UniProt: Quick tour

This quick tour provides a brief introduction to the Universal Protein Resource (UniProt). A full tutorial of UniProt can be found here [2].

Learning objectives:

- Basic understanding of UniProt and what it can do
- Know where to find out more about UniProt

What is UniProt?

UniProt [3] provides the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

Why do we need UniProt?

As the number of completely sequenced genomes continues to increase, huge efforts are being made in the research community to understand as much as possible about the proteins encoded by these genomes. This work is critical to many areas of science including biology, medicine and biotechnology - and is generating a wealth of data. UniProt provides an up-to-date, comprehensive body of protein information. The resource facilitates scientific discovery by collecting, interpreting and organising this information, which saves researchers countless hours of work.
What can I do with UniProt?

You can use UniProt for a wide range of tasks, from finding out about your protein of interest and comparing its protein sequence with other proteins, to mapping a list of identifiers from an external database to UniProtKB [4] or vice versa.

Who produces UniProt?

UniProt is produced by the UniProt Consortium (Figure 1), a collaboration between the European Bioinformatics Institute (EMBL-EBI), the SIB Swiss Institute of Bioinformatics and the Protein Information Resource (PIR). Around 90 people are involved across the three groups through a range of tasks such as database curation [5], software development and user support.

Figure 1. The UniProt Consortium.

The UniProt databases

UniProt comprises three different databases (Figure 2) which are optimised for different uses:

1. The UniProt Knowledgebase (UniProtKB) is used to access functional information on proteins. A subset of UniProtKB entries forms the Proteomes dataset. This consists of the set of proteins thought to be expressed by an organism whose genome has been completely sequenced;
2. The UniProt Reference Clusters (UniRefs) provide clustered sets of sequences at several resolutions;
3. The UniProt Archive (UniParc) is a sequence archive which contains all protein sequences from the main publicly available protein sequence databases.
UniProtKB

The UniProt Knowledgebase (UniProtKB), the centrepiece of the UniProt Consortium's activities, is an expertly and richly curated protein database, consisting of two sections called UniProtKB/Swiss-Prot and UniProtKB/TrEMBL (Figure 3).

UniProtKB/Swiss-Prot

UniProtKB/Swiss-Prot contains high-quality expertly curated and non-redundant protein sequence records. Expert curation [5] consists of a critical review of experimental and predicted data for each protein by a team of biologists, as well as manual verification of each protein sequence. UniProt curators extract biological information from the literature and perform numerous computational analyses. UniProtKB/Swiss-Prot aims to provide all known relevant information about a particular protein. Data captured from the scientific literature includes information on protein and gene names, function, catalytic activity, cofactors, subcellular location, protein-protein interactions and much more.

UniProtKB/TrEMBL

UniProtKB/TrEMBL contains high-quality computationally analysed records enriched with automatic annotation [6] and classification. Records are selected for full manual curation and integration into UniProtKB/Swiss-Prot according to defined priorities. You can find more information about UniProt curation priorities and processes on the UniProt website [7].
Figure 3 Relationship between manual curation and automatic annotation in the UniProt Knowledgebase

Sequence sources

The default raw sequence data for UniProtKB are:

- DDBJ/ENA/GenBank coding sequence (CDS [8]) translations;
- sequences of PDB [9] structures;
- sequences from Ensembl [10] and RefSeq [11];
- data derived from amino acid [12] sequences that are directly submitted to UniProtKB or scanned from the literature.

What is not included

The following protein sequence types are not included in UniProtKB but are stored in the UniProt Archive (UniParc [13]):

- small fragments;
- synthetic sequences;
- most non-germline immunoglobulins and T-cell receptors;
- most patent sequences;
- pseudogenes;
- sequences from redundant proteomes [14];

UniRef

Three UniProt Reference Cluster (UniRef) databases provide clustered sets of sequences from UniProtKB and selected UniParc records. This hides redundant sequences and obtains complete coverage of the sequence space at three resolutions for faster similarity searches:

UniRef100 combines identical sequences and sub-fragments with 11 or more residues from any organism into a single UniRef entry.
UniRef90 is built by clustering [15] UniRef100 sequences such that each cluster is composed of sequences that have at least 90% sequence identity to, and 80% overlap with, the longest sequence in the cluster (the seed sequence).

UniRef50 is built by clustering UniRef90 seed sequences that have at least 50% sequence identity to, and 80% overlap with, the longest sequence in the cluster.

