D214–D219 *Nucleic Acids Research, 2011, Vol. 39, Database issue* *Published online 4 November 2010*
doi:10.1093/nar/gkq1020

Ongoing and future developments at the Universal Protein Resource

The UniProt Consortium[1,2,3,4,*†]

[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, [3]Protein Information Resource, Georgetown University Medical Center, 3300 Whitehaven St. NW, Suite 1200, Washington, DC 20007 and [4]University of Delaware, 15 Innovation Way, Suite 205, Newark, DE 19711, USA

Received October 3, 2010; Accepted October 10, 2010

ABSTRACT

The primary mission of Universal Protein Resource (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 4 weeks and can be accessed online for searches or download at http://www.uniprot.org.

INTRODUCTION

Universal Protein Resource (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 numerous sources and is the most comprehensive catalog of protein sequences 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) (1) is a comprehensive sequence repository, reflecting the history of all protein sequences. UniProt Reference Clusters (UniRef) (2) merge closely related sequences based on sequence identity to facilitate sequence similarity searches while the UniProt Metagenomic and Environmental Sequences Database (UniMES) was created to cater for the expanding area of metagenomic data. UniProt is freely and easily accessible by researchers and provides interactive customized search facilities for proteins of interest to facilitate hypothesis generation and knowledge discovery.

ONGOING AND NEW DEVELOPMENTS

Annotation changes

Demerging of multi-gene entries derived from a single species in UniProtKB/Swiss-Prot. UniProtKB/Swiss-Prot has historically 'merged' 100% identical protein sequences from different genes in the same species into one single record. The aim of this approach was to reduce sequence redundancy within the proteome of individual species, facilitating protein identification and the functional annotation of protein sequences. As the availability and usage of genomic information has greatly increased in recent years, UniProtKB is modifying its merging policy. We have begun to 'demerge' entries containing multiple individual genes coding for 100% identical protein sequences into individual UniProtKB/Swiss-Prot entries containing a single gene. This will give a gene-centric view of protein space, where the same protein sequence can be represented multiple times by distinct UniProtKB/Swiss-Prot entries. It will allow a cleaner and more logical mapping of gene and genomic resources to UniProtKB, which provide the major point of entry to the resulting proteome for many users. One consequence of this change in annotation policy is that the level of protein sequence redundancy in UniProtKB will increase, as multiple identical instances

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

of a given protein sequence may appear. The increase in redundancy should however be minor. In *Homo sapiens*, the number of entries currently selected for demerge represents <1% of the total (~20 300 human entries) and although this number will certainly increase over time it will still represent only a very small fraction of the total.

Entries annotated via the HAMAP pipeline are integrated into UniProtKB/TrEMBL. The growth in the number of completely sequenced microbial genomes generated a need for a procedure that provides high-quality annotation for as many protein sequences as possible. UniProt's response was to develop the semi-automated system, HAMAP (high-quality automated and manual annotation of microbial proteomes), that uses manually built annotation templates for protein families to propagate annotation to all members of defined protein families, using very strict criteria. In UniProtKB, entries semi-automatically annotated via the HAMAP pipeline used to be integrated into UniProtKB/Swiss-Prot. To allow users to unambiguously discriminate between manually and automatically annotated records, our policy has changed and HAMAP is being integrated into the automatic annotation system applied to UniProtKB/TrEMBL. This approach has several advantages. First, it reinstates UniProtKB/Swiss-Prot policy of only containing manually reviewed entries. Second, it enhances UniProtKB/TrEMBL's quality by adding new, manually reviewed rules for automatic annotation. Third, the UniProtKB/TrEMBL's automatic annotation system allows quick and extensive updates at every release. Fourth, HAMAP-derived annotation displayed in these entries is tagged to indicate the source rule. It should be emphasized that entries used as templates to create HAMAP family rules undergo thorough manual annotation, not only at the sequence level, but also at the functional level, including in-depth literature mining. Consequently, these entries are in UniProtKB/Swiss-Prot and manually updated when required. This helps guarantee the high quality of the resulting HAMAP profiles.

