Methods in Molecular Biology (2016) 1375: 25–39
DOI 10.1007/7651_2015_236

Published online: 12 April 2015

Bioinformatics and Microarray Data Analysis on the Cloud

Barbara Calabrese and Mario Cannataro

Abstract

High-throughput platforms such as microarray, mass spectrometry, and next-generation sequencing are producing an increasing volume of omics data that needs large data storage and computing power. Cloud computing offers massive scalable computing and storage, data sharing, on-demand anytime and anywhere access to resources and applications, and thus, it may represent the key technology for facing those issues. In fact, in the recent years it has been adopted for the deployment of different bioinformatics solutions and services both in academia and in the industry. Although this, cloud computing presents several issues regarding the security and privacy of data, that are particularly important when analyzing patients data, such as in personalized medicine. This chapter reviews main academic and industrial cloud-based bioinformatics solutions; with a special focus on microarray data analysis solutions and underlines main issues and problems related to the use of such platforms for the storage and analysis of patients data.

Keywords: Cloud computing, Bioinformatics, Microarray data analysis

1 Introduction

High-throughput platforms for the investigation of the cell machinery, such as mass spectrometry, microarray, and next-generation sequencing, yielded to the so-called "omics" sciences. In particular, genomics regards the study of the activity of genes, proteomics the study of the activity of proteins, and interactomics the study of protein interactions inside a cell. Pharmacogenomics is an important branch of genomics that studies the impact of genetic variation (e.g., Single Nucleotide Polymorphisms—SNPs) on drug response in patients and is at the basis of the so-called "personalized medicine," where drugs are chosen or optimized to meet the genetic profile of each patient.

The availability of such high-throughput technologies and the application of genomics and pharmacogenomics studies of large populations, are producing an increasing amount of experimental and clinical data, as well as specialized databases spread over the Internet. However, the storage, preprocessing, and analysis of experimental data are becoming the main bottleneck of the analysis pipeline.

Managing omics data requires both space for data storing and services for data preprocessing, analysis, and sharing. The resulting scenario comprises a set of bioinformatics tools, often implemented as Web services, for the management and analysis of data stored in geographically distributed biological databases.

Cloud computing is a computing model that has spread very rapidly in recent years for the supply of IT resources (hardware and software) of different nature, through services accessible via the network. The resources that a cloud system provides to users include: CPU, memory, networks, operating systems, middleware, and applications. The cloud resources are dynamically scalable, virtualized, and accessible on the Internet (1). This model provides new advantages related to massive and scalable computing resources available on demand, virtualization technology, and payment for use as needed (2).

Thus, cloud computing may play an important role in many phases of the bioinformatics analysis pipeline, from data management and processing, to data integration and analysis, including data exploration and visualization.

Despite the many benefits associated with cloud computing, there are also several management, technology, security, and legal issues to be addressed. In fact, cloud computing currently presents some issues and open problems such as privacy and security, geographical localization of data, legal responsibilities in the case of data leaks, that are particularly important when managing sensitive data such as the patients data stored and processed in genomics and pharmacogenomics studies, and more in general when clinical data are transferred to the cloud.

The aim of this chapter is to describe and discuss the most significant applications of cloud computing in the bioinformatics with special focus on microarray data analysis. The chapter focuses on specific requirements and issues of such applications on cloud computing. The chapter is organized as follows: in Section 2 cloud computing definition is discussed. Service and delivery models are presented in order to define the cloud-related background. Successively, in Section 3 the chapter focuses on the application of cloud computing in bioinformatics and microarray data analysis. Section 4 summarizes the main problems to be faced when moving bioinformatics applications on the cloud and underlines open problems related to the full adoption of cloud computing in the bioinformatics data analysis pipeline.

2 Materials

Even though cloud computing is now becoming the key technology for the storage and analysis of large data sets both in academia and industry, it is not a totally new concept. In fact, it has some

relations with grid computing and other technologies such as utility computing, clustering, virtualization systems, and distributed systems. There are many definitions of cloud computing. The first definition comes from the work of Mell and Grance (1) and is a popular working definition of cloud computing from the National Institute of Standards and Technology, US Department of Commerce. Their definition focuses on computing resources that can be accessed from anywhere and may be provisioned online. It also specifies five characteristics of cloud computing (i.e., on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service), three service models (i.e., Software as a Service, Platform as a Service, and Infrastructure as a Service) and four deployment methods (i.e., private cloud, community cloud, public cloud, and hybrid cloud). Most of the other definitions do not mention deployment methods. In contrast to other definitions, this one does not explicitly mention virtualization as a key technology.

