InterMine

An Open Source Data Warehouse and Query Interface

Fengyuan Hu, Alex Kalderimis, Daniela Butano, Radek Stepan, Sergio Contrino, Julie Sullivan, Dan Tomlinson, Mike Lyne, Adrian Carr, Richard Smith and Gos Micklem (PI)
University of Cambridge, UK.

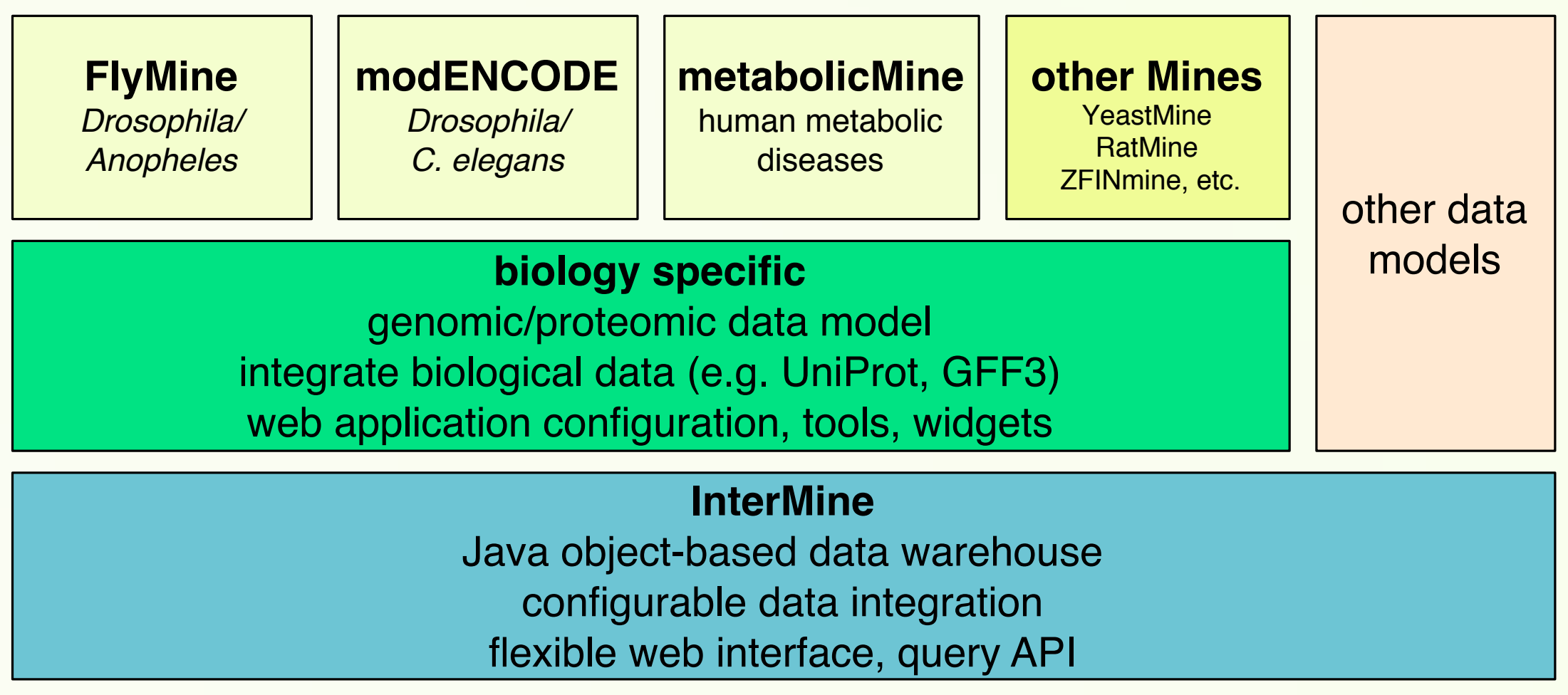
Introduction

InterMine (www.intermine.org) is a powerful open source system for building query-optimised data warehouses. It supports data integration from standard biological formats and makes it easy to add your own data. A sophisticated web application provides flexible query access for any data model. Data can be programmatically accessed by building and executing queries via the InterMine web-service API whose code can be generated within the web application in various programming languages.

