High-quality protein knowledge resource: SWISS-PROT and TrEMBL

O'Donovan, Claire; Martin, Maria Jesus; Gattiker, Alexandre; Gasteiger, Elisabeth; Bairoch, Amos Marc; Apweiler, Rolf

High-quality protein knowledge resource: SWISS-PROT and TrEMBL

Claire O'Donovan, Maria Jesus Martin, Alexandre Gattiker, Elisabeth Gasteiger, Amos Bairoch and Rolf Apweiler

Date received (in revised form): 25th June 2002

Claire O'Donovan
is the large-scale annotation coordinator and is responsible for the TrEMBL database production at the EMBL Outstation – EBI.

Maria Jesus Martin
coordinates software development and is responsible for the TrEMBL database production at the EMBL Outstation – EBI.

Alexandre Gattiker
is undertaking a PhD in the SWISS-PROT group at the SIB.

Elisabeth Gasteiger
coordinates software development in the SWISS-PROT group at the SIB and is in charge of the ExPASy server.

Amos Bairoch
heads the SWISS-PROT group at the SIB and is a professor at the Medical Biochemistry Department of the University of Geneva.

Rolf Apweiler
heads the SWISS-PROT, TrEMBL and InterPro database activities at the EMBL Outstation – EBI.

Keywords: *evidence attribution, protein sequence, functional annotation, automatic annotation*

Abstract

SWISS-PROT is a curated protein sequence database which strives to provide a high level of annotation (such as the description of the function of a protein, its domain structure, post-translational modifications, variants, etc.), a minimal level of redundancy and a high level of integration with other databases. Together with its automatically annotated supplement TrEMBL, it provides a comprehensive and high-quality view of the current state of knowledge about proteins. Ongoing developments include the further improvement of functional and automatic annotation in the databases including evidence attribution with particular emphasis on the human, archaeal and bacterial proteomes and the provision of additional resources such as the International Protein Index (IPI) and XML format of SWISS-PROT and TrEMBL to the user community.

INTRODUCTION

SWISS-PROT

SWISS-PROT[1] is an annotated protein sequence database maintained by the Swiss Institute of Bioinformatics (SIB) and the European Bioinformatics Institute (EBI).

The SWISS-PROT database distinguishes itself from other protein sequence databases by three distinct criteria: (i) annotations, (ii) minimal redundancy and (iii) integration with other databases.

Annotation

In SWISS-PROT two classes of data can be distinguished: the core data and the annotation. For each sequence entry the core data consist of the sequence data, the citation information and the taxonomic data, while the annotation consists of the description of the following items:

- function(s) of the protein;
- post-translational modification(s), eg glycosylation, phosphorylation, acetylation, glycosylphosphatidylinositol (GPI)-anchor;
- domains and sites, eg calcium-binding regions, ATP-binding sites, zinc fingers, homeoboxes, SH2 and SH3 domains;
- secondary structure, eg alpha helix, beta sheet;
- quaternary structure, eg homodimer, heterodimer;
- similarities to other proteins;
- disease(s) associated with deficiency(ies) in the protein;
- sequence conflicts, variants, etc.

As much annotation information as possible is included in SWISS-PROT. To obtain this information we use, in addition to the publications reporting new sequence data, review articles to periodically update the annotations of families or groups of proteins. We also make use of external

Claire O'Donovan,
EMBL Outstation – European Bioinformatics Institute,
Wellcome Trust Genome Campus,
Hinxton,
Cambridge CB10 1SD, UK

Tel: +44 (0) 1223 494 460
Fax: +44 (0) 1223 494 468
E-mail: odonovan@ebi.ac.uk

experts who have been recruited to send us their comments and updates concerning specific groups of proteins.[2]

The systematic recourse, both to publications other than those reporting the core data and to subject referees, represents a unique and beneficial feature of SWISS-PROT. In SWISS-PROT, annotation is mainly found in the comment lines (CC), in the feature table (FT) and in the keyword lines (KW). Comments are classified by 'topics' to facilitate the easy retrieval of specific categories of data from the database.

