D190–D195 *Nucleic Acids Research, 2008, Vol. 36, Database issue* *Published online 27 November 2007*
doi:10.1093/nar/gkm895

The Universal Protein Resource (UniProt)

The UniProt Consortium*

The EMBL Outstation, The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK, Protein Information Resource, Georgetown University Medical Center, 3300 Whitehaven St. NW, Suite 1200, Washington, DC 20007, USA and Swiss Institute of Bioinformatics, Centre Medical Universitaire 1 rue Michel Servet, 1211 Geneva 4, Switzerland

Received September 17, 2007; Revised and Accepted October 3, 2007

ABSTRACT

The Universal Protein Resource (UniProt) provides a stable, comprehensive, freely accessible, central resource on protein sequences and functional annotation. The UniProt Consortium is a collaboration between the European Bioinformatics Institute (EBI), the Protein Information Resource (PIR) and the Swiss Institute of Bioinformatics (SIB). The core activities include manual curation of protein sequences assisted by computational analysis, sequence archiving, development of a user-friendly UniProt website, and the provision of additional value-added information through cross-references to other databases. UniProt is comprised of four major components, each optimized for different uses: the UniProt Knowledgebase, the UniProt Reference Clusters, the UniProt Archive and the UniProt Metagenomic and Environmental Sequences database. UniProt is updated and distributed every three weeks, and can be accessed online for searches or download at http://www.uniprot.org.

INTRODUCTION

For the rapid and ongoing accumulation of predicted protein sequences by high-throughput genome sequencing for numerous and increasingly diverse organisms, the expansion of large-scale proteomics (e.g. gene expression profiling and protein–protein interactions) and the advent of structural genomics have combined to provide a wealth of data to analyze and use. There is a widely recognized need for a centralized repository of protein sequences with comprehensive coverage and a systematic approach to protein annotation, incorporating, integrating and standardizing data from these various sources.

UniProt is the central resource for storing and interconnecting information from large and disparate sources, and the most comprehensive catalog of protein sequence and functional annotation. It has four components optimized for different uses. The UniProt Knowledgebase (UniProtKB) is an expertly curated database, a central access point for integrated protein information with cross-references to multiple sources. The UniProt Archive (UniParc) is a comprehensive sequence repository, reflecting the history of all protein sequences (1). UniProt Reference Clusters (UniRef) merge closely related sequences based on sequence identity to speed up searches. The UniProt Metagenomic and Environmental Sequences (UniMES) database is a repository specifically developed for the newly expanding area of metagenomic and environmental data. UniProt is built upon the extensive bioinformatics infrastructure and scientific expertise at European Bioinformatics Institute (EBI), Protein Information Resource (PIR) and Swiss Institute of Bioinformatics (SIB). It is freely and easily accessible to researchers.

CONTENT

The UniProt Knowledgebase (UniProtKB)

UniProtKB consists of two sections, UniProtKB/Swiss-Prot and UniProtKB/TrEMBL. The former contains manually annotated high quality records with information extracted from literature and curator-evaluated computational analysis. Sequences for which novel functional, structural and/or biochemical data have been published are assigned priority. To achieve accuracy, annotations are performed by biologists with specific expertise. In UniProtKB, annotation consists of the description of the following: function(s), enzyme-specific information, biologically relevant domains and sites, post-translational modifications, subcellular location(s), tissue specificity, developmentally specific expression, structure, interactions, splice isoform(s), diseases associated with deficiencies or abnormalities, etc. Another important part of the annotation process involves the merging of different reports for a single protein. After a careful inspection of the sequences, the annotator selects the reference sequence, does the corresponding

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

merging, and lists the splice and genetic variants along with disease information when available. Any discrepancies between the different sequence sources are also annotated. Cross-references are provided to the underlying nucleotide sequence sources as well as to many other useful databases including organism-specific, domain, family and disease databases. UniProtKB/TrEMBL contains computationally analyzed records enriched with automatic annotation and classification. The computer-assisted annotation is created using automatically generated rules as in Spearmint (2), or manually curated rules based on protein families, including HAMAP family rules (3), RuleBase rules (4) and PIRSF classification-based name rules and site rules (5,6). UniProtKB/TrEMBL contains the translations of all coding sequences (CDS) present in the EMBL/GenBank/DDBJ Nucleotide Sequence Databases, 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 pseudogenes, small nucleotide fragments, synthetic sequences, most non-germline immunoglobulins and T-cell receptors, most patent sequences, some highly over-represented data and open reading frames (ORFs) which have been wrongly predicted to code for proteins. Records are selected for full manual annotation and integration into UniProtKB/Swiss-Prot according to defined annotation priorities.

