2018 CHPC National Conference

Contribution ID: **87** Type: **Talk**

An Evaluation of Galaxy and Ruffus Workflows System for DNA-seq Analysis

Wednesday, 5 December 2018 14:30 (20 minutes)

Functional genomics determines the biological functions of genes on a global scale by using large volumes of data obtained through techniques including next-generation sequencing (NGS). The application of NGS in biomedical research is gaining in momentum, and with its adoption becoming more widespread, there is an increasing need for access to customizable computational workflows that can simplify, and offer access to, computer-intensive analyses of genomic data. In this study, the Galaxy and Ruffus frameworks were designed and implemented with a view to addressing the challenges faced in biomedical research. Galaxy, a graphical web-based framework, allows researchers to build a graphical NGS data analysis pipeline for accessible, reproducible, and collaborative data-sharing. Ruffus, a UNIX command-line framework used by bioinformaticians as Python library to write scripts in an object-oriented style, allows for building a workflow in terms of task dependencies and execution logic. In this study, a dual data analysis technique was explored which focuses on a comparative evaluation of Galaxy and Ruffus frameworks that are used in composing analysis pipelines. To this end, we developed an analysis pipeline in Galaxy, and Ruffus, for the analysis of Mycobacterium tuberculosis sequence data. Furthermore, this study aimed to compare the Galaxy framework to Ruffus with preliminary analysis revealing that the analysis pipeline in Galaxy displayed a higher percentage of load and store instructions. In comparison, pipelines in Ruffus tended to be CPU bound and memory intensive. The CPU usage, memory utilization, and runtime execution are graphically represented in this study. Our evaluation suggests that workflow frameworks have distinctly different features from an ease of use, flexibility, and portability, to architectural designs. Therefore, in this CHPC Conference, I will discuss how we composed the NGS bioinformatics data analysis pipeline in the Galaxy and Ruffus workflow framework and the use of each framework in the analysis of Mycobacterium tuberculosis sequence data.

Presenter Biography

Ajayi Olabode is a bioinformatics masters student at the University of the Western Cape. Prior to joining the South Africa National Bioinformatics Institute (SANBI) for master degree, Ajayi Olabode holds both his BSc and Honour degree in computer science at the same institution. He is currently working as a bioinformatics and Database analyst at the Centre for Proteomic and Genomic Research (CPGR).

Primary author: Mr AJAYI, Olabode (South Africa National Bioinformatics Institute)

Co-authors: Mr VAN HEUSDEN, Peter (South Africa National Bioinformatics Institute); Prof. ALAN , Christoffels (South Africa National Bioinformatics Institute)

Presenter: Mr AJAYI, Olabode (South Africa National Bioinformatics Institute)

Session Classification: HPC Applications

Track Classification: Bioinformatics and Biological Sciences