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 so as to minimise the redundancy of the database. If conflicts exist between various sequencing reports, they are indicated in the feature table of the corresponding SWISS-PROT entry.

Integration with other databases

Central hub of information

Each SWISS-PROT entry should be seen as a central hub for the data available about each protein. It provides the core data directly, but additionally links to all relevant third party databases to provide access to the most comprehensive annotation for each protein. SWISS-PROT provides exhaustive cross-references to more than 43 external databases and is committed to increasing this as more databases are developed.[3] In addition to cross-references to the underlying DNA sequence database entries in the DDBJ/EMBL/GenBank nucleotide sequence databases, cross-references are derived from a number of different resources. These include model organism databases, genome databases, signature databases, protein family characterisation databases, post-translational modification (PTM) databases, 2D and 3D protein structure databases, National Center for Biotechnology Information (NCBI) taxonomy and the PubMed literature resource.

Exhaustive cross-referencing

TrEMBL

Owing to the increased data flow from genome projects to the sequence databases, SWISS-PROT faced a number of challenges to its time- and labour-intensive way of database annotation. The rate-limiting step in the production of SWISS-PROT is the careful and detailed annotation of every entry with information retrieved from the scientific literature and from rigorous sequence analysis. We do not wish to compromise on these standards but do wish to make new sequences available as soon as possible. To address this, the EBI introduced TrEMBL (Translation of EMBL nucleotide sequence database) in 1996. TrEMBL[1] consists of computer-annotated entries derived from the translation of all coding sequences (CDS) in the EMBL Nucleotide Sequence Database,[4] which are not yet integrated into SWISS-PROT. It is subdivided into two sections: SP-TrEMBL, which contains those sequences that will eventually be incorporated into SWISS-PROT, and REM-TrEMBL, which contains the sequences that will not. These include immunoglobulins and T-cell receptors, synthetic sequences, patent application sequences, fragments of fewer than eight 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 from new nucleotide sequences deposited in the EMBL nucleotide sequence database. At each TrEMBL release, the annotation of the TrEMBLnew entries is upgraded, any entries redundant against TrEMBL/SWISS-PROT[5] are merged and the remainder then progress into TrEMBL. The same approach to extensive cross-referencing described above for SWISS-PROT is implemented in TrEMBL.

There is also a serious commitment to enhancement of the annotation present in the TrEMBL entries through automatic annotation, which is described in more detail below.

SWISS-PROT and TrEMBL are available from the web sites.[6] Figure 1 shows how SWISS-PROT and TrEMBL have grown.

ONGOING DEVELOPMENTS

InterPro and automatic functional annotation in TrEMBL

Automatic functional information

With the rapid growth of sequence databases, there is an increasing need for reliable functional characterisation and annotation of newly predicted proteins. To cope with such large data volumes, faster and more effective means of protein sequence characterisation and annotation are required. One promising approach is automatic large-scale functional characterisation and annotation, which is generated with limited human interaction. To enhance the annotation of uncharacterised protein sequences in TrEMBL, the SWISS-PROT/TrEMBL group at the EBI developed a novel method for the prediction of functional information.[7] This methodology for the automated large-scale functional annotation of proteins requires three components:

- A reference database must serve as the source of annotation. SWISS-PROT is used as the reference database because of its reliable, well-annotated and standardised information.

- The highly diagnostic protein family signature database InterPro[8] supplies the means to assign proteins to groups. InterPro allows the reliable classification of proteins into families and the recognition of the domain structure of multi-domain proteins.

