UniProt: the Universal Protein Resource

www.uniprot.org

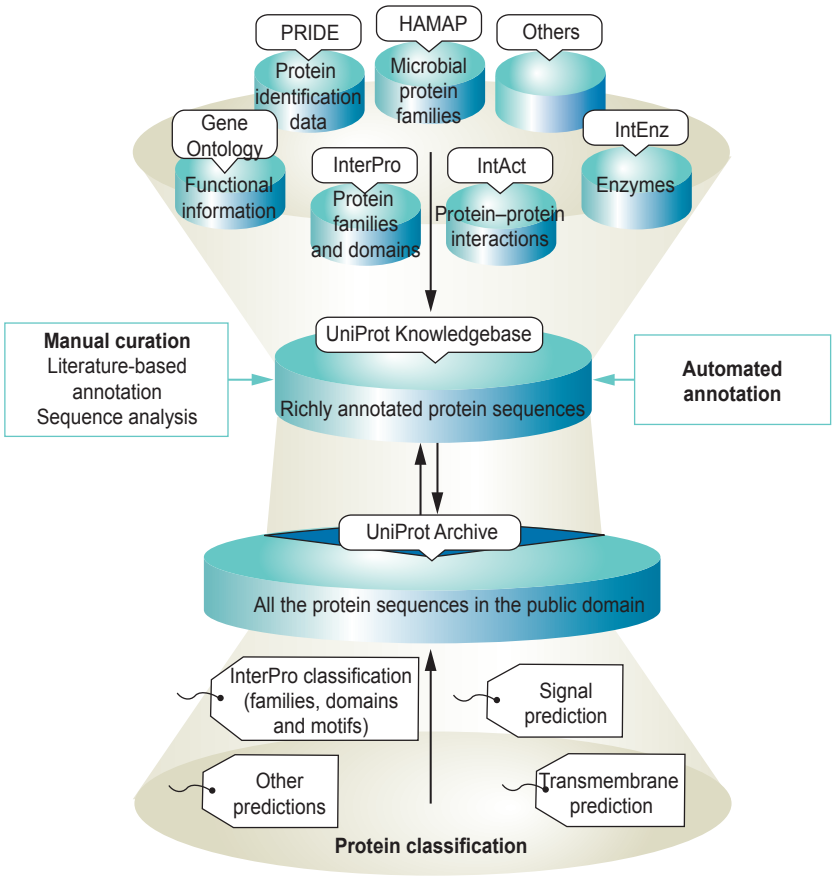
We can find out much more about the function of proteins by studying properties such as the biological processes they are involved in, their post-translational modifications, their interaction with other molecules, and their location in cells and organisms, than we could ever learn by studying the DNA that encodes them. As the number of completely sequenced genomes increases, the research community is refocusing on collecting information about all the proteins encoded in these genomes. UniProt allows biologists to access and rationalise this wealth of data.

What is UniProt?

UniProt is produced by the UniProt Consortium, a collaboration between the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR). UniProt comprises four components:

The UniProt Knowledgebase (UniProtKB)

The UniProt Knowledgebase, 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.



Sources of annotation for the UniProt Knowledgebase.

About the UniProt Consortium

The UniProt Consortium comprises the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR). EBI hosts a large resource of bioinformatics databases and services. SIB is the founding centre of the Swiss-Prot group and maintains the ExPASy (Expert Protein Analysis System) server – a central resource for proteomics databases and tools. PIR is heir to the oldest protein sequence database, Margaret Dayhoff's *Atlas of Protein Sequence and Structure*, and provides bioinformatics tools for protein sequence analysis and classification.

