EMBL-EBI Tutorial

(V14 26/10/11)

UniProt

Universal Protein Resource

www.uniprot.org

Introduction

The UniProt (Universal Protein Resource) Consortium is comprised of the European Bioinformatics Institute, the Swiss Institute of Bioinformatics and the Protein Information Resource. The UniProt consortium aims to support biological research by maintaining a high quality database that serves as a stable, fully classified, richly and accurately annotated protein sequence knowledgebase, with extensive cross-references and querying interfaces freely accessible to the scientific community. All data stored in UniProt can be downloaded in bulk from the Download Centre at http://www.uniprot.org/downloads. Specific data sets can be downloaded from the UniProt website.

UniProt website is the world's most comprehensive catalogue of information on proteins. It is a central repository of protein sequence and function. The **UniProt Knowledgebase** (UniProtKB) is the central access point for extensive curated protein information, including function, classification, and cross-reference. The database is divided into two section UniProtKB/Swiss-Prot which is manually curated and UniProtKB/TrEMBL which is automatically maintained.

During this tutorial you will learn how to search for entries in the database and navigate within an entry, find out what information we annotate and how to extract the maximum amount of information from them.

Tutorial

Note – some of the questions ask for numerical answers. Uniprot is an active database with on-going data input and curation. Thus numbers are accurate at the time of writing but may vary over time.

Exploring UniProtKB – searching, BLAST & align.

Q1 (i) Using the simple search do a search for Human. Find out how many UniProtKB entries are hit?

Q1 (ii) To find the number of Human proteins in UniProtKB do another search for human but this time search exclusively in the Species fields by using the Advanced Search link.

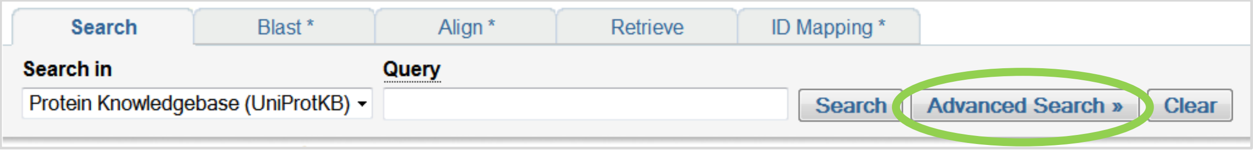


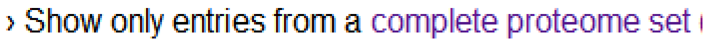
Figure 1. The UniProt search bar.

Q2. How many Homo sapiens (Human) entries are in UniProtKB/Swiss-Prot?

Q3. Why do you think there is such a big difference between the answers answer for Q 1 ii and Q 2?

(Hint – Think about the different sections of UniProtKB. – TrEMBL and Swiss-Prot)

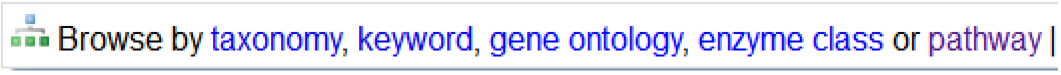
Click on the link to go to the complete proteome set:



› Show only entries from a complete proteome set

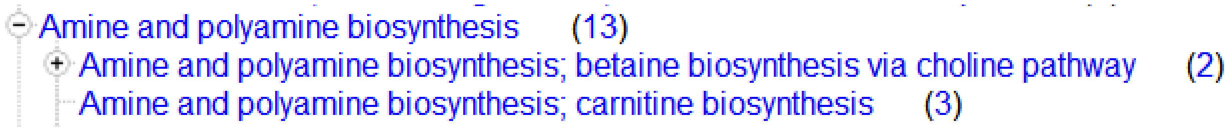
Q4. How many UniProtKB entries are there in the Human Complete Proteome Set?

Browse the various pathways which the proteins have been annotated as being involved in by clicking on the pathway link:



Browse by taxonomy, keyword, gene ontology, enzyme class or pathway |

You should get a list of pathways that look like this:

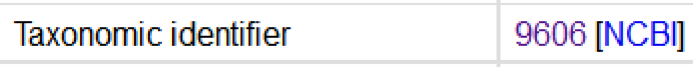


Amine and polyamine biosynthesis (13)
Amine and polyamine biosynthesis; betaine biosynthesis via choline pathway (2)
Amine and polyamine biosynthesis; carnitine biosynthesis (3)

The (+) symbol indicates that there are sub-pathways. The blue pathway name links to UniPathway which give more information about the pathway. The blue number links to those entries in UniProtKB which have been annotated to that pathway.

