CLOUD COMPUTING IN BIOINFORMATICS: CURRENT SOLUTIONS AND CHALLENGES

B. Calabrese, M. Cannataro

Department of Medical and Surgical Sciences, University Magna Graecia of Catanzaro, 88100 Catanzaro, ITALY

{calabreseb, cannataro}@unicz.it

MOTIVATIONS

The availability of 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. The storage, preprocessing and analysis of experimental data is becoming the main bottleneck of the analysis pipeline. Managing omics data requires both space for data storing as well as 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 [1].

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 because it offers massive scalable computing and storage, data sharing, on-demand anytime and anywhere access to resources and applications, thus it may represent the key technology for facing those issues [2].

METHODS

This work reviews main academic and industrial cloud-based bioinformatics solutions developed in the recent years; moreover, it underlines main issues and problems related to the use of such platforms for the storage and analysis of patients' data.

Specifically, the analysed solutions regard:

- Data as a Service (DaaS): it 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.
- Software as a Service (SaaS): several cloud-based tools to execute different bioinformatics tasks, e.g. mapping applications, sequences alignment, gene expression analysis have been proposed and made available.
- Platform as a Service (PaaS): unlike SaaS solutions, PaaS solutions allow users to customize the deployment of bioinformatics applications as well as to retain complete control over their instances and associated data.
- 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 deployed storage resources, operating systems and bioinformatics applications.

For each analysed solution, main technical characteristics as well as security and privacy issues arising when storing and analysing patients data, are reported.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2261v1 | CC BY 4.0 Open Access | rec: 6 Jul 2016, publ: 6 Jul 2016

RESULTS

The application of cloud computing in bioinformatics regards the efficient storage, retrieval and integration of experimental data and their efficient and high-throughput preprocessing and analysis.
Main results of this review are: 1) a thorough analysis of cloud-based bioinformatics tools and platforms, including DaaS, SaaS, PaaS and IaaS solutions; 2) a preliminary description of the challenges and possible solutions arising when using those tools in healthcare and biomedicine.

Among the others, the article reviews the main cloud-based SaaS solutions for genome resequencing (i.e. eCEO, StormSeq, Crossbow), short-read aligner (i.e. CloudAligner, CloudBurst), variant annotation (i.e. VAT) and RNA-seq (i.e. FX, Myrna). They use the open-source Hadoop implementation of MapReduce to parallelize execution using multiple compute nodes and support access through a user-friendly web interface. CloudMan, Galaxy Cloud and Eoulsan are the main examples of PaaS solutions for bioinformatics applications, allowing users to easily set up a cloud computing cluster and automate the analysis of several samples at once using various available software solutions. IaaS solutions, used by a variety of projects to process genomics and phenotypic data, include Bionimbus, CloVR and CloudBioLinux.

Finally, main challenges analysed in the article regard security and privacy (e.g. integrity, confidentiality, authenticity, accountability, audit, non-repudiation, anonymity, unlinkability), economics opportunities, and legal aspects, e.g. the US Health Insurance Portability and Accountability Act (HIPAA) limits companies from disclosing personal health data to third parties, while the Canadian Personal Information Protection and Electronic Documents Act (PIPEDA) prohibits organizations to collect, use, or disclose personal information in commercial activities.

Bibliography

[1] B Calabrese, M Cannataro. Bioinformatics and Microarray Data Analysis on the Cloud. Microarray Data Analysis Volume 1375 of the series Methods in Molecular Biology, 2015, pp 25-39
[2] B Calabrese, M Cannataro. Cloud computing in healthcare and biomedicine. Scalable Computing: Practice and Experience, 2015, 16 (1), 1-18. DOI: 10.12694/scpe.v16i1.1057

Acknowledgements

This work has been partially funded by the PON ICaRe (PON3PE_00009_1) and BA2KNOW (PON03PE_00001_1) projects, funded by the Italian Ministry of Research and Education (MIUR).