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

UniProtKB/Swiss-Prot contains high-quality manually annotated and non-redundant protein sequence records. Manual annotation consists of analysis, comparison and merging of all available sequences for a given protein, as well as a critical review of associated experimental and predicted data. 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. It describes, in a single record, the different protein products derived from a certain gene from a given species, including each protein derived by alternative splicing, polymorphisms and/or post-translational modifications. Protein families and groups are regularly reviewed to keep up with current scientific findings. UniProt curation priorities and processes are documented at: www.uniprot.org/help/biocuration

UniProtKB/TrEMBL contains high-quality computationally analysed records enriched with automatic annotation and classification. Records are selected for full manual annotation and integration into UniProtKB/Swiss-Prot according to defined annotation priorities.

The default raw sequence data for UniProtKB are:

- DDBJ/ENA/GenBank coding sequence (CDS) translations,
- the sequences of PDB structures,
- sequences from Ensembl and RefSeq,
- data derived from amino acid sequences that are directly submitted to UniProtKB or scanned from the literature.

We exclude some types of data such as DDBJ/ENA/GenBank entries that encode 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). We regularly review UniParc data to ensure that no valuable data are missing.

UniProt Reference Clusters (UniRef)

Three UniRef databases – UniRef100, UniRef90 and UniRef50 – merge sequences automatically across species. UniRef100 is based on all UniProtKB records. It also contains selected UniParc records, including Ensembl protein translations from chicken, cow, dog, fly, *Fugu*, human, mouse, rat, *Tetraodon*, *Xenopus* and zebrafish. UniRef100 is produced by clustering all these records by sequence identity. Identical sequences and sub-fragments are presented as a single UniRef100 entry with accession numbers of all the merged entries, the protein sequence, links to the corresponding UniProtKB and archive records. UniRef90 and UniRef50 are built from UniRef100 to provide records with mutual sequence identity of 90% or more, or 50% or more, respectively, with links to the corresponding UniProtKB records. All the sequences in each cluster are ranked to facilitate the selection of a representative sequence.

UniProt Archive (UniParc)

UniParc 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 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. 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.

UniParc data sources	
Database(s)	**Data type**
DDBJ, ENA and Genbank CDS translations	Coding sequences from the three public nucleotide sequence databases
Ensembl and VEGA	Predicted coding sequences from vertebrate genomes
FlyBase	Coding sequence for species from the Drosophilidae family
H-Invitational Database (H-Inv)	Human protein sequences
International Protein Index (IPI)	Protein sequences of higher eukaryotes
Patent Offices in Europe, US and Japan	Coding sequences associated with patents from the listed Patent Offices
PIR-PSD	Curated protein sequences
Protein Data Bank (PDB)	Sequences of proteins whose 3D structures are in the PDB
Protein Research Foundation (PRF)	Protein sequences from literature and predictions
RefSeq	Coding sequences from the NCBI's set of genomic, transcript and protein reference sequences
Saccharomyces Genome database (SGD)	Coding sequences for *Saccharomyces cerevisiae*
The Arabidopsis Information Resource (TAIR)	Coding sequences for *Arabidopsis thaliana*
TROME	Predicted protein sequences
UniProtKB/Swiss-Prot	Manually curated protein sequences mostly derived from TrEMBL
UniProtKB/TrEMBL	Automatically curated protein sequences derived from coding sequences in the nucleotide sequence databases
WormBase	Coding sequences for the nematode *Caenorhabditis elegans*

You can always trace the source database because UniParc cross-references their accession 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 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.

UniProt Metagenomic and Environmental Sequences (UniMES)

The availability of metagenomic data has necessitated the creation of a separate database, UniMES, to store sequences which are recovered directly from environmental samples. The predicted proteins from this dataset are combined with automatic classification by InterPro, an integrated resource for protein families, domains and functional sites, to enhance the original information with further analysis.

