InterMine as a Science Gateway for Systems Biology

Aaron Golden
School of Mathematics, Statistics & Applied Mathematics
National University of Ireland, Galway

IWSG 2017 – Poznan, Poland
Talk Roadmap

(1) Desktop technologies & systems biology

(2) Problem – how to leverage resulting data space

(3) Sequence Ontology & Data Schemas

(4) InterMine – how it works as a Science Gateway
We finally have a means of fully linking all of the active components that make up the machinery of a living organism. This enables us to perform Systems Biology.
The Problem...

Each `step' in this living workflow is easily examined using modern 'off the shelf' devices.

The problem is **integrating** the resulting processed data products!
Ontologies to the Rescue...

Solution lies in the
- **the genome**

Genomic features described as parts of gene models, assembly components, experimental results generating annotations defined as being associated with a specific location...

Forms an excellent means to bind the various systems biology datasets together in the same data space
The basis for building a schema...

List of names, locations, properties

Genomics

- Genes (~25,000 genes, 23 pairs of chromosomes)
- Gene Expression
- RNA
  - Translation
  - Proteins (>1,000,000)
    - Proteomics
      - List of names, quantity
- Molecular Interactions
- Cellular Function
- Functional End Products
- Metabolites (~2,500 small molecules)
  - List of names, quantity
- Lab work yields different types of data files, with differing information content – yet common theme is a name or, failing that, an originating location
Sequence Ontology – e.g. mapping experimental results

- **Genome annotation**
  - Sequence feature
    - biological region
    - gene
    - regulatory region
    - transcription regulatory region

- **ChIP-chip data**
  - Binding site
  - Protein binding site
  - Histone binding site

InterMine
Conceptual Solution...

Idealized system ultimately integrates data products – provides means for researchers to access & query data space.
Introducing InterMine...

InterMine has three parts:

1. Database
   a. Load data from different data sets into a single database

2. Webapp
   a. Mine the data
   b. Visualise!

3. Web services
InterMine - Backend

Originally developed by Gos Micklem's group for FlyMine (2007)

Uses read-only **ObjectStore** methodology to implement fast queries on linked dataspace using Seq. Ont. schema.

Web-enabled access also via RESTful Perl, Python, Java, Ruby, R APIs

Java/PostgreSQL core code base, Tomcat for web services
Main step - Constructing & Loading a Data Model (using SO)

Data Model for ObjectStore based on
- experimental data space
- organism's genome(s)
- literature references
- published/archived 'omics data
- Data loaded from parsers to XML InterMine format
The InterMine `Science Gateway' web portal
The InterMine `Science Gateway' web portal

HumanMine loads Gene models and sequence from ENSEMBL. We additional cross-reference Gene information from many additional sources. Read more

Query for genes:
- GO term ➔ Genes
- Gene ➔ GO terms.
- Gene ➔ Chromosomal location.
- Human Disease ➔ [Human +] Orthologue Gene(s)
- Region ➔ Genes
- Chromosomal Location ➔ All Genes + Transcripts + Exons
- Gene + GO term ➔ Genes by GO term
- Gene ➔ Overlapping genes.

Perl, Python, Ruby and Java API
Access our HumanMine data via our Application Programming Interface (API) tool! We provide client libraries in the following languages:
- Perl
- Python
- Ruby

News & Updates
16 Out and about where to find InterMiners over June and July 2017
Jun We recently added a public google calendar you can subscribe to if you...
13 InterMine community roundup: June 2017
Jun Here are some of the exciting things that have been happening in the I...
12 Bioschemas Summer Progress and InterMine
Jun A couple of weeks ago we took part in the May ELIKIR Bioschemas meetin...

More news
Using the QueryBuilder tool to explore dataspace
Using Web Services remotely...

List of Identifiers → List of database records → Homologues in X → Their protein domains

use InterMine::Webservice;

my $flymine = Webservice::InterMine->get_service('www.flymine.org/query', 'TOKEN');
my $list = $flymine->new_list(type => 'Gene', content => "some/file.txt");

my $query = $list->to_query;
$query->add_views('homologues.homologue.symbol');
$query->add_constraint('homologues.homologue.organism.shortName', '=', 'R norvegicus');
my $rat_symbols = [map {$_->{homologues.homologue.symbol}} $query->results('hashrefs')];

my $ratmine = Webservice::InterMine->get_service('ratmine.mcw.edu/ratmine');
my $rat_query = $ratmine->new_query(class => 'Gene');
$rat_query->add_views(qw/symbol primaryIdentifier proteins.proteinDomains.name/);
$rat_query->add_constraint('symbol', 'ONE OF', $rat_symbols);
$rat_query->show;
InterMines, Intermines...

