Published online 4 October 2008 *Nucleic Acids Research, 2009, Vol. 37, Database issue* ***D169–D174***
doi:10.1093/nar/gkn664

The Universal Protein Resource (UniProt) 2009

The UniProt Consortium[1,2,3,*]

[1]The EMBL Outstation, The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK, [2]Protein Information Resource, Georgetown University Medical Center, 3300 Whitehaven St NW, Suite 1200, Washington, DC 20007, USA and [3]Swiss Institute of Bioinformatics, Centre Medical Universitaire 1 rue Michel Servet, 1211 Geneva 4, Switzerland

Received September 15, 2008; Accepted September 19, 2008

ABSTRACT

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information that is essential for modern biological research. UniProt is produced by the UniProt Consortium which consists of groups from the European Bioinformatics Institute, the Protein Information Resource and the Swiss Institute of Bioinformatics. The core activities include manual curation of protein sequences assisted by computational analysis, sequence archiving, 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 Archive, the UniProt Knowledgebase, the UniProt Reference Clusters and the UniProt Metagenomic and Environmental Sequence Database. One of the key achievements of the UniProt consortium in 2008 is the completion of the first draft of the complete human proteome in UniProtKB/Swiss-Prot. This manually annotated representation of all currently known human protein-coding genes was made available in UniProt release 14.0 with 20 325 entries. UniProt is updated and distributed every three weeks and can be accessed online for searches or downloaded at www.uniprot.org.

INTRODUCTION

High-throughput genome sequencing and the exponential growth of proteomics data is providing a rapid and accelerating accumulation of predicted protein sequences and associated data for a large number of organisms. There is a widely recognized need for a centralized repository of protein sequences with comprehensive coverage and a systematic approach to protein annotation, incorporation, integration and standardization of data from these various sources and UniProt strives to provide this.

UniProt is the central resource for storing and interconnecting information from large and disparate sources, and the most comprehensive catalogue 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.

CONTENT

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. Biologists with specific expertise do the annotation to achieve accuracy. 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, developmental specific expression, structure, interactions, splice isoform(s), associated diseases or deficiencies or abnormalities, etc. Another important part of the annotation process

*To whom correspondence should be addressed. Tel: +44 1223 494435; Fax: +44 1223 494468; Email: apweiler@ebi.ac.uk involves the merging of different reports for a single protein. After a careful 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. Any discrepancies between the different sequence sources are also annotated. Cross-references are provided to the underlying nucleotide sequence sources as well as many other useful databases including organism-specific, domain, family and disease databases. UniProtKB/TrEMBL contains high quality computationally analysed 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-Bank/GenBank/DDBJ Nucleotide Sequence Databases (7) and sequences from The Arabidopsis Information Resource (TAIR) (8), SGD (9) and Ensembl *Homo sapiens* (10). We exclude some types of data such as EMBL-Bank/GenBank/DDBJ entries that encode small fragments, synthetic sequences, most non-germline immunoglobulins and T-cell receptors, most patent sequences and some highly over-represented data. 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 (11). 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 database, protein name and taxonomy organism information on each member sequence 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 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 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. The UniRef clusters are updated with every release of UniProtKB.

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 Vertebrate Genome Annotation Database (VEGA), the International Protein Index (IPI), Protein Research Foundation (PRF), the Protein Data Bank (PDB), NCBI's Reference Sequence Collection (RefSeq), model organism databases FlyBase, SGD, TAIR and WormBase, TROME and protein sequences from the American, European, Korean and Japanese Patent Offices. To avoid redundancy, sequences are handled as strings—all sequences 100% identical over the entire length are merged, regardless of the 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 Swiss-Prot and TrEMBL sections of the UniProt Knowledgebase contain 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 was originally submitted to the International Nucleotide Sequence Database Collaboration (INSDC). 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 UniProtKB and UniRef 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

New UniProt unified website

The UniProt consortium released its new official unified website: a new interface, a new search engine and many new options to serve its user community better. The individual mirrors (www.ebi.uniprot.org, www.expasy.uniprot.org, www.pir.uniprot.org and parts of www.expasy.org) are no longer maintained. User feedback and the analysis of the use of our previous sites have 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 we also provide the possibility of using ontologies to browse search results. The viewing of database entries is 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. Access to the following most common bioinformatics tools have been simplified: 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 (www.uniprot.org/help/technical). In addition to the existing formats that support the different data sets (e.g. plain text, FASTA and XML for UniProtKB), now it also provides (configurable) tab-delimited, RSS and GFF downloads where possible, and all data is available in RDF (www.w3.org/RDF/), a W3C standard for publishing data on the Semantic Web. Extensive documentation on how to best use this new resource is available at: www.uniprot.org/help/.

