# ORGANISATION AND STANDARDISATION OF INFORMATION IN SWISS-PROT AND TREMBL

***Michele Magrane* and Rolf Apweiler.***

*EMBL Outstation – European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, U.K.*
*Email:* magrane@ebi.ac.uk

## ABSTRACT

*SWISS-PROT is a curated, non-redundant protein sequence database which provides a high level of annotation and is integrated with a large number of other biological databases. It is supplemented by TrEMBL, a computer-annotated database which contains translations of all coding sequences in the EMBL Nucleotide Sequence Database which are not yet in SWISS-PROT. Each fully curated SWISS-PROT entry contains as much up-to-date information as possible from a variety of sources and the high quality of the annotation in SWISS-PROT provides the basis for the procedure which is used to automatically annotate the TrEMBL database. The large amounts of different data types found in both databases are stored in a highly structured and uniform manner and this structured organisation means that SWISS-PROT and TrEMBL together provide a comprehensive resource with data that are readily accessible for users and easily retrievable by computer programs.*

**Keywords:** Bioinformatics, Protein sequence, Database, Standardisation, Annotation.

## 1. INTRODUCTION

SWISS-PROT (Bairoch & Apweiler, 2000) is a curated protein sequence database which is maintained collaboratively by the Swiss Institute of Bioinformatics (SIB) and the European Bioinformatics Institute (EBI), an outstation of the European Molecular Biology Laboratory (EMBL). The database distinguishes itself from other protein sequence databases by three distinct criteria:
(i) It provides a high level of annotation. The entries are annotated by a team of biologists who use a variety of sources such as scientific literature, other databases, prediction programs and the help of external experts to add as much accurate and up-to-date information as possible to each entry.
(ii) It is non-redundant which means that all reports for a given protein are merged into a single entry, thus summarising many pages of scientific literature into a concise but comprehensive report.
(iii) It provides a high level of integration with other databases. Cross-references are provided to other sequence databases as well as to specialised data collections. Currently, there are cross-references to more than 40 different databases and this allows users to access a large amount of additional information related to a particular protein.

SWISS-PROT is supplemented by TrEMBL (Bairoch & Apweiler, 2000), a computer-annotated database which contains translations of all coding sequences in the EMBL Nucleotide Sequence Database (Stoesser, Baker, van den Broek, Camon, Garcia-Pastor, Kanz et al., 2002) which are not yet in SWISS-PROT. TrEMBL was created in 1996 due to the dramatic increase in data from genome sequencing projects and allows these sequences to be made publicly available as quickly as possible without diluting the high quality annotation found in SWISS-PROT.

The SWISS-PROT and TrEMBL databases can be accessed using either the EBI server at http://www.ebi.ac.uk/swissprot/ (SWISS-PROT Protein Knowledgebase, 1986) or the ExPASy server at SIB at http://www.expasy.ch/sprot/ (TrEMBL, n.d.).

The information in SWISS-PROT and TrEMBL is highly organised and structured and this is achieved through standardisation in a number of different areas including data storage and database management, data format, syntax and semantics of data items, and data analysis and automation of annotation and each of these areas will be discussed in detail below.

## 2. DATA STORAGE AND DATABASE MANAGEMENT

Although SWISS-PROT and TrEMBL are currently distributed as flatfile databases, both databases are now stored in ORACLE and, in the near future, production of the databases will switch to this system. This relational version of the databases is based on the relational schema used by the EMBL Nucleotide Sequence Database and shares as many parts of the EMBL schema as possible. Around the database, there is a C++ enwrapping which allows for basic operations on the data such as loading and unloading of entries, creation of releases, and updates, and this is a modified version of the code used in the EMBL database. So, using the EMBL schema and code which caters for nucleic acid entries, a modified schema and code have been designed to accommodate SWISS-PROT and TrEMBL protein entries which keeps as much compatibility of code with EMBL as possible and allows for easier maintenance.

