 Nucleic Acids Research, 2000, Vol. 28, No. 1 45–48

# The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000

**Amos Bairoch* and Rolf Apweiler[1]**

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 October 12, 1999; Accepted October 13, 1999

## 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 domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases. Recent developments of the database include format and content enhancements, cross-references to additional databases, new documentation files and improvements to TrEMBL, a computer-annotated supplement to SWISS-PROT. TrEMBL consists of entries in SWISS-PROT-like format derived from the translation of all coding sequences (CDSs) in the EMBL Nucleotide Sequence Database, except the CDSs already included in SWISS-PROT. We also describe the Human Proteomics Initiative (HPI), a major project to annotate all known human sequences according to the quality standards of SWISS-PROT. SWISS-PROT is available at: http://www.expasy.ch/sprot/ and http://www.ebi.ac.uk/swissprot/**

## INTRODUCTION

SWISS-PROT (1) is an annotated protein sequence database, which was created at the Department of Medical Biochemistry of the University of Geneva and has been a collaborative effort of the Department and the European Molecular Biology Laboratory (EMBL), since 1987. SWISS-PROT is now an equal partnership between the EMBL and the Swiss Institute of Bioinformatics (SIB). The EMBL activities are carried out by its Hinxton Outstation, the European Bioinformatics Institute (EBI) (2).

The SWISS-PROT protein sequence database consists of sequence entries. Sequence entries are composed of different line types, each with their own format. For standardisation purposes the format of SWISS-PROT (see http://www.expasy.ch/txt/userman.txt ) follows as closely as possible that of the EMBL Nucleotide Sequence Database. A sample SWISS-PROT entry is shown in http://www.expasy.ch/cgi-bin/niceprot.pl?P29965

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 consists of the sequence data; the citation information (bibliographical references) and the taxonomic data (description of the biological source of the protein), while the annotation consists of the description of the following items:

- Function(s) of the protein
- Post-translational modification(s). For example carbohydrates, phosphorylation, acetylation, GPI-anchor, etc.
- Domains and sites. For example calcium binding regions, ATP-binding sites, zinc fingers, homeoboxes, SH2 and SH3 domains, etc.
- Secondary structure. For example alpha helix, beta sheet, etc.
- Quaternary structure. For example homodimer, heterotrimer, etc.
- Similarities to other proteins
- Disease(s) associated with deficiencie(s) in the protein
- Sequence conflicts, variants, etc.

We try to include as much annotation information as possible 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 experts who have been recruited to send us their comments and updates concerning specific groups of proteins (see http://www.expasy.ch/cgi-bin/experts ).

We believe that 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). Most comments are classified by 'topics'; this approach permits 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

*To whom correspondence should be addressed. Tel: +41 22 702 5477; Fax: +41 22 702 5502; Email: amos.bairoch@medecine.unige.ch

indicated in the feature table of the corresponding SWISS-PROT entry.

### Integration with other databases

It is important to provide the users of biomolecular databases with a degree of integration between the three types of sequence-related databases (nucleic acid sequences, protein sequences and protein tertiary structures) as well as with specialised data collections. Cross-references are provided in the form of pointers to information related to SWISS-PROT entries and found in data collections other than SWISS-PROT. For example the sample sequence mentioned above contains, among others, DR (Databank Reference) lines that point to EMBL, PDB, OMIM, Pfam and PROSITE. In this particular example it is therefore possible to retrieve the nucleic acid sequence(s) that codes for that protein (EMBL), the description of genetic disease(s) associated with that protein (OMIM), the 3D structure (PDB) or information specific to the protein family to which it belongs (PROSITE and Pfam).

## RECENT DEVELOPMENTS

### Model organisms

We have selected a number of organisms that are the target of genome sequencing and/or mapping projects and for which we intend:

(i) to be as complete as possible;
(ii) to provide a higher level of annotation;
(iii) to provide cross-references to specialised database(s) that contain, among other data, some genetic information about the genes that code for these proteins; and
(iv) to provide specific indices or documents.