Vaquero et al. (3) collected 22 excerpts from previous works and fused these into a single definition by studying the common properties of cloud computing. This definition emphasizes the importance of Service Level Agreements (SLA) in order to increase confidence in the cloud environment and defines virtualization as the key enabler of cloud computing.

2.1 Service Models

Cloud services can be classified into three main models:

- Infrastructure as a Service (IaaS): this service model is offered in a computing infrastructure that includes servers (typically virtualized) with specific computational capability and/or storage. The user controls all the storage resources, operating systems, and applications deployed to, while he/she has limited control over the network settings. An example is Amazon's Elastic Compute Cloud (EC2), which allows the user to create virtual machines and manage them, and Amazon Simple Storage Service (S3), which allows storing and accessing data, through a Web-service interface.
- Platform as a Service (PaaS): it allows the development, installation and execution on its infrastructure of user-developed applications. Applications must be created using programming languages, libraries, services, and tools supported by the provider that constitute the development platform provided as a service. An example is Google Apps Engine, which allows developing applications in Java and Python and provides for both languages the SDK (Software Development Kit) and uses a plugin for the Eclipse development environment.
- Software as a Service (SaaS): customers can use the applications provided by the cloud provider infrastructure. The applications are accessible through a specific interface. Customers do not

manage the cloud infrastructure or network components, servers, operating systems, or storage. In some cases, it is possible to manage specific configurations of the application.

2.2 Delivery Models

Cloud services can be made available to users in different ways. In the following, a brief description of the delivery models is presented:

- Public Cloud: vendors who provide the users/customers the hardware and software resources of their data centers offer public cloud services. Examples of public clouds are Amazon, Google Apps, and Microsoft Azure.
- Private Cloud: private cloud is configured by a user or by an organization for its exclusive use. Computers that are in the domain of the organization supply services. To install a private cloud, several commercial and free tools are available (e.g., OpenStack, Eucalyptus, Open Nebula, Terracotta, and VMware Cloud).
- Community Cloud: it is an infrastructure on which are installed cloud services shared by a community or by a set of individuals, companies and organizations that share a common purpose and that have the same needs. The cloud can be managed by the community itself or by a third party (typically a cloud service provider).
- Hybrid Cloud: the cloud infrastructure is made up of two or more different clouds using different delivery models, which, while remaining separate entities, are connected by proprietary or standard technology that enables the portability of data and applications.

3 Methods

High-throughput platforms for the investigation of the cell machinery, such as mass spectrometry, microarray, and next-generation sequencing, are producing an overwhelming amount of the so-called "omics" data (4). In particular, genomics regards the study of the activity of genes, proteomics the study of the activity of proteins, and interactomics the study of protein interactions inside a cell (5).

Pharmacogenomics is an important branch of genomics that studies the impact of genetic variation (e.g., Single Nucleotide Polymorphisms—SNPs) on drug response in patients and is at the basis of the so-called "personalized medicine," where drugs are chosen or optimized to meet the genetic profile of each patient.

Pharmacogenomics correlates gene expression or SNPs with the toxicity or efficacy of a drug, with the aim to improve drug therapy with respect to the patients' genotype, to ensure maximum efficacy with minimal adverse effects. Many works demonstrated a correlation between the presence/absence of SNPs and the development of diseases, as well as the effectiveness of drugs (6).

Thus the presence (or the absence) of specific SNPs may be used as a clinical marker for the prediction of drug effectiveness, foreseeing the response of individuals with different SNPs to drugs.

The availability of such high-throughput technologies and the application of genomics and pharmacogenomics studies of large populations, are producing an increasing amount of experimental and clinical data, as well as specialized databases spread over the Internet. However, the storage, preprocessing, and analysis of experimental data are becoming the main bottleneck of the analysis pipeline.