## 3. DATA FORMAT

The SWISS-PROT and TrEMBL databases share a common data format which means that all line types used in SWISS-PROT are also used in TrEMBL and, wherever possible, a particular line type has the same format in both databases. There are some necessary exceptions to this shared format. For example, the data class used in the ID (identification) line of a SWISS-PROT entry is “STANDARD” which shows that the entry has been fully curated whereas in TrEMBL, it is “PRELIMINARY” which shows that the entry has not yet been manually curated. However, apart from a small number of such differences, the format in both databases is identical.

The format of the SWISS-PROT and TrEMBL databases follows that of the EMBL Nucleotide Sequence Database as closely as possible so that the general structure of an entry is identical in all three databases. This means that many line types found in the EMBL database are also present in SWISS-PROT and TrEMBL and have the same format as that used in EMBL. There are some differences such as line types defined in one database but not in the others or slight differences between the databases within a given line type but, where possible, all three databases share the same format.

## 4. SYNTAX AND SEMANTICS OF DATA ITEMS

### 4.1 Data types

The SWISS-PROT and TrEMBL databases consist of sequence entries (Figure 1) which are composed of different line types, each one having its own specified format. A full list of the line types used can be found in the SWISS-PROT user manual (SWISS-PROT, 2001). In SWISS-PROT, two classes of data can be distinguished, core data and annotation.

### 4.2 Core data

The core data is generally provided by the submitter of the sequence and consists of sequence data which come either from the translation of the corresponding nucleotide sequence in the EMBL Nucleotide Sequence Database or from submissions to SWISS-PROT in the case of peptide sequences; citation information which shows where the data has been published or, if unpublished, to which database it has been submitted; and taxonomic data which shows the biological source of the protein.

