# The Grinder and Gene Sets

## Master's Thesis

Greg Dougherty

Dept of BioMolecular Engineering, UCSC

November 28, 2007

# Points of this talk

- What is the Grinder.
- How do you use it.
- How do you add data and mappings to it.
- Gene Sets.

# What is the Grinder?

The Grinder is a Web accessible database that tracks mappings between Keyspaces.

# What is a Keyspace?

- A set of **unique** identifiers for biological information that can be **mapped** to other biological information.
- Allowed mappings: 1-1, 1-Many, Many-1, Many-Many.
- Characteristics:
  - Name
  - Description
  - Species
  - Type
  - Source: URL and / or FTP location for origin of the data
  - Download date

# What is a Keyspace?

- CREATE TABLE KeySpaces (
- id INT UNSIGNED NOT NULL AUTO_INCREMENT DEFAULT NULL,
- name VARCHAR(30) NOT NULL UNIQUE,
- species VARCHAR(30) NOT NULL,
- description VARCHAR(255) NOT NULL DEFAULT ' ',
- type INT UNSIGNED DEFAULT NULL REFERENCES KeySpaceTypes (id),
- url VARCHAR(10000) DEFAULT NULL,
- ftp VARCHAR(10000) DEFAULT NULL,
- lastDownloaded TIMESTAMP DEFAULT CURRENT_TIMESTAMP,
- PRIMARY KEY (id),
- INDEX (species))
- ENGINE = MyISAM;

# What is a Mapping?

- A one way or bidirectional equivalence between two bits of biological information.
- Example: GenBank Accession Number U48705 (discoidin domain receptor) maps to LocusLink accession number 780
- Characteristics:
  - Name
  - Description / Provenance
  - Quality
  - Source: URL and / or FTP location for origin of the data
  - Download date

# What is a Mapping?

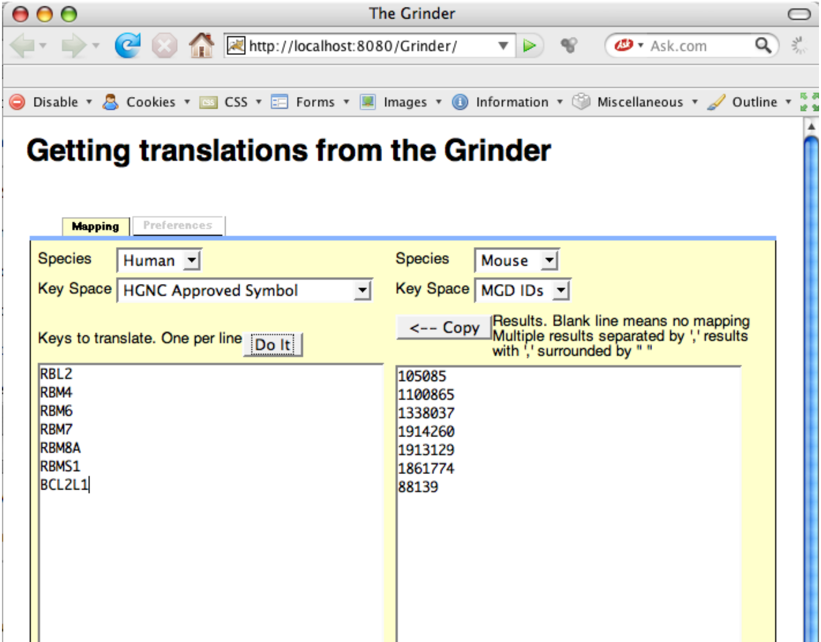
- CREATE TABLE Mappings (
- id INT UNSIGNED NOT NULL AUTO_INCREMENT DEFAULT NULL,
- name VARCHAR(30) NOT NULL UNIQUE,
- description VARCHAR(255) NOT NULL DEFAULT ' ', # AKA provenance. Who did this mapping?
- type INT UNSIGNED DEFAULT NULL REFERENCES MappingTypes (id),
- url VARCHAR(10000) DEFAULT NULL,
- ftp VARCHAR(10000) DEFAULT NULL,
- lastDownloaded TIMESTAMP DEFAULT CURRENT_TIMESTAMP,
- quality INT UNSIGNED NOT NULL DEFAULT 1,
- PRIMARY KEY (id))
- ENGINE = MyISAM;

# What is a Mapping?

Mapping sources

- Hugo
- Geo
- InParanoid
- MultiInparanoid
- OrthoMCL
- Rio

# How do you use the Grinder?



## Getting translations from the Grinder

Species: Human

Key Space: HGNC Approved Symbol

Species: Mouse

Key Space: MGD IDs

Keys to translate. One per line

Results. Blank line means no mapping Multiple results separated by ',' results with ',' surrounded by " "

