Nucleic Acids Research, 2003, Vol. 31, No. 1 365–370
DOI: 10.1093/nar/gkg095

The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003

Brigitte Boeckmann*, Amos Bairoch, Rolf Apweiler[1], Marie-Claude Blatter, Anne Estreicher, Elisabeth Gasteiger, Maria J. Martin[1], Karine Michoud, Claire O'Donovan[1], Isabelle Phan, Sandrine Pilbout and Michel Schneider

Swiss Institute of Bioinformatics, Centre Medical Universitaire, 1 rue Michel Servet, 1211 Geneva 4, Switzerland and [1]The EMBL Outstation—The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received September 16, 2002; Revised and Accepted October 23, 2002

ABSTRACT

The SWISS-PROT protein knowledgebase (http://www.expasy.org/sprot/ and http://www.ebi.ac.uk/swissprot/) connects amino acid sequences with the current knowledge in the Life Sciences. Each protein entry provides an interdisciplinary overview of relevant information by bringing together experimental results, computed features and sometimes even contradictory conclusions. Detailed expertise that goes beyond the scope of SWISS-PROT is made available via direct links to specialised databases. SWISS-PROT provides annotated entries for all species, but concentrates on the annotation of entries from human (the HPI project) and other model organisms to ensure the presence of high quality annotation for representative members of all protein families. Part of the annotation can be transferred to other family members, as is already done for microbes by the High-quality Automated and Manual Annotation of microbial Proteomes (HAMAP) project. Protein families and groups of proteins are regularly reviewed to keep up with current scientific findings. Complementarily, TrEMBL strives to comprise all protein sequences that are not yet represented in SWISS-PROT, by incorporating a perpetually increasing level of mostly automated annotation. Researchers are welcome to contribute their knowledge to the scientific community by submitting relevant findings to SWISS-PROT at swiss-prot@expasy.org.

INTRODUCTION

SWISS-PROT (1) is a protein sequence and knowledge database that is valued for its high quality annotation, the usage of standardized nomenclature, direct links to specialized databases and minimal redundancy. The format of SWISS-PROT follows as closely as possible that of the EMBL Nucleotide Sequence Database (2) for standardization purposes. A description of the distinct line types and their format is available at http://www.expasy.org/sprot/userman.html. A sample SWISS-PROT entry can be viewed at http://www.expasy.org/cgi-bin/niceprot.pl?P49810.

Core data and annotation

The core data, which is mandatory to each SWISS-PROT entry, consists principally of the amino acid sequence, the protein name (description), taxonomic data and citation information. If further information on the protein is available, the entries contain detailed annotation on items such as the function(s) of the protein, enzyme-specific information (catalytic activity, cofactors, metabolic pathway, regulation mechanisms), biologically relevant domains and sites, posttranslational modification(s), molecular weight determined by mass spectrometry, subcellular location(s) of the protein, tissue-specific expression, developmentally-specific expression of the protein, secondary structure, quaternary structure, splice isoform(s), polymorphism(s), similarities to other proteins, use of the protein in a biotechnological process, diseases associated with deficiencies in the protein, use of the protein as a pharmaceutical drug, sequence conflicts, etc.

To acquire a maximum of up-to-date knowledge regarding a protein, information is not only obtained from publications reporting new sequence data, but also from review articles with an aim to revise periodically the annotations of families or groups of proteins. Furthermore, we have enlisted external experts, who have been recruited to send us their comments and updates concerning specific groups of proteins (see http://www.expasy.org/cgi-bin/experts).

Standardized nomenclature and controlled vocabularies

Consistent nomenclature is indispensable for communication and literature search. Right from the start, SWISS-PROT aimed to standardize the nomenclature for a given protein and its isoforms across related organisms. For various SWISS-PROT items, we use controlled vocabularies, e.g. for tissues, plasmids

*To whom correspondence should be addressed. Email: brigitte.boeckmann@isb-sib.ch

and keywords, which are listed in documents distributed with SWISS-PROT (see http://www.expasy.org/sprot/sp-docu.html). Whenever available, we make use of the official nomenclature defined by international committees while still providing the published synonyms.

