Nucleic Acids Research, 2004, Vol. 32, Database issue ***D115–D119***
DOI: 10.1093/nar/gkh131

UniProt: the Universal Protein knowledgebase

Rolf Apweiler*, Amos Bairoch[1], Cathy H. Wu[2], Winona C. Barker[3], Brigitte Boeckmann[1], Serenella Ferro[1], Elisabeth Gasteiger[1], Hongzhan Huang[2], Rodrigo Lopez, Michele Magrane, Maria J. Martin, Darren A. Natale[2], Claire O'Donovan, Nicole Redaschi[1] and Lai-Su L. Yeh[3]

The EMBL Outstation—European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK, [1]Swiss Institute of Bioinformatics, Centre Medical Universitaire, 1 rue Michel Servet, 1211 Geneva 4, Switzerland, [2]Department of Biochemistry and Molecular Biology and [3]National Biomedical Research Foundation, Georgetown University Medical Center, 3900 Reservoir Road, NW, Box 571414, Washington, DC 20057-1414, USA

Received August 25, 2003; Revised and Accepted October 27, 2003

ABSTRACT

To provide the scientific community with a single, centralized, authoritative resource for protein sequences and functional information, the Swiss-Prot, TrEMBL and PIR protein database activities have united to form the Universal Protein Knowledgebase (UniProt) consortium. Our mission is to provide a comprehensive, fully classified, richly and accurately annotated protein sequence knowledgebase, with extensive cross-references and query interfaces. The central database will have two sections, corresponding to the familiar Swiss-Prot (fully manually curated entries) and TrEMBL (enriched with automated classification, annotation and extensive cross-references). For convenient sequence searches, UniProt also provides several non-redundant sequence databases. The UniProt NREF (UniRef) databases provide representative subsets of the knowledgebase suitable for efficient searching. The comprehensive UniProt Archive (UniParc) is updated daily from many public source databases. The UniProt databases can be accessed online (http://www.uniprot.org) or downloaded in several formats (ftp://ftp.uniprot.org/pub). The scientific community is encouraged to submit data for inclusion in UniProt.

INTRODUCTION

Until recently, Swiss-Prot + TrEMBL (1) and PIR-PSD (2) coexisted as protein databases with differing sequence coverage and annotation priorities. In 2002, the Swiss-Prot + TrEMBL group at the Swiss Institute of Bioinformatics (SIB) and European Bioinformatics Institute (EBI) and the Protein Information Resource (PIR) group at the Georgetown University Medical Center and National Biomedical Research Foundation joined forces as the UniProt consortium. 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 freely accessible to the scientific community. UniProt will build upon the solid foundations laid by the consortium members over many years.

The UniProt databases consist of three database layers:

(i) The UniProt Archive (UniParc) provides a stable, comprehensive, non-redundant sequence collection by storing the complete body of publicly available protein sequence data.

(ii) The UniProt Knowledgebase (UniProt) provides the central database of protein sequences with accurate, consistent and rich sequence and functional annotation.

(iii) The UniProt NREF databases (UniRef) provide non-redundant data collections based on the UniProt knowledgebase in order to obtain complete coverage of sequence space at several resolutions.

THE UNIPROT ARCHIVE (UNIPARC)

The UniProt Archive (UniParc) is the most comprehensive publicly accessible non-redundant protein sequence collection available. It contains publicly available protein sequences from many different sources, including Swiss-Prot, TrEMBL, PIR-PSD, EMBL (3), Ensembl (4), IPI (http://www.ebi.ac.uk/IPI), PDB (5), RefSeq (6), FlyBase (7), WormBase (8), and European, American and Japanese patent offices. While a protein sequence may exist in multiple databases and more than once in a given database, UniParc stores each unique sequence only once and assigns a unique UniParc identifier. Furthermore, UniParc provides cross-references to the source databases (accession numbers), sequence versions and status (active or obsolete). A UniParc sequence version is also provided, and incremented each time the underlying sequence changes, thus making it possible to observe sequence changes in all source databases. An example UniParc report can be found at http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-noSession+-e+[UNIPARC:UPI0000133132] (SRS view) and http://www.

*To whom correspondence should be addressed. Tel: +44 1223 494435; Fax: +44 1223 494468; Email: apweiler@ebi.ac.uk

pir.uniprot.org/cgi-bin/upEntry?id=UPI0000133132 (PIR view).

