Universal Protein Resource (UniProt)

http://www.uniprot.org

UniProt is probably the most comprehensive and up-to-date collection of protein sequence and annotation data. It should be the first stop for any researcher looking for the available information about a protein, as it is very comprehensive and saves the effort of integrating data from multiple sources.

Its major database is the UniProt KnowledgeBase (UniProtKB), where the Swiss-Prot and TrEMBL data sets are kept. Swiss-Prot is a manually annotated and reviewed protein resource where high-quality information can be found. TrEMBL is a compendium of automatically annotated information on proteins that were mostly (but not exclusively) obtained from the translation of nucleotide coding sequences (CDS) available at GenBank. TrEMBL provides the raw data for Swiss-Prot curators to revise; therefore, while TrEMBL has more entries than Swiss-Prot, it lacks the expert manual annotation.

UniProt is also home of the UniProt Archive (UniParc), a non-redundant database of almost all publicly available protein sequences in the world, and the UniProt Reference Clusters (UniRef), collections of UniProt sequences clustered at specific sequence similarity thresholds (such as 90% for UniRef90).

Some useful things to know about UniProt:

- The UniProt website allows browsing of its data sets by clicking the tiles on the homepage and through the links in the footer of every webpage. Searches can be performed using the top search bar at the homepage, with a drop-down menu to select the target database and an 'Advanced' option for further refinement. It also provides direct access to perform similarity searches on the UniProtKB database with BLAST and pairwise alignments with Clustal Omega.

- The 'Retrieve/ID mapping' link at the homepage can be used to search for a list of protein identifiers and retrieve their individual UniProt entries at once. It also serves for converting the input identifiers to their equivalent in GenBank, PDB and many other external databases.

- Searching UniProtKB leads to the results page where selected information about the retrieved sequences is presented in table format. A 'Column' button on top allows customization of the different fields shown. The results can be limited to a certain database, organism, search term, etc. using the 'Filter by' options on the left-hand bar. There is also a 'View by' section to cluster the results by taxonomy, pathway or other criteria.

- By default, UniProt entries are displayed in the 'Entry' view on the left-hand bar, with links to the different entry sections below. (Click on the title of each section for more information.) The 'Feature viewer' is a useful alternative that provides an overview of all annotations along the protein's sequence.

- Each UniProt entry can be referred to two unique identifiers. The 'Accession number' is a sequence of 6 to 10 alphanumerical characters (e.g. P04637) that is kept in all updates and thus should be used in all publications. The 'Entry name' is a more readable ID often selected to reflect biological properties like the protein name or the organism (e.g. P53_HUMAN), which may change if more information about the entry or related sequences becomes available.

- An UniProt entry can have both experimental and predicted data. Manual assertions, taken from published experiments, transferred from experiments in similar proteins or imported from other databases, are coloured in gold. Data retrieved by the automatic annotation process is coloured in blue.

- The annotation content of an UniProt entry is ranked by a discrete 5-stars system. The score of each entry is computed from the presence or the number of annotation types. More stars reflect a more extensively annotated entry; however, the system does not represent annotation correctness.

- UniProt is updated every four weeks. The latest data sets can be retrieved by following the 'Download latest release' link on the 'UniProt data' section of the homepage.

Exercises

1. Do you have a protein you are working on? Try to access its UniProt entry by browsing or searching the database. You can use your protein to perform most of the following exercises too.

2. Search for **CDC7**. ¿What is the name of this protein? ¿How many entries exist in different organisms, and how many in human? How many of these are high-quality entries?

3. When searching for **CDC7**, how many of the resulting entries are the product of the cdc7 gene? (*Hint*: use the filtering tools at the left-hand bar.) Why is Q8NEY8 among the results? What about B1AMW7?

4. **Lysine-specific demethylase 3B** exists in human (Q7LBC6) and mice (Q6ZPY7). Although its function is known, it has been annotated from different sources in each organism. Can you identify these sources?