### 4.3 Annotation

The SWISS-PROT database strives to provide a high level of annotation and this is achieved through extraction of relevant information from scientific literature and rigorous sequence analysis by a team of biologists. Use is also made of external experts who have been recruited to send us their comments and updates concerning specific groups of proteins. This process allows the addition of as much correct and up-to-date information as possible about each protein including descriptions of properties such as function(s) of the protein, post-translational modifications, domains and sites, secondary and quaternary structure, similarities to other proteins, diseases associated with deficiencies in a protein, developmental stages in which the protein is expressed, in which tissues the protein is found, pathways in which the protein is involved, and sequence conflicts and variants. The annotation is stored mainly in the comment or CC lines, the feature table or FT lines, and the keyword or KW lines. There are currently (Release 40) more than 300,000 CC lines, 470,000 FT lines and 300,000 keywords in SWISS-PROT.

```
ID   GCDH_MOUSE     STANDARD;      PRT;   438 AA.
AC   Q60759;
DT   01-NOV-1997 (Rel. 35, Created)
DT   01-NOV-1997 (Rel. 35, Last sequence update)
DT   01-OCT-2000 (Rel. 40, Last annotation update)
DE   GLUTARYL-COA DEHYDROGENASE, MITOCHONDRIAL PRECURSOR (EC 1.3.99.7)
DE   (GCD).
GN   GCDH.
OS   Mus musculus (Mouse).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
OX   NCBI_TaxID=10090;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=129/SV; TISSUE=LIVER;
RX   MEDLINE=96039264; PubMed=7490088;
RA   Koeller D.M., Digiulio K.A., Angeloni S.V., Dowler L.L., Frerman F.E.,
RA   White R.A., Goodman S.I.;
RT   "Cloning, structure, and chromosome localization of the mouse
RT   glutaryl-CoA dehydrogenase gene.";
RL   Genomics 28:508-512(1995).
CC   -!- FUNCTION: CATALYZES THE OXIDATIVE DECARBOXYLATION OF GLUTARYL-COA
CC       TO CROTONYL-COA AND CO(2) IN THE DEGRADATIVE PATHWAY OF L-LYSINE,
CC       L-HYDROXYLYSINE, AND L-TRYPTOPHAN METABOLISM. IT USES ELECTRON
CC       TRANSFER FLAVOPROTEIN AS ITS ELECTRON ACCEPTOR.
CC   -!- CATALYTIC ACTIVITY: GLUTARYL-COA + ACCEPTOR = CROTONOYL-COA +
CC       CO(2) + REDUCED ACCEPTOR.
CC   -!- COFACTOR: FAD.
CC   -!- PATHWAY: DEGRADATIVE PATHWAY OF L-LYSINE, L-HYDROXYLYSINE,
CC       AND L-TRYPTOPHAN METABOLISM.
CC   -!- SUBUNIT: HOMOTETRAMER.
CC   -!- SUBCELLULAR LOCATION: MITOCHONDRIAL MATRIX.
CC   -!- SIMILARITY: BELONGS TO THE ACYL-COA DEHYDROGENASES FAMILY.
DR   EMBL; U18992; AAB04679.1; -.
DR   HSSP; Q06319; 1BUC.
DR   MGD; MGI:104541; Gcdh.
DR   InterPro; IPR001552; Acyl-CoA_dh.
DR   Pfam; PF00441; Acyl-CoA_dh; 2.
DR   PROSITE; PS00072; ACYL_COA_DH_1; FALSE_NEG.
DR   PROSITE; PS00073; ACYL_COA_DH_2; 1.
KW   Oxidoreductase; Flavoprotein; FAD; Mitochondrion; Transit peptide.
FT   TRANSIT       1     44       MITOCHONDRION (POTENTIAL).
FT   CHAIN        45    438       GLUTARYL-COA DEHYDROGENASE.
FT   ACT_SITE    414    414       BASE (POTENTIAL).
SQ   SEQUENCE   438 AA;  48646 MW;  B91D78319067753E CRC64;
     MSLRGVSAQL LSRRSGLRFP RFPRTWSSAA AHTEKTQIRP AKSSRPVFDW KDPLILEEQL
     TADEKLIRDT FRNYWQERLM SQILLANRNE VFHRDIVYEM GELGVLGPTI KGYGCAGVSS
     VAYGLLTREL ERVDSGYRSM MSVQSSLVMH PIYTYGSEEQ RQKYLPRLAK GELLGCFGLT
     EPNHGSDPGG METRARHNPS NQSYTLSGTK TWITNSPVAD LFIVWARCED NCIPGFILEK
     GMRGSSAPRI EGKFSLRASA TGMIIMDSVE VPEENVLPNV SSLAGPFGCL NTARYGITWG
     VLGAAEFCLH TARQYALDRI QFGVPLARNQ LVQKKLADML TEITLGLHAC LQLGRLKDQD
     KATPEMVSML KRNNCGKALD IARQARDILG GNGISDEYHV IRHAMNLEAV NTYEGTHDIH
     ALILGRAITG IQAFTVGK
//
```

**Figure 1.** Example of a SWISS-PROT entry

## 4.4 CC lines

The comment or CC lines are free text comments which are used to convey any useful information about a protein. The information in the CC lines is contained in a number of defined topics which allows the easy retrieval of specific categories of data from the database. A full list of the currently used comment topics and their definitions is shown in Table 1.

## 4.5 FT lines

The feature table or FT lines provide a way of annotating position-specific data relating to the sequence. The lines have a fixed format and a defined set of feature keys which may be used. These feature keys describe domains and sites of interest within a sequence such as post-translationally modified residues, binding sites, enzyme active sites, secondary structure, and any other regions of interest. The full list of currently defined feature keys can be found in the SWISS-PROT user manual (SWISS-PROT, 2001).

