

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[SLLEN]

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

- ```
(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/O60500/entry>

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

U.S. National Library of Medicine

NCBI National Center for Biotechnology Information

BLAST® » blastp suite

Standard F

blastnblastpblastxtblastntblastx

BLASTP programs search protein c

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#)

Clear

NP\_004637.1

Query subrange [?](#)

From

To

Or, upload file

Choose File no file selected [?](#)

Job Title

NP\_004637:nephrin precursor [Homo sapiens]

Enter a descriptive title for your BLAST search [?](#)

☐ Align two or more sequences [?](#)

Choose Search Set

Database

Non-redundant protein sequences (nr) [?](#)

Organism

Optional

human (taxid:9606) ☒ exclude [?](#)

Primates (taxid:9443) ☐ exclude

Enter organism common name, binomial, or tax id. Only 30 top taxa will be shown. [?](#)

Exclude

Optional

☐ Models (XM/XP) ☐ Non-redundant RefSeq proteins (WP) ☐ Uncultured/environmental sample sequences

Program Selection

Algorithm

☐ Quick BLASTP (Accelerated protein-protein BLAST)

☒ blastp (protein-protein BLAST)

☐ PSI-BLAST (Position-Specific Iterated BLAST)

☐ PHI-BLAST (Pattern Hit Initiated BLAST)

☐ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm [?](#)

BLAST

Search database nr using Blastp (protein-protein BLAST)

☒ Show results in a new window

[+ Algorithm parameters](#)

**i** Your search is limited to records that include: Primates (taxid:9443) ; and exclude: human (taxid:9606)

Job Title **ref|NP\_004637.1|**

RID [TEF3UWCB014](#) Search expires on 10-05 20:35 pm [Download All](#) ▼

Program **BLASTP** [?](#) [Citation](#) ▼

Database **nr** [See details](#) ▼

Query ID **NP\_004637.1**

Description **nephrin precursor [Homo sapiens]**

Molecule type **amino acid**

Query Length **1241**

Other reports [Distance tree of results](#) [Multiple alignment](#) [MSA viewer](#) [?](#)

**Filter Results**

**Organism** *only top 20 will appear* ☐ exclude

Type common name, binomial, taxid or group name

[+ Add organism](#)

**Percent Identity**  to

**E value**  to

[Filter](#) [Reset](#)

- Descriptions** Graphic Summary Alignments Taxonomy

| Sequences producing significant alignments                            |                                                   |           |             |                         |                          |                                          | Download                           | Manage Columns | Show | 100 | ? |
|-----------------------------------------------------------------------|---------------------------------------------------|-----------|-------------|-------------------------|--------------------------|------------------------------------------|------------------------------------|----------------|------|-----|---|
| <input checked="" type="checkbox"/> select all 100 sequences selected |                                                   |           |             |                         |                          |                                          |                                    |                |      |     |   |
|                                                                       |                                                   |           |             | <a href="#">GenPept</a> | <a href="#">Graphics</a> | <a href="#">Distance tree of results</a> | <a href="#">Multiple alignment</a> |                |      |     |   |
|                                                                       | Description                                       | Max Score | Total Score | Query Cover             | E value                  | Per. Ident                               | Accession                          |                |      |     |   |
| <input checked="" type="checkbox"/>                                   | nephrin isoform X2 [Pan troglodytes]              | 2506      | 2506        | 100%                    | 0.0                      | 99.27%                                   | <a href="#">XP_016791216.2</a>     |                |      |     |   |
| <input checked="" type="checkbox"/>                                   | PREDICTED: nephrin [Gorilla gorilla gorilla]      | 2502      | 2502        | 100%                    | 0.0                      | 99.11%                                   | <a href="#">XP_018869931.1</a>     |                |      |     |   |
| <input checked="" type="checkbox"/>                                   | nephrin isoform X1 [Pan troglodytes]              | 2497      | 2497        | 100%                    | 0.0                      | 98.80%                                   | <a href="#">XP_016791215.2</a>     |                |      |     |   |
