EMBL-EBI Tutorial

Jennifer McDowall (V1.00, June 2008)

UniProt - *the protein sequence database*

http://beta.uniprot.org

UniProt is the world's most comprehensive catalogue of information on proteins. It is a central repository of protein sequence and function created by joining the information contained in UniProt/Swiss-Prot, UniProt/TrEMBL and PIR.

UniProt is comprised of three components. The *UniProt Knowledgebase* (UniProtKB) is the central access point for extensive curated protein information, including function, classification and cross-reference. The *UniProt Reference cluster* *UniRef) databases combine closely related sequences into a single record to speed searches. The *UniProt Archive* (UniParc) is a comprehensive repository, reflecting the history of all protein sequences. The *UniProt Metagenomic and Environmental Sequences* (UniMES) database is a repository specifically developed for metagenomic and environmental data.

This tutorial will introduce you to the wealth of annotated protein data available within the UniProt database, how to extract this information, and how to use the tools associated with UniProt to align and analyse protein sequences as well as to perform sequence searches using the web interface.

Exercise 1:

Searching UniProt using a text search

UniProt can be searched in a number of different ways. The Text Search allows the database to be searched using keywords, similar to how one searches *Google* (logical operators such as "and" and "butnot" can be used to restrict search).

ACCESS THE INTERPRO DATABASE AT: **http://www.uniprot.org/**

- From the UniProt homepage (http://www.uniprot.org/), type in "Myosin light chain kinase" in the "Query" box at the top of the page. Make sure the "Search in" box states "Protein Knowledgebase UniProtKB)". Click on the "Search" button.

? How many results did you retrieve?

To cut down on the volume of results, we can restrict our search:

- Click on the "Fields" tab on the right-hand side of the query box.

You should now have expanded search boxes.

- From the "Fields" box drop-down menu, select "Organism [OS]". In the "Term" box, type "Human".

A drop-down box should appear.

- Select "Human [9606]".

This selects only Homo sapiens; otherwise, if simply "Human" was put down, the search would retrieve all entries mentioning the word human (as opposed to human proteins), including human viruses.

- Click on "Add & Search".

Note: the "Query" box now has the extra search term added.

- Click on the "Search" button.

Now we only have myosin light chain kinases that occur in humans.

Note: UniProt/SwissProt entries have a gold star and UniProt/TrEMBL entries have a grey star.

? Now how many results did you retrieve?

Exercise 2:

Exploring a UniProt/SwissProt entry: General Information

- Click on the hyperlink to Q15746 to view the UniProt entry.

This is a UniProt/SwissProt entry (denoted by its gold star), which has been manually annotated by a curator.

- Scroll down to the "Entry Information" section. Click on "Show" on the grey bar if it happens to be collapsed.

You can then move the section to the top of the page by depressing your left mouse button over the grey bar and dragging it to the top.

? How many TrEMBL entries went to make up this SwissProt entry?

HINT: count both primary and secondary accession number.

- Collapse the "Entry Information" section using the "Hide" tab on the right of the grey bar (this may not work on all browsers).
- Scroll down to the "Names and Origin" section (you can try moving this section to the top of the page, however this will only work in certain browsers).

? What other names is this protein known by?

An "EC" number tells you the protein has catalytic activity.

- Click on "EC 2.7.11.18".

This takes you to the ENZYME database.

? Are there additional names for this enzyme in the ENZYME database?

Note: the names given in the UniProt/SwissProt entry are specific to a particular species (here, Homo sapiens), whereas the additional names in the ENZYME database are collectively for all species.

? What reaction is catalysed by this enzyme?

Go back to the Q15746 entry page.

Some proteins in the UniProt database are predicted from analysing the sequence, while others have direct evidence that they really do exist.

? Is there any evidence that this protein exists at the protein level?

Collapse the "Names & Origin" section, if you can.

Exercise 3:

Exploring a UniProt/SwissProt entry: Sequence

Scroll down to the "Sequences" section.

We can compute the molecular weight (MW) and isoelectric point of our protein.