Integration with other databases

SWISS-PROT provides cross-references to external data collections such as the underlying DNA sequence entries in the DDBJ/EMBL/GenBank nucleotide sequence databases, 2D and 3D protein structure databases, various protein domain and family characterization databases, posttranslational modification (PTM) databases, species-specific data collections, variant databases and disease databases.

Minimal redundancy

Many sequence databases contain, for a given protein sequence, separate entries which correspond to different literature reports. In SWISS-PROT, we try as much as possible to merge all these data in order to minimise 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 SWISS-PROT entry. Splice isoforms may differ considerably from one another, with potentially less than 50% sequence similarity between isoforms. The tool VARSPLIC (3), which is freely available (ftp://ftp.expasy.org/databases/sp_tr_nrdb/varsplic.txt), enables the recreation of all annotated splice variants from the feature table of a SWISS-PROT entry, or for the complete database. A fasta-formatted file containing all splice variants annotated in SWISS-PROT and TrEMBL (1) can be downloaded for use with similarity search programs. Most sequence analysis and proteomic tools on ExPASy (http://www.expasy.org/tools/), e.g. BLAST or PeptIdent, have been adapted to take into account, in addition to all SWISS-PROT and TrEMBL entries, all annotated splice isoforms.

TrEMBL: A computer-annotated supplement to SWISS-PROT

Due to the increased data flow from genome projects to the sequence databases, the SWISS-PROT protein knowledgebase faced a number of challenges in its time- and labour-intensive way of manual database annotation. While it is necessary to maintain the high annotation quality as described above, it is also vital to make sequences available as quickly as possible. To address this, we introduced TrEMBL (translation of EMBL nucleotide sequence database) in 1996. TrEMBL consists of computer-annotated entries derived from the translation of all coding sequences (CDS) in the nucleotide sequence databases, except for CDS already included in SWISS-PROT. It also contains protein sequences extracted from the literature and protein sequences submitted directly by the user community.

It is subdivided into two sections: SP-TrEMBL contains sequences, which will eventually be incorporated into SWISS-PROT and REM-TrEMBL contains those, which will not. These include immunoglobulins and T-cell receptors, synthetic sequences, patent application sequences, fragments of less than 8 amino acids and coding sequences where there is strong experimental evidence that the sequence does not code for a real protein. In addition, there is a weekly update to TrEMBL called TrEMBLnew. TrEMBLnew is produced weekly from new nucleotide sequences deposited in the EMBL nucleotide sequence database. At each TrEMBL release, the TrEMBLnew entries are processed; any entries redundant against SWISS-PROT/TrEMBL (4) are merged and the remainder then progressed into TrEMBL (5).

PROGRESS REPORT ON PROJECTS

SWISS-PROT is regularly enhanced in its content and format to adequately mirror new findings. The distinct line types are continuously overhauled, their content adapted to the current knowledge and the structure standardized to facilitate easy retrieval of related data. Details on recent changes and forthcoming developments are available in the release notes at http://www.expasy.org/sprot/relnotes/. Various projects with considerable progress are described in more detail below.

The Human Proteomics Initiative (HPI)

The goal of the HPI project (6) is to annotate all known human protein sequences and their mammalian orthologs. HPI places a special emphasis on actors playing a role in generating high levels of protein diversity.

The majority of proteins are the target of PTMs. Close to two hundred different PTMs are currently known: e.g. various proteolytic cleavages, the additions of simple chemical groups or complex molecules, disulfide bond formation. We add and update the annotation of PTMs according to experimental evidence.

Up to 60% of human genes have alternatively spliced isoforms (7,8), with an estimated average of 2.75 alternative splice isoforms per gene (7). All known alternatively spliced isoform sequences for a given protein are described in the feature table of the SWISS-PROT entry, and each publication concerning the described sequence is referenced. The existence of various alternative splicing isoforms is validated by comparison with genomic and EST sequences and by careful analysis of the gene structure.

