Future Generation Computer Systems 29 (2013) 2205–2219

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

Designing a parallel cloud based comparative genomics workflow to improve phylogenetic analyses

Kary A.C.S. Ocaña [a], Daniel de Oliveira [b], Jonas Dias [a], Eduardo Ogasawara [a,c], Marta Mattoso [a,*]

[a] PESC/COPPE - UFRJ - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

[b] UFF - Fluminense Federal University, Niterói, Brazil

[c] CEFET/RJ - Federal Center of Technological Education, Rio de Janeiro, Brazil

HIGHLIGHTS

- Scientific workflows are an attractive option in comparative genomics and phylogeny.
- SciHmm compute-intensive genomic workflow executed in 128 cores Amazon EC2 clouds.
- Muscle MSA method provided the best data quality, although other input data may point to different MSA methods.
- Muscle speedup factor was 28 on 32 cores, compared to a single core computation.
- Executing SciHmm before the phylogenetic analyses improved the performance up to 80%.

ARTICLE INFO

Article history:
Received 31 March 2012
Received in revised form
30 January 2013
Accepted 6 April 2013
Available online 17 April 2013

Keywords:
Profile hidden Markov models
Scientific workflows
Cloud computing

ABSTRACT

Over the last years, comparative genomics analyses have become more compute-intensive due to the explosive number of available genome sequences. Comparative genomics analysis is an important *a priori*step for experiments in various bioinformatics domains. This analysis can be used to enhance the performance and quality of experiments in areas such as evolution and phylogeny. A common phylogenetic analysis makes extensive use of Multiple Sequence Alignment (MSA) in the construction of phylogenetic trees, which are used to infer evolutionary relationships between homologous genes. Each phylogenetic analysis aims at exploring several different MSA methods to verify which execution produces trees with the best quality. This phylogenetic exploration may run during weeks, even when executed in High Performance Computing (HPC) environments. Although there are many approaches that model and parallelize phylogenetic analysis as scientific workflows, exploring all MSA methods becomes a complex and expensive task to be performed. If scientists determine *a priori*the most adequate MSA method to use in the phylogenetic analysis, it would save time, and, in some cases, financial resources. Comparative genomics analyses play an important role in optimizing phylogenetic analysis workflows. In this paper, we extend the SciHmm scientific workflow, aimed at determining the most suitable MSA method, to use it in a phylogenetic analysis. SciHmm uses SciCumulus, a cloud workflow execution engine, for parallel execution. Experimental results show that using SciHmm considerably reduces the total execution time of the phylogenetic analysis (up to 80%). Experiments also show that trees built with the MSA program elected by using SciHmm presented more quality than the remaining, as expected. In addition, the parallel execution of SciHmm shows that this kind of bioinformatics workflow has an excellent cost/benefit when executed in cloud environments.

1. Introduction

Comparative genomics is one of many bioinformatics fields that aim at computationally comparing hundreds of different genomes [1]. Many types of bioinformatics applications associated to this field, such as multiple sequence alignment (MSA), homologous detection and phylogenetic analysis are continuously increasing in scale and complexity [2]. Managing comparative genomics experiments is far from trivial, as they are computing-intensive and process large volumes of data. Accordingly, considering that these experiments are based on a pipeline of scientific programs, they are assisted by scientific workflows [3,4]. Comparative genomics workflows are a fundamental step to enhance the execution of other evolutionary analyses such as phylogenomics and phylogenetics [5]. To illustrate the benefits

* Correspondence to: COPPE, Federal University of Rio de Janeiro, P.O. Box 68511, 21941-972 Rio de Janeiro, RJ, Brazil. Tel.: +55 21 2562 8694; fax: +55 21 2562 8080.
E-mail addresses: kary@cos.ufrj.br (K.A.C.S. Ocaña), danielc@cos.ufrj.br (D. de Oliveira), jonasdias@cos.ufrj.br (J. Dias), ogasawara@cos.ufrj.br (E. Ogasawara), marta@cos.ufrj.br (M. Mattoso).

0167-739X/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.future.2013.04.005

of executing comparative genomics workflows to improve other bioinformatics analyses, let us take phylogenetic analysis as an example. In our previous work, in phylogenetic analysis workflows using SciPhy [3], scientists perform a set of activities to produce phylogenetic trees, which are used to infer evolutionary relationships between homologous genes represented in the genomes of divergent species. Phylogenetic scientific workflows are both computing and data-intensive. Depending on the amount of input data and complexity of the MSA method used, which corresponds to an activity in these workflows, each workflow execution may require weeks to complete.

Furthermore, according to the results presented by Bernardes et al. [6], different MSA methods produce results with different quality levels to detect distant homologues. This result can be extrapolated to phylogenetic analyses where different MSA methods can have a certain impact (positive or negative) in the final phylogenetic tree. Based on that, scientists need to try different variations of a phylogenetic workflow to analyze the quality of its produced results. These variations are basically related to the exploration of different MSA methods, algorithms, and parameters. Depending on the amount of input data to process, the execution of each variation of a phylogenetic workflow using a specific MSA method may demand many days to complete. The exploration of several MSA methods using the same input data is likely to be time-consuming, difficult to manage, and scientists without a high performance infrastructure have to reduce the scale of the analysis. For example, if scientists want to explore all the available MSA methods in SciPhy using 200 input multi-fasta with genomics data, they will need approximately 52 days when running it on a desktop machine. Even when SciPhy is executed in parallel in High Performance Computing (HPC) environments, the exploration may still need a couple of days to finish when using 32 cores in a multi-core machine or in a cloud, as detailed by Ocaña et al. [3].

