Practical 1

Protein sequence databases and sequence annotation

Swiss Institute of Bioinformatics

http://education.expasy.org/cours/SIB_UniProtKB_2018/

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/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 and NM_000799.2?

Discover the content of a UniProtKB entry, the differences between Swiss-Prot and TrEMBL 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? Have a look at the types of evidence that supports the existence of a 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?
How many in UniProtKB/Swiss-Prot? How many in UniProtKB/TrEMBL (computationaly mapped)?

- '**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?

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

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

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

Look at the DNA binding domain in the 3D structure *(Hint: click on the domain 418-493)*

- 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 UniProtKB entry (F1D8N4)

Compare this sequence protein with the canonical sequence P04159 (Align)

Back to the entry F1D8N4:

- **'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?

- **'Cross references' section**

Look (in GenBank) for the data available on the nucleic acid sequence.
Can you find a CoDing Sequence (CDS)? Where does the protein sequence come from?
How many RefSeq entries that have the same protein sequences?

Look at this UniProtKB entry (A0A0C5ESB7)

- **'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 ?
What is the source of these name assignments?

- **'Function' section**

What is the function of the protein?
Where does the information come from?

- **'Cross references' section**

Look at the GenBank entry: BGMY01000014
What are the inferences available for the CDS?

-**Feature viewer**

What are the evidences for the existence of the 2 active sites?

Look at this entry (NP_000167.1)

Which database? What is the main type of data?
Phosphorylation (Phosphothreonine): where does the information come from?
Look for the annotation derived from 'CDD'

Discover the UniProt Search tool

(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?

Proteomes

Mus musculus : How many proteome(s)?

What is the Proteome ID of the reference Proteome?
To which strain does it correspond?

- How many estimated gene number?
- How many estimated protein sequences?
- How many records describe more than one protein sequence?
  Hint: how many records with 'alternative products'

- ...download the mouse proteome sequences in fasta format, including the sequences of additional isoforms.

How many records are associated with the GO term ‘nucleus’

- % of records in UniProtKB/Swiss-Prot?
- % of records in UniProtKB/TrEMBL?
- Why this difference?

Further information:
'What are proteomes?'
'What are reference proteomes?'
'How to retrieve sets of protein sequences?'

@NCBI proteins: Mus musculus
How many protein sequences for Mus musculus @NCBI protein?
How many protein sequences in RefSeq?
Why this difference compared to UniProtKB?

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:
In the mouse 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

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 and RefSeq

Practical – BONUS

Proteomic results: true stories

(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 ?
Sequence section: look at the 'sequence conflict'

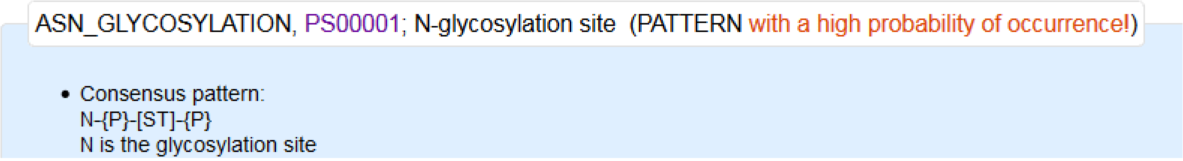
(3) Use "Peptide search"

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

ETMQFLNDRLASYLEKVRQLE and SENARLVVQIDNAKLAADDFRTKY were identified as human peptides.
Could they be typical 'false positive' ?

Glycosylation sites and proteomics

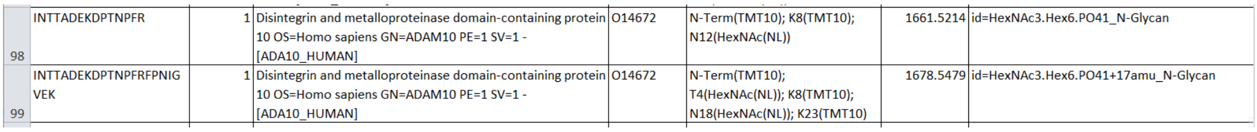
Look at the information available concerning the glycosylation site
https://prosite.expasy.org/PDOC00001



ASN_GLYCOSYLATION, PS00001; N-glycosylation site (PATTERN with a high probability of occurrence!)

- Consensus pattern:
  N-{P}-[ST]-{P}
  N is the glycosylation site

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 "INTTADEKDPTNPFR" the following "N12(HexNAc(NL))" modification.
and to the peptide "INTTADEKDPTNPFRFPNIGVEK" the following "N18(HexNAc(NL))" modification.



98	INTTADEKDPTNPFR	1	Disintegrin and metalloproteinase domain-containing protein 10 OS=Homo sapiens GN=ADAM10 PE=1 SV=1 - [ADA10_HUMAN]	O14672	N-Term(TMT10); K8(TMT10); N12(HexNAc(NL))	1661.5214	id=HexNAc3.Hex6.PO41_N-Glycan
99	INTTADEKDPTNPFRFPNIGVEK	1	Disintegrin and metalloproteinase domain-containing protein 10 OS=Homo sapiens GN=ADAM10 PE=1 SV=1 - [ADA10_HUMAN]	O14672	N-Term(TMT10); T4(HexNAc(NL)); K8(TMT10); N18(HexNAc(NL)); K23(TMT10)	1678.5479	id=HexNAc3.Hex6.PO41+17amu_N-Glycan

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 protein 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

Elisabeth Gasteiger, Marie-Claude Blatter
SIB Swiss Institute of Bioinformatics, November 2018