BIOINFORMATICS 1:

BIOLOGICAL DATABASES AND PROGRAMMATIC ACCESS TUTORIAL WALKTHROUGH

Lysimachos Zografos, PhD

lzografos@parkure.co.uk

Friday 3rd November 2017

BROWSING AND QUERYING BIOLOGICAL DATABASES

1 PROTEIN SEQUENCES EXAMPLE: UNIPROT

Open a browser and navigate to Uniprot (http://www.uniprot.org/). Enter a query for the product of the DLG4 gene (PSD-95). The search is done in various entry fields (gene/protein name, synonyms, organism etc) so you can try searching for "DLG4", "SAP90", "Postsynaptic density protein 95" or any other synonym. Observe the list of results. There are a) entries from all organisms b) entries of other proteins. This is because somewhere is the entry the queried protein is mentioned somewhere in the general annotation text section.

Open the DLG4 human entry (*DLG4_HUMAN*), note how there is another human entry (*B9EGL1_HUMAN*) which however is not reviewed (by a human data curator). Go through the entry and see the type of information displayed. Once you're done open it in text and xml format in two new tabs (use the buttons on the top right corner of the entry page). These are the (older and newer) machine readable formats.

A note on accession numbers and IDs: UniProt indexes entries based on accessions (e.g. P78352). Accessions are tracked through releases and are more or less stable. One the other hand IDs (e.g. DLG4_HUMAN) are more readable but more prone to change.

Databases (specially central data repositories that function as service providers) tend to exchange data and interlink a lot. Go to the cross-references section and explore. Try to answer questions such as:

- What is DLG4 human gene Ensembl accession number?
- Is there a solved 3D structure for it?

- Can we find information about the expression patterns of DLG4?

Then go to the Sequence annotation (Features) and Sequences sections and explore:

- Why are there >1 sequences for the same gene?
- How many PDZ domains are there?
- How many beta-strands are there in the folded protein?
- Is this a transmebrane protein?

The Gene Ontology is a controlled vocabulary collection that maps each gene to specific biological processes, molecular functions and cellular components. It's a very useful way to describe the products of genetic entities in a rich yet controlled (and machine readable) manner. Look up P78352 (that is the UniProt accession of DLG4). What biological processes is it associated with? Explore the term "Learning" (GO:0007612).

2 Nucleic acid sequences example: Ensembl

Open a *new tab or window* and navigate to Ensembl (http://www.ensembl.org/index.html). Look up the human DLG4. You can use its accession number from task 1 for the query (or just follow the link from UniProt). Follow the transcripts and protein entry links.

- How many alternative splice forms of DLG4 are there? How many are expressed as proteins?
- Is there a DLG4 orthologue in Zebrafish?

3 Common search interface

BioMart is a deployable database schema and API that is used in some biological databases. The largest database that uses it is Ensembl (EnsMart). With BioMart one can ask (not very) complicated questions involving large and heterogeneous datasets, e.g. "what are all the gene accessions for all known mouse genes?", "what are all the gene names and transcript accessions for all known mouse genes whose protein has protein domain X?" or "which are my gene list's orthologs across sequenced species?".

Pocklington et. al (Mol Syst Biol., 2006) identified and analysed the protein complex found in the synapse known as the postsynaptic density (PSD). This complex, called the NRC/MASC underlies basic neuronal plasticity which in turn underlies basic neuronal network computational properties, i.e. basic cognition / learning. A big part of the analysis of this complex and it's properties was based on annotation of the individual proteins involved. The process of going from sequence (or a database accession number) to annotating the properties of a protein can be assisted and automated with such search interfaces.

3.1 Biomart - find all NRC/MASC human UniProt accessions

Navigate to EnsMart (http://www.ensembl.org/biomart/martview). Select the Ensembl database and the human dataset. Download the NRC/MASC data file, open it and copy the column with the Ensembl ids. Paste the list in the Filters section, under GENE. Go to the attributes section and select the UniProt accession under "EXTERNAL". View the results.

3.2 Biomart - find all NRC/MASC orthologs in fly and yeast.

One would expect that gene orthologs would exist in organisms with a nervous system. Go back to the Attributes section, select Homologs from the top menu and add the Yeast Ensembl Id and C. elegans Ensembl Gene ID attributes. Retrieve the results. What type of NRC/MASC proteins are found all the way back to yeast?

4 Data warehouse examples (explore at home)

- Genes2Cognition: http://www.genes2cognition.org
- Flybase: http://flybase.org
- GeneCards: http://www.genecards.org
- InterMine (not a warehouse database but a system to build warehouse databases): http://intermine.org e.g. MouseMine http://www.mousemine.org/mousemine/begin.do

5 Programmatic access

EMERGENCY LINK (ONLY IN CASE THERE IS A PROBLEM WITH THE LINK ONLINE):

https://www.dropbox.com/s/tvp7ktzekqifr8r/WebservicesTeaching.tgz

A key component of every bioinformatic workflow is programmatic access and manipulation of the data. The traditional approach to this would be to download a flat file containing each database of interest and writing scripts to parse these and isolate the interesting data. Although this approach has its pros, some alternatives have been starting to prevail. One of these is the use of webservice APIs. These are servers which "listen" to specific queries and requests in a programming language independent format (e.g. XML) usually via the visit to a specific URL. Although the technical details are beyond the scope of this tutorial we will look at some simple examples.

5.1 Querying EnsMart

Please see examples under WebservicesTeaching/EnsMart