*D142–D148* *Nucleic Acids Research, 2010, Vol. 38, Database issue* *Published online 20 October 2009*
*doi:10.1093/nar/gkp846*

# The Universal Protein Resource (UniProt) in 2010

**The UniProt Consortium*,†**

The EMBL Outstation, The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK, Swiss Institute of Bioinformatics, Centre Medical Universitaire, 1 rue Michel Servet, 1211 Geneva 4, Switzerland, Protein Information Resource, Georgetown University Medical Center, 3300 Whitehaven St. NW, Suite 1200, Washington, DC 20007 and University of Delaware, 15 Innovation Way, Suite 205, Newark, DE 19711, USA

Received September 15, 2009; Accepted September 22, 2009



## ABSTRACT

**The primary mission of UniProt is to support biological research by maintaining 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 is produced by the UniProt Consortium which consists of groups from the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR). UniProt is comprised of four major components, each optimized for different uses: the UniProt Archive, the UniProt Knowledgebase, the UniProt Reference Clusters and the UniProt Metagenomic and Environmental Sequence Database. UniProt is updated and distributed every 3 weeks and can be accessed online for searches or download at http://www.uniprot.org.**

## INTRODUCTION

UniProt strives to provide a centralized repository of protein sequences with comprehensive coverage and a systematic approach to protein annotation, incorporating, interpreting, integrating and standardizing data from large and disparate sources and is 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 while the UniProt Metagenomic and Environmental Sequences database (UniMES) was created to respond to the expanding area of metagenomic data. UniProt is freely and easily accessible by researchers to conduct interactive and custom-tailored analyses for proteins of interest to facilitate hypothesis generation and knowledge discovery.

## THE UNIPROT DATABASES

### The UniProt Knowledgebase (UniProtKB)

UniProtKB consists of two sections, UniProtKB/Swiss-Prot and UniProtKB/TrEMBL. The former contains manually annotated records with information extracted from literature and curator-evaluated computational analysis. Annotation is done by biologists with specific expertise to achieve accuracy. In UniProtKB/Swiss-Prot, 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, developmental specific expression, structure, interactions, splice isoform(s), associated diseases or deficiencies or abnormalities, etc. Another important part of the annotation process involves merging of different reports for a single protein. After an inspection of the sequences, the curator selects the reference sequence, does the corresponding merging, and lists the splice and genetic variants along with disease information when available. UniProtKB/TrEMBL contains high quality 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 (UniRule) (3–6) based on protein families. UniProtKB/TrEMBL contains the translations of all coding sequences (CDS) present

*Correspondence should be addressed to Rolf Apweiler. Tel: +44 1223 494435; Fax: +44 1223 494468; Email: apweiler@ebi.ac.uk
†The members of the UniProt Consortium are given in the Acknowledgements.

in the EMBL/GenBank/DDBJ Nucleotide Sequence Databases (7) and sequences from TAIR Arabidopsis thaliana (8), SGD (9) and Ensembl *Homo sapiens* (10) with some defined exclusions. 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 UniProtKB (including splice forms as separate entries) and selected records from UniParc to achieve complete coverage of sequence space at identity levels of 100, 90 and 50% while hiding redundant sequences (11). The UniRef clusters are generated in a hierarchical manner; 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. Each individual member sequence can exist in only one UniRef cluster at each identity level and have only one parent or child cluster at another identity level. UniRef100, UniRef90 and UniRef50 yield database size reductions of ~11, 40 and 72%, respectively. Each cluster record contains source database, protein name and taxonomy information on each member sequence but is represented by a single selected representative protein sequence and name; the number of members and the lowest common taxonomy node are also included. UniRef100 is one of the most comprehensive non-redundant protein sequence datasets available. The reduced size of the UniRef90 and UniRef50 datasets provide faster sequence similarity searches and reduce the research bias in similarity searches by providing a more even sampling of sequence space. UniRef is 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.

### UniProt Archive (UniParc)

UniParc is the main sequence storehouse and is a comprehensive repository that reflects the history of all protein sequences (1). UniParc contains all new and revised protein sequences from all publicly available sources (http://www.uniprot.org/help/uniparc) to ensure that complete coverage is available at a single site. To avoid redundancy, 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 lacks a cross-reference to a UniProtKB entry, the reason for its exclusion 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.

