UniProt: the universal protein resource

www.uniprot.org

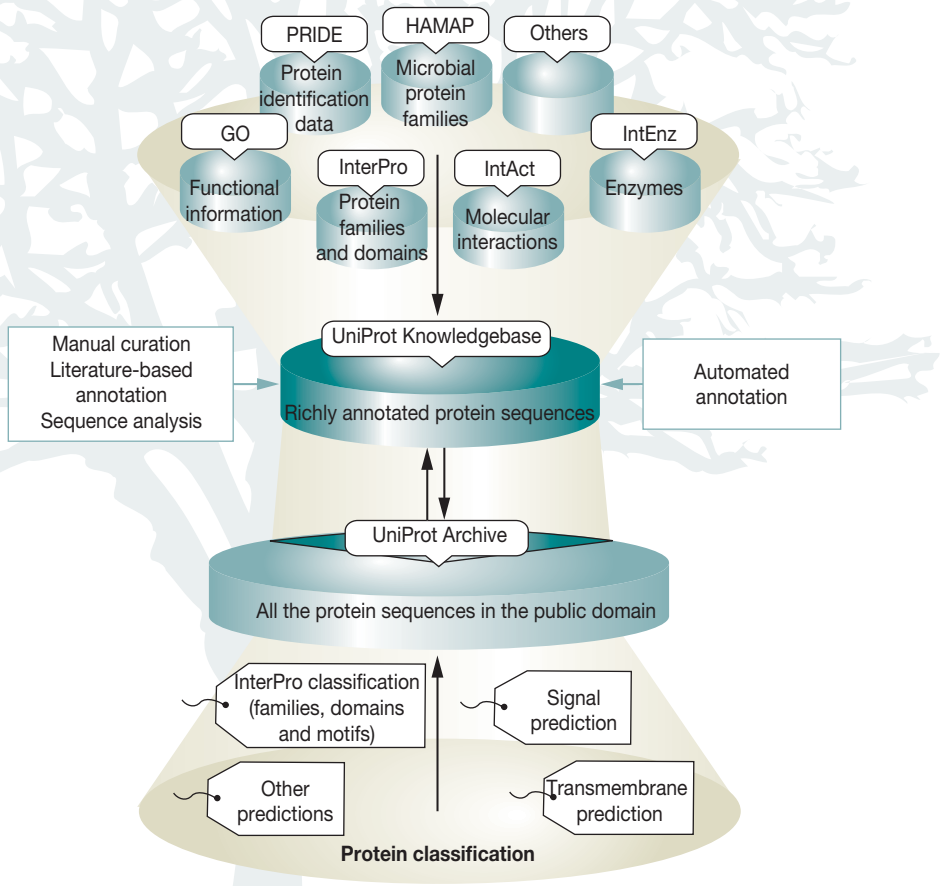
We can find out much more about the function of proteins by studying properties such as the biological processes they are involved in, their post-translational modifications, their interaction with other molecules, and their location in cells and organisms, than we could ever learn by studying the DNA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

What is UniProt?

UniProt is produced by the UniProt Consortium, a collaboration among the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB) and the Georgetown University Medical Center's Protein Information Resource (PIR). UniProt comprises three components:

The UniProt Knowledgebase (UniProtKB)

The UniProt Knowledgebase, the centrepiece of the UniProt Consortium's activities, continues the work of Swiss-Prot, TrEMBL and PIR-PSD (see side panel) by providing an expertly and richly curated protein database consisting

