InterMine Webservices for Phytozome Rev2

Authors: Joseph Carlson[1], David Goodstein[1], Daniel Rokhsar[1]

[1] *U.S. Department of Energy Joint Genome Institute // LBNL - Walnut Creek, CA*

* *To whom correspondence may be addressed. Joseph Carlson, Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA, 94598, USA* - JWCarlson@lbl.gov@lbl.gov

July 10, 2014

ACKNOWLEDGMENTS:

Work by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

DISCLAIMER:

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California.

InterMine Webservices for Phytozome

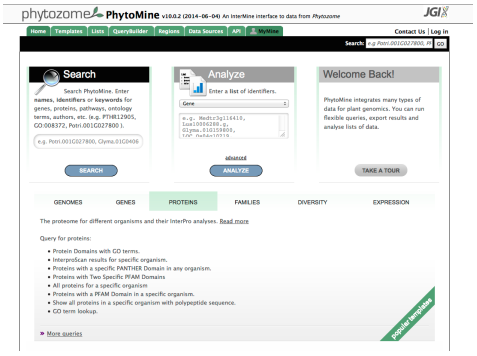
Joseph Carlson, David Goodstein and Daniel Rokhsar

DOE Joint Genome Institute

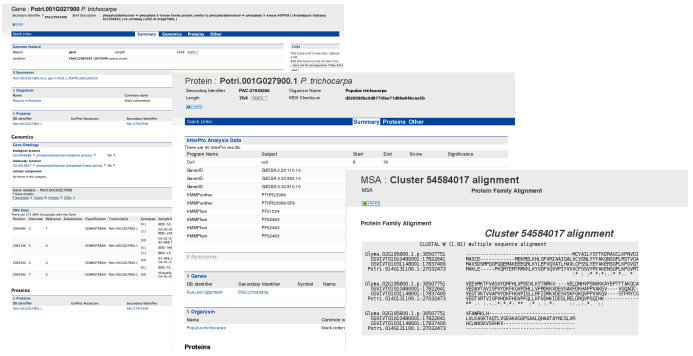
http://phytozome.jgi.doe.gov/phytomine

A datawarehousing framework for information provides a useful infrastructure for providers and users of genomic data. For providers, the infrastructure give them a consistent mechanism for extracting raw data. While for the users, the web services supported by the software allows them to make complex, and often unique, queries of the data. Previously, phytozome.net used BioMart to provide the infrastructure. As the complexity, scale and diversity of the dataset as grown, we decided to implement an InterMine web service on our servers. This change was largely motivated by the ability to have a more complex table structure and richer web reporting mechanism than BioMart. For InterMine to achieve its more complex database schema it requires an XML description of the data and an appropriate loader. Unlimited one-to-many and many-to-many relationship between the tables can be enabled in the schema.

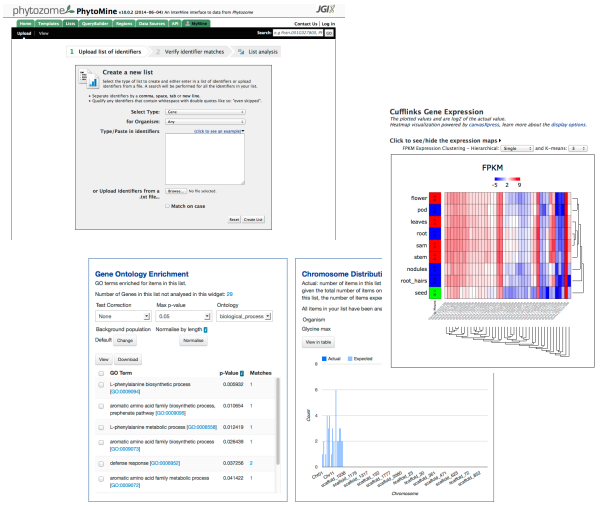
We have implemented support for:1.) Genomes and annotations for the data in Phytozome. This set is the 48 organisms currently stored in a back end CHADO datastore. The data loaders are modified versions of the CHADO data adapters from FlyMine. 2.) Interproscan results from all proteins in the Phytozome database. 3.) Clusters of proteins into a grouped heirarchically by similarity. 4.) Cufflinks results from tissue-specific RNA-Seq data of Phytozome organisms. 5.) Diversity data (GATK and SnpEFF results) from a set of individual organism. The last two datatypes are new in this implementation of our web services. We anticipate that the scale of these data will increase considerably in the near future.



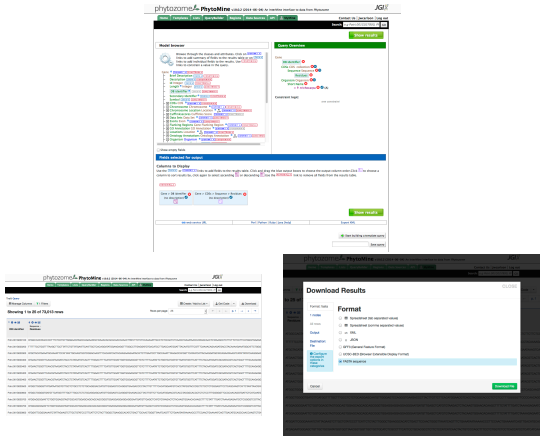
The opening page for the InterMine interface to PhytoWeb (PhytoMine) maintains the look and feel of other InterMine-based web applications. From the front page, you may conduct searches on database identifiers (gene names, GO terms, …), analyze a group of things, or access one of the existing template queries.



If a gene entered in a search, there is a report page available. All information present in the database is displayed or linked from this page, including transcripts, proteins, RNA-seq expression, SNPs, GO terms, and so on (depending on the data set).



Entering a list of genes, mRNAs, proteins, etc., displays a report page on the collection. This includes many of the InterMine reports including RNA-seq data (where available), GO enrichment, and so on. By creating an account with PhytoWeb, people can save lists permanently on our server.



The most powerful aspect of the web app is the ability to formulate and execute custom queries on the database. A graphical interface accessible from the QueryBuilder tab allow people to select which fields from the database are selected and impose constraints on the query. In this example, the gene name and sequence of all CDS's from that gene are displayed for all genes in poplar. The results are shown first in a paged view of tables, but can downloaded as a fasta file.

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.