Advanced Science and Technology Letters
Vol.33 (Bioscience and Medical Research 2013), pp.45-48
http://dx.doi.org/10.14257/astl.2013.33.11

A Scalable Computing Framework for Large-Scale Bioinformatics Analysis

Youngmahn Han*[1], Hyunsik Kim[1]

[1] Korea Institute of Science and Technology Information, Korea
{hans,wbstory}@kisti.re.kr

Abstract. Next generation sequencing (NGS) technologies, which play a crucial role in realizing the personal medicine, have generated unprecedented volumes of DNA sequence data. Big data problem of NGS data analysis leads to requirements of cost-effective computations using limited resources. The cloud computing environment can be considered as an alternative computing architecture that enables fast, large-scale, and cost-effective bioinformatics analysis. "Workflow" is a well-defined computing framework for bioinformatics analysis processes which are performed as a chain of interlinked data process tasks. Here, we describe a workflow-based computing approach for large-scale bioinformatics analysis using cloud computing resources.

Keywords: Large-scale bioinformatics analysis, Workflow, Cloud Computing, Next Generation Sequencing

1 Introduction

The identification of causal genomic variants that alter human phenotypes, particularly those associated with disease, is a central goal of research on human genetics. During the past decade, genome-wide association studies (GWASs) have identified several hundred common variants associated with complex human diseases and traits [1]. Despite these successes, most of the common variants contribute individually only to a small portion of their estimated heritability. Many explanations for this missing heritability have been suggested, including a large number of variants with unclear effects that are poorly detected by available genotyping arrays. Other possible factors include rare variants as well as structural, regulatory, and epigenetic variants [2]. Recent advances in next-generation sequencing (NGS) technology have created the possibility of identifying larger numbers of individual variants in protein-coding and non-coding regions of individual genomes without the limits of rate and cost; the dramatic reduction of DNA sequencing cost has caused unprecedented increase of DNA sequence data and the lack of data storage capacity [3]. Stein postulated that cloud computing and associated computing-as-a-service technologies would help researchers deal with the huge volume of NGS data [4].

Since many bioinformatics analysis processes are performed as a chain of interlinked data process tasks, pipeline or "workflow" can be a well-defined model for bioinformatics analysis processes [5]. A typical example for bioinformatics workflow

ISSN: 2287-1233 ASTL

given in [6] describes chained tasks for identification of the protein family from given DNA sequence input as shown in Figure 1. We have developed a workflow-based system for automation of complex bioinformatics analysis, namely Bioworks which is descried in our previous work [7].

Here, we describe a scalable computing framework for complex and large-scale bioinformatics analysis using cloud computing technology.

2 Workflow-based automation for complex bioinformatics analysis

Workflows are considered as glue to orchestrate heterogeneous biological information and data analysis tools for complex bioinformatics analysis. Several workflow-based systems including Galaxy [8], Taverna[9], Kepler[10], Triana[11] and BioWMS[12] have been published. There are several requirements for bioinformatics-capable workflow system as follows:

- Since there are many and heterogeneous tools and databases, Standardization and extensible integration of distributed tools is necessary for providing seamless access to them.
- Bioinformatics tools are highly heterogeneous in their input/output data types. These heterogeneity leads to be difficult to make links among tool tasks according to data flow. A workflow system should provide flexible integration methods to resolve the heterogeneity in data types.
- Reproducibility of scientific analyses and processes is at the core of the scientific method, in that it enables researchers to evaluate the validity of each other's hypothesis and to repeat techniques and analysis methods to obtain scientifically similar results. In order to support reproducibility, WMS should capture and generate provenance information as a critical part of the workflow-generated data. Provenance information can be referred as a historical metadata that provides explanations on how a particular intermediate result data has been generated from the given input data.

Effective implementation methods to fulfil these requirements have been discussed in previous publication of Bioworks system [7]. Figure 1 shows the user interface of Bioworks client program for typical workflow automation of differential expressed gene (DEG) analysis from RNA-seq data.

3 Scalable computing framework for large-scale bioinformatics analysis

To achieve the goal of large-scale, cost-effective bioinformatics analysis, the computing storage capacity needs to be scalable and extensible on demand of analysis

scale. Cloud computing architecture can be as an alternative solution due to its advantages including scalability, extensibility and isolated provision capability.