Evidence attribution

Each UniProtKB entry combines information from a wide range of sources including data imported from ENA/DDBJ/GenBank (3,4,5) nucleotide records, data imported from other databases, automatic annotation and manually curated information added from the scientific literature and sequence analysis programs. The UniProt Consortium utilizes an evidence attribution system which attaches an evidence tag to each data item in a UniProtKB entry identifying its source(s) and/or the method(s) used to generate it. This system allows users to easily trace the source of each annotation, and thereby to assess data reliability. It also allows for automated correction and updating of imported or predicted annotations in response to changes in their underlying sources, while preserving data which has been manually curated.

Finer-grained tagging will be implemented to allow a more detailed breakdown of source information that will enable users to more easily use the evidence tags to identify all experimentally characterized annotations for a given protein from a particular strain of a particular species as well as all experimentally characterized annotations/proteins from a proteome or protein family. This has the added benefit of providing the experimental annotation for the benchmarking of sequence analysis tools such as predictors of post-translational modifications or topology. Retrofitting will be ongoing for those manually curated entries created before the evidence attribution system was introduced. The UniProt Consortium is currently modifying UniProt evidence tags to make them compatible with the widely used gene ontology (6) evidence codes, and is also involved in discussions with other curated databases such as ENA/DDBJ/GenBank and RefSeq (7) to ensure consistency across our resources. The evidence attribution system is available in the XML version of UniProtKB and has also been partially implemented in the entry view on the UniProt web site allowing the user to link through to the source resource or appropriate SOP. Future plans include the provision of additional web services based on the evidence tags such as access to the source annotation rules (UniRules/SAAS) responsible for each predicted annotation. It is planned to extend the scope of the evidence attribution system to enable users to use the evidence tags to identify all experimentally characterized annotations for a given protein from a particular strain of a particular species as well as all experimentally characterized annotations/proteins from a proteome or protein family. This has the added benefit of providing the experimental annotation for the benchmarking of sequence analysis tools such as predictors of post-translational modifications or topology.

Complete proteomes

The International Protein Index (IPI) (8), an integrated database providing non-redundant complete datasets of proteins for featured higher eukaryotic organisms, has been extensively used in proteomics experiments. IPI was launched in 2001 when information about proteins was stored in diverse formats across many different databases. The situation has improved for many of the best-studied genomes and UniProtKB is now working in collaboration with Ensembl (9) and RefSeq to provide complete protein sequence coverage of IPI organisms. The UniProt Knowledgebase provides comprehensive coverage of the human, mouse and rat genomes. It has also incorporated the *Arabidopsis thaliana* data from TAIR (10), and sequence predictions for cow, chick, zebrafish and dog. UniProt will produce specific data sets for each of these species and, shortly after this, production of IPI will be discontinued. A further development in the provision of complete proteome sets by UniProt concerns those species which are represented by more than one complete proteome in UniProtKB. At the current time there are relatively few such species—some notable exceptions include *Escherichia coli* and *Streptococcus agalactiae*. However, their number is likely to grow sharply in the near future, as continuing developments in high-throughput sequencing technologies

render the genome sequencing of multiple strains or isolates of a particular species routine. Given the potentially large differences in protein composition that may exist between different strains of the same species, it is essential that users be able to identify the actual strain for which a particular protein or group of proteins has been experimentally characterized. This makes the provision and annotation of complete proteomes for individual strains or isolates essential. Historically, UniProtKB merged sequence and annotation data from different strains and complete proteomes into a single entry. Such merging will no longer be performed, existing merged proteomes will be actively demerged, and functional annotation assigned to the correct strain where possible. Individual complete proteomes that currently have a species-level taxonomic assignment will be reassigned to the correct strain-level taxonomic node, where that information is available. The final outcome of these changes will be the provision of clearly delineated and correctly annotated proteomes for individual strains or isolates in UniProtKB. In those cases where multiple complete genome and proteome sequences are available for a single strain or isolate, users will be provided with access to the individual proteomes from separate projects. To further help users select more easily among multiple proteomes, UniProt will define a single reference or representative proteome for each species—this will be the proteome that includes the greatest amount of relevant functional annotation, and, where resources permit, will be the proteome which is preferentially selected for manual curation. The selection and definition of reference strain proteomes in UniProtKB has already begun for viruses (11), where 355 reference strains have been selected in accordance with RefSeq. The manual selection of reference proteomes for these and other species will be supplemented by an automatic procedure that assesses annotation content. Access to data on complete proteomes including downloads will be provided via a new portal which is currently under development. This portal will provide users with information and simple statistics for both complete proteomes and their individual components, such as chromosomes and plasmids.

