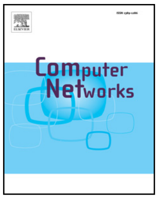
Computer Networks 145 (2018) 27–51

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet



Review article

Genomics as a service: A joint computing and networking perspective

G. Reali[a], M. Femminella[a,d,*], E. Nunzi[b], D. Valocchi[c]

[a] *Department of Engineering, University of Perugia, Perugia, Italy*
[b] *Department of Experimental Medicine, University of Perugia, Perugia, Italy*
[c] *Department of Electrical and Electronic Engineering, UCL, London, United Kingdom*
[d] *Consorzio Nazionale Interuniversitario per le Telecomunicazioni (CNIT), Italy*

ARTICLE INFO

Article history:
Received 22 December 2017
Revised 25 June 2018
Accepted 9 August 2018
Available online 11 August 2018

Keywords:
Genomic
Pipeline
Cloud computing
Big data
Network virtualization

ABSTRACT

This paper shows a global picture of the deployment of networked processing services for genomic data sets. Many current research and medical activities make an extensive use of genomic data, which are massive and rapidly increasing over time. They are typically stored in remote databases, accessible by using Internet connections. For this reason, the quality of the available network services could be a significant issue for effectively handling genomic data through networks. A first contribution of this paper consists in identifying the still unexploited features of genomic data that could allow optimizing their networked management. The second and main contribution is a methodological classification of computing and networking alternatives, which can be used to deploy what we call the Genomics-as-a-Service (GaaS) paradigm. In more detail, we analyze the main genomic processing applications, and classify both the computing alternatives to run genomics workflows, in either a local machine or a distributed cloud environment, and the main software technologies available to develop genomic processing services. Since an analysis encompassing only the computing aspects would provide only a partial view of the issues for deploying GaaS systems, we present also the main networking technologies that are available to efficiently support a GaaS solution. We first focus on existing service platforms, and analyze them in terms of service features, such as scalability, flexibility, and efficiency. Then, we present a taxonomy for both wide area and datacenter network technologies that may fit the GaaS requirements. It emerges that virtualization, both in computing and networking, is the key for a successful large-scale exploitation of genomic data, by pushing ahead the adoption of the GaaS paradigm. Finally, the paper illustrates a short and long-term vision on future research challenges in the field.

I. Introduction

THIS paper gives a comprehensive description of the ongoing initiatives aiming at increasing the usability and effectiveness of genomic computing by leveraging networking technologies. The motivations that have stimulated a fruitful trait-union between the genomics and networking essentially are represented by the need of supporting the modern medical activities making an extensive use of a massive and rapidly increasing genomic data, stored in repositories accessible through the Internet. These activities have been established over the last fifteen years, since the successful completion of the Human Genome project, in 2003, which required years of intense research. At that time, although the importance of results was clear, the possibility of handling the human genome as a commodity was far from imagination due to costs and complexity of sequencing and analyzing complex genomes. Today the situation is different. The progress of DNA (deoxyribonucleic acid) sequencing technologies has reduced the cost of sequencing a human genome, down to the order of 1000 € [1]. Since the decrease of these costs is faster that the Moore's law [3], two main consequences are expected. First, it is easy to predict that in few years a lot of applicative and societal fields, including academia, business, and public health (e.g., see [5]), will make an intensive use of the information present in DNA sequences. For this purpose, it is necessary to leverage interdisciplinary expertise from different disciplines, including biological science, medical research, and information and communication technology (ICT), which embraces data networking, software engineering, storage and database technologies, and bioinformatics. The impact of this process is significant in many application areas such as medicine, food industry, environmental monitoring, and others. The execution of genomic analyses requires significant efforts in terms of manpower and computing resources. Unfortunately, the cost of setting up large computing clusters and grids to efficiently perform

[*] Corresponding author at: Department of Engineering, University of Perugia, Perugia, Italy.
E-mail address: mauro.femminella@unipg.it (M. Femminella).

https://doi.org/10.1016/j.comnet.2018.08.005
1389-1286/© 2018 Elsevier B.V. All rights reserved.

such data analyses can be afforded by just few specialized research centers [12]. Since it cannot be assumed that any potential user owns the infrastructure for massive genome analysis, a cloud approach has been envisaged [17,4].

The second consequence is that, under a practical viewpoint, the cost to produce a unit of genomic data decreases more rapidly than the cost for storing the same unit and distributing it. Thus, given this trend, the bottleneck for handling genomics data will reside on the ICT side [4]. In other words, the most critical element of a networked genomic service is not the sequencing capability of machines, but the capacity of processing large data sets efficiently due to the limitations of accessing and exchanging data remotely. In fact, it is expected that genomics will be more demanding than astronomy, YouTube, and Twitter in terms of data acquisition, storage, distribution, and analysis [6]. The urgency of finding suitable networked solutions for managing such a huge amount of data is also witnessed by the fact that the Beijing Genomic Institute is compelled to ship hard drives for delivering genomic data [13,21].

