Journal of Biomedical Informatics xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Cloud-based bioinformatics workflow platform for large-scale next-generation sequencing analyses

Bo Liu [a,*], Ravi K Madduri [b,c], Borja Sotomayor [b], Kyle Chard [b], Lukasz Lacinski [b], Utpal J Dave [b], Jianqiang Li [d], Chunchen Liu [a], Ian T Foster [b,c]

[a] NEC Labs China, Beijing 100084, China
[b] Computation Institute, University of Chicago, Chicago, IL, USA
[c] Mathematics and Computer Science Division, Argonne National Lab, IL, USA
[d] School of Software Engineering, Beijing University of Technology, Beijing 100022, China

ARTICLE INFO

Article history:
Received 8 August 2013
Accepted 15 January 2014
Available online xxxx

Keywords:
Bioinformatics
Scientific workflow
Sequencing analyses
Cloud computing
Galaxy

ABSTRACT

Due to the upcoming data deluge of genome data, the need for storing and processing large-scale genome data, easy access to biomedical analyses tools, efficient data sharing and retrieval has presented significant challenges. The variability in data volume results in variable computing and storage requirements, therefore biomedical researchers are pursuing more reliable, dynamic and convenient methods for conducting sequencing analyses. This paper proposes a Cloud-based bioinformatics workflow platform for large-scale next-generation sequencing analyses, which enables reliable and highly scalable execution of sequencing analyses workflows in a fully automated manner. Our platform extends the existing Galaxy workflow system by adding data management capabilities for transferring large quantities of data efficiently and reliably (via Globus Transfer), domain-specific analyses tools preconfigured for immediate use by researchers (via user-specific tools integration), automatic deployment on Cloud for on-demand resource allocation and pay-as-you-go pricing (via Globus Provision), a Cloud provisioning tool for auto-scaling (via HTCondor scheduler), and the support for validating the correctness of workflows (via semantic verification tools). Two bioinformatics workflow use cases as well as performance evaluation are presented to validate the feasibility of the proposed approach.

1. Introduction

With the emergence of NGS (next-generation sequencing), various genome informatics ecosystems are now facing a potential tsunami of genome data that will swamp their storage systems and crush their computing clusters. Human DNA is comprised of approximately 3 billion base pairs with a personal genome representing approximately 100 gigabytes (GB) of data. By the end of 2011, the global annual sequencing capacity was estimated to be 13 quadrillion bases and counting [1]. The upcoming data deluge forces researchers to find reliable and convenient methods for storage and computing.

In the bioinformatics community, acquiring sequence data is always followed by large-scale computational analysis to process the data, validate experiment results and draw scientific insights. Therefore, investment in a sequencing instrument would normally be accompanied by substantial investment in computer hardware, skilled informatics support, and bioinformaticians competent in configuring and using specific software to analyze the data [2].

However, the need for storing and processing large-scale genome data, providing easy access to data analysis tools, enabling efficient data sharing and retrieval, integrating imaging, electrophysiological and clinical data, and supporting cross-institutional collaboration still has significant challenges.

Existing tools, such as Bioconductor [3], Bioperl [4], and EMBOSS [5], improve the accessibility of computation and facilitate bioinformatics research by decreasing IT efforts and automating data analyses workflows. But these approaches have difficulties when dealing with large datasets, which is generally common in NGS analyses; besides the software installation and programming efforts needed are often error-prone and time consuming for biomedical researchers. Moreover, most research institutes implement their applications on laboratory-hosted servers [6], and as data volume varies greatly, the capabilities and efficiency in storing and analyzing genome data are not enough to fulfill the dynamic requirements of different workflows.

* Corresponding author. Address: 11F, Bld. A, Innovation Plaza, Tsinghua Science Park, HaiDian District, Beijing 100084, China.
E-mail address: liu_bo@nec.cn (B. Liu).

http://dx.doi.org/10.1016/j.jbi.2014.01.005
1532-0464/

To address these problems, the authors propose a Cloud-based bioinformatics workflow platform for large-scale NGS analyses [7]. This platform integrates Galaxy, a scientific workflow system for biomedical analyses, Globus Provision (GP), a tool for deploying distributed computing clusters on Cloud, and a set of supporting tools and modules to provide an overall solution for biomedical researchers. This combination of tools implements an easy to use, high performance and scalable workflow environment that addresses the needs of data-intensive applications through dynamic cluster configuration, automatic user-defined node provisioning, high speed data transfer, and automated deployment and configuration of domain-specific software.

More specifically, the contributions of this paper are summarized as follows.