### The UniProt Metagenomic and Environmental Sequences database (UniMES)

The UniProt Knowledgebase contains entries with a known taxonomic source. However, the expanding area of metagenomic data has necessitated the creation of a separate database, the UniProt Metagenomic and Environmental Sequences database (UniMES). UniMES currently contains data from the Global Ocean Sampling Expedition (GOS) which predicts nearly 6 million proteins, primarily from oceanic microbes. 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. UniMES is available on the ftp site in FASTA format along with a UniMES matches to InterPro methods file.

## MANUAL ANNOTATION IN UNIPROTKB

UniProtKB/Swiss-Prot contains manually annotated records with information extracted from literature and curator-evaluated computational analysis. Manual annotation consists of a critical review of experimentally proven or computer-predicted data about each protein, including the protein sequences. Records are continuously created and updated by an expert team of biologists.

The annotation activities of the UniProtKB/Swiss-Prot can be divided into two parts.

### Model organism-oriented annotation

UniProtKB/Swiss-Prot provides annotated entries for many species, but concentrates on the annotation of entries from model organisms of distinct taxonomic groups to ensure the presence of high quality annotation for representative members of all protein families:

- Human and other mammals
- Non-mammalian vertebrates: *Xenopus*, Zebrafish
- Bacteria and Archaea
- Plants
- Fungi
- Viruses
- Toxins
- Dictyostelium
- Arthropods: *Drosophila, mosquito*
- *C. elegans and C. briggsae*

### Transversal annotation

Transversal annotation focuses on issues common to all organisms, such as post-translational modifications (PTMs), structural information and protein–protein interactions.

For more information on UniProt's annotation programs, please see http://www.uniprot.org/help/projects.

## PROGRESS REPORT

### UniProtKB annotation

*Revisiting the human proteome*. Following the publication in UniProt release 14.1 of the first draft of the complete human proteome, an intensive review and update of the 20 325 records was initiated. Our main objectives are to increase the depth and quality of human protein annotation and to continue to update and correct all associated protein sequences.

As part of the review process, we are using information extraction tools such as the STRING database (12) to identify UniProt entries which are candidates for re-annotation. STRING is a meta-database that integrates and assigns reliability scores to information on functional protein interactions and as such provides a useful first pass filter for re-annotation prioritization. Propagation of the annotation from well-characterized orthologs in closely related species (e.g. *Mus musculus*) to an uncharacterized human protein is another approach used. Sequence update and review includes the merging of previously undescribed splice isoforms and polymorphisms and the correction or removal of erroneous sequences by comparison to the reference human genome. We also continue to create records for newly discovered protein sequences and to delete spurious records which may correspond to pseudogenes or cloning artifacts. UniProt recently joined the Consensus CDS (CCDS) project (13), a collaborative effort to identify a core set of consistently annotated and high quality human and mouse protein-coding regions. The long-term goal is to support convergence towards a standard set of gene and protein annotations. To date, UniProt has investigated 700 records in close collaboration with the RefSeq annotation group at the NCBI and the HAVANA team at the WTSI. UniProtKB/Swiss-Prot release 57.6 contains 20 330 human proteome entries. More than one third of these contain additional sequences representing isoforms generated by alternative splicing, alternative promoter usage and/or alternative translation initiation, resulting in close to 34 000 human protein sequences. Approximately 58 000 single amino acid polymorphisms (SAPs), mostly disease-linked, are also described as well as 69 000 PTMs. This release of UniProtKB/Swiss-Prot also includes over 80 000 vertebrate proteins including 16 163 mouse proteins.

*Complete Schizosaccharomyces pombe proteome in UniProtKB/Swiss-Prot*. The annotation effort towards proteins from model organisms recently led to the integration into UniProtKB/Swiss-Prot of the complete set of proteins encoded by *S. pombe*, the sixth eukaryotic organism to be completely sequenced (14). It is the third eukaryotic complete proteome (after *S. cerevisiae* and human) to be integrated in UniProtKB/Swiss-Prot. Since UniProt release 15.4, users can access 4958 sequence-verified, manually curated protein entries, including a link to GeneDB_Spombe, the fission yeast community database. The *S. pombe* proteome set will not be a static one but will be revisited and updated as science progresses in this field.

The availability of the complete set of both *S. cerevisiae* and *S. pombe* proteins and the phylogenetic position of these two species in the fungal tree of life will aid in the identification and annotation of orthologous proteins in many other organisms.