Genomic data management can be classified as a Big Data problem [7,8], according to the classical 3 V (Volume, Velocity and Variety) model [48]. The size of a single human raw genome is roughly 3.2 GB and the global production rate is increasing over time with an exponential growth rate. Moreover, the bioinformatic processing tools, which are typically organized in software pipelines, make large use of metadata having a total volume sometimes even larger than raw data. Even these metadata, retrievable from reference databases, have to be distributed through the available networks for implementing networked genomic services [13]. The suitable handling of genomic data sets requires re-considering some aspects of data management already developed for managing other data types, such as the content growth rate, the content popularity variations over time, and the mutual relationships between genomic data. These aspects are illustrated in this paper, along with the relevant data management solutions.

These three aspects, namely the need to (i) resort to a cloud computing model for processing genomic data sets, (ii) design new networked solutions for accessing and exchanging huge amount of genomic data, and (iii) design novel data management policies to address their specific features, all together contribute to the definition of Genomics-as-a-Service (GaaS). Thus, GaaS is a novel paradigm that is rapidly gaining ground for processing genomic data sets based on the cloud computing technologies. It includes not only networking aspects, which could be either a bottleneck or a flywheel for a widespread usage of genomics in multiple fields, but also the specific features of datasets and their usage. The latter aspect could both generate significant issues and offer great exploitation potentials for network and service management.

To sum up, the main contributions of this paper are:

- Illustrating the technical problems and the still unexploited features that could allow optimizing the networked management of genomic data. In particular, the aim is to overcome or integrate the typical solutions already used to manage other types of big data, in order to improve the effectiveness of use of the GaaS instances.
- Giving an overview of widely used genomic processing applications, for medical and research activities, with a particular emphasis on open-source components and their impact on the network resource management, including a critical evaluation of the computing alternatives for GaaS implementation.
- Presenting the main ongoing activities related to the networked management of genomic data, together with a discussion on the most suitable networking alternatives for GasS deployment.
- Giving both short and long-term visions on future research challenges in the field, with a special emphasis on computing and networking issues and potential future implementation venues of GaaS in the upcoming fifth generation mobile services (5G) architectures [197].

The structure of the paper is as follows. In Section 2, we give a comprehensive view of the background, emphasizing the use of genomes and related challenges. In Section 3, we present the related works in the field and review other surveys in the genomics and Big Data applied to medicine, highlighting the original contributions of this paper. In Section 4, we present the peculiarities of genome content management and their potential impact on optimization of network and data management policies. The subsequent Section 5 focuses on genomic computing alternatives for GaaS systems. In particular, it deals with genomic applications and tools used for genomics processing. These findings are summed up in two taxonomies for genomics computing, one about computing infrastructures used for genomics computing, and the other about software technologies for implementing genomics pipelines. Finally, we also present two specific genomics processing case studies, analyzed to show main peculiarities of two real genomic pipelines, highlighting computing and networking requirements. Section 6 mainly focuses on classifying networking approach to support GaaS. In this regard, we present two taxonomies, one relevant to wide area network techniques, and another to datacenter networking. For each technique, we discuss pros and cons in the light of the application framework and ease of usage. Section 7 describes open research challenges, with emphasis on the aspects related to networking, computing, and privacy, both in the short and long term. Finally, Section 8 draws some final considerations.

2. Background

DNA and RNA (ribonucleic acid) are macromolecules that store the genetic information of any living body. They have a periodic helicoidal structure, which is analyzed by biologists for extracting information related to multiple aspects of life, including growth, reproduction, health, food production, evolution of species, more recently even for exploiting these molecules as a medium for storing information [46], and many others.

In more details, the DNA, is formed by two strands of nucleotides, or bases, commonly indicated by using the initial letter of their name: A (adenosine), C (cytosine), G (guanine) and T (thymine). Subsequent nucleotides in each strand are joint by covalent bonds while nucleotides of two separated strands are bound together with hydrogen bonds thus making the double DNA strand. The identification of significant combination of these bases, commonly referred to as *genes*, and their mutual relation (*genotypes*), is the research focus of genomic scientists, which are still struggling to associate them with any macroscopic features of bodies (*phenotypes*). The overall sequence of nucleotides encodes roughly 27,000 genes and is organized in 23 chromosomes. This research field is still in its early stage, since most of the genetic information stored within DNA is still unknown [47], even if the mere binary size of a human DNA is about 3.2 GB.

2.1. DNA sequencing

The increasing usage of genomes has been eased by the technical progresses of sequencing machines since the sequencing costs decreased more quickly than the Moore's Law during the last 15 years [3].

A comprehensive survey and comparison of modern sequencing techniques, referred to as Next Generation Sequencing (NGS) techniques, can be found in [41]. Different sequencers can offer different performance in terms of sequencing time, accuracy of results, output size, throughput, due to different sequencing mechanisms and hardware configurations. Table 1 reports a summary compar-