Automatic annotation

The current rate of production of genome sequence data far exceeds the rate at which such data can be manually annotated. UniProt has developed two complementary systems to automatically annotate protein sequences in UniProtKB/TrEMBL, providing accurate annotations to protein sequences that might never be experimentally characterized. The first system, UniRule, which incorporates the HAMAP(12), RuleBase(13) and Protein Information Resource (PIR) (14) systems, uses annotation rules created and monitored by experienced curators. Each annotation rule specifies a number of annotations, and conditions which must be satisfied for that annotation to be applied. These conditions may include family membership [as indicated by a match to a family defined by InterPro (15)], taxonomic constraints and the presence of particular sequence features. Rules are created by curators based on information from experimentally characterized template entries, and their predictions evaluated against the content of manually annotated UniProtKB/Swiss-Prot entries which serve as the gold standard. Those rules that are inconsistent with current UniProtKB/Swiss-Prot annotation are sent to curators for review. This validation step occurs at each UniProtKB release, and ensures that only high-quality predictions are added and prevents propagation of potentially erroneous data.

The second system, the Statistical Automatic Annotation System [SAAS, previously named Spearmint (16)] supplements the labor-intensive-UniRule system and to ensure scalability of computational annotation. SAAS generates automatic rules for functional annotation from UniProtKB/Swiss-Prot entries using the C4.5 decision-tree algorithm. This algorithm uses entropy gain to find the most concise rule for an annotation based on the criteria of sequence length, InterPro-group membership and taxonomy. SAAS employs a data-exclusion set that censors data not suitable for computational annotation (such as specific biophysical or chemical properties) and generates human readable rules for each release. Generating rules 'on the fly' ensures their evolution along with the UniProtKB with little or no manual intervention while providing seed rules for exploitation in the UniRule system. UniRule and SAAS currently predict both protein properties, such as function, catalytic activity, pathways, sub-cellular location and sequence-specific information, such as location of active sites. This combined approach produces annotation for ~38% of UniProtKB/TrEMBL entries at the current time. All predictions are refreshed with each UniProtKB release to ensure the latest state-of-knowledge predictions.

Integration with other databases

Cross-references. UniProtKB acts as a central hub for biomolecular information by connecting to >140 databases that provide additional or complementary information. Establishing and maintaining these cross-references is the result of a collaborative effort with the scientific community and contact with the resource developers is maintained to ensure access to up-to-date, reliable and comprehensive data. UniProKB provides cross-references to a number of resource types including such as the nucleotide sequence databases, protein structure databases, protein domain and family databases and species-specific and function/feature-specific data collections. Users can search for entries having particular cross-references in UniProtKB via the web interface.

ID mapping. ID mapping is essential to support data interoperability among disparate data sources and enables the integration and querying of data from heterogeneous molecular biology databases. UniProt provides 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. ID mapping can be of two types: first, mapping between similar objects such as NCBI RefSeq and UniProtKB accession

numbers and second, from identifiers to annotations such as a UniProtKB accession number to a Pathway or GO ID. UniProt currently provides mappings for >100 unique ID types at http://www.uniprot.org/mapping and the mapping table downloads at ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping.

Additional bibliography. The UniProt additional bibliography service, through close collaboration with providers of public-curated databases such as Entrez Gene and model organism databases (MODs), maps additional bibliographic data for UniProtKB protein entries to help users better explore existing protein knowledge. This service exploits the collaborative connections with the wider biocuration community to provide users access to additional publications absent from the UniProtKB record, and assists UniProtKB curators to identify and prioritize candidate entries for manual curation. Currently, the service provides literature annotations from 15 external gene or protein databases, including nine MODs. These external sources contribute ~475 000 unique PubMed citations not annotated in UniProtKB, covering ~230 000 UniProtKB entries. The additional bibliography is directly linked from the protein entry view on the UniProt web site.

New data sources

Most of the UniProtKB protein sequences are derived from translations in the ENA/DDBJ/GenBank nucleotide sequence databases. UniProtKB also contains sequences from PDB (17), TAIR and Ensembl (as described earlier). UniProt will extend its collaboration with Ensembl and will work closely with RefSeq to integrate relevant data into the UniProt databases and to identify core sets of protein coding regions that are consistently annotated and of high quality. The long-term goal is to support convergence towards a standard set of protein-coding gene annotations, especially for mammalian and model organisms.