| <input checked="" type="checkbox"/>                                   | nephrin [Pongo abelii]                            | 2466      | 2466        | 100%                    | 0.0                      | 97.82%                                   | <a href="#">XP_024093229.1</a>     |                |      |     |   |
| <input checked="" type="checkbox"/>                                   | LOW QUALITY PROTEIN: nephrin [Pan paniscus]       | 2463      | 2463        | 100%                    | 0.0                      | 97.60%                                   | <a href="#">XP_003816143.1</a>     |                |      |     |   |
| <input checked="" type="checkbox"/>                                   | nephrin [Nomascus leucogenys]                     | 2444      | 2444        | 100%                    | 0.0                      | 96.86%                                   | <a href="#">XP_030677702.1</a>     |                |      |     |   |
| <input checked="" type="checkbox"/>                                   | NPHS1 isoform 2 [Pan troglodytes]                 | 2407      | 2407        | 100%                    | 0.0                      | 96.05%                                   | <a href="#">PNI95818.1</a>         |                |      |     |   |
| <input checked="" type="checkbox"/>                                   | NPHS1 isoform 2 [Pongo abelii]                    | 2374      | 2374        | 100%                    | 0.0                      | 94.76%                                   | <a href="#">PNJ11235.1</a>         |                |      |     |   |
| <input checked="" type="checkbox"/>                                   | PREDICTED: nephrin [Colobus angolensis palliatus] | 2354      | 2354        | 100%                    | 0.0                      | 95.17%                                   | <a href="#">XP_011803737.1</a>     |                |      |     |   |
| <input checked="" type="checkbox"/>                                   | nephrin isoform X2 [Macaca mulatta]               | 2352      | 2352        | 100%                    | 0.0                      | 95.33%                                   | <a href="#">XP_028695880.1</a>     |                |      |     |   |
| <input checked="" type="checkbox"/>                                   | nephrin isoform X1 [Macaca mulatta]               | 2345      | 2345        | 100%                    | 0.0                      | 94.87%                                   | <a href="#">XP_028695879.1</a>     |                |      |     |   |
| <input checked="" type="checkbox"/>                                   | nephrin [Rhinopithecus roxellana]                 | 2339      | 2339        | 100%                    | 0.0                      | 94.76%                                   | <a href="#">XP_030768721.1</a>     |                |      |     |   |
| <input checked="" type="checkbox"/>                                   | PREDICTED: nephrin [Chlorocebus sabaeus]          | 2339      | 2339        | 100%                    | 0.0                      | 94.68%                                   | <a href="#">XP_007994626.1</a>     |                |      |     |   |
| <input checked="" type="checkbox"/>                                   | LOW QUALITY PROTEIN: nephrin [Macaca nemestrina]  | 2311      | 2311        | 100%                    | 0.0                      | 93.51%                                   | <a href="#">XP_011763437.1</a>     |                |      |     |   |

In other mammalian species?

**BLAST** U.S. National Library of Medicine National Center for Biotechnology Information

**BLAST** blastp suite

**Standard Protein BLAST**

Enter Query Sequence

Enter accession number(s), g(i), or FASTA sequence(s)

NP\_004637.1

Clear

Query subrange

From

To

Or, upload file

Choose File (no file selected)

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database

Non-redundant protein sequences (nr)

Organism

Optional

Primates (taxid:9443) exclude

human (taxid:9606) exclude

Mammalia (taxid:40674) exclude

Enter organism common name, binomial, or tax id. (Only 20 top taxa will be shown)

Exclude

Optional

Models (MMOP) Non-redundant RefSeq proteins (WP) Unstructured/environmental sample sequences

Program Selection

Algorithm

Quick BLASTP (Accelerated protein-protein BLAST)

blastp (protein-protein BLAST)

PSI-BLAST (Position-Specific Iterated BLAST)

PIR-BLAST (Protein-18 Iterated BLAST)

DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

**BLAST** Search database nr using Blastp (protein-protein BLAST)

Show results in a new window

Algorithm parameters

**!** Your search is limited to records that include: Mammalia (taxid:40674) ; and exclude: Primates (taxid:9443), human (taxid:9606)

Job Title **ref[NP\_004637.1]**

RID **TEF59VHHQ14** Search expires on 10-05 20:35 pm [Download All](#)

Program **BLASTP** [Citation](#)

Database **nr** [See details](#)

Query ID **NP\_004637.1**

Description **nephrin precursor [Homo sapiens]**

