SWISS-PROT

https://iop.vast.ac.vn/theor/conferences/smp/1st/kaminuma/SWISSPROT/index.html

Dr. Kuldeep Sharma

Assistant Professor

Department of Botany

M.L.S. University, Udaipur

Rajasthan, India

- SWISS-PROT established in 1986 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 (PTM), variants, etc.),
  - ✓ a minimal level of redundancy and
  - ✓ high level of integration with other databases.
- It was initiated and maintained by A. Bairoch at the Department of Medical Biochemistry of the University of Geneva in collaboration of the EMBL Data Library, since 1987.
- SWISS-PROT is now an equal partnership between the EMBL and Swiss Institute of Bioinformatics (SIB).
- The EMBL activities are carried out by The European Bioinformatics Institute (EBI) at Hinxton outstation, Cambridge UK.
- The SWISS-PROT group is headed by Rolf Apweiler.

- SWISS-PROT is a curated, added-value database, not a repository of primary information.
- SWISS-PROT's curation team adds detailed annotation and organisation to protein sequences, the overwhelming majority of which come from translations from the public nucleotide sequence databases. The value of SWISS-PROT to the academic and commercial research community is widely accepted. It is the gold standard for scientific databases and must be rendered secure.

Access to Swiss Prot:

- SRS is the easiest and simplest method available to quickly access the SWISS-PROT sequence database.
- Release 40.0 of SWISS-PROT contains 101'602 sequence entries, comprising 37'315'215 amino acids abstracted from 91'880 references. This represents an increase of 18% over release 39.
- SWISS-PROT is accompanied by TrEMBL, a computer-annotated supplement to SWISS-PROT. TrEMBL contains the translations of all coding sequences (CDS) present in the DDBJ/EMBL/GenBank Nucleotide Sequence Database and also protein sequences extracted from the literature or submitted to SWISS-PROT, which are not yet integrated into SWISS-PROT.

- No license fee will be charged to academic users, nor will any restrictions be imposed on their use or reuse of the data.
- Nothing will change in the methods by which academic or commercial users can access SWISS-PROT, but commercial users will be informed that their company is liable to pay a license fee irrespective of the method by which they access the database.
- Third party organisations providing services which make use of SWISS-PROT need not change those services at all, but will be asked to provide lists of commercial users of their services. Companies using these "secondary services" will be approached for license fees.
- SWISS-PROT is a curated, added-value database, not a repository of primary information.
- SWISS-PROT's curation team adds detailed annotation and organisation to protein sequences, the overwhelming majority of which come from translations from the public nucleotide sequence databases. The value of SWISS-PROT to the academic and commercial research community is widely accepted. It is the gold standard for scientific databases and must be rendered secure.

Access to Swiss Prot:

- SRS is the easiest and simplest method available to quickly access the SWISS-PROT sequence database.
- Release 40.0 of SWISS-PROT contains 101'602 sequence entries, comprising 37'315'215 amino acids abstracted from 91'880 references. This represents an increase of 18% over release 39.
- SWISS-PROT is accompanied by TrEMBL, a computer-annotated supplement to SWISS-PROT. TrEMBL contains the translations of all coding sequences (CDS) present in the DDBJ/EMBL/GenBank Nucleotide Sequence Database and also protein sequences extracted from the literature or submitted to SWISS-PROT, which are not yet integrated into SWISS-PROT.
- No license fee will be charged to academic users, nor will any restrictions be imposed on their use or reuse of the data.
- Nothing will change in the methods by which academic or commercial users can access SWISS-PROT, but commercial users will be informed that their company is liable to pay a license fee irrespective of the method by which they access the database.
- Third party organisations providing services which make use of SWISS-PROT need not change those services at all, but will be asked to provide lists of commercial users of their services. Companies using these "secondary services" will be approached for license fees.

