TUTORIAL Databases

Exercise 1.

Search for the prion gene on the Genbank home page:

http://www.ncbi.nlm.nih.gov/gene/

"prion protein"

"prion protein" AND (“Homo sapiens”[orgn])

Choose the best hit for your search criteria.
Find the answers on the result page:

- What is the name of the gene?
- On which chromosome it is located?
- What other names this gene have?

- What is the function of the gene?
- What kind of phenotype it is associated with ?

- What is the status of the associated Refseq enrty?
- How many splice variants does it have?

- How many SNP variations belong to this gene?
(Go to the SNP link on the right hand side)

- Select a mutation that is likely to be a pathogenic germline mutation.
Does it effect a coding regions? What type of amino acid change does the mutation lead to?

Exercise 2.

Go to the Uniprot page (http://www.uniprot.org/) and type:
Myosin light chain kinase
in the search box on the top of the page. Make sure you selected the "Protein Knowledgebase UniProtKB" option. Hit the search button.

How many hits did you get?

Not place the expression into quotation marks.

How many hits did you get this way?

Now narrow the search to human proteins. From the "Fields" box drop-down menu, select "Organism [OS]". In the "Term" box, type "Human".

Note: UniProt/SwissProt entries are marked with golden, a UniProt/TrEMBL entries are grey.

How many hits did you get this way? How many of them are reviewed?

Exercise 3.

More into a specific UniProt/SwissProt Entry : General information

Click on the link:
http://www.uniprot.org/uniprot/Q15746
This will show you an uniprot entry.

Is this from Swissprot or Trembl database? What does it mean?

You can get to different section by clicking on the different menu items on the left.

What is the accession number and identifier for this proteins?

Search for the "Entry Information" section.

How many different entries are merged into this entry ?
(Count the primary and secondary accessions codes)

Go back to the "Names and Taxonomy" part.

What are names this protein have?

Some uniprot entries are only predicted while for others there is direct evidence that this protein indeed expressed in cells.

Is there direct evidence for the existence of this protein?

Explore the Feature viewer!

Can you find a disease mutation in this protein? What other features does this position overlap with?

Exercise 4.

Go to the "Function" section.

What is the function of this protein?

Based on the keywords, what type of ligands bind to this protein?

Based on the GO annotation, what type of Biological processes is it involved in?
Follow the link to the "Complete GO annotation".

What type of process is inferred from mutation phenotype? What type of experiment is this based on?
What kind of generic slim term it this processed map to?

Exercise 5.

Go back to the Uniprot page. What kind of disease is linked to this protein?
What kind of mutation and effect causing this disease?
Find the general phenotypic description if the disease. What other genes are linked with this disease?

Exercise 6.

You can find many cross-links to other databases. Find to the link to the ENSEMBL database.

The ENSG, ENST és ENSP identifiers correspond to genes, transcript and protein entries, respectively.
What is the ENSEMBL gene id in for this protein?

Follow the gene link. Which chromosome is this gene located on?

How many transcripts this gene has? How many of them are actually transcribed into protein?
Which is the shortest protein?

Choose the principal isoform. How many exons is this protein composed of?
(! There is a trick here!)

Go to the Variation table link.

How many variations have been described for this proteins? How many of them are pathogenic?

Exercise 7.

Uniprot

- Website: http://www.uniprot.org
- REST help: FAQ on Programmatic access.

The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data consisting of the UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters (UniRef), and the UniProt Archive (UniParc) and more.

To access a full Uniprot entry by its unique id (as HTML), you would type http://www.uniprot.org/uniprot/P12931 The same data is available in different formats:

- http://www.uniprot.org/uniprot/P12931.txt Text only
- http://www.uniprot.org/uniprot/P12931.xml XML format
- http://www.uniprot.org/uniprot/P12931.rdf Special form of XML (Resource Description Framework)
- http://www.uniprot.org/uniprot/P12931.fasta only the sequence in FASTA format
- http://www.uniprot.org/uniprot/P12931.gff only the protein features in GFF (General Feature Format)

Exercise:

1. Try these links to get a feeling for the different formats.
2. Substitute the id (P12931) with the name or id of a protein of interest ('ABL1_HUMAN', 'Epsin'). Which of these works, which doesn't?

Performing queries

While the FASTA and GFF format only show a limited set of data of any protein dataset, the HTML and TEXT representations demonstrate the large amount of data that is annotated at Uniprot. Chances are not everybody needs all the data all the time. Therefore Uniprot allows to specify which fields to query and to return only selected fields. Use the Uniprot help page and in particular pay attention to the list of possible query fields for the following exercises!

In order to get yourself acquainted with the Uniprot syntax, please use the button 'Advanced Search' at the uniprot homepage to construct your uniprot queries (yielding HTML results).

1. Search for all proteins with the name `GRB2` which have the status `reviewed`
2. Next, copy this: `name:GRB2 AND reviewed:yes` from the search field into the url: `http://www.uniprot.org/uniprot/?query=` so that your URL looks like this: `http://www.uniprot.org/uniprot/?query=name:GRB2 AND reviewed:yes`
3. Add the following to your query to turn the output format from HTML to TAB:

`&format=tab`: `http://www.uniprot.org/uniprot/?query=name:GRB2 AND reviewed:yes&format=tab`

4. Repeat this query with protein name `ABL1`. How many resulting lines do you get?

   Different formats

   The uniprot output format can be one of:

   - html
   - tab
   - xls
   - fasta
   - gff
   - txt
   - xml
   - rdf
   - list
   - rss

5. Try the previous exercises (eg. searching for GRB2 or ABL1) with at least three different output formats.
6. (OPTIONAL) Instead of GRB2 or ABL1, try your favorite protein of interest.

Limiting results

While testing your queries, it's a good idea to limit the number of results (in order not to stress the server or block your browser) `&limit=10`.

Example:

Retrieving the first ten human sequences as fasta:

http://www.uniprot.org/uniprot/?query=reviewed:yes+AND+organism:9606&limit=10&format=fasta

You can 'walk' along a set of results by using limit and offset, eg to retrieve the next 10 results of previous query:

http://www.uniprot.org/uniprot/?query=reviewed:yes+AND+organism:9606&limit=10&offset=10&format=fasta

1. Try to get all proteins which have 'ABL2' in their name in tab format.
2. How many are there?
3. Try to 'walk' along these results by using a limit of 3 and an appropriate offset until you've seen all results.

Selecting columns

1. Read the FAQ (table in section ‘Retrieving entries via queries’) and find out which columns can be selected as returnvalues.
2. Using the tab format, retrieve all proteins named ‘Proepiregulin’ and select ‘id’, ‘entry name’, and ‘genes’ as output columns.
3. Using the same format and protein name, try at least three different column types as return (first indivdually, then all three at once)
4. Get all ‘ABL1’ proteins and use ‘entry name’ and ‘interactor’ as output.
5. Which of these proteins have interactors annotated?
6. Use the additional column ‘taxon’

Congratulations, you entered the world of programming!