Molecule type **amino acid**

Query Length **1241**

Other reports [Distance tree of results](#) [Multiple alignment](#) [MSA viewer](#)

**Filter Results**

Organism only top 20 will appear ☐ exclude

Type common name, binomial, taxid or group name

[+ Add organism](#)


Percent Identity  to


E value  to

**Filter** **Reset**

| Descriptions                                                          | Graphic Summary                                              | Alignments | Taxonomy    |             |         |            |                |
|-----------------------------------------------------------------------|--------------------------------------------------------------|------------|-------------|-------------|---------|------------|----------------|
| Sequences producing significant alignments                            |                                                              |            |             |             |         |            |                |
| Download Manage Columns Show 100                                      |                                                              |            |             |             |         |            |                |
| <input checked="" type="checkbox"/> select all 100 sequences selected |                                                              |            |             |             |         |            |                |
| GenPept Graphics Distance tree of results Multiple alignment          |                                                              |            |             |             |         |            |                |
|                                                                       | Description                                                  | Max Score  | Total Score | Query Cover | E value | Per. Ident | Accession      |
| <input checked="" type="checkbox"/>                                   | PREDICTED: LOW QUALITY PROTEIN: nephrin [Vicugna pacos]      | 2234       | 2234        | 100%        | 0.0     | 87.76%     | XP_006217154.2 |
| <input checked="" type="checkbox"/>                                   | PREDICTED: nephrin [Ursus maritimus]                         | 2227       | 2227        | 100%        | 0.0     | 87.60%     | XP_008710365.1 |
| <input checked="" type="checkbox"/>                                   | nephrin isoform X2 [Ursus arctos horribilis]                 | 2227       | 2227        | 100%        | 0.0     | 87.60%     | XP_026340116.1 |
| <input checked="" type="checkbox"/>                                   | PREDICTED: nephrin [Camelus dromedarius]                     | 2224       | 2224        | 100%        | 0.0     | 87.44%     | XP_010977151.1 |
| <input checked="" type="checkbox"/>                                   | hypothetical protein PANDA_009777 [Alluropoda melanoleuca]   | 2223       | 2223        | 100%        | 0.0     | 87.18%     | EFB29879.1     |
| <input checked="" type="checkbox"/>                                   | nephrin isoform X1 [Ursus arctos horribilis]                 | 2222       | 2222        | 100%        | 0.0     | 87.53%     | XP_026340115.1 |
| <input checked="" type="checkbox"/>                                   | nephrin isoform X1 [Balaenoptera acutorostrata scammoni]     | 2220       | 2220        | 100%        | 0.0     | 88.00%     | XP_007185225.1 |
| <input checked="" type="checkbox"/>                                   | PREDICTED: LOW QUALITY PROTEIN: nephrin [Lipotes vexillifer] | 2218       | 2218        | 98%         | 0.0     | 87.99%     | XP_007471073.1 |
| <input checked="" type="checkbox"/>                                   | PREDICTED: LOW QUALITY PROTEIN: nephrin [Camelus fetus]      | 2214       | 2214        | 100%        | 0.0     | 87.01%     | XP_006184629.2 |
| <input checked="" type="checkbox"/>                                   | PREDICTED: nephrin [Alluropoda melanoleuca]                  | 2213       | 2213        | 100%        | 0.0     | 86.42%     | XP_002920929.3 |
| <input checked="" type="checkbox"/>                                   | nephrin isoform X1 [Monodon monoceros]                       | 2212       | 2212        | 100%        | 0.0     | 87.60%     | XP_029065899.1 |
| <input checked="" type="checkbox"/>                                   | PREDICTED: nephrin [Orcinus orca]                            | 2212       | 2212        | 98%         | 0.0     | 88.29%     | XP_012393818.1 |
| <input checked="" type="checkbox"/>                                   | PREDICTED: nephrin [Ceratotherium simum simum]               | 2212       | 2212        | 100%        | 0.0     | 87.84%     | XP_004439771.1 |
| <input checked="" type="checkbox"/>                                   | PREDICTED: nephrin isoform X2 [Manis javanica]               | 2212       | 2212        | 100%        | 0.0     | 87.20%     | XP_017523001.1 |

In non-mammalian vertebrate species?

