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)

The protein encoded by nucleotide sequence in entry NM 005318. (NP 005309.1)

The human reference sequence encoded by gene *tnf*. (NP_000585.2) tnf[Gene Name] AND "Homo sapiens"[Primary Organism] AND Refseq[filter]

The largest isoform of the epidermal growth factor receptor in the fruit fly in refseq. (NP 476759.1)

epidermal growth factor receptor AND txid7227[Primary Organism] AND refseq[filter]
here we cannot order by size, but we could look at the largest and add NOT 1:1425[SLEN]

A Mus musculus protein with a molecular weight of 40000 Da (BAE28525.1, EDL28143.1) "Mus musculus"[Primary Organism] AND 40000[Molecular Weight]

How many rice reference proteins are with a signal peptide annotated in its GenPept record as a feature (507)

((rice[Primary organism]) AND "refseq"[Filter]) AND "sig peptide"[Feature key]

A reference sequence for an Escherichia coli beta-lactamase with known 3D structure

Escherichia coli[porgn] AND protein structure[filter] AND beta-lactamase[TI] AND refseq[filter]

WP_063860054.1 WP_063860093.1 WP_139524977.1 NP_313158.1 NP_418574.1

Last year was also returning these results: Can you guess why they are not here now?

YP_007447512.1 YP_009062843.1 YP_009061308.1 YP_006940092.1 YP_003829170.1 YP_003829069.1 NP_943295.1 YP_190222.1

The human and the mouse (*Mus musculus musculus*) "cytochrome b" reference protein YP 001686710.1. YP 003024038.1

(txid9606[Primary Organism] OR txid39442[Primary Organism]) 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? 10 Which one is considered the reference record? P01375 (organism_id:9606) AND (gene:tnf)

(organism_id:9606) AND (gene:tnf) AND (reviewed:true)

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? 42 What is the PDB accession code for the structure obtained with the highest (best) resolution? 5UUI (1.4A)

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

(taxonomy id:40674) AND (gene:tnf) AND (reviewed:true)

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.

Q2KIJ4 Hypothetical protein with evidence at transcript level.

Q3SXR2 Uncharacterized protein with clear experimental evidence for its existence.

C4AMC7 Putative function but with clear experimental evidence for its existence.

P06310 Known function inferred from homology.

Q8N0Y7 Probable function inferred from homology.

Q5T870 Function has been predicted but there is no evidence at protein, transcript, or homology levels.

Q0D2H9 The existence of the protein is unsure.

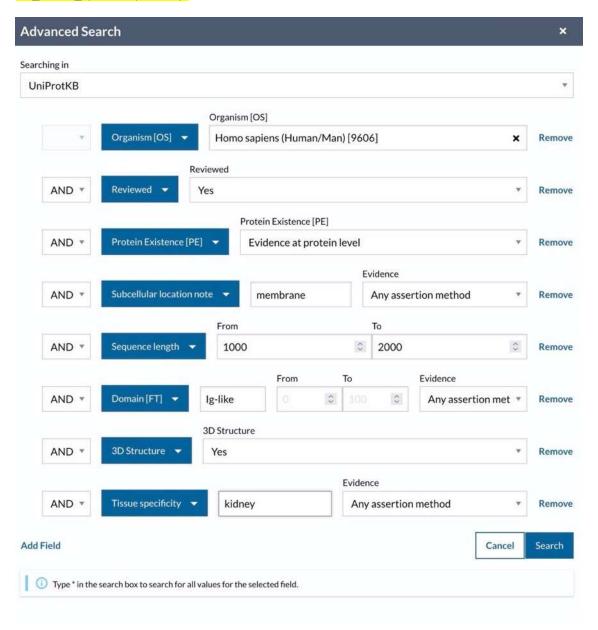
P11216 A clearly identified protein with clear function.

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

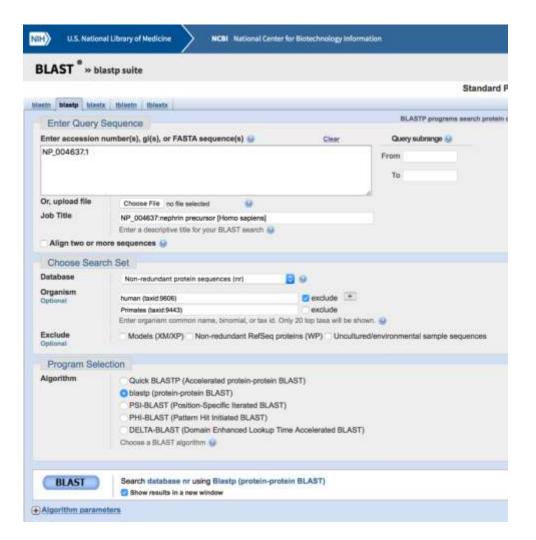
- 7. It has an structure in the PDB database
- 8. It is expressed in the kidney

