D154–D159 *Nucleic Acids Research, 2005, Vol. 33, Database issue*
doi:10.1093/nar/gki070

The Universal Protein Resource (UniProt)

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

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

Received September 14, 2004; Revised and Accepted October 5, 2004

ABSTRACT

The Universal Protein Resource (UniProt) provides the scientific community with a single, centralized, authoritative resource for protein sequences and functional information. Formed by uniting the Swiss-Prot, TrEMBL and PIR protein database activities, the UniProt consortium produces three layers of protein sequence databases: the UniProt Archive (UniParc), the UniProt Knowledgebase (UniProt) and the UniProt Reference (UniRef) databases. The UniProt Knowledgebase is a comprehensive, fully classified, richly and accurately annotated protein sequence knowledgebase with extensive cross-references. This centrepiece consists of two sections: UniProt/Swiss-Prot, with fully, manually curated entries; and UniProt/TrEMBL, enriched with automated classification and annotation. During 2004, tens of thousands of Knowledgebase records got manually annotated or updated; we introduced a new comment line topic: TOXIC DOSE to store information on the acute toxicity of a toxin; the UniProt keyword list got augmented by additional keywords; we improved the documentation of the keywords and are continuously overhauling and standardizing the annotation of post-translational modifications. Furthermore, we introduced a new documentation file of the strains and their synonyms. Many new database cross-references were introduced and we started to make use of Digital Object Identifiers. We also achieved in collaboration with the Macromolecular Structure Database group at EBI an improved integration with structural databases by residue level mapping of sequences from the Protein Data Bank entries onto corresponding UniProt entries. For convenient sequence searches we provide the UniRef non-redundant sequence databases. The comprehensive UniParc database stores the complete body of publicly available protein sequence data. 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 two weeks.

INTRODUCTION

Previously, Swiss-Prot + TrEMBL (1) and PIR-PSD (2) coexisted as protein databases with differing sequence coverage and annotation priorities. In 2002, the Swiss-Prot + TrEMBL groups at the SIB (Swiss Institute of Bioinformatics) and EBI (European Bioinformatics Institute) and the PIR (Protein Information Resource) group at Georgetown University Medical Center and National Biomedical Research Foundation joined forces as the UniProt consortium (3).

The UniProt consortium maintains three database layers:

(i) The UniProt Archive (UniParc) provides a stable, comprehensive, non-redundant sequence collection by storing the complete body of publicly available protein sequence data.
(ii) The UniProt Knowledgebase (UniProt) provides the central database of protein sequences with accurate, consistent and rich sequence and functional annotation.
(iii) The UniProt Reference (UniRef) databases provide non-redundant data collections based on the UniProt Knowledgebase and UniParc in order to obtain complete coverage of sequence space at several resolutions.

*To whom correspondence should be addressed: Tel: +44 0 1223 494435; Fax: +44 0 1223 494468; Email: apweiler@ebi.ac.uk

THE UniProt ARCHIVE

Although most protein sequence data are derived from the translation of DDBJ/EMBL/GenBank (4) sequences, primary protein sequence data are also submitted directly to UniProt or appear in patent applications or in entries from the Protein Data Bank (PDB) (5). The UniParc (6) is designed to capture all available protein sequence data—not just from the aforementioned databases, but also from sources such as Ensembl (7), the International Protein Index (IPI) (8), RefSeq (9), FlyBase (10) and WormBase (11). This combination of sources makes UniParc the most comprehensive publicly accessible, non-redundant protein sequence database available.

UniParc represents each protein sequence once and only once, assigning it a unique UniParc identifier. The UniParc release 2.6 from September 2004 contained 4 375 775 unique sequences from 11 978 094 original source records. UniParc cross-references the accession numbers of the source databases, using flags to indicate the status of the entry in the original source database, with 'active' indicating that the entry is still present in the source database and 'obsolete' indicating that the entry no longer exists in the source database. A UniParc sequence version is incremented each time the underlying sequence changes, making it possible to observe sequence changes in all source databases. A sample UniParc report can be found at http://www.uniprot.org/entry/UPI0000000C37. UniParc records carry no annotation, but this information can be found in the UniProt Knowledgebase or other underlying databases.