The organisms currently selected are: *Arabidopsis thaliana* (mouse-ear cress), *Bacillus subtilis*, *Caenorhabditis elegans* (worm), *Candida albicans*, *Dictyostelium discoideum* (slime mold), *Drosophila melanogaster* (fruit fly), *Escherichia coli*, *Haemophilus influenzae*, *Helicobacter pylori*, *Homo sapiens* (human), *Methanococcus jannaschii*, *Mus musculus* (mouse), *Mycobacterium tuberculosis*, *Mycoplasma genitalium*, *Saccharomyces cerevisiae* (budding yeast), *Salmonella typhimurium*, *Schizosaccharomyces pombe* (fission yeast), *Sulfolobus solfataricus* and *Synechocystis* sp. PCC 6803.

Collectively these organisms represent ~40% of the total number of sequence entries in SWISS-PROT. We are currently attempting to finish the integration into SWISS-PROT of all the predicted proteins from *E.coli*, *B.subtilis*, *M.jannaschii* and yeast.

### The Human Proteomics Initiative (HPI)

In a few months the combined efforts of a number of sequencing centers and companies will produce a first draft of the human genome sequence. Such an endeavor is only a very preliminary step in the understanding of human biological processes. The first pitfall to overcome is the detection of all coding regions on the genomic sequence. Current algorithms, while being very powerful, are not capable of detecting with certainty all exons, are not well equipped to distinguish different splice variants and are unable to detect small proteins (which are numerous and crucial to many biological processes). Even when all potential coding regions have been predicted, the user community will have at its disposition the sequences of between 80 000 and 100 000 'naked' proteins. We call these proteins 'naked' because genomic information does not allow the efficient prediction of all the post-translational modifications (PTM) of which the majority of proteins are the target. Proteins, once synthesized on the ribosomes, are subject to a multitude of modification steps. The complexity due to all these modifications is compounded by the high level of diversity that alternative splicing can produce at the level of sequence. Thus the number of different protein molecules expressed by the human genome is probably closer to a million than to the hundred thousand generally considered by genome scientists. Another factor of complexity to take into account is the amount of polymorphism at the protein sequence level. While some of these polymorphisms are linked to disease states, most are not, yet have in many cases a direct or indirect effect on the activities of the proteins.

We are therefore initiating a major project to annotate all known human sequences according to the quality standards of SWISS-PROT. This means providing, for each known protein, a wealth of information that includes the description of its function, its domain structure, subcellular location, post-translational modifications, variants, similarities to other proteins, etc. There are currently slightly more than 5400 annotated human sequences in SWISS-PROT. These entries are associated with ~14 500 literature references, 16 000 experimental or predicted PTMs, 800 splice variants and 8000 polymorphisms (most of which are linked with disease states). We will use the current information as the ground basis for what we call the 'Human Proteomics Initiative' (HPI).

The HPI project contains a number of sub-components, which are briefly described below:

- Annotation of all known human proteins. In the course of the next 9 months (up to April 2000) the human protein sequences that are not yet in SWISS-PROT will be fully annotated. We will also review and complete the annotation of the human sequences currently in SWISS-PROT. At the end of this 9-month period we expect to be complete and up-to-date and to hereafter keep up with the appearance of new data relevant to human proteins.
- Annotation of mammalian orthologs of human proteins. We will make sure that for any human proteins, existing orthologs in other mammalian species will also be annotated at a level equivalent to that of the cognate human sequences.
- Annotation of all known human polymorphisms at the protein sequence level. As mentioned above, SWISS-PROT already holds information on a sizeable amount of such polymorphisms, and it will significantly expand its effort to store and annotate all 'small' variations at the protein level.
- Annotation of all known post-translational modifications in human proteins. During the next 9 months a major effort will be made to supplement the already quite comprehensive description of known post-translational modifications in human proteins currently provided in SWISS-PROT.
- Tight links to structural information. SWISS-PROT is tightly linked to the PDB/RCSB 3D-structure database and already includes many features useful to structural biologists. These tight links will be further expanded by providing homology-derived models for all human proteins for which such an approach is scientifically relevant.

For all aspects of the HPI projects, we would 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 to speed up the comprehensive annotation of the human proteome.

The HPI project has two different time-related aspects: one of which is a 9-month 'marathon' to catch up with the current state of research, the other one is a long-term commitment to keep such a project alive as long as it is necessary. For a detailed description of the HPI project and its current status please consult http://www.expasy.ch/sprot/hpi/

### Format and content enhancements

Data stored in SWISS-PROT used to be represented exclusively in upper case. We initiated a process to convert the data into mixed case. This process is already well under way and will be completed during the year 2000. In the last 12 months we have added new comment topics ('Miscellaneous' and 'Pharmaceutical') as well as a new feature key ('Se_Cys'). Major changes to standardize the use and content of the 'Similarity' and 'Alternative product' comments topics are nearing completion. The formats used to store book and patent references have been modified so as to make this information computer parsable.