The UniProt Reference Clusters (UniRef)

UniRef provides clustered sets of all sequences from the UniProt Knowledgebase (including splice forms as separate entries) and selected UniProt Archive records to obtain complete coverage of sequence space at resolutions of 100%, 90% and 50% identity while hiding redundant sequences (7). The UniRef clusters provide a hierarchical set of sequence clusters where each individual member sequence can exist in only one UniRef cluster at each resolution and have only one parent or child cluster at another resolution. The UniRef100 database combines identical sequences and sub-fragments into a single UniRef entry. UniRef90 is built from UniRef100 clusters and UniRef50 is built from UniRef90 clusters. UniRef100, UniRef90 and UniRef50 yield a database size reduction of ~10%, 40% and 70%, respectively. Each cluster record contains source DB, protein name and taxonomy organism information on each member sequences but is represented by a single selected representative protein sequence and name, the number of members and highest common taxonomy node for the membership is included. UniRef100 is the most comprehensive and non-redundant protein sequence dataset available. The reduced size of the UniRef90 and 50 datasets provides for faster sequence similarity searches and reduces the research bias in similarity searches by providing a more even sampling of sequence space. UniRef is currently being used for a broad range of applications in the areas of automated genome annotation, family classification, systems biology, structural genomics, phylogenetic analysis and mass spectrometry (7). The UniRef clusters are updated with every release of UniProtKB.

The UniProt archive (UniParc)

UniParc is the main sequence storehouse and is a comprehensive repository that reflects the history of all protein sequences (1). UniParc houses all new and revised protein sequences from various sources to ensure that complete coverage is available at a single site. It includes not only UniProtKB but also translations from the EMBL-Bank/DDBJ/GenBank Nucleotide Sequence Databases, the Ensembl database of eukaryotic genomes, the H-Invitational Database (H-Inv), the International Protein Index (IPI), the Protein Data Bank (PDB), Protein Research Foundation (PRF), NCBI's Reference Sequence Collection (RefSeq), model organism databases FlyBase, SGD, TAIR Arabidopsis thaliana and WormBase, TROME and protein sequences from the European, American and Japanese Patent Offices. 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 within each UniParc entry is the identifier, the sequence, cyclic redundancy check number, source database(s) with accession and version numbers, and a time stamp. If a UniParc entry does not have a cross-reference to a UniProtKB entry, the reason for the exclusion of that sequence from UniProtKB is provided (e.g. pseudogene). 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 in the source database and cross-references to NCBI GI and TaxId if appropriate. UniParc records are designed to be without annotation since the annotation will be only true in the real biological context of the sequence: proteins with the same sequence may have different functions depending on species, tissue, developmental stage, etc.

The UniProt Metagenomic and Environmental Sequences database (UniMES)

The UniProt Knowledgebase contains entries with a known taxonomic source. A new development in sequence production—namely, the availability of metagenomic data—has necessitated the creation of a separate database, UniProt Metagenomic and Environmental Sequences database (UniMES). Metagenomics is the large-scale genomic analysis of microbes recovered from environmental samples as opposed to laboratory-grown organisms, which represent only a small proportion of the microbial world. UniMES currently contains data from the Global Ocean Sampling Expedition (GOS) (8) which was originally submitted to the International Nucleotide Sequence Databases (INSDC) (9). The initial GOS dataset is composed of 25 million DNA sequences primarily from oceanic microbes and predicts

nearly 6 million proteins. By combining the predicted protein sequences with automatic classification by InterPro, the integrated resource for protein families, domains and functional sites, UniMES uniquely provides free access to the array of genomic information gathered from the sampling expeditions, enhanced by links to further analytical resources. The environmental sample data contained within this database is not present in the UniProt Knowledgebase or the UniProt Reference Clusters but is integrated into UniParc. UniMES is available on the ftp site in FASTA format with a UniMES matches to InterPro methods file.