THE UniProt KNOWLEDGEBASE

The UniProt Knowledgebase merges Swiss-Prot, TrEMBL and PIR-PSD to provide a central database of protein sequences with annotations and functional information. All suitable PIR-PSD sequences missing from Swiss-Prot + TrEMBL were incorporated into UniProt and bi-directional cross-references were created to allow the easy tracking of PIR-PSD entries. The transfer into UniProt of references and experimentally verified data present in PIR but missing from Swiss-Prot + TrEMBL is ongoing.

The UniProt Knowledgebase has two parts: a section of fully, manually annotated records resulting from literature information extraction and curator-evaluated computational analysis, and a section with computationally analysed records awaiting full manual annotation. The two sections are referred to as 'UniProt/Swiss-Prot' (158 337 records in UniProt release 2.6 from September 2004) and 'UniProt/TrEMBL' (1 400 776 records in UniProt release 2.6 from September 2004), respectively. An example UniProt report can be found at http://www.uniprot.org/entry/P57727.

In the following paragraphs, we will explain the main principles of the UniProt Knowledgebase and enhancements introduced recently.

High-quality annotation

In addition to capturing the core data mandatory to each UniProt entry (consisting principally of the amino acid sequence, the protein name or description, taxonomic data and citation information), we attach other annotation information both manually and automatically.

Manual annotation is performed by biologists and is based on literature curation and sequence analysis. The annotation principles were described in detail previously (3,12). During 2004, tens of thousands of records were manually annotated or updated. We also have introduced a new comment (CC) line topic: TOXIC DOSE. This topic is used to store information on the poisoning potential (acute toxicity) of a toxin. Generally this topic holds information on the LD_{50} and PD_{50}. LD stands for 'Lethal Dose'. LD_{50} is the amount of a toxin, given all at once, which causes the death of 50% (one-half) of a group of test animals. PD_{50} stands for 'Paralytic dose'. It is the amount of a toxin, which causes the paralysis of 50% of a group of test animals.

Examples:

CC -!- TOXIC DOSE: PD_{50} is 1.72 mg/kg by injection in blowfly larvae.
CC -!- TOXIC DOSE: LD_{50} is 0.015 mg/kg by intravenous injection for sarafotoxin-A and sarafotoxin-B, and 0.3 mg/kg for sarafotoxin-C.

Automatic classification and annotation. Much progress was made during 2004 in our attempt to provide automatic large-scale functional characterization and annotation, which is generated with limited human interaction.

InterPro classification. We use InterPro (13) to recognize domains and to classify all the protein sequences in UniProt into families and superfamilies. InterPro is an integrated resource of protein families, domains and sites that amalgamates the efforts of the member databases: Pfam (14), PROSITE (15), PRINTS (16), ProDom (17), SMART (18), PIRSF (19), Superfamily (20) and TIGRFAMs (21). Approximately 80% of all UniProt Knowledgebase records are classified according to their InterPro domains and familes.

Automatic functional annotation of UniProt/TrEMBL. For automatic annotation, systems for standardized transfer of annotation from well-characterized proteins in the UniProt/Swiss-Prot to non-annotated UniProt/TrEMBL entries have been implemented. RuleBase (22) uses a semi-automatic approach, while the Spearmint approach is completely automated and is based on decision trees (23). InterPro is then used to assign UniProt entries into groups. The annotation shared by the functionally characterized UniProt/Swiss-Prot proteins of a group is then extracted and assigned to the non-annotated UniProt/TrEMBL entries of this group. These systems have been used to improve the annotation in 32% (RuleBase) and 55% (Spearmint) of UniProt/TrEMBL entries.

However, a part of the automatically added data will be erroneous, as are parts of the information coming from other sources. Therefore, we introduced a post-processing system called Xanthippe, which is based on a simple exclusion mechanism and a decision tree approach using the C4.5 data-mining algorithm. Xanthippe detects and flags a large part of the annotation errors and considerably increases the reliability of both automatically generated data and pre-existing annotation inherited from the underlying nucleotide sequence source data (24).