*Viral protein annotation program (VPAP)*. Started in 2004, the viral protein annotation program seeks to provide a detailed review on viral proteins from a representative subset of strains for each virus. For this purpose, we focused on the NCBI Reference Sequences (RefSeq) listed strains, which have the benefit of being both fully sequenced and representative.

UniProtKB/Swiss-Prot release 57.6 contains 14 233 annotated viral entries. Viral entries are created or updated to describe the protein's functionality and characteristics, such as 3D structure, functional domains, localization in the host cell, or post-translational modifications. The annotation also includes data concerning the infectious cycle or interactions with the host proteins (e.g. intracellular machinery, host cell immunity, host entry receptors), as well as a precise description of the host organism, often leading to taxonomy updates. In order to have the most up-to-date annotation, we frequently collaborate with virologists. Special emphasis is put on viruses of public health importance, especially those that are causative agents of human epidemics. As a result, we have fully annotated proteins from HIV, influenza, SARS, hepatitis A, hepatitis C, hepatitis E, Ebolavirus, caliciviruses, Chikungunya virus, Dengue virus, lyssaviruses and Rubella virus. In 2009, we completed the annotation of rotaviruses, Epstein-Barr virus, Varicella-zoster virus and Herpes simplex virus and have added a set of representative strains of the H1N1 swine flu 2009 outbreak.

*Dictyostelium annotation program*. UniProt and dictyBase, the model organism database (MOD) for *Dictyostelium discoideum*, have established a collaboration to improve data sharing. This collaboration, started in 2008, took a major step forward with a jointly organized annotation marathon. The 1-week marathon led to the annotation into UniProtKB/Swiss-Prot of more than 1000 *D. discoideum* proteins, plus the updating of a large number of gene symbols, protein names and gene models (15). This collaboration continues in a new annotation program that was established at the end of 2008. The program's main priority is to annotate *D. discoideum* proteins that have been characterized or whose gene models have been manually verified by dictyBase. It will also include work on gene symbols, protein names and gene model updates.

*Integration with other databases*. Integration between the three types of sequence-related databases (nucleic acid sequences, protein sequences and protein tertiary

structures) as well as with specialized data collections is important to our users. UniProtKB is currently cross-referenced with more than 10 million links to 114 different databases with regular update cycles. Table 1 lists the 17 new and diverse databases added in the past year. Database cross-references are stored in the DR (Database Reference) lines and allow access to related information in other databases. This extensive network of cross-references allows UniProt to act as a focal point of biomolecular database interconnectivity. All cross-referenced databases are documented at http://www.uniprot.org/docs/dbxref and if appropriate are included in the UniProt ID mapping tool at http://www.uniprot.org/help/mapping with the file for download at ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping.

| Database name | Database information |
|---|---|
| Bgee | dataBase for Gene Expression Evolution |
| CTD | Comparative Toxicogenomics Database |
| CAZy | Carbohydrate-Active enzymes |
| GeneCards | GeneCards: human genes, protein and diseases |
| IPI | International Protein Index |
| NextBio | NextBio gene-centric data for human, mouse, rat, fly, worm and yeast |
| OMA | Identification of Orthologs from Complete Genome Data |
| Pathway_Interaction_DB | NCI-Nature Pathway Interaction Database |
| PMAP_CutDB | CutDB - Proteolytic event database |
| PRIDE | PRoteomics IDEntifications database |
| STRING | STRING: functional protein association networks |
| TCDB | Transport Classification Database |
| UCSC | University of California Santa Cruz Genome Browser |
| XenBase | *Xenopus laevis* and *tropicalis* biology and genomics resource |

## Controlled vocabularies

Controlled vocabularies (CVs) (http://www.uniprot.org/docs/#vocabulary) are used to describe various UniProt annotation items such as keywords, plasmids or subcellular locations.