 U.S. National Library of Medicine

 National Center for Biotechnology Information


**BLAST®** » blastp suite

**Standard Protein BLAST**

blastnblastpblastxtblastntblastx


Enter Query Sequence

BLASTP programs search protein databases using a protein query. [more...](#)

Enter accession number(s), gi(s), or FASTA sequence(s) 

Clear


NP\_004637.1

Query subrange 


From


To

Or, upload file

Choose File no file selected 


Job Title

Enter a descriptive title for your BLAST search 


☐ Align two or more sequences 

Choose Search Set

Database

Non-redundant protein sequences (nr) 


Organism   
 Optional

Primates (taxid:9443) ☒ exclude 

human (taxid:9606) ☒ exclude

Mammalia (taxid:40674) ☒ exclude

Vertebrata (taxid:1261581) ☐ exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. 

Exclude   
 Optional

☐ Models (XM/XP) ☐ Non-redundant RefSeq proteins (WP) ☐ Uncultured/environmental sample sequences

Program Selection

Algorithm


☐ Quick BLASTP (Accelerated protein-protein BLAST)

☒ blastp (protein-protein BLAST)

☐ PSI-BLAST (Position-Specific Iterated BLAST)

☐ PHI-BLAST (Pattern Hit Initiated BLAST)


☐ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm 

BLAST

Search database nr using Blastp (protein-protein BLAST)

☒ Show results in a new window

 Algorithm parameters



**!** Your search is limited to records that include: Vertebrata (taxid:7742) ; and exclude: Primates (taxid:9443), human (taxid:9606), Mammalia (taxid:40674)

**Job Title** sp|O60500|NPHN\_HUMAN Nephrin OS=Homo sapiens...

**RID** JSPH683B013 [Search expires on 09-23 20:50 pm](#) [Download All](#) ▼

**Program** BLASTP [?](#) [Citation](#) ▼

**Database** nr [See details](#) ▼

**Query ID** Icl|Query\_40868

**Description** sp|O60500|NPHN\_HUMAN Nephrin OS=Homo sapiens C ...

**Molecule type** amino acid

**Query Length** 1241

**Other reports** [Distance tree of results](#) [Multiple alignment](#) [MSA viewer](#) [?](#)

**Filter Results**

**Organism** *only top 20 will appear* ☐ exclude

Type common name, binomial, taxid or group name

[+ Add organism](#)

**Percent Identity**  to  **E value**  to  **Query Coverage**  to

[Filter](#) [Reset](#)

Compare these results against the new Clustered nr database [?](#) [BLAST](#) [×](#)

**Descriptions** Graphic Summary Alignments Taxonomy

**Sequences producing significant alignments** [Download](#) ▼ [Select columns](#) ▼ Show  [?](#)

☒ select all 100 sequences selected [GenPept](#) [Graphics](#) [Distance tree of results](#) [Multiple alignment](#) [MSA Viewer](#)

|                                     | Description ▼                                               | Scientific Name ▼                      | Max Score ▼ | Total Score ▼ | Query Cover ▼ | E value ▼ | Per. Ident ▼ | Acc. Len ▼ | Accession                      |
|-------------------------------------|-------------------------------------------------------------|----------------------------------------|-------------|---------------|---------------|-----------|--------------|------------|--------------------------------|
| <input checked="" type="checkbox"/> | <a href="#">nephrin [Dermochelys coriacea]</a>              | <a href="#">Dermochelys coriacea</a>   | 940         | 940           | 97%           | 0.0       | 46.64%       | 1256       | <a href="#">XP_043357908.1</a> |
| <input checked="" type="checkbox"/> | <a href="#">nephrin isoform X2 [Mauremys reevesii]</a>      | <a href="#">Mauremys reevesii</a>      | 936         | 936           | 97%           | 0.0       | 46.55%       | 1250       | <a href="#">XP_039369280.1</a> |
| <input checked="" type="checkbox"/> | <a href="#">nephrin isoform X1 [Mauremys reevesii]</a>      | <a href="#">Mauremys reevesii</a>      | 932         | 932           | 97%           | 0.0       | 46.25%       | 1256       | <a href="#">XP_039369275.1</a> |
