Chapter 59
Bioinformatics Clouds for High-Throughput Technologies

Claudia Cava
National Research Council, Italy

Christian Salvatore
National Research Council, Italy

Francesca Gallivanone
National Research Council, Italy

Pasquale Anthony Della Rosa
National Research Council, Italy

Isabella Castiglioni
National Research Council, Italy

ABSTRACT

Bioinformatics traditionally deals with computational approaches to the analysis of big data from high-throughput technologies as genomics, proteomics, and sequencing. Bioinformatics analysis allows extraction of new information from big data that might help to better assess the biological details at a molecular and cellular level. The wide-scale and high-dimensionality of Bioinformatics data has led to an increasing need of high performance computing and repository. In this chapter, the authors demonstrate the advantages of cloud computing in Bioinformatics research for high-throughput technologies.

INTRODUCTION

High-throughput technologies produces an enormous amount of data that comes from the use of gene expression microarrays (Schena et al., 1995; Lipshutz et al., 1995), proteomics (Mann et al., 1999), and DNA sequencing (Lander et al., 2001; Venter et al., 2001).

Laboratories submit and archive their data to big archival databases such as GenBank at the National Center for Biotechnology Information (NCBI) (Benson et al., 2005), the European Bioinformatics Institute EMBL database (Brooksbank et al., 2010), the DNA Data Bank of Japan (DDBJ) (Sugawara et al., 2010), the Short Read Archive (SRA) (Shumway et al., 2010), the Gene Expression Omnibus (GEO) (Barrett et al., 2009) and the microarray database ArrayExpress (Kapushesky et al., 2010). These databases maintain, organize and distribute big data to the scientific community for Bioinformatics analysis. For instance, the public data repository GEO contains hundreds of thousands of microarray samples and supports many billions of analysis. So, in the traditional current

DOI: 10.4018/978-1-4666-6539-2.ch059

Praxis, Bioinformatics researchers download data from these databases and run analyses on in-house computer resources.

With significant advances in high-throughput technologies and consequently the exponential growth of biological data, Bioinformatics encounters difficulties in storage and analysis of these immense volumes of data. Mainly, the gap between high-throughput experimental technologies and computer capabilities in dealing with such big data is increasing.

At present, a promising solution to obtain the power and scale of computation is cloud computing, which uses the full potential of multiple computers and delivers analysis and repository as dynamically allocated virtual resources via the Internet.

The present chapter deals with cloud-based services and presents the advantages (and in some case disadvantages) for big data storage and analysis issues in Bioinformatics, such as data sharing, applications and time-critical calculations.

Data Sharing and Security: Public datasets change frequently and dynamically, causing problems in both archiving and sharing data for a long time. Data repositories often disappears from the public domain (e.g. due to cancelation policies for limited space) allowing users to perform partial analysis. Cloud Computing can be a solution for permanent resources where big data sets can be archived and easily accessed without necessarily copying it to another computer resources.

Bioinformatics Applications: Public datasets may be analyzed with standard tools for Bioinformatics, such as Significance Analysis of Microarrays (SAM) (Tusher et al., 2001), TM4 Multiple Expression Viewer (Saeed et al., 2006), GenePattern (Reich et al., 2006), and Bioconductor (Gentleman et al., 2004). In many cases it requires local installation and problem of maintenances and updates. Cloud Computing escapes it.

Time-critical calculations and scalability. Complex tasks that require data management are critical on clouds. Two framework 'MapReduce and Hadoop Distributed File System (HDFS)' (Taylor et al., 2010) are capable of performing time critical calculation using parallelized analysis.

In particular, cloud computing services in Bioinformatics belong to four major categories.

Cloud Software (Software as a Service, SaaS) covers applications like online software services. As a consequence, softwares are not tied to local computing resources, but are used remotely and are tied in large and often geographically distant clusters of computing hardware. Cloud software tools include sequence alignment and analysis, expression analysis, pathway annotation, machine learning method (Taylor et al., 2010). As representative examples of such software tools, Matsunaga et al. (2008) proposed a virtual machine (VM) integrating Hadoop, network Virtualization and one of the most useful Bioinformatics tools NCBI BLAST; Langmead et al. (2010) proposed Myrna, a cloud computing tool for processing differential gene expression in big RNA-Seq datasets; Zhang et al. (Zhang et al., 2012) developed a gene set analysis algorithm for biomarker identification in cloud software; Kim et al. (Kim et al., 2011) developed a cloud-computing software tool for single-nucleotide polymorphism (SNP) identification and visualization interface called Sequence Analyzer.

Cloud Platform (Platform as a Service, PaaS) involves frameworks to develop and to share Web applications and databases. Applications that use distributed algorithm are not common in Bioinformatics and there are currently few PaaS platforms in Bioinformatics. Among these applications, Jourdren et al. (2012) developed Eoulsan, a cloud platform dedicated to high throughput sequencing data analyses, Afgan et al. (2011) implemented Galaxy Cloud, a cloud platform for large-scale data analyses of next generation sequencing.

Cloud Infrastructure (Infrastructure as a Service, IaaS) provides access to servers and storage in terms of hardware components. Krampis et al.