Nucleic Acids Research, 2006, Vol. 34, Database issue ***D187–D191***
doi:10.1093/nar/gkj161

The Universal Protein Resource (UniProt): an expanding universe of protein information

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

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

Received September 17, 2005; Revised and Accepted October 31, 2005

ABSTRACT

The Universal Protein Resource (UniProt) provides a central resource on protein sequences and functional annotation with three database components, each addressing a key need in protein bioinformatics. The UniProt Knowledgebase (UniProtKB), comprising the manually annotated UniProtKB/Swiss-Prot section and the automatically annotated UniProtKB/TrEMBL section, is the preeminent storehouse of protein annotation. The extensive cross-references, functional and feature annotations and literature-based evidence attribution enable scientists to analyse proteins and query across databases. The UniProt Reference Clusters (UniRef) speed similarity searches via sequence space compression by merging sequences that are 100% (UniRef100), 90% (UniRef90) or 50% (UniRef50) identical. Finally, the UniProt Archive (UniParc) stores all publicly available protein sequences, containing the history of sequence data with links to the source databases. UniProt databases continue to grow in size and in availability of information. Recent and upcoming changes to database contents, formats, controlled vocabularies and services are described. New download availability includes all major releases of UniProtKB, sequence collections by taxonomic division and complete proteomes. A bibliography mapping service has been added, and an ID mapping service will be available soon. UniProt databases can be accessed online at http://www.uniprot.org or downloaded at ftp://ftp.uniprot.org/pub/databases/.

INTRODUCTION

The amount of information available about proteins continues to increase at a rapid pace. Protein interactions, expression profiles and structures are being discovered on a large scale, while completely sequenced genomes cover the taxonomic tree with both breadth and depth. Biological and biochemical functions of individual proteins continue to be elucidated. Furthermore, improved analytical tools are available to make intelligent predictions about function, localization, secondary structure and other important protein properties.

The ability to store and interconnect this expanding universe of protein information is crucial to modern biological research. Accordingly, the Universal Protein Resource (UniProt) plays an ever more important role by providing a central resource on protein sequences and functional annotation for biologists and for scientists active in functional proteomics and genomics research. The broad, long-term objective of UniProt can be summarized as the creation and maintenance of stable, comprehensive and high-quality protein databases, coupled with efficient and unencumbered access mechanisms, to enable rich protein information retrieval and scientific querying across multiple databases containing complementary information. The core activities in UniProt include sequence archiving, manual curation of protein sequences assisted by automated annotation, development of a user-friendly UniProt website and interaction with other protein-related databases for expanded cross-references. The resource builds upon the

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

Table 1. Names and sizes of the UniProt databases

Database name		Database size[a]
Abbreviation	Full name/meaning	
UniProt	Universal Protein Resource	
UniProtKB	UniProt Knowledgebase	2 299 834
UniProtKB/ Swiss-Prot	Swiss-Prot section of the UniProt Knowledgebase	194 317
UniProtKB/ TrEMBL	TrEMBL section of the UniProt Knowledgebase	2 105 517
UniParc	UniProt Archive	5 025 587
UniRef	UniProt Reference Clusters	
UniRef100	UniProt Reference Clusters: 100% identity	2 939 066
UniRef90	UniProt Reference Clusters: 90% identity	1 730 689
UniRef50	UniProt Reference Clusters: 50% identity	907 983

[a]Based on Release 6.0 (September 13, 2005).

solid foundations laid by the three UniProt Consortium members, the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR).

CONTENTS

UniProt comprises three database components, each of which addresses a key need in protein bioinformatics. The UniProt Knowledgebase (UniProtKB) provides protein sequences with extensive annotation and cross-references. The UniProt Archive (UniParc) is the main sequence storehouse. The UniProt Reference Clusters (UniRef) condense sequence information and annotation to facilitate both sequence similarity searches and analyses of the results. Table 1 summarizes the three UniProt databases and their sizes in the current release.

