Istanbul Journal of Innovation in Education

Volume 1 Issue 1 January, 2015 pp. 29-38

Free Drops from Cloud in Bioinformatics

Dr. Murat GEZER*, Istanbul University Informatics Department, Istanbul-TURKEY
Serra ÇELİK, Istanbul University Informatics Department, Istanbul-TURKEY
Dr. Çiğdem Selçukcan EROL, Istanbul University Informatics Department, Istanbul-TURKEY

Abstract: The need for the benefits of the cloud technology is in almost every discipline, which data size is gradually increasing. Bioinformatics is a field that can produce more data every passing day as a result of emerging scientific advances (high-throughput technologies, etc.). Processing and sharing data is as much important as storing data which can produce results affecting all creatures, particularly human being. In today's technologies, the road to the light passes through cloud. It is seen that many cloud solutions special to bioinformatics have been created in recent years. These can emerge as software, platform, or infrastructure solutions. In this study, it is aimed to determine positive and negative sides by comparing free cloud infrastructure systems used for bioinformatics data. For this purpose, cloud solutions that can meet the needs of bioinformatics field will be briefly mentioned by giving information about cloud information technologies and free infrastructure solutions will be compared. Consequently, the infrastructure to be established should have support through web in order to make a selection between compared systems. Apart from this, if the software needed in bioinformatics is found as predefined, this will be seen as an important reason for preference for the cloud infrastructure system to be used.

Keywords: Cloud Computing, Bioinformatics, Cloud Biolinux, Cloudman CloVR.

1. Introduction

As information technologies evolve, the needs are increasing and a cycle encouraging development will be created. Growth rate of the amount of the data that is produced, resulting in the need of further improvement in data processing and storage capacity of computers. When these needs are evaluated from different perspectives (like financial, place, etc.), it will be seen that Cloud Computing-based solution systems are offered. There are some alternative cloud structures in bioinformatics as one of the

* Contact Author: murat.gezer@istanbul.edu.tr, Istanbul University Informatics Department, 34134, Istanbul-TURKEY

areas where these systems are used. Some of these are software solutions, platform solutions, or infrastructure solutions.

In the field of bioinformatics, different data types and data analysis may require different compute need. For example, while analysis of next-generation sequencing (NGS) requires more RAM, CPU can be a more important limit for search with BLAST (Afgan et al., 2012). There are cloud infrastructure systems offered free (Cloud Biolinux, CloVR, etc.) or paid (Amazın EC2, Windows Azure, etc.) for meeting infrastructure requirements. In this study, information regarding cloud computing technologies is provided, cloud solutions that can meet the needs in the field of bioinformatics are briefly mentioned, and free infrastructure systems are compared.

2. Cloud Computing

In recent years, large-scaled projects can be carried out with super computers having high computing capability. High performance computing can be made with less expensive methods known as commodity cluster or grid in contrast to high costs of super computers. Although it doesn't have a standard definition, Cloud Computing is a datacenter hardware, software, and systems distributed for massive data processing (Armbrust et al., 2009). Cloud computing provides computing alternative to the researcher without the need to establish a physical infrastructure. Cloud computing offers less expensive alternative to super computer and customized clusters, a more reliable platform compared to grids and more scalability compared to biggest commodity cluster or weld pools (Ostermann et al., 2009).

Grid computing, the data processing structure before the development of cloud computing, can be considered as a revolution coming after internet and World Wide Web. Grid computing provides a combination of sources (super computers, storage systems, data sources, and devices customized for different fields) distributed geographically (Chetty & Buyya, 2002). Cloud computing, which is a reliable service distributed with new generation data centers based on virtualized computing and storing technologies, has emerged. Users can access data and applications from a cloud in any part of the world (Weiss, 2007). Security and privacy are important issues. Interoperability, multi-platform, multi-application, and multi-provider support services are also important.

Building and operating of cloud systems can be possible with expensive ventures. Since the beginning of 2000s, with the development of web services, important internet companies like Amazon, eBay, Google, and Microsoft using these systems have developed scalable infrastructure software (such as MapReduce, the Google File System, BigTable, and Dynamo) (Armbrust et al., 2009).

Cloud computing provides the most appropriate use for many computers in terms of offering proper and optional access to the sources (computing, storing, serves, and etc.) offered with Web Application Programming Interfaces (API) (Buyya et al., 2009). Cloud computing has a computing infrastructure accessed with a network like internet or where some actions such as managing, sharing application, or developing platform (McDonald, 2010). Cloud is seen as a single access point for the computing need of the user and the most widespread services are as follows (Rimal et al., 2009):

1) Software as a Service (SaaS): It is a multi-tenant platform. It uses object code sample under database that supports common sources and a large number of customers simultaneously. SaaS is generally defined as ASP (Application Service Provider) model. Examples for important providers: SalesForce.com (SFDC), NetSuite, Oracle, IBM, Microsoft.