*Plastid annotation controlled vocabulary*. In the OG (OrGanelle) line (Encoded on subsection of the 'Names and origin' section, (http://www.uniprot.org/manual/encoded_on), six general terms for plastids are used:

- 'Chloroplast' indicates the organism is photosynthetic.
- 'Non-photosynthetic plastid' is used when the organism is from a photosynthetic lineage but genetically unable to photosynthesize, as happens with some parasitic plants (*Epifagus virginiana*, *Aneura mirabilis*), a parastic 'green' algae (Helicosporidium sp. Subsp. *Simulium jonesii*) and a euglenoid (*Astasia longa*).
- 'Cyanelle' is used for the plastid of the glaucophyte algae. It has the remnants of a cell wall between its surrounding membranes.
- 'Apicoplast' is used for plastids from the non-photosynthetic Apicocomplexan parasites such as Plasmodium, Toxoplasma and Eimeria which cause malaria, toxoplasmosis and coccidian diseases respectively. Although the plastid remnant has a reduced coding capacity, it is essential for cell survival and as such is interesting as a drug target.
- 'Organellar chromatophore' is used for the plastid of the thecate amoeba Paulinella chromatophora, which has a very large endosymbiont genome (1.0 Mb, encoding almost 900 proteins).
- 'Plastid' (without any qualifier) is used for some parasitic plants (mostly from the genus Cuscuta) which may be briefly photosynthetic when very young.

*Structuring of the metabolic pathway topic using UniPathway*. The metabolism of living organisms can be understood as a network of biochemical reactions, generally catalyzed by enzymes. Dealing with this network as a whole is a complex task and a classical approach is to divide it into more manageable segments, called pathways. How this approach is applied is always somewhat arbitrary and depends upon the final usage of the data. Usually, a first level of segmentation is achieved on the basis of biological criteria. For instance, divisions could be achieved by considering the sub-network of all reactions involved in amino-acid biosynthesis or, more specifically, in L-lysine biosynthesis only, or even more specifically, in L-lysine biosynthesis via the AAA pathway. This results in a series of coarse- to fine-grained divisions (the coarsest is called a 'super-pathway').

We have followed this classical approach in the UniPathway project (a collaborative project with the INRIA and the LECA), and have, whenever possible, further refined this first-level segmentation to a second-level one, in order to split the pathways into linear segments (i.e. sub-networks without branches) called 'sub-pathways'. Such a fine-grained segmentation allows representation of pathway variants. Indeed, depending on the organism (or set of organisms), the chemical route from one compound to another can be performed in different ways. It is important to represent these variations within the same pathway since UniProtKB covers a large number of species. In addition, it offers a convenient way to label the enzymatic reactions that constitute a metabolic pathway by their relative position ('step') in the sub-pathway.

The role of a protein in metabolism is described in the 'Pathway' subsection of the 'General annotation (Comments)' section. The syntax is 'super-pathway; pathway; sub-pathway: step n/m'

Examples:
**P49367**: *S. cerevisiae* homoacotinase (EC 4.2.1.36) catalyzes the second step of the sub-pathway L-α-aminoadipate from 2-oxoglutarate (transformation of 2-oxoglutarate into L-α-aminoadipate through 4 enzymatic reactions), a component of the L-lysine biosynthesis via AAA pathway

```
CC   -!- PATHWAY: Amino-acid biosynthesis;
         L-lysine biosynthesis via AAA
CC       pathway; L-alpha-aminoadipate from
         2-oxoglutarate: step 2/4.
```

**Q980X0**: *S. solfataricus* acetylglutamate/acetylaminoadipate kinase (EC 2.7.2.8 & EC 2.7.2.-) catalyzes two reactions involved in two independent pathways.

```
CC   -!- PATHWAY: Amino-acid biosynthesis;
         L-arginine biosynthesis; N(2)-
CC       acetyl-L-ornithine from L-
         glutamate: step 2/4.
CC   -!- PATHWAY: Amino-acid biosynthesis;
         L-lysine biosynthesis via AAA
CC       pathway; L-lysine from L-alpha-
         aminoadipate (Thermus route): step
CC       2/5.
```

The UniProt web site supplies direct links to UniPathway (http://www.grenoble.prabi.fr/obiwarehouse/unipathway), which provides more detailed information on pathways, sub-pathways and biochemical reactions. By making use of UniProtKB/Swiss-Prot richness, UniPathway is able to offer several perspectives to understand metabolism: a protein perspective, a genomic perspective and a taxonomic perspective. The chemical perspective is based on KEGG LIGAND compounds and reactions with the kind permission of the Kanehisa Laboratory (16).

UniProtKB/Swiss-Prot release 57.6 contains more than 105 000 distinct proteins (~155 000 PATHWAY annotations) annotated with the UniPathway controlled vocabulary.