Table 1. PIR site rules for automated annotation of functional sites

Rule ID	Template ID	Rule condition	Feature for propagation	Reference
PIRSR000259-1	UniProt:P17846 PDB:1AOP	PIRSF000259 member Site match: Arg82, Arg152, Lys214, Lys216, Cys433, Cys439, Cys478, Cys482	ACT_SITE (catalytic): Arg82, Arg152, Lys214, Lys216, Cys482	CSA:1AOP PMID:9315848, 9315849
PIRSR000259-2	UniProt:P17846 PDB:1AOP	PIRSF000259 member Site match: Arg82, Arg152, Lys214, Lys216, Cys433, Cys439, Cys478, Cys482	BINDING (4Fe-4S cluster): Cys433, Cys439, Cys478, Cys482	PMID:7569952
PIRSR000259-3	UniProt:P17846 PDB:1AOP	PIRSF000259 member Site match: Arg82, Arg152, Lys214, Lys216, Cys433, Cys439, Cys478, Cys482	BINDING (Siroheme iron): Cys 482	PMID:7569952

The PIRSF classification serves as the basis for a rule-based approach to automatically provide standardized and rich functional annotation for position-specific sequence features, protein names, Enzyme Commission (EC) name and number, keywords and Gene Ontology (GO) terms (25). Position-specific site rules are developed for annotating active site residues, binding site residues, modified residues or other functionally important amino acid residues. To exploit known structure information, site rules are defined starting with PIRSF families that contain at least one known three-dimensional (3D) structure with experimentally verified site information. The rules are defined using appropriate syntax and controlled vocabulary for site description and evidence attribution. As shown in Table 1, each rule consists of the rule ID, template sequence (a representative sequence with known 3D structure), rule condition, feature for propagation (denoting site feature to be propagated) and reference. The rules are family-specific and there may be more than one site rule per family. Site rule curation involves manually editing a multiple sequence alignment of representative family members (including the template PDB entry), visualizing site residues in the 3D structure, and building hidden Markov models for the conserved regions containing the functional site residues (referred to as 'site HMMs'). The HMM thus built allows one to map functionally important residues from the template structure to other members of the PIRSF family that do not have a solved structure.

For site feature propagation, the entire rule condition is examined by PIRSF membership checking, site HMM matching and site residue matching. To avoid false positives, site features are only propagated automatically if all site residues match perfectly in the conserved region by aligning both the template and query sequences to the profile HMM using HmmAlign. Potential functional sites missing one or more residues or containing conservative substitutions are only annotated after expert review with evidence attribution. For accurate site propagation, it is sometimes necessary to match more residues in the rule condition than those to be propagated. For example, a total of eight catalytic and binding residues in sulfite reductase need to be matched in order to correctly propagate the sirohaem-ion binding Cys residue (PIRSR000259-3, Table 1).

The highly reliable automatic annotation has already been incorporated into the UniProt/TrEMBL flat files, while additional automatic annotation is available from the extended UniProt view at http://www.ebi.uniprot.org/.

The HAMAP project, or 'High-quality Automated and Manual Annotation of microbial Proteomes', aims to integrate manual and automatic annotation methods in order to enhance the speed of the curation process while preserving the quality of the database annotation (26). Automatic annotation is only applied to entries that belong to manually defined orthologous families and to entries with no identifiable similarities (ORFans). Many checks are enforced in order to prevent the propagation of wrong annotation and to spot problematic cases, which are channelled to manual curation. The results of this annotation are integrated in UniProt/Swiss-Prot.

Standardized nomenclature and controlled vocabularies. Whenever available, we make use of the official nomenclature defined by international committees while still providing the published synonyms. For various other UniProt items we use controlled vocabularies, e.g. for tissues, plasmids and keywords, which are listed in UniProt documents. The UniProt keyword list was augmented by additional keywords. We improved the documentation of the keywords by adding, to the list of keywords, the definition of their usage in the UniProt knowledgebase and additional information such as synonyms or relevant GO terms. The UniProt curators also contribute to the work of the GOA project (27) by assigning GO terms from each of the GOs, i.e. the function of a protein, what processes it is involved in and where in the cell it is located. A major effort was started to continuously overhaul and standardize the annotation of post-translational modifications (PTMs). Furthermore, we introduced a new documentation file of the strains and their synonyms together with the mnemonic species identification code representing the biological source of the protein in the knowledgebase. These and other documents can be found at http://www.uniprot.org/support/documents.shtml.