2) Platform as a Service (PaaS): Cloud system provides a platform developer covering all system and environments in improving, testing, applying, and hosting the last life cycle of web applications developing service from end to end. Google App Engine GAE and Microsoft's Azure can be given as examples.

3) Hardware as a Service (HaaS): It offers service to users without forcing them to make building and datacenter management investment.

4) Infrastructure as a Service (IaaS): IaaS distributes the computer infrastructure. Its high flexibility is its most important benefit. It has a user-based payment structure. Customers pay as much as they use. It always uses the latest technology. Customers gain faster service distribution and

more time. GoGrid, Flexiscale, Layered Technologies, Joyent, and Mosso/ Rackspace can be given as examples.

3. Cloud Systems in Bionformatics

The developments occurring in science world trigger each other. Therefore, in today's world where interdisciplinary studies are becoming more important, bioinformatics includes all processes such as storing data in the fields of biology and especially molecular biology and converting them into information in cooperation with disciplines such as computer sciences, mathematics, and statistics. Thanks to technological advances, sequencing is becoming faster, less expensive, and produced data size is rapidly increasing. However, analysis rate of these data is limited with the capacity of computers. Cloud service solutions are useful in removing this limitation. There are many free or paid cloud resources developed for bioinformatics in recent years. These services are given in Table 1 by combining from Dai et al. (2012) and Lin et al. (2013)'s reviews.

Resource	Description	References
	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.; http://aws.amazon.com/publicdatasets	
	Software as a Service (SaaS):	
BGI Cloud	Cloud-based implementations of various genomic analysis applications; http://cloud.genomics.cn	
CloudAligner	Fast and full-featured MapReduce-based tool for sequence mapping; http://cloudaligner.sourceforge.net	Nguyen et al., 2011
CloudBLAST	A cloud-based implementation of NCBI BLAST; http://ammatsun.acis.ufl.edu/amwiki/index.php/CloudBLAST_Project	Matsunaga et al., 2008
CloudBurst	Highly sensitive short read mapping with MapReduce; http://cloudburst-bio.sourceforge.net	Schatz, 2009
Contrail	Cloud-based de novo assembly of large genomes; http://contrail-bio.sourceforge.net	Schatz et al., 2010
Crossbow	Read Mapping and SNP calling using cloud computing; http://bowtie-bio.sf.net/crossbow	Langmead et al., 2009
EasyGenomics	Cloud-based NGS pipelines for whole genome resequencing, exome resequencing, RNA-Seq, small RNA and de novo assembly; http://www.easygenomics.org	
eCEO	Cloud-based identification of large-scale epistatic interactions in genome-wide association study (GWAS); http://www.comp. nus.edu.sg/~wangzk/eCEO.html	Wang et al., 2011

FX	RNA-Seq analysis tool; http://fx.gmi.ac.kr	Hong et al., 2012
Gaea	Cloud-based genome re-sequencing assembly; http://bgiamericas.com/data-analysis/cloud-computing	
Hecate	Cloud-based de novo assembly; http://bgiamericas.com/data-analysis/cloud-computing	
Jnomics	Cloud-scale sequence analysis suite based on Apache Hadoop; http://sourceforge.net/apps/mediawiki/jnomics	
Myrna	Differential gene expression tool for RNA-Seq; http://bowtie-bio.sourceforge.net/myrna	Langmead et al., 2010
PeakRanger	Cloud-enabled peak caller for ChIP-seq data; http://ranger.sourceforge.net/	Feng et al., 2011
VAT	Variant annotation tool to functionally annotate variants from multiple personal genomes at the transcript level; http://vat.gersteinlab.org	Habegger et al., 2012
YunBe	Pathway-based or gene set analysis of expression data; http://tinyurl.com/yunbedownload	Zhang et al., 2012
Cloud-Cofee	Multiple sequence alignment http://www.tcofee.org/	Tommaso et al., 2010
SEAL	Short read mapping and duplicate removal http://biodoop-seal.sourceforge.net/	Pireddu et al., 2011
Quake	Quality-aware detection and correction of sequencing errors http://www.cbcb.umd.edu/sotware/quake/	Kelley et al., 2010
ArrayExpressHTS	RNA-seq process and quality assessment http://www.ebi.ac.uk/services	Goncalves et al., 2011
BioVLab	A virtual collaborative lab for biomedical applications https://sites.google.com/site/biovlab/	Lee et al., 2012
Hadoop-BAM	Directly manipulate NGS data http://sourceforge.net/projects/hadoop-bam/	Niemenmaa et al., 2012
SeqWare	A scalable NoSQL database for NGS data http://seqware.sourceforge.net	O’Connor et al., 2010
GATK	Genome analysis toolkit http://www.broadinstitute.org/gatk/	McKenna et al., 2010
	Platform as a Service (PaaS):	
Eoulsan	Cloud-based platform for high throughput sequencing analyses; http://transcriptome.ens.fr/eoulsan	Jourdren et al., 2012
Galaxy Cloud (CloudMan)	Cloud-scale Galaxy for large-scale data analysis; http://galaxy.psu.edu	Afgan et al., 2010 Afgan et al., 2011
	Infrastructure as a Service (IaaS):	
Cloud BioLinux	A publicly accessible virtual machine for high performance bioinformatics computing using cloud platforms; http://cloudbiolinux.org	Krampis et al., 2012
CloVR	A portable virtual machine for automated sequence analysis using cloud computing; http://clovr.org	Angiuoli et al., 2011