To contribute to the standardization of the taxonomies used in molecular sequence databases we have switched to the NCBI taxonomy, which is used by the DDBJ/EMBL/GenBank nucleotide sequence databases. The taxonomic classification maintained at the NCBI is available at: http://www.ncbi.nlm.nih.gov/Taxonomy/

### 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 have been created recently and we are continuously adding new files. See http://www.expasy/sprot/sp_docu.html for a list of all the documents that are currently available.

### New cross-references

We have recently added cross-references that link SWISS-PROT to the Zebrafish Information Network (ZFIN) database (3) (see http://zfish.uoregon.edu/ZFIN/ ). We also have started to add cross-references from SWISS-PROT to the CarbBank Complex Carbohydrate Structure Database (CCSD) (see http://128.192.9.29/carbbank/ ).

Currently, SWISS-PROT is linked to 31 different databases and has consolidated its role as the major focal point of biomolecular databases interconnectivity. In release 38, there is an average of 4.5 cross-references for each sequence entry.

## TrEMBL: A COMPUTER ANNOTATED SUPPLEMENT TO SWISS-PROT

### Introduction

Due to the increased data flow from genome projects to the sequence databases we face a number of challenges to our way of database annotation. Maintaining the high quality of sequence and annotation in SWISS-PROT requires careful sequence analysis and detailed annotation of every entry. This is the rate-limiting step in the production of SWISS-PROT. On one hand we do not wish to relax the high editorial standards of SWISS-PROT and it is clear that there is a limit to how much we can accelerate the annotation procedures. On the other hand, it is also vital that we make new sequences available as quickly as possible. To address this concern, we introduced in 1996 TrEMBL (Translation of EMBL nucleotide sequence database). TrEMBL consists of computer-annotated entries derived from the translation of all coding sequences (CDSs) in the EMBL database, except for CDSs already included in SWISS-PROT.

### Current status

In July 1999, TrEMBL release 11 was produced. Release 11 was based on the translation of all 379 000 CDSs in the EMBL Nucleotide Sequence Database release 58. Around 119 000 of these CDSs were already as sequence reports in SWISS-PROT and thus excluded from TrEMBL. The remaining 260 000 sequence entries have been automatically merged whenever possible to reduce redundancy in TrEMBL. This step has led to 245 761 TrEMBL entries.

We have split TrEMBL into two main sections; SP-TrEMBL and REM-TrEMBL: SP-TrEMBL (SWISS-PROT TrEMBL) contains the entries (199 794 in release 11) which should be incorporated into SWISS-PROT. SWISS-PROT accession numbers have been assigned to these entries. SP-TrEMBL is partially redundant against SWISS-PROT, since ~60 000 of these entries are only additional sequence reports of proteins already in SWISS-PROT. For TrEMBL to act as a computer-annotated supplement to SWISS-PROT, new procedures have been introduced to remove redundancy (4) and to automatically add highly reliable annotation (5).

A rule-based system making use of the existing SWISS-PROT annotation as the gold standard is applied to improve the TrEMBL annotation. Right now this process affects only 15% of all TrEMBL entries. The creation of additional rules will be one of the priorities for TrEMBL over the next year. This should lead to a drastic increase in coverage by automatic annotation.

REM-TrEMBL (REMaining TrEMBL) contains the entries (~46 000 in release 11) that we do not want to include in SWISS-PROT.

## PRACTICAL INFORMATION

The use of SWISS-PROT is free for academic users. However, in September 1998 we implemented a system of annual subscription fee for commercial users of the database. The SIB and the EMBL/EBI mandated a new company, Geneva Bioinformatics (GeneBio) (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 will be used at SIB and the EBI to bring SWISS-PROT up to date, to keep it up to date, and to further enhance its quality. Further information on this new system is available from http://www.expasy.ch/announce/

### Content of the current SWISS-PROT release

Currently (October 1999), SWISS-PROT contains ~81 000 sequence entries, comprising 30 million amino acids

abstracted from ~65 000 references. The data file (sequences and annotations) requires 185 Mb of disk storage space. The documentation and index files require ~65 Mb of disk space.