UniProtKB additional protein bibliography information

UniProt strives to provide comprehensive literature citations on which UniProtKB protein annotations are based, e.g. currently, there are ~218 000 PubMed citations annotated in ~4.1 million UniProtKB sequences and 66% of the citations are in UniProtKB/Swiss-Prot. Various other public databases such as Entrez Gene and model organism databases (MODs), e.g. SGD, MGI also provide curated literature information for respective gene or protein entries. For genes commonly annotated in different databases, each database often provides unique literature annotations reflecting the bias or the different priorities of the databases. Therefore, it is of great benefit to the scientific community to integrate additional sources of curated literature into UniProtKB. We have now integrated literature annotations from five external curated gene or protein databases covering human, mouse, yeast and other organisms:

GeneRIF of Entrez Gene (www.ncbi.nlm.nih.gov/projects/GeneRif),
SGD (www.yeastgenome.org), MGI (www.informatics.jax.org),
GAD (geneticassociationdn.nih.gov) and PDB (www.rcsb.org/pdb/).

The five external sources contribute ~244 000 unique PubMed citations not annotated in UniProtKB, covering ~110 000 UniProtKB entries. The additional bibliography is directly linked from the protein entry view on the UniProt website. We will continue to identify more sources of bibliography information from other MODs and databases of protein functions to enhance the UniProtKB bibliography. The additional bibliography information will not only facilitate the curation of UniProtKB entries, but also benefit the scientific users to better explore the existing knowledge on proteins of their interest.

Format changes

UniProt format changes occur in order to improve data consistency and usability. We strongly urge our users to monitor our newsfeeds in order to maximize the full benefit of these changes. Below are some of the major changes from recent months and those planned for the near future. Full details are available at www.uniprot.org/

Recent format changes.

(i) The UniProtKB FASTA headers were unfortunately incompatible with the -o option of the NCBI's program formatdb. We have been working with the NCBI to remedy this and changes were required on both sides. The new version of formatdb now accepts a database code for UniProtKB/TrEMBL and we have modified our UniProtKB FASTA headers accordingly. For consistency reasons, we also changed the FASTA headers of the other UniProt databases.

(ii) We have structured the UniProtKB DE lines. The new format includes three categories:

- 'RecName' is the protein name recommended by the UniProt Consortium;

- ‘AltName’ represents synonyms found in the literature or in other databases;
- ‘SubName’ is the name provided by the submitters of the underlying nucleotide sequence. It is found in UniProtKB/TrEMBL only.

Three subcategories allow the fine-tuning of the nomenclature:

- Abbreviations and acronyms are available in the ‘Short’ subcategory;
- WHO INN (International Nonproprietary Names) are found in the ‘INN’ subcategory;
- EC (Enzyme nomenclature) numbers are located in the ‘EC’ subcategory.

Each block of DE lines may also contain the sections: ‘Includes’ or ‘Contains’ and the field ‘Flags’, which indicate, for instance, whether the sequence shown is a fragment and/or a precursor.

Forthcoming format changes.

(i) The CC line topic INTERACTION conveys information about binary protein-protein interactions. Currently, all interaction data are automatically derived from the IntAct database. In the future, we will start to add manually curated binary protein-protein interactions to this topic (these are currently described in the CC line topic SUBUNIT). In order to represent isoform- and chain-specific interactions (e.g. for viral polyproteins) and to add interactor-specific comments (e.g. PTMs and binding regions), we are going to modify the format of the INTERACTION lines. Each binary interaction will be represented by a block of three to four lines:

- The first line of a block indicates the experimental evidence for the interaction and the data source (literature reference or ‘By similarity’ and/or cross-reference to the database from which the data was derived).
- The next line is an optional comment about the interaction.
- The last two lines give details on the interacting proteins: the Protein1 = line represents the currently displayed entry, the Protein2 = line the other interacting protein. If Protein2 is from a different species than Protein1, its species or taxonomic range is indicated.

Example:
CC -!- INTERACTION:
CC Interact = Yes (PubMed:11533489);
CC Comment = HDAC3 mediates the deacetylation of RELA;
CC Protein1 = RELA [Q04206];
CC Protein2 = HDAC3 [O15379].

(ii) We are going to introduce the new CC line topic DISRUPTION PHENOTYPE to describe the effects caused by the disruption of the gene coding for a protein. Note that we only describe effects caused by the complete absence of a gene and thus of a protein *in vivo* (null mutants caused by random or target deletions, insertions of a transposable element, etc.) To avoid description of phenotypes due to partial or dominant negative mutants, mis-sense mutations will not be described in this topic, but in FT MUTAGEN instead. Not all defects caused by transient inactivation using methods such as RNA interference or blockage by antibodies will be described due to the difficulty of interpreting results.

UniProtKB ANNOTATION

UniProtKB consists of two sections, Swiss-Prot and TrEMBL.

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. Data are continuously 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 (HPI);
- Bacteria and Archaea (HAMAP);
- Plants (PPAP);
- Fungi (FPAP);
- Viruses;
- Toxins (Tox-Prot);
- *Drosophila, Xenopus*, Zebrafish and *C. elegans*.

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, please see www.uniprot.org/help/projects.

First draft of the complete human proteome. A recent result of the annotation approach outlined above is the first draft of the complete human proteome in UniProtKB/Swiss-Prot. This manually annotated representation of all currently known human protein-coding genes was made available in UniProt release 14.1. At the time of release, it represents 20 325 entries. More than a 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 46 000 single amino acid polymorphisms

(SAPs), mostly disease-linked, are also described as well as 60 000 PTMs.