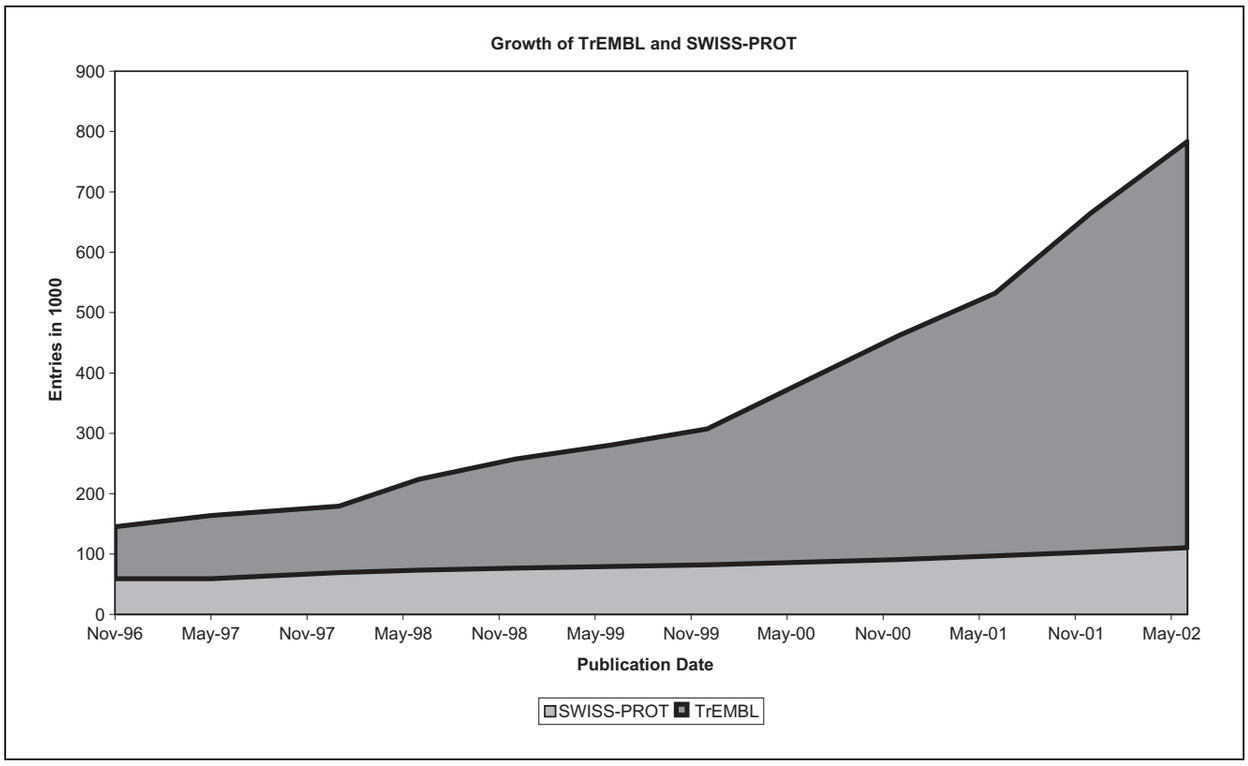

Figure 1: Growth of TrEMBL and SWISS-PROT

InterPro can classify currently around 70 per cent of all known protein sequences. This information is incorporated into SWISS-PROT and TrEMBL in the form of database cross-references to InterPro and its member databases.

- The final component is a database called RuleBase that stores and manages the annotation rules, their sources and their usage.

The actual flow of information during the automatic annotation can be divided into five steps:

Flow of information from known to unknown proteins

- Use InterPro to extract the information necessary to assign proteins to groups ('conditions') and store the conditions in RuleBase.
- Group the proteins in SWISS-PROT by the conditions.
- Extract from SWISS-PROT the common annotation shared by all functionally characterised proteins of each group and store this common annotation together with its conditions in RuleBase. Now every rule consists of conditions and the annotation common to all proteins of this group characterised by these conditions.
- Group the unannotated TrEMBL entries by the conditions stored in RuleBase.

Annotation of all known human proteins

- Add the common annotation to the unannotated TrEMBL entries. The predicted annotation is flagged with evidence tags, which will allow users to recognise the predicted nature of the annotation as well as the original source of the inferred annotation.

The RuleBase system used in TrEMBL production includes more than 500 rules that are frequently applied on TrEMBL proteins. At each run, all the rules are reviewed in the context of the current SWISS-PROT release, all annotation present in TrEMBL applied by the RuleBase system in the previous run is deleted and the rules are reapplied. Around 25 per cent of all TrEMBL entries receive additional annotation by applying RuleBase. We are committed to further extending the scope of and improving the functional information provided by this methodology to give users a comprehensive view over the output of sophisticated and well-maintained data-mining routines at a single glance.