### 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 WWW molecular biology server ExPASy (6) as well as the one developed by the EBI. The ExPASy Web server was made available to the public in September 1993. In October 1999 a cumulative total of 60 million connections was attained. Its address is:

http://www.expasy.ch/

Complete and up to date mirror sites of ExPASy are available in Australia, Canada and Taiwan:

http://expasy.proteome.org.au/ (at Australian Proteome Analysis Facility, Sydney)

http://expasy.cbr.nrc.ca/ (at Canadian Bioinformatics Resource, Halifax)

http://expasy.nhri.org.tw/ (at National Health Research Institutes, Taipei)

The EBI server is accessible at:

http://www.ebi.ac.uk/

On both the ExPASy and the EBI Web servers, you can use the Sequence Retrieval System (SRS) (6) software package to query and retrieve sequence entries. The EBI and SIB also offer a range of search services (see http://www2.ebi.ac.uk/ or http://www.expasy.ch/tools/ ) to run Smith–Waterman, FASTA and BLAST sequence similarity searches against SWISS-PROT + TrEMBL.

### How to obtain the full SWISS-PROT and/or TrEMBL releases

SWISS-PROT + TrEMBL is distributed on CD-ROM by the EBI (2). The CD-ROMs also contain some database query and retrieval software for MS-DOS and Apple Macintosh computers. For all enquiries contact: The EMBL Outstation—The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. Tel: +44 1223 494 444; Fax: +44 1223 494 468; Email: datalib@ebi.ac.uk

From a computer linked to the Internet you can obtain SWISS-PROT and TrEMBL using anonymous FTP (File Transfer Protocol) from the following servers: ftp.expasy.ch and ftp.ebi.ac.uk

### How to submit data or updates/corrections to SWISS-PROT

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 1223 494 457; Fax: +44 1223 494 468; Email: datasubs@ebi.ac.uk (for submission), or datalib@ebi.ac.uk (for enquiries).

To submit updates and/or corrections to SWISS-PROT you can either use the Email address: swiss-prot@expasy.ch or the WWW address http://www.expasy.ch/sprot/sp_update_form.html

### Release frequency, weekly updates and non-redundant data sets

The current distribution frequency is four releases per year. Weekly updates are also available; these updates are available by anonymous FTP. For SWISS-PROT, three files are updated every week:

- new_seq.dat Contains all the new entries since the last full release.
- upd_seq.dat Contains the entries for which the sequence data has been updated since the last release.
- upd_ann.dat Contains the entries for which one or more annotation fields have been updated since the last release.

For TrEMBL, a file containing all the new entries since the last full release (trembl_new.dat) is updated every week.

These files are available on the EBI and ExPASy servers, whose Internet addresses are listed above.

Every week we also produce a complete non-redundant protein sequence collection by providing three compressed files (in the directory /databases/sp_tr_nrdb on the ExPASy FTP server and in /pub/databases/sp_tr_nrdb on the EBI server): sprot.dat.Z, trembl.dat.Z and trembl_new.dat.Z.

### Swiss-Shop

Swiss-Shop is an automated sequence alerting system which allows users to obtain, by Email, new sequence entries relevant to their field(s) of interest. Keyword-based and sequence/pattern-based requests are possible. Every time a weekly SWISS-PROT release is performed, all new database entries matching the user-specified search keywords or patterns and the entries showing sequence similarities to the user-specified sequence will be sent automatically to the user by Email. Swiss-Shop requests can be submitted at http://www.expasy.ch/swisshop/

## REFERENCES


1. Bairoch,A. and Apweiler,R. (1999) *Nucleic Acids Res.*, **27**, 49–54.
2. Stoesser,G., Tuli,M.A., Lopez,R. and Sterk,P. (1999) *Nucleic Acids Res.*, **27**, 18–24. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 19–23.
3. Westerfield,M., Doerry,E., Kirkpatrick,A.E. and Douglas,S.A. (1999) *Methods Cell Biol.*, **60**, 339–355.
4. Fleischmann,W., Moeller,S., Gateau,A. and Apweiler,R. (1999) *Bioinfomatics*, **15**, 228–233.
5. Appel,R.D., Bairoch,A. and Hochstrasser,D.F. (1994) *Trends Biochem. Sci.*, **19**, 258–260.
6. Etzold,T. and Argos,P. (1993) *Comput. Appl. Biosci.*, **9**, 49–57.