**Table 1.** Comment topics used in the SWISS-PROT database

| Comment topic | Description |
|---|---|
| ALTERNATIVE PRODUCTS | Description of the existence of protein sequences produced by alternative splicing of the same gene or by the use of alternative initiation codons. |
| BIOTECHNOLOGY | Description of the biotechnological use(s) of a protein |
| CATALYTIC ACTIVITY | Description of the reaction(s) catalyzed by an enzyme |
| CAUTION | Warns about possible errors and/or grounds for confusion |
| COFACTOR | Description of an enzyme cofactor |
| DATABASE | Description of a cross-reference to a database for a specific protein |
| DEVELOPMENTAL STAGE | Description of the developmental-specific expression of a protein |
| DISEASE | Description of disease(s) associated with a deficiency of a protein |
| DOMAIN | Description of the domain structure of a protein |
| ENZYME REGULATION | Description of an enzyme regulatory mechanism |
| FUNCTION | Description of the function(s) of a protein |
| INDUCTION | Description of compound(s) which stimulate the synthesis of a protein |
| MASS SPECTROMETRY | Reports the exact molecular weight of a protein or part of a protein as determined by mass spectrometric methods |
| MISCELLANEOUS | Any comment which does not belong to any of the other defined topics |
| PATHWAY | Description of the metabolic pathway(s) with which a protein is associated |
| PHARMACEUTICAL | Description of the use of a protein as a pharmaceutical drug |
| POLYMORPHISM | Description of polymorphism(s) |
| PTM | Description of a post-translational modification |
| SIMILARITY | Description of the similarity (sequence or structural) of a protein with other proteins |
| SUBCELLULAR LOCATION | Description of the subcellular location of the mature protein |
| SUBUNIT | Description of the quaternary structure of a protein |
| TISSUE SPECIFICITY | Description of the tissue specificity of a protein |

### 4.6 Keywords

The keywords are found in the keyword or KW lines of an entry. They serve as a subject reference for each sequence and assist in the retrieval of specific categories of data from the database. A controlled list of approximately 800 keywords, each with a definition to clarify its biological meaning and intended usage, is maintained. The full list of currently defined keywords is available at http://www.expasy.org/cgi-bin/keywlist.pl.

## 5. AUTOMATION OF ANNOTATION

### 5.1 Annotation bottleneck

To produce a fully curated SWISS-PROT entry containing all of the above types of data is a highly labour-intensive process. This is the rate-limiting step in the production of SWISS-PROT as entries come into TrEMBL more quickly than they can be manually annotated and integrated into SWISS-PROT, thus creating a bottleneck of entries awaiting annotation. While it is necessary to maintain the high standard of annotation in SWISS-PROT, it is also vital to enhance the annotation of the proteins in TrEMBL, many of which are uncharacterised and about which very little functional information is known. This problem can be partly overcome by automatic annotation of TrEMBL entries (Apweiler, 2001).

### 5.2 Overcoming the annotation bottleneck by automatic annotation

For automatic annotation, a novel system of standardised transfer of annotation from well-characterised proteins in SWISS-PROT to unannotated TrEMBL entries has been developed (Fleischmann, Moeller, Gateau & Apweiler, 1999). Using this system, a TrEMBL entry is reliably recognised by a given method as being a member of a certain group of proteins. The annotation shared by the functionally characterised SWISS-PROT proteins of the group is then extracted and is assigned to the unannotated TrEMBL entry.

For such a system to work successfully, a number of requirements must be met. Firstly, a well-annotated reference database is needed from which annotation can be extracted for transfer to unannotated entries. For the automatic annotation of TrEMBL, the SWISS-PROT database is used as the source of high quality annotation because of its well-annotated and standardised content.

Secondly, there needs to be a system to store and manage the annotation rules used in the system and for this, RuleBase (Apweiler, 2001), a database which contains the rules as well their sources and usage, has been developed.

Thirdly, a highly diagnostic protein family signature database is necessary to supply the means to assign proteins to groups. To assign TrEMBL entries into groups, InterPro (Apweiler, Attwood, Bairoch, Bateman, Birney, Biswas et al., 2001) 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 (Falquet, Pagni, Bucher, Hulo, Sigrist, Hofmann et al., 2002), PRINTS (Attwood, Blythe, Flower, Gaulton, Mabey, Maudling et al., 2002), Pfam (Bateman, Birney, Cerruti, Durbin, Etwiller, Eddy et al., 2002), ProDom (Corpet, Servant, Gouzy & Kahn, 2000), SMART (Letunic, Goodstadt, Dickens, Doerks, Schultz, Mott et al., 2002) and TIGRFAMs (Haft, Loftus, Richardson, Yang, Eisen, Paulsen et al., 2001).