THE UNIPROT KNOWLEDGEBASE (UNIPROT)

The UniProt knowledgebase is the centrepiece of the consortium activities. We have merged Swiss-Prot, TrEMBL and PIR-PSD to form the UniProt knowledgebase in order to provide a central database of protein sequences with annotations and functional information. All suitable PIR-PSD sequences missing from Swiss-Prot + TrEMBL were incorporated into UniProt. Bidirectional cross-references between Swiss-Prot + TrEMBL and PIR-PSD were created to allow the easy tracking of the PIR-PSD entries. The transfer into UniProt of references and experimentally verified data present in PIR but missing from Swiss-Prot + TrEMBL is ongoing.

The UniProt knowledgebase consists of two parts: a section containing fully manually annotated records resulting from literature information extraction and curator-evaluated computational analysis, and a section with computationally analysed records awaiting full manual annotation. For the sake of continuity and name recognition, the two sections are referred to as 'Swiss-Prot' and 'TrEMBL'. An example UniProt report can be found at http://www.expasy.org/cgi-bin/niceprot.pl?P57727 (NiceProt view), http://www.pir.uniprot.org/cgi-bin/upEntry?id=P57727 (iProClass view) or http://srs.ebi.ac.uk/cgi-bin/wgetz?-e+[swall-acc:P57727] (SRS view).

In the following paragraphs we will explain the main principles of the UniProt knowledgebase.

High-quality annotation

We will curate UniProt knowledgebase entries to an even higher level of detail than that already achieved in Swiss-Prot + TrEMBL and PIR-PSD. In addition to capturing the core data mandatory to each UniProt entry (consisting principally of the amino acid sequence, the protein name or description, taxonomic data and citation information), we strive to attach as much annotation information as possible to the protein. This is achieved in two ways: manually and automatically.

Manual annotation by curators based on literature and sequence analysis

Sequences for which novel functional, structural, and/or biochemical data have been published are assigned high manual annotation priority. In UniProt, annotation consists of the description of the following items:

- function(s) of the protein;
- enzyme-specific information (catalytic activity, cofactors, metabolic pathway, regulation mechanisms);
- biologically relevant domains and sites;
- post-translational modification (PTM)(s);
- molecular weight determined by mass spectrometry;
- subcellular location(s) of the protein;
- tissue-specific expression of the protein;
- developmentally specific expression of the protein;
- secondary structure;
- quaternary structure;
- interactions;
- splice isoform(s);
- mature protein products;
- polymorphism(s);
- similarities to other proteins;
- use of the protein in a biotechnological process;
- diseases associated with deficiencies or abnormalities of the protein;
- use of the protein as a pharmaceutical drug;
- sequence conflicts, etc.

This annotation is found in the comment lines (CC), feature table (FT) and keyword lines (KW). Comments are classified according to topics to allow easy retrieval of specific categories of data from the database.

To acquire the most up-to-date and wide-ranging knowledge regarding a protein, information is obtained not only from publications reporting new sequence data, but also from review articles to facilitate the periodic revision of protein families or groups of proteins. Furthermore, we have enlisted external experts to send us comments and updates concerning specific groups of proteins.

In order to provide the high level of annotation described above, all UniProt curators read a large amount of scientific literature related to each protein. This enables them to contribute to the work of the gene ontology (GO) consortium (9) by assigning GO terms during the annotation process as they extract information related to each of the GO ontologies, i.e. the function of a protein, what processes it is involved in and where in the cell it is located.

Automatic classification and annotation

With the rapid growth of sequence databases, there is an increasing need for reliable functional characterization and annotation of newly predicted proteins. To cope with such large data volumes, faster and more effective means of protein sequence characterization and annotation are required. One promising approach is automatic large-scale functional characterization and annotation, which is generated with limited human interaction.

InterPro classification. We use InterPro (10) to recognize domains and to classify all protein sequences in UniProt into families and superfamilies. InterPro is an integrated resource of protein families, domains and sites that amalgamates the efforts of the member databases: Pfam (11), PROSITE (12), PRINTS (13), ProDom (14), SMART (15), PIRSF (16), Superfamily (17) and TIGRFAMs (18). The comprehensive InterPro classification is a prerequisite for improving the quality and quantity of our annotation using highly structured, classification-driven, rule-based, automated procedures.

Automatic functional annotation of the TrEMBL section of UniProt. For automatic annotation, a novel system of standardized transfer of annotation from well-characterized proteins in the Swiss-Prot section of UniProt to non-annotated TrEMBL entries has been developed (19). Using this system, the Swiss-Prot section is used as the source to generate the annotation rules, which are then stored and managed in RuleBase. InterPro is then used to assign TrEMBL entries into groups. The annotation shared by the functionally characterized Swiss-Prot proteins of the group is then extracted and is assigned to the unannotated TrEMBL entry. This system has

been used to improve the annotation in 25% of TrEMBL entries. A new data mining approach to automatic annotation is also being developed to complement this system, which will increase coverage by automatic annotation over the next year and will bring the standard of annotation in the TrEMBL section of UniProt closer to that of the Swiss-Prot section.

