

InterMine Webservices for Phytozome

Authors: Joseph Carlson¹, Richard Hayes¹, David Goodstein¹, Daniel Rokhsar¹

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*

January 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 W Carlson, Richard D Hayes, David M Goodstein and Daniel S Rokhsar

DOE Joint Genome Institute, Walnut Creek, CA



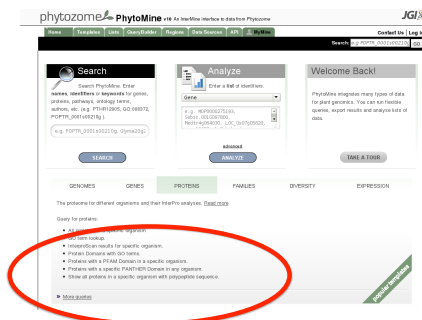
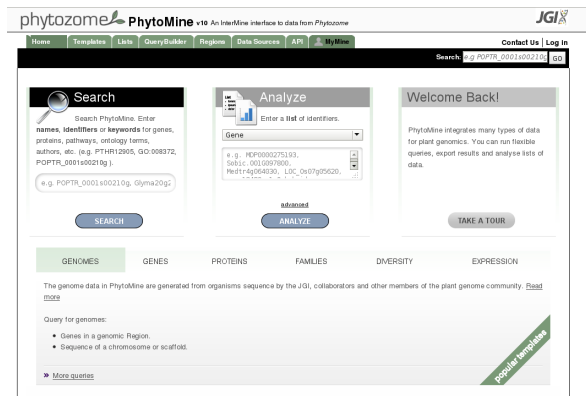
Visit www.phytozome.net for current status and deployment URL

A data warehousing framework for biological 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 either simple and common, or complex and 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 (www.intermine.org) 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.

Our implementation of InterMine is used to serve data for the next version of the web app on phytozome.net through their API. But it is also available for general use for information beyond what is available on our web pages.

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. Within this framework, complex SQL queries are generated which can return results in a highly optimized manner, with performance exceeding what we saw with BioMart.



One interface to phytomine relies on templated queries for generating results. A set of templates is linked from the start page. These can be used directly, or used as starting points for more complex queries by editing the templates.