FTP servers offered by the SWISS-PROT group:

Link	Explanation
/pub/databases/swissprot/	SWISS-PROT release
/pub/databases/swissprot/updates/	SWISS-PROT updates
/pub/databases/sp_tr_nrdb/	SPTR_nrdb

- SWISS-PROT Home
- Information
- Access
- Submissions
- Tools
- FTP
- Group info
- Documents
- Contact

SWISS-PROT

The SWISS-PROT Protein Knowledgebase is an annotated protein sequence database established in 1986.

The SWISS-PROT Protein Knowledgebase is a curated protein sequence database that provides a high level of annotation, a minimal level of redundancy and high level of
integration with other databases.

It is maintained collaboratively by the Swiss Institute for Bioinformatics (SIB) and the European Bioinformatics Institute (EBI).

The SWISS-PROT group is headed by: Rolf Apweiler.

Access the SWISS-PROT Database

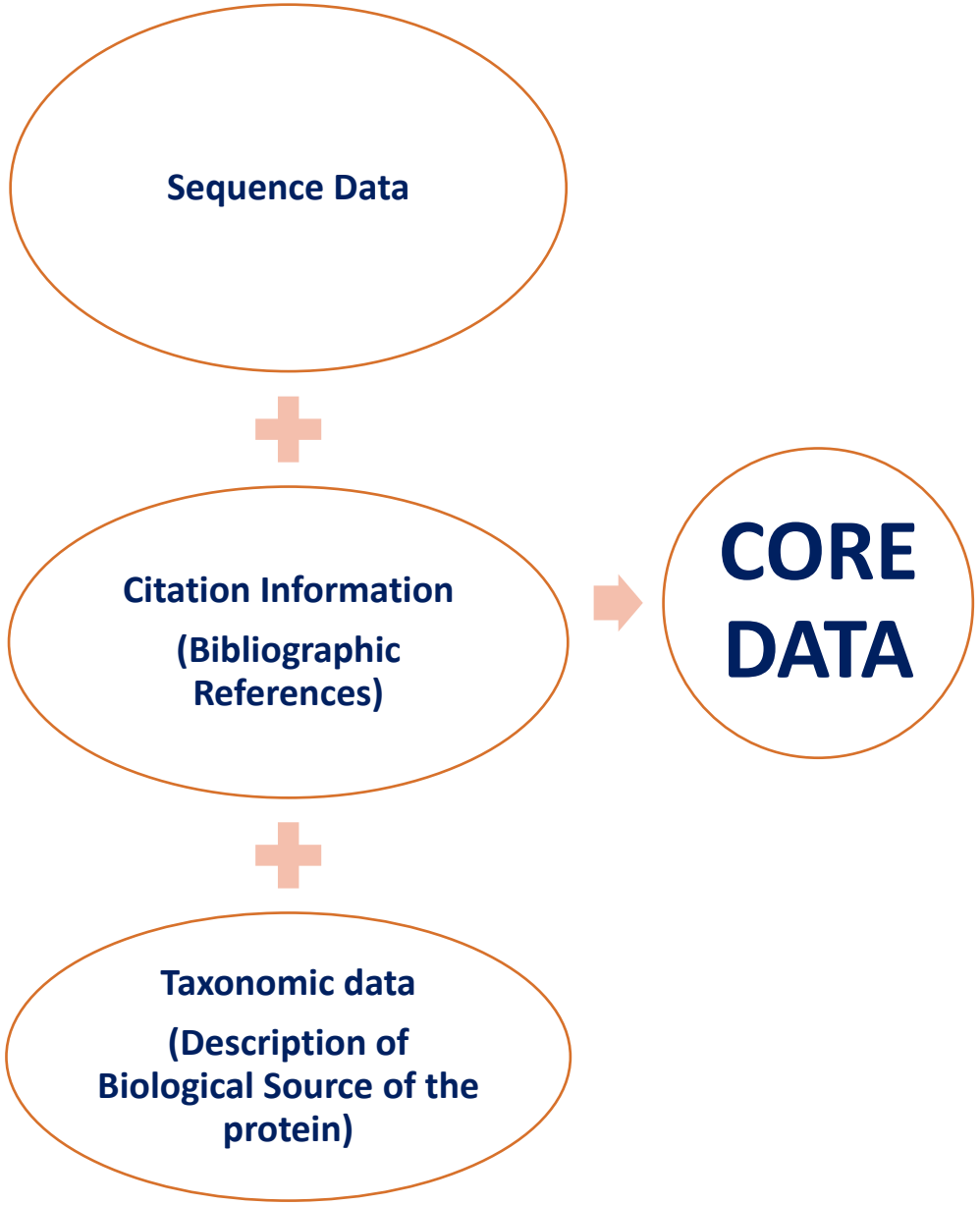
SRS is the easiest and simplest method available