Human Proteomics Initiative (HPI)

The goal of the HPI project[9] is to annotate all known human protein sequences according to the high-quality standards of SWISS-PROT. This also includes concomitant annotation of orthologues of human proteins in other mammalian species, mainly rat and mouse.[10] Each item of information, coming mostly from the scientific literature and predictions by sequence analysis software, is manually verified before integration into the database. The work carried out currently in the framework of the HPI project is therefore still highly dependent on the expert manual annotation skills of SWISS-PROT curators, who are assisted by external scientific experts, and sometimes by the authors of the original journal publications.

For each known protein, a wealth of information is provided that includes the description of its function, domain structure, subcellular location, similarities to other proteins, etc. Particular emphasis is placed on PTMs, isoforms produced by alternative splicing, polymorphisms and disease-linked variants. The majority of proteins are the target of PTMs, most of which are not efficiently predicted on the basis of genomic information alone. More than a hundred different types of PTM are currently known. In SWISS-PROT entries, the high level of diversity produced by PTM is mainly reflected at

the level of feature tables where cleavage sites, modified residues and bonds are indicated. An additional level of complexity is introduced by differential alternative splicing. All known splice variants of a given protein are described. As far as possible, the splice variant sequences are supplied and the effect on the protein function, subcellular location, etc. is discussed. This annotation is to be found in the feature table and in the comment lines. The sequences of such isoforms in SWISS-PROT and TrEMBL, which can be reconstructed from the feature tables are extracted and made available to the users in separate files.[11]

While all human proteins are, in the context of this project, equally important, a special effort has been made to annotate proteins encoded on chromosome 21, in a fruitful collaboration with the group of Prof. Stylianos Antonarakis (Division of Medical Genetics of the University of Geneva Medical School) who has been involved in chromosome 21 sequencing[12] and is now pursuing the analysis of the chromosome 21 transcriptome.[13] At the end of 2001, SWISS-PROT was completely synchronised with the current state of knowledge of proteins encoded on this chromosome. Currently, annotation efforts are focused on proteins encoded by the other two chromosomes that have been fully sequenced and whose encoded genes have been partially annotated, namely chromosomes 20 and 22, while keeping chromosome 21 up to date. SWISS-PROT release 40.16 of 2nd May, 2002, contains 8,110 annotated human sequences, and 2,216 additional alternatively spliced isoforms. The annotation of these proteins includes information about 13,174 variants and 20,832 PTMs. Up-to-date statistics are available.[14]

HAMAP

Complete microbial proteomes

In July 1995, the complete genomic sequence of a bacterium, *Haemophilus influenzae*, became available. That of an archaeon, *Methanoccocus jannaschii*, quickly followed it. It was the prelude to a flood of microbial genome sequences. Today close to a hundred of these genomes are available in public databases. Collectively they encode over 160,000 different protein sequences. Such a large number of sequences makes classical manual annotation an intractable task. The SWISS-PROT group at SIB therefore initiated a project that aims to annotate automatically a significant percentage of proteins originating from microbial genome sequencing projects. This project is termed HAMAP (High-quality Automated Microbial Annotation of Proteomes). It differs from other currently existing automatic annotation systems in that it does not try to hunt for distant similarity or to annotate all potential proteins originating from a microbial genome. Rather, it is being developed to deal specifically with two subsets of bacterial and archaeal proteins:

- It will automatically annotate proteins that have no recognisable similarity to any other microbial or non-microbial proteins (these are generally called 'ORFans'). This task mainly implies automatic recognition and annotation of features such as signal sequences, transmembrane domains, coiled-coil regions, inteins and ATP/GTP-binding sites.

- The most challenging part of the project is aimed at automatically annotating proteins that are part of well-defined families or subfamilies. In most cases, these are well-characterised protein families where it is possible, using software tools, to automatically build a SWISS-PROT entry of a quality identical to that produced manually by an expert curator. In order to do this, a rule system is being built that describes, for each well-defined (sub)family, the level and extent of annotation 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.

