Swiss Institute of Bioinformatics

Inside protein databases, March 9th, 2017
Practical 1 (morning session)

From nucleic acid sequence databases to protein sequence databases

Look at this entry (http://www.ncbi.nlm.nih.gov/nuccore/X02158)

- Which server? Which database? Which accession number?
- What is the main type of data?
- Look at the CoDing Sequence (CDS): click on 'CDS': how many exons? Can you find any information on the accuracy of the CDS?
- Follow the link to UniProtKB
- From UniProtKB: go back to the GenBank entry (*Cross-reference section*)

Look at this entry (http://www.ncbi.nlm.nih.gov/nuccore/X02157)

- Which server? Which database? Which accession number?
- What is the main type of data?
- Look at the CDS: how many exons?
- Follow the link to/protein_id="CAA26095.1": in which database are you?

Look at this entry (http://www.ncbi.nlm.nih.gov/nuccore/NM_000799.2)

- Which server? Which database? Which accession number?
- What is the main type of data?
- Look at the ##Evidence-Data-START##: How is this entry related to the first 2 entries from this exercise?
- Follow the link to the protein sequence entry (NP_); from there, follow the link to the corresponding UniProtKB entry
- From the UniProtKB entry (NCBI view), go to the original view of this entry on the UniProt website, and from there, return to the RefSeq entry (*Cross-reference section*).

Discover the content of a UniProtKB entry and the source(s) of annotation

Look at this UniProtKB entry (P04150)

- 'Header section'

What is the status of this entry: reviewed by a biocurator, or unreviewed?
What is the evidence for the existence of the protein? What is its annotation score? Use the context-sensitive help links for background information on these concepts.

- 'Names &Taxonomy' section

What are the name(s) of the gene and the name(s) of the protein?
NB: the source of protein and gene names is not always available: see another example entry where the source is available .
Does this entry belong to a proteome, and if so, what is the identifier of that proteome?

- 'Sequences' section

How many different protein sequences (isoforms) are available for this gene?

For isoform alpha, follow the link to the UniParc database. How many RefSeq entries are available for this isoform? Are the sequences identical?

UniParc is a protein sequence archive, used to keep track of sequences and their identifiers.
A UniParc entry lists all the entries, active or inactive, in different databases, which contain exactly the same protein sequence.

- 'Function' section

What is the function of the protein?
Where does the information come from?
Compare the keywords and the Gene Ontology terms relative to the protein's function
Note that the 3 GO ontologies Molecular Function, Biological Process and Cellular Component, as well as UniProt Keywords are dispatched to the relevant sections of the entry, not all GO terms can be found under "Function").

What is the source of the Gene Ontology and keyword annotations?
FAQ: What are the differences between UniProtKB keywords and the GO terms?

- 'PTM/Processing' section

How many phosphorylated sites?
How many sites have been **experimentally proven** to be phosphorylated?

- Look at the different tracks of the **Feature viewer** (including 'proteomics' and 'variants')

- 'Structure' section (3D structure databases):

How many 3D structures are available for this protein? Do they 'cover' the complete protein sequence?

- Look at the **'Complete history'** of the entry.

Has the protein sequence been updated since its integration in the database?

- Look at the same UniProt entry in the 'txt' format (via the "Format" button).

- Look at the same UniProtKB entry but displayed on the NCBI server. Can you go back to the UniProt web site?

Look at this entry (NP_000167.1)

Which database? What is the main type of data?

Look at this entry (NX_P04150)

Which database? What is the main type of data?

Look at this UniProtKB entry (X7PSG6)

- 'Header section'

What is the status of this entry: reviewed by a biocurator or unreviewed?
What is the annotation score?
What is the evidence for the existence of the protein?

- 'Names&Taxonomy' section

What are the name(s) of the gene and the name(s) of the protein?
What is the source of these name assignments?

- 'Function' section

What is known about the function of the protein? What is the source of these annotations?
Look at some rules used for automatic annotation: UniRule (i.e. RuleBase or HAMAP), and SAAS.

- Compare the keywords and the Gene Ontology terms relative to the protein's function

What is the source of the Gene Ontology and keyword annotations?

- 'Cross references' section

Look (in GenBank) for the data available on the DNA sequence.
Can you find any information on the accuracy of the CoDing Sequence (CDS)?

- 'Sequence' section: follow the link to the UniParc database.

How many entries that have the same protein sequences?

Discover the UniProt Search tool

(1) Simple query

(a) Look for the gene PRSS1 in UniProtKB
Restrict 'PRSS1' to exact gene name.

(b) Customize your results
Configure your column layout by adding and removing columns, until you have columns for 'Protein names', 'Reviewed/Unreviewed', 'Organism', '3D', 'Gene Ontology (Biological process / Molecular function)', 'Function' (Function CC).

