UniProt Knowledgebase
Swiss-Prot Protein Knowledgebase
TrEMBL Protein Database

User Manual
Release 2018_02 of 28-Feb-2018

Table of contents

Contact details

You can send us general questions and suggestions, updates or corrections to UniProt, submit new protein sequence data, or e-mail us directly.

Notice

Citation

If you want to cite UniProtKB in a publication, please use one of the references listed here.

Table of contents

1. What is the UniProt Knowledgebase?

Table of contents

Until 2002, the EBI/SIB Swiss-Prot + TrEMBL databases and the PIR Protein Sequence Database (PIR-PSD) coexisted as protein databases with differing protein sequence coverage and annotation priorities. In 2002, EBI, SIB, and PIR (at the Georgetown University Medical Center and National

Biomedical Research Foundation) joined forces as the UniProt consortium. The primary mission of the consortium is to support biological research by maintaining a high quality database that serves as a stable, comprehensive, fully classified, richly and accurately annotated protein sequence knowledgebase, with extensive cross-references and querying interfaces freely accessible to the scientific community.

The UniProt Knowledgebase (UniProtKB) provides the central database of protein sequences with accurate, consistent, rich sequence and functional annotation.

The UniProt Knowledgebase consists of two sections: Swiss-Prot - a section containing manually-annotated records with information extracted from literature and curator-evaluated computational analysis, and TrEMBL - a section with computationally analyzed records that await full manual annotation.

1.1. The Swiss-Prot Protein Knowledgebase

Table of contents

Swiss-Prot is an annotated protein sequence database. It was established in 1986 and maintained collaboratively, since 1987, by the group of Amos Bairoch first at the Department of Medical Biochemistry of the University of Geneva and now at the SIB Swiss Institute of Bioinformatics and the EMBL Data Library (now the EMBL Outstation - The European Bioinformatics Institute (EBI)). The Swiss-Prot Protein Knowledgebase consists of sequence entries. Sequence entries are composed of different line types, each with their own format. For standardization purposes the format of Swiss-Prot follows as closely as possible that of the EMBL Nucleotide Sequence Database.

Swiss-Prot distinguishes itself from protein sequence databases by four distinct criteria:

a) Annotation

In Swiss-Prot, as in many sequence databases, 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);
- The taxonomic data (description of the biological source of the protein).

The annotation consists of the description of the following items:

- Function(s) of the protein;
- Posttranslational modification(s) such as carbohydrates, phosphorylation, acetylation and GPI-anchor;
- Domains and sites, for example, calcium-binding regions, ATP-binding sites, zinc fingers, homeoboxes, SH2 and SH3 domains and kringle;
- Secondary structure, e.g. alpha helix, beta sheet;
- Quaternary structure, i.g. homodimer, heterotrimer, etc.;
- Similarities to other proteins;
- Disease(s) associated with any number of deficiencies 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 that report 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.

We believe that 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.

b) 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 minimize the redundancy of the database. If conflicts exist between various sequencing reports, they are indicated in the feature table of the corresponding entry.

c) 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. Swiss-Prot is currently cross-referenced to more than 100 different databases. 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. This extensive network of cross-references allows Swiss-Prot to play a major role as a focal point of biomolecular database interconnectivity.

d) Documentation

Swiss-Prot is distributed with a large number of index files and specialized documentation files. Some of these files have been available for a long time (this user manual, the release notes, the various indices for authors, citations, keywords, etc.), but many have been created recently and we are continuously adding new files. 'Documentation files' section contains an up-to-date descriptive list of all distributed document files.

1.2. The computer-annotated supplement TrEMBL

Table of contents

TrEMBL is the computer-annotated section of the UniProt Knowledgebase. It contains translations of all coding regions in the DDBJ/EMBL/GenBank nucleotide databases, and protein sequences extracted from the literature or submitted to UniProtKB, which are not yet integrated into Swiss-Prot. TrEMBL allows these sequences to be made publicly available quickly without diluting the high quality annotation found in Swiss-Prot.

The information in a TrEMBL entry is initially derived directly from the underlying DDBJ/EMBL/GenBank nucleotide entry and the quality of data is directly dependent on the information provided by the submitter of the nucleotide entry. This information may be enhanced later by automatic annotation procedures (see below) but if not, it remains as provided by the submitter until the entry is manually annotated and added to Swiss-Prot.

After creation of a TrEMBL entry, a number of steps are taken to improve the data quality for users:

a) Automatic annotation

Records waiting in TrEMBL for full manual annotation are enhanced by automatic annotation. Information is transferred from well-characterised entries in Swiss-Prot to unannotated entries in TrEMBL which belong to groups defined by InterPro, a database of protein families, domains and functional sites. This process 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.

b) Redundancy removal

Sequences from the same organism which are full-length and which have 100% identity are merged into a single entry to reduce redundancy.

2. Conventions used in the database

Table of contents

The following sections describe the general conventions used in the knowledgebase to achieve uniformity of presentation. Experienced users of the EMBL Database can skip these sections and directly refer to this document, which lists the minor differences in format between the two data collections.

2.1. General structure of the database

The UniProt Knowledgebase is composed of sequence entries. Each entry corresponds to a single contiguous sequence as contributed to the bank or reported in the literature. In some cases, entries have been assembled from several papers that report overlapping sequence regions. Conversely, a single paper can provide data for several entries, e.g. when related sequences from different organisms are reported.

References to positions within a sequence are made using sequential numbering, beginning with 1 at the N-terminal end of the sequence.

The sequence data correspond to the precursor form of a protein before posttranslational modifications and processing.

2.2. Status

To distinguish the fully annotated entries in the Swiss-Prot section of the UniProt Knowledgebase from the computer-annotated entries in the TrEMBL section, the 'status' of each entry is indicated in the first (ID) line of each entry. The two defined classes are:

Reviewed Entries that have been manually reviewed and annotated by UniProtKB curators (Swiss-Prot section of the UniProt Knowledgebase).

Unreviewed Computer-annotated entries that have not been reviewed by UniProtKB curators (TrEMBL section of the UniProt Knowledgebase).

2.3. Structure of a sequence entry

The entries in the UniProt Knowledgebase are structured so as to be usable by human readers as well as by computer programs. The explanations, descriptions, classifications and other comments are in ordinary English. Wherever possible, symbols familiar to biochemists, protein chemists and molecular biologists are used.