NEW DEVELOPMENTS

Recent format changes

(i) Introduction of the new line type PE (Protein Existence)

Most protein sequences are derived from translations of gene predictions. Some of them exhibit strong sequence similarity to known proteins in closely related species. For other proteins, there is experimental evidence such as Edman sequencing, clear identification by mass spectrometry (MSI), X-ray or NMR structure, detection by antibodies, etc. To indicate these different levels of evidence for the existence of a protein, we have introduced the PE (Protein Existence) line to all UniProtKB entries. The criteria for assigning a particular PE level is described in the document pe_criteria.txt, available both by ftp and on the website. Note that the PE line does not describe the accuracy or correctness of a sequence displayed in UniProtKB but the evidence for the existence of a protein. It may happen that the protein sequence is not entirely accurate, especially for sequences derived from gene predictions from genomic sequences.

The PE line appears between the DR and KW lines of the UniProtKB entries and the format is:

PE *Level: Evidence*;

With the following values:

- 1:Evidence at protein level;
- 2:Evidence at transcript level;
- 3:Inferred from homology;
- 4:Predicted;
- 5:Uncertain;

(ii) The format of the ID line was changed to better reflect the annotation status of an entry. The STANDARD and PRELIMINARY data classes were replaced by 'Reviewed' (entries that have been manually reviewed and annotated by UniProtKB curators) and 'Unreviewed' (computer-annotated entries that have not been reviewed by UniProtKB curators), respectively.

(iii) The feature key INIT_MET is used to indicate that the initiator methionine has been cleaved off. Previously, the initiator methionine was not included in the sequence of a UniProtKB entry in such a case and the INIT_MET sequence coordinates were therefore 0. The initiator methionine has now been added back to such protein sequences and the sequence coordinates of the feature key INIT_MET accordingly changed to 1.

(iv) A new CC line topic 'SEQUENCE CAUTION' was introduced to specifically describe reported sequences that differ substantially from that which is displayed, and where the underlying cause of this difference cannot be clearly described in FT CONFLICT lines. Typical examples of such sequence discrepancies may include frameshifts, erroneous gene model predictions and the presence of contaminating vector sequence or sequence of unknown origin. This type of information was previously reported in the CC line topic CAUTION, together with other types of warnings that are unrelated to sequence differences between the submitted sequences contained in the entry.

(v) Evidence tags in UniProtKB XML

In UniProtKB/TrEMBL, the evidence attribute and the evidence element are used to indicate the source of an annotation. This has now been extended to UniProtKB/Swiss-Prot. In the initial phase, automatic procedures are used to infer the evidence from the existing data (mainly the contents of the scope element). It will gradually also become part of the manual curation process. The completion of the retrofit of existing UniProtKB/Swiss-Prot with evidence information will be an ongoing process. The evidences are also visible on the UniProt website.

Forthcoming format change

The protein names contained in the description (DE) lines of reviewed UniProtKB entries are widely used by scientists to unambiguously identify a protein and provide a definitive source nomenclature for the annotation of homologous proteins in new genomic sequences. It is hence essential to provide description lines from which the recommended name(s) of a given protein, and all known synonyms, may be easily identified and extracted. To achieve this, the DE line format has been revised to clearly distinguish the recommended name of the protein as well as any commonly used synonyms, abbreviations, EC numbers for proteins with enzymatic activity, and to indicate obsolete or erroneous names.

This also allowed UniProt to undertake a major clean up of the content of these lines and to implement strict protein-naming guidelines. Whenever possible, the official nomenclature defined by the appropriate expert international committees are used, although UniProt also attempts to create standard nomenclature for proteins where no such recommendations currently exist and to establish naming conventions that can be consistently applied across the largest spectrum of species as possible. The guidelines are described in the document nameprot.txt, available both by ftp and on the website. In addition, the subset of these guidelines pertinent to microbial organisms has been ratified for use by the American Society for Microbiology, and discussions with the Joint Commission on Biochemical

Nomenclature of IUPAC and IUBMB (JCBN) are underway for likely endorsement as an official protein nomenclature document.