UniProt Knowledgebase

The centerpiece UniProt database is the UniProtKB—a richly annotated protein sequence database with extensive cross-references. Much of the annotation data are buried within the ever-increasing volume of scientific publications or spread among individual databases stored at different locations with differing formats. The UniProtKB provides an integrated and uniform presentation of these disparate data, including annotations such as protein name and function, taxonomy, enzyme-specific information (catalytic activity, cofactors, metabolic pathway, regulatory mechanisms), domains and sites, post-translational modifications, subcellular locations, tissue-specific or developmentally specific expression, interactions, splice isoforms, polymorphisms, diseases and sequence conflicts. Literature citations provide evidence for experimental data. Entries connect to various external data collections such as the underlying DNA sequence entries, protein structure databases, protein domain and family databases, and species- and function-specific data collections. As a result, UniProtKB acts as a central hub connecting biomolecular information archived in ~100 cross-referenced databases.

The UniProtKB contains two sections. UniProtKB/Swiss-Prot contains records with full manual annotation or computer-assisted, manually-verified annotation performed by biologists and based on published literature and sequence analysis. UniProtKB/TrEMBL contains records with computationally generated annotation and large-scale functional characterization. The computer-assisted annotation may employ automatically generated rules as in Spearmint (1), or manually curated rules based on protein families, including HAMAP family rules (2), PIRSF classification-based name rules and site rules (3) and Rulebase rules (4).

UniProt Reference Clusters

The UniRef are three separate datasets that compress sequence space at different resolutions, achieved by merging sequences and sub-sequences that are 100% (UniRef100), ⩾90% (UniRef90) or ⩾50% (UniRef50) identical, regardless of source organism. Reduction of sequence redundancy speeds sequence similarity searches while rendering such searches more informative.

To maximize the chances of biological discovery, homology searches are performed using up-to-date collections of sequences. However, with the accelerated growth of the number of sequences, similarity searching has become increasingly computationally intensive and prohibitive for resource providers and their users. Furthermore, there is an uneven distribution of sequences in sequence space (5). An overabundance of very closely related sequences (e.g. >90% identity) slows down database searches, and long lists of similar or identical alignments can obscure novel matches in the output. A more even sampling of sequences will shorten and clean output listings without repetition of redundant hits. The compression of UniRef100 into UniRef90 and UniRef50 yielded size reductions of ~40 and 65%, respectively.

UniProt Archive

Protein sequences are publicly available from several sources that largely—but not completely—overlap in coverage. The UniParc houses all new and revised protein sequences from these various sources to ensure that comprehensive coverage is available at a single site. A simple collection of sequences from disparate sources can potentially lead to redundancy in the archive, since the same sequence may be found in many sources (UniProt, GenPept, RefSeq, etc.). To avoid redundancy, each unique sequence is assigned a unique identifier and is stored only once. The basic information stored with each UniParc entry is the identifier, the sequence, cyclic redundancy check number (CRC64), source database(s) with accession and version numbers, and a time stamp. In addition, each source database accession number is tagged with its status in that database, indicating if the sequence still exists or has been deleted at that source. The archive thus provides a history of protein sequences.

NEW FEATURES

Availability of major releases, taxonomic divisions and complete proteome sets

Initially, the UniProt FTP site (ftp://ftp.uniprot.org/pub/databases) contained only the latest biweekly release of the complete UniProtKB and UniRef (under /uniprot/current_release/). We now make available every major release (1.0, 2.0, etc.) by FTP (/uniprot/previous_major_releases/). In addition to the possibility of downloading the complete databases, we

provide UniProtKB data divided into taxonomic divisions for archaea, bacteria, vertebrates, mammals, humans, plants, rodents, invertebrates, fungi, viruses and unclassified (/uniprot/current_release/knowledgebase/taxonomic_divisions/). Furthermore, complete proteomes from >200 organisms are also available for download (/uniprot/current_release/knowledgebase/complete_proteomes/).