Also to be incorporated into the RuleBase annotation pipeline are the PIR classification-driven rule-based procedures, which will provide standardized and rich UniProt annotation for position-specific features, protein names and keywords. New feature rules are being defined systematically for fully curated PIRSF families that contain at least one known 3D structure with experimentally verified functional/active/binding site information. The PIRSF classification, based on the evolutionary relationships of whole proteins, have also been used to detect and correct numerous genome annotation errors that have resulted from identifications based only on local domain similarities and subsequently propagated based on transitivity (20).

High-quality Automated and Manual Annotation of microbial Proteomes (HAMAP)

A combined approach of automated and manual annotation for prokaryotic genomes in Swiss-Prot has resulted in the development of the HAMAP project (21). The HAMAP project, or 'High-quality Automated and Manual Annotation of microbial Proteomes' aims to integrate manual and automatic annotation methods in order to enhance the speed of the curation process while preserving the quality of the database annotation. Automatic annotation is only applied to entries that belong to manually defined orthologous families and to entries with no identifiable similarities (ORFans).

Annotation of ORFans. Various prediction tools are applied to proteins that show no similarity to known protein families. Possible transmembrane regions, signal sequence, coiled coils, ATP/GTP binding sites, LPXTG motifs and some defined repeats are automatically annotated using rules of consistency and dependency, and without any further manual verification.

Annotation of members of well-characterized (sub)families. Proteins belonging to well-characterized protein (sub)families can be annotated automatically using a rule system that describes the extent and nature of annotations that can be assigned by similarity to a prototype manually annotated entry. Such a rule system also includes a carefully edited multiple alignment of the (sub)family, which is used both to propagate feature annotation from a model entry and to generate profiles used to identify new members of the family. Species-specific rules and rules specific to the biochemical pathways are used to develop a system able to spot inconsistencies at the level of the entire proteome.

Standardized nomenclature and controlled vocabularies

Consistent nomenclature is indispensable for communication and literature search. UniProt aims to standardize the nomenclature for a given protein and its isoforms across related organisms. For various other UniProt items we use controlled vocabulary, e.g. for tissues, plasmids and keywords, which are listed in UniProt documents. The unified UniProt keyword list is based on Swiss-Prot keywords augmented by the addition of selected PIR keywords that represent new concepts or new parent/child nodes of existing Swiss-Prot keywords. Whenever available, we make use of the official nomenclature defined by international committees while still providing the published synonyms. Collaborations and regular data exchange with other databases and organizations allow the implementation of community-specific nomenclatures.

Integration with other databases

UniProt provides cross-references to external data collections such as the 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, PTM databases, species-specific data collections, variant databases and disease databases. As a result of this, UniProt acts as a central hub for biomolecular information archived in more than 50 cross-referenced databases. A document listing all databases cross-referenced in UniProt (http://www.uniprot.org/support/docs/dbxref.shtml) is available and contains, for each database, a short description and the server URL. This interconnectivity is achieved almost exclusively via Database cross-Reference (DR) lines. In addition, links from subsequences or particular sites to databases specializing in certain types of PTMs or mutations are provided. Unique and stable feature identifiers (FTId) allow reference to a position-specific annotation item in the feature table. Currently these are systematically attributed to FT VARIANT lines of human sequence entries, to alternative splicing events (VARSPLIC) and to certain glycosylation sites (CARBOHYD), but will ultimately be assigned to all types of FT lines.

Minimal redundancy

Many sequence databases contain, for a given protein sequence, separate entries that correspond to different literature reports. In UniProt we try as much as possible to merge all these data in order to minimize the redundancy of the database. Differences between sequencing reports due to splice variants, polymorphisms, disease-causing mutations, experimental sequence modifications or simply sequencing errors are indicated in the feature table of the corresponding UniProt entry.

Splice isoforms may differ considerably from one another, with potentially <50% sequence similarity between isoforms. The tool VARSPLIC (22), which is freely available enables the recreation of all annotated splice variants from the FT of a UniProt entry, or for the complete database. A FASTA-formatted file containing all splice variants annotated in UniProt can be downloaded for use with similarity search programs.