Originally developed for FlyMine, there are now many InterMine implementations in operation. In particular, InterMine is used by several Model Organism Databases (MODs), including SGD (YeastMine), RGD (RatMine), and ZFIN (ZFINmine), the founding members of the InterMOD project (MOD mines for worm, mouse and *S. pombe* have been planned). InterMine is also the engine behind several dedicated data-mining projects, including metabolicMine and the modEncode project. A number of private and independent mines also exist.

Software Overview

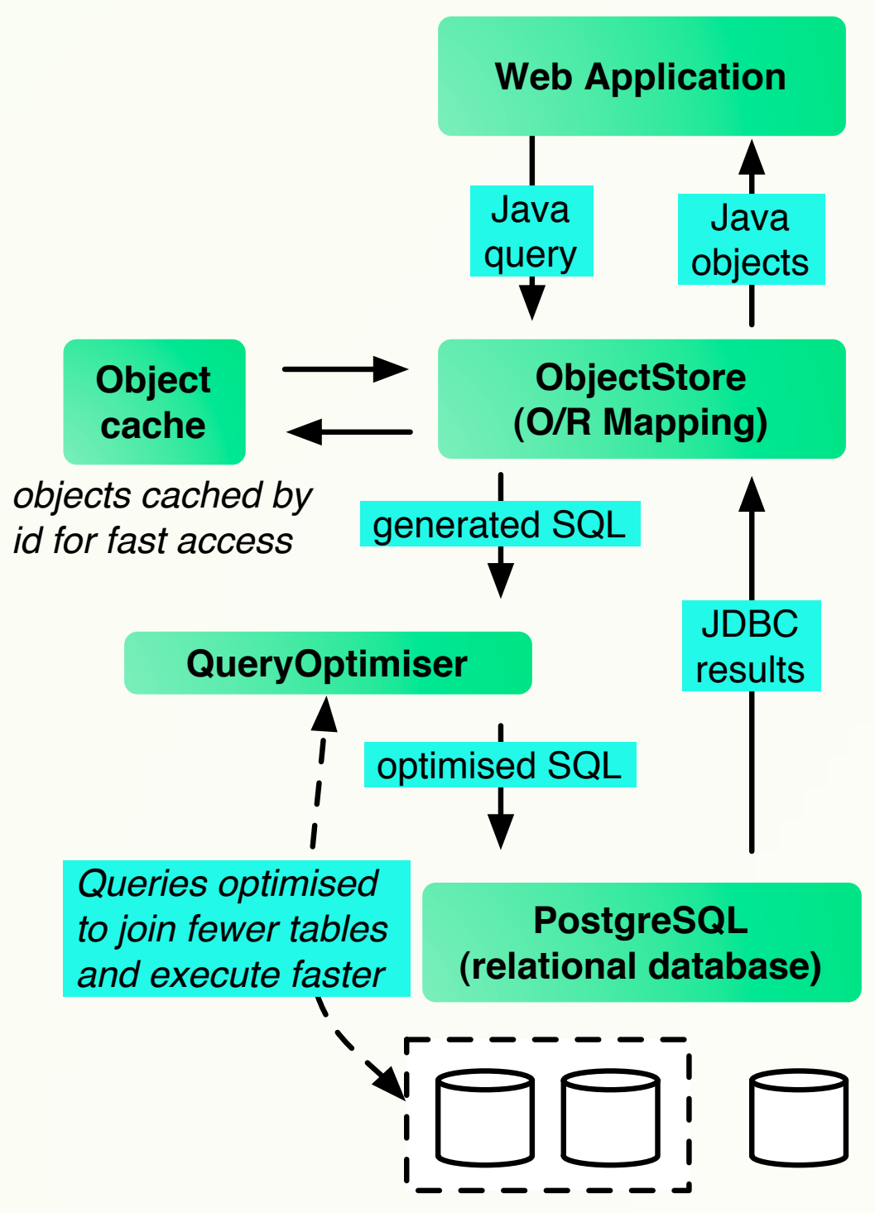
All InterMine code is freely available under the LGPL license.


FlyMine
Drosophila/
Anopheles
modENCODE
Drosophila/
C. elegans
metabolicMine
human metabolic
diseases
other Mines
YeastMine
RatMine
ZFINmine, etc.
other data
models
biology specific
genomic/proteomic data model
integrate biological data (e.g. UniProt, GFF3)
web application configuration, tools, widgets
InterMine
Java object-based data warehouse
configurable data integration
flexible web interface, query API

Architecture

InterMine is written in Java and built around a custom **object/relational mapping** tool optimised for read-only performance. The data model is object-based and defined by an XML file from which a database schema and Java classes are automatically generated.