5. Go to the UniProt entry for human **Lysine-specific demethylase 3B**. Inspect the different annotated features available (fields can be shown or hidden with the left-hand bar) and then answer the following:

   a. What is the entry name?

   b. Which UniProtKB database does this entry belong to?

   c. Which are some of the molecular functions and biological activities associated with this protein?

   d. Where is this protein located? Where does this information come from?

   e. How many proteins are known to interact with it?

   f. Does it have any protein interaction motif?

   g. Is there any known structure of this protein? Does it comprise the whole sequence?

   h. How many isoforms are annotated of this protein? How do they differ from the canonical sequence?

   i. Can you find when was this entry created and when is the date of its last modification? How many times was it modified?

   j. Find out which information is available for position 773 at the main 'Entry' page. Then, go to the 'Feature viewer' and inspect the same position. (*Hint*: the sliding semi-circles at the top can be used to focus the display on the desired region). You will probably find some extra information; what does it tell you about your protein? Now that you are aware of this information, can you find it in the main 'Entry' page too?

6. The UniProt entry Q9UBU3 corresponds to **GHSR**, an appetite-regulating hormone. It is expressed as a preprotein that is later cleaved into ghrelin and obestatin. Can you locate the position of these mature proteins in the full-length preprotein? (*Hint*: the 'Feature viewer' can be helpful). Do they have any natural sequence variant? How many proteins share 90% identity with the full-length protein?

7. Isoform 6 of **Prelamin-A/C** (P02545), also known as progerin, has a mutant variant associated with disease. What is the name of this syndrome? How does this variant differ from the canonical sequence?

8. **Lysine-specific demethylase 5C** (P41229) is another human histone demethylase that specifically targets the lysine 4 of histone H3. Can you find the Kcat and Km of this reaction on its UniProt entry?

9. Can you download the sequence of the **Lysine-specific demethylase 5C** isoform 1 in FASTA format? (Briefly, a FASTA file starts with the ">" (greater-than) symbol plus a brief description of the sequence, followed by the actual sequence in standard one-letter code in the second line. Multi-FASTA files have many of these combinations, one after another and with an optional blank line after each sequence).

10. UniProt entries have a specific section named 'Family & Domains'. It describes the identity, position and length of domains that were annotated in the protein. A domain is defined as 'a specific combination of secondary structures organized into a characteristic three-dimensional structure or fold.' How many of these domains can you find in **Lysine-specific demethylase 5C**? Are there any other relevant regions of the proteins described? If so, can you find references to these domains cross-linked from other databases?

BLAST

http://blast.ncbi.nlm.nih.gov

BLAST (acronym of Basic Local Alignment Search Tool) is arguably the most important contribution of Bioinformatics to science. Over the last 20 years it has been the *de facto* tool for finding similar sequences to a protein or nucleotide sequence of interest, a useful first step towards its characterization.

As its name suggests, BLAST perform local alignments. Contrary to global approaches, which are designed to identify the overall similarity between two or more sequence from start to end, a local alignment can find segments of similarity even if they appear on different order in the compared sequences. Given that sequence conservation is a usual indicator of significant functional similarity, the locally aligned sequences may correspond to conserved motifs, domains or other biologically relevant regions. Local conservation that extends to full-length sequences may indicate homology (evolution from a common ancestor).

The success of BLAST lies in its algorithm. Briefly, it splits the input sequence in small fragments (or 'words') of defined size and scans a database of sequences to find all identical occurrences of each word. When a match is found, BLAST extends the comparison forwards and backwards, allowing gaps until the score falls below a certain threshold. (This score is computed from a substitution matrix, like BLOSUM62, penalizing inclusion of gaps.) All similar sequences are presented in the results by increasing '*E*-value', an estimation of the number of times you may expect to find a similar or better alignment score just by chance.

