Pipengine: an ultra light YAML-based pipeline execution engine

Francesco Strozzi[1] and Raoul Jean Pierre Bonnal[2]

[1]Enterome Bioscience, Paris - France
[2]INGM - Isitituto Nazionale Genetica Molecolare "Romeo ed Enrica Invernizzi": Milan, Italy

25 July 2017

Paper DOI: http://dx.doi.org/10.21105/joss.00341
Software Repository: https://github.com/fstrozzi/bioruby-pipengine
Software Archive: http://dx.doi.org/10.5281/zenodo.851892

Summary

This is an ultra light YAML-based pipeline execution engine. The tool allows defining a pipeline template in YAML, specifying command lines, resources and software to be used along with pipeline steps dependencies. Pipengine is a sample-centric tool, so the pipeline can then be applied over a single sample or multiple samples data, generating actual runnable bash scripts which can then be submitted automatically to a scheduling system or run locally.

The bash scripts generated by Pipengine includes a list of features such as:

- error controls and logging for each step
- the automated generation of directories based on sample and pipeline steps names
- the moving of input and output data across original and temporary folders if needed
- a simple checkpoint strategy to avoid re-running already completed steps in a pipeline.

All these features prevent the users to write boiler plate code to perform all these necessary accessory tasks.

Moreover, Pipengine creates a stable and reproducible working and output tree for each analysis, which transparently stores all the results of each step of a pipeline for each sample analyzed. In this way pipelines' intermediate or final results can be predictably accessed by the analysts and/or easily parsed with other tools.

The software was developed back in 2012, when more generalized schemas such as for instance the Common Workflow Language (Language 2017) were not yet defined, and thus was among the firsts utilities to introduce the concept of using simple YAML as a template format to define reusable bioinformatics pipelines.

Pipengine has been used across several research groups and bioinformatics core facilities since its first appearance. It directly supports the PBS/Torque scheduler (Inc. 2017) for submission of jobs, but given that the support for a scheduler is based on specific options written automatically inside the bash scripts generated by the tool, it can be easily adapted to work with other schedulers, if needed.

Pipengine is written in Ruby and is available for download as a BioRuby Gem (Goto N 2010; Bonnal RJP 2012).

References

Bonnal RJP, Githinji G, Aerts J. 2012. "Biogem: An Effective Tool-Based Approach for Scaling up Open Source Software Development in Bioinformatics." *Bioinformatics*. doi:doi.org/10.1093/bioinformatics/bts080.

Goto N, Nakao M, Prins P. 2010. "Bioinformatics Software for the Ruby Programming Language." *Bioinformatics*. doi:doi.org/10.1093/bioinformatics/btq475.

Inc., Adaptive Computing. 2017. "TORQUE Resource Manager." http://www.adaptivecomputing.com/products/open-source/torque/.

Language, Common Workflow. 2017. "Common Workflow Language." doi:dx.doi.org/10.6084/m9.figshare.3115156.v2.