*Enzyme nomenclature in UniProt*. EC numbers are used to describe enzyme reactions and are based on the recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB). Unfortunately, not all known enzyme reactions have an EC number assigned yet. Until recently, these reactions were assigned so-called partial EC numbers where part of the numbers were replaced by dashes (e.g. EC 3.4.24.-). Partial EC numbers were used whenever the catalytic activity of the protein was not known exactly, or when the protein catalyzes a reaction which is known but not yet included in the IUBMB EC list. To distinguish between these two meanings, we have started to use the letter 'n' with a serial number instead of a dash '-'for enzymes with known catalytic activities. The serial numbers are included to ensure that each preliminary EC number is unique.

Examples:
The catalytic activity of the protein is not known exactly:
**Q9VAC5**: DE ADAM 17-like protease precursor (EC 3.4.24.-).

The catalytic reaction is known, but not yet included in the IUBMB's EC list:
**Q01468**: DE 4-oxalocrotonate tautomerase (EC 5.3.2.n1) (4-OT).

Preliminary EC numbers are used in UniProtKB/Swiss-Prot and are also included in the ENZYME database. They are thus available through same channels as the ENZYME database, i.e. the ENZYME site (http://www.expasy.org/enzyme/) and from our ftp server.

### UniProtKB additional protein bibliography information

UniProt strives to provide comprehensive literature citations on which UniProtKB protein annotations are based. Currently, there are ~228 000 distinct PubMed citations associated with ~4.2 million UniProtKB sequences and 67% of these citations are in UniProtKB/Swiss-Prot. Databases such as Entrez Gene and MODs (e.g. dictyBase, SGD, and MGI) also provide curated literature information, which reflect their priorities and focus. We have now integrated literature annotations from 11 external gene or protein databases, including GeneRIF of Entrez Gene (http://www.ncbi.nlm.nih.gov/projects/GeneRif), PDB (http://www.rcsb.org/pdb) and 9 MODs: SGD (http://www.yeastgenome.org), MGI (http://www.informatics.jax.org), GAD (geneticassociationdn.nih.gov), dictyBase (http://www.dictybase.org), ZFIN (http://www.zfin.org), WormBase (http://www.wormbase.org), TAIR (http://www.arabidopsis.org), RGD (rgd.mcw.edu) and FlyBase (http://www.flybase.org). These 11 external sources contribute ~350 000 unique PubMed citations not yet annotated in UniProtKB, covering ~188 000 UniProtKB entries. The additional bibliography is directly linked from the protein entry view on the UniProt website. We continue to identify more sources of bibliography information to enhance the UniProtKB bibliography and to allow scientific users to better explore the existing knowledge on their proteins of interest.

## DATABASE ACCESS AND FEEDBACK

The http://www.uniprot.org website [17] is the primary access point to our data and documentation and to tools such as full text and field-based text search, sequence similarity search, multiple sequence alignment, batch retrieval and database identifier mapping. These tools can be accessed directly through a tool bar that appears at the top of every page. Most data (including documentation and help) can be searched through the full text search, which allows searches requiring no prior knowledge of our data or search syntax. Results are sorted by relevance and, where possible, suggestions are provided to help refine searches that yield too many or no results. The field-based text search supports more complex queries. These can be built iteratively with the tool bar's query builder or entered manually in the query field, which can be faster and more powerful (http://www.uniprot.org/help/text-search). Searching with ontology terms is assisted by auto-completion, and we also provide the possibility of using ontologies to browse search results. Viewing of result sets, as well as database entries, is configurable. The site has a simple and consistent URL scheme and all searches can be bookmarked to be repeated at a later time. The home page features a site-tour as a quick introduction for novice users. In response to user requests for various downloadable data sets (e.g. all reviewed human entries in FASTA format), we have removed all download limits to allow this functionality by directly querying the website. However, large

downloads are given low priority in order to ensure that they do not interfere with interactive queries, and they can therefore be slow compared to downloads from the UniProt FTP server. We therefore recommend downloading complete datasets from ftp.uniprot.org/pub/databases. The website offers various download formats which depend on the chosen dataset (e.g. plain text, XML, RDF, FASTA, GFF for UniProtKB). The columns of result tables can be configured for customized downloads in tab-delimited or Excel format. All data is available in RDF (http://www.w3.org/RDF/), a W3C standard for publishing data on the Semantic Web. All search results can be retrieved as RSS feeds for integration with external tools such as news feed readers or Yahoo Pipes. Programmatic access to data and search results is possible via simple HTTP (REST) requests (http://www.uniprot.org/help/technical). Java applications can also make use of our Java API (UniProtJAPI) (18).