TrEMBL

A protein sequence database of nucleotide translated sequences.

The TrEMBL sequence database contains the translations of all coding sequences (CDS) present in the DDBJ/EMBL/GenBank Nucleotide Sequence Database and also protein sequences extracted from the literature or submitted to SWISS-PROT,

- In SWISS-PROT, as in most other sequence databases, two classes of data can be distinguished:
  - CORE DATA
  - ANNOTATION



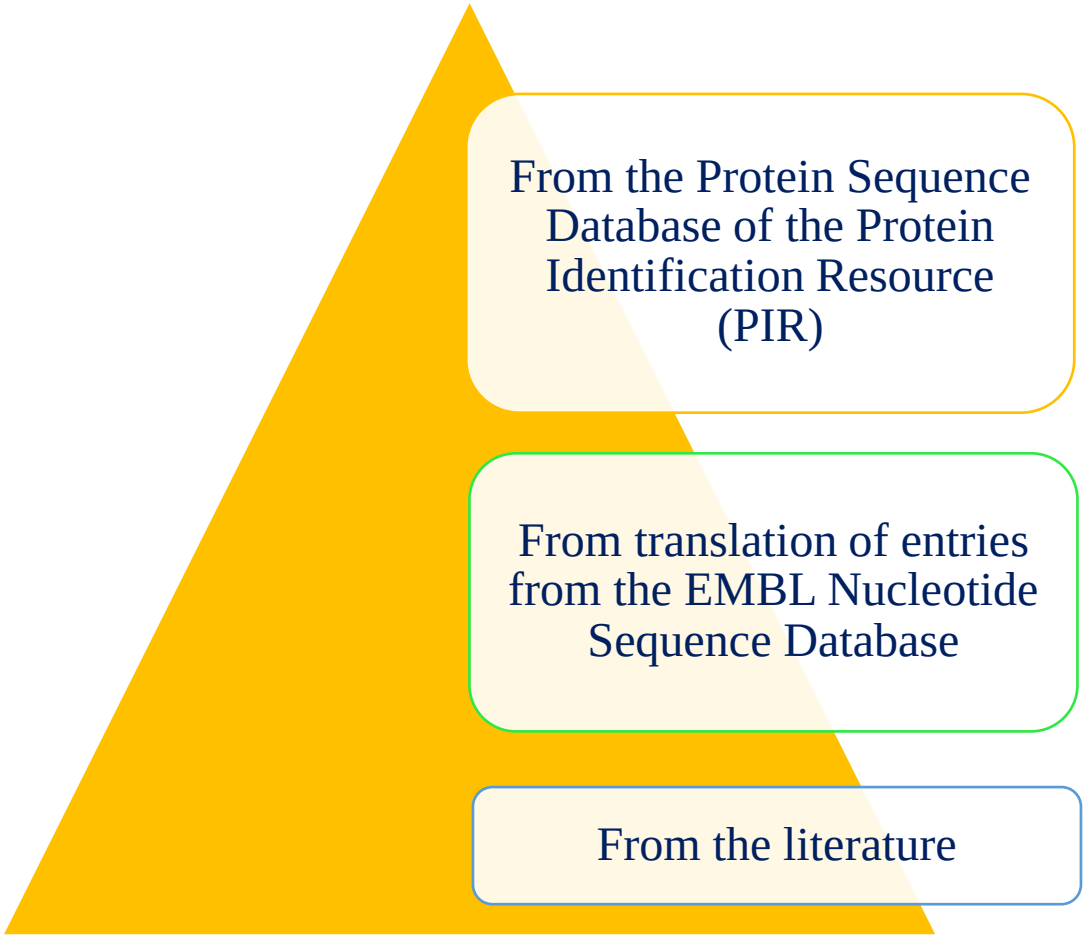
ANNOTATION information in SWISS-PROT:

- Function(s) of the protein
- Post-translational modification(s) (carbohydrates, phosphorylation, acetylation, GPI-anchor, etc.)
- Domains and sites (calcium binding regions, ATP binding sites, zinc fingers, homeobox, kringle, etc.)
- Disease(s) associated with deficiencie(s) in the protein Sequence conflicts, variants, etc.
- Secondary structure (using information from the DSSP database)
- Quaternary structure Similarities to other proteins
- Research Articles and Review Articles (who send their comments and updates concerning specific groups of proteins about which are also acknowledged and included)
- Advices and suggestions of external experts (publications that reports new sequence data, review articles to periodically update the annotations of families or groups of proteins are also used)

- Having 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.
- SWISS-PROT annotations include descriptions of the function of a protein, its domain structure, post-translational modifications, variants, reactions catalysed by this protein, similarities with other sequences, etc.
- The enzyme entries contain Enzyme Commission (EC) numbers and are cross-referenced with the ENZYME database (www.expasy.ch/sprot/enzjmie.html).
- to minimize the redundancy following data are merged:
  - A fragment of the protein sequenced at the level of the polypeptide
  - one or more reports reflecting the results of laboratories that have sequenced that protein at the cDNA level,
  - and finally reports from data provided by genomic sequencing.

are merged and, if conflicts exist between various sequencing reports, these are indicated in the feature table.

SOURCES OF THE SEQUENCE DATA:



Sequence data in SWISS-PROT originates from three different sources:

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 specialized data collections.
- So as to provide tools that will allow software developers to implement such an integrated approach SWISS-PROT has been cross-referenced with many other databases as follows:
  - EMBL Nucleotide Sequence Database.
  - PDB, the Brookhaven Protein Data Bank which stores crystallographic coordinates of proteins
  - PIR, the protein sequence database of the Protein Identification Resource.
  - EcoGene, from the EcoSeq/ EcoMap integrated *Escherichia coli* database.

Integration with other databases:

- FlyBase, the Drosophila Genetic database prepared by Michael Ashburner at the Department of Genetics, University of Cambridge.
- Gene-protein database of *Escherichia coli* K-12 (2D-gel spots).
- OMIM, the on-line version of the book 'Mendelian Inheritance in Man’.
- PROSITE, the Dictionary of Protein Sites and Patterns.
- REBASE, the restriction enzymes database.
- TFD, the transcription factors data bank.
- 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. They are implemented using a specific type of line, the 'DR' (for Data bank Reference) line.