Managing omics data requires both space for data storing and procedures for data preprocessing, analysis, and sharing. The resulting scenario comprises a set of bioinformatics tools, often implemented as Web services, for the management and analysis of data stored in geographically distributed biological databases.

The main challenges regard: (1) the efficient storage, retrieval, and integration of experimental data; (2) their efficient and high-throughput preprocessing and analysis; (3) the building of reproducible "in silico" experiments; (4) the annotation of omics data with preexisting knowledge stored into ontologies (e.g., Gene Ontology) or specialized databases; (5) the integration of omics and clinical data.

Cloud computing may play an important role in many phases of the analysis pipeline, from data management and processing, to data integration and analysis, including data exploration and visualization.

Currently, high performance computing is used to face the large processing power required when processing omics data, while Web services and workflows are used to face the complexity of the bioinformatics pipeline that comprises several steps. Cloud computing may be the glue that put together those mainstreams technologies already used in bioinformatics (parallelism, service orientations, knowledge management), with the elasticity and ubiquity made available by the cloud.

Cloud computing represents a cost-effective solution for the problems of storing and processing data in the context of bioinformatics. Classical computational infrastructure for data processing has become ineffective and difficult to maintain (7, 8). Dudley and his colleagues (9) demonstrated that cloud computing is a viable and cheaper technology that enables large-scale integration and analysis for studies in genomic medicine.

On the other hand, cloud computing presents some issues and open problems such as privacy and security, geographical localization of data, legal responsibilities in the case of data leaks, that are particularly important when managing sensitive data as the patients data stored and processed in genomics and pharmacogenomics studies and more in general when clinical data are transferred to the cloud.

In the following sections, the main cloud-based applications proposed in the fields of bioinformatics are illustrated and open problems related to the full adoption of cloud computing in bioinformatics are underlined.

3.1 Cloud-Based Bioinformatics Solutions

The traditional bioinformatics analysis involves downloading of public datasets (e.g., NCBI, Ensembl), installing software locally and analysis in-house. By entering the data and software in the cloud and providing them as a service, it is possible to get a level of integration that improves the analysis and the storage of bioinformatics big-data. In particular, as a result of this unprecedented growth of data, the provision of data as a service (*Data as a Service*, *DaaS*) is of extreme importance. DaaS provides data storage in a dynamic virtual space hosted by the cloud and allows to have updated data that are accessible from a wide range of connected devices on the Web. An example is represented by the DaaS of Amazon Web Services (AWS, http://aws.amazon.com/publicdatasets), which provides a centralized repository of public data sets, including archives of GenBank, Ensembl, 1000 Genomes Project, Unigene, Influenza Virus (10).

In the following subsections, examples of SaaS, PaaS, and IaaS for several tasks in bioinformatics domain are presented.

3.2 Bioinformatics Tools Deployed as SaaS

In recent years, there have been several efforts to develop cloud-based tools to execute different bioinformatics tasks (11), e.g., mapping applications, sequences alignment, gene expression analysis (12). Some examples of SaaS bioinformatics tools are reported in the following.

In ref. (13), the authors propose an efficient Cloud-based Epistasis cOmputing (*eCEO*) model for large-scale epistatic interaction in genome-wide association study (GWAS). Given a large number of combinations of SNPs (Single-nucleotide polymorphism), eCEO model is able to distribute them to balance the load across the processing nodes. Moreover, eCEO model can efficiently process each combination of SNPs to determine the significance of its association with the phenotype. The authors have implemented and evaluated eCEO model on their own cluster of more than 40 nodes. The experiment results demonstrate that the eCEO model is computationally efficient, flexible, scalable, and practical. In addition, the authors have also deployed the eCEO model on the Amazon Elastic Compute Cloud.

STORMSeq (Scalable Tools for Open—Source Read Mapping) (14), is a graphical interface cloud computing solution that performs read mapping, read cleaning and variant calling and annotation with personal genome data. At present, STORMSeq costs approximately 2 dollars and 510 h to process a full exome sequence and 30 dollars and 38 days to process a whole genome sequence. The authors provide this open-access and open-source resource as a user-friendly interface in Amazon EC2.