Using the scroll-down menu in the "Tools" box, select "Compute pI/MW" and press "Go". Click on "CHAIN 1-1914".

? What is the molecular weight and isoelectric point of this protein?

Go back to the Q15746 entry page and scroll down to the "Isoforms" information.

? How many different isoforms are known for this protein?

Let's try aligning some of the different isoforms.

Check the left-hand box for "Isoform 1" (main sequence) and for "Isoform 5". On the green bar at the foot of the screen, click on "Align".

? Are there any sequence differences between the two isoforms?

Let's look at some of the sequence features available for this protein.

Go back to the Q15746 entry page. Scroll down to the "Sequence annotation (Features)" section. Look at the "Regions" characterised for this protein.

? What are the residue positions of the "Fibronectin type-III" domain?

Look at the "Sites" characterised for this protein.

? At what position is the active site proton acceptor, and what amino acid is it?

HINT: after clicking on either the "graphical view" or the "position" data, you may need to scroll down the sequence to see it.

- Look at the "Amino acid modifications" characterised for this protein.

? How many residues are modified in this protein, and how are they modified?

- Look at the "Natural variations" characterised for this protein. Find the information for "Isoform 5".

? Now can you summarise the sequence differences between isoforms 1 (main sequence) and isoform 5 more easily than from the alignment?

Let's look at the alternative products this protein can produce.

- Scroll down to the "Alternative products section.

This section gives additional information about the isoforms.

? What other name is "Isoform 5" known by?

? What alternative initiation site does it use?

Exercise 4:

Exploring a UniProt/SwissProt entry: Structure

- Scroll down to the "Secondary structure" section. Expand the section by clicking on "Details".

? Is this protein primarily composed of beta-strands or alpha-helices?

- Scroll down to the "Cross-references" section. Find the "3D structure databases" subsection.

This section gives both experimentally determined structures from the PDB, and predicted structures from the homology database ModBase, where they exist.

- Set the drop-down menu to PDB (usually default), click on "2CQV".

This takes you to the Protein Data bank at the RCSB. On the right-hand side, there should be a picture of this protein.

- Go back to the Q15746 entry page. Set the drop-down menu to MSD and click on "2CQV".

This takes you to the Molecular Structure Database at the EBI. The primary information is the same, but these databases have different tools for analysing the structure.

- Go back to the Q15746 entry page. To look at the homology models if this protein, click on "Search" under the "ModBase" information.

As there is only experimentally determined structure for a small region of this protein (for 114 residues (positions 1238-1338) - *you can find this out by looking under the "Sequence Details" tab in the PDB page for 2CQV*), it is useful to see what the predicted structure is for the rest of the protein.

- Go back to the Q15746 entry page.

Exercise 5:

Exploring a UniProt/SwissProt entry: General Annotation

- Scroll down to the "Ontologies" section.

These terms can be used to get a rough idea of what this protein does. In addition, the GO terms use controlled vocabulary and are very useful for accurately cross-comparing data between databases.

? What ligands might this protein bind according to the keywords list?

? According to the "GO" information, what "Biological process" is this protein involved in?

- Expand the GO information by clicking on "Complete GO annotation".

This gives you all the evidence used to assign the various GO terms.

- Go back to the Q15746 entry page. Scroll down to the "General annotation (Comments)" section.

General annotation is manually curated from the literature.

? We already know this protein is an enzyme, but what is its function?

? We already know from the "Sequence annotation" section that this protein has amino acid modifications. Using the information given here, how do these post-translational modifications affect this protein?

- Scroll down to the "Sequence similarities" subsection.

This tells you what family/superfamily this protein belongs to, and what domains it shares with other proteins.

? What domains does this protein have?

- Click on "fibronectin type-III domain".

This gives you a list of all the other proteins in UniProtKB that also have a "fibronectin type-III" domain.

- Go back to the Q15746 entry page.

Exercise 6:

Searching UniProt using a BLAST search

UniProt can also be searched using BLAST (Basic Local Alignment Search Tool), which takes a protein or nucleotide sequence (which is translates) and compares it with those contained in the UniProt database.