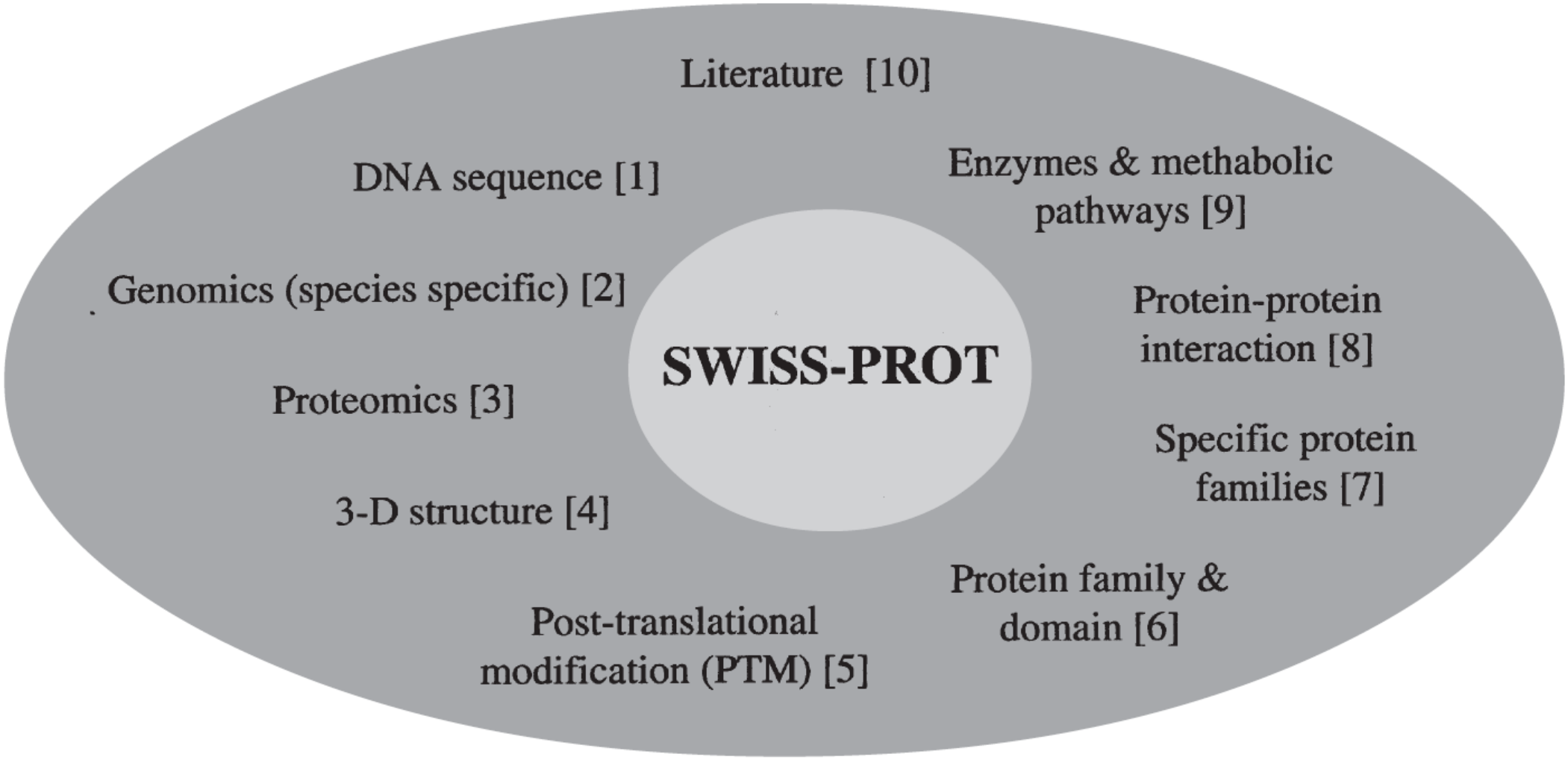


Figure: SWISS-PROT and cross-references to other databases.

DISTRIBUTION:

- SWISS-PROT is distributed on magnetic tape and on CD-ROM by the EMBL Data Library.
- The CD-ROM contains both SWISSPROT and the EMBL Nucleotide Sequence Database as well as other data collections and some database query and retrieval software for MS-DOS PC compatible computers.
- For all enquiries regarding the subscription and distribution of SWISS- PROT one should contact:

EMBL Data Library
European Molecular Biology Laboratory
Postfach 10.2209, Meyerhofstrasse 1
6900 Heidelberg, Germany
Telephone: (+49 6221) 387 258
Telefax : (+49 6221) 387 519 or 387 306
Electronic network address: datalib@EMBL-heidelberg.de