Single amino acid polymorphism (SAP) variants, primarily those linked to disease states, are continuously integrated into SWISS-PROT. Currently, more than 18% of the human entries describe at least one SAP. The total number of SAPs is 13 697. Whenever it is possible, we provide a direct link to dbSNP (http://www.ncbi.nlm.nih.gov/SNP/) at the National Center for Biotechnology Information (NCBI).

SWISS-PROT is gradually being enhanced by the addition of a number of features that are specifically intended for researchers working on human genetic diseases, such as links to human gene databases [OMIM (9), GeneCards (10), GeneLynx (11), Genew (12)] as well as to many gene-specific mutation databases. Medically relevant keywords are created continuously and information relevant to the use of specific proteins as therapeutic agents is stored. Brand names, as well as names of companies developing and selling the drug are also indicated.

A special effort is being made to annotate proteins encoded on chromosomes 20, 21 and 22, which were the first chromosomes to be fully sequenced and partially annotated (13–15). Currently, SWISS-PROT is nearly synchronised with the current state of knowledge of proteins encoded on these chromosomes.

SWISS-PROT contains 8398 annotated human sequences. These sequences are associated with 26 897 literature references, 21 563 experimental or predicted PTMs. 2463 splice variants are described in 1371 entries. Up-to-date statistics are available at http://www.expasy. org/sprot/hpi/hpi_stat.html. For all aspects of the HPI project, we appreciate the help and collaboration of the scientific community. Information concerning the human proteome is highly critical to a large section of the life science community. We, therefore, appeal to the user community to fully participate in this initiative by providing all the necessary information to help and accelerate the comprehensive annotation of the human proteome (hpi@isb-sib.ch).

The International Protein Index (IPI)

IPI (http://www.ebi.ac.uk/IPI) provides a top-level guide to the main databases that describe the human proteome, namely SWISS-PROT, TrEMBL, RefSeq and Ensembl. IPI maintains a database of cross-references between the primary data sources in order to provide a set of human proteins with minimal redundancy yet maximal completeness (one sequence per transcript). Stable identifiers (with incremental versioning) are maintained within IPI facilitating the tracking of sequences between IPI releases. IPI is produced automatically through mapping on the basis of protein similarity between the different data sets. Each IPI entry consists of a cluster of related entries from the constituent databases, together with a sequence and a description line taken from a master entry. These data are presented in FASTA format and in SWISS-PROT format. The latter contains additional cross-references linking IPI to GO (16), Genew, LocusLink (17) and InterPro (18), and identifies the chromosome on which the gene encoding each IPI entry is found.

High-quality Automated and Manual Annotation of microbial Proteomes (HAMAP)

More than 80 archaeal and bacterial genomes have been sequenced and many more are under way. In order to be able to handle this huge number of microbial proteins, we have set up the HAMAP project (http://www.expasy.org/sprot/hamap/), which aims to automatically annotate a significant percentage of proteins from complete bacterial and archaeal proteomes while maintaining the same level of quality that we obtain through manual annotation. The targets of automated annotation are proteins with no similarity to other proteins (ORFans) and proteins that are members of protein (sub)families.

Various prediction tools are applied to proteins that show no similarity to known protein families. Possible transmembrane regions, signal sequence, coiled coils, ATP/GTP binding-sites, LPXTG motifs and some defined repeats are automatically annotated using rules of consistency and dependency, and without any further manual verification.

Proteins belonging to well-characterized protein (sub)families can be annotated automatically using a rule system that describes the extent and nature of annotations that can be assigned by similarity with a prototype manually-annotated entry. Such a rule system also includes a carefully edited multiple alignment of the (sub)family, which is used both to propagate feature annotation from a model entry and to generate identification profiles. Species-specific rules and rules specific to the biochemical pathways are used to develop a system able to spot inconsistencies at the level of the entire proteome.

Currently, we have developed 650 (sub)family rules, each with at least one multiple sequence alignment and corresponding profile for identification of further family members. These profiles are scanned nightly against the TrEMBL entries from complete proteomes. Based on their profile score, entries with high-confidence matches are selected for automated annotation according to the family rule. These entries are further checked to make sure that no errors are introduced into SWISS-PROT. More than 36 000 proteins from complete proteomes have been annotated (manually or semi-automatically) and integrated into SWISS-PROT so far.