In both cases described above, the idea is to annotate proteins to the highest level of quality. The programs in development are specifically designed to track down 'eccentric' proteins. Among the peculiarities recognised by the programs are: size discrepancy, absence or divergence of regions involved in activity or binding (to metals, nucleotides, etc.), presence of paralogues, inconsistencies with the biological context (ie if a protein belongs to a pathway apparently absent in a particular organism), etc. Such 'problematic' proteins are not annotated automatically and are flagged for further analysis by SWISS-PROT expert curators. This allows curators in the SWISS-PROT groups at SIB and EBI to concentrate on the proteins that really need careful manual annotation.

Complete proteome prioritisation and processing

While the manual preparation and automatic annotation in the framework of HAMAP are limited to SWISS-PROT, there are a number of steps that need to be carried out before a genome is run through the HAMAP pipeline on the level of the nucleotide sequence databases and TrEMBL. In particular, the entry of complete microbial genomes from TrEMBLnew into TrEMBL is prioritised. Although usually proteins move from TrEMBLnew into TrEMBL only during quarterly TrEMBL releases, complete microbial genomes are included in TrEMBL whenever their sequence becomes available. It ensures that HAMAP curators and programs are able to work on the most recent data, with all the intrinsic automated TrEMBL annotation (accession numbers, taxonomy, domain-related cross-references, etc.) already present, and several operations based on the individual analysis of a genome and its annotation in EMBL are then carried out.

- **TrEMBL processing**. The nucleotide sequences are translated into TrEMBLnew entries, enter the TrEMBL protein database, and receive an accession number. At this point they are recognisable by a reference to the genome paper and the 'Complete proteome' keyword. Any annotation improvements required to conform to the SWISS-PROT standard for each proteome are implemented at this time.

- **Redundancy clean-up**. For most complete genomes, a number of protein sequences (or even a complete genome of a different strain) will have already been submitted to the databases. In this case, the entries are manually merged, the appropriate references kept and possible conflicts in the sequence highlighted. In some cases where the evolutionary distance between sequences in two strains is too high, they are not merged and are considered to be different proteomes (eg *Helicobacter pylori* strains J99 and 26695, *Neisseria meningitidis* serovars A and B).

- **Similarity searches**. At this point a number of similarity searches in the protein databases are necessary to assert whether each protein is (i) an ORF with no similarities except in very close species (ORFan), (ii) a member of a characterised family or (iii) a protein with similarities to proteins in the database for which a function is known or predicted, but that do not belong to a family *yet characterised in the context of HAMAP*. In case (i) no annotation can be propagated to it, until further biochemical characterisation has given an insight into its function. However, a number of prediction programs are run on the entry to detect putative signal sequences, transmembrane regions, and other types of domains. In case (ii) the decision must be taken with caution, as it should not produce false positive results if it is aimed to maintain the quality and consistency of the database. The method will take into account similarity to most or all members of a family throughout the length of the protein. It will also use pattern-based detection of family signatures to detect the functional class of a protein. However, not all families have a known

function, hence the concept of UPFs (uncharacterised protein families) that are interesting candidates for functional or structural studies. In case (iii) the entry undergoes manual annotation. The curators assert the relevance of the match, and if a set of reciprocally homologous proteins is found, may use them as the seed of a new family.

- **Automated sequence annotation**. The program PicoHamap has been developed to carry out the annotation automatically by reading instructions from a family rule, calling external prediction and characterisation programs, running multiple sequence alignments and writing a modified entry.

In addition to bacterial and archaeal proteomes, the use of HAMAP in the context of organelles such as chloroplast and mitochondrion is envisaged. The technologies and expertise that will be developed thanks to the HAMAP project will also be important for future automatic annotation projects in the framework of eukaryotic proteomes. Already 570 rules containing alignments and extensive annotation of protein families have been created. Even though the HAMAP project is not yet fully operational, it has facilitated the annotation of about 6,000 new microbial sequence entries in 2001. Thus the proportion of archaeal and bacterial proteins in SWISS-PROT increased from 37 to 39 per cent. In 2002, it is planned to further improve the annotation of TrEMBL, to increase the number of protein family rules to 1,000 and to complete the automatic annotation pipeline so that a significant fraction of every new microbial proteome can be annotated confidently by programs. For more information, please see the web site.[15]

Evidence attribution

Evidence tagging

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 DDBJ/EMBL/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, the EBI started, in June 2000, to add evidence tags to the internal version of TrEMBL for two purposes:

- Data tracing to allow users to see where each data item comes from and to assess its reliability.
- Data correction to allow database staff to automatically correct or update data.