Evidence attribution

The UniProt consortium emphasizes the use of an evidence attribution mechanism for protein annotation that will include, for all data, the data source, the types of evidence and methods for annotation. This is essential as the UniProt knowledgebase will contain data automatically imported from the underlying nucleotide sequence databases, data imported from other databases, data from specific programs, the results of

automatic annotation systems and most important of all, expert manual curation. The implementation of evidence tags will allow the user to distinguish between all these data sources and to easily identify particular classes of data of interest such as the experimentally proven protein annotation.

To further improve the quality of protein annotation by increasing the amount of experimentally verified data with source attribution, UniProt has developed a bibliography submission system and is conducting retrospective attribution of literature data. The submission page allows submission and categorization of literature citations for experimental annotations, and displays comprehensive bibliographic data collected from many curated databases for each UniProt entry. A systematic manual attribution of experimental features is being carried out with computer-assisted mapping to existing protein bibliographic information. So far, a few thousand experimental features have been associated with publications and cross-referenced to the corresponding PMIDs for direct incorporation into the UniProt knowledgebase.

THE UNIREF DATABASES

Automatic procedures have been developed to create three UniProt NREF (UniRef) databases, NREF100, NREF90 and NREF50, from the UniProt knowledgebase as representative protein sequence databases with high information content. The databases provide complete coverage of sequence space while hiding redundant sequences from view. The non-redundancy facilitates sequence merging in the UniProt knowledgebase (based on NREF100) and allows faster sequence similarity searches (by using NREF90 and NREF50).

NREF100 provides, as a modified extension of the PIR-NREF database (2), a comprehensive non-redundant sequence collection clustered by sequence identity and taxonomy with source attribution. As in PIR-NREF, identical sequences and subfragments from the same source organism (species) are presented as a single NREF entry with accession numbers of all the merged UniProt entries, the protein sequence, taxonomy, bibliography, links to the corresponding UniProt knowledgebase and archive records, as well as close sequence neighbors (with at least 95% sequence identity) from the same source organism. An example NREF100 report can be found at http://www.pir.uniprot.org/cgi-bin/unipEntry?id=URI0000E815.

NREF90 and NREF50 are built from NREF100 using the CD-HIT algorithm (23) to provide non-redundant sequence collections for the scientific user community to perform faster homology searches. All records from all source organisms with mutual sequence identity of >90% or >50%, respectively, are merged into a single record that links to the corresponding UniProt knowledgebase records. NREF90 and NREF50 yield a size reduction of ~40% and 65%, respectively.

PRACTICAL INFORMATION

Interactive access to UniProt

The most efficient and user-friendly way to browse the UniProt databases is via the UniProt web site (http://www.uniprot.org), which serves as a portal to all aspects developed in the framework of the UniProt project, and contains detailed documentation about the background and scope of UniProt. It provides database query and data mining mechanisms, user support and communication, file download capabilities and links to consortium resources (SIB: http://www.expasy.org, EBI: http://www.ebi.ac.uk and PIR: http://pir.georgetown.edu). The UniProt Help Desk (help@uniprot.org) provides access to UniProt curators and database maintainers.

UniProt data availability and submission

UniProt and UniRef entries, and supporting documentation can be retrieved in various formats (Swiss-Prot/TrEMBL flat file, FASTA, XML) via anonymous FTP from ftp://ftp.uniprot.org/pub/.

UniProt accepts submissions of new sequences, entry updates and corrections, and annotated bibliographic information for protein entries. Directions for submission are available at http://www.uniprot.org/support/submissions.shtml.

CONCLUSIONS

Complete and up-to-date databases of biological knowledge are vital for information-dependent biological and biotechnological research. With the rapid accumulation of genome sequences for many organisms, attention is turning to the identification and function of proteins encoded by these genomes. With the increasing volume and variety of protein sequences and functional information, UniProt serves as a central resource of protein sequence and function, providing a cornerstone for scientists active in modern biological research, especially in the field of proteomics. The resource provides rich, consistent and non-redundant protein information by combining reliable automated annotation approaches with literature-based expert manual curation. UniProt will facilitate knowledge discovery by allowing researchers to integrate the enormous amount of data from the Human Genome Project and from structural and functional genomics and proteomics.

ACKNOWLEDGEMENTS

UniProt is supported mainly by the National Institutes of Health (NIH) grant 1 U01 HG02712-01. Minor support for the EBI's involvement in UniProt comes from the two European Union contracts BioBabel (QLRT-2000-00981) and TEMBLOR (QLRI-2001-00015) and from the NIH grant 1R01HGO2273-01. Swiss-Prot activities at the SIB are supported by the Swiss Federal Government through the Federal Office of Education and Science. PIR activities are also supported by the National Science Foundation (NSF) grants DBI-0138188 and ITR-0205470.