Each sequence entry is composed of lines. Different types of lines, each with their own format, are used to record the various data that make up the entry. A sample sequence entry is shown below.

```
ID   GRAA_HUMAN              Reviewed;         262 AA.
AC   P12544; A4PHN1; Q6IB36;
DT   01-OCT-1989, integrated into UniProtKB/Swiss-Prot.
DT   11-JAN-2011, sequence version 2.
DT   07-JUN-2017, entry version 186.
DE   RecName: Full=Granzyme A;
DE            EC=3.4.21.78;
DE   AltName: Full=CTL tryptase;
DE   AltName: Full=Cytotoxic T-lymphocyte proteinase 1;
DE   AltName: Full=Fragmentin-1;
DE   AltName: Full=Granzyme-1;
DE   AltName: Full=Hanukkah factor;
DE            Short=H factor;
DE            Short=HF;
DE   Flags: Precursor;
GN   Name=GZMA; Synonyms=CTLA3, HFSP;
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
OC   Catarrhini; Hominidae; Homo.
```

```
OX   NCBI_TaxID=9606;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [MRNA] (ISOFORM ALPHA), AND VARIANT THR-121.
RC   TISSUE=T-cell;
RX   PubMed=3257574; DOI=10.1073/pnas.85.4.1184;
RA   Gershenfeld H.K., Hershberger R.J., Shows T.B., Weissman I.L.;
RT   "Cloning and chromosomal assignment of a human cDNA encoding a T cell-
RT   and natural killer cell-specific trypsin-like serine protease.";
RL   Proc. Natl. Acad. Sci. U.S.A. 85:1184-1188(1988).
RN   [2]
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE MRNA] (ISOFORM ALPHA), AND VARIANT
RP   THR-121.
RA   Ebert L., Schick M., Neubert P., Schatten R., Henze S., Korn B.;
RT   "Cloning of human full open reading frames in Gateway(TM) system entry
RT   vector (pDONR201).";
RL   Submitted (JUN-2004) to the EMBL/GenBank/DDBJ databases.
RN   [3]
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].
RX   PubMed=15372022; DOI=10.1038/nature02919;
RA   Schmutz J., Martin J., Terry A., Couronne O., Grimwood J., Lowry S.,
RA   Gordon L.A., Scott D., Xie G., Huang W., Hellsten U., Tran-Gyamfi M.,
RA   She X., Prabhakar S., Aerts A., Altherr M., Bajorek E., Black S.,
RA   Branscomb E., Caoile C., Challacombe J.F., Chan Y.M., Denys M.,
RA   Detter J.C., Escobar J., Flowers D., Fotopulos D., Glavina T.,
RA   Gomez M., Gonzales E., Goodstein D., Grigoriev I., Groza M.,
RA   Hammon N., Hawkins T., Haydu L., Israni S., Jett J., Kadner K.,
RA   Kimball H., Kobayashi A., Lopez F., Lou Y., Martinez D., Medina C.,
RA   Morgan J., Nandkeshwar R., Noonan J.P., Pitluck S., Pollard M.,
RA   Predki P., Priest J., Ramirez L., Retterer J., Rodriguez A.,
RA   Rogers S., Salamov A., Salazar A., Thayer N., Tice H., Tsai M.,
RA   Ustaszewska A., Vo N., Wheeler J., Wu K., Yang J., Dickson M.,
RA   Cheng J.-F., Eichler E.E., Olsen A., Pennacchio L.A., Rokhsar D.S.,
RA   Richardson P., Lucas S.M., Myers R.M., Rubin E.M.;
RT   "The DNA sequence and comparative analysis of human chromosome 5.";
RL   Nature 431:268-274(2004).
RN   [4]
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE MRNA] (ISOFORM ALPHA), AND VARIANT
RP   THR-121.
RC   TISSUE=Blood;
RX   PubMed=15489334; DOI=10.1101/gr.2596504;
RG   The MGC Project Team;
RT   "The status, quality, and expansion of the NIH full-length cDNA
RT   project: the Mammalian Gene Collection (MGC).";
RL   Genome Res. 14:2121-2127(2004).
RN   [5]
RP   NUCLEOTIDE SEQUENCE [GENOMIC DNA] OF 1-72, NUCLEOTIDE SEQUENCE [MRNA]
RP   OF 1-34 (ISOFORM BETA), ALTERNATIVE PROMOTER USAGE, AND INDUCTION.
RX   PubMed=17180578; DOI=10.1007/s10038-006-0099-9;
RA   Ruike Y., Katsuma S., Hirasawa A., Tsujimoto G.;
```

```
RT   "Glucocorticoid-induced alternative promoter usage for a novel 5'
RT   variant of granzyme A.";
RL   J. Hum. Genet. 52:172-178(2007).
RN   [6]
RP   NUCLEOTIDE SEQUENCE [GENOMIC DNA] OF 1-23.
RA   Goralski T.J., Krensky A.M.;
RT   "The upstream region of the human granzyme A locus contains both
RT   positive and negative transcriptional regulatory elements.";
RL   Submitted (NOV-1995) to the EMBL/GenBank/DDBJ databases.
RN   [7]
RP   PROTEIN SEQUENCE OF 29-53.
RX   PubMed=3047119;
RA   Poe M., Bennett C.D., Biddison W.E., Blake J.T., Norton G.P.,
RA   Rodkey J.A., Sigal N.H., Turner R.V., Wu J.K., Zweerink H.J.;
RT   "Human cytotoxic lymphocyte tryptase. Its purification from granules
RT   and the characterization of inhibitor and substrate specificity.";
RL   J. Biol. Chem. 263:13215-13222(1988).
RN   [8]
RP   PROTEIN SEQUENCE OF 29-40, AND CHARACTERIZATION.
RX   PubMed=3262682;
RA   Hameed A., Lowrey D.M., Lichtenheld M., Podack E.R.;
RT   "Characterization of three serine esterases isolated from human IL-2
RT   activated killer cells.";
RL   J. Immunol. 141:3142-3147(1988).
RN   [9]
RP   PROTEIN SEQUENCE OF 29-39, AND CHARACTERIZATION.
RX   PubMed=3263427;
RA   Kraehenbuhl O., Rey C., Jenne D.E., Lanzavecchia A., Groscurth P.,
RA   Carrel S., Tschopp J.;
RT   "Characterization of granzymes A and B isolated from granules of
RT   cloned human cytotoxic T lymphocytes.";
RL   J. Immunol. 141:3471-3477(1988).
RN   [10]
RP   FUNCTION AS SET PROTEASE.
RX   PubMed=11555662; DOI=10.1074/jbc.M108137200;
RA   Beresford P.J., Zhang D., Oh D.Y., Fan Z., Greer E.L., Russo M.L.,
RA   Jaju M., Lieberman J.;
RT   "Granzyme A activates an endoplasmic reticulum-associated caspase-
RT   independent nuclease to induce single-stranded DNA nicks.";
RL   J. Biol. Chem. 276:43285-43293(2001).
RN   [11]
RP   FUNCTION AS SET PROTEASE.
RX   PubMed=12628186; DOI=10.1016/S0092-8674(03)00150-8;
RA   Fan Z., Beresford P.J., Oh D.Y., Zhang D., Lieberman J.;
RT   "Tumor suppressor NM23-H1 is a granzyme A-activated DNase during CTL-
RT   mediated apoptosis, and the nucleosome assembly protein SET is its
RT   inhibitor.";
RL   Cell 112:659-672(2003).
RN   [12]
```

```
RP   FUNCTION, AND INTERACTION WITH APEX1.
RX   PubMed=12524539; DOI=10.1038/ni885;
RA   Fan Z., Beresford P.J., Zhang D., Xu Z., Novina C.D., Yoshida A.,
RA   Pommier Y., Lieberman J.;
RT   "Cleaving the oxidative repair protein Ape1 enhances cell death
RT   mediated by granzyme A.";
RL   Nat. Immunol. 4:145-153(2003).
RN   [13]
RP   FUNCTION AS SET PROTEASE.
RX   PubMed=16818237; DOI=10.1016/j.molcel.2006.06.005;
RA   Chowdhury D., Beresford P.J., Zhu P., Zhang D., Sung J.S., Demple B.,
RA   Perrino F.W., Lieberman J.;
RT   "The exonuclease TREX1 is in the SET complex and acts in concert with
RT   NM23-H1 to degrade DNA during granzyme A-mediated cell death.";
RL   Mol. Cell 23:133-142(2006).
RN   [14]
RP   GLYCOSYLATION [LARGE SCALE ANALYSIS] AT ASN-170.
RC   TISSUE=Liver;
RX   PubMed=19159218; DOI=10.1021/pr8008012;
RA   Chen R., Jiang X., Sun D., Han G., Wang F., Ye M., Wang L., Zou H.;
RT   "Glycoproteomics analysis of human liver tissue by combination of
RT   multiple enzyme digestion and hydrazide chemistry.";
RL   J. Proteome Res. 8:651-661(2009).
RN   [15]
RP   3D-STRUCTURE MODELING OF 29-262.
RX   PubMed=3237717; DOI=10.1002/prot.340040306;
RA   Murphy M.E.P., Moult J., Bleackley R.C., Gershenfeld H.,
RA   Weissman I.L., James M.N.G.;
RT   "Comparative molecular model building of two serine proteinases from
RT   cytotoxic T lymphocytes.";
RL   Proteins 4:190-204(1988).
RN   [16]
RP   X-RAY CRYSTALLOGRAPHY (2.4 ANGSTROMS) OF 29-262 IN COMPLEX WITH A
RP   TRIPEPTIDE CMK INHIBITOR.
RX   PubMed=12819769; DOI=10.1038/nsb944;
RA   Bell J.K., Goetz D.H., Mahrus S., Harris J.L., Fletterick R.J.,
RA   Craik C.S.;
RT   "The oligomeric structure of human granzyme A is a determinant of its
RT   extended substrate specificity.";
RL   Nat. Struct. Biol. 10:527-534(2003).
RN   [17]
RP   X-RAY CRYSTALLOGRAPHY (2.5 ANGSTROMS) OF 29-262 IN COMPLEX WITH
RP   SUBSTRATE.
RX   PubMed=12819770; DOI=10.1038/nsb945;
RA   Hink-Schauer C., Estebanez-Perpina E., Kurschus F.C., Bode W.,
RA   Jenne D.E.;
RT   "Crystal structure of the apoptosis-inducing human granzyme A dimer.";
RL   Nat. Struct. Biol. 10:535-540(2003).
CC   -!- FUNCTION: Abundant protease in the cytosolic granules of cytotoxic
CC       T-cells and NK-cells which activates caspase-independent cell
CC       death with morphological features of apoptosis when delivered into
CC       the target cell through the immunological synapse. It cleaves
CC       after Lys or Arg. Cleaves APEX1 after 'Lys-31' and destroys its
CC       oxidative repair activity. Cleaves the nucleosome assembly protein
CC       SET after 'Lys-189', which disrupts its nucleosome assembly
CC       activity and allows the SET complex to translocate into the
CC       nucleus to nick and degrade the DNA. {ECO:0000269|PubMed:11555662,
CC       ECO:0000269|PubMed:12524539, ECO:0000269|PubMed:12628186,
CC       ECO:0000269|PubMed:16818237}.
CC   -!- CATALYTIC ACTIVITY: Hydrolysis of proteins, including fibronectin,
CC       type IV collagen and nucleolin. Preferential cleavage: -Arg-|-
CC       Xaa-, -Lys-|-Xaa- >> -Phe-|-Xaa- in small molecule substrates.
CC   -!- SUBUNIT: Homodimer; disulfide-linked. Interacts with APEX1.
CC       {ECO:0000269|PubMed:12524539, ECO:0000269|PubMed:12819769,
CC       ECO:0000269|PubMed:12819770}.
CC   -!- SUBCELLULAR LOCATION: Isoform alpha: Secreted. Cytoplasmic
CC       granule.
CC   -!- ALTERNATIVE PRODUCTS:
CC       Event=Alternative promoter usage; Named isoforms=2;
CC       Name=alpha;
CC         IsoId=P12544-1; Sequence=Displayed;
CC       Name=beta;
CC         IsoId=P12544-2; Sequence=VSP_038571, VSP_038572;
CC   -!- INDUCTION: Dexamethasone (DEX) induces expression of isoform beta
CC       and represses expression of isoform alpha. The alteration in
CC       expression is mediated by binding of glucocorticoid receptor to
CC       independent promoters adjacent to the alternative first exons of
CC       isoform alpha and isoform beta. {ECO:0000269|PubMed:17180578}.
CC   -!- SIMILARITY: Belongs to the peptidase S1 family. Granzyme
CC       subfamily. {ECO:0000255|PROSITE-ProRule:PRU00274}.
CC   -!- CAUTION: Exons 1a and 1b of the sequence reported in
CC       PubMed:17180578 are of human origin, however exon 2 shows strong
CC       similarity to the rat sequence. {ECO:0000305}.
CC   -!- WEB RESOURCE: Name=Atlas of Genetics and Cytogenetics in Oncology
CC       and Haematology;
CC       URL="http://atlasgeneticsoncology.org/Genes/GZMAID51130ch5q11.html";
DR   EMBL; M18737; AAA52647.1; -; mRNA.
DR   EMBL; CR456968; CAG33249.1; -; mRNA.
DR   EMBL; AC091977; -; NOT_ANNOTATED_CDS; Genomic_DNA.
DR   EMBL; BC015739; AAH15739.1; -; mRNA.
DR   EMBL; AB284134; BAF56159.1; -; mRNA.
DR   EMBL; U40006; AAD00009.1; -; Genomic_DNA.
DR   CCDS; CCDS3965.1; -. [P12544-1]
DR   PIR; A31372; A31372.
DR   RefSeq; NP_006135.1; NM_006144.3.
DR   UniGene; Hs.90708; -.
DR   PDB; 1HF1; Model; -; A=29-262.
DR   PDB; 1OP8; X-ray; 2.50 A; A/B/C/D/E/F=29-262.
DR   PDB; 1ORF; X-ray; 2.40 A; A=29-262.
DR   PDBsum; 1HF1; -.
DR   PDBsum; 1OP8; -.
DR   PDBsum; 1ORF; -.
DR   ProteinModelPortal; P12544; -.
DR   SMR; P12544; -.
DR   BioGrid; 109256; 17.
DR   IntAct; P12544; 4.
DR   STRING; 9606.ENSP00000274306; -.
DR   ChEMBL; CHEMBL4307; -.
DR   MEROPS; S01.135; -.
DR   iPTMnet; P12544; -.
DR   PhosphoSitePlus; P12544; -.
DR   BioMuta; GZMA; -.
DR   DMDM; 317373360; -.
DR   MaxQB; P12544; -.
DR   PaxDb; P12544; -.
DR   PeptideAtlas; P12544; -.
DR   PRIDE; P12544; -.
DR   DNASU; 3001; -.
DR   Ensembl; ENST00000274306; ENSP00000274306; ENSG00000145649. [P12544-1]
DR   GeneID; 3001; -.
DR   KEGG; hsa:3001; -.
DR   UCSC; uc003jpm.4; human. [P12544-1]
DR   CTD; 3001; -.
DR   DisGeNET; 3001; -.
DR   GeneCards; GZMA; -.
DR   HGNC; HGNC:4708; GZMA.
DR   HPA; HPA054134; -.
DR   MIM; 140050; gene.
DR   neXtProt; NX_P12544; -.
DR   OpenTargets; ENSG00000145649; -.
DR   PharmGKB; PA29086; -.
DR   eggNOG; KOG3627; Eukaryota.
DR   eggNOG; COG5640; LUCA.
DR   GeneTree; ENSGT00760000118895; -.
DR   HOGENOM; HOG000251820; -.
DR   HOVERGEN; HBG013304; -.
DR   InParanoid; P12544; -.
DR   KO; K01352; -.
DR   OMA; KEFPYPC; -.
DR   OrthoDB; EOG091G0AH5; -.
DR   PhylomeDB; P12544; -.
DR   TreeFam; TF333630; -.
DR   BRENDA; 3.4.21.78; 2681.
DR   SIGNOR; P12544; -.
DR   EvolutionaryTrace; P12544; -.
DR   GeneWiki; GZMA; -.
DR   GenomeRNAi; 3001; -.
DR   PRO; PR:P12544; -.
DR   Proteomes; UP000005640; Chromosome 5.
DR   Bgee; ENSG00000145649; -.
DR   CleanEx; HS_GZMA; -.
DR   Genevisible; P12544; HS.
DR   GO; GO:0005576; C:extracellular region; IEA:UniProtKB-SubCell.
DR   GO; GO:0001772; C:immunological synapse; TAS:UniProtKB.
DR   GO; GO:0005634; C:nucleus; TAS:UniProtKB.
DR   GO; GO:0042803; F:protein homodimerization activity; IDA:UniProtKB.
DR   GO; GO:0004252; F:serine-type endopeptidase activity; IDA:UniProtKB.
DR   GO; GO:0006915; P:apoptotic process; TAS:UniProtKB.
DR   GO; GO:0019835; P:cytolysis; IEA:UniProtKB-KW.
DR   GO; GO:0006955; P:immune response; TAS:UniProtKB.
DR   GO; GO:0043392; P:negative regulation of DNA binding; IDA:UniProtKB.
DR   GO; GO:0032078; P:negative regulation of endodeoxyribonuclease activity; IDA:UniProtKB.
DR   GO; GO:0051354; P:negative regulation of oxidoreductase activity; IDA:UniProtKB.
DR   GO; GO:0043065; P:positive regulation of apoptotic process; IDA:UniProtKB.
DR   GO; GO:0051603; P:proteolysis involved in cellular protein catabolic process; IDA:UniProtKB.
DR   CDD; cd00190; Tryp_SPc; 1.
DR   InterPro; IPR009003; Peptidase_S1_PA.
DR   InterPro; IPR001314; Peptidase_S1A.
DR   InterPro; IPR001254; Trypsin_dom.
DR   InterPro; IPR018114; TRYPSIN_HIS.
DR   InterPro; IPR033116; TRYPSIN_SER.
DR   Pfam; PF00089; Trypsin; 1.
DR   PRINTS; PR00722; CHYMOTRYPSIN.
DR   SMART; SM00020; Tryp_SPc; 1.
DR   SUPFAM; SSF50494; SSF50494; 1.
DR   PROSITE; PS50240; TRYPSIN_DOM; 1.
DR   PROSITE; PS00134; TRYPSIN_HIS; 1.
DR   PROSITE; PS00135; TRYPSIN_SER; 1.
PE   1: Evidence at protein level;
KW   3D-structure; Alternative promoter usage; Apoptosis;
KW   Complete proteome; Cytolysis; Direct protein sequencing;
KW   Disulfide bond; Glycoprotein; Hydrolase; Polymorphism; Protease;
KW   Reference proteome; Secreted; Serine protease; Signal; Zymogen.
FT   SIGNAL        1     26       In isoform alpha.
FT   PROPEP       27     28       Activation peptide (in isoform alpha).
FT                                {ECO:0000269|PubMed:3047119,
FT                                ECO:0000269|PubMed:3262682,
FT                                ECO:0000269|PubMed:3263427}.
FT                                /FTId=PRO_0000027393.
FT   CHAIN        29    262       Granzyme A.
FT                                /FTId=PRO_0000027394.
FT   DOMAIN       29    259       Peptidase S1. {ECO:0000255|PROSITE-
FT                                ProRule:PRU00274}.
FT   ACT_SITE     69     69       Charge relay system.
FT   ACT_SITE    114    114       Charge relay system.
FT   ACT_SITE    212    212       Charge relay system.
FT   CARBOHYD    170    170       N-linked (GlcNAc...) asparagine.
FT                                {ECO:0000269|PubMed:19159218}.
FT   DISULFID     54     70
FT   DISULFID    148    218
FT   DISULFID    179    197
FT   DISULFID    208    234
FT   VAR_SEQ       1     17       Missing (in isoform beta).
FT                                {ECO:0000303|PubMed:17180578}.
FT                                /FTId=VSP_038571.
FT   VAR_SEQ      18     23       LLLIPE -> MTKGLR (in isoform beta).
FT                                {ECO:0000303|PubMed:17180578}.
FT                                /FTId=VSP_038572.
FT   VARIANT     121    121       M -> T (in dbSNP:rs3104233).
FT                                {ECO:0000269|PubMed:15489334,
FT                                ECO:0000269|PubMed:3257574,
FT                                ECO:0000269|Ref.2}.
FT                                /FTId=VAR_024291.
FT   CONFLICT     33     34       NE -> DT (in Ref. 5; no nucleotide
FT                                entry). {ECO:0000305}.
FT   CONFLICT     36     36       T -> V (in Ref. 5; no nucleotide entry).
FT                                {ECO:0000305}.
FT   CONFLICT     47     47       S -> K (in Ref. 5; no nucleotide entry).
FT                                {ECO:0000305}.
FT   CONFLICT     49     52       DRKT -> KPDS (in Ref. 5; no nucleotide
FT                                entry). {ECO:0000305}.
FT   CONFLICT     62     62       D -> N (in Ref. 5; no nucleotide entry).
FT                                {ECO:0000305}.
FT   CONFLICT     71     72       NL -> IP (in Ref. 5; no nucleotide
FT                                entry). {ECO:0000305}.
FT   STRAND       43     47       {ECO:0000244|PDB:1ORF}.
FT   STRAND       49     51       {ECO:0000244|PDB:1ORF}.
FT   STRAND       53     60       {ECO:0000244|PDB:1ORF}.
FT   STRAND       63     66       {ECO:0000244|PDB:1ORF}.
FT   STRAND       77     81       {ECO:0000244|PDB:1ORF}.
FT   STRAND       83     87       {ECO:0000244|PDB:1ORF}.
FT   STRAND       93     95       {ECO:0000244|PDB:1ORF}.
FT   STRAND       97    102       {ECO:0000244|PDB:1ORF}.
FT   TURN        108    110       {ECO:0000244|PDB:1ORF}.
FT   STRAND      116    122       {ECO:0000244|PDB:1ORF}.
FT   STRAND      127    130       {ECO:0000244|PDB:1ORF}.
FT   STRAND      147    154       {ECO:0000244|PDB:1ORF}.
FT   STRAND      156    160       {ECO:0000244|PDB:1ORF}.
FT   STRAND      167    174       {ECO:0000244|PDB:1ORF}.
FT   HELIX       176    179       {ECO:0000244|PDB:1ORF}.
FT   TURN        182    189       {ECO:0000244|PDB:1ORF}.
FT   STRAND      195    199       {ECO:0000244|PDB:1ORF}.
FT   STRAND      215    218       {ECO:0000244|PDB:1ORF}.
FT   STRAND      221    228       {ECO:0000244|PDB:1ORF}.
FT   STRAND      241    245       {ECO:0000244|PDB:1ORF}.
FT   HELIX       248    258       {ECO:0000244|PDB:1ORF}.
SQ   SEQUENCE   262 AA;  28999 MW;  FD773628BA6F301B CRC64;
     MRNSYRFLAS SLSVVVSLLL IPEDVCEKII GGNEVTPHSR PYMVLLSLDR KTICAGALIA
     KDWVLTAAHC NLNKRSQVIL GAHSITREEP TKQIMLVKKE FPYPCYDPAT REGDLKLLQL
     MEKAKINKYV TILHLPKKGD DVKPGTMCQV AGWGRTHNSA SWSDTLREVN ITIIDRKVCN
     DRNHYNFNPV IGMNMVCAGS LRGGRDSCNG DSGSPLLCEG VFRGVTSFGL ENKCGDPRGP
     GVYILLSKKH LNWIIMTIKG AV
//
```

