UniProt: Quick tour

Michele Magrane [1]

- Proteins
- Beginner
- 0.5 hour

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.

The UniProt databases

The UniProt databases

UniProt [6] comprises three databases (Figure 2) which are optimised for different uses:

- **UniProt Knowledgebase** [7] (UniProtKB) is used to access functional information on proteins. A subset of UniProtKB entries form the Proteomes dataset. This consists of the set of proteins thought to be expressed by an organism whose genome has been completely sequenced.
- **UniProt Reference Clusters** [8] (UniRef) provide clustered sets of sequences at several resolutions;
• **UniProt Archive** [9] (UniParc) is a sequence archive which contains all protein sequences from the main publicly available protein sequence databases;

![Diagram of UniProt data flow](image)

Figure 2 Sources and flow of data for UniProt’s component databases.

**The UniProt Knowledgebase (UniProtKB)**

The **UniProt Knowledgebase** [7], the centrepiece of the UniProt Consortium’s activities, is an expertly and richly curated protein database, consisting of two sections called **UniProtKB/Swiss-Prot** [10] and **UniProtKB/TrEMBL** [11] (Figure 3).

**UniProtKB/Swiss-Prot**

UniProtKB/Swiss-Prot contains high-quality manually annotated and non-redundant protein sequence records. **Manual annotation** [12] consists of a critical review of experimental and predicted data for each protein, as well as manual verification of each protein sequence. UniProt curators **extract** [13] biological information from the literature and perform numerous computational analyses. UniProtKB/Swiss-Prot aims to provide all known relevant information about a particular protein. Annotation captured from the scientific literature includes information on protein and gene names, function, catalytic activity, cofactors, **subcellular** [14] location, protein-protein interactions and much more.
UniProtKB/TrEMBL [11]

UniProtKB/TrEMBL contains high-quality computationally analysed records enriched with automatic annotation [15] and classification. Records are selected for full manual annotation [12] and integration into UniProtKB/Swiss-Prot [10] according to defined annotation priorities. See the UniProt curation priorities and processes documentation [16].

Sequence sources

The default raw sequence data for UniProtKB are:

- **DDBJ** [17]/**ENA** [18]/**GenBank** [19] coding sequence (**CDS**) [20]) translations;
- the sequences of **PDB** [21] structures;
- sequences from **Ensembl** [22] and **RefSeq** [23];
- data derived from **amino acid** [24] sequences that are directly submitted to UniProtKB or scanned from the literature.

What is not included

We exclude some types of data such as DDBJ/ENA/GenBank entries that **encode** [25] small
fragments, synthetic sequences, most non-germline immunoglobulins and T-cell receptors, most patent sequences and highly over-represented data (e.g. viral antigens). However, these excluded data are stored in the UniProt Archive (UniParc [26]). We regularly review UniParc data to ensure that no valuable data are missing.

**UniProt Reference Clusters (UniRef)**

Three UniRef [8] databases – UniRef100, UniRef90 and UniRef50 – merge sequences automatically across species. UniRef100 is based on all UniProtKB [4] records. It also contains selected UniParc [26] records, including Ensembl [27] protein translations from chicken, cow, dog, fly, Fugu [28], human, mouse, rat, Tetraodon [29], Xenopus and zebrafish.

UniRef100 combines identical sequences and sub-fragments with 11 or more residues from any organism into a single UniRef entry. The entries display the sequence of a representative protein, the accession [30] numbers of all the merged entries and links to the corresponding UniProtKB and UniParc records.

UniRef90 and UniRef50 are built from UniRef100, providing records with mutual sequence identity of ≥90% or ≥50%, respectively. The records also contain links to the corresponding UniProtKB records. All the sequences in each cluster are ranked to help you select a representative sequence.

**UniProt Archive (UniParc)**

UniParc [9] is designed to capture all publicly available protein sequence data and contains all the 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 [9].

A protein sequence may exist in several databases and more than once in a given database, thus creating redundant information. UniParc overcomes this problem by storing each unique sequence only once, and assigning it a unique UniParc identifier [31]. UniParc handles all sequences simply as text strings – 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 [30] 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 annotation [32] 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 [4].

**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 [33] 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.
**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).

**Other services**

In addition to text searches, you can also:

- run sequence-based searches of the UniProt databases;
- perform multiple sequence alignments;
- upload lists of identifiers in batches to retrieve multiple entries in one go;
- map identifiers from an external database to UniProtKB [4] or vice versa.