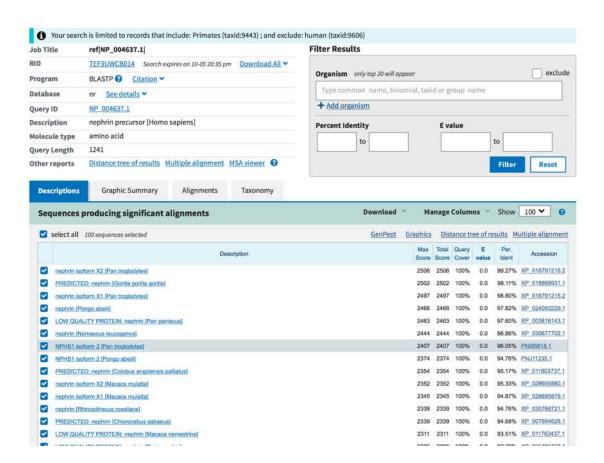
(organism_id:9606) AND (reviewed:true) AND (existence:1) AND (cc_scl_note:membrane) AND (length:[1000 TO 2000]) AND (ft_domain:lg-like) AND (structure_3d:true) AND (cc_tissue_specificity:kidney)



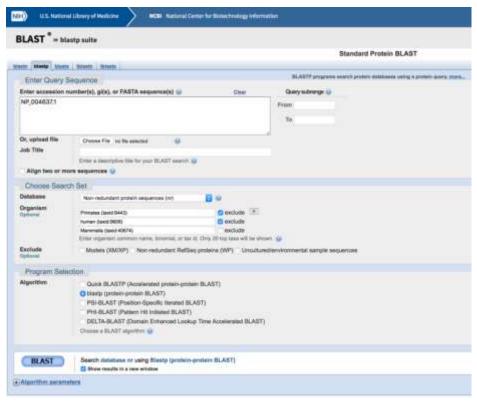
result: https://www.uniprot.org/uniprotkb/060500/entry

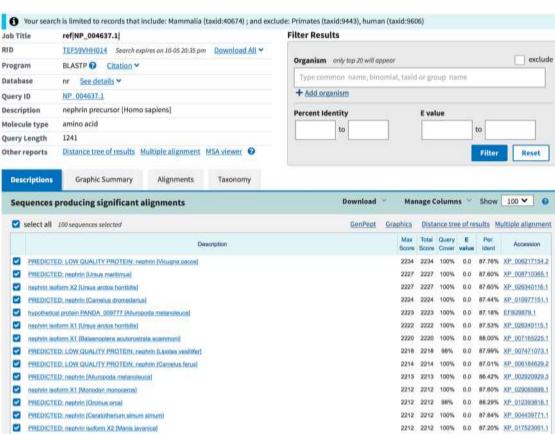
4.1 Will there be a possible homolog for this protein in non-human primates?



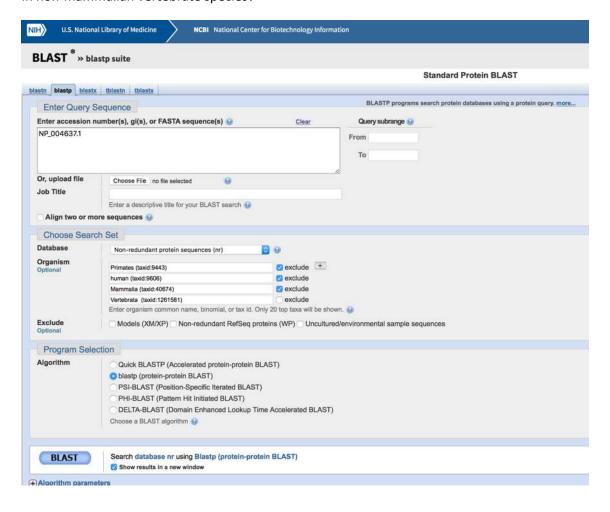


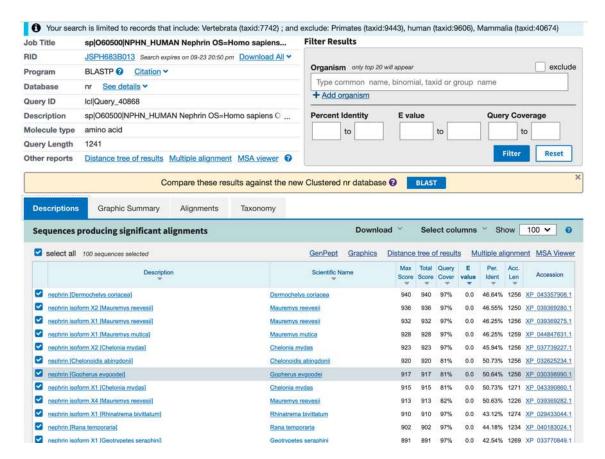
In other mammalian species?





In non-mammalian vertebrate species?