Bibliography mapping service

The laborious nature of extracting information from the literature often hampers the ability to provide links to reports of experimentally verified data directly within database entries. Moreover, while literature citation is extensive for curated entries, the same is not true for non-curated entries or those not recently curated. We therefore provide annotated bibliography pages (http://www.uniprot.org/bibliography/biblioretrieve.shtml) that list, for each UniProt entry, both curated bibliography and computationally mapped bibliography. The curated bibliography includes, in addition to UniProtKB citations, references from other curated databases [such as SGD (6), MGD (7) and GeneRIF (8)] that are mapped to UniProtKB entries. Brief descriptions of information contained in a citation about protein features and/or functions are included with source attributions whenever available. Also linked is the tagged text evidence describing experimental features (e.g. phosphorylation). To further assist literature mining, we also provide a link to the BioThesaurus of protein and gene names for query expansion using synonyms for PubMed searches.

New documents

A number of documents have been added that collect information previously found only within individual UniProtKB/Swiss-Prot entries. Proteins involved in annotated biochemical pathways are given in pathway.txt (http://www.uniprot.org/support/docs/pathway.html). Sequence variations noted for human proteins are collected in humsavar.txt (http://www.uniprot.org/support/docs/humsavar.html), which presents the sequence position and amino acid substitution, the variation type (e.g. polymorphism or disease mutation) and the associated disease, if any. Finally, proteins exhibiting sequence or structural similarity are presented in similar.txt (http://www.uniprot.org/support/docs/similar.html), listed by family or domain.

RECENT CHANGES

A full account of recent and forthcoming changes can be found in the files http://www.uniprot.org/support/docs/sp_news.html and http://www.uniprot.org/support/docs/sp_soon.html, respectively. Some highlights are presented below.

Database contents

To avoid over-representation of certain sequences in the UniProtKB, immunoglobulins and T-cell receptors were specifically excluded in UniProtKB/TrEMBL and could be found only within the UniParc. This policy has been changed. With the exception of those identified as non-germline, such sequences are now included. In addition, whole genome shotgun entries—previously excluded due to instability—have been included (except for those derived from environmental samples). The UniRef no longer contain sequences from IPI, but instead contain RefSeq sequences. In addition, Ensembl sequences for human, mouse, rat, *Arabidopsis* and zebrafish have been added, as have PDB sequences that are not otherwise represented in the UniProtKB. The major sources of the UniProt databases are depicted in Figure 1.

UniProt Knowledgebase format changes

Meaningful entry IDs are designed to facilitate at-a-glance identification of the protein and the species of origin, and may change as new information becomes available about the protein (this contrasts with entry accessions, which are designed for stability of reference). The Swiss-Prot section