The speed of the BLAST algorithm serves well for an efficient search within huge databases, such as those of all known proteins and nucleotides. The National Center for Biotechnology Information implements different versions of BLAST in its dedicated web server:

- BLASTN can be used to search nucleotide databases for similarity to a nucleotide query. It's the only BLAST tool that aligns nucleotides to nucleotides.
- BLASTP can be used to search protein databases for similarity to a protein query.

It is also possible to use translated nucleotide sequences to look for potential coding regions with similarity:

- BLASTX takes a nucleotide query, translates it in the six possible frames and uses these to search protein databases.
- TBLASTN uses a protein query to search all possible translated sequences from a nucleotide database.
- TBLASTX translates a nucleotide query in all six frames and searches a translated nucleotide database.

The NCBI BLAST web server provides many databases to search for similarity, such as the (mostly) non-redundant and comprehensive 'nr' (protein) and 'nt' (nucleotide) databases; the 'refseq' data set of genomic, transcript and protein sequences; the 'pdb' database of protein sequences with known structure, and many organism-specific genomic databases.

Some useful things to know about BLAST:

- The protein you search with is referred to as the 'query'. Every segment of similarity returned by the search is a 'hit', because you 'hit' this on a 'subject' sequence of the target database.
- BLAST is, above all, an algorithm for pairwise alignment of sequences. You can align two sequences of interest with the option 'Align two or more sequences' below the query input box in BLASTP.
- It is generally difficult to find remote homologues which may share low sequence similarity with the protein of interest. In these cases, PSI-BLAST (accessible from the 'Program Selection' section in BLASTP) can be used to perform a deep search by incorporating similar sequences in further search rounds.
- In the BLASTN tool you can opt for the MEGABLAST program to optimize the search for highly similar sequences. This can be useful, for example, when your goal is to identify an unknown transcript or genome.

Exercises

1. Go to the NCBI BLAST homepage and select Protein BLAST. The search interface will load, with 'blastp' active on the top and several sections below (clic 'Algorithm parameters' to see more).

   a. The main input box allows searching with a sequence, an 'Accession.Version' identifier (a stable code that increases the version number when the sequence is updated) or a 'GI' number (a numeric code that is replaced by a new GI when the sequence changes). Can you find this information in the FASTA sequence copied in the following page? (It corresponds to the green algae *C. reinhardtii* protein **BLD10**, involved in creating the centrosome/basal body complex.) You can also also upload the sequence as a file and limit the search to a part of it using the 'From' and 'To' boxes on the right.

   b. Which databases can you search? How many sequences does the 'swissprot' database have?

   c. Can you restrict your search to *Eukaryota* but removing *Homo sapiens* from the search?

   d. Increase the number of 'Max target sequences' and set a lower 'Expectation threshold' (e.g. 1e-4) so that you get many results while avoiding most of those hit sequences without statistical significance.

   e. The choice of substitution matrix (and gap costs) affect calculation of the similarity score between the 'query' and each 'hit'. Higher PAM and lower BLOSUM matrices (e.g. PAM250, BLOSUM45) are better for identifying distant homologues. Which are the different matrices you can choose?

   f. Which would be the effect of activating the 'Low complexity regions' filter?

   g. Which options change if you opt for running PSI-BLAST instead?

2. Search for similarity to *C. reinhardtii* **BLD10** with BLASTP using default parameters (use the 'Reset page' top right button if you changed anything). You'll get the BLAST output page, divided in four main sections.

   a. Job details are shown on the top. The 'RID' link can be used to retrieve the results from any computer until the expiration date. Also, 'Search summary' is useful for reviewing the parameters used for running BLAST. Which is the alternative name given here to the 'Max target sequences' input option?

   b. In the same section, use 'Taxonomy reports' to get an overview of the evolutionary extent of BLD10.

   c. For protein searches, the 'Graphic Summary' section maps superfamilies of conserved domains. Which domains are found in BLD10, and where?