(c) Compare the annotations of reviewed and unreviewed entries.
What is the source of the GO annotation in most unreviewed entries?

(2) Advanced query

(a) Look for yeast (Saccharomyces cerevisiae) proteins located in the nucleus by using the search box at the top.
Refine your 'full text' query according to the suggestions and filters provided by the query tool.
Restrict to the 'Reference proteome' (Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast))

How many of them have been 'experimentally proven' to be phosphorylated on a threonine ?
(Advanced search: PTM/Processing - modified residue - phosphothreonine)

Note: retrofitting the source attribution ('evidence') to older reviewed entries is an ongoing process.
For additional info look at 'Why do not all UniProtKB/Swiss-Prot annotations have evidence?'
In the advanced query tool, you have the choice between 'Any experimental assertion' and 'Manual assertion: Experimental':
- 'Any experimental assertion': used for all experimental data which were historically considered as 'experimental', including the data which we were not able to automatically attribute to a publication.
- 'Manual assertion: Experimental': used for experimental data which are linked with a publication (link established by an expert curator).
-> We recommend to use 'Any experimental assertion'.

Add the corresponding column (PTM/processing - modified residue)
Do you think that you get a comprehensive set? Why?

(b) What is the query corresponding to this URL:
http://www.uniprot.org/uniprot/?query=annotation%3A%28type%3Atransmem+count%3A[5+TO+*]%29

(3) Frequency of N-glycosylation

N-linked glycosylation is the most important form of post-translational modification for proteins synthesized and folded in the Endoplasmic Reticulum.

Imagine you have been working on a mouse protein family where 20% of all members are N-glycosylated. Try to find out whether your protein family is glycosylated significantly more frequently than other proteins.

Hints:
Look at the mouse proteome. In this proteome:

- What is the percentage of N-glycosylated entries?
- What is the percentage of reviewed entries?
- What is the percentage of N-glycosylated entries in reviewed and unreviewed entries, respectively?

Relevant UniProt help pages:
- query syntax
- Glycosylation

Discover the UniProt BLAST tool

Look for plant protein sequences similar to human hemoglobin (HBB):
BLAST the human hemoglobin HBB sequence against 'Plant' sequences in UniProtKB.

- *Customize your 'BLAST results'*
Add columns for protein names, gene names, function, keywords, gene ontology (some are already there by default).

- Open the first binary alignment
Look at the conservation of the iron binding sites ('Metal binding' in the “Highlight” options on the left).

Discover the UniProt ID mapping tool

Which database do these identifiers correspond to?
NP_001018084 NP_001018085 NP_001018086 NP_001191191 NP_001191192 NP_001191193 XP_005268476 XP_005268477 XP_016864886 XP_016864887

Find the corresponding UniProtKB entries, using UniProt's ID mapping tool.
Do a multiple alignment of the UniProtKB entries.

(2) Using PICR : find the accession numbers corresponding to this sequence in UniProtKB, RefSeq and UniParc:

```
MEEPQAGDAARFSCPPNFTAKPPASESPRFSLEALTGPDTELWLIQAPADFAPECFNGRHVPLSGSQIVK
GKLAGKRHRYRVLSSCPQAGEATLLAPSTEAGGGLTCASAPQGTLRILEGPQQSLSGSPLQPIPASPPPQ
IPPGLRPRFCAFGGNPPVTGPRSALAPNLLTSGKKKKEMQVTEAPVTQEAVNGHGALEVDMALGSPEMDV
RKKKKKKKNQQLKEPEAAGPVGTEPTVETLEPLGVLFPSTTKKRKKPKGKETFEPEDKTVKQEQINTEPLE
DTVLSPTKKRKRQKGTEGMEPEEGVTVESQPQVKVEPLEEAIPLPPTKKRKKEKGQMAMMEPGTEAMEPV
EPEMKPLESPGGTMAPQQPEGAKPQAQAALAAPKKKTKKEKQQDATVEPETEVVGPELPDDLEPQAAPTS
TKKKKKKKERGHTVTEPIQPLEPELPGEGQPEARATPGSTKKRKKQSQESRMPETVPQEEMPGPPLNSES GEEAPTGRDKKRKQQQQQPV
```

Proteomes

Human proteome

- Look for human proteins in UniProtKB: how many entries?

- Look for human proteins in UniProtKB/Swiss-Prot: how many entries?

- Look for human proteins in UniProtKB/TrEMBL: how many entries?

- Look for the human proteome in UniProtKB: how many entries?
  How many entries from the proteome are in UniProtKB/Swiss-Prot, i.e. are reviewed?

- Look for the human proteome set in neXtProt: how many entries?