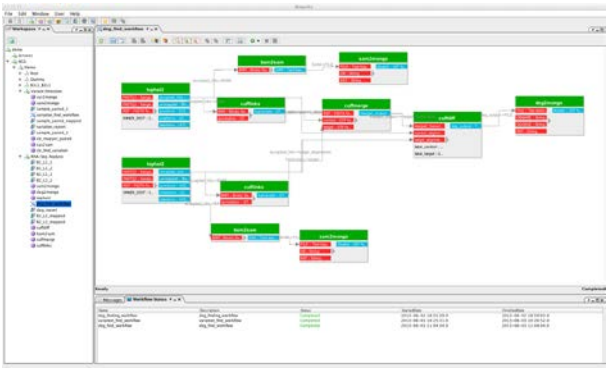

Fig. 1. Typical example of workflow automation for DEG analysis using Bioworks.

Figure 2 shows the entire computing framework for large-scale bioinformatics analysis using cloud computing resources. The system administrator creates a specific virtual machine (VM) template and store within the cloud infrastructure. This template, contains an OS, tools, and several metadata, such as a workflow, data exchange rules, and resource assignment information. Next, an analytics requires a VM template via the VM Package Handler. This loads a proper VM template image from the infrastructure and installs it within actual virtual machines. The template can install more than one VM. This is controlled automatically via the Cloud Resource Provider. An analytics pushes data and runs VM. The VM runs through the execution pipeline. The Pipeline Execution Handler assists with organization. Distributed computing and scalable storage is provided via the Cloud Service Infrastructure. This provides a VM container, common data storage, and load balancing. Distributed input and output (I/O) is applied through the VMs, which facilitates data exchange between VMs.

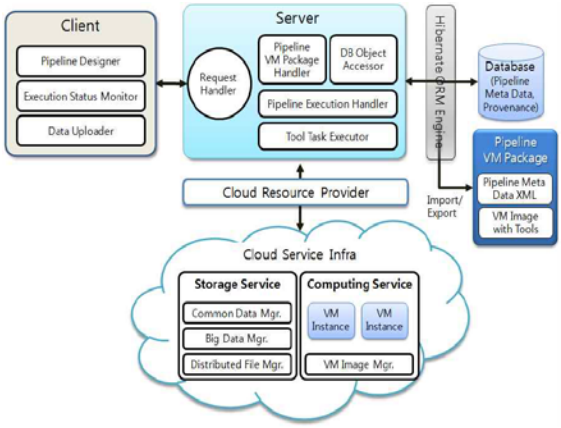

Fig. 2. Cloud-based computing framework for large-scale bioinformatics analysis workflows

4 Conclusion

As DNA sequencing becomes cheaper, more and more data are being generated. This rapid increase in data is outstripping data storage capacity and computing resource development. Moreover, many bioinformatics analysis processes are too complex to be handled by researchers. A workflow-based computing approach employing cloud computing resources can be an alternative solution for large-scale bioinformatics analysis by using a pre-configured pipeline-installed VM template, automatic resource distribution, and isolated provisioning capability. In this paper, we have described how to build an integrated bioinformatics analysis environment for life science researchers by using workflow-based automation and cloud computing technology.

References

1.Hardy J, Singleton A. "Genome-wide association studies and human disease". N. Engl. J. Med, 360, pp. 1759-1768(2009)
2.Teri A. Manolio, "Finding the missing heritability of complex diseases", Nature. vol. 461(7265), pp. 747–753(2009)
3.Elaine R. Mardis, "The impact of next-generation sequencing technology on genetics", Trends in Genetics, vol. 24(3), pp. 133-141(2008)
4.Lincoln D Stein, "The case for cloud computing in genome informatics", Genome Biology, vol. 11, Issue 5(2010)
5.Yolanda Gil, et al., "Examining the Challenges of Scientific Workflows", IEEE Computer, vol. 40, pp. 24-32(2007)
6.Malika Mahoui, et al., "A Dynamic Workflow Approach for the Integration of Bioinformatics Services", Cluster Computing, vol. 8, pp. 279–291(2005)
7.Y. Han, "Bioworks: a workflow system for automation of bioinformatics analysis processes", International Journal of Bio-Science and Bio-Technology, vol. 3, no. 4, pp. 59–68(2011)
8.Jeremy Goecks, et al., "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences", Genome Biology, 11:R86(2010)
9.Oinn T, Addis M, et al., "Taverna: a tool for the composition and enactment of bioinformatics workflows", Bioinformatics, vol. 20, pp. 3045–54(2004)
10.Bertram Ludäscher, et al., "Scientific Workflow Management and the Kepler System", Concurr. Comput.: Pract. Exp., vol. 18, pp. 1039 – 1065(2005)
11.Matthew Shields, et al., "Programming Scientific and Distributed Workflow with Triana Services, Concurrency and Computation, vol. 18, pp. 1021–37(2006)
12.Bartocci E, et al., "BioWMS: a web based workflow management system for bioinformatics", BMC Bioinformatics, vol. 8, S2(2007)