Previous format:

DE GDP-fucose protein O-fucosyltransferase 1 precursor (EC 2.4.1.221)
DE (Peptide-O-fucosyltransferase) (O-fucosyltransferase 1) (O-FucT-1)
DE (Neurotic protein).

New format:

DE RecName: Full = GDP-fucose protein O-fucosyltransferase 1;
DE EC = 2.4.1.221;
DE AltName: Full = Peptide-O-fucosyltransferase;
DE AltName: Full = O-fucosyltransferase 1;
DE Short = O-FucT-1;
DE AltName: Full = Neurotic protein;
DE Flags: Precursor;

UniProtKB cross-references

The UniProt Knowledgebase cross-references 118 external databases, 92 with explicit links and 26 via implicit links. These resources provide additional or complementary information to what is available at UniProt and can be valuable for biological discovery. Further documentation is present in dbxref.txt, available both on ftp and the website. Recently, a collaboration between UniProt and National Center for Biotechnology Information (NCBI) teams was established to provide bi-directional cross-references between Entrez Gene/RefSeq and UniProtKB and this was achieved in our databases in September 2007.

ID Mapping enhancements

UniProt provides a mapping service to convert common gene IDs and protein IDs to UniProtKB AC/ID and vice versa. Mappings are either inherited from cross-references within UniProtKB entries, or make use of cross-references obtained from the iProClass database (10). This service is available at http://www.uniprot.org/search/idmapping.shtml, where users can map between UniProtKB and about 100 other data sources, such as NCBI (e.g. gi numbers, RefSeq accession numbers, Entrez Gene IDs, PubMed IDs), GO (www.geneontology.org/), PFAM (www.sanger.ac.uk/Software/Pfam/) and PIRSF (pir.georgetown.edu/pirsf.shtml). Users can enter a set of IDs or the name of an ID file to retrieve the mappings. A HTTP client is also provided for programmatic access. In addition, users can download selected mappings in the form of a tab-delimited table. To facilitate the large-scale proteomic and gene expression data analysis, the ID mapping services will be further enhanced to: (i) provide downloadable files for mappings between UniProtKB AC/IDs and commonly used IDs (such as a NCBI gi number); (ii) provide downloadable mapping files for the most commonly used organisms (such as model organisms); (iii) provide a SOAP web services for programmatic access.

Enhancement of bibliography information

The UniProtKB bibliography information provides protein entries with additional and up-to-date curated literature from several other sources including GeneRIF, SGD, MGI and more recently GAD (Genetic Association Database) (11). New sources of curated literature information continue to be added to the protein bibliography from model organism databases (MOD) such as ZFIN, RGD, Wormbase, dictyBase and Flybase. The bibliography information provides not only the source attribution of each reference, but also succinct information about the reference annotated by the source database. The bibliography information is available via the website.

UniProt website developments

The UniProt website (http://www.uniprot.org/) has been thoroughly redeveloped and the individual mirrors (http://www.ebi.uniprot.org, http://www.expasy.uniprot.org, http://www.pir.uniprot.org and parts of http://www.expasy.org) are no longer maintained. User feedback and the analysis of the use of our previous sites has led us to put more emphasis on supporting the most frequently used functionalities: Database searches with simple (and sometimes less simple) queries that often consist of only a few terms have been enhanced by a good scoring system and a suggestion mechanism. Searching with ontology terms is assisted by auto-completion, and there is also the possibility of using ontologies to browse search results. The viewing of database entries was improved with configurable views, a simplified terminology and a better integration of documentation. Medium-to-large sized result sets can now be retrieved directly on the site, so people no longer need to be referred to commercial, third party services. We have also simplified access to the most common bioinformatics tools: sequence similarity searches, multiple sequence alignments, batch retrieval and a database identifier mapping tool can now be launched directly from any page, and the output of these tools can be combined, filtered and browsed like normal database searches. Programmatic access to all data and results is possible via simple HTTP (REST) requests (http://www.uniprot.org/help/technical). In addition to the existing formats that we support for our different data sets (e.g. plain text, FASTA and XML for UniProtKB), we now also provide (configurable) tab-delimited, RSS and GFF downloads where possible, and all data is available in RDF (http://www.w3.org/RDF/), a W3C standard for publishing data on the Semantic Web.