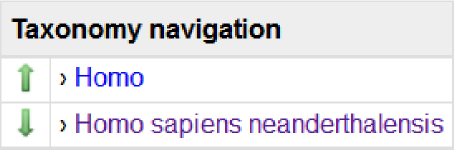
✍ Click on the BACK button on your browser to get back to the list of proteins

✍ Click on any entry in the results section and go to the Names and Origin section. Then click on the numerical taxonomic identifier.



Taxonomic identifier	9606 [NCBI]

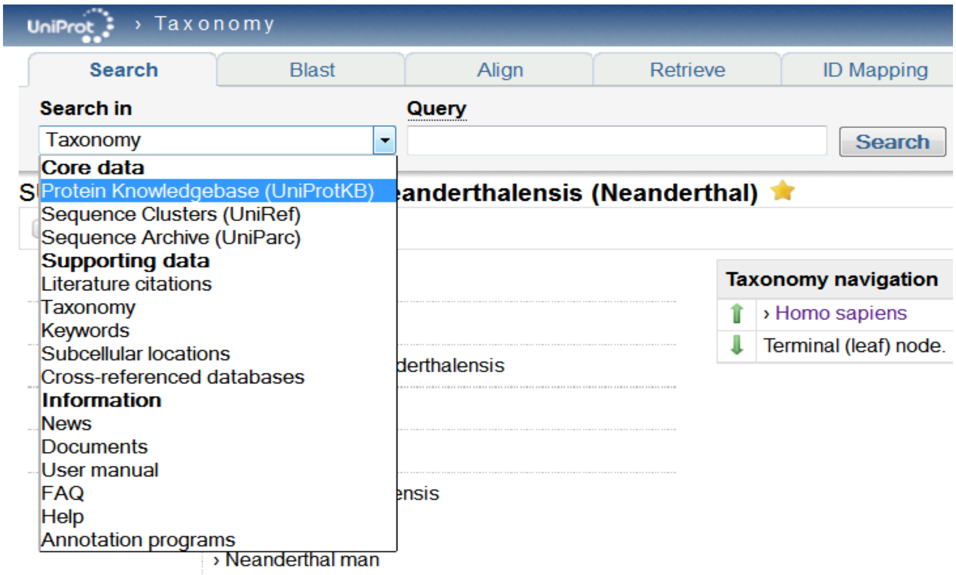
On the new page there is Taxonomic navigation table to go up and down the taxonomic lineage:



Taxonomy navigation	
↑	› Homo
↓	› Homo sapiens neanderthalensis

Q5. In UniProtKB how many proteins are there from Homo sapiens neanderthalensis?

✍ In the Search tab, drop down the Search in menu and select Protein Knowledgebase (UniProtKB).