The UniProt Consortium will also extend collaborations with the International Nucleotide Sequence Database Collaboration (INSDC) (18), RefSeq and Ensembl to promote the integration of next-generation-sequencing (NGS) data into the UniProt databases. A special emphasis will be placed on the wealth of variation data being generated for key species of biomedical interest, including of course *H. sapiens*. This will benefit users of protein and nucleotide resources through improved interoperability and data synchronization.

In the context of new annotation sources, UniProt has collaborated for some time with the wwPDB to produce the UniProtKB/PDB mappings on a weekly basis which is made available through the SIFTS project (19). This is now used as the basis of the automated annotation of UniProtKB/TrEMBL entries using data from the PDBeMotif database (20). The derived annotations provide positional small molecule–protein interaction data in the feature section of the UniProt records, their associated literature citations, keywords and cross-references. Given the broad range of chemical entities contained in the PDB archive, only a small manually chosen subset deemed to be unambiguously biologically relevant is included.

DATABASE ACCESS AND FEEDBACK

The www.uniprot.org website (21) 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. Searches 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 (www.uniprot.org/help/text-search). 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 web site 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 (www.w3.org/RDF/), a W3C standard for publishing data on the Semantic Web. Programmatic access to data and search results is possible via simple HTTP (REST) requests (www.uniprot.org/faq/28). Java applications can also make use of our Java API (UniProtJAPI) (22). While the UniProt web site provides a query interface for all UniProt data, users frequently require the facility to search across related data in different databases. BioMart is an open source query-oriented data-management system that allows for integrated querying of biological-data resources regardless of their geographical locations. A UniProt Biomart is available which allows complext queries between UniProt and other data resources such as PRIDE, Ensembl and InterPro (http://www.ebi.ac.uk/uniprot/biomart/martview). 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 www.uniprot.org/contact or email us directly at help@uniprot.org. The page www.uniprot.org/help/submissions provides information about data submissions and updates. Extensive documentation on how to best use our resource is available at www.uniprot.org/help/. UniProt is freely available for both commercial and non-commercial use (www.uniprot.org/help/license). New releases are published every 4 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.

ACKNOWLEDGEMENTS

Universal Protein Resource 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, Francesco

Fazzini, Alexander Fedotov, Rebecca Foulger, John Garavelli, Leyla Garcia Castro, Rachael Huntley, Julius Jacobsen, Michael Kleen, Kati Laiho, Duncan Legge, Quan Lin, Wudong Liu, Jie Luo, Sandra Orchard, Samuel Patient, Klemens Pichler, Diego Poggioli, Nikolas Pontikos, Manuela Pruess, Steven Rosanoff, Tony Sawford, Harminder Sehra, Edward Turner, Matt Corbett, Mike Donnelly and Pieter van Rensburg at the European Bioinformatics Institute (EBI); Ioannis Xenarios, Lydie Bougueleret, Andrea Auchincloss, Ghislaine Argoud-Puy, Kristian Axelsen, Amos Bairoch, Delphine Baratin, Marie-Claude Blatter, Brigitte Boeckmann, Jerven Bolleman, Laurent Bollondi, Emmanuel Boutet, Silvia Braconi Quintaje, Lionel Breuza, Alan Bridge, Edouard deCastro, Elisabeth Coudert, Isabelle Cusin, Mikael Doche, Dolnide Dornevil, Severine Duvaud, Anne Estreicher, Livia Famiglietti, Marc Feuermann, Sebastien Gehant, 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, Vicente Lara, Philippe Lemercier, Damien Lieberherr, Xavier Martin, Patrick Masson, Madelaine Moinat, Anne Morgat, Salvo Paesano, Ivo Pedruzzi, Sandrine Pilbout, 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 at the Swiss Institute of Bioinformatics (SIB); Cathy H. Wu, Cecilia N. Arighi, Leslie Arminski, Winona C. Barker, Chuming Chen, Yongxing Chen, Pratibha Dubey, Hongzhan Huang, Raja Mazumder, Peter McGarvey, Darren A. Natale, Thanemozhi G. Natarajan, Jules Nchoutmboube, Natalia V. Roberts, Baris E. Suzek, Uzoamaka Ugochukwu, C. R. Vinayaka, Qinghua Wang, Yuqi Wang, Lai-Su Yeh and Jian Zhang at the Protein Information Resource (PIR).

