Poster H-82

An Integrated System for High Throughput Genome Annotation and Comparative Genome Analysis

Authors:
Samuel V Angiuoli (*TIGR*)
Jonathan Crabtree (*TIGR*)
Joshua Orvis (*TIGR*)
Jennifer R Wortman (*TIGR*)
Owen R White (*TIGR*)

Short Abstract: We describe an architecture for large scale, high throughput sequence analysis and highlight two components: Ergatis, a workflow system for running computational analysis in a distributed computing environment and Sybil, a suite of web interfaces for comparative genome analysis that utilizes the Chado database schema and community ontologies.

Long Abstract:

We present an integrated system for large scale, high throughput genome annotation and comparative genome analysis that incorporates community file standards, ontologies, and a variety of open source software packages. The system also includes two novel open-source software packages: Sybil and Ergatis. Ergatis is a data management and workflow system targeted towards high throughput data generation in a distributed computing environment. Sybil is a suite of web interfaces and an associated API targeted for use in comparative genome analysis. We will describe the system architecture, algorithms used for genome annotation and comparative analysis, as well as performance and results from a number of recent projects that have utilized the system at The Institute for Genomic Research (TIGR).

The growth of available genome data for a wide variety of species has provided opportunities for comparative genomics studies while also raising challenges in large scale data management and effective exploration of large, diverse data sets. Additional challenges arise when attempting to incorporate data from a diverse set of sources due to inconsistent data models and conflicting semantics. The system presented here addresses these challenges by coupling generic systems, imposing strict use of ontologies and adherence to usage documentation. The system incorporates existing open standards including several ontologies from Open Biomedical Ontologies [http://obo.sourceforge.net], BSML (Bioinformatics Sequence Markup Language) [http://bsml.org], and the Chado database schema [http://gmod.sourceforge.net]. The system is presented in three sections: data generation, data storage, and data retrieval and exploration. This system is being used in a production environment at The Institute for Genomic Research (TIGR).

Data generation:

The Ergatis software package is meant to address many of the challenges of large-scale, high-throughput data generation. In particular, Ergatis provides a set of applications used to create and manage highly scalable computational pipelines that can be run on a distributed computing grid. The workflow backend to Ergatis provides an XML language and processing

engine for specifying the steps of a computational pipeline. For example a computational pipeline that identifies orthologous gene families may perform a series of discrete steps that includes: an all_vs_all BLAST search, parsing of BLAST output, clustering BLAST results, and building multiple sequence alignments. By richly defining the computation process in XML, Ergatis provides well-defined analysis pipelines that allow traceability and reproducibility. In addition, Ergatis 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, such as Condor and Sun Grid Engine. The XML format employed enables commands to be run serially, in parallel, and in any combination or nesting level.

As a front-end to the workflow engine, Ergatis is a web-based utility that is used to create, run, and monitor reusable computational analysis pipelines. It contains pre-built components for over 50 common bioinformatics analysis tasks, including numerous sequence alignment programs and gene finders. These components can be arranged in a graphical interface to form highly-configurable pipelines. The performance of several large comparative analysis pipelines executed using Ergatis on a 300+ node compute cluster will be presented. These pipelines include pipelines for ortholog, syntenic region, and SNP identification. We will provide an overview of the pipeline as well as a description of the algorithms used in these pipelines.

Data storage:
In order to integrate a diverse set of data types, we use a generic XML format, BSML (Bioinformatics Sequence Markup Language), to provide a standard encoding for the output of computational pipelines. BSML provides a rich set of XML elements that have allowed us to capture a diverse set of data types for all the pipelines supported in Ergatis. The Ergatis package includes converters for many common file formats to BSML, including converters for the file output formats from most commonly used sequence alignment programs. By documenting usage conventions and employing ontologies such as the Sequence Ontology, we are able to use BSML to encode a consistent data model for the output of all the computational pipelines supported by Ergatis. A mapping of BSML to the Chado relational database schema and an associated database loader provides the capability to store pipeline outputs in a community supported relational database schema. The Chado relational database scheme provides a generic, modular relational database schema that tightly integrates ontologies/controlled vocabularies. We will describe an overview of the data model used in BSML and Chado, focusing on development of data types required for comparative genomes. We will also describe community ontologies employed and a number of custom ontologies we've needed to develop to fully describe the computational data types.

Data retrieval/visualization:
Operating on top of the Chado relational database schema, Sybil is a web-based software package for the visualization and analysis of comparative genome data. We have developed Sybil to provide a rich set of web interfaces for browsing a diverse set of comparative data types. In addition, Sybil provides an API for querying complex data sets directly from the Chado relational database, including SNPs, orthologous families, and syntenic regions. Sybil has been used as an exploratory tool for looking at many comparative data types, including orthologous gene families, gene synteny, and SNPs within a diverse set of prokaryotic and eukaryotic taxa. We will highlight recent applications of Sybil interfaces to aid with analysis,

including a pan-genome analysis of Streptococci [PNAS, 102], the Pathema web site [http://pathema.tigr.org], and the comparative analysis of classes of parasites, such as Apicomplexa and Trypanomatids[Science, 309].