Annotation developments

Subcellular location annotation. In order to facilitate the standardization of the CC line (Comment) topic SUBCELLULAR LOCATION, the free text content of these lines was analyzed. It was decided to describe concepts for the location, the topology and the orientation with respect to a membrane, with a controlled vocabulary. This controlled vocabulary is described in the new

document subcell.txt, available both by ftp and on the website. For a given concept, the preferred term for the controlled vocabulary is provided with a precise definition, its synonyms and other relevant information. To allow optimal use of this controlled vocabulary, the format of the SUBCELLULAR LOCATION comment has been modified. The line starts with the controlled vocabulary(ies) optionally followed by a 'Note=' containing additional relevant information. If there are multiple isoforms/peptides/chains for which location information is available, they are described in distinct SUBCELLULAR LOCATION comments. An integral part of developing this controlled vocabulary was a collaboration with the Gene Ontology Annotation Database (GOA) (12) to ensure mapping of our terminologies. This collaboration is ongoing to ensure synchronization and the further development of the controlled vocabularies. In September 2007, subcell.txt contains 314 'location', 10 'topology' and 8 'orientation' terms and there are 158'596 SUBCELLULAR LOCATION comments in UniProtKB/Swiss-Prot with 3329 specific annotations for isoforms/peptides/chains.

Virus annotation. For the last three years, a special effort has been ongoing in UniProtKB/Swiss-Prot to annotate viruses in the framework of the new virus annotation program. Viruses are highly specialized organisms that often display unusual molecular functions. Their diversity is enormous with more than 73 families, each having a different replication cycle. Because of this, it became an annotation priority to integrate specific viral information into UniProtKB/Swiss-Prot. The initial focus is on important human pathogens such as HIV, Influenza, Hepatitis C, Rabies, SARS, Ebola, Dengue and Yellow fever viruses.

For each family, the taxonomy is standardized and updated according to recent publications and the International Committee for Taxonomy of Viruses (ICTV) guidelines. There are a huge number of sequences available for some viruses like HIV and Hepatitis C (HIV is the organism with the most entries in the EMBL-Bank/DDBJ/GenBank nucleotide sequence databases). For these viruses, representative strains of each subtype or genotype are chosen for annotation in order to cover the whole diversity. Annotation for a well-studied strain (often called: type species) can then be propagated as appropriate to each of the related strains or isolates. Since all viral proteins are synthesized within an infected host, a new line was introduced in viral entries to display the host organism(s). Virus–host proteins interactions are critical for virus replication. These interactions are now annotated in the virus protein entries as well as in the concerned host entries, and a specific keyword has been created (Host–virus interaction) in order to further facilitate access to this information.

Annotation propagation using PIRSF-based manually curated rules. PIRSF classification-based manually curated rules are used for annotation propagation (5,6). Specifically, family-specific site rules are created for annotating structural features such as active sites, binding sites, modified residues and other functionally important residues. This information is then propagated to the rest of the members of a given PIRSF based on a structure-guided multiple alignment and profile HMM created for each PIRSF. Only residues that satisfy the rule criteria are ultimately propagated to entries within the PIRSF that lack an experimentally derived structure. For curation of ligand-binding sites, a systematic 'ligand-centric' approach is being followed. This involves selecting a biologically relevant ligand and then mapping all the structures bound to this ligand from the PDB on to PIRSFs. The liganded structure within each PIRSF serves as a template for curation of the binding sites. For example, all available structures bound to the ligand S-adenosyl-*L*-methionine are mapped to about 90 PIRSFs. Site rules have been created for each of these families and the information is being integrated into UniProtKB records. This systematic approach will enable proper naming and error-free propagation of these sites. This will eventually cover the ligand space and will enable the identification of conserved motifs and patterns.