Each evidence tag contains:

- Type: the description of the data source. Examples: Curator (information added by a curator according to judgement), Similarity (added by a curator due to sequence similarity to another entry), Rulebase (TrEMBL system for automatic annotation), EMBL (directly copied from the annotation of the DDBJ/EMBL/GenBank coding sequence), MGD_ADD (data imported from MGD).
- Category: evidence types are grouped into four major categories: curation, data import, automatic annotation and predicted data from programs.
- Initials: initials of the person who has last touched the data item. If a program has made the change, a dash is used.
- Attributes: one or more attributes, which describe the data source, eg the

accession number of an entry in the external data source or the version number of a program for transmembrane prediction.

- Date: The date of the last update or confirmation of the data item.

This method allows each piece of information to be easily traced back to its source and, as each piece of data may have more than one evidence tag, this system caters for data items, which are derived from multiple sources. This method will be further extended to allow tagging of known false information, eg if a keyword generated by an automated annotation system is known to be wrong by curator judgment. This will allow the incorporation of feedback from users and curators to improve the rules for automatic annotation. All data items tagged with a 'negative' evidence tag will be removed from the entry before publication.

Distinguishing between experimental and predicted data

The use of evidence tags during the annotation process will allow users to trace the source of each data item added by a curator and to readily distinguish between experimental and predicted data. It will also prevent the overwriting by a program of any data, which has been edited by a curator. This means that it will be possible to use programs to add information to a curated entry without touching manually curated data items. During the annotation process, curators will assess any information that has been added by a program, and if it is correct, they will confirm it using the appropriate evidence tag. This will increase the reliability of all program-added information. It will also allow for improvements to programs through feedback from curators to programmers in cases where curators disagree with the results from a program.

The addition of evidence tags by computer programs and during the literature-based curation process will allow users to view the source of all data items in each record and to distinguish between experimentally and computationally derived data. This is already possible, in part, both in SWISS-PROT and TrEMBL[16] through the use of a number of qualifiers or status tags that differentiate between experimental and non-experimental data. However, the introduction of an evidence attribution system that allows the source of each data item to be easily traced while also easily distinguishing between experimental and computational data will increase the value of the SWISS-PROT/TrEMBL resource for external users and will also be of use to us for effective error tracking and error correction. It is intended that an evidence-tagged version of the TrEMBL database will be available within the year. Please see the web site[17] for more information. We would welcome any feedback from the user community.

The International Protein Index (IPI)

IPI[18] 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 with the aim of providing a minimally redundant yet maximally complete set of human proteins (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. The data are presented in FastA format. IPI is also available in SWISS-PROT format. The data in SWISS-PROT format contains additional cross-references linking IPI to GO, HUGO, LocusLink and InterPro, and identifies the chromosome on which the gene encoding each IPI entry is found.

SP-ML: the SWISS-PROT and TrEMBL XML format

SP-ML provides the users with an easily parsable view on the rich data in these two databases. By using standards such as XML, XML Schema and Xlink, it relieves developers from the task of creating their own parsers and conversion tools. The first draft release was made available to the bioinformatics community in February 2002. For more information, please see the web site.[19]

Documentation files

SWISS-PROT and TrEMBL are 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 have been created recently and new files are always being added. For a list of all documents that are currently available, please see the web sites.[20]

PRACTICAL INFORMATION

Every week we also produce a complete non-redundant protein sequence collection SPTR by providing three compressed files sprot.dat.gz, trembl.dat.gz and trembl_new.dat.gz (in the directory /pub/databases/sp_tr_nrdb on the EBI server and /databases/sp_tr_nrdb on the ExPASy FTP). The files in this directory contain the weekly release versions of SWISS-PROT and TrEMBL. They immediately reflect the improvements that are continuously being done to SWISS-PROT and TrEMBL. We recommend using these versions instead of SWISS-PROT/TrEMBL releases plus updates as they also reflect merges between entries and deletions of entries. In addition, if entries are moved from TrEMBL to SWISS-PROT, a combination of the SWISS-PROT release, SWISS-PROT updates and the TrEMBL release will contains these entries twice, while they are appropriately removed from the sp_tr_nrdb versions.

