# UniProt

**UniProt** is a freely accessible database of protein sequence and functional information, many entries being derived from genome sequencing projects. It contains a large amount of information about the biological function of proteins derived from the research literature. It is maintained by the UniProt consortium, which consists of several European bioinformatics organisations and a foundation from Washington, DC, United States.

## 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, located at the Wellcome Trust Genome Campus in Hinxton, UK, hosts a large resource of bioinformatics databases and services. SIB, located in Geneva, Switzerland, maintains the ExPASy (Expert Protein Analysis System) servers that are a central resource for proteomics tools and databases. PIR, hosted by the National Biomedical Research Foundation (NBRF) at the Georgetown University Medical Center in Washington, DC, US, is heir to the oldest protein sequence database, Margaret Dayhoff's Atlas of Protein Sequence and Structure, first published in 1965.[2] In 2002, EBI, SIB, and PIR joined forces as the UniProt consortium.[3]

## The roots of the UniProt databases

Each consortium member is heavily involved in protein database maintenance and annotation. Until recently, EBI and SIB together produced the Swiss-Prot and TrEMBL databases, while PIR produced the Protein Sequence Database (PIR-PSD).[4][5][6] These databases coexisted with differing protein sequence coverage and annotation priorities.

Swiss-Prot was created in 1986 by Amos Bairoch during his PhD and developed by the Swiss Institute of Bioinformatics and subsequently developed by Rolf Apweiler at the European Bioinformatics Institute.[7][8][9] Swiss-Prot aimed to provide reliable protein sequences associated with a high level of annotation (such as the description of the function of a protein, its domain structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases. Recognizing that sequence data were being generated at a pace exceeding Swiss-Prot's

**UniProt**

| | |
|---|---|
| **Content** | |
| **Description** | **UniProt** is the **Uni**versal **Prot**ein resource, a central repository of protein data created by combining the Swiss-Prot, TrEMBL and PIR-PSD databases. |
| **Data types captured** | Protein annotation |
| **Organisms** | All |
| **Contact** | |
| **Research center** | EMBL-EBI, UK; SIB, Switzerland; PIR, US. |
| **Primary citation** | UniProt Consortium[1] |
| **Access** | |
| **Data format** | Custom flat file, FASTA, GFF, RDF, XML. |
| **Website** | www.uniprot.org (http://www.uniprot.org) www.uniprot.org/news/ (http://www.uniprot.org/news/) |
| **Download URL** | www.uniprot.org/downloads (http://www.uniprot.org/d |

ability to keep up, TrEMBL (Translated EMBL Nucleotide Sequence Data Library) was created to provide automated annotations for those proteins not in Swiss-Prot. Meanwhile, PIR maintained the PIR-PSD and related databases, including iProClass, a database of protein sequences and curated families.

The consortium members pooled their overlapping resources and expertise, and launched UniProt in December 2003.[10]

# Organization of the UniProt databases

UniProt provides four core databases: UniProtKB (with sub-parts Swiss-Prot and TrEMBL), UniParc, UniRef and Proteome.

## UniProtKB

UniProt Knowledgebase (UniProtKB) is a protein database partially curated by experts, consisting of two sections: UniProtKB/Swiss-Prot (containing reviewed, manually annotated entries) and UniProtKB/TrEMBL (containing unreviewed, automatically annotated entries).[11] As of 22 February 2023, release "2023_01" of UniProtKB/Swiss-Prot contains 569,213 sequence entries (comprising 205,728,242 amino acids abstracted from 291,046 references) and release "2023_01" of UniProtKB/TrEMBL contains 245,871,724 sequence entries (comprising 85,739,380,194 amino acids).[12]

### UniProtKB/Swiss-Prot

UniProtKB/Swiss-Prot is a manually annotated, non-redundant protein sequence database. It combines information extracted from scientific literature and biocurator-evaluated computational analysis. The aim of UniProtKB/Swiss-Prot is to provide all known relevant information about a particular protein. Annotation is regularly reviewed to keep up with current scientific findings. The manual annotation of an entry involves detailed analysis of the protein sequence and of the scientific literature.[13]