Entries from the TrEMBL section follow the same format. For format differences see the description of the distinct line types.

Each line begins with a two-character line code, which indicates the type of data contained in the line. The current line types and line codes and the order in which they appear in an entry, are shown in the table below.

Line code	Content	Occurrence in an entry
ID	Identification	Once; starts the entry
AC	Accession number(s)	Once or more
DT	Date	Three times
DE	Description	Once or more
GN	Gene name(s)	Optional
OS	Organism species	Once or more
OG	Organelle	Optional
OC	Organism classification	Once or more
OX	Taxonomy cross-reference	Once
OH	Organism host	Optional
RN	Reference number	Once or more
RP	Reference position	Once or more
RC	Reference comment(s)	Optional
RX	Reference cross-reference(s)	Optional
RG	Reference group	Once or more (Optional if RA line)
RA	Reference authors	Once or more (Optional if RG line)
RT	Reference title	Optional
RL	Reference location	Once or more
CC	Comments or notes	Optional
DR	Database cross-references	Optional
PE	Protein existence	Once
KW	Keywords	Optional
FT	Feature table data	Once or more in Swiss-Prot, optional in TrEMBL
SQ	Sequence header	Once
(blanks)	Sequence data	Once or more
//	Termination line	Once; ends the entry

As shown in the above table, some line types are found in all entries, others are optional. Some line types occur many times in a single entry. Each entry must begin with an identification line (ID) and end with a terminator line (//).

A detailed description of each line type is given in the next section of this document. It must be noted that, with the exception of GN, all line types exist in the EMBL Database. A description of the format differences between the UniProt Knowledgebase and EMBL databases is given in this document.

The two-character line-type code that begins each line is always followed by three blanks, so that the actual information begins with the sixth character. In general, information is not extended beyond character position 75, there are however a few exceptions where lines may be longer (e.g. OH lines, CC lines that contain the 'WEB RESOURCE' topic (see section 3.22), etc.).

2.4. Evidence attributions

The evidence for annotations are available in UniProtKB entries. An individual evidence description consists of a mandatory evidence type, represented by a code from the Evidence Codes Ontology (ECO) and, where applicable, the source of the data which is usually another database record that is represented by the database name and record identifier, but in the case of publications that are not in PubMed we indicate instead the corresponding UniProtKB reference number.

Examples:

a) An evidence type without source

```
{ECO:0000305}
{ECO:0000250}
```

```
{ECO:0000255}
```

b) An evidence type with source

```
{ECO:0000269|PubMed:10433554}

{ECO:0000303|Ref.6}

{ECO:0000305|PubMed:16683188}

{ECO:0000250|UniProtKB:Q8WUF5}

{ECO:0000312|EMBL:BAG16761.1}

{ECO:0000313|EMBL:BAG16761.1}

{ECO:0000255|HAMAP-Rule:MF_00205}

{ECO:0000256|HAMAP-Rule:MF_00205}

{ECO:0000244|PDB:1K83}

{ECO:0000213|PDB:1K83}
```

c) Several evidences

```
{ECO:0000269|PubMed:10433554, ECO:0000303|Ref.6}
```

3. The different line types

Table of contents

3.1. The ID line

Table of contents

The ID (IDentification) line is always the first line of an entry. The general form of the ID line is:

```
ID   EntryName Status; SequenceLength.
```

3.1.1. Entry name

The first item on the ID line is the entry name of the sequence. This name is a useful means of identifying a sequence, but it is not a stable identifier as is the accession number (see 3.2).

a) Swiss-Prot entry names

The Swiss-Prot entry name consists of up to **11** uppercase alphanumeric characters. Swiss-Prot uses a general purpose naming convention that can be symbolized as **X_Y**, where:

- **X** is a mnemonic code of at most 5 alphanumeric characters representing the protein name. Examples: B2MG is for Beta-2-microglobulin, HBA is for Hemoglobin alpha chain and INS is for Insulin, CAD17 for Cadherin-17;
- The '**_**' sign serves as a separator;
- **Y** is a mnemonic species identification code of at most 5 alphanumeric characters representing the biological source of the protein. This code is generally made of the first three letters of the genus and the first two letters of the species.

Examples:

PSEPU is for *Pseudomonas putida* and NAJNI is for *Naja nivea*.

However, for species most commonly encountered in the database, self-explanatory codes are used. There are 16 of those codes: BOVIN for Bovine, CHICK for Chicken, ECOLI for *Escherichia coli,* HORSE for Horse, HUMAN for Human, MAIZE for Maize (*Zea mays*), MOUSE for Mouse, PEA for Garden pea (*Pisum sativum*), PIG for Pig, RABIT for Rabbit, RAT for Rat, SHEEP for Sheep, SOYBN for Soybean (*Glycine max*), TOBAC for Common tobacco (*Nicotina tabacum*), WHEAT for Wheat (*Triticum aestivum*), and YEAST for Baker's yeast (*Saccharomyces cerevisiae*).

As it was not possible to apply the above rules to viruses, they were given arbitrary, but generally easy-to-remember identification codes.

Examples of complete protein sequence entry names are: RL1_ECOLI for ribosomal protein L1 from *Escherichia coli,*, AFTIN_HUMAN for Aftiphilin from *human*, SODC_DROME for Superoxide dismutase [Cu-Zn] from *Drosophila melanogaster*.

The names of all the presently-defined species identification codes are listed in the document file speclist.txt.

b) TrEMBL entry names

The TrEMBL entry name consists of up to **16** uppercase alphanumeric characters. TrEMBL uses a general purpose naming convention similar to that of Swiss-Prot, where:

- **X** is identical to the **accession number** of the entry
- The '**_**' sign serves as a separator;
- **Y** is a mnemonic species identification code.

As it is not possible in a reasonable timeframe to manually assign organism codes to all species represented in TrEMBL, "virtual" codes have been defined that regroup organisms at a certain taxonomic level. Such codes are prefixed by the number "9" and generally correspond to a "pool" of organisms, which can be 'wide' as a kingdom. Here are some examples of such codes:

```
9BACT B      2: N=Bacteria

9CNID E   6073: N=Cnidaria

9FUNG E   4751: N=Fungi

9REOV V  10880: N=Reoviridae

9TETR E  32523: N=Tetrapoda

9VIRI E  33090: N=Viridiplantae
```

These type of "virtual" codes are also listed in the document file speclist.txt.

Examples of complete TrEMBL entry names are O95417_HUMAN, Q9VVG0_DROME, P71025_BACSU or Q9SR52_ARATH.

3.1.2. Status

The second item on the ID line indicates the status of the entry (see section 2.2).

3.1.3. Length of the molecule

The third and last item of the ID line is the length of the molecule, which is the total number of amino acids in the sequence. This number includes the positions reported to be present but which have not been determined (coded as 'X'). The length is followed by the letter code 'AA' (Amino Acids).