UniParc

The UniProt Archive (UniParc) contains all protein sequences from the main publicly available protein sequence databases. This makes UniParc the most comprehensive publicly accessible non-redundant protein sequence database. A complete list of the source databases is available on the [UniProt website][16].

Proteins may exist in different source databases and in multiple copies in the same database. UniParc avoids such redundancy by storing each unique sequence only once and giving it a stable and unique UniParc identifier [17], making it possible to identify the same protein from different source databases. UniParc handles all sequences simply as text strings and sequences that are 100% identical over their entire length are merged regardless of whether they are from the same or different species. You can always trace the source database because UniParc cross-references their accession [18] numbers. UniParc also provides sequence versions which are incremented every time the underlying sequence changes. This allows you to observe sequence changes in all the source databases.

UniParc records are not annotated because annotated is context-dependent. Proteins with the same sequence can have different functions depending on species, tissue, developmental stage or other variables. This context-dependent information is the scope of UniProtKB.

Searching data from UniProt

Quick search on the UniProt website

The [UniProt website][3] features a search bar at the top of each page which allows you to search each of the UniProt databases [19] as well as supporting data and various information documents. Simply select a dataset from the drop-down menu to the left of the search box and enter your query (Figure 4). There is also an advanced search option on the right hand side of the search bar and this is described below.

![Search bar on the UniProt website](image-url)

**Figure 4.** Search bar on the UniProt website.
Advanced search on the UniProt website

The advanced search allows you to progressively build up complex queries by adding one new search term at a time to drill down to a particular data set of interest. Click on the ‘Advanced’ button to the right of the search box to access this option. For example, you can start off with an initial query for insulin and then further refine it by specifying that you only want to retrieve mouse proteins (Figure 5).

![Advanced search on the UniProt website](image)

Figure 5. Advanced search on the UniProt website. An initial query for insulin can be further refined using the query builder to include a taxonomic restriction.

Other services

In addition to text searches, you can also:

- run sequence-based searches [20] of the UniProt databases;
- perform multiple sequence alignments [21];
- upload [22] lists of identifiers in batches to retrieve multiple entries in one go;
- map identifiers [22] from an external database to UniProtKB or vice versa;
- run a peptide search [23] to find all UniProtKB sequences that exactly match a query peptide sequence.

Getting data from UniProt

Data downloads


You can also download small data sets in a range of formats directly from the UniProt website [3] by following the download link on any search result page.

Programmatic access

UniProt provides several application programming interfaces (APIs [26]) to query and access its data programmatically. You can find more information on how to query UniProt programatically on the UniProt website [27].
Submitting data to UniProt

Submitting sequence data

Directly sequenced protein sequences and associated biological information can be submitted to UniProt [28] using SPIN [29], a web-based submission tool. All of the information required to create a database entry is collected during the submission process. Accession [18] numbers are provided for submitted sequences which can be used in publications and data can be held as confidential until publication.

Submitting updates and corrections

We welcome feedback regarding updates or corrections of existing data which can be sent to the help [at] uniprot.org (UniProt help desk)

Your feedback

Please tell us what you thought about this course. Your feedback is invaluable and helps us to improve our courses and thus enhance your learning experience.

Get help and support on UniProt

Support

For all support enquiries, please contact the help [at] uniprot [28].org (UniProt help desk)

Keep up with the latest news

You can keep up with the latest UniProt developments by:

- help [at] uniprot.org (reading the )
  - UniProt blog [30]
  - visiting the UniProt Facebook page [31]
  - following UniProt on Twitter [32]

Find out more

If you want to find out more about UniProt:

- visit the UniProt YouTube channel [33] for tutorials to help you use the website effectively
- look at the help pages [34] on the UniProt website
- read the UniProt FAQs [35] to find the answers to some frequent questions
References

These papers give an overview of the UniProt project:


You can find additional papers about the UniProt project on our [website] [41].

Funding

UniProt is mainly funded by the [European Molecular Biology Laboratory] [42] (EMBL), the US [National Institutes of Health] [43], the [European Union] [44] and the [Swiss Federal Government] [45].

Contributors

[Michele Magrane] [1]

EMBL-EBI
Annotation Coordinator, Protein Function Content team

Michele Magrane is the UniProt annotation coordinator within the Protein Function Content team at EMBL-EBI, managing a team of biocurators who contribute to the expert curation of the UniProt Knowledgebase. She studied at University College Dublin and has more than 15 years’ experience in the field of protein curation. She was a founding member of the International Society for Biocuration and has served on the Nominating and Award Committees of the society.

Source URL: https://www.ebi.ac.uk/training/online/course/uniprot-quick-tourversion-0

Links