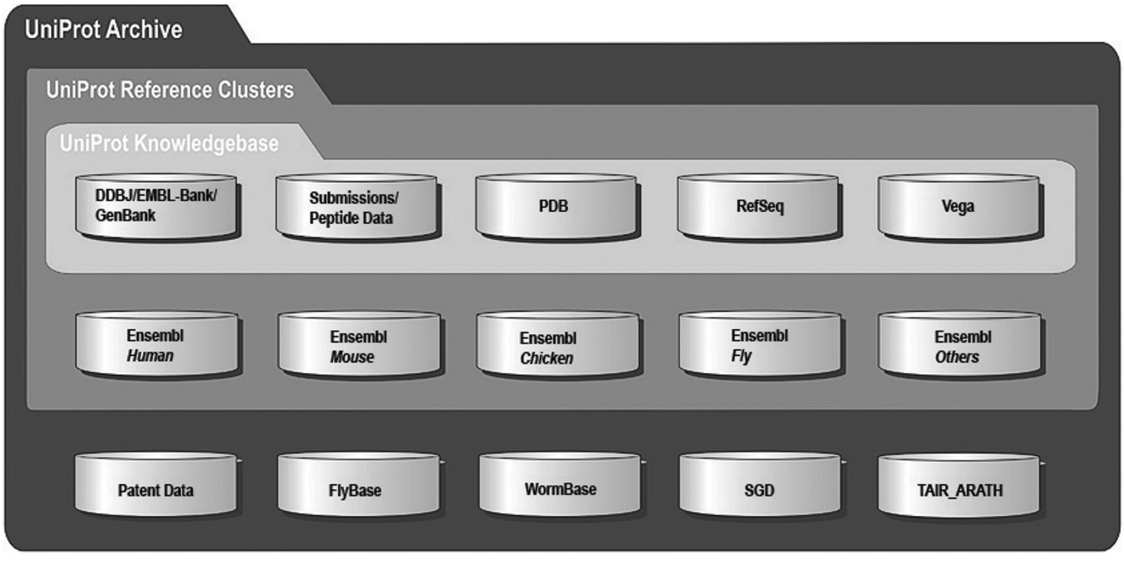

Figure 1. Overview of the major data sources of the UniProt databases.

of the UniProtKB uses a general purpose naming convention for entry IDs that can be symbolized as X_Y, where X is a mnemonic code of alphanumeric characters representing the protein name and Y is a species identification code of at most five alphanumeric characters representing the biological source of the protein. We have elongated the mnemonic code for the protein name to up to five characters. The naming convention for TrEMBL entries has been changed by appending a biological source indicator to the six-character accession number using the UniProtKB/Swiss-Prot-like format. Programmatically, a UniProtKB/Swiss-Prot entry can be distinguished from a UniProtKB/TrEMBL entry by the number of characters preceding the underscore (six for the latter, up to five for the former). The IDtracker tool (http://www.expasy.org/cgi-bin/idtracker) is available to trace protein entry identifiers.

To increase the readability of UniProtKB/Swiss-Prot entries, most fields have been converted to mixed case. This does not apply to the ID or accession (AC) lines, the reference position (RP) line, the sequence header (SQ) line or the sequence itself.

Controlled vocabulary

A number of changes have occurred to keyword, feature, organelle and comment fields. Of special note are the new comment (CC) lines INTERACTION, which is used to convey information relevant to protein–protein interactions, and BIOPHYSICOCHEMICAL PROPERTIES, which conveys information on pH and temperature dependence, kinetic parameters, redox potentials and maximal absorption. The INTERACTION comment is derived automatically from the IntAct database (9) and is updated on a monthly basis. Interactions can be derived by any appropriate experimental method, but must be confirmed by a second experiment if the evidence derives from a single yeast two-hybrid experiment. Many new terms have been added to features such as MODIFIED RESIDUE (MOD_RES; for post-translational modification), CROSSLNK and LIPID. In addition, to describe distinct types of regions in a protein sequence, we redefined two feature keys, DOMAIN and SITE, and introduced five new keys, COILED, COMPBIAS, MOTIF, REGION and TOPO_DOM (Table 2).

Table 2. Addition and redefinition of UniProt feature keys

Feature key	Definition
COILED	A coiled-coil region
COMPBIAS	A compositionally biased region
MOTIF	A short (≤20 amino acids) sequence of biological interest
REGION	A region of interest in the sequence
TOPO_DOM	A topological domain
DOMAIN[a]	A specific combination of secondary structures organized into a characteristic three-dimensional structure or fold
SITE[a]	A single amino acid residue; can also apply to an amino acid bond represented by the positions of the two flanking amino acids

[a]Revised.

Clarification of 'UniProt'

The use of 'UniProt' to refer to the Resource, the Consortium, and the Knowledgebase has created some confusion and consistency issues. Accordingly, we have clarified the use of the following terms: 'UniProt' refers to the 'Universal Protein Resource,' while the UniProt Knowledgebase is abbreviated as 'UniProtKB' and the Consortium is referred to by 'UniProt Consortium.'