3.1.4. Examples of identification lines

Two examples of Swiss-Prot ID lines are shown below:

```
ID   CYC_BOVIN               Reviewed;         104 AA.


ID   GIA2_GIALA              Reviewed;         296 AA.
```

Example of a TrEMBL ID line:

```
ID   Q5JU06_HUMAN            Unreviewed;       268 AA.
```

3.2. The AC line

Table of contents

The AC (ACcession number) line lists the accession number(s) associated with an entry. The format of the AC line is:

```
AC   AC_number_1;[ AC_number_2;]...[ AC_number_N;]
```

An example of an accession number line is shown below:

```
AC   P00321;
```

Semicolons separate the accession numbers and a semicolon terminates the list. If necessary, more than one AC line can be used. Example:

```
AC   Q16653; O00713; O00714; O00715; Q13054; Q13055; Q14855; Q92891;

AC   Q92892; Q92893; Q92894; Q92895; Q93053; Q96KU9; Q96KV0; Q96KV1;

AC   Q99605;
```

The purpose of accession numbers is to provide a stable way of identifying entries from release to release. It is sometimes necessary for reasons of consistency to change the names of the entries, for example, to ensure that related entries have similar names. However, an accession number is always conserved, and therefore allows unambiguous citation of entries.

Researchers who wish to cite entries in their publications should **always cite the first accession number**. This is commonly referred to as the 'primary accession number'. 'Secondary accession numbers' are sorted alphanumerically.

We strongly advise those users who have programs performing mappings of Swiss-Prot to another data resource to use Swiss-Prot accession numbers to identify an entry.

Entries will have more than one accession number if they have been merged or split. For example, when two entries are merged into one, the accession numbers from both entries are stored in the AC line(s).

If an existing entry is split into two or more entries (a rare occurrence), the original accession numbers are retained in all the derived entries and a new primary accession number is added to all the entries.

An accession number is dropped only when the data to which it was assigned have been completely removed from the database. Accession numbers deleted from Swiss-Prot are listed in the document file delac_sp.txt and those deleted from TrEMBL are listed in delac_tr.txt.

UniProtKB accession numbers consist of 6 or 10 alphanumerical characters in the format:

1	2	3	4	5	6	7	8	9	10
[A-N,R-Z]	[0-9]	[A-Z]	[A-Z, 0-9]	[A-Z, 0-9]	[0-9]				
[O,P,Q]	[0-9]	[A-Z, 0-9]	[A-Z, 0-9]	[A-Z, 0-9]	[0-9]				
[A-N,R-Z]	[0-9]	[A-Z]	[A-Z, 0-9]	[A-Z, 0-9]	[0-9]	[A-Z]	[A-Z,0-9]	[A-Z,0-9]	[0-9]

The three patterns can be combined into the following regular expression:

```
[OPQ][0-9][A-Z0-9]{3}[0-9]|[A-NR-Z][0-9]([A-Z][A-Z0-9]{2}[0-9]){1,2}
```

Here are some examples of valid accession numbers: P12345, Q1AAA9, O456A1, P4A123 and A0A022YWF9.

3.3. The DT line

Table of contents

The DT (DaTe) lines show the date of creation and last modification of the database entry.

The format of the DT line in Swiss-Prot is:

```
DT   DD-MMM-YYYY, integrated into UniProtKB/database_name.
DT   DD-MMM-YYYY, sequence version x.
DT   DD-MMM-YYYY, entry version x.
```

Where 'DD' is the day, 'MMM' the month and 'YYYY' the year, respectively. The dates shown in DT lines correspond to the date of the biweekly release at which an entry was integrated or updated. There are always three DT lines in each entry, each of them is associated with a specific comment:

- The first DT line indicates when the entry first appeared in the database. The associated comment, 'integrated into UniProtKB/database_name', indicates in which section of UniProtKB, Swiss-Prot or TrEMBL, the entry can be found;
- The second DT line indicates when the sequence data was last modified. The associated comment, 'sequence version', indicates the sequence version number. The sequence version number of an entry is incremented by one when the amino acid sequence shown in the sequence record is modified;
- The third DT line indicates when data other than the sequence was last modified. The associated comment, 'entry version', indicates the entry version number. The entry version number is incremented by one whenever any data in the flat file representation of the entry is modified.

Example of a block of Swiss-Prot DT lines:

```
DT   01-OCT-1996, integrated into UniProtKB/Swiss-Prot.
DT   01-OCT-1996, sequence version 1.
DT   07-FEB-2006, entry version 49.
```

Example of a block of TrEMBL DT lines:

```
DT   01-FEB-1999, integrated into UniProtKB/TrEMBL.
DT   15-OCT-2000, sequence version 2.
DT   15-DEC-2004, entry version 5.
```

Whenever the sequence of an entry is updated there is always also an annotation update. The date in the third DT line is thus always at least as recent as the one in the second DT line.

Note that sequence and entry versions are not reset when an entry moves from Swiss-Prot to TrEMBL. The date of integration into Swiss-Prot can be more recent than the last sequence update.

```
DT   25-OCT-2005, integrated into UniProtKB/Swiss-Prot.
DT   01-NOV-1996, sequence version 1.
DT   07-FEB-2006, entry version 35.
```

A comprehensive archive of UniProtKB/Swiss-Prot and UniProtKB/TrEMBL entry versions is available: the UniProtKB Sequence/Annotation Version Database (UniSave) is a repository of UniProtKB/Swiss-Prot and UniProtKB/TrEMBL entry versions. Unlike the UniProt Knowledgebase, which contains only the latest Swiss-Prot and TrEMBL entry and sequence versions, the UniProtKB Sequence/Annotation Version Database provides access to all versions of these entries. This allows to track sequence changes, to find out when a given annotation appeared in an entry and how it evolved.

3.4. The DE line

Table of contents

The DE (DEscription) lines contain general descriptive information about the sequence stored. This information is generally sufficient to identify the protein precisely.

a) The DE line in Swiss-Prot

The description always starts with the recommended name (RecName) of the protein. Alternative names (AltName) are indicated thereafter.

The DE line contains 3 categories, as well as several subcategories, of protein names:

Category Field	Subcategory Field	Cardinality	Description
RecName:		1 in UniProtKB/Swiss-Prot 0-1 in UniProtKB/TrEMBL	The name recommended by the UniProt consortium.
	Full=	1	The full name.
	Short=	0-n	An abbreviation of the full name or an acronym.
	EC=	0-n	An Enzyme Commission number.
AltName:		0-n	A synonym of the recommended name.
	Full=	0-1	The full name.
	Short=	0-n	An abbreviation of the full name or an acronym.
	EC=	0-n	An Enzyme Commission number.
AltName:	**Allergen=**	0-1	See allergen.txt.
AltName:	**Biotech=**	0-1	A name used in a biotechnological context.
AltName:	**CD_antigen=**	0-n	See cdlist.txt.
AltName:	**INN=**	0-n	The international nonproprietary name: A generic name for a pharmaceutical substance or active pharmaceutical ingredient that is globally recognized and is a public property.
SubName:		0 in UniProtKB/Swiss-Prot 0-n in UniProtKB/TrEMBL	A name provided by the submitter of the underlying nucleotide sequence.
	Full=	1	The full name.
	EC=	0-n	An Enzyme Commission number.

Each name is shown on a separate line; lines may therefore exceed 75 characters.

Protein naming guidelines are described in the document file nameprot.txt.

```
DE   RecName: Full=Annexin A5;

DE            Short=Annexin-5;

DE   AltName: Full=Annexin V;

DE   AltName: Full=Lipocortin V;

DE   AltName: Full=Endonexin II;

DE   AltName: Full=Calphobindin I;

DE   AltName: Full=CBP-I;

DE   AltName: Full=Placental anticoagulant protein I;

DE            Short=PAP-I;

DE   AltName: Full=PP4;

DE   AltName: Full=Thromboplastin inhibitor;

DE   AltName: Full=Vascular anticoagulant-alpha;

DE            Short=VAC-alpha;

DE   AltName: Full=Anchorin CII;
```

```
DE   RecName: Full=Granulocyte colony-stimulating factor;

DE            Short=G-CSF;

DE   AltName: Full=Pluripoietin;

DE   AltName: Full=Filgrastim;

DE   AltName: Full=Lenograstim;

DE   Flags: Precursor;
```

A block of DE lines may further contain multiple **Includes:** and/or **Contains:** sections and a separate field **Flags:** to indicate whether the protein sequence is a precursor or a fragment:

Field	Cardinality	Value
Includes:	0-n	A block of protein names as described in the table above.
Contains:	0-n	A block of protein names as described in the table above.

Flags: 0-1 Precursor and/or Fragment or Fragments

If a protein is known to be cleaved into multiple functional components, the description starts with the name of the precursor protein, followed by 'Contains:' section(s). Each individual component is described in a separate 'Contains:' section Alternative names (AltName) are allowed for each individual component. Example:

```
DE   RecName: Full=Corticotropin-lipotropin;
DE   AltName: Full=Pro-opiomelanocortin;
DE            Short=POMC;
DE   Contains:
DE     RecName: Full=NPP;
DE   Contains:
DE     RecName: Full=Melanotropin gamma;
DE     AltName: Full=Gamma-MSH;
DE   Contains:
DE     RecName: Full=Potential peptide;
DE   Contains:
DE     RecName: Full=Corticotropin;
DE     AltName: Full=Adrenocorticotropic hormone;
DE              Short=ACTH;
DE   Contains:
DE     RecName: Full=Melanotropin alpha;
DE     AltName: Full=Alpha-MSH;
DE   Contains:
DE     RecName: Full=Corticotropin-like intermediary peptide;
DE              Short=CLIP;
DE   Contains:
DE     RecName: Full=Lipotropin beta;
DE     AltName: Full=Beta-LPH;
DE   Contains:
DE     RecName: Full=Lipotropin gamma;
DE     AltName: Full=Gamma-LPH;
DE   Contains:
DE     RecName: Full=Melanotropin beta;
DE     AltName: Full=Beta-MSH;
DE   Contains:
DE     RecName: Full=Beta-endorphin;
DE   Contains:
DE     RecName: Full=Met-enkephalin;
DE   Flags: Precursor;
```

If a protein is known to include multiple functional domains each of which is described by a different name, the description starts with the name of the overall protein, followed by 'Includes:' section(s). All the domains are listed in a separate 'Includes:' section. Alternative names (AltName) are allowed for each individual domain. Example:

```
DE   RecName: Full=CAD protein;
DE   Includes:
DE     RecName: Full=Glutamine-dependent carbamoyl-phosphate synthase;
DE              EC=6.3.5.5;
DE   Includes:
DE     RecName: Full=Aspartate carbamoyltransferase;
DE              EC=2.1.3.2;
DE   Includes:
```

```
DE     RecName: Full=Dihydroorotase;
DE              EC=3.5.2.3;
```

In rare cases, the functional domains of an enzyme are cleaved, but the catalytic activity can only be observed, when the individual chains reorganize in a complex. Such proteins are described in the DE line by a combination of both 'Includes:' and 'Contains:', in the order given in the following example:

```
DE   RecName: Full=Arginine biosynthesis bifunctional protein argJ;
DE   Includes:
DE     RecName: Full=Glutamate N-acetyltransferase;
DE              EC=2.3.1.35;
DE     AltName: Full=Ornithine acetyltransferase;
DE              Short=OATase;
DE     AltName: Full=Ornithine transacetylase;
DE   Includes:
DE     RecName: Full=Amino-acid acetyltransferase;
DE              EC=2.3.1.1;
DE     AltName: Full=N-acetylglutamate synthase;
DE              Short=AGS;
DE   Contains:
DE     RecName: Full=Arginine biosynthesis bifunctional protein argJ alpha chain;
DE   Contains:
DE     RecName: Full=Arginine biosynthesis bifunctional protein argJ beta chain;
```

When the mature form of a protein is derived by processing of a precursor, we indicate this fact using the Flag 'Precursor'; in such cases the sequence displayed does not correspond to the mature form of the protein.

```
DE   RecName: Full=Chondroitin proteoglycan 3;
DE   Flags: Precursor;
```

If the complete sequence is not determined, we indicate it in the 'Flags' section with 'Fragment' or 'Fragments'. Example:

```
DE   RecName: Full=Dihydrodipicolinate reductase;
DE            Short=DHPR;
DE            EC=1.3.1.26;
DE   Flags: Fragment;
```

b) The DE line in TrEMBL

The format of the DE line in TrEMBL follows closely the format used in Swiss-Prot. However, as TrEMBL is not manually annotated, the description is derived directly from the underlying nucleotide entry and its accuracy relies on the information provided by the submitter of the nucleotide entry. It is why TrEMBL entries usually have submitted name (SubName) instead of a recommended name (RecName). The description may later be improved by automatic annotation procedures (see section Automatic annotation) but if not, it remains as provided by the submitter until the entry is manually annotated and added to Swiss-Prot.