Integration with other databases

UniProt provides cross-references to external data collections such as the underlying DNA sequence entries in the DDBJ/EMBL/GenBank nucleotide sequence databases, two dimensional (2D) PAGE and 3D protein structure databases, various protein domain and family characterization databases, PTM databases, species-specific data collections, variant databases and disease databases. Many new cross-references were included over the last year. Accordingly, UniProt acts as a central hub for biomolecular information with now more than four million cross-references to more than 60 databases. A document listing all databases cross-referenced in UniProt

(http://www.uniprot.org/support/docs/dbxref.shtml) is available and contains, for each database, a short description and the server URL.

UniProt achieved in 2004 in collaboration with the Macromolecular Structure Database (MSD) group at EBI an improved integration with structural databases by residue level mapping of sequences from the PDB entries onto corresponding UniProt entries (28). This work led to an overhaul of the format of the UniProt cross-references to PDB to reflect the mappings. The UniProt–PDB mappings are available at ftp://ftp.ebi.ac.uk/pub/databases/msd/sifts/.

We also started to make use of Digital Object Identifiers (DOIs). The DOI system is used for identifying and exchanging intellectual property in the digital environment. We introduced the new optional identifier 'DOI' in the RX line to store the DOI of a cited document.

Minimal redundancy

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

The UniProt Knowledgebase is therefore by design non-redundant, with the goal of representing all known information regarding a particular protein. The definition of non-redundancy here is different from that employed in UniParc: in UniParc, all sequences that are 100% identical over their entire length are merged into a single entry, regardless of species; the UniProt Knowledgebase aims to describe in a single record all protein products derived from a certain gene (or genes if the translation from different genes in a genome leads to indistinguishable proteins) from a certain species and to give not only the whole record an accession number but to assign to each protein form derived by alternative splicing, proteolytic cleavage and post-translational modification Isoform identifiers, which are accession numbers for the isoforms. The underlying reason for giving each of these isoforms a unique identifier is that each of these may have a different function or biological role or may only exist during specific developmental stages or under certain environmental conditions, even when all these isoforms are derived from a single gene. Isoform identifiers have been so far only introduced for splice isoforms. Splice isoforms may differ considerably from one another, with potentially <50% sequence similarity between isoforms. The tool VARSPLIC (29), which is freely available, enables the recreation of all annotated splice variants from the feature table of a UniProt entry, or for the complete database. A FASTA-formatted file containing all splice variants annotated in UniProt can be downloaded for use with similarity search programs.

Evidence attribution

The UniProt consortium emphasizes the use of an evidence attribution mechanism for protein annotation that will include, for all data, the data source, the types of evidence and methods for annotation. This is essential as the UniProt Knowledgebase will contain data automatically imported from the underlying nucleotide sequence databases, data imported from other databases, data from specific programs, the results of automatic annotation systems and, most importantly, expert manual curation. The implementation of evidence tags will allow the user to distinguish between these data sources and to easily identify particular classes of data of interest such as experimentally proven protein annotation. Evidence tags for the annotation present in UniProt/TrEMBL records are already available in the UniProt XML distribution.

THE UniProt REFERENCE DATABASES

Automatic procedures have been developed to create three UniRef databases, such as UniRef100, UniRef90 and UniRef50, from the UniProt Knowledgebase and UniParc as representative protein sequence databases with high information content. The databases provide complete coverage of sequence space while hiding redundant sequences from view. The non-redundancy facilitates sequence merging in the UniProt Knowledgebase (based on UniRef100) and allows faster sequence similarity searches (by using UniRef90 and UniRef50).

UniRef100 provides a comprehensive non-redundant sequence collection clustered by sequence identity. UniRef merges sequences automatically across different species and also adds some data from UniParc, such as translations from highly unstable gene predictions; while merging in the Knowledgebase is restricted to curator-assisted inclusion of reliable and stable sequence data for a single species. UniRef100 is based on all UniProt Knowledgebase records, as well as UniParc records that represent sequences deemed over-represented in the Knowledgebase, DDBJ/EMBL/GenBank Whole Genome Shotgun coding sequence translations, Ensembl protein translations from various organisms, as well as IPI data. The production of UniRef100 begins with the clustering of all records by sequence identity. Identical sequences and subfragments are presented as a single UniRef100 entry, containing the accession numbers of all merged entries, and the protein sequence. The UniRef100 release 2.6 from September 2004 contained 2 611 612 records derived from the corresponding UniProt knowledgebase and UniParc releases.