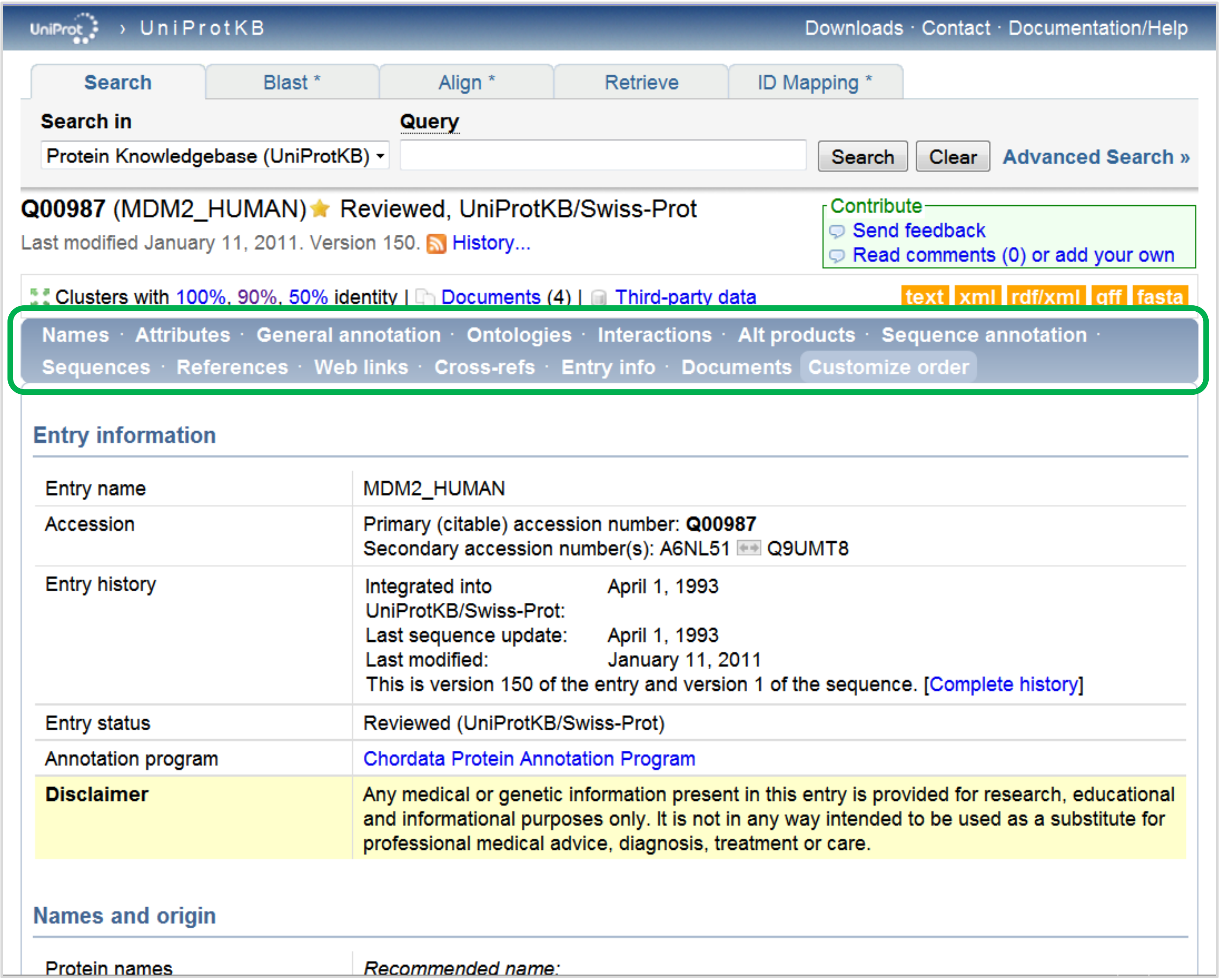


Taxonomy navigation	
↑	› Homo sapiens
↓	Terminal (leaf) node.

Type mdm2 in to the query box and hit Search.

Open the UniProtKB/Swiss-Prot entry Q00987.

At this point I’ll describe the layout of a UniProtKB entry. The annotation of an entry is divided up to in to sections. Links to these sections are provided in a grey bar. The bar remains at the top as you scroll down. Furthermore, in the bar is a link to customize the order of the sections. By using a cookie the website remembers which order you prefer and shows all entries in that order.


UniProt › UniProtKB
Downloads · Contact · Documentation/Help
Search
Blast *
Align *
Retrieve
ID Mapping *
Search in
Query
Protein Knowledgebase (UniProtKB)
Search
Clear
Advanced Search »
Contribute
Send feedback
Read comments (0) or add your own
Clusters with 100%, 90%, 50% identity | Documents (4) | Third-party data
text xml rdf/xml gff fasta
Names · Attributes · General annotation · Ontologies · Interactions · Alt products · Sequence annotation · Sequences · References · Web links · Cross-refs · Entry info · Documents Customize order

Q00987 (MDM2_HUMAN) ★ Reviewed, UniProtKB/Swiss-Prot

Last modified January 11, 2011. Version 150. History...