3.5. The GN line

Table of contents

The GN (Gene Name) line indicates the name(s) of the gene(s) that code for the stored protein sequence. The GN line contains three types of information:

1. **Gene names** (a.k.a gene symbols). The name(s) used to represent a gene. As there can be more than one name assigned to a gene, we make a distinction between the one which we believe should be used as the official gene name and the other names which are listed as "Synonyms".
2. **Ordered locus names** (a.k.a. OLN, ORF numbers, CDS numbers or Gene numbers). A name used to represent an ORF in a completely sequenced genome or chromosome. It is generally based on a prefix representing the organism and a number which usually represents the sequential ordering of genes on the chromosome. Depending on the genome sequencing center, numbers are only attributed to protein-coding genes, or also to pseudogenes, or also to tRNAs and other features. If two predicted genes have been merged to form a new gene, both gene identifiers are indicated, separated by a slash (see last example). Examples: HI0934, Rv3245c, At5g34500, YER456W, YAR042W/YAR044W.
3. **ORF names** (a.k.a. sequencing names or contig names or temporary ORFNames). A name temporarily attributed by a sequencing project to an open reading frame. This name is generally based on a cosmid numbering system. Examples: MtCY277.28c, SYGP-ORF50, SpBC2F12.04, C06E1.1, CG10954.

The format of the GN line is:

```
GN   Name=<name>; Synonyms=<name1>[, <name2>...]; OrderedLocusNames=<name1>[, <name2>...];
GN   ORFNames=<name1>[, <name2>...];
```

None of the above four tokens are mandatory. But a **"Synonyms"** token can only be present if there is a **"Name"** token.

If there is more than one gene, GN line blocks for the different genes are separated by the following line:

```
GN   and


Example:


GN   Name=Jon99Cii; Synonyms=SER1, SER5, Ser99Da; ORFNames=CG7877;
GN   and
GN   Name=Jon99Ciii; Synonyms=SER2, SER5, Ser99Db; ORFNames=CG15519;
```

Wrapping is done preferentially at a semicolon, otherwise at a comma.

It often occurs that more than one name has been assigned to an individual locus, in which case all the synonyms will be listed alphabetically and case-insensitively. Example:

```
GN   Name=hns; Synonyms=bglY, cur, drdX, hnsA, msyA, osmZ, pilG, topS;
GN   OrderedLocusNames=b1237, c1701, z2013, ECs1739;
```

3.6. The OS line

Table of contents

The OS (Organism Species) line specifies the organism which was the source of the stored sequence. In the rare case where all the species information will not fit on a single line, more than one OS line is used. The last OS line is terminated by a period.

The species designation consists, in most cases, of the Latin genus and species designation followed by the English name (in parentheses). For viruses, only the common English name is given.

Examples of OS lines are shown here:

```
OS   Escherichia coli.


OS   Homo sapiens (Human).


OS   Solanum melongena (Eggplant) (Aubergine).


OS   Rous sarcoma virus (strain Schmidt-Ruppin A) (RSV-SRA) (Avian leukosis
OS   virus-RSA).
```

The names (official name, common name, synonym) concerning one species are cut across lines when they do not fit into a single line:

```
OS   Epizootic hemorrhagic disease virus 2 (strain Alberta) (EHDV-2).
```

3.7. The OG line

Table of contents

The OG (OrGanelle) line indicates if the gene coding for a protein originates from mitochondria, a plastid, a nucleomorph or a plasmid.

The format of the OG line is:

```
OG   Hydrogenosome.
OG   Mitochondrion.
OG   Nucleomorph.
OG   Plasmid name.
OG   Plastid.
OG   Plastid; Apicoplast.
OG   Plastid; Chloroplast.
OG   Plastid; Organellar chromatophore.
OG   Plastid; Cyanelle.
OG   Plastid; Non-photosynthetic plastid.
```

Where 'name' is the name of the plasmid.

Hydrogenosomes are membrane-enclosed redox organelles found in some anaerobic unicellular eukaryotes which contain hydrogenase and produce hydrogen and ATP by glycolysis. They are thought to have evolved from mitochondria; most hydrogenosomes lack a genome, but some like (e.g. the anaerobic ciliate Nyctotherus ovalis) have retained a rudimentary genome.

Mitochondria are redox-active membrane-bound organelles found in the cytoplasm of most eukaryotic cells. They are the site of sthe reactions of oxidative phosphorylation, which results in the formation of ATP.

Nucleomorphs are reduced vestigal nuclei found in the plastids of cryptomonad and chlorachniophyte algae. The plastids originate from engulfed eukaryotic phototrophs.

Plastids are classified based on either their taxonomic lineage or in some cases on their photosynthetic capacity.

Apicoplasts are the plastids found in Apicocomplexa parasites such as Eimeria, Plasmodium and Toxoplasma; they are not photosynthetic.

Chloroplasts are the plastids found in all land plants and algae with the exception of the glaucocystophyte algae (see below). Chloroplasts in green tissue are photosynthetic; in other tissues they may not be photosynthetic and then may also have secondary information relating to subcellular location (e.g. amyloplasts, chromoplasts).

Organellar chromatophores are the photosynthetic plastids found in Paulinella chromatophora, a photosynthetic thecate amoeba of the Cercozoa lineage. At 1 Mb this plastid is 3 times larger than the largest known plastid; it is not clear if it is derived from the same endosymbiotic event that is thought to have led to all other plastids.

Cyanelles are the plastids found in the glaucocystophyte algae. They are also photosynthetic but their plastid has a vestigial cell wall between the 2 envelope membranes.

Non-photosynthetic plastid is used when the plastid in question derives from a photosynthetic lineage but the plastid in question is missing essential genes. Some examples are Aneura mirabilis, Epifagus virginiana, Helicosporidium (a liverwort, higher plant and green alga respectively).

The term Plastid is used when the capacities of the organism are unclear; for example in the parasitic plants of the Cuscuta lineage, where sometimes young tissue is photosynthetic.

If an entry reports the sequence of a protein identical in a number of plasmids, the names of these plasmids will all be listed in the OG lines of that entry. The plasmid names are separated by commas, the last plasmid name is preceded by the word 'and'. Plasmid names are never written across two lines. Example:

```
OG   Plasmid R6-5, Plasmid IncFII R100 (NR1), and
OG   Plasmid IncFII R1-19 (R1 drd-19).
```

The document plasmid.txt lists all the plasmid names that are used in the database in the context of the OG line.

3.8. The OC line

Table of contents

The OC (Organism Classification) lines contain the taxonomic classification of the source organism. The taxonomic classification used is that maintained at the NCBI (see https://www.ncbi.nlm.nih.gov/Taxonomy/) and used by the nucleotide sequence databases (EMBL/GenBank/DDBJ). The NCBI's taxonomy reflects current phylogenetic knowledge. It is a sequence-based taxonomy as much as possible and based on published authorities wherever possible. Because of the inherent ambiguity of evolutionary classification and the specific needs of database users (e.g. trying to track down the phylogenetic history of a group of organisms or to elucidate the evolution of a molecule), this taxonomy strives to accurately reflect current phylogenetic knowledge. The NCBI's taxonomy is intended to be informative and helpful; no claim is made that it is the best or the most exact.

The classification is listed top-down as nodes in a taxonomic tree in which the most general grouping is given first. The classification may be distributed over several OC lines, but nodes are not split or hyphenated between lines. Semicolons separate the individual items and the list is terminated by a period.

The format of the OC line is:

```
OC   Node[; Node...].
```

For example the classification lines for a human sequence would be:

```
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Euarchontoglires; Primates; Catarrhini; Hominidae;
OC   Homo.
```

3.9. The OX line

Table of contents

The OX (Organism taxonomy cross-reference) line is used to indicate the identifier of a specific organism in a taxonomic database. The format of the OX line is:

```
OX   Taxonomy_database_Qualifier=Taxonomic code;
```

Currently the cross-references are made to the taxonomy database of NCBI, which is associated with the qualifier 'TaxID' and a taxonomic code.

Examples:

```
OX   NCBI_TaxID=9606;
```

```
OX   NCBI_TaxID=562;
```

3.10. The OH line

Table of contents

The OH (Organism Host) line is optional and appears only in viral entries. It indicates the host organism(s) that are susceptible to be infected by a virus.

A virus being an inert particle outside its hosts, the virion has neither metabolism, nor any replication capability, nor autonomous evolution. Identifying the host organism(s) is therefore essential, because features like virus-cell interactions and posttranslational modifications depend mostly on the host.

The format of the OH line is:

```
OH   NCBI_TaxID=TaxID; HostName.
```

The *HostName* consists of the official name and, optionally, a common name and/or synonym. The length of an OH line may exceed 75 characters.

Example for Simian hepatitis A virus:

```
OH   NCBI_TaxID=9481; Callithrix.
OH   NCBI_TaxID=9536; Cercopithecus hamlyni (Owl-faced monkey) (Hamlyn's monkey).
OH   NCBI_TaxID=9539; Macaca (macaques).
OH   NCBI_TaxID=9598; Pan troglodytes (Chimpanzee).
```

3.11. The reference (RN, RP, RC, RX, RG, RA, RT, RL) lines

Table of contents

These lines comprise the literature citations. The citations indicate the sources from which the data has been abstracted. The reference lines for a given citation occur in a block, and are always in the order RN, RP, RC, RX, RG, RA, RT and RL. Within each such reference block, the RN line occurs once, the RC, RX and RT lines occur zero or more times, and the RP, RG/RA and RL lines occur one or more times. If several references are given, there will be a reference block for each.

An example of a complete reference is:

```
RN   [1]
RP   NUCLEOTIDE SEQUENCE [MRNA] (ISOFORMS A AND C), FUNCTION, INTERACTION
RP   WITH PKC-3, SUBCELLULAR LOCATION, TISSUE SPECIFICITY, DEVELOPMENTAL
RP   STAGE, AND MUTAGENESIS OF PHE-175 AND PHE-221.
RC   STRAIN=Bristol N2;
RX   PubMed=11134024; DOI=10.1074/jbc.M008990200;
RA   Zhang L., Wu S.-L., Rubin C.S.;
RT   "A novel adapter protein employs a phosphotyrosine binding domain and
RT   exceptionally basic N-terminal domains to capture and localize an
RT   atypical protein kinase C: characterization of Caenorhabditis elegans
RT   C kinase adapter 1, a protein that avidly binds protein kinase C3.";
RL   J. Biol. Chem. 276:10463-10475(2001).
```

The formats of the individual lines are explained below.

3.11.2. The RN line

The RN (Reference Number) line gives a sequential number to each reference citation in an entry. This number is used to indicate the reference in comments and feature table notes. The format of the RN line is:

```
RN   [n]
```

where 'n' denotes the n^{th} reference for this entry. The reference number is always between square brackets.

3.11.3. The RP line

The RP (Reference Position) lines describe the extent of the work relevant to the entry carried out by the authors. The format of the RP line is:

```
RP   COMMENT.
```

It should contain a description of the information that has been propagated in the Swiss-Prot entry.

A typical comment is "NUCLEOTIDE SEQUENCE". This item might be tagged with a qualifier, indicating the origin of the sequence data. Valid names of this qualifiers are:

- GENOMIC DNA: the individual gene has been sequenced
- GENOMIC RNA: the individual gene has been sequenced
- MRNA: the individual cDNA has been sequenced
- LARGE SCALE GENOMIC DNA: the gene has been sequenced as part of a genome project
- LARGE SCALE MRNA: the cDNA has been sequenced as part of a large-scale cDNA project

If 2 qualifiers apply, both are indicated, separated by a '/'.