For information, comments and/or suggestions, please use any of the following contact details:

By E-mail

swissprot@ebi.ac.uk - (for general information)

datasubs@ebi.ac.uk - (for data submissions)

- Individual sequence entries can be obtained from the EMBL File Server.
- Detailed instructions on how to make best use of this service, and in particular on how to obtain protein sequences, can be obtained by sending to the network address netserv@EMBL-heidelberg.de the following message:
  - HELP
  - HELP PROT
- It can also be obtained using FTP (File Transfer Protocol), from the following file servers:
  - GenBank On-line Service; Internet address: genbank.bio.net (134.172.1.160)
  - NCBI (National Library of Medicine, NIH, Washington D.C., U.S.A.); Internet address: ncbi.nlm.nih.gov (130.14.20.1)
  - ExPASy (Expert Protein Analysis System server, University of Geneva, Switzerland), Internet address: expasy.hcuge.ch (129.195.254.61)
- The present distribution frequency is four releases per year.
- No restrictions are placed on use or redistribution of the data.

FORMAT:

The SWISS-PROT contains the information about the name and origin of the protein, protein attributes, general information, ontologies, sequence annotation, amino acid sequence, bibliographic references, cross-references with sequence, structure and interaction databases, and entry information.

The SWISS-PROT protein sequence data bank is composed of sequence entries. Each sequence entry is composed of lines.

Different types of lines, each with their own format, are used to record the various data which make up the entry.

For standardization purposes the format of SWISS-PROT follows as closely as possible that of the EMBL Nucleotide Sequence Database.

A SWISS-PROT entry is composed of different line types, and each line is introduced by a two-letter code indicating the type of data following on that line.

```
ID   CAH2_HUMAN      STANDARD;      PRT;   259 AA.
AC   P00918;
DT   21-JUL-1986 (REL. 01, CREATED)
DT   21-JUL-1986 (REL. 01, LAST SEQUENCE UPDATE)
DT   01-MAR-1992 (REL. 21, LAST ANNOTATION UPDATE)
DE   CARBONIC ANHYDRASE II (EC 4.2.1.1) (CARBONATE DEHYDRATASE II).
GN   CA2.
OS   HOMO SAPIENS (HUMAN).
OC   EUKARYOTA; METAZOA; CHORDATA; VERTEBRATA; TETRAPODA; MAMMALIA;
OC   EUTHERIA; PRIMATES.
RN   [1]
RP   SEQUENCE FROM N.A.
RM   87231043
RA   MONTGOMERY J.C., VENTA P.J., TASHIAN R.E., HEWETT-EMMETT D.;
RL   NUCLEIC ACIDS RES. 15:4687-4687(1987).
RN   [2]
RP   SEQUENCE FROM N.A.
RM   88085190
RA   MURAKAMI H., MARELICH G.P., GRUBB J.H., KYLE J.W., SLY W.S.;
RL   GENOMICS 1:159-166(1987).
RN   [3]
RP   SEQUENCE.
RM   77006079
RA   HENDERSON L.E., HENRIKSSON D., NYMAN P.O.;
RL   J. BIOL. CHEM. 251:5457-5463(1976).
RN   [4]
RP   SEQUENCE.
RM   74143468
RA   LIN K.-T.D., DEUTSCH H.F.;
RL   J. BIOL. CHEM. 249:2329-2337(1974).
RN   [5]
RP   SEQUENCE OF 1-76 FROM N.A.
RM   86077780
RA   VENTA P.J., MONTGOMERY J.C., HEWETT-EMMETT D., TASHIAN R.E.;
RL   BIOCHIM. BIOPHYS. ACTA 826:195-201(1985).
RN   [6]
RP   X-RAY CRYSTALLOGRAPHY, 2.0 ANGSTROMS.
RM   72111787
RA   LILJAS A., KANNAN K.K., BERGSTEN P.-C., WAARA I., FRIDBORG K.,
RA   STRANDBERG B., CARLBOM U., JARUP L., LOVGREN S., PETEF M.;
RL   NATURE NEW BIOL. 235:131-137(1972).
RN   [7]
RP   X-RAY CRYSTALLOGRAPHY, 2.0 ANGSTROMS.
RM   89315726
RA   ERIKSSON A.E., JONES T.A., LILJAS A.;
RL   PROTEINS 4:274-282(1988).
RN   [8]
RP   X-RAY CRYSTALLOGRAPHY, 2.0 ANGSTROMS.
RM   89315727
RA   ERIKSSON A.E., KYLSTEN P.M., JONES T.A., LILJAS A.;
RL   PROTEINS 4:283-293(1988).
RN   [9]
RP   VARIANT JOGJAKARTA.
RM   83100296
RA   JONES G.L., SOFRO A.S.M., SHAW D.C.;
RL   BIOCHEM. GENET. 20:979-1000(1982).
RN   [10]
RP   VARIANT MELBOURNE.
RM   83236368
RA   JONES G.L., SHAW D.C.;
RL   HUM. GENET. 63:392-399(1983).
```