(1) We propose a novel approach for automatically deploying and configuring bioinformatics workflows in Cloud environments. The integration of scientific workflows and Cloud computing provides fast provisioning of computational and storage resources, elastic scaling and pay-as-you-go pricing. Our approach builds on GP, and supports automated deployment of all prerequisite tools and software packages required for Galaxy along with additional domain specific tools. The deployed workflow environment can respond to workload changes by adding or removing nodes from the cluster and changing instance types to balance cost and performance.
(2) The variability in data volume results in variable computing and storage requirements for data processing. HTCondor [8] is a tool for High Throughput Computing (HTC) on large collections of distributive computing resources. By integrating Galaxy with the HTCondor scheduler, specified Galaxy jobs are executed in parallel using distributed computing nodes in a dynamic HTCondor pool. The proposed auto-scaling strategy significantly improves resource utilization and processing speed, especially for compute-intensive tools, like alignment and SNP calling.
(3) When dealing with large-scale datasets that are common in NGS, Galaxy's file upload and download capabilities via HTTP and FTP are often unreliable and inefficient. To meet the need for large-scale data transfer, we have integrated Galaxy with Globus Transfer, a service that provides high performance, secure and reliable data transfer, to enable efficient upload and download of large quantities of data in and out of Galaxy. Globus Transfer provides not only powerful Grid transfer capabilities to automate the task of moving files across administrative domains [9,10], but also superior and easy-to-use data management capabilities for transferring big datasets from geographically distributed sequencing centers into Cloud computing infrastructure.
(4) To demonstrate the flexibility of our approach we have extended this framework to meet the requirements of a specific domain, by adding a set of domain-specific tools to the deployment. This paper introduces two different tools that we have wrapped and integrated into the Galaxy platform: CRData tools for executing R scripts, and CummeRbund [11] tool for analyzing Cufflinks RNA-Seq output. These new tools complement the functionality of Galaxy, and have been integrated into our forked Galaxy repository so it's convenient to deploy a user-specific Galaxy with additional tools.
(5) Galaxy's workflow canvas provides a platform for assembling tools and building workflows; however building a workflow, especially a complex computational workflow, still requires a lot of domain-specific knowledge and understanding of Galaxy tools. This process is both error-prone and time consuming. Moreover it is often impossible to identify possible errors until the workflow is running. Consequently, we propose semantic verification approaches to facilitate the generation of workflows. By using semantic representations to describe the parameters, tools and workflows, and maintaining an ontology to identify the semantic annotations and appropriate constraints among them, the parameter consistency, functional consistency and reachability of workflows are validated.
(6) The Cloud-based bioinformatics workflow platform integrates all the aforementioned tools, and provides an overall solution for deploying and configuring Galaxy system on Clouds, auto-scaling Cloud resources, enabling high-performance data transfer capabilities, providing customization of user-specific tools, and leveraging a semantic verification mechanism. The platform reduces the considerable usage barriers that existed previously, leverages Amazon EC2 with its pay-as-you-go billing model for resource usage, and provides a scalable and elastic execution environment for sequencing analyses. To validate the effectiveness of our proposed approaches, two bioinformatics workflow use cases as well as performance evaluation are presented, including CRData workflow and RNA-Seq analysis workflow.

The rest of the paper is organized as follows: Section 2 describes an RNA-Seq workflow as a motivating scenario. Section 3 briefly introduces Galaxy. Section 4 describes the tools we have integrated into Galaxy, including Globus Transfer, CRData, CummeRbund, and semantic verification tools. In Section 5, a Globus Provision-based method is proposed to automatically deploy Galaxy on Amazon Cloud. Then the system implementation, use cases and performance evaluation are depicted in Section 6. Section 7 reviews the related work of scientific workflow and Cloud computing. Finally the conclusions and future work are given in Section 8. This paper is an extension of our previous work that describes the methods used to deploy bioinformatics workflows on the Cloud [7].

2. Motivating scenario

In this section, we first introduce an RNA-Sequencing analysis workflow as a motivating scenario. RNA-Sequencing (RNA-Seq) [12] is a deep-sequencing technique used to explore and profile the entire transcriptome of any organism. Fig. 1 shows a sketch map of an RNA-Sequencing analysis workflow downloaded from the public Galaxy website (https://usegalaxy.org/workflow/list_published) for understanding the functional elements of the genome.

This workflow mainly contains 6 kinds of tools: FASTQ Groomer, TopHat for Illumina, Map with Bowtie for Illumina, Map with BWA for Illumina, Cufflinks and Flagstat. *FASTQ Groomer* offers several conversion options relating to the FASTQ format if a quality score falls outside of the target score range. *TopHat for Illumina* is a fast splice junction mapper for RNA-Seq reads, which aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons. *Map*

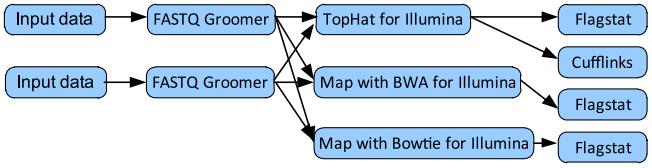

Fig. 1. RNA-Sequencing analysis workflow.