SNP ID	Start	End	Reference	Gene	Type	Score
20811	10	10	A	BEUC-100	DOWNSTREAM	10.0
20812	10	10	A	BEUC-100	DOWNSTREAM	10.0
20813	10	10	A	BEUC-100	DOWNSTREAM	10.0
20814	10	10	A	BEUC-100	DOWNSTREAM	10.0
20815	10	10	A	BEUC-100	DOWNSTREAM	10.0
20816	10	10	A	BEUC-100	DOWNSTREAM	10.0
20817	10	10	A	BEUC-100	DOWNSTREAM	10.0
20818	10	10	A	BEUC-100	DOWNSTREAM	10.0
20819	10	10	A	BEUC-100	DOWNSTREAM	10.0
20820	10	10	A	BEUC-100	DOWNSTREAM	10.0
20821	10	10	A	BEUC-100	DOWNSTREAM	10.0
20822	10	10	A	BEUC-100	DOWNSTREAM	10.0
20823	10	10	A	BEUC-100	DOWNSTREAM	10.0
20824	10	10	A	BEUC-100	DOWNSTREAM	10.0
20825	10	10	A	BEUC-100	DOWNSTREAM	10.0
20826	10	10	A	BEUC-100	DOWNSTREAM	10.0
20827	10	10	A	BEUC-100	DOWNSTREAM	10.0
20828	10	10	A	BEUC-100	DOWNSTREAM	10.0
20829	10	10	A	BEUC-100	DOWNSTREAM	10.0
20830	10	10	A	BEUC-100	DOWNSTREAM	10.0
20831	10	10	A	BEUC-100	DOWNSTREAM	10.0
20832	10	10	A	BEUC-100	DOWNSTREAM	10.0
20833	10	10	A	BEUC-100	DOWNSTREAM	10.0
20834	10	10	A	BEUC-100	DOWNSTREAM	10.0
20835	10	10	A	BEUC-100	DOWNSTREAM	10.0
20836	10	10	A	BEUC-100	DOWNSTREAM	10.0
20837	10	10	A	BEUC-100	DOWNSTREAM	10.0
20838	10	10	A	BEUC-100	DOWNSTREAM	10.0
20839	10	10	A	BEUC-100	DOWNSTREAM	10.0
20840	10	10	A	BEUC-100	DOWNSTREAM	10.0
20841	10	10	A	BEUC-100	DOWNSTREAM	10.0
20842	10	10	A	BEUC-100	DOWNSTREAM	10.0
20843	10	10	A	BEUC-100	DOWNSTREAM	10.0
20844	10	10	A	BEUC-100	DOWNSTREAM	10.0
20845	10	10	A	BEUC-100	DOWNSTREAM	10.0
20846	10	10	A	BEUC-100	DOWNSTREAM	10.0
20847	10	10	A	BEUC-100	DOWNSTREAM	10.0
20848	10	10	A	BEUC-100	DOWNSTREAM	10.0
20849	10	10	A	BEUC-100	DOWNSTREAM	10.0
20850	10	10	A	BEUC-100	DOWNSTREAM	10.0
20851	10	10	A	BEUC-100	DOWNSTREAM	10.0
20852	10	10	A	BEUC-100	DOWNSTREAM	10.0
20853	10	10	A	BEUC-100	DOWNSTREAM	10.0
20854	10	10	A	BEUC-100	DOWNSTREAM	10.0
20855	10	10	A	BEUC-100	DOWNSTREAM	10.0
20856	10	10	A	BEUC-100	DOWNSTREAM	10.0
20857	10	10	A	BEUC-100	DOWNSTREAM	10.0
20858	10	10	A	BEUC-100	DOWNSTREAM	10.0
20859	10	10	A	BEUC-100	DOWNSTREAM	10.0
20860	10	10	A	BEUC-100	DOWNSTREAM	10.0
20861	10	10	A	BEUC-100	DOWNSTREAM	10.0
20862	10	10	A	BEUC-100	DOWNSTREAM	10.0
20863	10	10	A	BEUC-100	DOWNSTREAM	10.0
20864	10	10	A	BEUC-100	DOWNSTREAM	10.0
20865	10	10	A	BEUC-100	DOWNSTREAM	10.0
20866	10	10	A	BEUC-100	DOWNSTREAM	10.0
20867	10	10	A	BEUC-100	DOWNSTREAM	10.0
20868	10	10	A	BEUC-100	DOWNSTREAM	10.0
20869	10	10	A	BEUC-100	DOWNSTREAM	10.0
20870	10	10	A	BEUC-100	DOWNSTREAM	10.0
20871	10	10	A	BEUC-100	DOWNSTREAM	10.0
20872	10	10	A	BEUC-100	DOWNSTREAM	10.0
20873	10	10	A	BEUC-100	DOWNSTREAM	10.0
20874	10	10	A	BEUC-100	DOWNSTREAM	10.0
20875	10	10	A	BEUC-100	DOWNSTREAM	10.0
20876	10	10	A	BEUC-100	DOWNSTREAM	10.0
20877	10	10	A	BEUC-100	DOWNSTREAM	10.0
20878	10	10	A	BEUC-100	DOWNSTREAM	10.0
20879	10	10	A	BEUC-100	DOWNSTREAM	10.0
20880	10	10	A	BEUC-100	DOWNSTREAM	10.0
20881	10	10	A	BEUC-100	DOWNSTREAM	10.0
20882	10	10	A	BEUC-100	DOWNSTREAM	10.0
20883	10	10	A	BEUC-100	DOWNSTREAM	10.0
20884	10	10	A	BEUC-100	DOWNSTREAM	10.0
20885	10	10	A	BEUC-100	DOWNSTREAM	10.0
20886	10	10	A	BEUC-100	DOWNSTREAM	10.0
20887	10	10	A	BEUC-100	DOWNSTREAM	10.0
20888	10	10	A	BEUC-100	DOWNSTREAM	10.0
20889	10	10	A	BEUC-100	DOWNSTREAM	10.0
20890	10	10	A	BEUC-100	DOWNSTREAM	10.0
20891	10	10	A	BEUC-100	DOWNSTREAM	10.0
20892	10	10	A	BEUC-100	DOWNSTREAM	10.0
20893	10	10	A	BEUC-100	DOWNSTREAM	10.0
20894	10	10	A	BEUC-100	DOWNSTREAM	10.0
20895	10	10	A	BEUC-100	DOWNSTREAM	10.0
20896	10	10	A	BEUC-100	DOWNSTREAM	10.0
20897	10	10	A	BEUC-100	DOWNSTREAM	10.0
20898	10	10	A	BEUC-100	DOWNSTREAM	10.0
20899	10	10	A	BEUC-100	DOWNSTREAM	10.0
20900	10	10	A	BEUC-100	DOWNSTREAM	10.0

The output of a query is normally a list. The example shows a list of SNPs near an annotated gene in poplar. The output of the list may be downloaded, or, by selecting one of the displayed values, further information can be extracted.

We have implemented a database for:

1. Genomes and annotations for the data in Phytozome. This set is the 45 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 hierarchically 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 QueryBuilder interface of InterMine enables browsing the database schema and graphical selection of fields extracted in a query. It also allows constraints on the query as well. This example shows the edits possible for the query that extract all proteins in poplar that contain the PFAM domain PF11779.



Full reports of individual items displays information about the item and other items related to it. The example in the upper left shows a protein report and includes the associated gene, transcript and Interproscan results. Alternatively, as shown on the right, a set of genes can be used to extract a heat map of gene or mRNA expression.