UniRef90 and UniRef50 are built from UniRef100 using the CD-HIT algorithm (30) to provide non-redundant sequence collections for the scientific user community to perform faster homology searches. All records from all source organisms with mutual sequence identity of >90 or >50%, respectively, are merged into a single record that links to the corresponding UniProt Knowledgebase records. UniRef90 and UniRef50 yield a size reduction of ~40 and 65%, respectively. A sample UniRef90 report can be found at http://www.uniprot.org/entry/uniref90_P57727.

PRACTICAL INFORMATION

Interactive access and linking to UniProt

The most efficient and user-friendly way to browse the UniProt databases is via the UniProt website (http://www.uniprot.org),

which serves as a portal to all aspects of the UniProt project, and contains detailed documentation about the background and scope of UniProt. It provides database query and data-mining mechanisms, user support and communication, file download capabilities, and links to consortium resources. The UniProt Help Desk (help@uniprot.org) provides access to UniProt curators and database maintainers.

The standard way of linking to UniProt, displaying the UniProt 'basic' view as HTML, is: http://www.uniprot.org/entry/entryname or accession number.

Examples:

http://www.uniprot.org/entry/cyc_human
http://www.uniprot.org/entry/P99999
http://www.uniprot.org/entry/UniRef100_P99999
http://www.uniprot.org/entry/UniRef90_P99999
http://www.uniprot.org/entry/UniRef50_P99999
http://www.uniprot.org/entry/UPI00000002E4

UniProt data availability and submission

UniProt, UniParc and UniRef entries, with supporting documentation, can be retrieved in various formats (Swiss-Prot/TrEMBL flat file, FASTA, XML) via anonymous FTP from ftp://ftp.uniprot.org/pub/. New UniProt, UniParc and UniRef releases are produced every two weeks.

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

CONCLUSIONS

Complete and up-to-date databases of biological knowledge are vital for information-dependent biological and biotechnological research. With the rapid accumulation of genome sequences for many organisms, attention is turning to the identification and functions of proteins encoded by these genomes. With the increasing volume and variety of protein sequences and functional information, UniProt serves as a central resource of protein sequence and function, providing a cornerstone for scientists active in modern biological research. The resource provides rich, consistent and non-redundant protein information by combining reliable automated annotation approaches with literature-based expert manual curation.

ACKNOWLEDGEMENTS

UniProt is mainly supported by the National Institutes of Health (NIH) grant 1 U01 HG02712-01. Minor support for the EBIs 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. UniProt/Swiss-Prot activities at the SIB are supported by the Swiss Federal Government through the Federal Office of Education and Science. PIR activities are also supported by the National Science Foundation (NSF) grants DBI-0138188 and ITR-0205470.