The 'LARGE SCALE ANALYSIS' is another typical tag added in references that report large screen results to indicate that results have not been extensively studied.

Typical examples of RP lines are shown below:

```
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].
```

```
RP   NUCLEOTIDE SEQUENCE [GENOMIC DNA / MRNA] (ISOFORM 1).
```

```
RP   NUCLEOTIDE SEQUENCE [GENOMIC RNA / MRNA].
```

```
RP   NUCLEOTIDE SEQUENCE [GENOMIC DNA], AND PROTEIN SEQUENCE OF 21-35.
```

```
RP   PROTEIN SEQUENCE OF 39-76; 95-118 AND 125-138, AND DISULFIDE BONDS.
```

```
RP   SEQUENCE REVISION TO 76-84 AND 129.
```

```
RP   STRUCTURE BY NMR.
```

```
RP   X-RAY CRYSTALLOGRAPHY (1.8 ANGSTROMS).
```

```
RP   CHARACTERIZATION.
```

```
RP   MUTAGENESIS OF TYR-65.
```

```
RP   REVIEW.
```

```
RP   VARIANT ALA-1368.
```

```
RP   VARIANTS HDLD1 ARG-597 AND ARG-1477, AND VARIANT HDLD2 LEU-693 DEL.
```

```
RP   NUCLEOTIDE SEQUENCE [GENOMIC DNA], PROTEIN SEQUENCE OF 1-22; 2-17;
RP   240-256; 318-339 AND 381-390, AND CHARACTERIZATION.
```

```
RP   NUCLEOTIDE SEQUENCE [MRNA], PROTEIN SEQUENCE OF 154-171; 302-308;
RP   312-328; 377-384 AND 419-431, FUNCTION, SUBCELLULAR LOCATION, AND
RP   MUTAGENESIS OF ARG-331; GLY-332 AND ARG-333.
```

```
RP   PHOSPHORYLATION [LARGE SCALE ANALYSIS] AT SER-387 AND SER-391, AND
RP   MASS SPECTROMETRY.
```

3.11.4. The RC line

The RC (Reference Comment) lines are optional lines which are used to store comments relevant to the reference cited. The format of the RC line is:

```
RC   TOKEN1=Text; TOKEN2=Text; ...
```

The currently defined tokens and their order in the RC line are:

STRAIN
PLASMID
TRANSPOSON
TISSUE

Reference comment line topics may span lines. Examples of RC lines:

```
RC   STRAIN=Sprague-Dawley; TISSUE=Liver;
```

```
RC   STRAIN=Holstein; TISSUE=Lymph node, and Mammary gland;
```

```
RC   STRAIN=301 / Serotype 2a;
```

```
RC   STRAIN=cv. SP753012-O; TISSUE=Leaf;
```

```
RC   PLASMID=R1 (R7268); TRANSPOSON=Tn3;
```

```
RC   STRAIN=AL.012, AZ.026, AZ.180, DC.005, GA.039, GA2181, IL.014, IL2.17,
RC   IN.018, KY.172, KY2.37, LA.013, MI.035, MN.001, MNb027, MS.040,
RC   NY.016, OH.036, TN.173, TN2.38, UT.002, and VA.015;
```

3.11.5. The RX line

The RX (Reference cross-reference) line is an optional line which is used to indicate the identifier assigned to a specific reference in a bibliographic database. The format of the RX line is:

```
RX   Bibliographic_db=IDENTIFIER[; Bibliographic_db=IDENTIFIER...];
```

Where the valid bibliographic database names and their associated identifiers are:

Name	Identifier
MEDLINE	Eight-digit MEDLINE Unique Identifier (UI)
PubMed	PubMed Unique Identifier (PMID)
DOI	Digital Object Identifier (DOI)
AGRICOLA	AGRICOLA Unique Identifier

Example of RX lines:

```
RX   MEDLINE=83283433; PubMed=6688356;
```

```
RX   PubMed=15626370; DOI=10.1016/j.toxicon.2004.10.011;
```

```
RX   MEDLINE=22709107; PubMed=12788972; DOI=10.1073/pnas.1130426100;
```

```
RX   AGRICOLA=IND20551642; DOI=10.1007/BF00224104;
```

3.11.6. The RG line

The Reference Group (RG) line lists the consortium name associated with a given citation. The RG line is mainly used in submission reference blocks, but can also be used in paper references, if the working group is cited as an author in the paper. RG line and RA line (Reference Author) can be present in the same reference block; at least one RG or RA line is mandatory per reference block. An example of the use of RG lines is shown below:

```
RG   The mouse genome sequencing consortium;
```

3.11.7. The RA line

The RA (Reference Author) lines list the authors of the paper (or other work) cited. The RA line is present in most references, but might be missing in references that cite a reference group (see RG line). At least one RG or RA line is mandatory per reference block.

All of the authors are included, and are listed in the order given in the paper. The names are listed surname first followed by a blank, followed by initial(s) with periods. The authors' names are separated by commas and terminated by a semicolon. Author names are not split between lines. An example of the use of RA lines is shown below:

```
RA   Galinier A., Bleicher F., Negre D., Perriere G., Duclos B.,
RA   Cozzone A.J., Cortay J.-C.;
```

As many RA lines as necessary are included in each reference.

An author's initials can be followed by an abbreviation such as 'Jr' (for Junior), 'Sr' (Senior), 'II', 'III' or 'IV' (2nd, 3rd and 4th). Example:

```
RA   Nasoff M.S., Baker H.V. II, Wolf R.E. Jr.;
```

3.11.8. The RT line

The RT (Reference Title) lines give the title of the paper (or other work) cited as exactly as possible given the limitations of the computer character set. The format of the RT line is:

```
RT   "Title.";
```

Example of a set of RT lines:

```
RT   "New insulin-like proteins with atypical disulfide bond pattern
RT   characterized in Caenorhabditis elegans by comparative sequence
RT   analysis and homology modeling.";
```

It should be noted that the format of the title is not always identical to that displayed at the top of the published work:

- Major title words are not capitalized;
- The text of a title ends with either a period '.', a question mark '?' or an exclamation mark '!';
- Double quotation marks ' " ' in the text of the title are replaced by single quotation marks;
- Titles of articles published in a language other than English have been translated into English;
- Greek letters are written in full (alpha, beta, etc.).

3.11.9. The RL line

The RL (Reference Location) lines contain the conventional citation information for the reference. In general, the RL lines alone are sufficient to find the paper in question.

a) Journal citations

The RL line for a journal citation includes the journal abbreviation, the volume number, the page range and the year. The format for such an RL line is:

```
RL   Journal_abbrev Volume:First_page-Last_page(YYYY).
```

Journal names are abbreviated according to the conventions used by the National Library of Medicine (NLM) and are based on the existing ISO and ANSI standards. A list of the abbreviations currently in use is given in the document file jourlist.txt

An example of an RL line is:

```
RL   J. Mol. Biol. 168:321-331(1983).
```

When a reference is made to a paper which is 'in press' at the time the database is released, the page range, and possibly the volume number, are indicated as '0' (zero). An example of such an RL line is shown here:

```
RL   Int. J. Parasitol. 0:0-0(2005).
```

b) Electronic publications

The RL line for an electronic publication includes an '(er)' prefix. The format is indicated below:

```
RL   (er) Free text.
```

Examples:

```
RL   (er) Plant Gene Register PGR98-023.
```

```
RL   (er) Worm Breeder's Gazette 15(3):34(1998).
```

c) Book citations

A variation of the RL line format is used for papers found in books or other types of publication, which are then cited using the following format:

```
RL   (In) Editor_1 I.[, Editor_2 I., Editor_X I.] (eds.);
RL   Book_name, pp.[Volume:]First_page-Last_page, Publisher, City (YYYY).
```

Examples:

```
RL   (In) Boyer P.D. (eds.);
RL   The enzymes (3rd ed.), pp.11:397-547, Academic Press, New York (1975).

RL   (In) Rich D.H., Gross E. (eds.);
RL   Proceedings of the 7th American peptide symposium, pp.69-72, Pierce
RL   Chemical Co., Rockford Il. (1981).

RL   (In) Magnusson S., Ottesen M., Foltmann B., Dano K., Neurath H.
RL   (eds.);
RL   Regulatory proteolytic enzymes and their inhibitors, pp.163-172,
RL   Pergamon Press, New York (1978).
```

d) Unpublished observations

For unpublished observations the format of the RL line is:

```
RL   Unpublished observations (MMM-YYYY).
```

Where 'MMM' is the month and 'YYYY' is the year.

We use the 'unpublished observations' RL line to cite communications by scientists to Swiss-Prot of unpublished information concerning various aspects of a sequence entry.

e) Thesis

For Ph.D. theses the format of the RL line is:

```
RL   Thesis (Year), Institution_name, Country.
```

An example of such a line is given here:

```
RL   Thesis (1977), University of Geneva, Switzerland.
```

f) Patent applications

For patent applications the format of the RL line is:

```
RL   Patent number Pat_num, DD-MMM-YYYY.
```

Where 'Pat_num' is the international publication number of the patent, 'DD' is the day, 'MMM' is the month and 'YYYY' is the year. Example:

```
RL   Patent number WO9010703, 20-SEP-1990.
```

g) Submissions

The final form that an RL line can take is that used for submissions. The format of such an RL line is:

```
RL   Submitted (MMM-YYYY) to Database_name.
```

Where 'MMM' is the month, 'YYYY' is the year and 'Database_name' is one of the following:

the EMBL/GenBank/DDBJ databases
UniProtKB
the PDB data bank
the PIR data bank

Two examples of submission RL lines are given here:

```
RL   Submitted (OCT-1995) to the EMBL/GenBank/DDBJ databases.
RL   Submitted (APR-2004) to UniProtKB.
```

3.12. The CC line

Table of contents

The CC lines are free text comments on the entry, and are used to convey any useful information. The comments always appear below the last reference line and are grouped together in comment blocks; a block is made up of 1 or more comment lines. The first line of a block starts with the characters '-!-'.

The format of a comment block is:

```
CC   -!- TOPIC: First line of a comment block;

CC       second and subsequent lines of a comment block.
```

The comment blocks are arranged according to what we designate as 'topics'. The current topics and their definitions are listed in the table below.

Topic	Description
ALLERGEN	Information relevant to allergenic proteins
ALTERNATIVE PRODUCTS	Description of the existence of related protein sequence(s) produced by alternative splicing of the same gene, alternative promoter usage, ribosomal frameshifting or by the use of alternative initiation codons; see 3.22.16
BIOPHYSICOCHEMICAL PROPERTIES	Description of the information relevant to biophysical and physicochemical data and information on pH dependence, temperature dependence, kinetic parameters, redox potentials, and maximal absorption; see 3.22.8
BIOTECHNOLOGY	Description of the use of a specific protein in a biotechnological process
CATALYTIC ACTIVITY	Description of the reaction(s) catalyzed by an enzyme [1]
CAUTION	Warning about possible errors and/or grounds for confusion
COFACTOR	Description of any non-protein substance required by an enzyme for its catalytic activity
DEVELOPMENTAL STAGE	Description of the developmentally-specific expression of mRNA or protein
DISEASE	Description of the disease(s) associated with a deficiency of a protein
DISRUPTION PHENOTYPE	Description of the effects caused by the disruption of the gene coding for the protein; see 3.22.28
DOMAIN	Description of the domain structure of a protein
ENZYME REGULATION	Description of an enzyme regulatory mechanism
FUNCTION	General description of the function(s) of a protein
INDUCTION	Description of the compound(s) or condition(s) that regulate gene expression
INTERACTION	Conveys information relevant to binary protein-protein interaction 3.22.12
MASS SPECTROMETRY	Reports the exact molecular weight of a protein or part of a protein as determined by mass spectrometric methods; see 3.22.24
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 any chemical alternation of a polypeptide (proteolytic cleavage, amino acid modifications including crosslinks). This topic complements information given in the feature table or indicates polypeptide modifications for which position-specific data is not available.
RNA EDITING	Description of any type of RNA editing that leads to one or more amino acid changes
SEQUENCE CAUTION	Description of protein sequence reports that differ from the sequence that is shown in UniProtKB due to conflicts that are not described in FT CONFLICT lines, such as frameshifts, erroneous gene model predictions, etc. See 3.22.37
SIMILARITY	Description of the similaritie(s) (sequence or structural) of a protein with other proteins
SUBCELLULAR LOCATION	Description of the subcellular location of the chain/peptide/isoform. See 3.22.14
SUBUNIT	Description of the quaternary structure of a protein and any kind of interactions with other proteins or protein complexes; except for receptor-ligand interactions, which are described in the topic FUNCTION.
TISSUE SPECIFICITY	Description of the tissue-specific expression of mRNA or protein
TOXIC DOSE	Description of the lethal dose (LD), paralytic dose (PD) or effective dose of a protein
WEB RESOURCE	Description of a cross-reference to a network database/resource for a specific protein; see 3.22.39

Note:

[1] For the 'CATALYTIC ACTIVITY' topic: To describe the catalytic activity of an enzyme we have used, whenever possible, the recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB) as published in Enzyme Nomenclature, NC-IUBMB, Academic Press, New-York, (1992).