A **cost-based query optimiser** attempts to re-write incoming queries to make use of pre-computed tables to join fewer tables and run faster. New pre-computed tables can be added at any time to adapt to actual usage. In contrast to traditional data warehouses performance optimisation is thus separated from schema design.



NOTE: ObjectStore is not related to commercial product of the same name

Data Sources

+

Custom Data Sources

FASTA GFF3 XML Tabular

Configurable
data integration

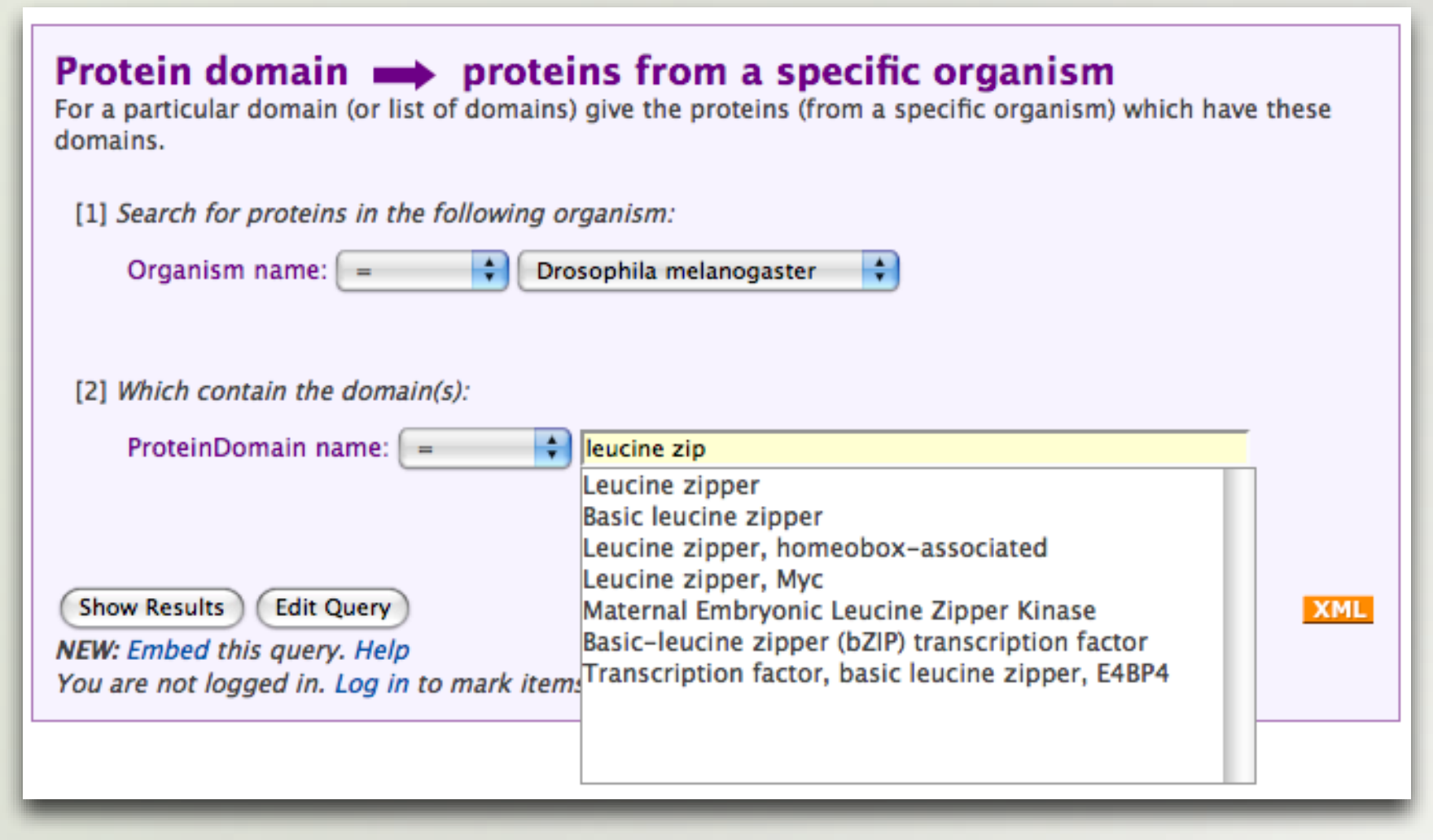
InterMine
Data Warehouse

InterMine makes it easy to integrate multiple data sources into a central data warehouse. It has a core data model based on the sequence ontology and supports many biological data formats. The object-based data model is defined as XML from which a database schema and Java classes are automatically generated. It is simple to add custom sources to extend the data model and integrate your own data. Java and Perl APIs are provided for converting data.