REFERENCES

1. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al*. (2003) The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*., **31**, 365–370.
2. Wu,C.H., Yeh,L.-S.L., Huang,H., Arminski,L., Castro-Alvear,J., Chen,Y., Hu,Z., Kourtesis,P., Ledley,R.S., Suzek,B.E. *et al*. (2003) The Protein Information Resource. *Nucleic Acids Res*., **31**, 345–347.
3. Apweiler,R., Bairoch,A., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al*. (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res*., **32**, D115–D119.
4. Kulikova,T., Aldebert,P., Althorpe,N., Baker,W., Bates,K., Browne,P., van den Broek,A., Cochrane,G., Duggan,K., Eberhardt,R. *et al*. (2004) The EMBL Nucleotide Sequence Database. *Nucleic Acids Res*., **32**, D27–D30.
5. Westbrook,J., Feng,Z., Chen,L., Yang,H. and Berman,H. (2003) The Protein Data Bank and structural genomics. *Nucleic Acids Res*., **31**, 489–491.
6. Leinonen,R., Diez,F.G., Binns,D., Fleischmann,W., Lopez,R. and Apweiler,R. (2004) UniProt Archive. *Bioinformatics*, **20**, 3236–3237.
7. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T. *et al*. (2002) The Ensembl genome database project. *Nucleic Acids Res*., **30**, 38–41.
8. Kersey,P.J., Duarte,J., Williams,A., Karavidopoulou,Y., Birney,E. and Apweiler,R. (2004) The International Protein Index: an integrated database for proteomics experiments. *Proteomics*, **4**, 1985–1988.
9. Pruitt,K. and Maglott,D. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res*., **29**, 137–140.
10. FlyBase Consortium (2003) The FlyBase database of the *Drosophila* genome projects and community literature. *Nucleic Acids Res*., **31**, 172–175.
11. Harris,T., Lee,R., Schwarz,E., Bradnam,K., Lawson,D., Chen,W., Blasier,D., Kenny,E., Cunningham,F., Kishore,R. *et al*. (2003) WormBase: a cross-species database for comparative genomics. *Nucleic Acids Res*., **31**, 133–137.
12. Apweiler,R., Bairoch,A. and Wu,C.H. (2004) Protein sequence databases. *Curr. Opin. Chem. Biol*., **8**, 76–80.
13. Mulder,N., Apweiler,R., Attwood,T., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al*. (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res*., **31**, 315–318.
14. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L.L. (2002) The Pfam Protein Families Database. *Nucleic Acids Res*., **30**, 276–280.
15. Hulo,N., Sigrist,C.J.A., Le Saux,V., Langendijk-Genevaux,P.S., Bordoli,L., Gattiker,A., De Castro,E., Bucher,P. and Bairoch,A. (2004) Recent improvements to the PROSITE database. *Nucleic Acids Res*., **32**, D134–D137.
16. Attwood,T.K., Bradley,P., Flower,D.R., Gaulton,A., Maudling,N., Mitchell,A.L., Moulton,G., Nordle,A., Paine,K., Taylor,P. *et al*. (2003) PRINTS and its automatic supplement, preprints. *Nucleic Acids Res*., **31**, 400–402.
17. Servant,F., Bru,C., Carrere,S., Courcelle,E., Couzy,J., Peyruc,D. and Kahn,D. (2002) Prodom: automated clustering of homologous domains. *Brief. Bioinformatics*, **3**, 246–251.
18. Letunic,I., Goodstadt,L., Dickens,N.J., Doerks,T., Schultz,J., Mott,R., Ciccarelli,F., Copley,R.R., Ponting,C.P. and Bork,P. (2002) Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res*., **30**, 242–244.
19. Wu,C.H., Nikolskaya,A., Huang,H., Yeh,L.-S., Natale,D., Vinayaka,C.R., Hu,Z., Mazumder,R., Kumar,S., Kourtesis,P. *et al*. (2004) PIRSF family classification system at the Protein Information Resource. *Nucleic Acids Res*., **32**, D112–D114.
20. Gough,J., Karplus,K., Hughey,R. and Chothia,C. (2001) Assignment of homology to genome sequences using a library of hidden markov models that represent all proteins of known structure. *J. Mol. Biol*., **313**, 903–919.
21. Haft,D.H., Loftus,B.J., Richardson,D.L., Yang,F., Eisen,J.A., Paulsen,I.T. and White,O. (2001) TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res*., **29**, 41–43.

22. Fleischmann,W., Moeller,S., Gateau,A. and Apweiler,R. (1999) A novel method for automatic and reliable functional annotation. *Bioinformatics*, **15**, 228–233.
23. 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.
24. Wieser,D., Kretschmann,E. and Apweiler,R. (2004) Filtering erroneous protein annotation. *Bioinformatics*, **20**, i342–i347.
25. 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.
26. Gattiker,A., Michoud,K., Rivoire,C., Auchincloss,A.H., Coudert,E., Lima,T., Kersey,P., Pagni,M., Sigrist,C.J.A., Lachaize,C. *et al*. (2003) Automatic annotation of microbial proteomes in Swiss-Prot. *Comput. Biol. Chem.*, **27**, 49–58.
27. Camon,E., Magrane,M., Barrell,D., Lee,V., Dimmer,E., Maslen,J., Binns,D., Harte,N., Lopez,R. and Apweiler R. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.*, **32**, D262–D265.
28. Velankar,S., McNeil,P., Mittard-Runte,V., Suarez,A., Barrell,D., Apweiler,R. and Henrick,K. (2005) E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res.*, **33**, 262–265.
29. Kersey,P., Hermjakob,H. and Apweiler,R. (2000) VARSPLIC: alternatively-spliced protein sequences derived from Swiss-Prot and TrEMBL. *Bioinformatics*, **11**, 1048–1049.
30. Li,W., Jaroszewski,L. and Godzik,A. (2002) Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics*, **18**, 77–82.