It is not the first time that UniProtKB/Swiss-Prot has provided a fully annotated proteome set for a model organism (e.g. *Escherichia coli* or *Saccharomyces cerevisiae*) and there are many more planned in the near and more distant future (*Arabidopsis thaliana*, *Bacillus subtilis*, *Dictyostelium discoideum*, mouse, rice, *Staphylococcus aureus*, *Schizosaccharomyces pombe*, etc.). However, there is unlikely to be anything as important as this proteome. For the first time, we can present to the life sciences community a clean set of what we believe to be a full (although still imperfect) representation of human proteins. It is the ultimate goal of the life sciences to fully understand *Homo sapiens* at the molecular level and we hope this set will significantly contribute to this. There are still many challenging tasks in front of us. We will create entries for newly discovered human proteins, review and update the existing set, increase the number of splice variants, explore the full range of PTMs and continue to build a comprehensive view of protein variation in the human population. The characterization at the molecular level will also need to be placed in its physiological context: subcellular location, tissue expression, protein-protein interaction, etc.

DATABASE ACCESS AND FEEDBACK

UniProt is freely available for both commercial and non-commercial use. Please see www.uniprot.org/help/license for details. The UniProt databases can be accessed online (www.uniprot.org) or downloaded in several formats (ftp.uniprot.org/pub/databases). 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 (www.uniprot.org/contact) or updates (www.uniprot.org/help/submissions) or email us directly at help@uniprot.org. You can also subscribe to e-mail alerts (www.uniprot.org/help/alerts) for the latest information on UniProt databases.

FUNDING

UniProt is mainly supported by the National Institutes of Health (NIH) grant (2U01HG02712-04). Additional support for the EBI's involvement in UniProt comes from the European Commission contract FELICS grant (021902RII3) and from the NIH grant (2P41HG02273-07). UniProtKB/Swiss-Prot activities at the SIB are supported by the Swiss Federal Government through the Federal Office of Education and Science. Additional support comes from the European Commission contract FELICS (091902RII3) and from the PATRIC BRC (NIH/NIAID contract HHSN 266200400035C). PIR activities are also supported by the NIH grants and contracts on proteomics (HHSN266200400061C), protein ontology (1R01GM080646-01) and grid enablement (NCI-1435-04-04-CT-73980). Funding for open access charges: 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., 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., Aldebert,P., Althorpe,N., Baldwin,A., Bates,K., Bhattacharyya,S., Bonfield,J., Bower,L., Browne,P. *et al*. (2008) Priorities for nucleotide trace, sequence and annotation data capture at the Ensembl Trace Archive and the EMBL Nucleotide Sequence Database. *Nucleic Acids Res*., **36**, D5–D12.
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.

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, 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, Dolnide Dornevil, Paula Duek Roggli, Severine Duvaud, Anne Estreicher, Livia Famiglietti, Marc Feuermann, Sebastian 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, Vivien Junker, Thomas Kappler, Guillaume Keller, Corinne Lachaize, Lydie Lane-Guermonprez, Petra Langendijk-Genevaux, Vicente Lara, Philippe Lemercier, Virginie Le Saux, Damien Lieberherr, Tania de Oliveira Lima, Veronique Mangold, Xavier Martin, Patrick Masson, Karine Michoud, Madelaine Moinat, Anne Morgat, Anais Mottaz, Salvo Paesano, Ivo Pedruzzi, Isabelle Phan, Sandrine Pilbout, Violaine Pillet, Sylvain Poux, Monica Pozzato, Nicole Redaschi, Sorogini Reynaud, Catherine Rivoire, Bernd Roechert, Michel Schneider, Christian Sigrist, Karin Sonesson, Sylvie Staehli, 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 Bioinformatics and Structural Biology Department of the University of Geneva; Rolf Apweiler, Yasmin Alam-Faruque, Ricardo Antunes, Daniel Barrell, David Binns, Lawrence Bower, Paul Browne, Wei Mun Chan, Emily Dimmer, Ruth Eberhardt, Alexander Fedotov, Rebecca Foulger, John Garavelli, Renato Golin, Alan Horne, Rachael Huntley, Julius Jacobsen, Michael Kleen, Paul Kersey, Kati Laiho, Rasko Leinonen, Duncan Legge, Quan Lin, Michele Magrane, Maria Jesus Martin, Claire O'Donovan, Sandra Orchard, John O'Rourke, Samuel Patient, Manuela Pruess, Andrey Sitnov, Eleanor Stanley, Matt Corbett, Giuseppe di Martino, Mike Donnelly, Jie Luo and Pieter van Rensburg at the European Bioinformatics Institute (EBI); Cathy Wu, Cecilia Arighi, Leslie Arminski, Winona Barker, Yongxing Chen, Zhang-Zhi Hu, Hsing-Kuo Hua, Hongzhan Huang, Raja Mazumder, Peter McGarvey, Darren A. Natale, Anastasia Nikolskaya, Natalia Petrova, Baris E. Suzek, Sona Vasudevan, C. R. Vinayaka, Lai Su Yeh and Jian Zhang at the Protein Information Resource (PIR).