Since many proteins are common to both prokaryotes and plastids (chloroplast and cyanelle) genomes, we have included complete plastid proteomes in the frame of HAMAP. For the moment, 24 such genomes have been completely sequenced and they consist of about 2500 proteins. The family rules can take into account special conditions required for the proper annotation of these proteins in their specific context.

More than 40% of the SWISS-PROT entries from microbial and plastids proteomes belong to a HAMAP family. Five complete proteomes have been fully annotated in SWISS-PROT: *Escherichia coli*, *Buchnera aphidicola subsp*. *Acyrtosiphon pisum*, *Haemophilus influenzae*, *Mycoplasma genitalium* and *Mycoplasma pneumoniae*. *Bacillus subtilis* and *Methanococcus jannaschii* will be completed shortly.

Model organisms

In addition to our efforts in the priority annotation of human proteins (see HPI project) and microbes (see HAMAP), 9 eukaryotic species that are the target of genome sequencing and/or mapping projects are considered as model organisms: *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Candida albicans*, *Danio rerio*, *Dictyostelium discoideum*, *Drosophila melanogaster*, *Mus musculus*, *Saccharomyces cerevisiae* and *Schizosaccharaomyces pombe*. Data exchange established with species-specific databases ensures that new and corrected data are incorporated and provided to the public as quickly as possible. Cross-references to these databases are regularly updated.

Collectively, the entries from all model organisms (including human) represent about a third of all SWISS-PROT (August 2002) entries. SWISS-PROT builds species-specific indices for most model organisms. Statistical and comparative analysis information for proteomes of fully sequenced genomes can be obtained at http://www.ebi.ac.uk/proteome/, a site that also provides access to non-redundant data sets for these organisms.

Plant Proteome Annotation Project (PPAP)

We have initiated the PPAP. Emphasis is currently given to the annotation of plant-specific protein families from *A. thaliana*. The main problem for plant sequences encountered to date is the unreliability of the gene prediction programs. About one third of the predicted genes need correction when compared with newly released full-length cDNAs. We will broaden our scope to other species when additional plant genomes become available. SWISS-PROT contains currently 8831 sequence entries from plants, of which 1675 are from *A. thaliana.* More information on PPAP can be found at http://www.expasy.org/sprot/ppap.

The New Taxonomy database (NEWT)

The NEWT database (http://www.ebi.ac.uk/newt/) serves as a taxonomic portal to SWISS-PROT and TrEMBL. For each species, NEWT displays the SWISS-PROT scientific name, SWISS-PROT common name and SWISS-PROT synonym(s), lineage, number of protein sequence entries in SWISS-PROT and TrEMBL as well as links to each entry. The taxonomic classification used in SWISS-PROT is that maintained at the NCBI (see http://www.ncbi.nlm.nih.gov/Taxonomy/). Species with protein sequences stored in the SWISS-PROT protein database are named according to SWISS-PROT nomenclature: following SWISS-PROT conventions, a systematic approach for naming viral and bacterial strains has been adopted and we endeavor to include both the teleomorph name and the anamorph name for fungi. NEWT is updated daily and stores the taxonomy tree structure, which enables users to navigate from one taxon to another and to access the lineage for each taxon. Currently (August 2002), we provide more than 16 000 links to relevant external pictures and sites presenting various scientific data.

Cross-references and unique feature identifiers

Since its inception, SWISS-PROT has placed a major emphasis on integration with other databases and thus became a central hub for biomolecular information archived in currently 66 databases (http://www. expasy. org/cgi-bin/lists?dbxref.txt) (20). For many years, this interconnectivity was achieved almost exclusively via SWISS-PROT DR (Database Cross-Reference) lines, i.e. *explicit* links [to 44 databases]. There are an average of 7.5 cross-references for each sequence entry. More recently, *implicit* links [to currently 22 databases] have been introduced (20), i.e. virtual DR lines created on the fly on the ExPASy server.