3.12.1. Examples for each comment line topic

We show here, for each of the defined topics, two examples of their usage:

```
CC   -!- ALLERGEN: Causes an allergic reaction in human. Binds to IgE.
CC       Partially heat-labile allergen that may cause both respiratory and
CC       food-allergy symptoms in patients with the bird-egg syndrome.

CC   -!- ALLERGEN: Causes an allergic reaction in human. Minor allergen of
CC       bovine dander.

CC   -!- ALTERNATIVE PRODUCTS:
CC       Event=Alternative splicing; Named isoforms=3;
CC         Comment=Additional isoforms seem to exist. Experimental
CC         confirmation may be lacking for some isoforms;
CC       Name=1; Synonyms=AIRE-1;
CC         IsoId=O43918-1; Sequence=Displayed;
CC       Name=2; Synonyms=AIRE-2;
CC         IsoId=O43918-2; Sequence=VSP_004089;
CC       Name=3; Synonyms=AIRE-3;
CC         IsoId=O43918-3; Sequence=VSP_004089, VSP_004090;

CC   -!- ALTERNATIVE PRODUCTS:
CC       Event=Alternative initiation; Named isoforms=2;
CC       Name=Alpha;
CC         IsoId=P51636-1; Sequence=Displayed;
CC       Name=Beta;
CC         IsoId=P51636-2; Sequence=VSP_018696;

CC   -!- BIOPHYSICOCHEMICAL PROPERTIES:
CC       pH dependence:
CC         Optimum pH is 8-10;
CC       Temperature dependence:
CC         Highly active at low temperatures, even at 0 degree Celsius.
CC         Thermolabile;

CC   -!- BIOPHYSICOCHEMICAL PROPERTIES:
CC       Kinetic parameters:
CC         KM=98 uM for ATP;
CC         KM=688 uM for pyridoxal;
CC         Vmax=1.604 mmol/min/mg enzyme;
CC       pH dependence:
CC         Optimum pH is 6.0. Active from pH 4.5 to 10.5;

CC   -!- BIOTECHNOLOGY: The effect of PG can be neutralized by introducing
CC       an antisense PG gene by genetic manipulation. The Flavr Savr
CC       tomato produced by Calgene (Monsanto) in such a manner has a
CC       longer shelf life due to delayed ripening.

CC   -!- BIOTECHNOLOGY: Used in the food industry for high temperature
CC       liquefaction of starch-containing mashes and in the detergent
CC       industry to remove starch. Sold under the name Termamyl by
CC       Novozymes.

CC   -!- CATALYTIC ACTIVITY: ATP + L-glutamate + NH(3) = ADP + phosphate +
CC       L-glutamine.

CC   -!- CATALYTIC ACTIVITY: (R)-2,3-dihydroxy-3-methylbutanoate + NADP(+)
CC       = (S)-2-hydroxy-2-methyl-3-oxobutanoate + NADPH.

CC   -!- CAUTION: It is uncertain whether Met-1 or Met-3 is the initiator.

CC   -!- COFACTOR: Pyridoxal phosphate.

CC   -!- COFACTOR: FAD. {ECO:0000255|HAMAP-Rule:MF_01202}.

CC   -!- WEB RESOURCE: Name=Wikipedia; Note=Amyloid beta entry;
CC       URL="http://en.wikipedia.org/wiki/Amyloid_beta";

CC   -!- DEVELOPMENTAL STAGE: Expressed early during conidial (dormant
CC       spores) differentiation.

CC   -!- DEVELOPMENTAL STAGE: Detected in embryonic skin (E12.5 and E14.5)
CC       during the formation of hair follicles and at E15.5 in the enamel
CC       knot of the developing tooth. Detected in the basal layer of the
CC       epidermis and hair follicles of P2 mice.

CC   -!- DISEASE: Defects in PHKA1 are linked to X-linked muscle
CC       glycogenosis [MIM:311870]. It is a disease characterized by slowly
CC       progressive, predominantly distal muscle weakness and atrophy.

CC   -!- DISEASE: Defects in ABCD1 are the cause of recessive X-linked
CC       adrenoleukodystrophy (X-ALD) [MIM:300100]. X-ALD is a rare
CC       peroxisomal metabolic disorder that occurs in boys and is
CC       characterized by progressive multifocal demyelination of the
CC       central nervous system and by adrenocortical insufficiency. It
CC       produces mental deterioration, corticospinal tract dysfunction,
CC       and cortical blindness. There is laboratory evidence of adrenal
CC       cortical dysfunction. Different clinical manifestations exist
CC       like: cerebral childhood ALD (CALD), adult cerebral ALD (ACALD),
CC       adrenomyeloneuropathy (AMN) and "Addison disease only" (ADO)
CC       phenotype.

CC   -!- DISRUPTION PHENOTYPE: Mice display impaired B-cell development
CC       which does not progress pass the progenitor stage.

CC   -!- DISRUPTION PHENOTYPE: Death before reaching adulthood, probably
CC       due to lethal epilepsy. Mice display severe defects in the
CC       olfactory bulbs, the hippocampus, and the cerebellum. These
CC       defects appear to result from impaired cytokinesis followed by the
CC       induction of apoptosis in specific neuroblast populations.

CC   -!- DOMAIN: Contains a coiled-coil domain essential for vesicular
CC       transport and a dispensable C-terminal region.

CC   -!- DOMAIN: The B chain is composed of two domains, each domain
CC       consists of 3 homologous subdomains (alpha, beta, gamma).

CC   -!- ENZYME REGULATION: The activity of this enzyme is controlled by
CC       adenylation under conditions of abundant glutamine. The fully
CC       adenylated enzyme complex is inactive (By similarity).
CC       {ECO:0000250}.

CC   -!- ENZYME REGULATION: Activated by Gram-negative bacterial
CC       lipopolysaccharides and chymotrypsin.

CC   -!- FUNCTION: Binds to actin and affects the structure of the
CC       cytoskeleton. At high concentrations, profilin prevents the
CC       polymerization of actin, whereas it enhances it at low
CC       concentrations. By binding to PIP2, it inhibits the formation of
CC       IP3 and DG.

CC   -!- FUNCTION: Inhibitor of fungal polygalacturonase. It is an
CC       important factor for plant resistance to phytopathogenic fungi.
CC       Substrate preference is polygalacturonase (PG) from A.niger >> PG
CC       of F.oxysporum, A.solani or B.cinerea. Not active on PG from
CC       F.moniliforme.

CC   -!- INDUCTION: By heat shock, salt stress, oxidative stress, glucose
CC       limitation and oxygen limitation.

CC   -!- INDUCTION: By infection, plant wounding, or elicitor treatment of
CC       cell cultures.

CC   -!- INTERACTION:
CC       Self; NbExp=1; IntAct=EBI-123485, EBI-123485;
CC       Q9W158:CG4612; NbExp=1; IntAct=EBI-123485, EBI-89895;
CC       Q9VYI0:fne; NbExp=1; IntAct=EBI-123485, EBI-126770;

CC   -!- INTERACTION:
CC       Q9W1K5-1:CG11299; NbExp=1; IntAct=EBI-133844, EBI-212772;

CC   -!- MASS SPECTROMETRY: Mass=24948; Mass_error=6; Method=MALDI;
CC       Range=1-228; Source=PubMed:11101899;

CC   -!- MASS SPECTROMETRY: Mass=13822; Method=MALDI; Range=19-140 (P15522-
CC       2); Source=PubMed:10531593;

CC   -!- MISCELLANEOUS: Binds to bacitracin.
CC   -!- MISCELLANEOUS: Called DUO because the encoded protein is closely
CC       related to but shorter than TRIO.

CC   -!- PATHWAY: Cofactor biosynthesis; porphyrin biosynthesis; 5-
CC       aminolevulinate from L-glutamyl-tRNA(Glu): step 2/2.

CC   -!- PATHWAY: Nucleotide metabolism; purine metabolism.

CC   -!- PHARMACEUTICAL: Available under the names Avonex (Biogen),
CC       Betaseron (Berlex) and Rebif (Serono). Used in the treatment of
CC       multiple sclerosis (MS). Betaseron is a slightly modified form of
CC       IFNB1 with two residue substitutions.

CC   -!- PHARMACEUTICAL: Available under the name Proleukin (Chiron). Used
CC       in patients with renal cell carcinoma or metastatic melanoma.

CC   -!- POLYMORPHISM: The allelic form of the enzyme with Gln-191
CC       (Allozyme A) hydrolyzes paraoxon with a low turnover number and
CC       the one with Arg-191 (Allozyme B) with a high turnover number.

CC   -!- POLYMORPHISM: In the human populations there are two major allelic
CC       forms, alpha-1 with 83 residues and alpha-2 with 142 residues.
CC       These alleles determine the 3 major phenotypes HP*1F/HP*1S and
CC       HP*2FS. The two main alleles of HP*1 are called HP*1F (fast) and
CC       HP*1S (slow).

CC   -!- PTM: N-glycosylated and probably also O-glycosylated.

CC   -!- PTM: A soluble short 95 kDa form may be released by proteolytic
CC       cleavage from the long membrane-anchored form.

CC   -!- RNA EDITING: Modified_positions=393, 431, 452, 495.

CC   -!- RNA EDITING: Modified_positions=59, 78, 94, 98, 102, 121; Note=The
CC       nonsense codon at position 59 is modified to a sense codon. The
CC       stop codon at position 121 is created by RNA editing.

CC   -!- SEQUENCE CAUTION:
CC       Sequence=CAI24940.1; Type=Erroneous gene model prediction;

CC   -!- SIMILARITY: Belongs to the annexin family.

CC   -!- SUBCELLULAR LOCATION: Bacterial cell inner membrane; Multi-pass
CC       membrane protein.

CC   -!- SUBUNIT: Homotetramer.
CC   -!- SUBUNIT: Disulfide-linked heterodimer of a light chain (L) and a
CC       heavy chain (H). The light chain has the pharmacological activity,
CC       while the N- and C-terminal of the heavy chain mediate channel
CC       formation and toxin binding, respectively.

CC   -!- TISSUE SPECIFICITY: Shoots, roots, and cotyledon from dehydrating
CC       seedlings.

CC   -!- TISSUE SPECIFICITY: Expressed at high levels in brain and ovary.
CC       Lower levels in small intestine. In brain regions, detected in all
CC       regions tested. Highest levels in the cerebellum and cerebral
CC       cortex.

CC   -!- TOXIC DOSE: PD(50) is 1.72 mg/kg by injection in blowfly larvae.

CC   -!- TOXIC DOSE: LD(50) is 0.015 mg/kg by intravenous injection for
CC       sarafotoxin-A and sarafotoxin-B, and 0.3 mg/kg for sarafotoxin-C.

CC   -!- WEB RESOURCE: Name=CD40Lbase; Note=CD40L defect database;
CC       URL="http://bioinf.uta.fi/CD40Lbase/";
```

3.12.2. Syntax of the topic 'BIOPHYSICOCHEMICAL PROPERTIES'

```
CC   -!- BIOPHYSICOCHEMICAL PROPERTIES:
CC       Absorption:
CC         Abs(max)=xx nm;
CC         Note=free_text;
CC       Kinetic parameters:
CC         KM=xx unit for substrate [(free_text)];
CC         Vmax=xx unit enzyme [free_text];
CC         Note=free_text;
CC       pH dependence:
CC         free_text;
CC       Redox potential:
CC         free_text;
CC       Temperature dependence:
CC         free_text;
```

A BIOPHYSICOCHEMICAL PROPERTIES block must contain at least one of the properties Absorption, Kinetic parameters, pH dependence, Redox potential, Temperature dependence and may have any combination of these properties (ordered as indicated above). The meaning of these subtopics is as follows:

Property	Description
Absorption	indicates the wavelength at which photoreactive proteins such as opsins and DNA photolyases show maximal absorption
Kinetic parameters	mentions the Michaelis-Menten constant (KM) and maximal velocity (Vmax) of enzymes
pH dependence	describes the optimum pH for enzyme activity and/or the variation of enzyme activity with pH variation
Redox potential	reports the value of the standard (midpoint) oxido-reduction potential(s) for electron transport proteins
Temperature dependence	indicates the optimum temperature for enzyme activity and/or the variation of enzyme activity with temperature variation; the thermostability/thermolability of the enzyme is also mentioned when it is known

3.12.3. The topic 'INTERACTION'

The CC line topic INTERACTION conveys information relevant to binary protein-protein interaction. It is automatically derived from the IntAct database and is updated on a monthly basis. The occurrence is one INTERACTION topic per entry, with each binary interaction being presented in a separate line. Each data line can be longer than 75 characters.

