BioMart

Data integration in four easy steps

Arek Kasprzyk
European Bioinformatics Institute
22 July 2006
BioMart

- A joint project
  - European Bioinformatics Institute (EBI)
  - Cold Spring Harbor Laboratory (CSHL)

- Funding
  - Wellcome Trust
  - European Commission
  - NIH
Synopsis

• **Higher level data management system**
  - Data mining type access to descriptive data
  - Query optimization
  - Data federation
  - Meta data support
BioMart

1. Source data
2. Transformation
3. Configuration
4. BioMart software
5. Querying
Transformation and Configuration Tools
Query interfaces
Programmatic access

• APIs
  – Perl (biomart-plib)
  – Java (martj)
  – R (biomaRt)

• Web service
Data federation

MySQL  ORACLE  PostgreSQL

REGISTRY

XML  XML  XML

XML  XML  XML
Dataset, Attribute and Filter

Dataset

Attribute

Filter

<table>
<thead>
<tr>
<th>GENE</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene_id(PK)</td>
</tr>
<tr>
<td>gene_stable_id</td>
</tr>
<tr>
<td>gene_start</td>
</tr>
<tr>
<td>gene_chrom_end</td>
</tr>
<tr>
<td>chromosome</td>
</tr>
<tr>
<td>gene_display_id</td>
</tr>
<tr>
<td>description</td>
</tr>
</tbody>
</table>
Joining two datasets

Dataset 1

Exportable

name = uniprot_id
attributes = uniprot_ac

Links

Dataset 2

Importable

name = uniprot_id
filters = uniprot_ac
Dataset linking
Third party software
Ensembl
The Generic Model Organism Database Project
Submitted by mod on Thu, 2005-11-17 15:40.

About GMOD
Submitted by mod on Wed, 2005-11-23 18:01.
The Generic Model Organism Database (GMOD) Project is a largely open source project to develop a complete set of software for creating and administering a model organism database. Components of this project include genome visualization and editing tools, literature curation tools, a robust database schema, biological ontology tools, and a set of standard operating procedures. This project is funded by the NIH and the USDA Agricultural Research Service, with participation from members of several database projects, including WormBase, FlyBase, Mouse Genome Informatics, Gramene, the Rat Genome Database, TAIR, Eucyc, and the Saccharomyces Genome Database. For more information, please see http://www.gmod.org/ or contact Scott Cain <cain@cshl.edu>.

> login or register to post comments | read more

Frequent GMOD Downloads

- GMOD Core (Chado database and tools)
  - **gmom-core**: the schema and tools to install it. [Download]
  - **XORT**: a tool for loading and dumping chado-xml [Download]

- **GMODTools**: [this needs a description, and a package for that matter]

- Genome Editing and Visualization
  - **Apollo**: a Java application for viewing and editing genome annotations [Download]
  - **GBrowse**: a CGI application for displaying genome annotations [Download]

- Comparative Genomics
  - Synbrowse

more
Distributed Annotation System
Welcome to Galaxy! Galaxy is a web-based data analysis framework that allows you to collect and work with biological data from many different sources.

Watch Galaxy screencasts to see how it works. Just click below...

Galaxy 2.2 has a slight problem when viewed with Microsoft Internet Explorer. It works best with Firefox or Safari. We are working on resolving this issue.

- Galaxy wiki page and documentation
- People and Technology behind Galaxy
- Report problems to Galaxy mailing list: galaxy-user@bx.psu.edu

The Galaxy team is a part of the Center for Comparative Genomics and Bioinformatics at Penn State.
Examples
Genomic data
Uniprot, MSD, ArrayExpress
Proteomic, structure, expression
Model organism databases

Genes
Expression
Phenotypes
Variations
Literature
Ontologies
Sequence
Zebra Fish models for human development and disease
Select the dataset for this query

Database
- ENSTMBL 37 (SANGER)
- SNP 37 (SANGER)

Dataset
- VEGA 37 (SANGER)
- UNIPROT 4-5 (EBI)
- MSW 4 (EBI)
- WORMBASE CURRENT (CSHL)

Using MartView, you can apply FILTERS on the next page and then, which data you want to output on the OUTPUT page. At any stage, the COUNT button will show the number of entries you can expect in the final output.

MartView can generate a number of different types of output, including sequence and tabulated list data. Multiple output formats, including HTML, text and Microsoft Excel, are also supported.