We have recently increased the level of depth and complexity of database cross-references, by allowing links also to and from subsequences or particular sites, rather than only to complete entries. This concept can be applied to databases specialised in certain types of PTMs, or in mutations. We have, therefore, introduced unique and stable feature identifiers (FTId), which allows referral to a position-specific annotation item in the feature table. Currently, these are systematically attributed to FT VARIANT lines of human sequence entries, and to certain glycosylation sites, but will ultimately be assigned to all types of FT lines. The FTIds of human variants are used to refer to a sequence variation in a unique and stable manner, and serve as anchors for specifically directed links to dbSNP (9). The same principle is used to further enhance the links to GlycoSuiteDB (21), an annotated database of glycan structures. An example for the new type of cross-references can be found in the SWISS-PROT entry P02765.

SWISS-PROT documentation files

SWISS-PROT is distributed with a large number of documentation files. Some of these files have been available for a long time (the user manual, release notes, the various indices for authors, citations, keywords, etc.), but many species-specific documents have been created recently and we are continuously adding new files. See http://www.expasy.org/sprot/sp-docu.html for a list of all the documents that are currently available.

Automatic functional annotation in TrEMBL

For automatic annotation, a novel system of standardised transfer of annotation from well-characterized proteins in SWISS-PROT to non-annotated TrEMBL entries has been developed (22). RuleBase (5) manages and stores more than 500 annotation rules, which are applied to defined protein groups in TrEMBL. To assign TrEMBL entries into protein groups, the highly diagnostic protein family signature database InterPro (18) is used. This is an integrated resource of protein families, domains and sites which amalgamates the efforts of the member databases which are currently PROSITE (23), PRINTS (24), Pfam (25), ProDom (26), SMART (27) and TIGRFAMs (28). This system has been used to improve the annotation in 25% of all TrEMBL entries. A new data mining approach to the automatic annotation is being developed to complement this approach, which should lead to an increased coverage by automatic annotation over the next year.

Evidence attribution in TrEMBL

Evidence attribution is of growing importance since large biomolecular databases usually combine data from a broad variety of sources. TrEMBL, in particular, contains data automatically imported from the underlying EMBL/DDBJ/GenBank coding sequences, partial manual curation, data imported from other databases, data from specific programs and the results of automatic annotation systems. Although every effort is made to ensure correct and consistent data, the data quality is often limited by the quality of the input data. Currently, it is often difficult for database users to recognise where individual data items come from. To address these issues, we started, in June 2000, to add evidence tags to the internal version of TrEMBL. Evidence tags will allow users to trace the source of each data item added by a curator and to readily distinguish between experimental and predicted data. An evidence-tagged version of the TrEMBL database will soon be available in XML format. Please see ftp://ftp.ebi.ac.uk/pub/databases/trembl/evidenceDocumentation.html for more information. We would welcome any feedback from the user community.

PRACTICAL INFORMATION

SWISS-PROT contains 113 470 sequence entries. Up-to-date statistics are available at http://www.expasy.org/sprot/relnotes/relstat.html. The data file (sequences and annotations) requires 377 Mb of disk storage space. The documentation and index files require 121 Mb of disk space. TrEMBL contains 755 169 sequence entries (SP-TrEMBL: 685 601; REM-TrEMBL: 79 568), TrEMBLnew contains 93 546 entries. Up-to-date statistics are available at http://www.ebi.ac.uk/ swissprot/sptr_stats/. The SP-TrEMBL and REM-TrEMBL data files require ~1.2 Gb and 82 Mb of disk storage space, respectively.

Interactive access to SWISS-PROT and TrEMBL

The most efficient and user-friendly way to browse interactively in SWISS-PROT or TrEMBL is to use the ExPASy web server at http://www.expasy.org/ (see http://www.expasy.org/doc/expasy.pdf), one of its complete and up-to-date mirror sites in Australia, Canada, China, Korea, Taiwan and the USA, or the EBI server (http://www.ebi.ac.uk/).

On both the ExPASy and the EBI Web servers, you can use the Sequence Retrieval System (SRS) (29) software package to query and retrieve sequence entries. The EBI and ExPASy also offer a range of search services (see http://www.ebi.ac.uk/Tools/ or http://www.expasy.org/tools/) to run Smith–Waterman, FASTA and BLAST sequence similarity searches or proteomic identification tools against SWISS-PROT and TrEMBL.

