InterMine Webservices for Phytozome

Authors: Joseph Carlson[1], Richard Hayes[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

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

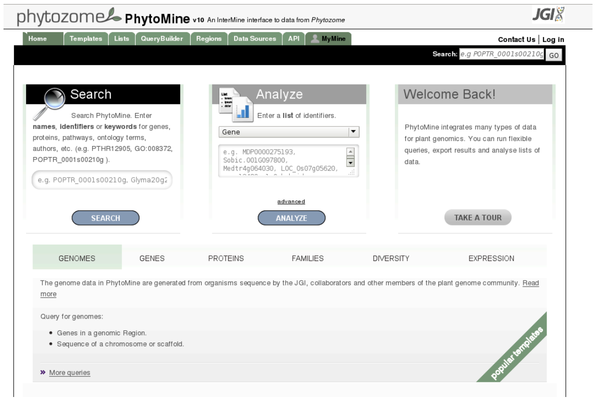
Visit www.phytozome.net for current status and deployment URL

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 performace exceeding what we saw with BioMart.



Search PhytoMine. Enter **names**, **identifiers** or **keywords** for genes, proteins, pathways, ontology terms, authors, etc. (e.g. PTHR12905, GO:008372, POPTR_0001s00210g).

e.g. POPTR_0001s00210g, Glyma20g2

Enter a **list** of identifiers.

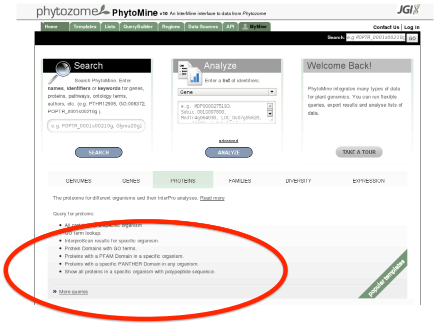
e.g. MDP0000275193, Sobic.001G097800, Medtr4g064030, LOC_Os07g05620,

PhytoMine integrates many types of data for plant genomics. You can run flexible queries, export results and analyse lists of data.

The genome data in PhytoMine are generated from organisms sequence by the JGI, collaborators and other members of the plant genome community. Read more

Query for genomes:

- Genes in a genomic Region.
- Sequence of a chromosome or scaffold.



Search PhytoMine. Enter **names**, **identifiers** or **keywords** for genes, proteins, pathways, ontology terms, authors, etc. (e.g. PTHR12905, GO:008372, POPTR_0001s00210g).

Enter a **list** of identifiers.

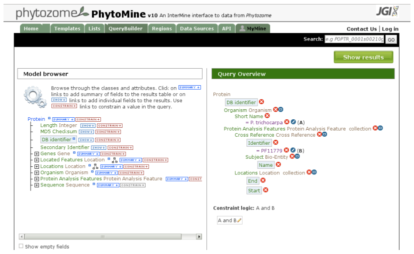
PhytoMine integrates many types of data for plant genomics. You can run flexible queries, export results and analyse lists of data.

The proteome for different organisms and their InterPro analyses. Read more

Query for proteins:

- All proteins in a specific organism
- GO term lookup.
- InterproScan results for specific organism.
- Protein Domains with GO terms.
- Proteins with a PFAM Domain in a specific organism.
- Proteins with a specific PANTHER Domain in any organism.
- Show all proteins in a specific organism with polypeptide sequence.

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.



Browse through the classes and attributes. Click on SUMMARY links to add summary of fields to the results table or on SHOW links to add individual fields to the results. Use CONSTRAIN links to constrain a value in the query.

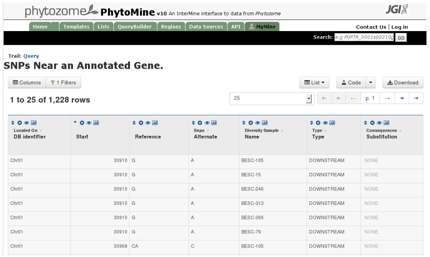
```
Protein
  Length Integer
  MD5 Checksum
  DB identifier
  Secondary Identifier
  Genes Gene
  Located Features Location
  Locations Location
  Organism Organism
  Protein Analysis Features Protein Analysis Feature
  Sequence Sequence
```

```
Protein
  DB identifier
  Organism Organism
    Short Name
      = P. trichocarpa (A)
  Protein Analysis Features Protein Analysis Feature  collection
    Cross Reference Cross Reference
      Identifier
        = PF11779 (B)
    Subject Bio-Entity
      Name
  Locations Location  collection
    End
    Start
```

Constraint logic: A and B

A and B

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.



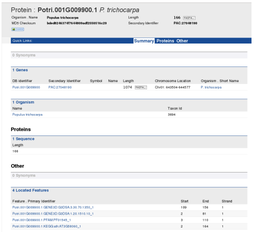
Trail: Query

SNPs Near an Annotated Gene.

1 to 25 of 1,228 rows

Located On DB identifier	Start	Reference	Snps Alternate	Diversity Sample Name	Type Type	Consequences Substitution
Chr01	30910	G	A	BESC-105	DOWNSTREAM	NONE
Chr01	30910	G	A	BESC-15	DOWNSTREAM	NONE
Chr01	30910	G	A	BESC-246	DOWNSTREAM	NONE
Chr01	30910	G	A	BESC-313	DOWNSTREAM	NONE
Chr01	30910	G	A	BESC-366	DOWNSTREAM	NONE
Chr01	30910	G	A	BESC-79	DOWNSTREAM	NONE
Chr01	30968	CA	C	BESC-105	DOWNSTREAM	NONE

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.



Protein : **Potri.001G009900.1** *P. trichocarpa*

Organism . Name	Populus trichocarpa	Length	166
MD5 Checksum	bded8246374f7644809adf2550516c29	Secondary Identifier	PAC:27048190

0 Synonyms

1 Genes

DB identifier	Secondary Identifier	Symbol	Name	Length	Chromosome Location	Organism . Short Name
Potri.001G009900	PAC:27048190			1074	Chr01: 643504-644577	P. trichocarpa

1 Organism

Name	Taxon Id
Populus trichocarpa	3694

Proteins

1 Sequence

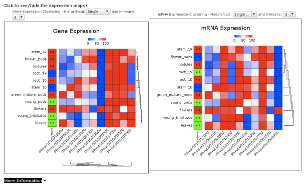
Length
166

Other

0 Synonyms

4 Located Features

Feature . Primary Identifier	Start	End	Strand
Potri.001G009900.1:GENE3D:G3DSA:3.30.70.1350_1	109	156	1
Potri.001G009900.1:GENE3D:G3DSA:1.20.1510.10_1	2	81	1
Potri.001G009900.1:PFAM:PF01545_1	3	110	1
Potri.001G009900.1:KEGG:ath:AT3G58060_1	2	164	1



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.

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 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 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.