-00301-w

[10] McAfee, A. and Brynjolfsson, E. (2012) Big Data: The Management Revolution. Harvard Business Review.

[11] Sandryhaila, Aliaksei & Moura, Jose. (2014). Big Data Analysis with Signal Processing on Graphs: Representation and processing of massive data sets with irregular structure. Signal Processing Magazine, IEEE. 31. 80-90. 10.1109/MSP.2014.2329213.

[12] Balazinska, Magdalena & Merlo, Ettore & Dagenais, Michel & Lague, B. & Kontogiannis, Kostas. (2000). Advanced Clone-Analysis to Support Object-Oriented System Refactoring. 98-107. 10.1109/WCRE.2000.891457.

[13] Dudley JT, Butte AJ. In silico research in the era of cloud computing. Nat Biotechnol. 2010;28(11):1181–1185. doi: 10.1038/nbt1110-1181.

[14] Stein LD. The case for cloud computing in genome informatics. Genome Biol. 2010;11(5):207. doi: 10.1186/gb-2010-11-5-207.

[15] Taylor RC. An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. BMC Bioinformatics. 2010;11(Suppl 12):S1. doi: 10.1186/1471-2105-11-S12-S1.

[16] Stanoevska-Slabeva K, Wozniak T. In: Grid and Cloud Computing: Business Perspective on Technology and Applications. Stanoevska K, Wozniak T, Ristol S, editor. Berlin: Springer; 2010. Cloud Basics - An Introduction to Cloud Computing; pp. 47–61.

[17] Truong HL, Dustdar S. On Analyzing and Specifying Concerns for Data as a Service. 2009 Ieee Asia-Pacific Services Computing Conference (Apscc 2009) 2009. pp. 83–90.
[18] DaaS: The New Information Goldmine.
[19] Fusaro VA, Patil P, Gafni E, Wall DP, Tonellato PJ. Biomedical cloud computing with Amazon Web Services. PLoS Comput Biol. 2011;7(8):e1002147. doi: 10.1371/journal.pcbi.1002147.
[20] Nguyen T, Shi W, Ruden D. CloudAligner: a fast and full-featured MapReduce based tool for sequence mapping. BMC Res Notes. 2011;4:171. doi: 10.1186/1756-0500-4-171.
[21] Schatz MC. CloudBurst: highly sensitive read mapping with MapReduce. Bioinformatics. 2009;25(11):1363–1369.
[22] Langmead B, Schatz MC, Lin J, Pop M, Salzberg SL. Searching for SNPs with cloud computing. Genome Biol. 2009;10(11):R134. doi: 10.1186/gb-2009-10-11-r134.
[23] Matsunaga A, Tsugawa M, Fortes J. Combining MapReduce and Virtualization on Distributed Resources for Bioinformatics Applications. Fourth IEEE International Conference on eScience. 2008. pp. 222–229.
[24] 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. doi: 10.1093/bioinformatics/bts023.
[25] Langmead B, Hansen KD, Leek JT. Cloud-scale RNA-sequencing differential expression analysis with Myrna. Genome Biol. 2010;11(8):R83. doi: 10.1186/gb-2010-11-8-r83.
[26] Zhang L, Gu S, Liu Y, Wang B, Azuaje F. Gene set analysis in the cloud. Bioinformatics. 2012;28(2):294–295.
[27] Wall DP, Kudtarkar P, Fusaro VA, Pivovarov R, Patil P, Tonellato PJ. Cloud computing for comparative genomics. BMC Bioinformatics. 2010;11:259. doi: 10.1186/1471-2105-11-259.
[28] Feng X, Grossman R, Stein L. PeakRanger: a cloud-enabled peak caller for ChIP-seq data. BMC Bioinformatics. 2011;12:139. doi: 10.1186/1471-2105-12-139.
[29] 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.
[30] 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. doi: 10.1093/bioinformatics/btr091.
[31] Jourdren L, Bernard M, Dillies M-A, Le Crom S. Eoulsan: a cloud computing-based framework facilitating high throughput sequencing analyses. Bioinformatics. 2012. published online April 5, 2012. 2010.1093/bioinformatics/bts2165.
[32] 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. doi: 10.1038/nbt.2028.
[33] 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. doi: 10.1186/1471-2105-11-S12-S4.
[34] 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. doi: 10.1186/1471-2105-13-42.
[35] 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. doi: 10.1186/1471-2105-12-356.
[36] Dudley JT, Pouliot Y, Chen R, Morgan AA, Butte AJ. Translational bioinformatics in the cloud: an affordable alternative. Genome Med. 2010;2(8):51. doi: 10.1186/gm172.
[37] Zhang Z, Bajic VB, Yu J, Cheung K-H, Townsend JP. In: Bioinformatics - Trends and Methodologies. Mahdavi MA, editor. Rijeka, Croatia: InTech - Open Access Publisher; 2011. Data Integration in Bioinformatics: Current Efforts and Challenges.
[38] Fox A. Cloud computing-what's in it for me as a scientist? Science. 2011;331(6016):406–407. doi: 10.1126/science.1198981.
[39] Deorowicz S, Grabowski S. Compression of DNA sequence reads in FASTQ format. Bioinformatics. 2011;27(6):860–862.
[40] Cox AJ, Bauer MJ, Jakobi T, Rosone G. Large-scale compression of genomic sequence databases with the Burrows-Wheeler transform. Bioinformatics. 2012;28(11):1415–1419. doi: 10.1093/bioinformatics/bts173.
[41] Langille MGI, Eisen JA. BioTorrents: a file sharing service for scientific data. PLoS One. 2010;5(4):e10071. doi: 10.1371/journal.pone.0010071.
[42] Sangket U, Phongdara A, Chotigeat W, Nathan D, Kim WY, Bhak J, Ngamphiw C, Tongsima S, Khan AM, Lin H. et al. Automatic synchronization and distribution of biological databases and software over low-bandwidth networks among developing countries. Bioinformatics. 2008;24(2):299–301.
[43] Bishop M. e-Science. Brief Bioinform. 2003;4(3):208–209. doi: 10.1093/bib/4.3.208.
[44] Zhang Z, Cheung KH, Townsend JP. Bringing Web 2.0 to bioinformatics. Brief Bioinform. 2009;10(1):1–10.
[45] Marx V. My data are your data. Nat Biotechnol. 2012;30(6):509–511. doi: 10.1038/nbt.2243.
[46] Rosenthal A, Mork P, Li MH, Stanford J, Koester D, Reynolds P. Cloud computing: a new business paradigm for biomedical information sharing. J Biomed Inform. 2010;43(2):342–353. doi: 10.1016/j.jbi.2009.08.014.
[47] Dillon T, Wu C, Chang E. Cloud Computing: Issues and Challenges. Int Con Adv Info Net. 2011. pp. 27–33.
[48] Parameswaran AV, Chaddha A. Cloud interoperability and standardization. SETLabs Briefings. 2009;7(7):19–26