Table 1: Cloud resources in bioinformatics (Dai et al. 2012; Lin et al., 2013)

4. Comparison of Free Cloud Systems in Bionformatics

There are Linux-based operating systems called Cloud BioLinux and CloVR as infrastructure systems offered with free software licensing in order to carry out the analyses of bioinformatics data on Cloud (Dai et al. 2012; Lin et al., 2013). Receiving a free or paid web service is necessary in order to use these infrastructure services. Amazon (free or paid) or Eucalyptus (free), web services found in market, can be used in order that infrastructure systems, which are the subjects of this study, can operate. Eucalyptus is an AWS compliant service that can be used to create our own cloud. Within the scope of this study, Amazon which is an installed system has been used because of hardware limitation. Cloud BioLinux was developed by J.Craig Venter Institute. CloVR is a Linux distribution developed by Maryland University. It has been stated on the websites of the two systems that the both infrastructures can be downloaded through internet. However, in the period in which the study was being carried out, virtualbox image couldn’t be reached because of an error message appearing on the link of the related website of Cloud BioLinux virtual box version (Cloud BioLinux 32-bit VirtualBox appliance). Therefore, Image was established by signing up to Free Usage Tier service of Amazon Web Services (AWS) through Machine Images connection found on the same website. Amazon Company offers this service freely within the limitations specified on their websites (AWS Free Usage Tier, 2014).

On CloVR website, there are both virtualbox and vmware images. By following the route map found on the website they can be run on CloVR AWS. In this study, the two infrastructure systems are compared.

Cloud Bio Linux runs on more current Ubuntu 12.04, 12.10, or 13.04 distributions, while CloVR runs on Ubuntu 10.04 Linux distribution, which doesn’t have an updating support by Ubuntu and has also lost its currency.

Both systems prefer offering their documents through web environment. CloVR offers more understandable and well-coordinated information to its end users. Communication between users and developers is provided via e-mail lists. On Cloud BioLinux, Documentation is carried out through the direction of different websites and the communication between users and developers is provided through Google groups.

Both systems have different software apart from the software they use commonly for bioinformatics researches. Systems can be compared from many aspects but in the comparison made in this study it is thought that issues such as installation difficulty or difficulty in accessing support can be solved by spending more time and effort. It is also thought that whether having the needed bioinformatics analysis tools or not is the key point.

5. Conclusion

Cloud systems offer different drops just like rain about the requirements of bioinformatics data. These drops can sometimes be used together. However, choosing the drops to be used can be a problem. From an external perspective, all drops resemble each other and it can be difficult to find the right drop without getting wet. In this study, a few drops offered freely are examined and it has been aimed to show the way to researchers which have hesitation about choosing the right cloud system for bioinformatics data set. Consequently, the infrastructure to be established should have a support through web in order to make a selection between compared systems. In addition, it can be seen as an important reason for preference for the cloud infrastructure system to be used if the software needed in the field of bioinformatics is predefined.

6. Acknowledgements

This work was supported by Research Fund of the Istanbul University. Project number UDP-27318.

References

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

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

Afgan, E., Chapman, B., Jadan, M., Franke, V., Taylor, J. (2012). Using cloud computing infrastructure with CloudBioLinux, CloudMan, and Galaxy. Current Protocols in Bioinformatics, 11-9.