   d. Also in the 'Graphic Summary' you get a graphical representation of the top scoring hits, coloured by score and mapped by their region of similarity with the query. Each block corresponds to an independent hit; blocks linked by thin grey lines are separate matches in the same subject protein.

   e. The 'Descriptions' section lists the sequences producing significant alignments. Can you recognize them all? Do they look like real homologues to BLD10?

   f. The results are first sorted from best to worse *E*-value, then by decreasing scores, etc. Reorder the list by decreasing query coverage. What does this field mean? What are the other fields in this list?

   g. In this example, if you were interested in downloading five sequences of significant similarity to the query, would you download the top one? Use the check boxes to select five hits of high identity, from different organisms, and covering as much as possible of the query. Download these as a FASTA file. What is the difference between downloading 'complete sequences' and 'aligned sequences'?

   h. In the 'Alignments' section you can find the alignments between the query and each hit. Individual hits of the same subject are listed together; this is the case of the BLD10 protein from *Gonium pectorale*. How long is this sequence? How long is the region of significant similarity with the query? How many of their residues are identical and how many are similar?

   i. Click the 'Sequence ID' link at the top of the G. pectoral BLD10 protein alignment (or to the right of the entry in the 'Descriptions' section) to get the GenPept entry of this protein, with information about it presented in a structured format usually known as 'GenBank'. Can you find the region of the outer membrane channels? Use the drop-down menu on the top right to view this region only.

>gi|158280659|EDP06416.1|basal body protein [Chlamydomonas reinhardtii]
MAIDVDRTLAVLRRKLEALGYSDPLEPASLQLVQKLVEDLVHTTDSYTAVKQQCAKQAQEIAAFDTRLESVRQDSVRLQS
ENSQLHVLVMQHAERHEREAREHYTAVKRLEDTIAELSYWKHAAAEKLASADKENAGLRKRCEELAKLTDRLASGAATPQ
SVAPKISSRSPIRVAPPPSPPRPRQATVDVLQAANGRILSLQRQLADATAELQELRQRVAEDEDQIRRRDVEIDRLGTRA
GTDTNVLALRARNEANESMILQLNGTVESLAARVRELEAVEVRCEELQGALRRAEMDRDQAEERYSRSARDHDALSREVL
GLRRDLAALQDTNNRAAGLLAADAAGASTPDTTAGAPALRQRLADSRADVERLSGQLAAADMERRNLAQQLSALRSELDD
TQFLLAEAQSRAAGLAAAQAVAESEARRLAGEAAAREGRLRELDSQLAVVLSDLEARQAGFAALEKDRAEANARAEELAR
RLDEVERSAASERAAAAAAQQSVSRLDSELRVVRGSAAALEAEAAALRQELQDVSVGKVRATSALSSTEDEAVRARQQAE
ALRMQLTAERRAAEELRAGHDTLQLEVDRLGGQLALQQQEAELLRQQLAAARGELAASEAAASGAEQKLSGLGALSQRLE
EMGEQARRAQAATAEAEAEAVRLRAAVSEAKEGQARAERGLREARREVEGAREAEALVRAQLREVEAQAEGTSKALKAAE
ADRDRALMDARLAAGDLASLRDQLAAETSSAADAGSTARQMAARLSAAERAAAAAQEERERAAVAAEEAEAAAAAARGRE
EEARAQGREWAERARRAEALVAEYEADVAQLRAARDSDAAALRSLEDTVAAARRDLDARRSEVEQLTTLSLRGDATVQEY
MANLKAMSTDLRAAEMRAADLAGEAAAAQDAAASWRSEAEQLRGLLRQMDADRDNLQHELDAKAERLVAQEQQLAGAQAA
EQEAARLLALAEGRLALTDNRAREGEAEAAAVRAQLAAALDSVRALSGEGEALREELRAVSEDLEALVRENQVVGGELAA
VAAQRDSAAEEARRLGGRRASAEQLLRAKEAEAEDLRRVYEALAAEHRRLQGGVGALEREGAMREAALQAKAAEVSSLAE
SQRAAQATINQYVMDLQAFERQVDSLSRQLSQAEADGEELVRQREALLEEIRAAQQVRLGLERHREELQRQVASLDSQVA
IGRARLEDSNSEAASLNQRLAMERSRVAELEGLLAGMRAREFRSDFASDRAGGQLAVMVDRNRALEEQVASLQHQVGALQ
ASREAQDRELSRLRGEALALAASTAASLEGRTAAAGGAAAGAAKDQAAALDRLTSERDAAQDEAARLRGALAAAEAAAAS
ASTAAAVSIPAAGSGSEAAAVLRVRCSELERRNTELMQELRTLQDTCRQQESLLSAAQNELSALQAEHRRLVELVARLDQ
DKAAAAAEAAAARQQVATATRRVATAEQEAAAGAARLQSQLRDEQGRRRQAERDFLELLSSIEGAGGEAAAAAVAAAHGE
GAAELASRRLRELQTQVDALEAEKAGLEEATQRTRATLGAMSGQMAAIQAEYDATNTALAGLAGAMAAGAQGQGQVQGPA
GTAPAAAAGAPGPQPGQAQAGGFGGAHGGGSISLSGGPRR