EXISTING MINES

A number of different data warehouses powered by InterMine already exist. These include:

- **FlyMine** - *Drosophila* genomics
- **modMine** - fly and worm modENCODE data
- **MouseMine** - at MGI
- **RatMine** - at RGD
- **WormMine** - at WormBase
- **YeastMine** - at SGD
- **ZebrafishMine** - at ZFIN
- **INDIGOMine** - microbes
- **ThaleMine** - Araport Project with data for Arabidopsis thaliana
- **ChickpeaMine** - Desi & Kabul chickpea
- **TargetMine** - drug target discovery
- **MitoMiner** - proteomic data for mitochondria
- **HumanMine** - human
- **FlyTF.org** - *Drosophila* transcription factors
- **PhytoMine** - plants
- **MedicMine** - *Medicago truncatula*
- **BovineMine** - *Bos Taurus*
- **HymenopteraMine** - Bees, Ants & Wasps
- **SoyMine** - Soybase soy bean data
- **CHOMine** - *Cricetulus griseus* and CHO cells
- **BeanMine** - LegFed chado bean data
- **LegumeMine** - String bean, Soy, and Peanut
- **PeanutMine** - Peanut chado/GFF data
- **Shaare** - Gene candidate prioritisation
- **PlanMine** - Planarian flatworms
- **Wheat3BMine** - Wheat chromosome 3B
- **GrapeMine** - Grapevine
- **RepetDB** - repetitive DNA elements
- **XenMine** - Xenopus
- **TetraMine** - *Tetrahymena thermophila*
What about the DCI thing?

InterMines' tend to be standalone 'silos' based one model organism/biological process.

ToxoMine:
- DB server (8 core, 32GB RAM, 2 TB)
- Web servers (4 core, 16GB, 256 GB)

However – **fundamental homology between all living organisms enables interoperability 'for free' through the Sequence Ontology.**

Offers a Virtual Observatory like experience across registered Mines.
Cross-species query can be performed using the InterMOD infrastructure

List Analysis for GCN5b protein interacting partners (41 Genes)

<table>
<thead>
<tr>
<th>DB identifier</th>
<th>Symbol</th>
<th>Description</th>
<th>Organism</th>
</tr>
</thead>
<tbody>
<tr>
<td>TGE49_203220</td>
<td>TGE49_203220</td>
<td>DEAD/DEAH box helicase domain-containing protein</td>
<td>T. gondii ME49</td>
</tr>
<tr>
<td>TGE49_204140</td>
<td>TGE49_204140</td>
<td>Ph-D finger-domain-containing protein</td>
<td>T. gondii ME49</td>
</tr>
<tr>
<td>TGE49_200280</td>
<td>TGE49_200280</td>
<td>hypothetical protein</td>
<td>T. gondii ME49</td>
</tr>
<tr>
<td>TGME40_213900</td>
<td>TGME40_213900</td>
<td>regulator of chromosome condensation RCC1</td>
<td>T. gondii ME49</td>
</tr>
<tr>
<td>TGME40_214240</td>
<td>TGME40_214240</td>
<td>bromodomain-containing protein</td>
<td>T. gondii ME49</td>
</tr>
<tr>
<td>TGME40_214960</td>
<td>TGME40_214960</td>
<td>AP2 domain transcription factor AP2K-II (AP2K8)</td>
<td>T. gondii ME49</td>
</tr>
<tr>
<td>TGME40_216750</td>
<td>TGME40_216750</td>
<td>Par/tRNA polymerase II complex subunit 3 family 1 ASC03L1, putative</td>
<td>T. gondii ME49</td>
</tr>
<tr>
<td>TGME40_223390</td>
<td>TGME40_223390</td>
<td>activating signal co/integrator 1 complex subunit 3 family 1 ASC03L1, putative</td>
<td>T. gondii ME49</td>
</tr>
<tr>
<td>TGME40_224260</td>
<td>TGME40_224260</td>
<td>PhD finger-domain-containing protein</td>
<td>T. gondii ME49</td>
</tr>
<tr>
<td>TGME40_225910</td>
<td>TGME40_225910</td>
<td>hypothetical protein</td>
<td>T. gondii ME49</td>
</tr>
</tbody>
</table>

Convert to a different type

Protein 1
Orthologues
D. melanogaster (12) H. sapiens (18) M. musculus (18) R. norvegicus (10)

View homologues in other Mines:
- FlyMine
- RatMine
- YeastMine
- NZWAmine
- MouseMine
- modMine
- D. melanogaster
- metabolicMine

External Links

NUI Galway O'Gaillimh

ToxoMine

FlyMine

RatMine

YeastMine
**Ongoing InterMine Community Activity**

**RDFization of InterMine**

- improve interoperability
- enable federated queries across larger RDF space
- better packaging for local/cloud deployment

**Enchanced Webservice functionality**

- move away from Java/Tomcat paradigm
- modularity to integrate third-party web apps
- Clojure, Angular, PostgreSQL, Web Services
Acknowledgements

In Galway
David Rhee
Pilib Ó Broin
Kevin Shieh

Gos Micklem's Group (Cambridge)
Yo Yehuda
Julie Sullivan
Xiaonan Wang

Kami Kim's Lab (Albert Einstein College of Medicine)
Matthew Croken
Sheila C Nardelli
Ravikant Ranjan
Natalie Silmon De Monerri
Zoi Tampaki
Li-Min Ting
Netha Ulahannan

Funding
National Institute of Allergy and Infectious Diseases (#RC4AI092801)
Wellcome Trust (#099133)
National Human Genome Research Institute (#R01HG004834)

And the National Science Foundation!