Entry information

Entry name	MDM2_HUMAN
Accession	Primary (citable) accession number: **Q00987** Secondary accession number(s): A6NL51 Q9UMT8
Entry history	Integrated into UniProtKB/Swiss-Prot: April 1, 1993 Last sequence update: April 1, 1993 Last modified: January 11, 2011 This is version 150 of the entry and version 1 of the sequence. [Complete history]
Entry status	Reviewed (UniProtKB/Swiss-Prot)
Annotation program	Chordata Protein Annotation Program
Disclaimer	Any medical or genetic information present in this entry is provided for research, educational and informational purposes only. It is not in any way intended to be used as a substitute for professional medical advice, diagnosis, treatment or care.

Names and origin

Protein names	*Recommended name:*

Figure 2. The UniProtKB Navigation bar

First of all we're going to search using the amino acid sequence of this protein to find proteins that have the same job in other organisms (likely to be orthologs) and see how the sequences compare

For this exercise, we will use the BLAST program on the UniProt website. BLAST (Basic Local Alignment Search Tool), finds regions of sequence similarity and gives functional and evolutionary clues about the structure and function of your novel sequence.

☞ Click on the BLAST tab to search for proteins similar to Q00987 in UniProtKB/Swiss-Prot.

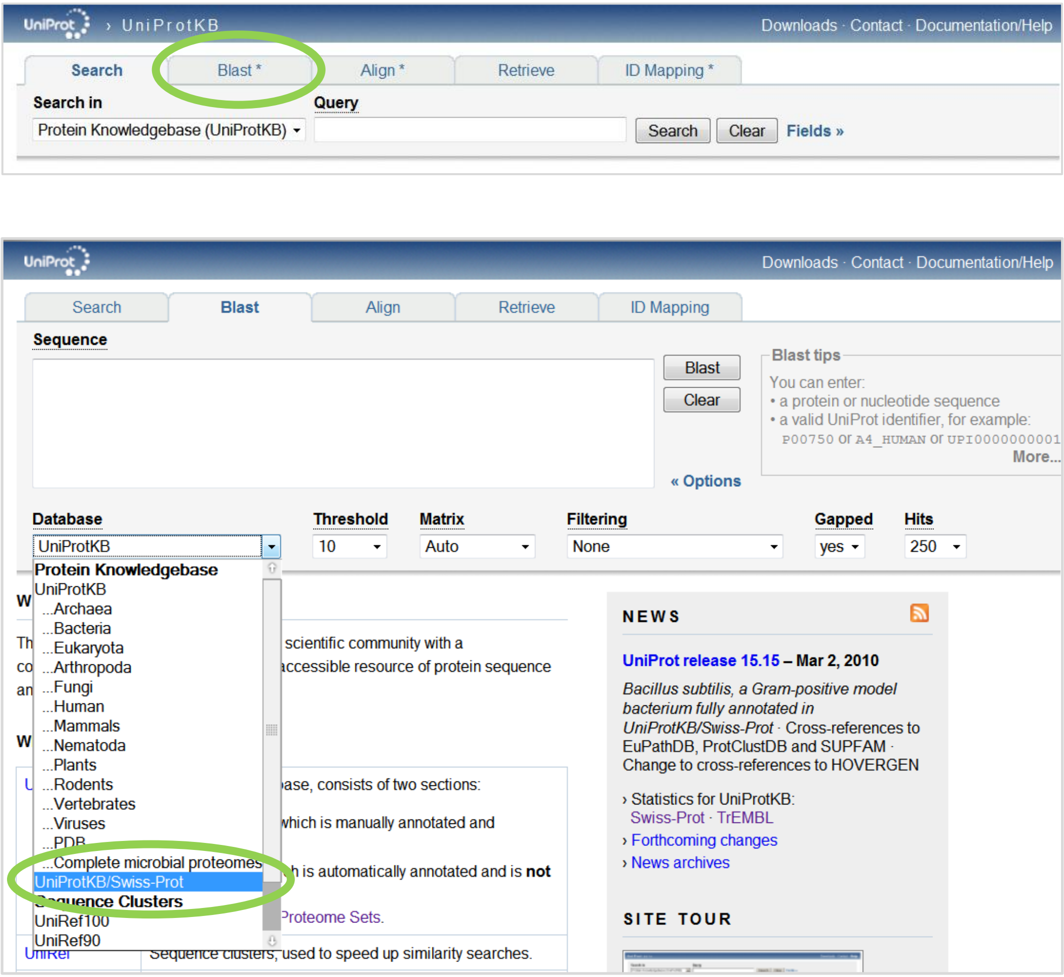

UniProt › UniProtKB
Downloads · Contact · Documentation/Help
Search
Blast *
Align *
Retrieve
ID Mapping *
Search in
Query
Protein Knowledgebase (UniProtKB)
Search
Clear
Fields »
UniProt
Downloads · Contact · Documentation/Help
Search
Blast
Align
Retrieve
ID Mapping
Sequence
Blast
Clear
« Options
Blast tips
You can enter:
• a protein or nucleotide sequence
• a valid UniProt identifier, for example: P00750 or A4_HUMAN or UPI0000000001
More...
Database
Threshold
Matrix
Filtering
Gapped
Hits
UniProtKB
10
Auto
None
yes
250
Protein Knowledgebase
UniProtKB
...Archaea
...Bacteria
...Eukaryota
...Arthropoda
...Fungi
...Human
...Mammals
...Nematoda
...Plants
...Rodents
...Vertebrates
...Viruses
...PDB
...Complete microbial proteomes
UniProtKB/Swiss-Prot
Sequence Clusters
UniRef100
UniRef90
NEWS
UniProt release 15.15 – Mar 2, 2010
Bacillus subtilis, a Gram-positive model bacterium fully annotated in UniProtKB/Swiss-Prot · Cross-references to EuPathDB, ProtClustDB and SUPFAM · Change to cross-references to HOVERGEN
› Statistics for UniProtKB: Swiss-Prot · TrEMBL
› Forthcoming changes
› News archives
SITE TOUR

Figure 3. The UniProt BLAST tool.

☞ Tick the check boxes next to the entries for MDM2_HUMAN (Q00987, not the isoform entry Q00987-11), MDM2_HORSE (P56951) and

