Abstract

The web-based bioinformatics platform Galaxy gained great popularity as a tool for allowing access to powerful compute clusters and sophisticated bioinformatics software with user-friendly point-and-click interfaces. In contrast, command line tools will always be more efficient for those who are comfortable working within a Unix-like environment. Likewise, biologists who understand the computing environment are more likely to understand what is and is not computationally efficient or possible. The obstacle lies in training users with little programming experience to be comfortable with the command line. We have explored alternate frameworks to Galaxy that allow users to design their workflows in a simplified web browser environment, and then automatically transfer these workflows into pipelines suitable for running at the command line.

Users design their workflows in Galaxy, execute at the terminal.

Data Mapping

Both Yabi and Galaxy use databases to store data on tools that have been integrated, histories that users have completed, workflows that users have saved and information on datasets. In order to transfer this data from one system to the other these databases need to be mapped and contingencies must be in place for data that is stored in one system but not the other.

Modular Design

As seen in Figure 2, above, the modular design is intended to allow flexible design of analysis pipelines that easily flows back and forth between the interactive Galaxy web server and the powerful Yabish shell implementation.

1. The user runs an exploratory, interactive data analysis through the Galaxy web server.
2. The user decides to turn the analysis they designed into a workflow using the Galaxy editor.
3. The user exports the Galaxy workflow, with automated conversion to Yabi format.
4. Yabi workflows can be run at the command line as a single pipeline with Yabish.

This enables the use of a Graphical User Interface for the analysis and design of the workflow, but the command line for the actual running of the newly created pipeline on additional samples. This scenario gently introduces advanced Galaxy users to the terminal environment.

Acknowledgements

We thank Tamas Szabo, Todd Heywood, Gerald McCloskey, Clever Tam, Osama El Demerdash, and Darian DeFalco for helpful discussions and help with compute cluster implementation design. This work was performed with assistance from the CSHL Bioinformatics Shared Resource, which is funded, in part, by an NIH/NCI Cancer Center Support Grant SP30CA45508.

References