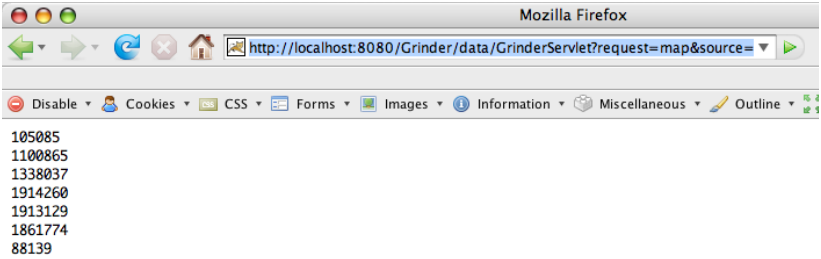
```
RBL2
RBM4
RBM6
RBM7
RBM8A
RBMS1
BCL2L1
```

```
105085
1100865
1338037
1914260
1913129
1861774
88139
```

# How do you use the Grinder? The web page

1. Choose the Starting Species
2. Choose the Starting Keyspace
3. Choose the Ending Species
4. Choose the Ending Keyspace
5. Enter keys to map
6. Hit “Do It”

# How do you use the Grinder?



```
105085
1100865
1338037
1914260
1913129
1861774
88139
```

# How do you use the Grinder? The Servlet

How to talk directly to the servlet
http://disco.cse.ucsc.edu:8089/Grinder/data/GrinderServlet?
request=map&source=HGNC%20Approved%20Symbol
&target=MGD%20IDs
&ids=RBL2,RBM4,RBM6,RBM7,RBM8A,RBMS1,BCL2L1

- Servlet Address: <Grinder Address>/data/GrinderServlet
- Query marker: '?'
- "request="
  - species: Get a list of all available species
  - keyspaces: Get a list of all keyspaces for a species, or all species
    Optional param: "species=" to restrict to one species
  - map: Map ids from one keyspace to another
    Parameters: "source=", "target=" : Source and target keyspaces
    "ids=" : Comma separated list of ids
- Result : Text file, one id per line

# Adding Information to the Grinder

Control Files

- XML Format
- Lets you add Keyspaces, Mappings, or Both
- Can add to existing ones, and / or create new ones
- One Data File per Control File
- Can define specific one way or bidirectional Mappings
- Can also define N-way bidirectional Mappings
- Format defined in the file ControlFile.xsd
- Requires User ID and Password to the Database

# What are GeneSets?

- A collection of items from one or more KeySpaces.
- The collection can be, but doesn't have to be, ordered.
- The items can have, but don't have to have, values.
- Can be part of a SetFamily.
- Can be loaded into the DB via a data file and an XML description file for the data.
- Characteristics:
  - Name
  - Description
  - Family

# What are GeneSets?

- CREATE TABLE GeneSets (
- id INT UNSIGNED NOT NULL AUTO_INCREMENT DEFAULT NULL,
- family INT UNSIGNED NOT NULL REFERENCES GeneSetFamilies (id),
- name VARCHAR(1000) NOT NULL UNIQUE,
- description VARCHAR(255) NOT NULL DEFAULT ' ',
- installTime TIMESTAMP DEFAULT CURRENT_TIMESTAMP, (User Sets: 2 weeks w/o use gets killed)
- PRIMARY KEY (id))
- ENGINE = MyISAM;

# What are GeneSets?

- CREATE TABLE GeneSetLinks (
- family INT UNSIGNED NOT NULL REFERENCES GeneSetFamilies (id),
- setID INT UNSIGNED NOT NULL REFERENCES GeneSets (id),
- geneKS INT UNSIGNED NOT NULL REFERENCES KeySpaces (id),
- geneID INT UNSIGNED NOT NULL,
- theOrder INT NOT NULL,
- value DOUBLE PRECISION DEFAULT NULL,
- INDEX fgg (family, geneKS, geneID),
- INDEX sgg (setID, geneKS, geneID),
- INDEX ki (geneKS, geneID),
- UNIQUE INDEX fsgg (family, setID, geneKS, geneID))
- ENGINE = MyISAM;

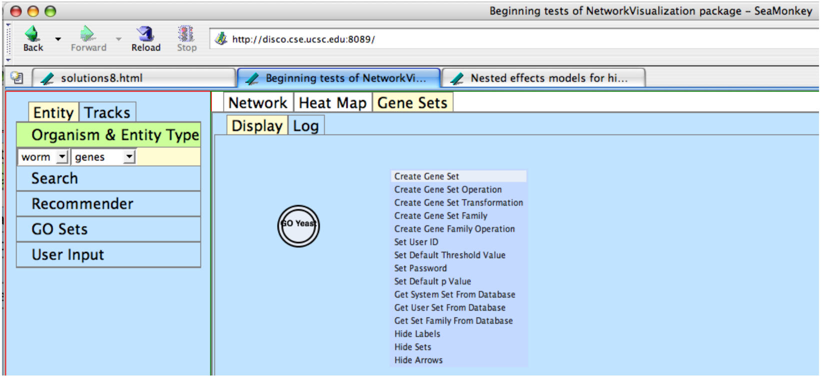
# What are GeneSetFamilies?