We are constantly trying to improve our databases and services in terms of accuracy and representation and hence, consider your feedback extremely valuable. Please contact us if you have any questions via http://www.uniprot.org/contact or email us directly at help@uniprot.org. The page http://www.uniprot.org/help/submissions provides information about data submissions and updates. You can also subscribe to e-mail alerts (http://www.uniprot.org/help/alerts) for the latest information on UniProt databases. Extensive documentation on how to best use our resource is available at http://www.uniprot.org/help/. UniProt is freely available for both commercial and non-commercial use. Please see http://www.uniprot.org/help/license for details. New releases are published every 3 weeks except for UniMES, which is updated only when the underlying source data are updated. Statistics are available with each release at http://www.uniprot.org.

## ACKNOWLEDGEMENTS

UniProt has been prepared by: Rolf Apweiler, Maria Jesus Martin, Claire O'Donovan, Michele Magrane, Yasmin Alam-Faruque, Ricardo Antunes, Daniel Barrell, Benoit Bely, Mark Bingley, David Binns, Lawrence Bower, Paul Browne, Wei Mun Chan, Emily Dimmer, Ruth Eberhardt, Alexander Fedotov, Rebecca Foulger, John Garavelli, Rachael Huntley, Julius Jacobsen, Michael Kleen, Kati Laiho, Rasko Leinonen, Duncan Legge, Quan Lin, Wudong Liu, Jie Luo, Sandra Orchard, Samuel Patient, Diego Poggioli, Manuela Pruess, Matt Corbett, Giuseppe di Martino, Mike Donnelly and Pieter van Rensburg at the European Bioinformatics Institute (EBI); Amos Bairoch, Lydie Bougueleret, Ioannis Xenarios, Severine Altairac, Andrea Auchincloss, Ghislaine Argoud-Puy, Kristian Axelsen, Delphine Baratin, Marie-Claude Blatter, Brigitte Boeckmann, Jerven Bolleman, Laurent Bollondi, Emmanuel Boutet, Silvia Braconi Quintaje, Lionel Breuza, Alan Bridge, Edouard deCastro, Luciane Ciapina, Danielle Coral, Elisabeth Coudert, Isabelle Cusin, Gwennaelle Delbard, Mikael Doche, Dolnide Dornevil, Paula Duek Roggli, Severine Duvaud, Anne Estreicher, Livia Famiglietti, Marc Feuermann, Sebastien Gehant, Nathalie Farriol-Mathis, Serenella Ferro, Elisabeth Gasteiger, Alain Gateau, Vivienne Gerritsen, Arnaud Gos, Nadine Gruaz-Gumowski, Ursula Hinz, Chantal Hulo, Nicolas Hulo, Janet James, Silvia Jimenez, Florence Jungo, Thomas Kappler, Guillaume Keller, Corinne Lachaize, Lydie Lane-Guermonprez, Petra Langendijk-Genevaux, Vicente Lara, Philippe Lemercier, Damien Lieberherr, Tania de Oliveira Lima, Veronique Mangold, Xavier Martin, Patrick Masson, Madelaine Moinat, Anne Morgat, Anais Mottaz, Salvo Paesano, Ivo Pedruzzi, Sandrine Pilbout, Violaine Pillet, Sylvain Poux, Monica Pozzato, Nicole Redaschi, Catherine Rivoire, Bernd Roechert, Michel Schneider, Christian Sigrist, Karin Sonesson, Sylvie Staehli, Eleanor Stanley, Andre Stutz, Shyamala Sundaram, Michael Tognolli, Laure Verbregue, Anne-Lise Veuthey, Lina Yip and Luiz Zuletta at the Swiss Institute of Bioinformatics (SIB) and the Biochemistry and Structural Biology Department of the University of Geneva; Cathy Wu, Cecilia Arighi, Leslie Arminski, Winona Barker, Chuming Chen, Yongxing Chen, Zhang-Zhi Hu, Hongzhan Huang, Raja Mazumder, Peter McGarvey, Darren A. Natale, Jules Nchoutmboube, Natalia Petrova, Nisha Subramanian, Baris E. Suzek, Uzoamaka Ugochukwu, Sona Vasudevan, C. R. Vinayaka, Lai Su Yeh and Jian Zhang at the Protein Information Resource (PIR).