CloudBurst (15) and *CloudAligner* (16) are parallel read-mapping algorithms optimized for mapping next-generation sequence (NGS) data to the human genome and other reference genomes, for use in a variety of biological analyses including SNP discovery, genotyping, and personal genomics. They use the open-source Hadoop implementation of MapReduce to parallelize execution using multiple compute nodes. Specifically, CloudAligner has been designed for more long sequences. An other Hadoop-based tool is *Crossbow* (17) that combines the speed of the short read aligner Bowtie with the accuracy of the SNP caller SOAPsnp to perform alignment and SNP detection for multiple whole-human datasets per day.

VAT (Variant Annotation Tool) (18) has been developed to functionally annotate variants from multiple personal genomes at the transcript level as well as to obtain summary statistics across genes and individuals. VAT also allows visualization of the effects of different variants, integrates allele frequencies and genotype data from the underlying individuals and facilitates comparative analysis between different groups of individuals. VAT can either be run through a command-line interface or as a Web application. Finally, in order to enable on-demand access and to minimize unnecessary transfers of large data files, VAT can be run as a virtual machine in a cloud-computing environment.

FX (19) is an RNA-Seq analysis tool, which runs in parallel on cloud computing infrastructure, for the estimation of gene expression levels and genomic variant calling. FX allows analysis of RNA-Seq data on cloud computing infrastructures, supporting access through a user-friendly Web interface. An other cloud-computing pipeline for calculating differential gene expression in large RNA-Seq datasets is *Myrna* (20). Myrna integrates short read alignment with interval calculations, normalization, aggregation, and statistical modeling in a single computational pipeline. After alignment, Myrna calculates coverage for exons, genes, or coding regions and differential expression using either parametric or nonparametric permutation tests. Myrna exploits the availability of multiple computers and processors where possible and can be run on the cloud using Amazon Elastic MapReduce, on any Hadoop cluster, or on a single computer (bypassing Hadoop entirely).

PeakRanger (21) is a software package for the analysis in Chromatin Immunoprecipitation Sequencing (ChIP-seq) technique. This technique is related to NGS and allows investigating the interactions between proteins and DNA. Specifically, PeakRanger is a peak caller software package that can be run in a parallel cloud computing environment to obtain extremely high performance on very large data sets.

For spectrometry-based proteomics research, *ProteoCloud* (22) is a freely available, full-featured cloud-based platform to perform computationally intensive, exhaustive searches using five different peptide identification algorithms. ProteoCloud is entirely open-source and is built around an easy-to-use and cross-platform software client with a rich graphical user interface. This client allows full control of the number of cloud instances to initiate and of the spectra to assign for identification. It also enables the user to track progress, and to visualize and interpret the results in detail.

An environment for the integrated analysis of microRNA and mRNA expression data is provided by *BioVLAB-MMIA* (23). Recently, a new version called BioVLAB-NGS, deployed on both Amazon cloud and on a high performance, publically available server called MAHA, has been developed (24). By utilizing next generation sequencing (NGS) data and integrating various bioinformatics tools and databases, BioVLAB-MMIA-NGS offers several advantages, such as a more accurate data sequencing for determining miRNA expression levels or the implementation of various computational methods for characterizing miRNAs.

Cloud4SNP (25) is a novel Cloud-based bioinformatics tool for the parallel preprocessing and statistical analysis of pharmacogenomics SNP microarray data. Cloud4SNP is able to perform statistical tests in parallel, by partitioning the input data set and using the virtual servers made available by the Cloud. Moreover, different statistical corrections such as Bonferroni, False Discovery Rate, or none correction, can be applied in parallel on the Cloud, allowing the user to choice among different statistical models, implementing a sort of parameter sweep computation.

3.3 Bioinformatics Platforms Deployed as PaaS

Currently, the most used platform (PaaS) for bioinformatics applications is *Galaxy Cloud*, which is a Galaxy cloud-based platform for the analysis of data at a large scale. It allows anyone to run a private Galaxy installation on the Cloud exactly replicating functionality of the main site, but without the need to share computing resources with other users. With Galaxy Cloud, unlike software service solutions, the user can customize their deployment as well as retain complete control over their instances and associated data; the analysis can also be moved to other cloud providers or local resources, avoiding concerns about dependence on a single vendor. Currently, a public Galaxy Cloud deployment, called *CloudMan*, is

provided on the popular Amazon Web Services (AWS) cloud; however, it is compatible with Eucalyptus and other clouds (26). CloudMan (27) enables individual bioinformatics researchers to easily deploy, customize, and share their entire cloud analysis environment, including data, tools, and configurations.