PFAM

http://pfam.xfam.org

Pfam is a useful resource to identify conserved functional regions in proteins. Pfam helps finding regions of similarity between a query sequence and its database of annotated protein families, with the goal of gaining knowledge about the architecture, function and relationships of the protein of interest.

A Pfam entry is built from a multiple sequence alignment of a curated set of sequences known to belong to the family. This is the 'seed' alignment that is used to train a 'profile Hidden Markov Model' (or 'profile HMM') that provides an extended representation of the family. This probabilistic model reflects the variability in each position of the family and is used for an exhaustive search of a large database (such as UniProtKB) for all homologous sequences. Those retrieved sequences showing significant similarity to the profile HMM are aligned to this model, resulting in a more comprehensive 'full alignment' of the family.

Although regions covered in Pfam are commonly called "domains", a Pfam entry does not necessarily represent a stretch of sequence that folds into a discrete tertiary structure, but rather a conserved evolutionary unit. In fact, each entry in Pfam is classified in one of six categories according to the length and characteristics of it. A 'Domain' is a collection of related sequence regions that behave as a distinct structural unit. A 'Family' of related sequence can have one or many domains. Sequences of a 'Repeat' can only form stable structural unit when multiple copies are present. A 'Motif' is a group of short sequences with a defined role, often in binding. 'Coiled-coil' sequences adopt alpha-helical structures that coil together. 'Disordered' regions show noticeable sequence similarity but do not adopt regular structures.

Large and divergent families may share significant sequence, structural of functional similarity with members of other families. Since these 'superfamilies' are hard to represent by a unique alignment or profile HMM, Pfam provides a higher-level grouping of these families into 'Clans'.

Some useful things to know about Pfam:

- Pfam acts as a database of families that can be browsed and as a tool for similarity searching with the sequence of interest. Searches can also be made with the accession or name of the entry (e.g. VAV_HUMAN), or a Pfam family name (e.g. PF00571), or using related keywords.
- Any residue in any given sequence can only belong to one Pfam family.
- Data in Pfam is mostly taken from UniProt. Information is also linked from Wikipedia when available.
- Just like BLAST, Pfam provides an E-value that reflect the significance of a search hit.
- When Pfam builds a protein family, a 'gathering threshold' is set by hand for each family. This GA score determines the minimum score that any sequence should get in the profile HMM search to include this sequence in the full alignment.
- Profile HMMs are built and used using the widely used HMMER3 software package, available at http://hmmer.org
- It is possible to download the seed and full alignments and the profile HMM.
- Sometimes it is useful to download large volumes of data in bulk. Although this is not straightforward to use, Pfam (like the NCBI and many others) provides access to its database via an FTP site and a RESTful interface.
- Although Pfam is highly relevant, it may be superseded by other, more integrative databases sometime soon. Keep an eye for updates!