The final requirement for the success of the above system is that all of the above should be stored in a proper database management system and this is met by the fact that SWISS-PROT, TrEMBL, RuleBase and InterPro are all stored in ORACLE.

This process of automatic annotation brings the standard of annotation in TrEMBL closer to that found in SWISS-PROT through the addition of accurate, high-quality information to TrEMBL entries, thus improving the quality of data available to the user.

## 6. CONCLUSIONS

The SWISS-PROT and TrEMBL databases together provide a complete collection of protein sequences with minimal redundancy and offer a high level of integration with a large number of other biological databases. Each SWISS-PROT entry is manually annotated by a biologist, thus ensuring that the quality of information in the database is as accurate and up-to-date as possible. Using this high quality annotation as a basis, a procedure has been developed to automatically annotate the TrEMBL database. This system adds information to entries which are awaiting manual curation and improves the quality of data available to users of the TrEMBL database. All information items in the SWISS-PROT and TrEMBL databases are stored in a highly structured and uniform manner which means that they are easily retrievable by users and by computer programs in a consistent manner.

## 7. REFERENCES

Apweiler, R. (2001) Functional information in SWISS-PROT: the basis for large-scale characterisation of protein sequences. *Briefings in Bioinformatics* 2(1),9-18.

Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N.J., Oinn, T.M., Pagni, M., Servant, F., Sigrist, C.J., Zdobnov, E.M. (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res*. 29(1),37-40.

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., Scordis, P. (2002) PRINTS and PRINTS-S shed light on protein ancestry. *Nucleic Acids Res*. 30(1),239-241.

Bairoch, A., Apweiler, R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* 28(1),45–48.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., Sonnhammer, E.L.L. (2002) The Pfam Protein Families Database. *Nucleic Acids Res.* 30(1),276-280.

Corpet, F., Servant, F., Gouzy, J., Kahn, D. (2000) ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res*. 28(1),267-269.

Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C.J.A., Hofmann, K., Bairoch, A. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res*. 30(1),235-238.

Fleischmann, W., Moeller, S., Gateau, A., Apweiler, R. (1999) A novel method for automatic and reliable functional annotation. *Bioinformatics* 15(3),228-233.

Haft, D.H., Loftus, B.J., Richardson, D.L., Yang, F., Eisen, J.A., Paulsen, I.T., White, O. (2001) TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res*. 29(1),41-43.

Letunic, I., Goodstadt, L., Dickens, N.J., Doerks, T., Schultz, J., Mott, R., Ciccarelli, F., Copley, R.R., Ponting, C.P., Bork P. (2002) Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res.* 30(1),242-244.

Stoesser, G., Baker, W., van den Broek, A., Camon, E., Garcia-Pastor, M., Kanz, C., Kulikova, T., Leinonen, R., Lin, Q., Lombard, V., Lopez, R., Redaschi, N., Stoehr, P., Tuli, M.A., Tzouvara, K., Vaughan, R. (2002) The EMBL Nucleotide Sequence Database. *Nucleic Acids Res*. 30(1),21-26.

SWISS-PROT. (2001) SWISS-PROT Knowledgebase User Manual (Release 40). Retrieved April 20, 2002 from the Swiss Institute of Bioinformatics Web site: http://www.expasy.org/sprot/userman.html

SWISS-PROT Protein Knowledgebase (1986) Available from the European Bioinformatics Institute Web site: http://www.ebi.ac.uk/swissprot/

TrEMBL (n.d) Available from the Swiss Institute of Bioinformatics Web site: http://www.expasy.ch/sprot/.