Ergatis: A web-based bioinformatics pipeline management and collaborative development system

Orvis J[1][*], Crabtree J[2], Galens K[1], Inman JM[2], Lee E[4], Riley D[1], Sundaram J[2], Whitty B[3] , White O[1], Wortman JR[1], Angiuoli SV[1]

[1] Institute for Genome Sciences, University of Maryland School of Medicine.
[2] J. Craig Venter Institute, Rockville MD. [3] Michigan State University, East Lansing MI. [4] Lawrence Berkeley National Laboratory
* corresponding author (jorvis@som.umaryland.edu)

Ergatis is a web-based utility that is used to create, run, and monitor reusable computational analysis pipelines. It contains pre-built components for common bioinformatics analysis tasks. These components can be arranged graphically to form highly-configurable pipelines. Each analysis component supports multiple output formats, including the Bioinformatic Sequence Markup Language (BSML). The current implementation includes support for data loading into project databases following the CHADO schema, a highly normalized, community-supported schema for storage of biological annotation data.

Ergatis uses the Workflow Engine to process its work on a compute grid. Workflow provides an XML language and processing engine for specifying the steps of a computational pipeline. It provides detailed execution status and logging for process auditing, facilitates error recovery from point of failure, and is highly scalable with support for distributed computing environments. The XML format employed enables commands to be run serially, in parallel, and in any combination or nesting level.

This open-source framework has been employed in the annotation of several large, eukaryotic organisms, including Aedes aegypti and Trichomonas vaginalis as well as published comparative genome analyses and primary processing of the CAMERA metagenomic datasets.

Recent development has focused on better code portability and ease of collaborative pipeline development. An Ubuntu-based VMware image is available with Ergatis and all dependencies pre-installed and configured. A JeOS distribution has also been ported for use on Amazon's EC2 cloud computing service and work is ongoing to make it an available resource on TeraGrid.

Ergatis is distributed under the Artistic License at available at http://ergatis.sf.net.