MDM2_MOUSE (P23804) and in the green bar at the bottom of the page click on the option to Align the sequences.

In the alignment page scroll down to the Sequence annotation (Features) section and tick the box by Zinc Fingers.

Would you say that the zinc finger domains are well conserved? This is indicative of proteins having a similar function. If you click on the tick box for Mutagenesis you'll see that many of the amino acids in the second zinc finger domains have been swapped. Now to go the UniProtKB entry for MDM2_HUMAN to see what affect this has on the activity

Click on the hyperlinked entry Q00987 to go to that protein, then use the navigation bar to get to the Sequence annotation section and look at the Mutagenesis lines in the Experimental Info area.

Notice what enzyme activity has been affected.

UniProtKB Annotation.

In the Names and origin section click on the EC number link in the protein names field.

Q6. What kind of enzyme is denoted by this EC number?

Return back to the UniProtKB website.

Q7. Use the advanced search to find out how many UniProtKB/Swiss-Prot entries have been curated (reviewed) as belonging to this enzyme category?

(Hint – there is a field in the advanced search for EC number searching use the term 6.3.2.-. Do this search then modify the results to get the number in Swiss-Prot)

Q8. How many UniProtKB/TrEMBL (unreviewed) entries have been automatically annotated as belonging to this enzyme category?

Return to the MDM2_HUMAN (Q00987) entry and browse to answer the questions below, most of the information will be in the General Annotation (Comments) section.

Q9. How many **species** have proteins which are members of the MDM2/MDM4 family?

(Hint – once you've clicked on the family link look at the different browsing options, where you found the pathway link. Another hint – expand the taxonomy and just count the end nodes with animal names at the end).

From the screen with the pie chart click on the 2 next to Homo sapiens (Human) then click on Q00987 to get back to our example entry, MDM2_HUMAN.

Q10. What induces this protein?

Q11. How many experiments have reported an Interaction with this protein and the Human protein TP53?

Q12. Which isoform has been reported not to interact with TP53 [Hint: in Subunit structure comment]?

Q13. Which type of post-translation modification results in proteasomal degradation?

Go to the Cross references (Cross-refs) section to answer the following question.

Q14. How many PDB structures have been reported using NMR?

Go to the Ontologies section to answer the following question.