How to obtain the SWISS-PROT and TrEMBL databases

SWISS-PROT and TrEMBL can be obtained by anonymous FTP from the ExPASy server ftp.expasy.org and EBI server ftp.ebi.ac.uk/pub/. Further information how to obtain weekly updates and complete data sets in various formats is available at http://www.expasy.org/sprot/download.html.

Format issues

Currently, SWISS-PROT and TrEMBL are maintained and distributed as flat files. An inherent problem of flat file databanks is that their maintenance becomes increasingly difficult when they grow large in size and many people are involved in the production of the data. To overcome these shortcomings, a Relational Database Management System has been developed and we are in the process of porting the production of SWISS-PROT and TrEMBL to this new system, as well as developing a new file format based on the Extensible Markup Language (XML): the SWISS-PROT Markup Language (SP-ML) see http://www.ebi.ac.uk/swissprot/SP-ML for documentation and samples. In order to develop a good representation of the data using either XML or a relational schema, we are designing a conceptual data model that describes the structure and constraints present in the data, using the Unified Modeling Language (UML) notation.

Submission of new data and updates

To submit updates and/or corrections to SWISS-PROT, you can either use the email address: swiss-prot@expasy.org or the WWW address http://www.expasy.org/sprot/sp_update_form.html. To submit new sequence data to SWISS-PROT and for all enquiries regarding the submission process contact: SWISS-PROT, The EMBL Outstation—The European Bioinformatics Institute Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. Tel: +44 1223494444; Fax: +44 1223494468; Email: datasubs@ebi.ac.uk (for submission); datalib@ebi.ac.uk (for enquiries).

No fees for academic users

The use of SWISS-PROT is free of charge for academic users. However, in September 1998 we implemented a system of annual subscription fees for commercial users of the database. The Swiss Institute of Bioinformatics (SIB) and the EMBL/EBI mandated the company Geneva Bioinformatics (GeneBio) (see http://www.genebio.com) to act as their representative for the purpose of concluding the necessary license agreements and levying the fees. The funds raised are used at the SIB and the EBI to bring and keep SWISS-PROT up-to-date and to further enhance its quality.

CONCLUSIONS

Over the past years SWISS-PROT could not only keep up with the high quality of annotation, but has continuously enhanced its format and content to adjust to the exploding knowledge in proteomics. With exemplary model organisms in view, we accomplished high annotation standards that were transferred to all the database entries. Automated annotation procedures are used for SWISS-PROT in a very conservative manner and are only applied where they allow the achievement of the same level of quality as obtained by manual annotation. The extensive integration of SWISS-PROT with specialized databases enables users to navigate through the current knowledge in the Life Sciences providing an insight into the universe of proteins.

ACKNOWLEDGEMENTS

We wish to thank Andrea Auchincloss, Livia Famiglietti and Michele Magrane for helpful discussions, and Vivienne Baillie Gerritsen for the correction of the manuscript. All statistical information given in this article is retrieved from SWISS-PROT release 40.27 (August 2002) and TrEMBL release 21.10 (September 2002), respectively.

REFERENCES

1. O'Donovan,C., Martin,M.J., Gattiker,A., Gasteiger,E., Bairoch,A. and Apweiler,R. (2002) High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Brief. Bioinform.*, **3**, 275–284.
2. Stoesser,G., Baker,W., van den Broek,A., Camon,E., Garcia-Pastor,M., Kanz,C., Kulikova,T., Leinonen,R., Lin,Q., Lombard,V. *et al.* (2002) The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.*, **30**, 21–26.
3. Kersey,P., Hermjakob,H. and Apweiler,R. (2000) VARSPLIC: alternatively-spliced protein sequences derived from SWISS-PROT and TrEMBL. *Bioinformatics*, **11**, 1048–1049.
4. O'Donovan,C., Martin,M.J., Glemet,E., Codani,J.-J. and Apweiler,R. (1999) Removing redundancy in SWISS-PROT and TrEMBL. *Bioinformatics*, **15**, 258–259.

