Practical Session #3: Protein primary databases

https://www.ncbi.nlm.nih.gov/books/NBK49540/

Note: When we ask for "3D structures" we mean 3D structures experimentally determined, not predicted

I. Find out the protein entries at NCBI which meet each of the following criteria. Try using the Advanced Search Builder. (Indicate only one Accession number for each one)

A. The protein encoded by nucleotide sequence in entry NM_005318.**H1.0 linker histone**

B. The human reference sequence encoded by gene *tnf*. **NP_000585**
+human[porgn] AND "tnf"[gene] AND RefSeq[filter]

C. The largest isoform of the epidermal growth factor receptor in the fruit fly in refseq.**NP_476759.1**
+"fruit fly"[porgn] AND RefSeq[filter] AND EGF[gene] >select largest

D. A mouse protein with a molecular weight of 40000 Da.**EDL28143.1**
>ADVANCED search

E. How many rice reference proteins are with a signal peptide annotated in its GenPept record as a feature?. **510**
+"rice"[porgn] AND "sig peptide"[fkey] AND refseq[filter]

F. A reference sequence for an *Escherichia coli* beta-lactamase with known 3D structure. **5 possible, NP_418574.1**
+"Escherichia coli"[porgn] AND protein_structure[filter] AND beta-lactamase[title] AND refseq[filter]

G. The human and the mouse (*Mus musculus musculus*) "cytochrome b" reference protein.**YP_003024038.1 (homo sapiens)**
+("mus musculus musculus"[porgn] OR "human"[porgn]) AND refseq[filter] AND "cytochrome b"[protein name]

II. Now look for the human sequence encoded by gene *tnf* at the UniProtKB database.

How many entries have you found? Which one is considered the reference record? **10 entries, P01375**

Open this reference/reviewed entry and get familiar with the structure of a sequence entry at UniProtKB and all the information and cross-references associated to each protein sequence. What can we learn about this protein from this UniProtKB entry? (for discussion in class)

In this UniProt entry check if there is any 3D structure for this human protein. How many entries in the database have you found?**1** What is the PDB accession code for the structure obtained with the highest (best) resolution?**5UUI 1.40**

How many mammalian proteins encoded by a gene whose name is tnf, are there at the Swiss-Prot database (reviewed records)? **35**

Tips: Try using the Advanced Search Builder

III. For the following proteins at UniprotKB match them with the type of evidence that supports their existence and how their functions were annotated.

UniProt Codes

Q2KIJ4 **D**

Q3SXR2 **B**

C4AMC7

P06310

Q8N0Y7

Q5T870

Q0D2H9

P11216

Annotation

A. Probable function inferred from homology.
B. Function has been predicted but there is no evidence at protein, transcript, or homology levels.
C. The existence of the protein is unsure.
D. A clearly identified protein with clear function. Hypothetical protein with evidence at transcript level.
E. Ribosome related protein with clear experimental evidence for its existence.
F. Nuclear protein with clear experimental evidence for its existence.
G. Known function inferred from homology.

IV. Find out the protein that meets all these requirements at the same time.

1. It is a human protein
2. It is stored at the SWISS-PROT protein sequence database
3. There is clear experimental evidence for its existence
4. In the cell it is localized in the plasma membrane (subcellular location note)
5. Its length is between 1000 and 2000 amino acids
6. It contains an Ig-like domain (feature)
7. It has a structure in the PDB database
8. It is expressed in the kidney

4.1 Will there be a possible homolog for this protein in non-human primates? In other mammalian species? In non-mammalian vertebrate species? (for discussion in class)

V. Protein 3D structures help to discover new drugs

In a scientific paper with PubMed ID 25970480 authors describe a series of already known chemical compounds that inhibit the activity of a human enzyme (ec:1.13.11.52) that is currently consider as an attractive target for anticancer therapy.

5.1 Search for this enzyme at UniProtKB and indicate its accession number, protein name and gene name.

5.2 Discover all the information associated with this entry at UniProtKB. Are there clear experimental evidences for the existence of the protein? Is its function known? What is the cofactor for this enzyme?

5.3 How many 3D structures are there associated with this UniProtKB entry? Which has been obtained with the highest(best) resolution? (PDB accession code)

5.4 Open the PDB file for the structure obtained with the highest resolution. Which method was used to solve this structure? How many polypeptide chains form this structure? If we wish to know which amino acids are thought to participate in the binding of the cofactor for this enzyme. Will this structure be useful? (for discussion in class)

5.6 Is this protein similar to any other protein at swissprot? Which ones? (for discussion in class)

5.7 Analyze the UniRef sets containing this protein. Are there putative orthologs in other organisms? How it could be possible to find remote homologs? (for discussion in class)

5.8 Will there be any protein similar to this in the four non-mammalian model organisms zebrafish, the yeast *Pichia pastoris*, *Caenorhabditis elegans* or *Escherichia coli*? (for discussion in class)