Supported formats include GFF3, FASTA, Chado, GO gene association, UniProt XML, PSI XML, PDB XML and Ensembl. A central configuration file defines the sources to load into an InterMine instance and specific organisms and data files to include.

Web Application

A web application works 'out of the box' for any data model to provide flexible query access to the data warehouse, designed for functionality beyond looking up an identifier and viewing a report. The application is highly customisable through configuration and development of new data displayers or integration of tools.



Protein domain ➡ proteins from a specific organism

For a particular domain (or list of domains) give the proteins (from a specific organism) which have these domains.

[1] *Search for proteins in the following organism:*

Organism name: = Drosophila melanogaster

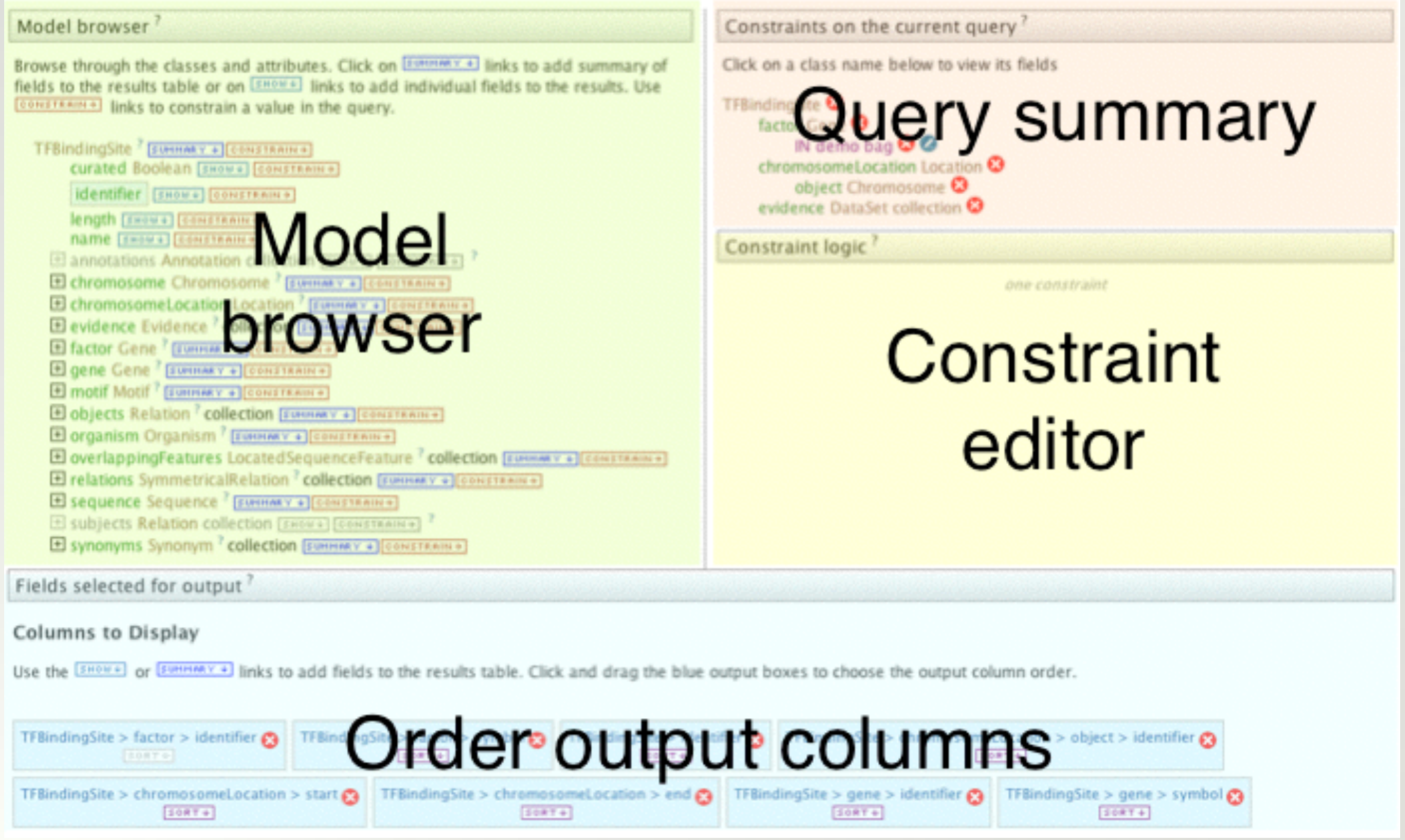
[2] *Which contain the domain(s):*

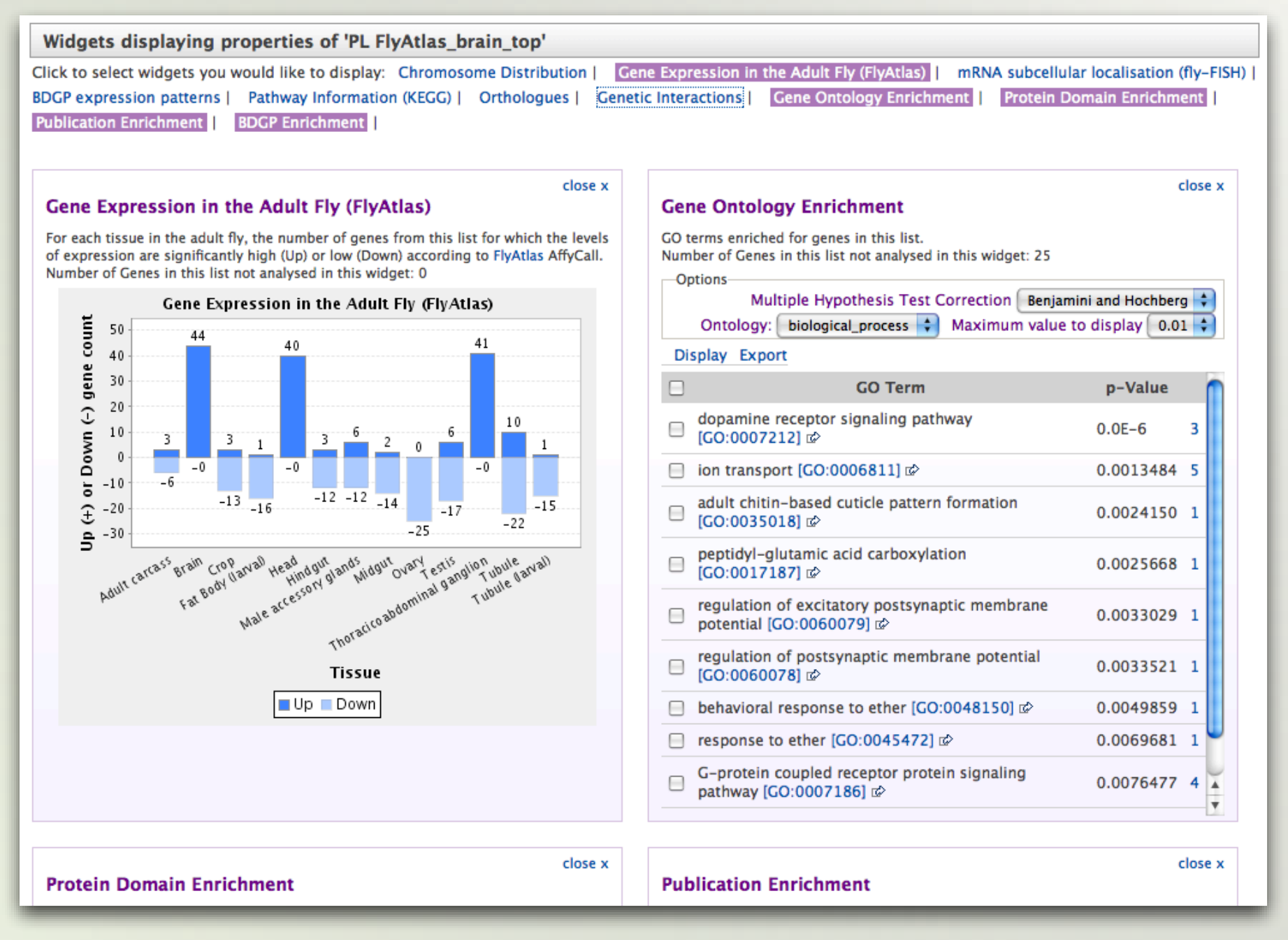
ProteinDomain name: = leucine zip

Leucine zipper
Basic leucine zipper
Leucine zipper, homeobox-associated
Leucine zipper, Myc
Maternal Embryonic Leucine Zipper Kinase
Basic-leucine zipper (bZIP) transcription factor
Transcription factor, basic leucine zipper, E4BP4

NEW: Embed this query. Help
You are not logged in. Log in to mark items

A **QueryBuilder** (right) allows non-programmers to create custom queries. Any query can be turned into a re-usable **query form** (above) providing a simple form with a description and editable constraints.