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

Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R. H., Konwinski, A., Lee, G., Patterson, D. A., Rabkin, A., Stoica, I., Zaharia, M., (2009), Above the Clouds: A Berkeley View of Cloud Computing, Electrical Engineering and Computer Sciences University of California at Berkeley, Technical Report No. UCB/EECS-2009-28 http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.html February 10, 2009.

Bradshaw, R., Desai, N., Freeman, T., Keahey, K ., (2007), “A scalable approach to deploying and managi ng appliances”, In: TeraGrid Conference 2007

Buyya R., Yeo C. S., Venugopal S, Broberg J, Brandic, (2009), “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.

Chetty, M., Buyya, R., (2002), Weaving Computational Grids: How Analogous Are They with Electrical Grids?, Computing in Science and Engineering, *4*(4), 61-71.

Dai, L., Gao, X., Guo, Y., Xiao, J., Zhang, Z., (2012), “Bioinformatics clouds for big data manipulation”, Biology Direct, 7, 43.

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

Goncalves, A., A. Tikhonov, A. Brazma, and M. Kapushesky, (2011) “A pipeline for RNA-seq data processing and quality assessment,” Bioinformatics, *27*(6), 867–869.

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

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

Jourdren L, Bernard M, Dillies M-A, Le Crom S (2012), Eoulsan: a cloud computing-based framework facilitating high throughput sequencing analyses. Bioinformatics. doi:2010.1093/bioinformatics/bts2165.

Kelley, D.R., M.C.Schatz,and S.L.Salzberg, (2010) ,“Quake:quality-aware detection and correction of sequencing errors,” Genome Biology, *11*(11), article R116.

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

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

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

Lee, H., Y. Yang, H. Chae, (2012) "BioVLAB-MMIA: a cloud environment for microRNA and mRNA integrated analysis (MMIA) on Amazon EC2," IEEE Transactions on Nanobio-science, *11*(3), 266–272.

Lin Y-C, Yu, C-S, Lin, Y-J, (2013), "Enabling Large-Scale Biomedical Analysis in the Cloud", BioMed Research International, 1-6, http://dx.doi.org/10.1155/2013/185679

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

McDonald, K. T., (2010), Above the Clouds Managing Risk in the World of Cloud Computing, IT Governance Publishing.

McKenna, A., M. Hanna, E. Banks, (2010) "genome analysis toolkit: aMapReduce framework for analyzing next-generation DNA sequencing data," Genome Research, *20*(9), 1297–1303.

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

Niemenmaa, M. A. Kallio, A. Schumacher, P. Klemel¨ a, E. Korpelainen, and K. Heljanko, (2012) "Hadoop-BAM: directly manipulating next generation sequencing data in the cloud," Bioinformatics, *28*(6), 876–877.

O'Connor, B.D., B.Merriman, and S. F.Nelson, (2010) "SeqWareQuery Engine: storing and searching sequence data in the cloud," BMC Bioinformatics, *11*(12), articleS2.

Ostermann, S., Iosup, A., Yigitbasi, N., Prod, R., Fahringer, T., Eperna, D., Avresky, D. R., et al., (2009), A Performance Analysis of EC2 Cloud Computing Services for Scientific Computing, Cloudcomp 2009, LNICST, 34, 115- 131.

Pireddu, L., S. Leo, and G. Zanetti, (2011), "Seal:adistributedshortread mapping and duplicate removal tool," Bioinformatics, *27*(15), 2159–2160.

Rimal, B. P., Choi, E., Lumb, I., (2009), "A Taxonomy and Survey of Cloud Computing Systems", Fifth International Joint Conference on INC, IMS and IDC.

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

Schatz, M.C., A.L.Delcher, and S.L.Salzberg, (2010), "Assembly of large genomes using second-generation sequencing," Genome Research, *20*(9), 1165–1173.

Tommaso, P.di, M.Orobitg,F.Guirado,F.Cores,T.Espinosa, and C.Notredame, (2010), "Cloud-Cofee: implementation of a parallel consistency-basedmultiple alignment algorithmin the T-cofee package and its benchmarking on the Amazon Elastic-Cloud," Bioinformatics, *26*(15), 1903–1904.

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*(8), 1045–1051.

Weiss A., (2007), Computing in the Clouds. netWorker, *11*(4), 16-25

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

URL-1: AWS Free Usage Tier (2014). Retrieved April 6,2014,from http://aws.amazon.com/free/

URL-2: CloudLinux Included Open Source Software Packages (2013). Retrieved April 6, 2014, from http://www.jcvi.org/cms/research/projects/jcvi-cloud-biolinux/included-software/

URL-3: ClovVR Edition Comparison (2013). Retrieved April 6, 2014, from http://clovr.org/developers/edition-comparison/