Q15. Which Molecular Function in the Gene Onotolgy has been added by a UniProt Curator using reference 15?

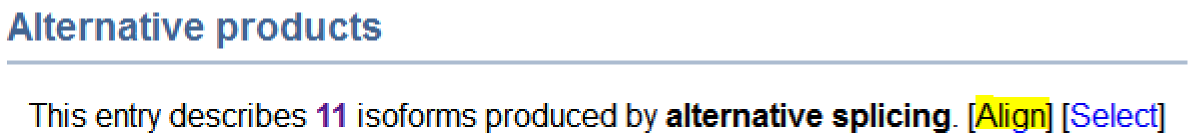
Alternative Products in UniProtKB

This part of the tutorial shows the benefits of a manually curated database that collates experimental results reported in the literature. By analyzing the missing regions of isoforms and using the information in UniProtKB/Swiss-Prot entries interesting biological hypothesizes can be deciphered.

In UniProtKB entry Q00987 go to the Alternative Products section.

Mdm2 is a protein for which multiple (11) isoforms have been identified. All of these are mapped within UniProtKB, and given stable identifiers.

Press the button "Align" to see how all the isoforms differ.



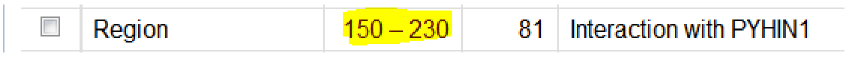
Alternative products

This entry describes **11** isoforms produced by **alternative splicing**. [Align] [Select]

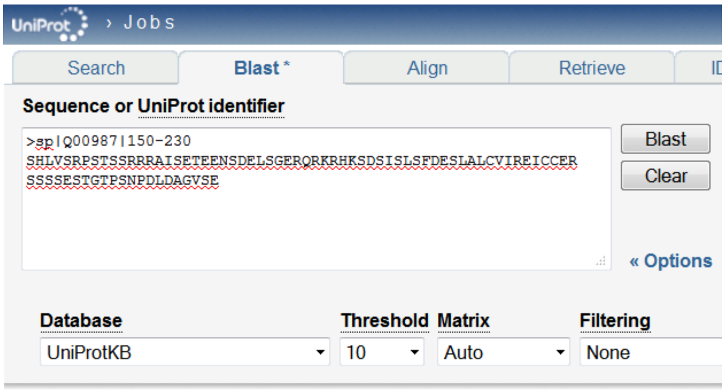
As before click the tick box by Zinc finger.

Q16. (i) Does isoform e (or isoform 7 (Q0087-7)) have the zinc finger regions?
(ii) Do you think isoform e will have ubiquitin ligase E3 activity?

Click on the Q00987 to go back to the main entry. Use the navigagtion bar to go to the Sequence annotation section. Click on the region for the interaction with PYHIN1.



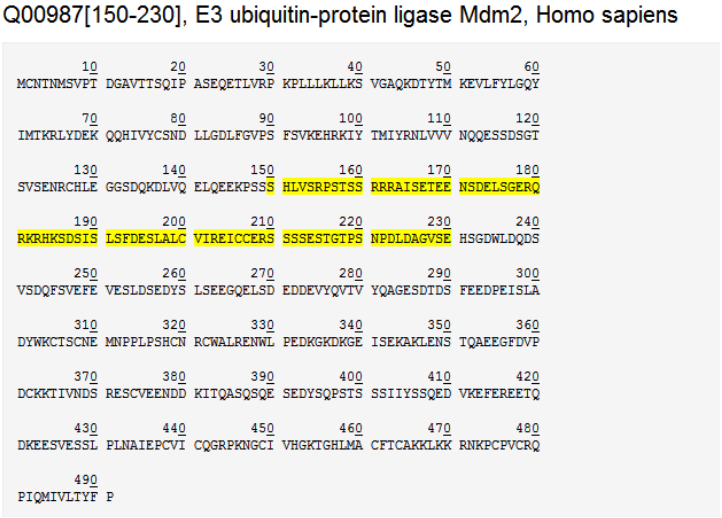
☐	Region	150 – 230	81	Interaction with PYHIN1