Exercises

1. Do you have a protein of interest? If so, try to find which Pfam domains it has and if it belongs to any Pfam clan.

2. Go to entry **PF00571**. Does it describe a protein, a domain, a family or a clan? How many sequences are linked with this entry?

3. On the same Pfam entry **PF00571**, observe the tabs on the left-hand menu.
   a. In the 'Summary' section, can you find the names of three proteins that carry this domain?
   b. 'Domain organisation' lists the domain architectures (specific arrangements of certain domains) where this family is found. How many proteins have the 'CBS x 2' architecture? Notice the graphical and colourful representation of the annotated domains; you can hold the mouse over them for an extended description. Can you find a 'disorder' region in this architecture? Discuss: Would you call this disordered region a domain?
   c. Another abundant architecture is 'IMPDH, CBS x 2'. What do the little colour dots indicate? Where does this information come from?
   d. Still in the 'Domain organisation' section, you'll find on the bottom of the page the less populated architecture. Click on 'Show' to see all 70+ sequences with this architecture. Do they all look the same? Why are some colour blocks displayed with rugged edges?
   e. In the 'Alignments' section the various formats in which the pre-calculated alignments for the family are presented. Try to download the 'seed' alignment in FASTA format, with sequences depicted in alphabetical order and showing gaps as dashes.
   f. The 'HMM logo' is a graphical summary of the profile HMM that provides a quick overview of its properties. For each position in the X-axis, a higher Y-axis value indications stronger conservation. Which is the most conserved position in this logo?
   g. Pfam provides evolutionary information about the proteins that belong to this family. Notice you can get a phylogenetic tree of the family in the 'Trees' section: which multiple sequence alignment was used to build this tree? There is also a menu option ('Species') for checking the distribution of this protein family across species. In which kingdom is this protein more abundant? How many eukaryotic species appear to have this protein?
   h. The section 'Interactions' lists 15 interactions for domains in this family. Use the 'More…' link to display further information about the source of this data: is it determined experimentally or based on computational predictions?
   i. In the 'Interactions' section, follow the link to the IMPDH family. What is the activity of this domain? Can you determine the structural fold of this family from the name of the clan it belongs to? (Notice that the 'Clan' section is active for this entry.)
   j. The 'Structures' section links the regions of UniProt entries where the CBS domain was found with their identifiers the PDB database of known structures. By following the 'PDB ID' link you can inspect the tertiary structures in the PDB database. How many different structures of the first 60 residues of the domain can you find?

4. 'Domain organisation' lists the domain architectures (specific arrangements of certain domains) You may recall that UniProt has at least three domains (JmjN, ARID and then JmjC) annotated for the human **Lysine-specific demethylase 5C**. Search Pfam with its UniProt entry name **KDM5C_HUMAN** to see the corresponding entry of the protein. Notice that the links on the left-hand menu are different than those for Pfam families.
   a. In the 'Summary' section, can you find any other domains and/or interesting regions?
   b. What does the ARID domain of this protein bind to? Can you find a PDB code of it?
   c. Are the zinc fingers listed as domains in Pfam? Were they also annotated as domains in UniProt?

Additional Resources

UniProt: Exploring protein sequence and functional information

https://www.ebi.ac.uk/training/online/course/uniprot-exploring-protein-sequence-and-functional

BLAST Help

https://blast.ncbi.nlm.nih.gov/blast/Blast.cgi?PAGE_TYPE=BlastDocs

BLAST homepage & selected search pages: How to BLAST

ftp://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_BLASTGuide.pdf

Pfam: Quick tour

https://www.ebi.ac.uk/training/online/course/pfam-quick-tour