UniProt: the Universal Protein Resource

Lai-Su Yeh[1], UniProt Consortium[2]
[1]Protein Informatics Resource, Georgetown University Medical Center, Washington, DC 20057
[2] European Bioinformatics Institute (EBI), Hinxton, UK; Protein Information Resource (PIR), Washington, DC, USA; Swiss Institute of Bioinformatics (SIB), Geneva, Switzerland

The Universal Protein Resource (UniProt) provides the scientific community with a single, centralized, authoritative resource for protein sequences and functional information. It is produced by the UniProt consortium, formed by European Bioinformatics Institute (EBI), Georgetown University Protein Information Resource (PIR) and Swiss Institute of Bioinformatics (SIB). UniProt comprises three components, each optimized for different uses: the UniProt Archive, the UniProt Knowledgebase and the UniProt Reference Clusters.

The **UniProt Archive (UniParc)** is a comprehensive repository that reflects the history of all protein sequences. UniParc handles all sequences as strings - all sequences 100% identical over the entire length are merged, regardless of source species. New and updated sequences are loaded on a daily basis, cross-referenced to the source database accession number, and provided with a sequence version that increments upon changes to the underlying sequence - making it possible to observe sequence changes in all source databases. UniParc records are without annotation since the annotation will be only true in the real context of the sequence: proteins with the same sequence may have different functions depending on species, tissue, developmental stage, etc. UniParc Release 4.0 (1.2.2005) contained 4,467,956 unique sequences from 11,544,956 different source database records.

The **UniProt Knowledgebase (UniProtKB)** is the central access point for extensively curated protein information. It continues the work of Swiss-Prot, TrEMBL and PIR-PSD by providing a comprehensive, expertly curated, fully classified, richly and accurately annotated protein sequence knowledgebase with extensive cross-references. UniProtKB is a protein-centric, non-redundant database aiming to provide everything that is known about a protein. This centerpiece consists of three sections: one contains fully manually annotated records resulting from literature information extraction and curator-evaluated computational analysis (UniProtKB/Swiss-Prot); the other contains computationally analyzed records awaiting full manual annotation (UniProtKB/TrEMBL). The third section (UniProtKB/ENV) is a newly created section that will contain computationally analyzed "Environmental and other taxonomically unassigned sequences." All suitable PIR-PSD sequences missing from Swiss-Prot + TrEMBL were incorporated into UniProt. Bi-directional cross-references between Swiss-Prot + TrEMBL records and PIR-PSD entries were created to allow easy tracking of the now historic and no longer updated PIR-PSD records.

UniProtKB represents a subset of all the sequences stored in UniParc. UniProtKB release 4.0 (1.2.2005) contained 163,235 Swiss-Prot and 1,449,374 TrEMBL records. The DDBJ/EMBL/GenBank CDS translations, sequences of the PDB structures, and data derived from direct amino acid sequencing are used, by default, as raw material for the UniProt Knowledgebase. However, some data from these sources (including CDS translations leading to small fragments or not coding for real proteins, synthetic sequences, non-germline immunoglobulins and T-cell receptors, and most patent application sequences) are actively excluded from UniProtKB. This process ensures that important sequences are not missed while minimizing the number of unstable and low quality data.

The UniProt Knowledgebase provides extensive cross-references to external data collections such as underlying DNA sequence entries in the DDBJ/EMBL/GenBank nucleotide sequence databases, 2D PAGE and 3D protein structure databases, various protein domain and family characterization databases, post-translational modification (PTM) databases, species-specific data collections, variant databases and disease databases. Accordingly, UniProtKB acts as a central hub for biomolecular information archived in 64 cross-referenced databases (release 4.0).

The main annotation principles of the UniProt Knowledgebase follow the established procedures used to annotate Swiss-Prot, TrEMBL, and PIR-PSD. Expert manual annotation involves extracting from the literature as much reliable information as possible about protein properties such as function(s), domains and sites, developmental stages, catalytic activity, prosthetic groups, bound metal ions and covalently modified residues, regulation, induction, pathways, tissue specificity, subcellular location, quaternary structure, diseases associated with deficiencies in the protein, processes in which the protein may be involved, and similarities to other proteins. Automatic annotation at UniProt is classification-driven, whereby annotation can be reliably propagated from sequences containing experimentally determined properties to closely-related homologous sequences based on sequence families. The accuracy of this process is ensured by using annotation rules (both manually and automatically created), which are family-specific and allow propagation of protein names, synonyms, EC numbers, GO terms, and comment fields such as Function, Pathway, and Caution.

For non-redundancy, UniProtKB aims to describe, in a single record, all protein products derived from a certain gene (or genes if the translation from different genes in a genome leads to indistinguishable proteins) from a certain species, giving not only the whole record an accession number but also assigning unique identifiers to each protein form derived by alternative splicing, proteolytic cleavage, and post-translational modification.

The **UniProt Reference Clusters (UniRef)** merge UniProtKB and select UniParc sequences (including certain Ensembl protein translations, RefSeq data, and other smaller data sets) at different resolutions based on sequence identity. UniRef100 clusters records such that identical sequences and subfragments are presented in a single entry containing a representative protein sequence and links to the accession numbers of the corresponding UniProtKB and UniParc records. UniRef90 and UniRef50, yielding size reductions of 40% and 65%, respectively, are built from UniRef100, and allow for faster homology searches.

The UniProt databases can be accessed online (http://www.uniprot.org) or downloaded in several formats (ftp://ftp.uniprot.org/pub). New releases are published every two weeks.