REFERENCES

1. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
2. Wu,C.H., Yeh,L.-S.L., Huang,H., Arminski,L., Castro-Alvear,J., Chen,Y., Hu,Z., Kourtesis,P., Ledley,R.S., Suzek,B.E. *et al.* (2003) The Protein Information Resource. *Nucleic Acids Res.*, **31**, 345–347.

3. Stoesser,G., Baker,W., van den Broek,A., Garcia-Pastor,M., Kanz,C., Kulikova,T., Leinonen,R., Lin,Q., Lombard,V., Lopez,R. *et al.* (2003) The EMBL Nucleotide Sequence Database: major new development. *Nucleic Acids Res.*, **30**, 21–26.
4. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.
5. Westbrook,J., Feng,Z., Chen,L., Yang,H. and Berman,H. (2003) The Protein Data Bank and structural genomics. *Nucleic Acids Res.*, **31**, 489–491.
6. Pruitt,K. and Maglott,D. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.
7. FlyBase Consortium (2003) The FlyBase database of the *Drosophila* genome projects and community literature. *Nucleic Acids Res.*, **31**, 172–175.
8. Harris,T., Lee,R., Schwarz,E., Bradnam,K., Lawson,D., Chen,W., Blasier,D., Kenny,E., Cunningham,F., Kishore,R. *et al.* (2003) WormBase: a cross-species database for comparative genomics. *Nucleic Acids Res.*, **31**, 133–137.
9. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.
10. Mulder,N., Apweiler,R., Attwood,T., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.
11. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.
12. Hulo,N., Sigrist,C.J.A., Le Saux,V., Langendijk-Genevaux,P., Bordoli,L., Gattiker,A., De Castro,E., Bucher,P. and Bairoch,A. (2004) Recent improvements to the PROSITE database. *Nucleic Acids Res.*, **32**, D134–D137.
13. Attwood,T.K., Bradley,P., Flower,D.R., Gaulton,A., Maudling,N., Mitchell,A.L., Moulton,G., Nordle,A., Paine,K., Taylor,P. *et al.* (2003) PRINTS and its automatic supplement, preprints. *Nucleic Acids Res.*, **31**, 400–402.
14. Servant,F., Bru,C., Carrere,S., Courcelle,E., Couzy,J., Peyruc,D.and Kahn,D. (2002) Prodom: Automated clustering of homologous domains. *Brief. Bioinform.*, **3**, 246–251.
15. Letunic,I., Goodstadt,L., Dickens,N.J., Doerks,T., Schultz,J., Mott,R., Ciccarelli,F., Copley,R.R., Ponting,C.P. and Bork,P. (2002) Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res.*, **30**, 242–244.
16. Wu,C.H., Nikolskaya,A., Huang,H., Yeh,L.-S., Natale,D.,Vinayaka,C.R., Hu,Z., Mazumder,R., Kumar,S., Kourtesis,P. *et al.* (2004) PIRSF: family classification system at the Protein Information Resource. *Nucleic Acids Res.*, **32**, D112–D114.
17. Gough,J., Karplus,K., Hughey,R. and Chothia,C. (2001) Assignment of homology to genome sequences using a library of Hidden Markov Models that represent all proteins of known structure. *J. Mol. Biol.*, **313**, 903–919.
18. Haft,D.H., Loftus,B.J., Richardson,D.L., Yang,F., Eisen,J.A., Paulsen,I.T. and White,O. (2001) TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res.*, **29**, 41–43.
19. Fleischmann,W., Moeller,S., Gateau,A. and Apweiler,R. (1999) A novel method for automatic and reliable functional annotation. *Bioinformatics*, **15**, 228–233.
20. Wu,C.H., Huang,H., Yeh,L.-S. and Barker,W.C. (2003) Protein family classification and functional annotation. *Comput. Biol. Chem.*, **27**, 37–47.
21. Gattiker,A., Michoud,K., Rivoire,C., Auchincloss,A.H., Coudert,E., Lima,T., Kersey,P., Pagni,M., Sigrist,C.J.A., Lachaize,C. *et al.* (2003) Automatic annotation of microbial proteomes in Swiss-Prot. *Comput. Biol. Chem.*, **27**, 49–58.
22. Kersey,P., Hermjakob,H. and Apweiler,R. (2000) VARSPLIC: alternatively-spliced protein sequences derived from Swiss-Prot and TrEMBL. *Bioinformatics*, **11**, 1048–1049.
23. Li,W., Jaroszewski,L. and Godzik,A. (2002) Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics*, **18**, 77–82.