In ref. (28), a modular and scalable framework called *Eoulsan*, based on the Hadoop implementation of the MapReduce algorithm dedicated to high-throughput sequencing data analysis, is presented. Eoulsan allows users to easily set up a cloud computing cluster and automate the analysis of several samples at once using various software solutions available. Tests with Amazon Web Services demonstrated that the computation cost is linear with the number of instances booked as is the running time with the increasing amounts of data. Eoulsan is implemented in Java, supported on Linux systems and distributed under the LGPL License.

Cloud-based bioinformatics applications

Project's name/ref	Services models	Task	URL
eCEO	SaaS	Sequencing (genome resequencing)	www.comp.nus.edu.sg
STORMSEQ	SaaS	Sequencing (genome resequencing)	http://www.stormseq.org
Crossbow	SaaS	Sequencing (genome resequencing)	http://bowtie-bio.sourceforge.net/crossbow/index.shtml
CloudBurst	SaaS	Sequencing: genome resquencing, short-read aligner	http://sourceforge.net/projects/cloudburst-bio/
CloudAligner	SaaS	Sequencing: genome resquencing, short-read aligner	http://sourceforge.net/projects/cloudaligner/
VAT	SaaS	Sequencing: genome resquencing, variant annotation	vat.gersteinlab.org

(continued)

(continued)

Cloud-based bioinformatics applications			
Project's name/ref	**Services models**	**Task**	**URL**
FX	SaaS	Sequencing: RNA-seq	fx.gmi.ac.kr
Myrna	SaaS	Sequencing: RNA-seq	http://bowtie-bio.sourceforge.net/myrna/index.shtml
PeakRanger	SaaS	Sequencing: ChIP SEQ	http://ranger.sourceforge.net
ProteoCloud	SaaS	Mass spectrometry: MS-based proteomics	https://code.google.com/p/proteocloud/
YunBE	SaaS	Transcriptomics: gene set analysis	http://lrcv-crp-sante.s3-website-us-east-1.amazonaws.com
BioVLAB-MMIA	SaaS	Analysis of microRNA and mRNA expression data	https://sites.google.com/site/biovlab/
Cloud4SNP	SaaS	Microarray: SNP Analysis	Not available
CloudMan	PaaS	A public Galaxy cloud deployment for bioinformatics	wiki.galaxyproject.org
Eoulsan	PaaS	A framework for high-throughput sequencing data analysis	http://transcriptome.ens.fr/eoulsan/
Bionimbus	IaaS	A cloud-based infrastructure for managing, analyzing and sharing genomics datasets.	bionimbus.opensciencedatacloud.org/opensciencedatacloud
CloVR	IaaS	A virtual machine for automated and portable microbial	http://clovr.org

(continued)

(continued)

Cloud-based bioinformatics applications			
Project's name/ref	**Services models**	**Task**	**URL**
		sequence analysis	
CloudBioLinux	IaaS	Genome analysis resources for cloud computing platforms	cloudbiolinux.org

3.4 Bioinformatics Tools Deployed as IaaS

Bionimbus (29) is an open-source cloud-computing platform used by a variety of projects to process genomics and phenotypic data. It is based primarily upon OpenStack, which manages on-demand virtual machines that provide the required computational resources, and GlusterFS, which is a high-performance clustered file system. Bionimbus also includes Tukey, which is a portal, and associated middleware that provides a single entry point and a single sign on for the various Bionimbus resources; and Yates, which automates the installation, configuration, and maintenance of the software infrastructure required.