Sequences from the same gene and the same species are merged into the same database entry. Differences between sequences are identified, and their cause documented (for example alternative splicing, natural variation, incorrect initiation sites, incorrect exon boundaries, frameshifts, unidentified conflicts). A range of sequence analysis tools is used in the annotation of UniProtKB/Swiss-Prot entries. Computer-predictions are manually evaluated, and relevant results selected for inclusion in the entry. These predictions include post-translational modifications, transmembrane domains and topology, signal peptides, domain identification, and protein family classification.[13][14]

| | |
|---|---|
| | ownloads) & for downloading complete data sets ftp.uniprot.org (http://ftp.uniprot.org) |
| **Web service URL** | Yes – JAVA API see info here (http://www.ebi.ac.uk/uniprot/remotingAPI/) & REST see info here (http://www.uniprot.org/faq/28) |
| **Tools** | |
| **Web** | Advanced search, BLAST, ClustalO, bulk retrieval/download, ID mapping |
| **Miscellaneous** | |
| **License** | Creative Commons Attribution-NoDerivs |
| **Versioning** | Yes |
| **Data release frequency** | 8 weeks |
| **Curation policy** | Yes – manual and automatic. Rules for automatic annotation generated by database curators and computational algorithms. |
| **Bookmarkable entities** | Yes – both individual protein entries and searches |

Relevant publications are identified by searching databases such as PubMed. The full text of each paper is read, and information is extracted and added to the entry. Annotation arising from the scientific literature includes, but is not limited to:[10][13][14]

- Protein and gene names
- Function
- Enzyme-specific information such as catalytic activity, cofactors and catalytic residues
- Subcellular location
- Protein-protein interactions
- Pattern of expression
- Locations and roles of significant domains and sites
- Ion-, substrate- and cofactor-binding sites
- Protein variant forms produced by natural genetic variation, RNA editing, alternative splicing, proteolytic processing, and post-translational modification

Annotated entries undergo quality assurance before inclusion into UniProtKB/Swiss-Prot. When new data becomes available, entries are updated.

### UniProtKB/TrEMBL

UniProtKB/TrEMBL contains high-quality computationally analyzed records, which are enriched with automatic annotation. It was introduced in response to increased dataflow resulting from genome projects, as the time- and labour-consuming manual annotation process of UniProtKB/Swiss-Prot could not be broadened to include all available protein sequences.[10] The translations of annotated coding sequences in the EMBL-Bank/GenBank/DDBJ nucleotide sequence database are automatically processed and entered in UniProtKB/TrEMBL. UniProtKB/TrEMBL also contains sequences from PDB, and from gene prediction, including Ensembl, RefSeq and CCDS.[15] Since 22 July 2021 it also includes predicted with AlphaFold tertiary and Alphafold-multimer can even do quaternary[16] structures.[17]

## UniParc

UniProt Archive (UniParc) is a comprehensive and non-redundant database, which contains all the protein sequences from the main, publicly available protein sequence databases.[18] Proteins may exist in several different source databases, and in multiple copies in the same database. In order to avoid redundancy, UniParc stores each unique sequence only once. Identical sequences are merged, regardless of whether they are from the same or different species. Each sequence is given a stable and unique identifier (UPI), making it possible to identify the same protein from different source databases. UniParc contains only protein sequences, with no annotation. Database cross-references in UniParc entries allow further information about the protein to be retrieved from the source databases. When sequences in the source databases change, these changes are tracked by UniParc and history of all changes is archived.

### Source databases

Currently UniParc contains protein sequences from the following publicly available databases:

- INSDC EMBL-Bank/DDBJ/GenBank nucleotide sequence databases
- Ensembl
- European Patent Office (EPO)

- FlyBase: the primary repository of genetic and molecular data for the insect family Drosophilidae (FlyBase)
- H-Invitational Database (H-Inv)
- International Protein Index (IPI)
- Japan Patent Office (JPO)
- Protein Information Resource (PIR-PSD)
- Protein Data Bank (PDB)
- Protein Research Foundation (PRF)[19]
- RefSeq
- Saccharomyces Genome Database (SGD)
- The Arabidopsis Information Resource (TAIR)
- TROME[20]
- US Patent Office (USPTO)
- UniProtKB/Swiss-Prot, UniProtKB/Swiss-Prot protein isoforms, UniProtKB/TrEMBL
- Vertebrate and Genome Annotation Database (VEGA)
- WormBase

### UniRef

The UniProt Reference Clusters (UniRef) consist of three databases of clustered sets of protein sequences from UniProtKB and selected UniParc records.[21] The UniRef100 database combines identical sequences and sequence fragments (from any organism) into a single UniRef entry. The sequence of a representative protein, the accession numbers of all the merged entries and links to the corresponding UniProtKB and UniParc records are displayed. UniRef100 sequences are clustered using the CD-HIT algorithm to build UniRef90 and UniRef50.[21][22] Each cluster is composed of sequences that have at least 90% or 50% sequence identity, respectively, to the longest sequence. Clustering sequences significantly reduces database size, enabling faster sequence searches.