FUNDING

National Institutes of Health [grants 2U01HG02712-04, 2P41HG02273-07 and 5R01GM080646-04, 3R01GM 080646-04S2, 1G08LM010720-01, 3P20RR016472-09S2 to UniProt and European Bioinformatics Institute (EBI)'s involvement in UniProt and Protein Information Resource activities, respectively]; European Commission contract SLING (grant 226073 to EBI's involvement in UniProt); Swiss Federal Government through the Federal Office of Education and Science and from the European Commission contracts GEN2PHEN (grant 200754 to UniProtKB/Swiss-Prot activities at the Swiss Institute of Bioinformatics); MICROME (grant 222886-2 to UniProtKB/Swiss-Prot activities at the Swiss Institute of Bioinformatics); SLING (grant 226073 to UniProtKB/Swiss-Prot activities at the Swiss Institute of Bioinformatics); NSF (grant DBI-0850319 to Protein Information Resource activities). Funding for open access charge: NIH (grant 2U01HG02712-04).

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. 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.
3. Leinonen,R., Akhtar,R., Birney,E., Bonfield,J., Bower,L., Corbett,M., Cheng,Y., Demiralp,F., Faruque,N., Goodgame,N. *et al*. (2010) Improvements to services at the European nucleotide archive. *Nucleic Acids Res*., **38**, D39–D45.
4. Kaminuma,E., Mashima,J., Kodama,Y., Gojobori,T., Ogasawara,O., Okuba,Y., Takagi,T. and Nakamuru,Y. (2010) DDBJ launches a new archive database with analytical tools for next-generation sequence data. *Nucleic Acids. Res*., **38**, D33–D38.
5. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2010) GenBank. *Nucleic Acids Res*., **38**, D46–D51.
6. The Gene Ontology Consortium. (2000) Gene ontology: tool for the unification of biology. *Nat. Genet*., **25**, 25–29.
7. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*., **35**, D61–D65.
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. Hubbard,T.J.P., Aken,B.L., Ayling,S., Ballester,B., Beal,K., Bragin,E., Brent,S., Chen,Y., Clapham,P., Clarke,L. *et al*. (2009) Ensembl 2009. *Nucleic Acids Res*., **37**, D690–D697.
10. Swarbreck,D., Wilks,C., Lamesch,P., Berardini,T., 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.
11. Hulo,C., de Castro,E., Masson,P., Bougueleret,L., Bairoch,A., Xenarios,I. and Le Mercier,P. (2010) ViralZone: a knowledge resource to understand virus diversity. *Nucleic Acids Res*., doi:10.1093/nar/gkq901.
12. 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.
13. Fleischmann,W., Moller,S., Gateau,A. and Apweiler,R. (1999) A novel method for automatic functional annotation of proteins. *Bioinformatics*, **15**, 228–233.
14. 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, NY.
15. Hunter,S., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Das,U., Daugherty,L., Duquenne,L. *et al*. (2009) InterPro: the integrative protein signature database (2009). *Nucleic Acids Res*., **37**, D224–D228.
16. 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.
17. Velankar,S., Best,C., Beuth,B., Boutselakis,C.H., Cobley,N., Sousa Da Silva,A.W., Dimitropoulos,D., Golovin,A., Hirshberg,M., John,M. *et al*. (2010) PDBe: Protein Data Bank in Europe. *Nucleic Acids Res*., **38**, D308–D317.

18. Brunak,S., Danchin,A., Hattori,M., Nakamura,H., Shinozaki,K., Matise,T. and Preuss,D. (2002) Nucleotide sequence database policies. *Science*, **298**, 1333.
19. Velankar,S., McNeil,P., Mittard-Runtel,V., Suarez,A., Barrell,D., Apweiler,R. and Henrick,K. (2005) E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res.*, **33**, D262–D265.
20. Golovin,A. and Henrick,D. (2008) MSDmotif: exploring protein sites and motifs. *BMC Bioinformatics*, **9**, 312.
21. 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.
22. 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.