- A collection of GeneSets.
- System Sets: GO (Yeast, what other species?), KEGG, Others?
- User Sets: Whatever you want to add. Has your user name attached to it.
- Characteristics:
  - Name
  - Description
  - Source: URL and / or FTP location for origin of the data
  - Kill Date

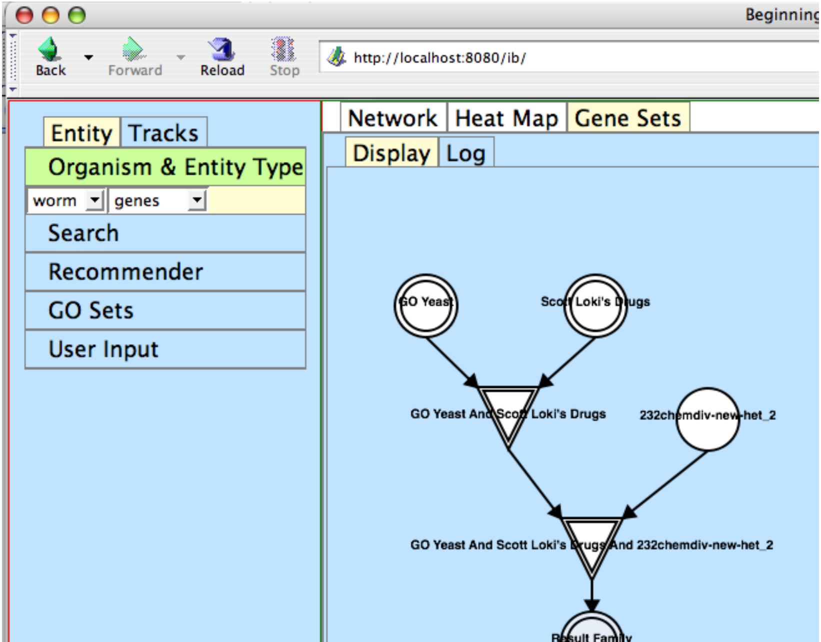
# What are GeneSetFamilies?

- CREATE TABLE GeneSetFamilies (
- id INT UNSIGNED NOT NULL AUTO_INCREMENT DEFAULT NULL,
- name VARCHAR(255) NOT NULL UNIQUE,
- description VARCHAR(255) NOT NULL DEFAULT ’ ’,
- url VARCHAR(10000) DEFAULT NULL,
- ftp VARCHAR(10000) DEFAULT NULL,
- killTime TIMESTAMP DEFAULT 0,
- PRIMARY KEY (id))
- ENGINE = MyISAM;

# Gene Sets Commands


Beginning tests of NetworkVisualization package – SeaMonkey
Back
Forward
Reload
Stop
http://disco.cse.ucsc.edu:8089/
solutions8.html
Beginning tests of NetworkVi...
Nested effects models for hi...
Entity
Tracks
Organism & Entity Type
worm
genes
Search
Recommender
GO Sets
User Input
Network
Heat Map
Gene Sets
Display
Log
GO Yeast
Create Gene Set
Create Gene Set Operation
Create Gene Set Transformation
Create Gene Set Family
Create Gene Family Operation
Set User ID
Set Default Threshold Value
Set Password
Set Default p Value
Get System Set From Database
Get User Set From Database
Get Set Family From Database
Hide Labels
Hide Sets
Hide Arrows

# Gene Sets in Action



# Gene Sets

- Gene Sets allow us to represent organizations of genetic information, such as GO categories.
- Circles are Gene Sets put together by the User.
- Triangles are Binary Set operations, and the results of those operations.
- Doubled figures involve families of sets, rather than individual sets.
- Will be able to group arbitrary Sets into a Family, or show the top 'n' Sets from a Family.
- Will be able to Load and Save layouts to / from your file system, and / or the DB.

# Where to go for more

The complete documentation for the Grinder is online at the Grinder Wiki page: http://wiki.soe.ucsc.edu/bin/view/SysBio/Grinder

# Points of this talk

- The Grinder is a one stop location for all your mapping needs
- It's easy to use
- It's easy for authorized users to add data to it
- **You** should use the Grinder
- Gene Set manipulation is also easy to use. Build a pipeline once, use it forever more.