UPCOMING DEVELOPMENTS

Annotation archive

UniParc allows tracking of 'historic' sequence data. However, UniParc entries contain no annotation and therefore do not enable tracking of annotation changes. UniProtKB entries are subject to change in both sequence and annotation, but only the most recent versions are currently preserved in the database. The UniProtKB entry version archive will retain earlier versions of entries, thus allowing retrieval of historic views of UniProtKB records.

ID mapping service

ID cross-referencing is fundamental to support data interoperability among disparate data sources and to allow integration and querying of data from heterogeneous molecular biology databases. UniProt will therefore provide a mapping service to convert common gene IDs and protein IDs (such as NCBI's gi number and Entrez Gene ID) to UniProtKB AC/ID and vice versa. A preview of this service is available at http://www.pir.uniprot.org/search/idmapping.shtml, which maps between UniProtKB and ~30 other data sources. Some of the mapping is inherited from cross-references within UniProtKB entries, some are based on the existing bridge between EMBL and GenBank entries, and others make use of cross-references obtained from the iProClass database (10). A subset of the latter (such as between UniProtKB and NCBI gi number) require matching based on sequence and taxonomy identity. Thus, it is possible to map between numerous databases using only a few sources for the mapping itself; these include UniProtKB, iProClass, RefSeq and Genbank nr.

caBIG Grid enablement

The cancer Biomedical Informatics Grid (caBIG), a National Cancer Institute initiative, is designed as an infrastructure that connects resources to enable the sharing of data and tools for cancer research. The UniProtKB is being grid-enabled in the caBIG architecture as a reference project for grid data service to provide query mechanisms for (i) database ID searches based on, for example, UniProtKB ID or RefSeq accession number, (ii) text searches for fields such as protein or gene name or keywords and (iii) boolean searches of two fields. Results are returned in XML and FASTA format for easy data exchange. The caBIG grid enablement will allow UniProtKB data to be interoperated and queried in connection with other cancer biomedical data and services on the grid.

SCIENTIFIC COMMUNITY INTERACTION AND DATABASE ACCESS

One challenge in life sciences research is the ability to integrate and exchange data coming from multiple research

groups. The UniProt Consortium is committed to fostering interaction and exchange with the scientific community, ensuring wide access to UniProt resources and promoting interoperability between resources.

External links

Established through close collaborations with the research community, the UniProtKB provides explicit and implicit links via DR (Database cross-Reference) lines to ~100 molecular databases and resources. Examples include genomic sequence repositories [e.g. GenBank/EMBL/DDBJ (11)], model organism genomes [e.g. TAIR (9)], mutation databases [e.g. dbSNP (8)], protein family and domain [e.g. InterPro (12), Pfam (13)], protein structure [e.g. PDB (14)], ontology [e.g. Gene Ontology (15)], enzyme and function [e.g. MEROPS (16)] and biological processes [e.g. Reactome (17)]. A document listing all databases cross-referenced in UniProt (http://www.uniprot.org/support/docs/dbxref.html) is available and contains, for each database, a short description and the server URL.

UniProt continually adds new database cross-references to UniProtKB records, thereby facilitating broader access to relevant online resources with complementary protein-related information. New cross-references should have: (i) stable accession numbers mapped to the latest UniProtKB accession numbers, (ii) an established update procedure, (iii) public availability of the data and (iv) reciprocal links to UniProt. External resources can easily link to individual protein entries in UniProt using a link URL. For example, UniProtKB entries can be referenced by http://www.uniprot.org/entry/AC (e.g. http://www.uniprot.org/entry/P99999).

Database access

UniProt database entries are available for searching, browsing and retrieval from the UniProt website (http://www.uniprot.org). In addition, the UniProtKB and the UniRef are available for download from the UniProt FTP site (ftp://ftp.uniprot.org/pub/databases/). Correspondence with UniProt Consortium scientists and programmers is facilitated through the Help Desk (http://www.uniprot.org/support/helpdesk.shtml).