Gene Expression in the Adult Fly (FlyAtlas)
Gene Ontology Enrichment
Protein Domain Enrichment
Publication Enrichment

Lists of any type (e.g. genes, proteins) can be uploaded or created from query results. Lists can be used in any query or template in place of a single identifier. A **list analysis** page (left) summarises properties of a list through interactive **widgets**, for example graphs of gene expression or statistical enrichment of GO terms or protein domains. A framework is provided for creating new widgets.

A **MyMine** account allows users to save lists and queries between sessions.

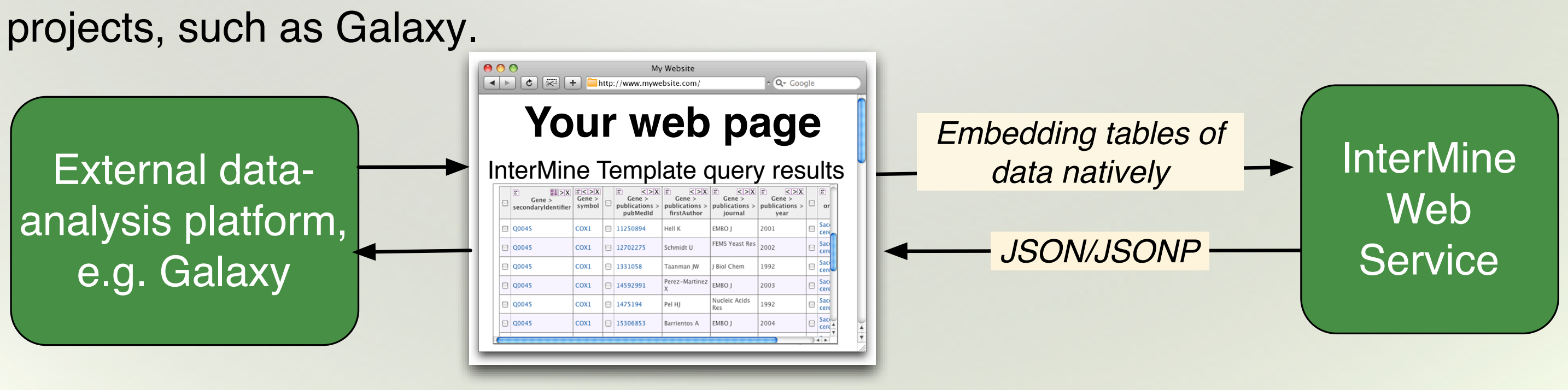
In the biology domain, mines have been automatically enabled to send data to external platforms such as Galaxy in different data formats.

An important feature is that much of the presentation can be controlled by non-programmers: an **admin user** can create and publish new template queries and lists at any time and can adapt report pages by applying tags.

Interoperability

The increasing number of InterMine implementations has allowed for greater interoperation, which is mediated through webservices over a published RESTful API. This allows sites to query a mine's data, and for any website to embed tables of data natively in their pages. The foundational technologies behind this are webservices that return JSON, JSONP for cross-domain communication, and a JavaScript library for web client access.

Webservice interoperability enables mines to display homology data and provide links to data in different species. End users can automate workflows to access public or private resources (authentication reqired) using client libraries in Java, Perl and Python. Mines can integrate with work-flow and data-analysis projects, such as Galaxy.


External data-analysis platform, e.g. Galaxy
My Website
http://www.mywebsite.com/
Your web page
InterMine Template query results
Embedding tables of data natively
JSON/JSONP
InterMine Web Service

UNIVERSITY OF CAMBRIDGE
Cambridge Systems Biology Centre
Tennis Court Road, Cambridge CB2 1QR, UK
Tel: +44 1223 760262 Email: info@intermine.org

genome.gov
National Human Genome Research Institute
National Institutes of Health

wellcome trust

www.intermine.org
info@intermine.org