Cloud Virtual Resource, *CloVR*, (30) is a new desktop application for push-button automated sequence analysis that can utilize cloud computing resources. CloVR is implemented as a single portable virtual machine (VM) that provides several automated analysis pipelines for microbial genomics, whole genome and metagenome sequence analysis. The CloVR VM runs on a personal computer, utilizes local computer resources, and requires minimal installation, addressing key challenges in deploying bioinformatics workflows. In addition CloVR supports use of remote cloud computing resources to improve performance for large-scale sequence processing.

Cloud BioLinux (31) is a publicly accessible Virtual Machine (VM) that enables scientists to quickly provision on-demand infrastructures for high-performance bioinformatics computing using cloud platforms. Users have instant access to a range of preconfigured command line and graphical software applications, including a full-featured desktop interface, documentation and over 135 bioinformatics packages for applications including sequence alignment, clustering, assembly, display, editing, and phylogeny. Besides the Amazon EC2 cloud, the authors started instances of Cloud BioLinux on a private Eucalyptus cloud and demonstrated access to the bioinformatics tools interface through a remote connection to EC2 instances from a local desktop computer.

4 Notes

Genomics data extracted by patients' samples, as in pharmacogenomics studies, as well as other clinical data and exams (e.g., bioimages), are sensitive data and present unprecedented requirements of privacy and security.

On the other hand, cloud computing presents some issues and open problems, such as privacy and security, geographical localization of data, legal responsibilities in the case of data leaks, that are particularly important when managing such sensitive data.

In general, genomics and clinical data managed through a cloud are susceptible to unauthorized access and attacks. Specifically, the chapter (32) claims that storing huge volumes of patients' sensitive medical data in third-party cloud storage is susceptible to loss, leakage, or theft. The privacy risk of cloud environment includes the failure of mechanisms for separating storage, memory, routing, and even reputation between different tenants of the shared infrastructure. The centralized storage and shared tenancy of physical storage space means the cloud users are at higher risk of disclosure of their sensitive data to unwanted parties.

Threats to the data privacy in the cloud include spoofing identity, tampering with the data, repudiation, and information disclosure. In spoofing identity attack, the attacker pretends to be a valid user, whereas data tampering involves malicious alterations and modification of the content. Repudiation threats are concerned with the users who deny after performing an activity with the data. Information disclosure is the exposure of information to the entities having no right to access information. The same threats prevail for the health data stored and transmitted on the third-party cloud servers.

Therefore, confidentiality and integrity of the stored health data are the most important challenges elevated by the health-care and biomedicine cloud-based systems.

A secure protection scheme will be necessary to protect the sensitive information of the medical record. There is considerable work on protecting data from privacy and security attacks. NIST (33) has developed guidelines to help consumers to protect their data in the Cloud. The work reported in ref. (34) evidences that using cryptographic storage significantly enhances security of the data. The chapter discusses the main mechanisms to be adopted in order to guarantee and satisfy the previous cited issues. Specifically, the authors present and discuss the utility of cryptographic and non-cryptographic approaches. The cryptographic approaches to mitigate the privacy risks utilize certain encryption schemes and cryptographic primitives. Conversely, non-cryptographic approaches mainly use policy based authorization infrastructure that allows the data objects to have access control policies. Particularly, in the public cloud environment operated by the commercial

service providers and shared by several other customers, data privacy and security are the most attended requirements.

Abbas and Khan (35) summarized the security and privacy requirements for cloud-based applications in the following way:

- Integrity: it is needed to ensure that the health data captured by a system or provided to any entity is true representation of the intended information and has not been modified in any way.
- Confidentiality: the health data of patients is kept completely undisclosed to the unauthorized entities.
- Authenticity: the entity requesting access is authentic. In the healthcare systems, the information provided by the healthcare providers and the identities of the entities using such information must be verified.
- Accountability: an obligation to be responsible in light of the agreed upon expectations. The patients or the entities nominated by the patients should monitor the use of their health information whenever that is accessed at hospitals, pharmacies, insurance companies etc.
- Audit: it is needed to ensure that all the healthcare data is secure and all the data access activities in the e-Health cloud are being monitored.
- Non-repudiation: repudiation threats are concerned with the users who deny after performing an activity with the data. For instance, in the healthcare scenario neither the patients nor the doctors can deny after misappropriating the health data.
- Anonymity: it refers to the state where a particular subject cannot be identified. For instance, identities of the patients can be made anonymous when they store their health data on the cloud so that the cloud servers could not learn about the identity.
- Unlinkability: it refers to the use of resources or items of interest multiple times by a user without other users or subjects being able to interlink the usage of these resources. More specifically, the information obtained from different flows of the health data should not be sufficient to establish linkability by the unauthorized entities.