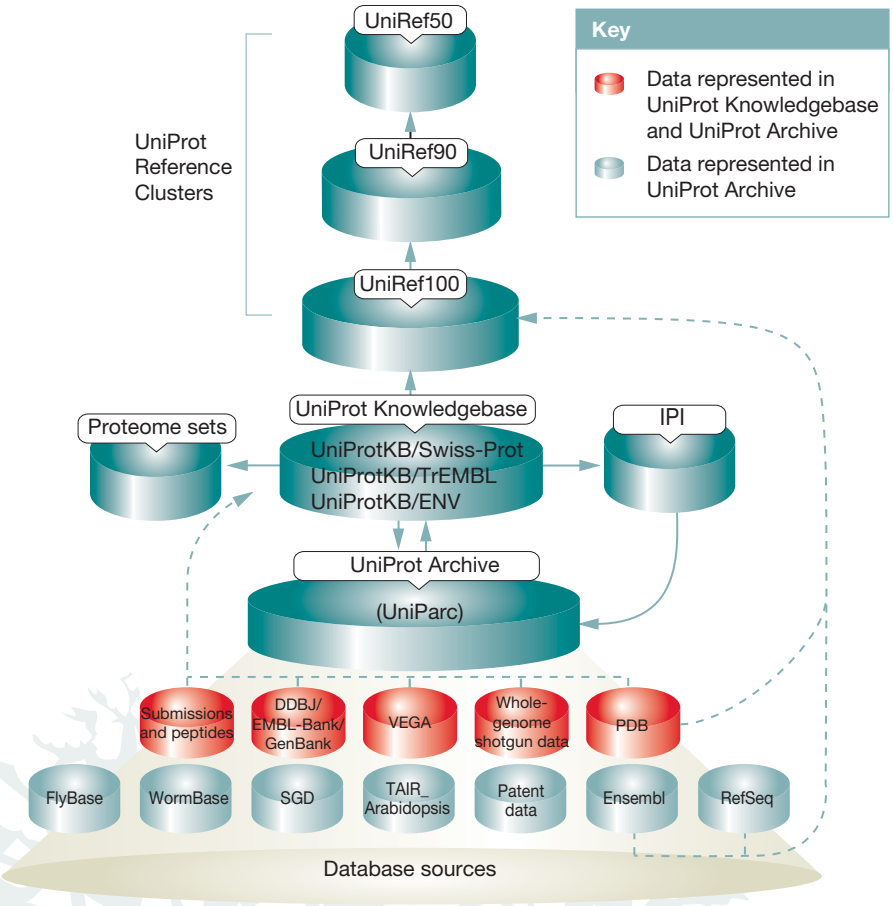


Sources of annotation for the UniProt Knowledgebase.

The UniProt Consortium

The UniProt Consortium comprises the European Bioinformatics Institute (EMBL-EBI), the Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR). EMBL-EBI hosts a comprehensive range of bioinformatics databases and services, as well as performing cutting-edge bioinformatics research, providing training and supporting industry. SIB is the founding centre of the Swiss-Prot group and maintains the ExPASy (Expert Protein Analysis System) servers – 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 primary mission of the consortium is to support biological research by maintaining a high-quality database that serves as a stable, comprehensive, fully classified, richly and accurately annotated protein sequence knowledgebase, with extensive cross-references and querying interfaces, that is freely accessible to the scientific community. UniProt is built upon the solid foundations laid by the consortium members over many years.

Published by the EMBL-EBI – a part of the European Molecular Biology Laboratory



Simplified representation of sources and flow of data for UniProt's component databases. Data flow from the database sources to IPI is omitted for the sake of clarity.

of two sections. For the sake of continuity and name recognition, these sections are referred to as UniProtKB/Swiss-Prot and UniProtKB/TrEMBL.

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 the known relevant information about a particular protein. It describes, in a single record, the different protein products derived from a certain gene (or genes if the translation from different genes in a genome leads to indistinguishable proteins) from a given species, including each protein form derived by alternative splicing and/or post-translational modifications. Protein families and groups of proteins are regularly reviewed to keep up with current scientific findings.

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 the knowledgebase are coding sequence (CDS) translations from the public nucleotide sequence databases (DDBJ/EMBL-Bank/GenBank), the sequences of PDB structures, and data derived from amino acid sequences that are directly submitted to the UniProt Knowledgebase or scanned from the literature. We exclude some types of data such as DDBJ/EMBL-Bank/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). These excluded data are stored in the UniProt Archive (UniParc; see later). We regularly review UniParc data to ensure that no valuable data are missing.