Sequence or UniProt identifier

```
>sp|Q00987|150-230
SHLVSRPSTSSRRRAISETEENSDELSGERQRKRHKSDSISLSFDESLALCVIREICCER
SSSSESTGTPSNPDLDAGVSE
```

Database: UniProtKB
Threshold: 10
Matrix: Auto
Filtering: None



Q00987[150-230], E3 ubiquitin-protein ligase Mdm2, Homo sapiens

```
        10         20         30         40         50         60
MCNTNMSVPT DGAVTTSQIP ASEQETLVRP KPLLLKLLKS VGAQKDTYTM KEVLFYLGQY

        70         80         90        100        110        120
IMTKRLYDEK QQHIVYCSND LLGDLFGVPS FSVKEHRKIY TMIYRNLVVV NQQESSDSGT

       130        140        150        160        170        180
SVSENRCHLE GGSDQKDLVQ ELQEEKPSSS HLVSRPSTSS RRRAISETEE NSDELSGERQ

       190        200        210        220        230        240
RKRHKSDSIS LSFDESLALC VIREICCERS SSSESTGTPS NPDLDAGVSE HSGDWLDQDS

       250        260        270        280        290        300
VSDQFSVEFE VESLDSEDYS LSEEGQELSD EDDEVYQVTV YQAGESDTDS FEEDPEISLA

       310        320        330        340        350        360
DYWKCTSCNE MNPPLPSHCN RCWALRENWL PEDKGKDKGE ISEKAKLENS TQAEEGFDVP

       370        380        390        400        410        420
DCKKTIVNDS RESCVEENDD KITQASQSQE SEDYSQPSTS SSIIYSSQED VKEFEREETQ

       430        440        450        460        470        480
DKEESVESSL PLNAIEPCVI CQGRPKNGCI VHGKTGHLMA CFTCAKKLKK RNKPCPVCRQ

       490
PIQMIVLTYF P
```

The region which is involved in the interaction with PYHIN1 is highlighted in yellow. Also the sequence that corresponds to the region has been placed in the BLAST text box.

🖱 **In the Database drop-down menu select Human and then click the Blast button**

Q17. Looking at the results of the BLAST search can you see Q00987-6 (isoform d), thus do you think this isoform would interact with PYHIN1?

🖱 **Click on Q00987 to get back to the entry page for MDM2_HUMAN. Using the navigation bar go to the General annotation section.**

Q18.What tissues are isoforms d and e found?

Q19.This protein has been shown to be over-expressed in which types of tumours?

🖱 **Navigate to the Web links section.**

Q20. Finally, looking in the Web resources section, which famous encyclopaedia website has information on this protein?

The Answers

Exploring UniProtKB – searching, BLAST & align.

Q1 (i) Using the simple search do a search for Human. Find out how many UniProtKB entries are hit?

A1 (i). 1,049,509

Q1 (ii) To find the number of Human proteins in UniProtKB do another search for human but this time search exclusively in the Species fields by using the Advanced Search link.

A1 (ii). 115,694.

Q2. How many Homo sapiens (Human) entries are in UniProtKB/Swiss-Prot?

A2. 20,256.

Q3. Why do you think there is such a big difference between the answers answer for Q 1 ii and Q 2?

A3. There are two parts to UniProtKB; the TrEMBL part and the Swiss-Prot part. Proteins arrive in TrEMBL, the unrevised section of UniProtKB, which contains multiple entries for the same protein resulting in redundancy. This is because UniProtKB entries for the same protein which come from different sources need to be kept separate so they can be merged manually. Swiss-Prot is the part of UniProtKB that contains just the entries that have been manually curated (reviewed). In Swiss-Prot all duplicate entries for a protein have been merged and annotated accordingly. The number of entries in Swiss-Prot is much lower than the rest of UniProtKB (TrEMBL) due to this merging and also because all Swiss-Prot entries have to be analyzed by a Curator which takes time.

Q4. How many UniProtKB entries are there in the Human Complete Proteome Set?

A4. 56,582.

There are over double the number of entries in the Human Complete Proteome Set than in UniProtKB/Swiss-Prot Human because many protein entries in the complete proteome are waiting to be annotated and so reside in UniProtKB/TrEMBL. Many of these protein entries are isoforms or sequence variants which have to be merged manually by a Curator.