**Getting data from UniProt**
Data downloads

You can download the entire UniProtKB [4], UniRef [35] and UniParc [26] databases.

You can also download small data sets in a range of formats directly from the UniProt [36] web site by following the download link on any search result page.

Programmatic access

- For information on how to access UniProt programmatically, see the related UniProt retrieval FAQ [37].
- We also provide a JAVA programming application interface, the UniProt JAPI [38], which provides remote access for JAVA applications.

UniProt BioMart

The UniProt BioMart [39] allows for querying of UniProtKB data and linking it to data in other resources such as InterPro [40], PRIDE [41] and Ensembl [27].

Submitting data to UniProt

Submitting sequence data

Directly sequenced protein sequences and associated biological information can be submitted to UniProt [6] using SPIN [42], a web-based submission tool. All of the information required to create a database entry is collected during the submission process. Accession [30] 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 UniProt help desk [43].

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 UniProt help desk [43].

Keep up with the latest news
You can keep up with the latest UniProt developments by:

- reading the [UniProt blog](http://www.ebi.ac.uk/training/online) [44]
- visiting the [UniProt Facebook page](http://www.ebi.ac.uk/training/online) [45]
- following [UniProt on Twitter](http://www.ebi.ac.uk/training/online) [46]

**Find out more**

If you want to find out more about UniProt:

- visit the [UniProt YouTube channel](http://www.ebi.ac.uk/training/online) [47] for tutorials to help you use the website effectively
- look at the [help pages](http://www.ebi.ac.uk/training/online) [48] on the UniProt website
- read the [UniProt FAQs](http://www.ebi.ac.uk/training/online) [49] 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](http://www.ebi.ac.uk/training/online) [55].

**Funding**

UniProt is mainly funded by the [European Molecular Biology Laboratory](http://www.ebi.ac.uk/training/online) [56] (EMBL), the US [National Institutes of Health](http://www.ebi.ac.uk/training/online) [57], the [European Union](http://www.ebi.ac.uk/training/online) [58] and the [Swiss Federal Government](http://www.ebi.ac.uk/training/online) [59].

**Contributors**

![Contributor Image]
Michele Magrane [1]

EMBL-EBI
Curation Coordinator, UniProt (EMBL-EBI)

Michele Magrane is the curation coordinator for the UniProt group at EMBL-EBI. She studied at University College Dublin and joined the group in 1997, specialising in curation of mammalian proteins. She oversees the running of the EMBL-EBI UniProt curation group with responsibility for activities such as quality control, curator training and documentation.

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

Links
[1] http://www.ebi.ac.uk/training/online/trainers/magrane
[5] http://www.ebi.ac.uk/training/online/glossary/curation
[14] http://www.ebi.ac.uk/training/online/glossary/subcellular
[18] http://www.ebi.ac.uk/ena/
[20] http://www.ebi.ac.uk/training/online/glossary/cds
[24] http://www.ebi.ac.uk/training/online/glossary/amino-acid
[25] http://www.ebi.ac.uk/training/online/glossary/encode
[26] http://www.ebi.ac.uk/training/online/glossary/uniparc
[27] http://www.ebi.ac.uk/training/online/glossary/ensembl
[28] http://www.ebi.ac.uk/training/online/glossary/fugu
[29] http://www.ebi.ac.uk/training/online/glossary/tetraodon
[31] http://www.ebi.ac.uk/training/online/glossary/identifier
[32] http://www.ebi.ac.uk/training/online/glossary/annotation
[33] http://www.ebi.ac.uk/training/online/course/uniprot-quick-tour/what-uniprot/uniprot-databases
[34] http://www.uniprot.org/downloads
[35] http://www.ebi.ac.uk/training/online/glossary/uniref
[36] http://www.ebi.ac.uk/training/online/glossary/term/707
[38] http://www.ebi.ac.uk/uniprot/remotingAPI/
[40] http://www.ebi.ac.uk/training/online/glossary/interpro
[41] http://www.ebi.ac.uk/training/online/glossary/pride
[42] https://www.ebi.ac.uk/swissprot/Submissions/spin/account/login;jsessionid=CA25632EC64FEFF7BB0CF761CE421791
[43] mailto:help@uniprot.org