UniParc's data sources	
Database(s)	**Data type**
UniProtKB/Swiss-Prot	Manually curated protein sequences mostly derived from TrEMBL
UniProtKB/TrEMBL	Automatically curated protein sequences derived from coding sequences in nucleotide sequence databases
PIR-PSD	Curated protein sequences
DDBJ, EMBL-Bank and Genbank CDS translations	Coding sequences from the three public nucleotide sequence databases
Ensembl and VEGA	Predicted coding sequences from vertebrate genomes
International Protein Index (IPI)	Protein sequences of higher eukaryotes
Protein Data Bank (PDB)	Sequences of proteins whose 3D structures are in the PDB
RefSeq	Coding sequences from the NCBI's set of genomic, transcript and protein reference sequences
FlyBase	Coding sequence for species from the Drosophilidae family (fruit flies)
WormBase	Coding sequences for the nematode *Caenorhabditis elegans*
Patent Offices in Europe, US and Japan	Coding sequences associated with patents from the listed Patent Offices

UniProt Reference Clusters (UniRef)

Three UniRef databases – UniRef100, UniRef90 and UniRef50 – merge sequences automatically across species. UniRef100 is based on all UniProt Knowledgebase records. It also contains selected UniParc records representing sequences that are actively excluded from the UniProt Knowledgebase. These include Ensembl protein translations from chicken, dog, fruit fly, *Fugu*, human, mouse, *Tetraodon*, rat and *Xenopus*. UniRef100 is produced by clustering all these records by sequence identity. Identical sequences and subfragments are presented as a single UniRef100 entry containing the accession numbers of all the merged entries, the protein sequence, and links to the corresponding UniProt Knowledgebase 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 UniProt Knowledgebase records. All the sequences in each cluster are ranked to facilitate the selection of a representative sequence. The sequences are ranked as follows: (1) quality of the entry: member entries from UniProtKB/Swiss-Prot are preferred; (2) meaningful name: entries with names that do not contain words such as 'hypothetical' or 'probable' are preferred; (3) organism: entries from widely used model organisms are preferred; (4) sequence length: the longest sequence is preferred.

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 (see table). 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. 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

Need help?

URL: www.ebi.ac.uk/uniprot

Support: www.ebi.uniprot.org/support/Support.shtml.

e-mail: help@uniprot.org
(general enquiries)

datasubs@ebi.ac.uk
(submission enquiries)

Tel: +44 (0) 1223 494444

Fax: +44 (0) 1223 494468

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

different functions depending on species, tissue, developmental stage or other variables. This context-dependent information is the scope of the UniProt Knowledgebase.

Different databases for different uses

Why have we built three different protein databases? Each is optimized for a different use.

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

The UniRef databases provide clustered sets of sequences from the UniProt Knowledgebase 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 for 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.

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/EMBL-Bank/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/support/submissions.shtml.

Retrieving data from UniProt databases

Browsing. UniProt offers a range of services that allow you to browse and analyse the data (www.uniprot.org/search/SearchTools.shtml). Depending on the complexity of your query you can choose from three different types of text-based search. You can perform sequence-based searches of any of the UniProt databases and some of their component data sets using a variety of sequence comparison tools, and you can search for families, domains and motifs. You can perform multiple sequence alignments, retrieve multiple entries, identify proteins from proteomics experiments and perform bibliographic searches.

Downloading. If you need to download entire databases, the UniProt Knowledgebase and UniRef databases are available at www.uniprot.org/database/download.shtml.

CD-ROM. 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/.

Web services. Programmatic access to UniProt is available through the EBI's web services at www.ebi.ac.uk/Tools/webservices.

Further reading

Wu, C. *et al.* The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.* 34, D187-191 (2006)

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.