- The first section of every SWISS-PROT entry contains:
- ID Line: the entry name
- AC Line: a unique primary accession number (AC), sometimes followed by several secondary accession numbers
- DT Line: Indicates dates when the entry was created and when its sequence and annotations were last updated
- DE Line: The description line (DE) lists all names, including synonyms, under which the protein has been known
- GN Line: This contains the name(s) of the gene(s) coding for it.
- The following section contains taxonomic data about the organism from which the protein originates, in particular the:
  - OS Line: organism name
  - OC Line: its classification in the taxonomic tree
  - OX Line: and a unique taxonomy identifier
- RN, RP, RX, RA, RT and RL Lines: depicts the reference section that contains all literature references consulted for the annotation of the protein.
- The list of references includes not only publications of the sequence itself, but also articles detailing post-translational modifications, 3-D structure, polymorphisms etc.

```
RL   MOL. GENET. 63:372-377(1983).
CC   -!- FUNCTION: REVERSIBLE HYDRATATION OF CARBON DIOXIDE.
CC   -!- CATALYTIC ACTIVITY: H(2)CO(3) = CO(2) + H(2)O.
CC   -!- THERE ARE AT LEAST 6 ENZYMATIC FORMS OF CARBONIC ANHYDRASE: CA-I
CC       (OR B), CA-II (OR C), CA-III (OR M), CA-IV, CA-V AND CA-VI.
CC   -!- DISEASE: DEFECTS IN CA2 ARE THE CAUSE OF OSTEOPETROSIS WITH RENAL
CC       TUBULAR ACIDOSIS (MARBLE BRAIN DISEASE).
DR   EMBL; Y00339; HSCA2.
DR   EMBL; X03251; HSCAII.
DR   EMBL; J03037; HSCAIIA.
DR   PIR; A01141; CRHU2.
DR   PIR; A23202; A23202.
DR   PIR; A27175; A27175.
DR   PDB; 1CA2; 15-JAN-90.
DR   PDB; 2CA2; 15-APR-90.
DR   PDB; 3CA2; 15-APR-90.
DR   MIM; 259730; NINTH EDITION.
DR   PROSITE; PS00162; CARBONIC_ANHYDRASE.
KW   LYASE; ACETYLATION; ZINC; 3D-STRUCTURE.
FT   INIT_MET      0      0
FT   MOD_RES       1      1       ACETYLATION.
FT   ACT_SITE     63     63
FT   ACT_SITE     66     66
FT   METAL        93     93       ZINC (CATALYTIC).
FT   METAL        95     95       ZINC (CATALYTIC).
FT   METAL       118    118       ZINC (CATALYTIC).
FT   ACT_SITE    126    126
FT   ACT_SITE    196    198
FT   VARIANT      17     17       K -> E (JOGJAKARTA).
FT   VARIANT     235    235       P -> H (MELBOURNE).
```

