TUTORIAL Databases

Exercise 1.
How can we find basic information for a given gene at the NCBI web site?

Go to the NCBI Website and from the Entrez interface and select a database called “Gene”.
Each record within Gene compiles information from numerous sources on a single gene.

To start with, search for the *prion* gene :

Type:

prion
or
prion AND (“Homo sapiens”[orgn])

into the search box.

Try the website:
https://www.ncbi.nlm.nih.gov/gene/

You can see several tips on how to specify you search terms.

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 does this gene have?

- What is the short summary of what is known about this gene (Check Summary)?
- Where is this gene mostly expressed?

Go to Reference Sequences. (If you cannot find it on the page, just search for it on the web page using CTRL F to open the search box)

- How many splice variants does it have?
- What is the special property of the first Refseq entry?
- What is the status of the associated Refseq enrty?
- What is the accession code for the this transcript?
- What is the accession code for the protein?

Follow the link to Refseq transcript entry (accession code usually starts with NM_ ...)
This page shows you the genbank format.

- How many basepairs does this transcript have?

Look at what type of features are annotated.

By clicking on the the features, the corresponding region is highlighted in the sequence.

Download the sequence corresponding to the CDS regions in FASTA format.
- Paste the sequence here:

Go back to the GENE page.
Explore variations associated with this gene.

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

Follow ClinVar link. Select a mutation that is likely to be a
pathogenic
germline
missense mutation.

- How many mutations did you find?
- Any of these affect coding regions? What type of amino acid change does the mutation lead to? What diseases are these associated with? (select one)

Exercise 2.

The main database for proteins is the UniprotKB. First you can familiarize youself with search options.

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?

Now place the expression into quotation marks.

-How many hits did you get this way?

Now narrow the search to human proteins. Select Advanced option, add "Term" box "Organism [OS]". , 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?

- How many have matching protein name?

Can you find an entry that does have only 1 corresponding 3D structure?
(Hint: You can customize columns, and select Structure and 3D within it.) Then you will be able to see the number of 3D structures corresponding to a given entry.

Exercise 3.

More into a specific UniProt/SwissProt Entry : General information

Click on the link:
https://www.uniprot.org/uniprot/O75874
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 other names does 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?

- What is the function of this protein? Describe the catalytic activity.
- What type of ligands does it have?
- Where is the protein located?

Explore the Feature viewer!

Can you find disease mutations in this protein?
Click on Variants, and select Disease and Uniprot reviewed mutations.
- Which position do they correspond to? What other features do they position overlap?
- In what type disease do they occur?

Go back to the main page.

- How does this mutation cause the disease?

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?