## FUNDING

UniProt is mainly supported by Award Number U01HG02712 from the National Human Genome Research Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Human Genome Research Institute or the National Institutes of Health (NIH). Additional support for the EBI's involvement in UniProt comes from the European Commission contract SLING grant (226073) and from the NIH grant (2P41HG02273-07). UniProtKB/Swiss-Prot activities at the SIB are supported in addition from the Swiss Federal Government through the Federal Office of Education and Science and from the European Commission contract SLING (226073). PIR activities are also supported by the NIH grants and contracts on HHSN266200400061C, NCI-caBIG and 5R01GM080646-04 and the Department of Defense grant W81XWH0720112. Funding for open access charge: European Bioinformatics Institute.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Leinonen,R., Diez,F.G., Binns,D., Fleischmann,W., Lopez,R. and Apweiler,R. (2009) 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., Lachaize,C.

*et al*. (2003) 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., Kourtesis,P. *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. (2004) Large-scale, classification-driven, rule-based functional annotation of proteins. In Subramaniam,S. (ed.), *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics. Bioinformatics Volume*. John Wiley & Sons, Ltd., West Sussex, England.
7. Cochrane,G., Akhtar,R., Bonfield,J., Bower,L., Demiralp,F., Faruque,N., Gibson,R., Hoad,G., Hubbard,T., Hunter,C. *et al*. (2009) Petabyte-scale innovations at the European Nucleotide Archive. *Nucleic Acids Res*., **37**, D19–D25.
8. Swarbreck,D., Wilks,C., Lamesch,P., Berardini,T.Z., Garcia-Hernandez,M., Foerster,H., Li,D., Meyer,T., Muller,R., Ploetz,L. *et al*. (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res*., **36**, D1009–D1014.
9. Hong,E.L., Balakrishnan,R., Dong,Q., Christie,K.R., Park,J., Binkley,G., Costanzo,M.C., Dwight,S.S., Engel,S.R., Fisk,D.G. *et al*. (2008) The Ontology annotations at SGD: new data sources and annotation methods. *Nucleic Acids Res*., **36**, D577–D581.
10. Flicek,P., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F., Cutts,T. *et al*. (2008) Ensembl 2008. *Nucleic Acids Res*., **36**, D707–D714.
11. 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.
12. Jensen,L.J., Kuhn,M., Stark,M., Chaffron,S., Creevey,C., Muller,J., Doerks,T., Julien,P., Roth,A., Simonovic,M. *et al*. (2009) STRING 8 – a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res*., **37**, D412–D416.
13. Pruitt,K.D., Harrow,J., Harte,R.A., Wallin,C., Diekhans,M., Maglott,D.R., Searle,S., Farrell,C.M., Loveland,J.E., Ruef,B.J. *et al*. (2009) The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res*., **19**, 1316–1323.
14. Wood,V., Gwilliam,R., Rajandream,M.A., Lyne,M.H., Lyne,R., Stewart,A., Sgouros,J.G., Peat,N., Hayles,J., Baker,S.G. *et al*. (2002) The genome sequence of Schizosaccharomyces pombe. *Nature*, **415**, 871–880.
15. Gaudet,P., Lane,L., Fey,P., Bridge,A., Poux,S., Auchincloss,A., Axelsen,K., Braconi-Quintaje,S., Boutet,E., Brown,P. *et al*. (2009) Collaborative annotation of genes and proteins between UniProtKB/Swiss-Prot and dictyBase. *Database*, doi:10.1093/database/bap013.
16. Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S., Tokimatsu,T. *et al*. (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res*., **36**, D480–D484.
17. Jain,E., Bairoch,A., Duvaud,S., Phan,I., Redaschi,N., Suzek,B.E., Martin,M.J., McGarvey,P. and Gasteiger,E. (2009) Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics*, **10**, 136.
18. Patient,S., Wieser,D., Kleen,M., Kretschmann,E., Martin,M.J. and Apweiler,R. (2008) UniProtJAPI: a remote API for accessing UniProt data. *Bioinformatics*, **24**, 1321–1322.