| <input checked="" type="checkbox"/> | <a href="#">nephrin isoform X1 [Mauremys mutica]</a>        | <a href="#">Mauremys mutica</a>        | 928         | 928           | 97%           | 0.0       | 46.25%       | 1259       | <a href="#">XP_044847631.1</a> |
| <input checked="" type="checkbox"/> | <a href="#">nephrin isoform X2 [Chelonia mydas]</a>         | <a href="#">Chelonia mydas</a>         | 923         | 923           | 97%           | 0.0       | 45.94%       | 1256       | <a href="#">XP_037739227.1</a> |
| <input checked="" type="checkbox"/> | <a href="#">nephrin [Chelonoidis abingdonii]</a>            | <a href="#">Chelonoidis abingdonii</a> | 920         | 920           | 81%           | 0.0       | 50.73%       | 1256       | <a href="#">XP_032625234.1</a> |
| <input checked="" type="checkbox"/> | <a href="#">nephrin [Gopherus evgodei]</a>                  | <a href="#">Gopherus evgodei</a>       | 917         | 917           | 81%           | 0.0       | 50.64%       | 1256       | <a href="#">XP_030398990.1</a> |
| <input checked="" type="checkbox"/> | <a href="#">nephrin isoform X1 [Chelonia mydas]</a>         | <a href="#">Chelonia mydas</a>         | 915         | 915           | 81%           | 0.0       | 50.73%       | 1271       | <a href="#">XP_043390860.1</a> |
| <input checked="" type="checkbox"/> | <a href="#">nephrin isoform X4 [Mauremys reevesii]</a>      | <a href="#">Mauremys reevesii</a>      | 913         | 913           | 82%           | 0.0       | 50.63%       | 1226       | <a href="#">XP_039369282.1</a> |
| <input checked="" type="checkbox"/> | <a href="#">nephrin isoform X1 [Rhinatremia bivittatum]</a> | <a href="#">Rhinatremia bivittatum</a> | 910         | 910           | 97%           | 0.0       | 43.12%       | 1274       | <a href="#">XP_029433044.1</a> |
| <input checked="" type="checkbox"/> | <a href="#">nephrin [Rana temporaria]</a>                   | <a href="#">Rana temporaria</a>        | 902         | 902           | 97%           | 0.0       | 44.18%       | 1234       | <a href="#">XP_040183024.1</a> |
| <input checked="" type="checkbox"/> | <a href="#">nephrin isoform X1 [Geotrvpetes serachini]</a>  | <a href="#">Geotrvpetes serachini</a>  | 891         | 891           | 97%           | 0.0       | 42.54%       | 1269       | <a href="#">XP_033770849.1</a> |

(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 considered 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:

**2D0U**

|      |   |     |    |     |   |     |     |   |     |     |   |     |     |   |     |
|------|---|-----|----|-----|---|-----|-----|---|-----|-----|---|-----|-----|---|-----|
| SITE | 1 | AC1 | 22 | TYR | A | 126 | PHE | A | 163 | SER | A | 167 | PHE | A | 214 |
| SITE | 2 | AC1 | 22 | PHE | A | 226 | SER | A | 263 | ALA | A | 264 | PHE | A | 270 |
| SITE | 3 | AC1 | 22 | ARG | A | 343 | HIS | A | 346 | ILE | A | 349 | VAL | A | 350 |
| SITE | 4 | AC1 | 22 | TYR | A | 353 | LEU | A | 384 | LEU | A | 388 | VAL | A | 391 |

**6e44**

|      |   |     |    |     |   |     |     |   |     |     |   |     |     |   |     |
|------|---|-----|----|-----|---|-----|-----|---|-----|-----|---|-----|-----|---|-----|
| SITE | 1 | AC1 | 21 | TYR | A | 126 | PHE | A | 163 | SER | A | 167 | VAL | A | 170 |
| SITE | 2 | AC1 | 21 | PHE | A | 214 | PHE | A | 226 | GLY | A | 262 | ALA | A | 264 |
| SITE | 3 | AC1 | 21 | GLY | A | 265 | PHE | A | 270 | MET | A | 295 | ARG | A | 343 |
| SITE | 4 | AC1 | 21 | HIS | A | 346 | ILE | A | 349 | VAL | A | 350 | ILE | A | 354 |
| SITE | 5 | AC1 | 21 | LEU | A | 384 | VAL | A | 391 | HOH | A | 609 | HOH | A | 736 |

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

## BLAST

Find a protein sequence to run BLAST sequence similarity search by UniProt ID (e.g. P05067 or A4\_HUMAN or UPI0000000001).