DATABASE ACCESS AND FEEDBACK

UniProt is freely available for both commercial and non-commercial use. Please see http://www.uniprot.org/terms for details. 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 three weeks except for UniMES, which is updated only when the underlying source data are updated. Statistics are available with each release at www.uniprot.org.We are constantly trying to improve our database in terms of accuracy and representation and hence, consider your feedback extremely valuable. Please contact us if you have any questions (http://www.uniprot.org/support/helpdesk.shtml) or comments (http://www.uniprot.org/support/feedback.shtml). You can also subscribe to e-mail alerts (http://www.uniprot.org/support/alerts.shtml) for the latest information on UniProt databases.

ACKNOWLEDGEMENTS

UniProt is mainly supported by the National Institutes of Health (NIH) grant 2 U01 HG02712-05. Additional support for the EBI's involvement in UniProt comes from the European Commission contract FELICS (021902RII3) and from the NIH grant 2 P41 HG002273-07. UniProtKB/Swiss-Prot activities at the SIB are supported by the Swiss Federal Government through the Federal Office of Education and Science, by the European Commission contract FELICS (021902RII3) and by the NIH/NIAID (HHSN 2662040035C ADB contract number N01-AI-40035). PIR activities are also supported by the NIH grants and contracts HHSN266200400061C, NCI-caBIG and 1R01GM080646-01, and the National Science Foundation (NSF) grant IIS-0430743. Funding to pay the Open Access publication charges for this article

was provided by the National Institutes of Health (NIH) (grant 2 U01 HG02712-05).

Conflict of interest statement. None declared.

REFERENCES

1. Leinonen,R., Diez,F.G., Binns,D., Fleischmann,W., Lopez,R. and Apweiler,R. (2004) UniProt archive. *Bioinformatics*, **20**, 3236–3237.
2. Wieser,D., Kretschmann,E. and Apweiler,R. (2004) Filtering erroneous protein annotation. *Bioinformatics*, **20**, i342–i347.
3. Gattiker,A., Michoud,K., Rivoire,C., Auchincloss,A.H., Coudert,E., Lima,T., Kersey,P., Pagni,M., Sigrist,C.J. *et al.* (2004) Automated annotation of microbial proteomes in SWISS-PROT. *Comput. Biol. Chem.*, **27**, 49–58.
4. Fleischmann,W., Moller,S., Gateau,A. and Apweiler,R. (1999) A novel method for automatic functional annotation of proteins. *Bioinformatics*, **15**, 228–233.
5. Wu,C.H., Nikolskaya,A., Huang,H., Yeh,L.S., Natale,D.A., Vinayaka,C.R., Hu,Z.Z., Mazumder,R., Kumar,S. *et al.* (2004) PIRSF: family classification system at the Protein Information Resource. *Nucleic Acids Res.*, **32**, D112–D114.
6. Natale,D.A., Vinayaka,C.R. and Wu,C.H. (2005) Large-scale, classification-driven, rule-based functional annotation of proteins. In Subramaniam,S. (ed), *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*, John Wiley & Sons, Ltd. West Sussex, England. Vol. 7, pp. 2993–3004.
7. Suzek,B.E., Huang,H., McGarvey,P., Mazumder,R. and Wu,C. H. (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, **23**, 1282–1288.
8. Yooseph,S., Sutton,G., Rusch,D.B., Halpern,A.L., Williamson,S.J., Remington,K., Eisen,J.A., Heidelberg,K.B., Manning,G. *et al.* (2007) The Sorcerer II Global Ocean Sampling Expedition: expanding the universe of protein families. *PLoS Biol.*, **5**, e16.
9. Brunak,S., Danchin,A., Hattori,M., Nakamura,H., Shinozaki,K., Matise,T. and Preuss,D. (2002) Nucleotide Sequence Database Policies. *Science*, **298**, 1333.
10. Wu,C.H., Huang,H., Nikolskaya,A., Hu,Z. and Barker,W.C. The iProClass integrated database for protein functional analysis. (2004) *Comput. Biol. Chem.*, **28**, 87–96.
11. Becker,K.G., Barnes,K.C., Bright,T.J. and Wang,S.A. (2004) The genetic association database. *Nat. Genet.*, **36**, 431–432.
12. Camon,E., Magrane,M., Barrell,D., Lee,V., Dimmer,E., Maslen,J., Binns,D., Harte,N., Lopez,R. *et al.* (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in UniProt with Gene Ontology. *Nucleic Acids Res.*, **32**, D262–D266.