SUMMARY

The combination of SWISS-PROT and TrEMBL provides a comprehensive and high-quality protein resource to the biological community. With the continuing commitment to expert manual functional curation in SWISS-PROT and the expansion of automatic annotation in TrEMBL, the databases will continue to contribute significantly to the exploitation of the sequence avalanche.

References

1. Bairoch, A. and Apweiler, R. (2000), 'The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000', *Nucleic Acids Res.*, Vol. 28(1), pp. 45–48.
2. URL: http://www.expasy.org/cgi-bin/experts/
3. Gasteiger, E., Jung, E. and Bairoch, A. (2001), 'SWISS-PROT: Connecting biological knowledge via a protein database', *Curr. Issues Mol. Biol.*, Vol. 3, pp. 47–55.
4. Stoesser, G., Baker, W., van der Broek, A. *et al.* (2001), 'The EMBL Nucleotide Sequence Database', *Nucleic Acids Res.*, Vol. 30(1), pp. 21–26.
5. O'Donovan, C., Martin, M. J., Glemet, E. *et al.* (1999), 'Removing redundancy in SWISS-PROT and TrEMBL', *Bioinformatics*, Vol. 15, pp. 258–259.
6. URLs: http://www.ebi.ac.uk/swiss-prot/and http://www.expasy.org/sprot/
7. Fleischmann, W., Moeller, S., Gateau, A. and Apweiler, R. (1999), 'A novel method for automatic functional annotation of proteins', *Bioinformatics*, Vol. 15, pp. 228–233.
8. Apweiler, R., Attwood, T. K., Bairoch, A. *et al.* (2001), 'The InterPro database, an integrated documentation resource for protein families, domains and functional sites', *Nucleic Acids Res.*, Vol. 29(1), pp. 37–40.
9. URL: http://www.expasy.org/sprot/hpi/
10. O'Donovan, C., Apweiler, R. and Bairoch, A. (2001), 'The human proteomics initiative', *Trends Biotechnol.*, Vol. 19(5), pp. 178–181.
11. Kersey, P., Hermjakob, H. and Apweiler, R. (2000), 'VARSPLIC: alternatively-spliced protein sequences derived from SWISS-PROT and TrEMBL', *Bioinformatics*, Vol. 16(11), pp. 1048–1049.
12. Hattori, M., Fujiyama, A., Taylor, T. D. *et al.* (2000), 'The DNA sequence of human chromosome 21', *Nature*, Vol. 405, pp. 311–319.
13. Reymond, A., Friedli, M., Neergaard Henrichsen, C. *et al.* (2001), 'From PREDs

and open reading frames to cDNA isolation: Revisiting the human chromosome 21 transcription map', *Genomics*, Vol. 78, pp. 46–54.

14. URL: http://www.expasy.org/sprot/hpi/hpi_stat.html
15. URL: http://www.expasy.org/sprot/hamap/
16. Junker, V., Contrino, S., Fleischmann, W. *et al.* (2000), 'The role SWISS-PROT and TrEMBL play in the genome research environment', *J. Biotechnol.*, Vol. 78, pp. 221–234.
17. URL: ftp:»ftp.ebi.ac.uk/pub/databases/trembl/evidenceDocumentation.html
18. URL: http://www.ebi.ac.uk/IPI/
19. URL: http://www.ebi.ac.uk/swissprot/SP-ML/index.html
20. URLs: http://www.expasy.org/sprot/sp_docu.html and http://www.ebi.ac.uk/swissprot/Documents/