OR

Enter one or more sequences (20 max). You may also [load from a text file](#).

```
>sp|P14902|I2301_HUMAN OS=Homo sapiens OX=9606 GN=IDO1 PE=1 SV=1
MAHAMENSWT ISKEYHIDEE VGFALPNPQE NLPDFYNDWM FIAKHLPLDI ESQQLRERVE
KLNLSDIHL TDHKSQRLAR LVLGCITMAY VWGKGHGDVR KVLPRNIAPV YCQLSKKLEL
PPILVYADCV LANKKKDPN KPLTYENHDV LFSFRDGDGS KGFFLVSLV EIAAASAIKV
IPTVFKAQWQ QERDTLLKAL LEIASCLEKA LQVFHQIHDH VHKAFPSVL RIYLSGNKGN
PQLSDGLVYE GFWEDPKEFA GGSAGQSSVF QCFDVLGIIQ QTAGGGHAAQ FLQDMRRYMP
PAHRNFLCSL ESNPSVREFV LSKGDAGLRE AYDACVKALV SLRSYHLQIV TKYILIPASQ
QPKENTSED PSKLEAKGTG GTDLMNPLKT VRSTTEKSL KEG
```

Your input contains 1 sequence

Target database: UniProtKB Swiss-Prot

Restrict by taxonomy:

Name your BLAST job:

**Advanced parameters**  
Sequence type: Protein  Program: blastp  E-Threshold: 10  Matrix: Auto - BLOSUM62  Filter: None  Gapped: yes  Hits: 250   
HSPs per hit: All

UniProt BLAST Align Peptide search ID mapping SPARQL **BLAST results** Advanced 1.0.0 Search Help

BLAST parameters: Identity: 23.1 (131) Score: 76 (2133) E-Value: 1.1e-13

### BLAST 19 results found in UniProtKB

Overview Taxonomy Hit Distribution Text Output Input Parameters API Request

BLAST [Help](#) [Help](#) [Download](#) [Help](#) [Results](#)

| Accession | Gene | Protein                       | Organism          |
|-----------|------|-------------------------------|-------------------|
| P14902    | IDO1 | Indoleamine 2,3-dioxygenase 1 | Homo sapiens      |
| Q7N82P    | IDO1 | Indoleamine 2,3-dioxygenase 1 | Rattus norvegicus |
| P18276    | IDO1 | Indoleamine 2,3-dioxygenase 1 | Macaca mulatta    |
| P11546    | IDO2 | Indoleamine 2,3-dioxygenase 2 | Rattus norvegicus |
| Q7N82Q    | IDO2 | Indoleamine 2,3-dioxygenase 2 | Macaca mulatta    |

**Q7N82P - indoleamine 2,3-dioxygenase 1 - Rattus norvegicus**

Highlight properties Select annotation View Overview Wrapped

Query: MAHAMENSWTISKEYHIDEEVGFALPNPQENLPDFYNDWMFIAKHLPLDI ESQQLRERVEKLNLSDIHLTDHKSQRLAREGRRRL E EYHIDEEVGFALPHRLLELPDTRPWLIVARNLKL E NGKLREEVEKLPTRTEELRGHRLQRLA 79 83

Q7N82P-Chain: RLVLCITMAYVWVGKHQDVVKVLPNNIAVETCOLSKKLELPPILVYADCVLANWKKKDPNKF LT TENMOVLFS 153

Query: HVALDYITMAYVWNRDDDTIKVLPNNIAVETCOLSKKLELPPILVYADCVLANWKKKDPNQRNITENMDILFS 157

Q7N82P-Chain: FRDDDCSGKFFLVSLVETIAAASATKVIPTVFKAQWQERDTLLKALLEIASCLEKALQVFHQIHDHVNPKAFF 227

Query: FRDDDCSGKFFLVSLVETIAAASPAIKATPTWSSAVEHDDPKALEKALCSIAASLEKAKEIKRMRDFVDDTFR 231

Q7N82P-Chain: SVLNIYLSQWKGNFQSLSDGLVYEFWFEDPKEFAGGSAGQSSVFQCFDVLGIIQQTAGGGHAAQFLQDMRRYMP 301

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

**Similar Proteins<sup>1</sup>**

100% identity 90% identity 50% identity

P14902-1  
UniRef100\_P14902

| Protein name                    | Organism                                                  | Length |
|---------------------------------|-----------------------------------------------------------|--------|
| ■ Indoleamine 2,3-dioxygenase 1 | Homo sapiens (Human)                                      | 403    |
| ■ Indoleamine 2,3-dioxygenase 1 | Homo sapiens (Human)                                      | 180    |
| ■ Indoleamine 2,3-dioxygenase 1 | Homo sapiens (Human)                                      | 47     |