UniRef is available from the UniProt FTP site (http://ftp.uniprot.org/pub/databases/uniprot/current_release/uniref/).

## Funding

UniProt is funded by grants from the National Human Genome Research Institute, the National Institutes of Health (NIH), the European Commission, the Swiss Federal Government through the Federal Office of Education and Science, NCI-caBIG, and the US Department of Defense.[11]

## References

1. UniProt, Consortium. (January 2015). "UniProt: a hub for protein information" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4384041). *Nucleic Acids Research*. **43** (Database issue): D204–12. doi:10.1093/nar/gku989 (https://doi.org/10.1093%2Fnar%2Fgku989). PMC 4384041 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4384041). PMID 25348405 (https://pubmed.ncbi.nlm.nih.gov/25348405).
2. Dayhoff, Margaret O. (1965). *Atlas of protein sequence and structure*. Silver Spring, Md: National Biomedical Research Foundation.

3. "2002 Release: NHGRI Funds Global Protein Database" (https://web.archive.org/web/20150924040602/http://www.genome.gov/page.cfm?pageID=10005283). *National Human Genome Research Institute (NHGRI)*. Archived from the original (http://www.genome.gov/page.cfm?pageID=10005283) on 24 September 2015. Retrieved 14 April 2018.
4. O'Donovan, C.; Martin, M. J.; Gattiker, A.; Gasteiger, E.; Bairoch, A.; Apweiler, R. (2002). "High-quality protein knowledge resource: SWISS-PROT and TrEMBL" (https://doi.org/10.1093%2Fbib%2F3.3.275). *Briefings in Bioinformatics*. **3** (3): 275–284. doi:10.1093/bib/3.3.275 (https://doi.org/10.1093%2Fbib%2F3.3.275). PMID 12230036 (https://pubmed.ncbi.nlm.nih.gov/12230036).
5. Wu, C. H.; Yeh, L. S.; Huang, H.; Arminski, L.; Castro-Alvear, J.; Chen, Y.; Hu, Z.; Kourtesis, P.; Ledley, R. S.; Suzek, B. E.; Vinayaka, C. R.; Zhang, J.; Barker, W. C. (2003). "The Protein Information Resource" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC165487). *Nucleic Acids Research*. **31** (1): 345–347. doi:10.1093/nar/gkg040 (https://doi.org/10.1093%2Fnar%2Fgkg040). PMC 165487 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC165487). PMID 12520019 (https://pubmed.ncbi.nlm.nih.gov/12520019).
6. Boeckmann, B.; Bairoch, A.; Apweiler, R.; Blatter, M. C.; Estreicher, A.; Gasteiger, E.; Martin, M. J.; Michoud, K.; O'Donovan, C.; Phan, I.; Pilbout, S.; Schneider, M. (2003). "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC165542). *Nucleic Acids Research*. **31** (1): 365–370. doi:10.1093/nar/gkg095 (https://doi.org/10.1093%2Fnar%2Fgkg095). PMC 165542 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC165542). PMID 12520024 (https://pubmed.ncbi.nlm.nih.gov/12520024).
7. Bairoch, A.; Apweiler, R. (1996). "The SWISS-PROT protein sequence data bank and its new supplement TREMBL" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC145613). *Nucleic Acids Research*. **24** (1): 21–25. doi:10.1093/nar/24.1.21 (https://doi.org/10.1093%2Fnar%2F24.1.21). PMC 145613 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC145613). PMID 8594581 (https://pubmed.ncbi.nlm.nih.gov/8594581).
8. Bairoch, A. (2000). "Serendipity in bioinformatics, the tribulations of a Swiss bioinformatician through exciting times!" (https://doi.org/10.1093%2Fbioinformatics%2F16.1.48). *Bioinformatics*. **16** (1): 48–64. doi:10.1093/bioinformatics/16.1.48 (https://doi.org/10.1093%2Fbioinformatics%2F16.1.48). PMID 10812477 (https://pubmed.ncbi.nlm.nih.gov/10812477).
9. Séverine Altairac, "Naissance d'une banque de données: Interview du prof. Amos Bairoch (http://expasy.org/prolune/pdf/prolune018_fr.pdf)". *Protéines à la Une (http://expasy.org/prolune/)*, August 2006. ISSN 1660-9824 (https://www.worldcat.org/search?fq=x0:jrnl&q=n2:1660-9824).
10. Apweiler, R.; Bairoch, A.; Wu, C. H. (2004). "Protein sequence databases". *Current Opinion in Chemical Biology*. **8** (1): 76–80. doi:10.1016/j.cbpa.2003.12.004 (https://doi.org/10.1016%2Fj.cbpa.2003.12.004). PMID 15036160 (https://pubmed.ncbi.nlm.nih.gov/15036160).
11. Uniprot, C. (2009). "The Universal Protein Resource (UniProt) in 2010" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2808944). *Nucleic Acids Research*. **38** (Database issue): D142–D148. doi:10.1093/nar/gkp846 (https://doi.org/10.1093%2Fnar%2Fgkp846). PMC 2808944 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2808944). PMID 19843607 (https://pubmed.ncbi.nlm.nih.gov/19843607).
12. "UniProtKB/Swiss-Prot Release 2023_01 statistics" (https://web.expasy.org/docs/relnotes/relstat.html). *web.expasy.org*. Retrieved 31 March 2023.
13. "How do we manually annotate a UniProtKB entry?" (https://www.uniprot.org/faq/45). *www.uniprot.org*. Retrieved 14 April 2018.