For a bookmarkable version of this page, click [here]
Behind closed doors ;)
Target SNP selection for the study of one autoimmune disease, type 1 diabetes (T1D), and infectious diseases, malaria and dengue

Laboratory of Genetics of Infectious and Autoimmune Diseases
<table>
<thead>
<tr>
<th>Name</th>
<th>Position</th>
<th>Allele</th>
<th>Strand</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP1</td>
<td>AL1392381659852</td>
<td>T/A</td>
<td>1</td>
</tr>
<tr>
<td>SNP2</td>
<td>NT 25698</td>
<td>C/T</td>
<td>-1</td>
</tr>
<tr>
<td>SNP3</td>
<td>chr13</td>
<td>C/G</td>
<td>1</td>
</tr>
</tbody>
</table>

Genetics of Infectious and Autoimmune Diseases, Pasteur Institute, INSERM U730, Paris, France.

Data conversion and integration

Diabetes-Gene Association Database

Combined proprietary and public data

UCSC  | HapMap  | Ensembl |

NCBI  | Proprietary data |
Using the Molecular Integration Database to Answer CAPRISA’s Questions

Research that contributes to understanding HIV pathogenesis and epidemiology as well as HIV/AIDS treatment and prevention
How is the MID populated?

Cellular Immunity
Humoral Immunity
HLA Typing
Sequence & Sequence Related
Clinical Data

Pipeline

MID
Select the dataset for this query

Dataset: MID (dev)

Using MartView

After choosing a DATASET above, select some FILTERS on the next page and then which data you want to EXPORT from the OUTPUT page. At any stage the COUNT button will calculate the number of entries you can expect in the final output.

MartView can generate a number of different types of output, including sequence and tabulated list data. Multiple output formats, including HTML, text and Microsoft Excel, are also supported.
What role for ‘Omics’?

- **Human study to evaluate Omics** in assessing safety indicators

- **Study of skin inflammation** in response to detergent
  - Skin samples taken and analyzed with multiple Omics techniques.
    - Blood
    - Skin biopsy
    - Microdialysis
• Requires an extensible file and metadata management system for omics data

• **Oracle 9i** database used for staging area and BioMarts
• **Database indexes files** on a separate file system
Adding Annotation

- Query Ensembl for details of genes measured or identified in experiments e.g. GeneSpring Annotation

- For example, we can link to Ensembl from Microarray Experiments by Gene ID
Four easy(?) steps
Step 1

Transformation
Step 2
Configuration
Step 3

Query
User interfaces
Web service

<Query virtualSchemaName = "default" count = "0" >
  <Dataset name = "hsapiens_gene_ensembl">
    <Attribute name = "gene_stable_id" />
    <Filter name = "chr_name" value = "22"/>
  </Dataset>
  <Dataset name = "uniprot">
    <Attribute name = "accession" />
    <Filter name = "pfam" value = "only"/>
  </Dataset>
</Query>
my $initializer = BioMart::Initializer->new('registryFile' =>$confFile);
my $registry = $initializer->getRegistry();
$registry->configure();

$query->addAttribute('hsapiens_gene_ensembl','ensembl_gene_id');
$query->addFilter('hsapiens_gene_ensembl','chromosome_name','[1]');
$query->addAttribute('uniprot','accession',);
$query->addFilter('uniprot', 'chromosome_name','[1]');
$query->formatter('HTML');

my $runner = BioMart::QueryRunner->new();
$runner->execute($query);
$runner->printResults();
Step 4

Ask for a pay rise :)}
Summary

• A generic data management system
• Provides building blocks for designing your own ‘tailor-made’ data management
  – A set of easily configurable user interfaces
  – Distributed Data federation
  – Query optimization
• Easy to install and manage
  – A project for bioinformatics students
• Open source software.
  – No restrictions for academics or commercial users
Credits

- **BioMart**
  - Syed Haider
  - Richard Holland
  - Damian Smedley
  - Gudmundur Thorisson

- **Contributors**
  - Steffen Durinck (NCI, NIH)
  - Eric Just (Northwestern University)
  - Don Gilbert (Indiana University)
  - Darin London (Duke University)
  - Will Spooner (CSHL)
  - Benoit Ballester (Universite de la Mediterranee)
  - James Smith (Ensembl)
  - Arne Stabenau (Ensembl)
  - Andreas Kahari (Ensembl)
  - Craig Melsopp (Ensembl)
  - Katerina Tzouvara (EBI)
  - Paul Donlon (Unilever)