| ■ Indoleamine 2,3-dioxygenase 1 | Homo sapiens (Human)                                      | 78     |
| ■ IDO1 isoform 9                | Pongo abelii (Sumatran orangutan) (Pongo pygmaeus abelii) | 180    |

Show all (5) UniProtKB entries

View all

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) **NCBI BLAST**

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

[Edit Search](#)[Save Search](#)[Search Summary](#)[How to read this report?](#)[BLAST Help Videos](#)[Back to Traditional Results Page](#)

**i** Your search is limited to records that include: zebrafish (taxid:7955), Pichia pastoris (taxid:4922), Caenorhabditis elegans (taxid:6239), Escherichia coli (taxid:562)

Job Title **sp|P14902|I23O1\_HUMAN Indoleamine 2,3-dioxygenase...**  
RID **JSZBF5TY013** Search expired on 09-23 23:31 pm [Download All](#)   
Program **BLASTP**  [Citation](#)   
Database **nr** [See details](#)   
Query ID **lcl|Query\_52175**  
Description **sp|P14902|I23O1\_HUMAN Indoleamine 2,3-dioxygenase ...**  
Molecule type **amino acid**  
Query Length **403**  
Other reports [Distance tree of results](#) [Multiple alignment](#) [MSA viewer](#) 

## Filter Results

**Organism** only top 20 will appear ☐ exclude




Type common name, binomial, taxid or group name

[+ Add organism](#)Use up and down arrows to scroll**Percent Identity** to **E value** to **Query Coverage** to **Filter****Reset**Compare these results against the new Clustered nr database **BLAST**

## Descriptions

[Graphic Summary](#)[Alignments](#)[Taxonomy](#)

## Sequences producing significant alignments

[Download](#) [Select columns](#) [Show](#) **100** ☒ select all 5 sequences selected[GenPept](#)[Graphics](#)[Distance tree of results](#)[Multiple alignment](#)[MSA Viewer](#)

|                                     | Description                                            | Scientific Name       | Max Score | Total Score | Query Cover | E value | Per. Ident | Acc. Len | Accession      |
|-------------------------------------|--------------------------------------------------------|-----------------------|-----------|-------------|-------------|---------|------------|----------|----------------|
| <input checked="" type="checkbox"/> | Indoleamine 2,3-dioxygenase 1 (Danio rerio)            | Danio rerio           | 370       | 370         | 95%         | 2e-123  | 44.64%     | 435      | AA83017.1      |
| <input checked="" type="checkbox"/> | Indoleamine 2,3-dioxygenase 1 (Danio rerio)            | Danio rerio           | 368       | 368         | 95%         | 2e-122  | 44.50%     | 435      | HP_001077323.1 |
| <input checked="" type="checkbox"/> | Indoleamine 2,3-dioxygenase 2 isoform X1 (Danio rerio) | Danio rerio           | 231       | 231         | 82%         | 3e-71   | 43.94%     | 289      | XP_021333878.1 |
| <input checked="" type="checkbox"/> | BATS_0099170 (Komagataella pastoris)                   | Komagataella pastoris | 212       | 212         | 95%         | 1e-61   | 34.15%     | 482      | AN274802.1     |
| <input checked="" type="checkbox"/> | Indoleamine 2,3-dioxygenase 2 isoform X2 (Danio rerio) | Danio rerio           | 190       | 190         | 91%         | 2e-55   | 42.51%     | 273      | XP_021333879.1 |