Finally, in the cloud, physical storages could be widely distributed across multiple jurisdictions, each of which may have different laws regarding data security, privacy, usage, and intellectual property. For example, the US Health Insurance Portability and Accountability Act (HIPAA) restricts companies from disclosing personal health data to nonaffiliated third parties. Similarly, the Canadian Personal Information Protection and Electronic Documents Act (PIPEDA) limits the powers of organizations to collect, use, or disclose personal information in the course of commercial activities. However, a provider may, without notice to a user, move

the users' information from jurisdiction to jurisdiction. Data in the cloud may have more than one legal location at the same time, with different legal consequences.

5 Conclusions

Applications and services in bioinformatics and microarray data analysis pose quite demanding requirements. The fulfillment of those requirements could results in the development of comprehensive bioinformatics data analysis pipeline easy to use, available through the Internet, that may increase the knowledge in biology and medicine. As shown and discussed in this chapter, cloud computing may play a key role in many phases of the bioinformatics and microarray data analysis pipeline. In particular, cloud computing may be the glue that put together the parallelism, service orientation, and knowledge management technologies already used in bioinformatics, with the elasticity, ubiquity, and pay-per-use characteristics of the cloud.

Naturally, the adoption of this technology with its benefits will determine a reduction of costs and the possibility of also providing new services. However, it is important to emphasize that the use of cloud in these fields is featured still by a number of open issues and problems, such as privacy and security, geographical localization of data, legal responsibilities in the case of data leaks, that are particularly important when managing patients' data stored and processed in the cloud.

References

1. Mell P, Grance T. The NIST definition of cloud computing. Recommendations of the National Institute of Standards and Technology, Special Publication, 800–145 http://csrc.nist.gov/publications/PubsSPs.html
2. Armbrust M, Fox A, Griffith R et al (2010) A view of cloud computing. Commun ACM 53 (4):50–58
3. Vaquero LM, Rodero-Merino L, Caceres J et al (2009) A break in the clouds: towards a cloud definition. Comput Comm Rev 39:50–55
4. Calabrese B, Cannataro M, Cloud Computing in Healthcare and Biomedicine, Scalable Computing: Practice and Experience 16(1):1–18. doi:10.12694/scpe.v16i1.1057
5. Cannataro M, Guzzi PH, Veltri P (2010) Protein-to-protein interactions: technologies, databases, and algorithms. ACM Comput Surv 43 (1):1–36
6. Phillips C (2009) SNP databases. In: Komar AA (ed) Single nucleotide polymorphisms, vol 578. Humana, Totowa, NJ, pp 43–71, ch. 3
7. Schadt EE, Linderman MD, Sorenson J et al (2011) Cloud and heterogeneous computing solutions exist today for the emerging big data problems in biology. Nat Rev Genet 12(3):224
8. Grossmann RL, White KP (2011) A vision for a biomedical cloud. J Intern Med 271(2): 122–130
9. Dudley JT, Pouliot Y, Chen JR et al (2010) Translational bioinformatics in the cloud: an affordable alternative. Genome Med 2:51
10. Fusaro VA, Patil P, Gafni E et al (2011) Biomedical cloud computing with Amazon web services. PLoS Comput Biol 7(8):e1002147. doi:10.1371/journal.pcbi.1002147
11. Dai L, Gao X, Guo Y et al (2012) Bioinformatics clouds for big data manipulation. Biol Direct 7:43. doi:10.1186/1745-6150-7-43
12. Zhang L, Gu S, Wang B et al (2012) Gene set analysis in the cloud. Bioinformatics 28 (2):294–295
13. Wang Z, Wang Y, Tan KL et al (2011) eCEO: an efficient Cloud Epistasis cOmputing model

