Practical 1 Protein sequence databases & annotation

From nucleic acid sequence databases to protein sequence databases

Look at this entry (https://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)
  - NB: notice that there are many cross-references to GenBank entries which have been used to 'construct' the UniProtKB/Swiss-Prot sequences (different isoforms)

Look at this entry (https://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 (https://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.

From this UniProtKB entry (Section Cross-References / Sequence databases)

- Can you find a link to the entries X02158, X02157 and NM_000799.2?

At NCBI Genome

Look for the HIV1 genome

- How many genes? How many proteins? Which INSDC accession number? Which RefSeq accession number for the complete genome?
- Look for the corresponding proteins in UniProtKB (query with the INSDC AC): how many proteins?
  - How many records in UniProtKB/Swiss-Prot? Add the column for RefSeq cross-references
  - Which entry is not ‘common’ to UniProtKB/Swiss-Prot and RefSeq? (click on ‘Protein’ in the previous Genome entry)
- Can you find the protein YP_009028572 in UniProtKB (perform a BLAST)? Is it identical?
NB: this predicted protein sequence is the translation of an Env antisense RNA (protein Asp). By analogy with other viruses (HTLV) some biologists suggest that this protein may also exist in HIV1. This ORF remains putative even if there are some publications on it.

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?
  What is the 'Taxonomic identifier' (TaxId)?
  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.

- 'Cross references' section (sequence databases)
  Look at the RefSeq cross references: How many?
  Why are not all isoforms cross-referenced to RefSeq?

- '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? Phosphorylation: where does the information come from? Look for the annotation derived from 'CDD'

**Look at this UniProtKB entry (F8VQZ7)**

- '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 a Coding Sequence (CDS)? Where does the protein sequence come from?

- 'Sequence' section: follow the link to the UniParc database.
  How many RefSeq entries that have the same protein sequences?
  Are there any UniProtKB entries that have exactly 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:
https://www.uniprot.org/uniprot/?query=annotation%3A%28type%3Atransmem+count%3A[5+TO+*]%29

(3) N-glycosylation: compare different datasets
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?
- Try to find out whether your protein family is glycosylated significantly more frequently than other proteins.

Relevant UniProt help pages:
- query syntax
- Glycosylation
Gene Ontology database: structure and validation of the data...

- On UniProt: look for pig and human insulin (query: "ins human" OR "ins pig")
- Compare the GO terms associated with pig and human insulin (add column 'Gene Ontology')
- Human insulin: Follow the link to QuickGO and have a look at the different GO evidence tags associated with the different GO terms
- Have a look at the hierarchy (tree view) of the GO term 'glucose metabolic process' (GO:0006006).

- On UniProt: have a look at the GO terms (category: cellular component) associated with histone H4 from *A.thaliana*.
- Look at the publication(s) associated with the GO term 'chloroplast'
- On UniProt: compare the GO terms associated with histone H4 from different plant species ('Viridiplantae').

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 (use Advanced BLAST).

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

- For the first matching UniProtKB/Swiss-Prot entry, open the pairwise 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

(1) 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. How many differences?

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

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?
- From the UniProt home page, follow the link to 'Proteomes' and...
  ...look for homo sapiens: how many proteomes?
Do the same query using the Taxid
  ...download the human proteome sequences in fasta format, including the sequences of additional isoforms.
Look for the Candida albicans proteome(s)

How many ‘non redundant’ proteome(s)?

Do the same query by using the taxonomy identifier (TaxId=237561)
How many ‘non redundant’ proteome(s)?

What is the Proteome ID of the Candida Albicans reference Proteome?
To which strain does it correspond?
How many proteins?
  How many Swiss-Prot entries?
  How many TrEMBL entries?

How many proteins have been proven to exist at the protein level? (Protein existence)
  % in Swiss-Prot, % in TrEMBL?

How many proteins with a phosphohistidine?
  % in Swiss-Prot, % in TrEMBL?
  Where does the information about the phosphohistidine sites come from?

How many proteins for Candida albicans in RefSeq? Why does this number differ from UniProtKB?

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 – Prosite and HAMAP

Go to this URL:
(1) You did a proteomics analysis in December 2007 without any match. You repeat the analysis in April 2008 and get entry with AC P0C6S9 as the best match. In 2014, you can no longer find the entry P0C6S9 anymore. What happened?

(2) QAGLTYAGPPPVGR is a unique human peptide identified by Peptide Atlas. Do a Blast against UniProtKB/Swiss-Prot. To which protein does it correspond? Look at the alignment Why is it not a perfect match with the UniProtKB/Swiss-Prot protein sequence? Look at the 'Highlight' annotation on the left

(3) Use “Peptide search”

DSCQGDSGPVPVCNGQLQG was identified as a human peptide. Could it be a typical 'false positive'?

ETMQFLNDRASYELKVQREL and SENARLVVQIDNAKLADDFRTKY were identified as human peptides. Could they be typical 'false positive'?

Glycosylation sites and proteomics

Look at the information available concerning the glycosylation site

https://prosite.expasy.org/PDOC00001

A publication (in Nature (PubMed:28959962)) gives the list of N-glycosylated peptides with the position of the N-glycosylation (HexNAc).

"We develop a novel quantitative approach to identify intact glycopeptides from comparative proteomic data sets, allowing us not only to infer complex glycan structures but also to directly map them to sites within the associated proteins at the proteome scale."

In Supplementary Table 6 dedicated to "N-glycoproteins" the authors attribute to the peptide "INTTAEKDPTNPRF" the following "N12(HexNAc(NL))" modification. and to the peptide "INTTAEKDPTNPFRPNIGVEK" the following "N18(HexNAc(NL))" modification.

Is that correct? What could have been done by the authors to avoid this?

Align the peptide with the corresponding UniProtKB entry:
- What is the position of the N-glycosylation in the protein?
- Where does the annotation come from?

- Could you get a set of proteins which are not glycosylated?

**User request**

A biologist has isolated a threonine phosphorylated human protein, by immunoaffinity. The monoclonal antibody recognises the following epitope: V-S-T-Q where the T is the phosphorylated threonine.
Could you help him to find a list of candidate proteins?
- Use ScanProsite against UniProtKB/Swiss-Prot (Option 2; select database = Swiss-Prot (without the splice variants); Exclude fragments; Maximum number of displayed matches: 1000) to retrieve the list of proteins containing the subsequence 'V-S-T-Q'
Click on 'Matched UniProtKB entries' at the bottom of the result page to get the corresponding UniProtKB entries.
- Perform a 'UniProt query' on our subset of proteins in order to select the correct candidate proteins (Homo sapiens; PTM/Processing, Modified residues, Phosphothreonine, Evidence: any manual assertion) (select 'Jobs' and then perform your query)

**Links to**
- [User manual](#)
- [FAQ](#)
- [Documentation/Help](#)