5. Apweiler,R. (2001) Functional information in SWISS-PROT: the basis for large-scale characterisation of protein sequences. *Brief. Bioinform.*, **2**, 9–18.
6. O'Donovan,C., Apweiler,R. and Bairoch,A. (2001) The human proteomics initiative (HPI). *Trends Biotechnol.*, **19**, 178–181.
7. Brett,D., Pospisil,H., Valcarcel,J., Reich,J. and Bork,P. (2002) Alternative splicing and genome complexity. *Nature Genet.*, **30**, 29–30.
8. Modrek,B. and Lee,C. (2002) A genomic view of alternative splicing. *Nature Genet.*, **1**, 13–19.
9. Wheeler,D.L., Church,D.M., Lash,A.E., Leipe,D.D., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Tatusova,T.A., Wagner,L. and Rapp,B.A. (2002) Database resources of the National Center for Biotechnology Information: 2002 update. *Nucleic Acids Res.*, **30**, 13–16.
10. Rebhan,M., Chalifa-Caspi,V., Prilusky,J. and Lancet,D. (1998) GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics*, **14**, 656–664.
11. Lenhard,B., Hayes,W.S. and Wasserman,W.W. (2001) GeneLynx: a gene-centric portal to the human genome. *Genome Res.*, **11**, 2151–2157.
12. Wain,H.M., Lush,M., Ducluzeau,F. and Povey,S. (2002) Genew: the human gene nomenclature database. *Nucleic Acids Res.*, **30**, 169–171.
13. Deloukas,P., Matthews,L.H., Ashurst,J., Burton,J., Gilbert,J.G., Jones,M., Stavrides,G., Almeida,J.P., Babbage,A.K., Bagguley,C.L. *et al.* (2001) The DNA sequence and comparative analysis of human chromosome 20. *Nature*, **414**, 865–871.
14. Hattori,M., Fujiyama,A., Taylor,T.D., Watanabe,H., Yada,T., Park,H.S., Toyoda,A., Ishii,K., Totoki,Y., Choi,D.K. *et al.* (2000) The DNA sequence of human chromosome 21. *Nature*, **405**, 311–319.
15. Dunham,I., Shimizu,N., Roe,B.A., Chissoe,S., Hunt,A.R., Collins,J.E., Bruskiewich,R., Beare,D.M., Clamp,M., Smink,L.J. *et al.* (1999) The DNA sequence of human chromosome 22. *Nature*, **402**, 489–495.
16. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–29.
17. Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.
18. Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Birney,E., Biswas,M., Bucher,P., Cerutti,L., Corpet,F., Croning,M.D. *et al.* (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **29**, 37–40.
19. Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.-F., Dougherty,B.A., Merrick,J.M. *et al.* (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, **269**, 496–512.
20. Gasteiger,E., Jung,E. and Bairoch,A. (2001) SWISS-PROT: connecting biomolecular knowledge via a protein database. *Curr. Issues Mol. Biol.*, **3**, 47–55.
21. Cooper,C.A., Harrison,M.J., Wilkins,M.R. and Packer,N.H. (2001) GlycoSuiteDB: a new curated relational database of glycoprotein glycan structures and their biological sources. *Nucleic Acids Res.*, **29**, 332–335.
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. Falquet,L., Pagni,M., Bucher,P., Hulo,N., Sigrist,C.J.A., Hofmann,K. and Bairoch,A. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res.*, **30**, 235–238.
24. Attwood,T.K., Blythe,M.J., Flower,D.R., Gaulton,A., Mabey,J.E., Maudling,N., McGregor,L., Mitchell,A.L., Moulton,G., Paine,K. and Scordis,P. (2002) PRINTS and PRINTS-S shed light on protein ancestry. *Nucleic Acids Res.*, **30**, 239–241.
25. 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.
26. Corpet,F., Servant,F., Gouzy,J. and Kahn,D. (2000) ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res.*, **28**, 267–269.
27. 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.
28. 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.
29. Etzold,T. and Argos,P. (1993) SRS—an indexing and retrieval tool for flat file data libraries. *Comput. Appl. Biosci.*, **9**, 49–57.