in genome-wide association study. Bioinformatics 27(8):1045–1051

14. Karczewski KJ, Fernald GH, Martin AR et al (2014) STORMSeq: an open-source, user-friendly pipeline for processing personal genomics data in the cloud. PLoS One 9(1): e84860. doi:10.1371/journal.pone.0084860
15. Schatz MC (2009) CloudBurst: highly sensitive read mapping with MapReduce. Bioinformatics 25(11):1363–1369
16. Nguyen T, Shi W, Ruden D (2011) CloudAligner: a fast and full-featured MapReduce based tool for sequence mapping. BMC Res Notes 4:171. doi:10.1186/1756-0500-4-171
17. Langmead B, Schatz MC, Lin J et al (2009) Searching for SNPs with cloud computing. Genome Biol 10:R134. doi:10.1186/gb-2009-10-11-r134
18. Habegger L, Balasubramanian S, Chen DZ et al (2012) VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. Bioinformatics 28(17):2267–2269
19. Hong D (2012) FX: an RNA-Seq analysis tool on the cloud. Bioinformatics 28(5):721–723
20. Langmead B, Hansen KD, Leek JT (2010) Cloud-scale RNA-sequencing differential expression analysis with Myrna. Genome Biol 11:R83. doi:10.1186/gb-2010-11-8-r83
21. Feng X, Grossman R, Stein L (2011) PeakRanger: a cloud-enabled peak caller for ChIP-seq data. BMC Bioinformatics 12:139. doi:10.1186/1471-2105-12-139
22. Muth T, Peters J, Blackburn J et al (2013) ProteoCloud: a full-featured open source proteomics cloud computing pipeline. J Proteomics 88:104–108
23. Lee H, Yang Y, Chae H et al (2012) BioVLAB-MMIA: a cloud environment for microRNA and mRNA integrated analysis (MMIA) on Amazon EC2. IEEE Trans Nanobioscience 11 (3):266–272
24. Chae H, Rhee S, Nephew KP et al (2014) BioVLAB-MMIA-NGS: MicroRNA-mRNA integrated analysis using high throughput sequencing data. Bioinformatics 31:265–267. doi:10.1093/bioinformatics/btu614
25. Agapito G, Cannataro M, Guzzi PH et al (2013) Cloud4SNP: distributed analysis of SNP microarray data on the cloud. In: Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics (BCB'13)
26. Afgan E, Baker D, Coraor N et al (2011) Harnessing cloud computing with Galaxy Cloud. Nat Biotechnol 29(11):972–974
27. Afgan E, Chapman B, Taylor J (2012) CloudMan as a platform for tool, data and analysis distribution. BMC Bioinformatics 13:315. doi:10.1186/1471-2105-13-315
28. Jourdren L, Bernard M, Dillies MA et al (2012) Eoulsan: a cloud computing-based framework facilitating high throughput sequencing analyses. Bioinformatics 11(28):1542–1543
29. Heath P, Greenway M, Powell R et al (2014) Bionimbus: a cloud for managing, analyzing and sharing large genomics datasets. Int J Med Inform 21(6):969–975. doi:10.1136/amiajnl-2013-002155
30. Angiuoli SV, Matalka M, Gussman A et al (2011) CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing. BMC Bioinformatics 12:356. doi:10.1186/1471-2105-12-356
31. Krampis K, Booth T, Chapman B et al (2012) Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community. Bioinformatics 13:42. doi:10.1186/1471-2105-13-42
32. Johnson ME (2009) Data hemorrhages in the health-care sector, Financial Cryptography and Data Security, Lecture Notes in Computer Science Volume 5628, pp. 71–89. doi:10.1007/978-3-642-03549-4_5
33. Guidelines on security and privacy in public cloud computing. National Institute of Standards and Technology (NIST), U.S. Department of Commerce. Special Publication, 800–144. http://csrc.nist.gov/publications/nistpubs/800- 144/SP800-144.pdf
34. Kamara S, Lauter K (2010) Cryptographic Cloud Storage, Financial Cryptography and Data Security, Lecture Notes in Computer Science Volume 6054, pp. 136–149. doi:10.1007/978-3-642-14992-4_13
35. Abbas A, Khan SU (2014) A review on the state-of-the-art privacy preserving approaches in the e-health clouds. IEEE J Biomed Health Inform 18(4):1431–1441