SwiftGECKO: a provenance-enabled parallel comparative genomics workflow

Maria Luiza Mondelli, Oscar Torreño, Kary A. C. S. Ocaña, Marta Mattoso, Michael Wilde, Ana Tereza Vasconcellos, Oswaldo Trelles, Luiz M. R. Gadelha

National Laboratory for Scientific Computing, Federal University of Rio de Janeiro

Abstract

Biology has moved from gene-gene to overall genomes analysis, both somehow pushed up for improvements in DNA sequencing technology and the availability of high performance computing (HPC) resources. Conducting computer-based comparative genomic experiments is a complex and time-consuming task since a huge quantity of data needs to be processed and a large set of programs are used as the income of another. This coherent flow can be designed as scientific workflows. Scientists may require technologies as parallel scientific workflow management (SWfMS) and HPC environments for improving the total processing time and assisting scientists at the management, treatment and analyses of the data. We propose the workflow SwiftGECKO, an updated version of the sequential application GECKO for genome comparisons based on the fast identification of high-scoring segment pairs (HSPs). SwiftGECKO was implemented in the Swift parallel functional dataflow system, providing benefits such as the intrinsic parallelism for execution and provenance data management. We tested SwiftGECKO using a dataset of 40 evolutionary related bacterial genomes. SwiftGECKO keeps a detailed domain data provenance trace of the experiment in a relational database. General benefits regard the capacity to retrieve domain data associated with computational ones. In the specific field of comparative genomics, we explain these benefits with a set of queries aimed to (i) exploring the biological information contained in resulting files, no prone to error manual manipulation is required from scientists; (ii) tracing the taxonomic lineage and inferring evolutionary relationships of genomes based on annotations of domain data provenance of the experiment (e.g. statistics of hits and fragments); and (iii) driving new task scheduling strategies for the execution of experiments, e.g. estimating a priori the demanded CPU/time for processing other genomes. Regarding computational results, we present performance improvements of up to 89.40% (128 cores) in the execution time when compared to its sequential execution (1 core), which drops from around 2 hours and 50 minutes to 18 minutes using a shared-memory computer with 128 processing cores. The process of multiple comparisons of genomes is considered as a time-consuming and costly experiment. In this article, we have addressed the problem of tracing and managing the provenance data and results information (both domain specific and general computational annotations) of the experiment. Additionally, we propose a parallel solution which significantly reduces the execution time of the sequential executions. The presented results demonstrate that SwiftGECKO is computationally-efficient for parallelizing massive tasks for complete genome comparisons.