Module 2: BioMart – datamining databases.

Aims

In this module, users will be introduced to the Biomart search interface, providing access to most of the main genomic and phenotype databases. Performing both simple and advanced queries, users will learn how to both refine their search parameters and define the result output to extract the information they require.

Introduction

The BioMart project (www.biomart.org) provides free software and data services to the international scientific community in order to foster scientific collaboration and facilitate the scientific discovery process. The Biomart project can be separated into two components:

BioMart Software

BioMart is a freely available, open source, federated database system that provides unified access to disparate, geographically distributed data sources. It is designed to be data agnostic and platform independent, such that existing databases can easily be incorporated into the BioMart framework.

There are currently two releases of Biomart available, an earlier version 0.7 and the current version 0.8. Differences between the two versions are largely related to the user interface, with the underlying query structure remaining the same.

Version 0.7 – this remains the most common implementation of the Biomart software and is used by most of the main databases, including Ensembl, MGI and Europhenome.

Version 0.8 – this is default version of the software used by the Biomart.org central portal.
BioMart Services

A large number of servers that provide access to a wide range of research databases have been set up by the BioMart community. Using BioMart's unique data federation technology, Data Portals were established to provide a convenient single point of access to all of these data, which is distributed worldwide.

In addition to the central portal, every database using Biomart provides its own interface, allowing independent access via a resource's own website. Example links are shown below;
Basic Query Structure

All mart searches adhere to a basic structure that allows the user to build search queries in a stepwise manner, without requiring knowledge of the database schema. At each stage, the available data is displayed in menus, allowing the user to select different options.

An important feature of biomart is the ability to import and export data files. This ability to import data allows the user to define a search list, either using data from other databases or external sources, such as publications. Search results can be exported in different formats, including text files, which can be opened in spreadsheets. These functions will be demonstrated in the worked examples that are to follow.
Biomart.org – a central portal for database access

Navigating to the Biomart.org homepage, we are presented with three portals, a general biological database portal, together with more specific Cancer and Mouse resource portals. Clicking on the Bio portal link, we are taken to the central Biomart portal, shown below;

The central portal provides details about associated databases (currently 41), simple search tools with preformatted parameters and links to specific databases for more advanced queries.
Simple Queries

Using the simple search tools, we can retrieve details for the Apoe gene. Go to the ‘Tools’ panel and click on Ensembl, under the ‘Gene retrieval’ tab (shown below).

Since Ensembl has gene information for many different genomes, the first step will be to select the dataset (1) by clicking on Mus musculus. Next we need to restrict the search parameters (2), so that only information about the Apoe gene is returned. This is done by specifying the Ensembl ID for the mouse Apoe gene, which is ENSMUSG0000002985.

Note. Hold shift to select multiple menu options.
Now that we have selected the dataset and refined our search parameters we can click ‘Go’ to get the search results.

As this is a simple query, the search result contains only basic information, such as the genomic location and the number of variants for the Apoe gene. To save this information, click on ‘Download data’ to save a text file.

Advanced query

To find more information about the different Apoe transcripts, we can create a more advanced query, specifying both the search criteria and the results output. Return to the central biomart portal and using the ‘Database search’ panel, select Ensembl from the ‘Genome’ menu located on the ‘Search by type’ tab.
Although we are using the same Ensembl database as before, we now have a different interface with many more options available. The three phases of the query build (Datasets>Filters>Output) are shown at the top of the page, with a summary of the selected criteria shown on the right.

As before we need to select the Mus musculus dataset to start the query process, once selected click ‘Next’.
As with a simple query, the ‘Filters’ step allows us to refine the search parameters. In the Gene section, add the Ensembl_ID for Apoe (ENSMUSG00000002985) to the ID list limit field, which should also appear under the query summary attributes.

Click ‘Next’ to proceed to the ‘Output’ step, which unlike a simple query allows us to select the information that is returned by the search query.
When selecting the ‘Output’ criteria, it’s important to note that the information fields will be displayed in the order that they are selected. Under the ‘Features’ tab, select the ‘Gene’ attributes in the order shown below:

1. Associated Gene Name
2. Ensembl Gene ID
3. Chromosome Name
4. Gene Start (bp)
5. Gene End (bp)
6. Gene Biotype
7. Ensembl Transcript ID
8. Transcript Start (bp)
9. Transcript End (bp)
10. Transcript Biotype
Click on ‘Results’

Unlike the earlier simple query, we now have information about each transcript variant for the Apoe gene. Note how the gene start and end positions relate to the transcript start and end positions. As before the results can be saved as a text file, by clicking on ‘Download Data’.

It is also easy to change the search criteria or results output for any biomart query, just click the ‘Back’ button.

External Database Biomarts

Biomarts for each of the associated databases can also be accessed individually via each resources own website. Currently, most of the individual database biomarts use version 0.7, which has a slightly different user interface. The query building process however, is almost identical to the advanced query method above.
Ensembl Biomart (www.ensembl.org)

Navigate to Ensembl and click on the BioMart link at the top of the page. As before, the first step is to select the database and dataset (Ensembl Genes 67; Mus musculus).

Step 2. Filter search parameters.
Step 3. Select results output.

Step 4. Results

The results should be identical to those returned using biomart.org.
Importing Data

Databases often use unique identifiers, for example MGI accession IDs or Ensembl Gene IDs, which can be used to extract related information from multiple sources.

The unique identifier type can be selected using the menu options, the identifiers can then be pasted into the text box or uploaded from a text file (useful for large gene lists).

Summary

Biomarts make it easy for users to extract information from databases without requiring any knowledge of the database structure or SQL scripts. Both the search parameters and the results output can easily be adjusted by returning to the relevant step, making it easy to refine a search query. Data can also be uploaded to form part of the search parameters, which is useful for gene lists extracted from other sources, such as publications or other databases.