Different databases for different uses

Why have we built four different databases? Each is optimised for a different use:

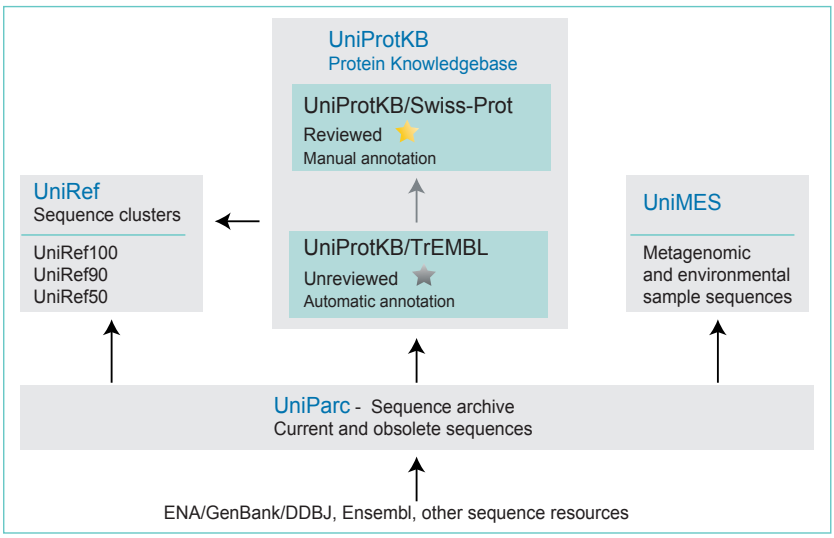
The UniProt Knowledgebase, and in particular UniProtKB/Swiss-Prot, is used to access functional information on proteins. Every UniProtKB entry contains the amino acid sequence, protein name or description, taxonomic data and citation information but in addition to this, we add as much annotation as possible. This includes widely accepted biological ontologies, classifications and cross-references, as well as clear indications

on the quality of annotation in the form of evidence attribution to experimental and computational data.

The UniRef databases provide clustered sets of sequences from UniProtKB and selected UniParc records to provide complete coverage of sequence space at several resolutions. UniRef90 and UniRef50 yield a database size reduction of approximately 40% and 65%, respectively, providing significantly faster sequence searches.

UniParc is the most comprehensive publicly accessible non-redundant protein sequence database available, providing links to all underlying sources and versions of these sequences. You can instantly find out whether a sequence of interest is already in the public domain and, if not, identify its closest relatives.

UniMES is a repository specifically for metagenomic and environmental data.



Sources and flow of data for UniProt's component databases.

Submitting data to the UniProt Knowledgebase

We provide accession numbers for proteins that have been directly sequenced. We do not provide, in advance, accession numbers for protein sequences that result from translation of nucleic acid sequences. These translations are automatically forwarded to us from the DDBJ/ENA/GenBank nucleotide sequence databases and are processed into UniProtKB/TrEMBL. All the information you need to submit sequence or annotation updates is available at www.uniprot.org/help/submissions

Retrieving data from UniProt databases

Browsing. UniProt offers a range of services at www.uniprot.org that allow you to browse and analyze data. You can perform both simple and complex text-based queries, run sequence-based searches of the UniProt databases, perform multiple sequence alignments, retrieve multiple entries and map identifiers from an external database to UniProtKB or vice versa.

Downloading. If you need to download entire databases, the UniProtKB, UniRef and UniMES databases are available at www.uniprot.org/downloads

CD-ROM. The UniProt Knowledgebase full releases are distributed on CD-ROM. If you would like to receive them, please send us an e-mail using the query form at www.ebi.ac.uk/support/.

Further reading

The UniProt Consortium. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res*. 39, D214-D219 (2011)

Support

UniProt is funded by the European Molecular Biology Laboratory (EMBL), the US National Institutes of Health, the European Union and the Swiss Federal Government.

Contacting UniProt

help@uniprot.org

www.uniprot.org

EMBL-EBI
Wellcome Trust Genome Campus
Cambridge, CB10 1SD, UK

Phone: +44 (0)1223 494444

SIB Swiss Institute of Bioinformatics
CMU - 1 rue Michel Servet
CH-1211 Geneva 4
Switzerland

Phone: +41 22 379 50 50

Protein Information Resource
Georgetown University Medical Center
3300 Whitehaven Street NW
Suite 1200
Washington, DC 20007
USA

Phone: +1 202 687 1432