Scientists usually avoid exploring different MSA methods in the phylogenetic analysis due to its high cost (execution time and financial, in some cases). In this paper, we extend the recently proposed SciHmm scientific workflow [7] to optimize the task of phylogenetic analysis implemented in SciPhy workflow.

SciHmm is a comparative genomics workflow, based on profile hidden Markov models (pHMMs) [8] to detect homologues. SciHmm aids scientists in choosing the most suitable MSA method, which produces alignments with better quality, to be used in the phylogenetic analysis in order to generate phylogenetic trees, also with better quality, without having to construct all trees, exploring all possible MSA variations. SciHmm consists of three main activities: (i) MSA construction; (ii) pHMMs building and scoring; and (iii) pHMMs comparison against a local database. The main idea behind SciHmm is to perform a cross-validation procedure to evaluate: (i) the Specificity and Sensitivity of each MSA method to detect homologues and (ii) the identification of the best MSA method and e-value for an *a posteriori* phylogenetic analysis workflow. The e-value is defined as the expected number of errors *per* query. In the context of one test, this means that one would expect on average E false positives *per* model with an e-value score better than E to occur by chance [9].

Many HPC environments can be used to execute SciHmm in parallel such as supercomputers or grids [10]. However, computing clouds [11,12] provide a new dimension for HPC workflows without having to acquire or to configure many pieces of software and hardware. Clouds have demonstrated feasibility for many bioinformatics problems [13,14], as they provide characteristics, such as, elasticity and high availability. SciHmm was developed on top of SciCumulus [15], a cloud workflow execution engine, as a solution for running SciHmm in parallel in clouds. SciCumulus executes SciHmm using parameter sweep [15] mechanisms, where each Virtual Machine (VM) – that is part of a virtual cluster – processes independent activities consuming different input data. By using SciCumulus [3,7,14,16], scientists can benefit from parallel execution coupled to distributed data provenance management. In this context, the term "Provenance" [17] represents the ancestry of an object within the workflow. It contains information about the process used to derive the object, in this case the data product. Provenance data is used to determine quality and authorship, and to reproduce the workflow as well as to interpret and validate the associated results.

We show in this paper, a thorough evaluation of SciHmm (both computational and biological) to compare genomics results with phylogenetic trees (constructed based on several MSA methods). By analyzing these phylogenetic trees, scientists are able to verify if the MSA method elected by SciHmm is the one that produces the best quality results on the *a posteriori* phylogenetic analysis workflow. This way, we use phylogenetic experiments [3] to validate the SciHmm results. In the experiments presented in this paper, SciHmm uses five MSA methods to perform the analysis: ClustalW [18], Kalign [19], MAFFT [20], Muscle [21], and ProbCons [22]. SciHmm was executed in the Amazon EC2 [23] cloud. SciHmm uses a cross-validation procedure to evaluate and identify the best MSA method for an *a posteriori* phylogenetic analysis. Experimental results reinforce the importance of optimizing a phylogenetic analysis using SciHmm to reduce costs (execution time and financial costs). The optimization of the phylogenetic analysis with SciHmm resulted in trees with more quality and more performance (up to 80%), when compared to an *ad hoc* MSA method exploration including all variations.

The remainder of this paper is organized as follows. Section 2 brings important biological and bioinformatics background. Section 3 gives a brief explanation of the SciCumulus architecture and implementation details on top of the Amazon EC2. Section 4 describes the conceptual specification of the SciHmm workflow. Section 5 describes experimental results. Section 6 brings related work and finally, Section 7 concludes this article.

2. Comparative genomics

Comparative genomics aims at inferring relationships of genome structure and function across several species [24]. It is a critical and enabling field for functional genomics, allowing researchers to access useful information about genes, their resulting proteins and the role played by these proteins in the organism's biochemical process. One of the most important goals of comparative genomics field is to identify existing mechanisms for eukaryotic genome evolution [25]. To process the large amount of information contained in modern genomes (the human genome alone having 3.2 GB) scientists need fine-tuned computational methods for comparative genomics. MSA methods, gene finding and profile hidden Markov models (pHMM) are important applications used in comparative genomics experiments. Although comparative genomics has attracted much attention from the scientific community in the last decade [26–28], it is still a new research field.

MSA is an important step in many bioinformatics experiments, such as in pHMMs generation [8]. Hidden Markov models (HMMs) [29] are probabilistic models used in pattern recognition problems, and are used in computational biology to analyze sequential data [2] e.g. in remote homology detection between protein sequences. HMMs can be used first for training HMM that represent a group of homologue sequences and for comparing these HMM against a target database of protein sequences, i.e., RefSeq [30]. The HMM that represent a group of homologue sequences are called pHMM, and are a probabilistic model built from a MSA of related sequences. One major program that applies