Use BLAST to find the proteins with the closest sequence identity to the protein Q15746.

- On the grey section at the very top of the page, click on the "BLAST" tab.

Note: the sequence for our protein (Q15746) should already be inserted into the BLAST box.

- Click on the "Options" button.

The default setting is for the UniProtKB database, but the search can be restricted to a specific organism, just to SwissProt entries, or to environmental samples.

- Leave at the default settings and press "Blast" button.

Exploring the BLAST results

The results page will provide a list of proteins that match the query sequence from the UniProt databases ordered by their scores, along with their description, length and scoring information.

? Which UniProtKB accession most closely matches this query sequence (ignoring the isoforms of Q15746)?

? What is the closest matching UniProt/Swiss-Prot entry (gold star), and what % identity does it have with this query sequence?

Looking at sequence alignments

Alignments can be made between any of the proteins in the search results.

- On the BLAST results page, pick one protein and click on the green bar under the heading "local alignment".

The alignment shows a perfect match between your selected protein and our query protein Q15746.

- GO BACK to the BLAST results page. Check the left-hand boxes for Q15746 and two other proteins, and then click on "Align" button on the green strip at the bottom of the page.

The alignments should show the match between your selected proteins. Matching residues are identified by "*" below the alignment, while similar amino acids (e.g. leucine vs. methionine) are identified by ".", and mismatches are shown by a space.

? Can you identify regions of high and low conservation?

- Scroll down to the bottom of the page, below the aligned sequences. On the grey bar marked 'Amino Acid Properties', click on 'Show' at the right-hand side (if it reads 'Hide', then the information is already displayed).

The 'Amino Acid Properties' section allows you to view differences in properties between aligned sequences. The default display is to view 'Conserved Regions'.

- Click on a couple of other properties, such as 'Hydrophobic' or 'Polar'.

? Do these proteins have distinct regions of hydrophobic or polar residues?

- Uncheck these properties when you are finished, so only 'Conserved Regions' is checked.
- Scroll down to the "Sequence annotation" section and click on "Show".

? Which sequence annotation features are available for these proteins?

? Can you find their positions in the alignments?

HINT: have only one property checked at a time; otherwise they may obscure each other from view.

- GO BACK to the BLAST results page.

Exercise 7:

Exploring a UniProt/TrEMBL entry

- On the grey section at the very top of the page, click on the "Search" tab. In the "Query" box, type in Q9VCL7 and click on the "Search" button.

This is a UniProt/TrEMBL entry (denoted by its grey star), which has only automatic annotation and additional cross-referencing added.

- Scroll down to the "Names & Origins" section.

The gene name for Q9VCL7 is an Orf (open reading frame) clone name, which will probably change when the gene/protein is characterised. However, there is evidence that the protein is expressed.

- Scroll down to the "Ontologies" section.

NOTE: there are only Keywords in a TrEMBL entry, as it has not yet been manually curated; however, these terms become very useful when there is less annotation available.

- Scroll down to the "Binary interactions" section.

Looking at protein interaction partners can provide valuable information on the function of unknown proteins, especially if the partners have annotated UniProt/SwissProt entries.

? What is the name of the protein with which Q9VCL7 interacts?

HINT: look in the first column under "With".

- Scroll down to the "Cross-references" section. Find the "Organism-specific databases" subsection.

Organism-specific databases, such as FlyBase, can provide additional information. FlyBase is a comprehensive database for information on the genetics and molecular biology of *Drosophila*, and includes data from the Drosophila Genome Project and data curated from the literature.

- Find the "Family and domain databases" subsection.

Using InterPro, we can look at the predicted domain organisation of the protein, as well as its family relationships with other proteins.

? How many InterPro entries are associated with this protein?

Using UniProt, we can uncover a lot of information about a protein in addition to its sequence, even for UniProt/TrEMBL entries; this includes functional annotation, sequence annotation, structure and isoform information, and much more.

This is the end of the short tour of the UniProt database, available at the EBI. Perhaps you might like to try it again with a more relevant sequence.