(for discussion in class) BLAST

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 P14902((ec:1.13.11.52) AND (organism_id:9606)) protein name Indoleamine 2,3-dioxygenase 1 and gene name IDO1, IDO, INDO
- 5.2 Discover all the information associated with this entry at UniProtKB. Are there clear experimental evidences for the existence of the protein? Yes(Experimental evidence at protein level) Is its function known? Yes What is the cofactor for this enzyme? heme group
- 5.3 How many 3D structures are there associated with this UniProtKB entry? 71 Which has been obtained with the highest(best) resolution? 8ABX(1.65A) (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? X-ray Crystallography. How many polypeptide chains form this structure? 1. 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)

In Uniprot, In the "Cofactor" annotation we can see that its cofactor is "heme"

8ABX does not have the heme group, we have to find one that has it, we can use PDB advanced search:

IDs and Keywords -> Accession codes uniprot: P14902

and

chemical components->synonyms: heme (chemical name does not return results) now we can refine by resolution and obtain only one structure: 6E44

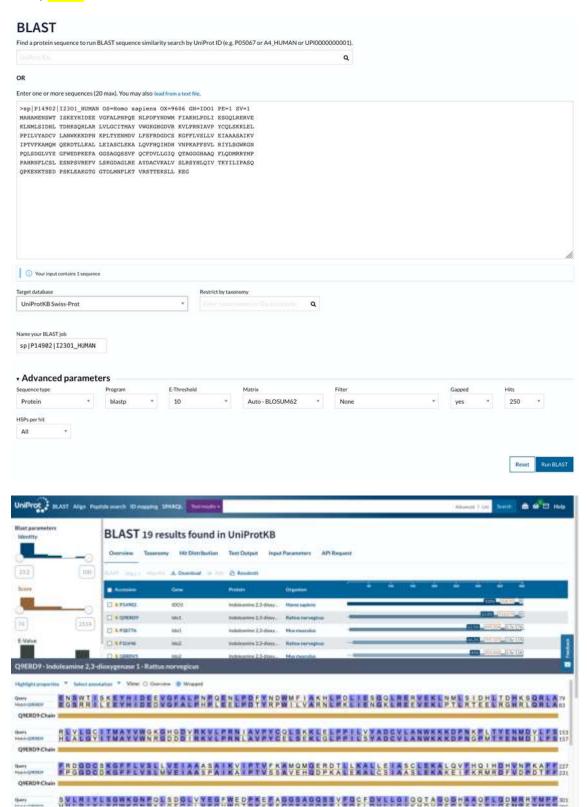
We can also search using the pubmed id of any of the papers that describe the cofactor (they can be found in the cofactor section in the P14902 uniprot entry.

This way we can use (6E44) (2D0U, 2D0T) or (4PK5 4PK6) (or many others):

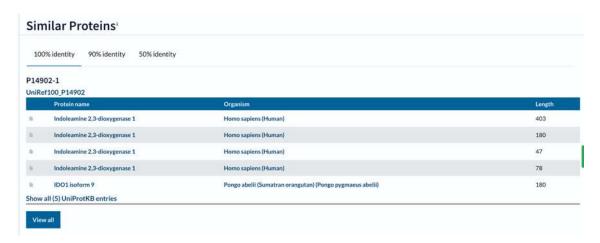
We will find certain differences, but it is normal Ex:

<mark>2D0U</mark>															
SITE	1	AC1	22	TYR	А	126	PHE	А	163	SER	A	167	PHE	А	214
SITE	2	AC1	22	PHE	A	226	SER	A	263	ALA	A	264	PHE	А	270
SITE	3	AC1	22	ARG	A	343	HIS	A	346	ILE	A	349	VAL	А	350
SITE	4	AC1	22	TYR	A	353	LEU	А	384	LEU	A	388	VAL	А	391
6e44															
SITE	1	AC1	21	TYR	A	126	PHE	A	163	SER	A	167	VAL	А	170
SITE	2	AC1	21	PHE	A	214	PHE	A	226	GLY	A	262	ALA	А	264
SITE	3	AC1	21	GLY	A	265	PHE	A	270	MET	A	295	ARG	А	343
SITE	4	AC1	21	HIS	A	346	ILE	А	349	VAL	A	350	ILE	А	354
SITE	5	AC1	21	LEU	A	384	VAL	Α	391	НОН	A	609	НОН	A	736

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



5.7 Analyze the UniRef sets containing this protein. Are there putative orthologs proteins in other organisms?



Uniref 100 (100% identical) contains a protein from *Pongo abelii*. Uniref90 contains proteins from several different species. Take care with uniref50, they are similar but could not be orthologs.

How it could be possible to find remote orthologs? (for discussion in class) BLAST

Remote homologous search (psi-blast)

Foldseek (structure seach)

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) NCBIBLAST

We can only find hits for zebrafish and *Pichia pastoris* (Komagataella pastoris is a synonym) perhaps a remote homolog, but ...