14. Apweiler, R.; Bairoch, A.; Wu, C. H.; Barker, W. C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; Martin, M. J.; Natale, D. A.; o'Donovan, C.; Redaschi, N.; Yeh, L. S. (2004). "UniProt: The Universal Protein knowledgebase" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC308865). *Nucleic Acids Research*. **32** (90001): 115D–1119. doi:10.1093/nar/gkh131 (https://doi.org/10.1093%2Fnar%2Fgkh131). PMC 308865 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC308865). PMID 14681372 (https://pubmed.ncbi.nlm.nih.gov/14681372).
15. "Where do the UniProtKB protein sequences come from?" (https://www.uniprot.org/faq/37). *www.uniprot.org*. Retrieved 14 April 2018.
16. Humphreys, Ian R.; Pei, Jimin; Baek, Minkyung; Krishnakumar, Aditya; Anishchenko, Ivan; Ovchinnikov, Sergey; Zhang, Jing; Ness, Travis J.; Banjade, Sudeep; Bagde, Saket R.; Stancheva, Viktoriya G. (2021). "Computed structures of core eukaryotic protein complexes" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7612107). *Science*. **374** (6573): eabm4805. doi:10.1126/science.abm4805 (https://doi.org/10.1126%2Fscience.abm4805). PMC 7612107 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7612107). PMID 34762488 (https://pubmed.ncbi.nlm.nih.gov/34762488).
17. "Putting the power of AlphaFold into the world's hands" (https://deepmind.com/blog/article/putting-the-power-of-alphafold-into-the-worlds-hands). *Deepmind*. Retrieved 24 July 2021.
18. Leinonen, R.; Diez, F. G.; Binns, D.; Fleischmann, W.; Lopez, R.; Apweiler, R. (2004). "UniProt archive" (https://doi.org/10.1093%2Fbioinformatics%2Fbth191). *Bioinformatics*. **20** (17): 3236–3237. doi:10.1093/bioinformatics/bth191 (https://doi.org/10.1093%2Fbioinformatics%2Fbth191). PMID 15044231 (https://pubmed.ncbi.nlm.nih.gov/15044231).
19. "Protein Research Foundation" (http://www.prf.or.jp/index-e.html).
20. ftp://ftp.isrec.isb-sib.ch/pub/databases/trome
21. Suzek, B. E.; Huang, H.; McGarvey, P.; Mazumder, R.; Wu, C. H. (2007). "UniRef: Comprehensive and non-redundant UniProt reference clusters" (https://doi.org/10.1093%2Fbioinformatics%2Fbtm098). *Bioinformatics*. **23** (10): 1282–1288. doi:10.1093/bioinformatics/btm098 (https://doi.org/10.1093%2Fbioinformatics%2Fbtm098). PMID 17379688 (https://pubmed.ncbi.nlm.nih.gov/17379688).
22. Li, W.; Jaroszewski, L.; Godzik, A. (2001). "Clustering of highly homologous sequences to reduce the size of large protein databases" (https://doi.org/10.1093%2Fbioinformatics%2F17.3.282). *Bioinformatics*. **17** (3): 282–283. doi:10.1093/bioinformatics/17.3.282 (https://doi.org/10.1093%2Fbioinformatics%2F17.3.282). PMID 11294794 (https://pubmed.ncbi.nlm.nih.gov/11294794).

## External links

- UniProt (https://www.uniprot.org)