APPENDIX

UniProt has been prepared by:

Amos Bairoch, Lydie Bougueleret, Severine Altairac, Valeria Amendolia, Andrea Auchincloss, Ghislaine Argoud Puy, Kristian Axelsen, Delphine Baratin, Marie-Claude Blatter, Brigitte Boeckmann, Laurent Bollondi, Emmanuel Boutet, Silvia Braconi Quintaje, Lionel Breuza, Alan Bridge, Virginie Bulliard-Le Saux, Edouard deCastro, Luciane Ciampina, Danielle Coral, Elisabeth Coudert, Isabelle Cusin, Fabrice David, Gwennaelle Delbard, Dolnide Dornevil, Paula Duek-Roggli, Severine Duvaud, Anne Estreicher, Livia Famiglietti, Nathalie Farriol-Mathis, Serenella Ferro, Marc Feuermann, Elisabeth Gasteiger, Alain Gateau, Sebastian Gehant, Vivienne Gerritsen, Arnaud Gos, Nadine Gruaz-Gumowski, Ursula Hinz, Chantal Hulo, Nicolas Hulo, Alessandro Innocenti, Janet James, Eric Jain, Silvia Jimenez, Florence Jungo, Vivien Junker, Guillaume Keller, Corinne Lachaize, Lydie Lane-Guermonprez, Petra Langendijk-Genevaux, Vicente Lara, Philippe Le Mercier, Damien Lieberherr, Tania de Oliveira Lima, Veronique Mangold, Xavier Martin, Karine Michoud, Madelaine Moinat, Anne Morgat, Marisa Nicolas, Salvo Paesano, Ivo Pedruzzi, David Perret, Isabelle Phan, Sandrine Pilbout, Violaine Pillet, Sylvain Poux, Monica Pozzato, Nicole Redaschi, Sorogini Reynaud, Catherine Rivoire, Bernd Roechert, Claudia Sapsezian, Michel Schneider, Christian Sigrist, Karin Sonesson, Sylvie Staehli, Andre Stutz, Shyamala Sundaram, Michael Tognolli, Laure Verbregue, Anne-Lise Veuthey, Claudia Vitorello, Lina Yip and Luiz Fernando Zuletta at the Swiss Institute of Bioinformatics (SIB) and the Medical Biochemistry Department of the University of Geneva.

Rolf Apweiler, Yasmin Alam-Faruque, Daniel Barrell, Lawrence Bower, Paul Browne, Wei Mun Chan, Louise Daugherty, Emilio Salazar Donate, Ruth Eberhardt, Alexander Fedotov, Rebecca Foulger, Gabriella Frigerio, John Garavelli, Renato Golin, Alan Horne, Julius Jacobsen, Michael Kleen, Paul Kersey, Kati Laiho, Duncan Legge, Michele Magrane, Maria Jesus Martin, Patricia Monteiro, Claire O'Donovan, Sandra Orchard, John O'Rourke, Samuel Patient, Manuela Pruess, Andrey Sitnov, Eleanor Whitfield, Daniela Wieser, Quan Lin, Mark Rynbeek, Giuseppe di Martino, Mike Donnelly and Pieter van Rensburg at the European Bioinformatics Institute (EBI).

Cathy Wu, Cecilia Arighi, Leslie Arminski, Winona Barker, Yongxing Chen, Daniel Crooks, Zhang-Zhi Hu, Hsing-Kuo Hua, Hongzhan Huang, Robel Kahsay, Raja Mazumder, Peter McGarvey, Darren Natale, Anastasia N. Nikolskaya, Natalia Petrova, Baris Suzek, Sona Vasudevan, C. R. Vinayaka, Lai Su Yeh, and Jian Zhang at the Protein Information Resource (PIR).