Q5. In UniProtKB how many proteins are there from Homo sapiens neanderthalensis?

A5. 23.

Q6. What kind of enzyme is denoted by this EC number?

A6. Ligase.

Q7. Use the advanced search to find out how many UniProtKB/Swiss-Prot entries have been curated (reviewed) as belonging to this enzyme category?

A7. 4,921 results for ec:6.3.2.- AND reviewed:yes in UniProtKB.

Q8. How many UniProtKB/TrEMBL (unreviewed) entries have been automatically annotated as belonging to this enzyme category?

A8. 24,974 results for ec:6.3.2.- AND reviewed:no in UniProtKB.

Q9. How many **species** have proteins which are members of the MDM2/MDM4 family?

A9. 10.

Q10. What induces this protein?
A 10. DNA damage.

Q11. How many experiments have reported an Interaction with this protein and the Human protein TP53 [Hint: in Interactions section]?
A11. 25.

Q12. Which isoform has been reported not to interact with TP53?
A12. Isoform f.

Q13. Which type of post-translation modification results in proteasomal degredation?
A13. Auto-ubiquitinated.

Q14. How many PDB structures have been reported using NMR?
A14. 4.

Q15. Which Molecular Function in the Gene Onotolgy has been added by a UniProt Curator using reference 15?
A15. Ubiquitin-protein ligase activity.

Q16. (i) Does isoform e (or isoform 7 (Q0087-7)) have the zinc finger regions? (ii) Do you think isoform e will have ubiquitin ligase E3 activity?
A16. (i) No. (ii) No.

Q17. Bearing in mind the sequence missing from isoform d (Q00987-6) do you think this isoform would interact with PYHIN1?
A17. No isoform d is missing from the search results indicating that it does not have the binding region for PYHIN1.

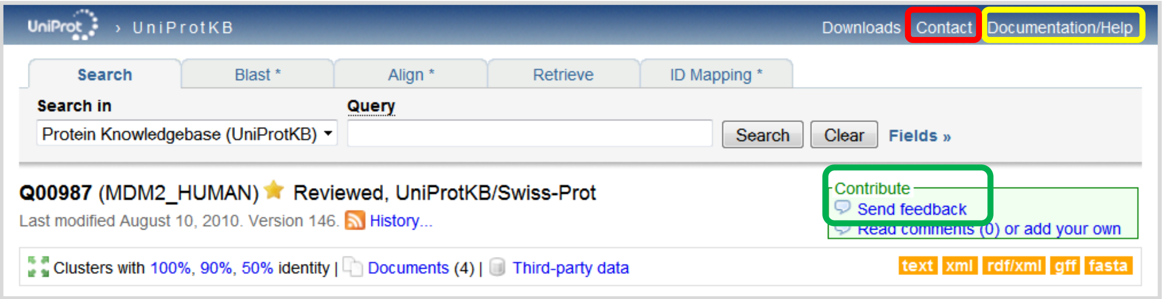
Q18. Looking in the General annotation section, what tissues are isoforms d and e found?
A18. Isoforms Mdm2-D and E are observed in a range of cancers but absent in normal tissues.

Q19.This protein has been shown to be over-expressed in which types of tumours?
A19. Soft tissue sarcomas, osteosarcomas and gliomas.

Q20. Finally, looking in the Web resources section, which famous encyclopedia website has information on this protein?
A20. Wikipedia.

I hope having done this introductory tutorial you feel that you now know a bit more of what UniProtKB has to offer; what information UniProtKB entries can contain and where it can be found. If you have questions or comments about this tutorial please feel free to contact us at help@uniprot.org.

Alternatively, general UniProtKB feedback and help can be sent using the Contact form (highlighted red) or the Send feedback link from each entry (highlighted green). For further information please browse the Documentation/Help section (highlighted yellow).



Search in: Protein Knowledgebase (UniProtKB)
Query:

Q00987 (MDM2_HUMAN) ★ Reviewed, UniProtKB/Swiss-Prot
Last modified August 10, 2010. Version 146. History...

Clusters with 100%, 90%, 50% identity | Documents (4) | Third-party data

Best regards,

Duncan Legge & the UniProt team.