Copperfish: A version control system for bioinformatic pipeline development and validation

Art F. Y. Poon

April 4, 2013

1 Project summary

Due to the cutting-edge nature of bioinformatics, genetic sequence data are usually processed using custom bioinformatic scripts that are constantly being modified and refined. As a result, different versions of methods can rapidly accumulate with sporadic documentation. More importantly, there is no convenient revision control system for linking the results of data processing with the methods (or more importantly, revisions of methods) that were used to produce them. Without such a system in place, tracking which version of a method was used to generate a result must be carried out on an *ad hoc* basis. This deficiency makes it extremely difficult to make bioinformatic analyses reproducible between lab members or to disseminate methods to the research community. Moreover, the increasing uptake of high-throughput technologies such as next-generation sequencing (NGS) in clinical laboratories is driving a growing demand for quality assurance; for example, version tracking, validation, and documentation of NGS bioinformatic pipelines are now all specifically required by the College of American Pathologists (CAP) Laboratory Accreditation Program. Our objective is to create an open platform for collaborative UDS data processing and revision control that is secure, easy to use, and highly accessible as a web application on our public webserver or on a laboratory's own local network.

1. Copperfish is a framework for revision control and validation of bioinformatic pipelines.

There are many revision control systems available for software development. However, none of these systems can track the application of bioinformatic pipelines to data sets. It is very easy for a single raw data set to proliferate into numerous processed versions. Copperfish is designed to track which pipeline (comprised of revisions of different scripts) was used to generate each result.

2. Copperfish is a framework for automating the execution of pipelines. The database will have all the necessary information to map one data set to another through the application of some method (*e.g.*, a Python script or call to an executable binary). A persistent filesystem will be populated with the end-products of pipeline revisions and the database tracks which pipelines were used. Executables will be required to pass unit tests, *e.g.*, is the binary installed on the system?

3. Copperfish is a framework for the immediate and transparent communication of bioinformatic development to end-users, *i.e.*, laboratory members generating data who need to apply methods to their data. It will always be clear how data were analyzed because the end-user will be able to look up which pipeline revision was used to process their raw data.