- From the UniProt home page, follow the link to 'Proteomes' and...
  ...look for homo sapiens: how many entries?
  ...download the sequences in fasta format, including the sequences of additional isoforms.

Further information:
'What are proteomes?'
'What are reference proteomes?'
'What is UniProt's human proteome?'
'How to retrieve sets of protein sequences?'
'Why do we keep dubious sequences in UniProtKB? How to discard them from a protein set?'

Practical 2 (afternoon session)

UniProtKB, neXtProt, PANTHER and GO

The PANTHER (protein annotation through evolutionary relationship) classification system (http://www.pantherdb.org/) is a comprehensive system that combines gene function, ontology, pathways and statistical analysis tools that enable biologists to analyze large-scale, genome-wide data from sequencing, proteomics or gene expression experiments (PMID:23868073). PANTHER allows to access all Gene Ontology annotations--updated monthly from the Gene Ontology database--in addition to the annotations that have been inferred through evolutionary relationships (PMID:26578592).

1. SIMPLE QUERIES / IDmapping / Batch retrieval UniProtKB

Perform a GO enrichment analysis (using www.pantherdb.org) on the 20 most frequently mutated human cancer genes (according to the STRING database):

CIC FLT3 APC IDH1 PIK3CA DNMT3A ERBB3 NCOR1 TP53 PIK3R1 FBXW7 BRAF LPHN2 SF3B1 SMAD4 CTNNB1 NF1 NRAS PTEN CDKN2A

You can use the UniProt ID Mapping tool to get the UniProtKB accession numbers corresponding to these gene names, and then submit the list of identifiers to the PANTHER classification system at http://www.pantherdb.org. Select “Functional classification viewed in pie chart”. The default ontology is “Molecular Function”, but also have a look at the Panther protein class.

See also: Can I convert gene symbols to UniProtKB identifiers? How can I map UniProtKB IDs or ACs to gene symbols?

2. SIMPLE QUERIES UniProtKB/neXtProt

Choose one of the 2 queries below:

- Proteasome human/mouse: KW-0647 (keyword)
- Cytokeleton human/mouse: SL-0090 (cv subcellular location)

Look at differences in number of entries between neXtProt, UniProtKB, proteome

Perform a GO enrichment on the proteome members from your selected query, using www.pantherdb.org

- Functional classification viewed in pie chart**.** Once you have your result, select "Protein class" from the "Select Ontology" dropdown.
- Statistical overrepresentation test (use the default parameters)

3. SPARQL QUERIES ('Federated query') - neXtProt

With neXtProt: Query proteins expressed in brain, but not in testis
GO TO: https://snorql.nextprot.org/

Find the corresponding query from the examples in the tutorial, and execute it.
NB: this type of query related to tissue expression is not yet possible with UniProtKB

Look at GO enrichment using www.pantherdb.org

- Functional classification viewed in pie chart
- Statistical overrepresentation test (use default parameters)

4. SPARQL QUERIES ('Federated query') - UniProtKB and neXtProt

With neXtProt: Find protein kinases that bind to drugs :

GO TO: https://snorql.nextprot.org/

ENTER QUERY:

```
PREFIX cco: <http://rdf.ebi.ac.uk/terms/chembl#>

select distinct ?entry ?gen where {
  SERVICE<http://www.ebi.ac.uk/rdf/services/chembl/sparql>{
         SELECT distinct ?protein WHERE {
  ?target cco:hasTargetComponent ?tarComp .
  ?target cco:taxonomy <http://identifiers.org/taxonomy/9606> .
  ?tarComp skos:exactMatch ?protein .
  ?activity a cco:Activity ; cco:hasMolecule ?drug ; cco:hasAssay ?assay .
  ?assay cco:hasTarget ?target .
                }
        }
  ?entry skos:exactMatch ?protein .
  ?entry :isoform / :keyword / :term cv:KW-0418. #kinase
  ?entry :gene / :name ?gen.
}
```

With neXtProt: Find ALL proteins that bind to drugs:
Modify the above query and remove the restriction to kinases.

With UniProtKB:

GO TO: http://sparql.uniprot.org/

Select the preferred gene name and disease annotation of all UniProtKB entries from mouse that are known to be involved in a disease.

5. Do a GO enrichment on the following dataset

We have to upload the file. The input file includes a column of numerical values for each gene/protein identifier (the corresponding expression coefficient). The file was created in 2013 (publication).

Upload the file at www.pantherdb.org

Look at the entries which are "unmapped ID": why are they unmapped ?

- Use the UniProt Retrieve/ID mapping tool

Look at GO enrichment using www.pantherdb.org

- Statistical enrichment test (use per default parameters)

Pascale Gaudet, Elisabeth Gasteiger, Marie-Claude Blatter
SIB Swiss Institute of Bioinformatics, March 2017