ACKNOWLEDGEMENTS

UniProt is mainly supported by the National Institutes of Health (NIH) grant 1 U01 HG02712-01. Additional support for the EBI's involvement in UniProt comes from the two European Union contracts BioBabel (QLRT-2000-00981) and TEMBLOR (QLRI-2001-00015) and from the NIH grant 1R01HGO2273-01. UniProtKB/Swiss-Prot activities at the SIB are supported by the Swiss Federal Government through the Federal Office of Education and Science. PIR activities are also supported by the NIH grants for NIAID proteomic resource (HHSN266200400061C) and grid enablement (NCI-caBIG-ICR-10-10-01) and National Science Foundation grants for iProClass (DBI-0138188) and BioThesaurus (ITR-0205470). Funding to pay the Open Access publication charges for this article was provided by EMBL.

Conflict of interest statement. None declared.

REFERENCES

1. Kretschmann,E., Fleischmann,W. and Apweiler,R. (2001) Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. *Bioinformatics*, **17**, 920–926.
2. 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.
3. Wu,C.H., Huang,H., Yeh,L.S. and Barker,W.C. (2003) Protein family classification and functional annotation. *Comput. Biol. Chem.*, **27**, 37–47.
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. Holm,L. and Sander,C. (1998) Dictionary of recurrent domains in protein structures. *Proteins*, **33**, 88–96.
6. Balakrishnan,R., Christie,K.R., Costanzo,M.C., Dolinski,K., Dwight,S.S., Engel,S.R., Fisk,D.G., Hirschman,J.E., Hong,E.L., Nash,R. *et al.* (2005) Fungal BLAST and Model Organism BLASTP Best Hits: new comparison resources at the Saccharomyces Genome Database (SGD). *Nucleic Acids Res.*, **33**, D374–D377.
7. Eppig,J.T., Bult,C.J., Kadin,J.A., Richardson,J.E., Blake,J.A., Anagnostopoulos,A., Baldarelli,R.M., Baya,M., Beal,J.S., Bello,S.M. *et al.* (2005) The Mouse Genome Database (MGD): from genes to mice—a community resource for mouse biology. *Nucleic Acids Res.*, **33**, D471–D475.
8. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S., Helmberg,W. *et al.* (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **33**, D39–D45.
9. Rhee,S.Y., Beavis,W., Berardini,T.Z., Chen,G., Dixon,D., Doyle,A., Garcia-Hernandez,M., Huala,E., Lander,G., Montoya,M. *et al.* (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res.*, **31**, 224–228.
10. Wu,C.H., Huang,H., Nikolskaya,A., Hu,Z. and Barker,W.C. (2004) The iProClass integrated database for protein functional analysis. *Comput. Biol. Chem.*, **28**, 87–96.
11. Kanz,C., Aldebert,P., Althorpe,N., Baker,W., Baldwin,A., Bates,K., Browne,P., van den Broek,A., Castro,M., Cochrane,G. *et al.* (2005) The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.*, **33**, D29–D33.
12. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bradley,P., Bork,P., Bucher,P., Cerutti,L. *et al.* (2005) InterPro, progress and status in 2005. *Nucleic Acids Res.*, **33**, D201–D205.
13. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
14. Deshpande,N., Addess,K.J., Bluhm,W.F., Merino-Ott,J.C., Townsend-Merino,W., Zhang,Q., Knezevich,C., Xie,L., Chen,L., Feng,Z. *et al.* (2005) The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res.*, **33**, D233–D237.
15. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
16. Rawlings,N.D., Tolle,D.P. and Barrett,A.J. (2004) MEROPS: the peptidase database. *Nucleic Acids Res.*, **32**, D160–D164.
17. Joshi-Tope,G., Gillespie,M., Vastrik,I., D'Eustachio,P., Schmidt,E., de Bono,B., Jassal,B., Gopinath,G.R., Wu,G.R., Matthews,L. *et al.* (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.*, **33**, D428–D432.