- CC Lines: The reference section is followed by the comment block (CC) containing textual information classified into different “topics” and describing the protein’s function, subcellular localisation, post-translational modifications, association with diseases etc.
- DR Lines: Database cross-references are stored in the DR lines and allow the user to access related information in other databases.
- KW Lines: The keyword section (KW line type) lists a number of terms from a controlled vocabulary, which can be used to retrieve subsets of the database.
- FT Lines: A very important part of a SWISS-PROT protein entry is the feature table (FT lines), which contains information about interesting sites or domains within the protein sequence, for which positional information is known.
- The feature table describes events such as post-translational modifications, sequence variants due to polymorphisms, domain structure, sequence conflicts, etc.
- Each feature line consists of a feature key, start and end positions of the described feature in the precursor sequence, and the feature description.

```
FT   VARIANT     251    251        N -> D.
FT   TURN          8     10
FT   HELIX        12     18
FT   HELIX        20     23
FT   TURN         25     26
FT   STRAND       31     32
FT   TURN         34     36
FT   STRAND       38     39
FT   TURN         41     42
FT   STRAND       46     49
FT   TURN         51     52
FT   STRAND       55     60
FT   STRAND       65     69
FT   STRAND       77     80
FT   TURN         81     82
FT   STRAND       87     96
FT   TURN        100    101
FT   STRAND      107    108
FT   TURN        109    110
FT   STRAND      111    111
FT   STRAND      115    123
FT   HELIX       124    126
FT   HELIX       129    132
FT   TURN        133    134
FT   TURN        136    137
FT   STRAND      139    148
FT   HELIX       153    165
FT   TURN        168    169
FT   STRAND      171    173
FT   HELIX       179    182
FT   STRAND      189    194
FT   TURN        199    200
FT   STRAND      205    210
FT   STRAND      214    216
FT   HELIX       218    224
FT   TURN        225    226
FT   STRAND      228    228
FT   TURN        232    233
FT   STRAND      238    238
FT   TURN        250    251
FT   STRAND      255    256
SQ   SEQUENCE   259 AA;  29115 MW;  365693 CN;
     SHHWGYGKHN GPEHWHKDFP IAKGERQSPV DIDTHTAKYD PSLKPLSVSY DQATSLRILN
     NGHAFNVEFD DSQDKAVLKG GPLDGTYRLI QFHFHWGSLD GQGSEHTVDK KKYAAELHLV
     HWNTKYGDFG KAVQQPDGLA VLGIFLKVGS AKPGLQKVVD VLDSIKTKGK SADFTNFDPR
     GLLPESLDYW TYPGSLTTPP LLECVTWIVL KEPISVSSEQ VLKFRKLNFN GEGEPEELMV
     DNWRPAQPLK NRQIKASFK
//
```

Figure 1. A sample entry from SWISS-PROT

- Finally, there is the amino acid sequence itself.
- The SWISS-PROT database was the first biomolecular database to include cross-references in its entries – long before the advent of the World Wide Web, which made navigation between data resources distributed all over the planet become second nature to all its users.
- There are five different types of cross-references available in SWISS-PROT:
  - explicit and implicit cross-references in the DR lines,
  - URL addresses under the comment (CC) topic
  - “DATABASE”, and cross-references departing from certain key types in the feature table (FT).
- Finally, the Medline/ PubMed identifiers of literature references are stored in RX (Reference Cross(X)reference) lines and thus allow direct access to these literature databases.
- There are a number of other annotation items in SWISS-PROT that might also be termed cross references and that are, in the World Wide Web version, enhanced with
  - active hypertext links, namely scientific journal references (RL lines)
  - taxonomy identifier (OX lines)
  - or enzyme classification numbers (DE lines)
- In addition to cross-references provided by SWISSPROT itself, SWISS-PROT also plays an important role for federated 2D-PAGE databases, which achieve much of the integration of data located and maintained at different sites through SWISS-PROT as their main index.

Thank You