Poster B-5

UniProt: the Universal Protein Resource.

Authors:
Raja Mazumder (*Protein Information Resource, Georgetown University*)
Claire ODonovan (*European Bioinformatics Institute*)

Short Abstract: UniProt is a comprehensive catalog of protein sequence and function, produced by EBI, PIR and SIB. It consists of the UniProt Knowledgebase (expertly curated database), the UniProt Reference Clusters (merged sequences based on sequence identity) and the UniProt Archive (comprehensive sequence repository). An additional component consisting of environmental sequences will be made available later this year.

Long Abstract:

The ability to store and interconnect all available information on proteins is crucial to modern biological research. Accordingly, the Universal Protein Resource (UniProt) plays an ever more important role by providing a stable, comprehensive, freely accessible central resource on protein sequences and functional annotation. UniProt is produced by the UniProt consortium, formed by European Bioinformatics Institute (EBI), Georgetown University Protein Information Resource (PIR) and Swiss Institute of Bioinformatics (SIB). The core activities in UniProt include manual curation of protein sequence assisted by automated annotation, sequence archiving, development of a user-friendly UniProt web site, and the provision of additional value-added information on proteins through cross-references to other databases. UniProt is comprised of three major components, each optimized for different uses: the UniProt Archive, the UniProt Knowledgebase and the UniProt Reference Clusters. An additional component consisting of metagenomic and environmental sequences will be made available later this year.

The UniProt Archive (UniParc) is the main sequence storehouse and is a comprehensive repository that reflects the history of all protein sequences. UniParc houses all new and revised protein sequences from various sources including not only the UniProt Consortium databases but also translations from the EMBL-Bank/DDBJ/GenBank nucleotide sequence databases, the EnsEMBL database of animal genomes, the International Protein Index (IPI), the Protein Data Bank (PDB), NCBI's Reference Sequence Collection (RefSeq), model organism databases such as FlyBase and WormBase, and protein sequences from the European, American and Japanese Patent Offices. to ensure that complete coverage is available at a single site. To avoid redundancy, sequences are handled as strings - all sequences 100% identical over the entire length are merged, regardless of source organism. 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. The basic information stored with each UniProt Archive entry is the identifier, the sequence, cyclic redundancy check number, source database(s) with accession and version numbers, and a time stamp. In addition, each source database accession number is tagged with its status in that database, indicating if the sequence still exists or has been deleted at that source. 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.

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, 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. UniProt Knowledgebase provides an integrated and uniform presentation of disparate data, including annotations such as protein name and function, taxonomy, enzyme-specific information (catalytic activity, cofactors, metabolic pathway, regulatory mechanisms), domains and sites, post-translational modifications, subcellular locations, tissue- or developmentally-specific expression, interactions, splice isoforms, polymorphisms, diseases, and sequence conflicts. Literature citations provide evidence for experimental data. Entries connect to various external data collections such as the underlying DNA sequence entries, protein structure databases, protein domain and family databases, and species- and function/feature-specific data collections. As a result, UniProtKB acts as a central hub connecting biomolecular information archived in about 100 cross-referenced databases. The UniProt Knowledgebase contains two sections. UniProtKB/Swiss-Prot contains records with full manual annotation (or computer-assisted, manually-verified annotation) performed by biologists and based on published literature and sequence analysis. UniProtKB/TrEMBL contains computationally generated records enriched with automatic classification and annotation.

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. With the accelerated growth of the number of sequences, similarity searching has become increasingly computationally intensive and prohibitive for resource providers and their users. Furthermore, there is an uneven distribution of sequences in sequence space. An overabundance of very closely related sequences (e.g., above 90% identity) slows down database searches, and long lists of similar or identical alignments can obscure novel matches in the output. UniRefs provide an even sampling of sequences, producing shorter and cleaner output listings without repetition of redundant hits in similarity searches. The UniProt Reference Clusters are three separate datasets that compress sequence space at different resolutions, achieved by merging sequences and sub-sequences that are 100% (UniRef100), 90% (UniRef90), or 50% (UniRef50) identical, regardless of source organism. The compression of UniRef100 into UniRef90 and UniRef50 yielded size reductions of approximately 40% and 65%, respectively.

OTHER FEATURES

Metagenomic and Environmental Sequences (UniMES)

Swiss-Prot and TrEMBL sections of the UniProt Knowledgebase contain entries with a known taxonomic source. However, a new development in sequence production—namely, the availability of metagenomic data—has necessitated the creation of a separate section, UniProt Metagenomic and Environmental Sequences (UniMES).

UniProtKB Sequence/Annotation Version Database (UniSave)
UniProt Knowledgebase entries are subject to change in both sequence and annotation, but only the most recent versions are currently preserved in the database. The UniProtKB Sequence/Annotation Version Database (UniSave) retains earlier versions of entries, thus allowing retrieval of historic views of UniProtKB records.

ID mapping service
ID cross-referencing is fundamental to support interoperability among disparate data sources and to allow integration and querying of data from heterogeneous molecular biology databases. UniProt therefore provides a mapping service to convert common gene IDs and protein IDs (such as NCBI's gi number and Entrez Gene ID) to UniProt Knowledgebase AC/ID and vice versa. Mapping is provided between UniProtKB and about 30 other data sources.

External links
Established through close collaborations with the research community, the UniProt Knowledgebase provides explicit and implicit links via DR (Database cross-Reference) lines to about 100 molecular databases and resources. UniProt continually adds new database cross-references to UniProtKB records, thereby facilitating broader access to relevant online resources with complementary protein-related information. External resources can easily link to individual protein entries in UniProt using a link URL. For example, UniProtKB entries can be referenced by http://www.uniprot.org/entry/AC.

Bibliography mapping service
We provide annotated bibliography pages (http://www.uniprot.org/bibliography/biblioretrieve.shtml) that list, for each UniProt entry, both curated bibliography and computationally-mapped bibliography.

Database access
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.