Interactions can be derived by any appropriate experimental method, but must be confirmed by a second experiment, if resulting from a single yeast-two-hybrid experiment. For large-scale experiments, interactions are considered if a high confidence is assigned from the authors.

The format of the CC line topic INTERACTION is:

```
CC   -!- INTERACTION:
CC       {{SP_Ac:identifier[ (xeno)]}|Self}; NbExp=n; IntAct=IntAct_Protein_Ac, IntAct_Protein_Ac;
```

where

SP_Ac is the Swiss-Prot or TrEMBL accession number of the interacting protein. If appropriate, the IsoId is used instead to specify the relevant interacting protein isoform.

identifier serves to describe the interacting protein. It is derived from the Swiss-Prot or TrEMBL GN line and thus presents either a "gene name", a "ordered locus name" or a "ORF name". When no GN line is available a dash is indicated instead.

(xeno) is an optional qualifier indicating that the interacting proteins are derived from different species. This may be due to the experimental set-up or may reflect a pathogen-host interaction.

Self reflects a self-association; the corresponding current entry's SP_Ac and 'identifier' are not given/repeated.

NbExp=n refers to the number of experiments in IntAct supporting the interaction.

IntAct_Protein_Ac is the IntAct accession number of a interacting protein. The first IntAct_Protein_Ac refers to the protein or an isoform of the current entry, the second refers to the interacting protein or isoform.

Within the CC INTERACTION topic, homomeric interactions are listed before the heteromeric interactions; latter are sorted alphanumerical according the 'identifier'.

"IntAct=IntAct_Protein_Ac, IntAct_Protein_Ac" identifies the interaction in IntAct by using the two IntAct protein identifiers.

Examples of interaction lines are given below. The CC INTERACTION topics are not complete; only explained interaction lines are indicated.

```
CC   -!- INTERACTION:
CC       P11450:fcp3c; NbExp=1; IntAct=EBI-126914, EBI-159556;
```

In the typical example the current protein is interacting with P11450 which is further characterized by "fcp3c" derived from its GN line and presents its gene name "Fcp3C". The interaction is supported by one experiment stored in IntAct. Experimental details for this interaction can be found by querying IntAct with "EBI-126914, EBI-159556".

```
CC   -!- INTERACTION:
CC       Q9W1K5-1:CG11299; NbExp=1; IntAct=EBI-133844, EBI-212772;
```

The current protein interacts with an isoform of Q9W1K5 defined by the IsoID Q9W1K5-1 .

```
CC   -!- INTERACTION:
CC       Q8NI08:-; NbExp=1; IntAct=EBI-80809, EBI-80799;
```

No gene name information for the interacting protein is available.

```
CC   -!- INTERACTION:
CC       Self; NbExp=1; IntAct=EBI-123485, EBI-123485;
```

The protein self-associates.

```
CC   -!- INTERACTION:
CC       Q8C1S0:2410018M14Rik (xeno); NbExp=1; IntAct=EBI-394562, EBI-398761;
```

The source organisms of the interacting proteins are different.

```
CC   -!- INTERACTION:
```

```
CC       P51617:IRAK1; NbExp=1; IntAct=EBI-448466, EBI-358664;

CC       P51617:IRAK1; NbExp=1; IntAct=EBI-448472, EBI-358664;
```

Different isoforms of the current protein are shown to interact with the same protein (P51617). This is reflected by different IntAct_Protein_Acs for the current protein.

Example entry with many interaction lines: Q02821.

3.12.4. Syntax of the topic 'SUBCELLULAR LOCATION'

The document subcell.txt, lists the controlled vocabularies used in the comment line (CC) topic SUBCELLULAR LOCATION, their definitions and further information such as synonyms or relevant GO terms in the following format:

Line code	Content	Occurrence in an entry
ID	Identifier (location)	Once; starts an entry
IT	Identifier (topology)	Once; starts a 'topology' entry
IO	Identifier (orientation)	Once; starts an 'orientation' entry
AC	Accession (SL-xxxx)	Once
DE	Definition	Once or more
SY	Synonyms	Optional; Once or more
SL	Content of subc. loc. lines	Once
HI	Hierarchy ('is-a')	Optional; Once or more
HP	Hierarchy ('part-of')	Optional; Once or more
KW	Associated keyword (accession)	Optional; Once or more
GO	Gene ontology (GO) mapping	Optional; Once or more
WW	Interesting links or references	Optional; Once or more
//	Terminator	Once; ends an entry

Example:

```
ID   Cyanelle.
AC   SL-0082
DE   A cyanelle is a photosynthetic organelle of glaucocystophyte algae.
DE   Cyanelles are surrounded by a double membrane and, in between, a
DE   peptidoglycan wall. Thylakoid membrane architecture and the presence
DE   of carboxysomes are cyanobacteria-like. Historically, the term
DE   cyanelle is derived from a classification as endosymbiotic
DE   cyanobacteria, and thus is not fully correct.
SY   Muroplast; Cyanoplast.
SL   Plastid, cyanelle.
HI   Plastid.
KW   KW-0194
GO   GO:0009842; cyanelle
//
```

The format of SUBCELLULAR LOCATION is:

```
CC   -!- SUBCELLULAR LOCATION:(( Molecule:)?( Location\.)+)?( Note=Free_text( Flag)?\.)?
```

Where:

- *Molecule*: Isoform, chain or peptide name
- *Location* = `Subcellular_location( Flag)?(; Topology( Flag)?)?(; Orientation( Flag)?)?`
  - *Subcellular_location*: SL-line of subcell.txt ID-record
  - *Topology*: SL-line of subcell.txt IT-record
  - *Orientation*: SL-line of subcell.txt IO-record

Note: Perl-style multipliers indicate whether a pattern (as delimited by parentheses) is optional (?) or may occur 1 or more times (+).

Examples:

When no chain/peptide/isoform is specified, the subcellular location corresponds to that of the mature protein.

```
CC   -!- SUBCELLULAR LOCATION: Cytoplasm. Endoplasmic reticulum membrane;
CC       Peripheral membrane protein. Golgi apparatus membrane; Peripheral
CC       membrane protein.
```

```
CC   -!- SUBCELLULAR LOCATION: Cell membrane {ECO:0000250}; Peripheral
CC       membrane protein {ECO:0000250}. Secreted {ECO:0000250}. Note=The
CC       last 22 C-terminal amino acids may participate in cell membrane
CC       attachment.
CC   -!- SUBCELLULAR LOCATION: Isoform 2: Cytoplasm {ECO:0000305}.
```

```
 CC   -!- SUBCELLULAR LOCATION: Golgi apparatus, trans-Golgi network
 CC       membrane; Multi-pass membrane protein. Note=Predominantly found in
 CC       the trans-Golgi network (TGN). Not redistributed to the plasma
 CC       membrane in response to elevated copper levels.
 CC   -!- SUBCELLULAR LOCATION: Isoform 1: Golgi apparatus membrane
 CC       {ECO:0000269|PubMed:9307043}; Multi-pass membrane protein
 CC       {ECO:0000269|PubMed:9307043}.
 CC   -!- SUBCELLULAR LOCATION: Isoform 2: Cytoplasm
 CC       {ECO:0000269|PubMed:9307043}.
 CC   -!- SUBCELLULAR LOCATION: WND/140 kDa: Mitochondrion.
```

3.12.5. Syntax of the topic 'ALTERNATIVE PRODUCTS'

The format of the CC line topic ALTERNATIVE PRODUCTS is:

```
 CC   -!- ALTERNATIVE PRODUCTS:
 CC       Event=Event(, Event)*; Named isoforms=Number_of_isoforms;
(CC         Comment=Free_text;)?
(CC       Name=Isoform_name;( Synonyms=Synonym(, Synonym)*;)?
 CC         IsoId=Isoform_identifier(, Isoform_identifer)*;
 CC         Sequence=(Displayed|External|Not described|Feature_identifier(, Feature_identifier)*);
(CC         Note=Free_text;)?)+
```

Note: Variable values are represented in italics. Perl-style multipliers indicate whether a pattern (as delimited by parentheses) is optional (?), may occur 0 or more times (*), or 1 or more times (+). Alternative values are separated by a pipe symbol (|).

Topic	Description

Event	Biological process that results in the production of the alternative forms. It lists one or a combination of the following values (Alternative promoter usage, Alternative splicing, Alternative initiation, Ribosomal frameshifting).
	Format: Event=controlled vocabulary;
	Example: Event=Alternative splicing;

Named isoforms	Number of isoforms listed in the topics 'Name' currently only for 'Event=Alternative splicing'.
	Format: Named isoforms=number;
	Example: Named isoforms=6;

Comment	Any comments concerning one or more isoforms; optional; Format: Comment=free text; Example: Comment=Experimental confirmation may be lacking for some isoforms;

Name	A common name for an isoform used in the literature or assigned by Swiss-Prot; currenty only available for spliced isoforms.
	Format: Name=common name;
	Example: Name=Alpha;

Synonyms	Synonyms for an isoform as used in the literature; optional; currently only available for spliced isoforms.
	Format: Synonyms=Synonym_1[, Synonym_n];
	Example: Synonyms=B, KL5;

IsoId	Unique identifier for an isoform, consisting of the Swiss-Prot accession number, followed by a dash and a number.
	Format: IsoId=acc#-isoform_number[, acc#-isoform_number];
	Example: IsoId=P05067-1;

Sequence	Information on the isoform sequence; the term '**Displayed**' indicates, that the sequence is shown in the entry; a lists of feature identifiers (VSP_#) indicates that the isoform is annotated in the feature table; the FTIds enable programs to create the sequence of a splice variant; if the accession number of the IsoId does not correspond to the accession number of the current entry, this topic contains the term '**External**'; '**Not described**' points out that the sequence of the isoform is unknown.
	Format: Sequence=VSP_#[, VSP_#]|Displayed|External|Not described;
	Example: Sequence=Displayed;
	Example: Sequence=VSP_000013, VSP_000014; Example: Sequence=External;
	Example: Sequence=Not described;

Note	Lists isoform-specific information; optional. It may specify the event(s), if there are several. Format: Note=Free text; Example: Note=No experimental confirmation available;

Example of the CC lines and the corresponding FT lines for an entry with alternative splicing:

```
CC   -!- ALTERNATIVE PRODUCTS:
CC       Event=Alternative splicing, Alternative initiation; Named isoforms=8;
CC         Comment=Additional isoforms seem to exist;
CC       Name=1; Synonyms=Non-muscle isozyme;
CC         IsoId=Q15746-1; Sequence=Displayed;
CC       Name=2;
CC         IsoId=Q15746-2; Sequence=VSP_004791;
CC       Name=3A;
CC         IsoId=Q15746-3; Sequence=VSP_004792, VSP_004794;
CC       Name=3B;
CC         IsoId=Q15746-4; Sequence=VSP_004791, VSP_004792, VSP_004794;
CC       Name=4;
CC         IsoId=Q15746-5; Sequence=VSP_004792, VSP_004793;
CC       Name=Del-1790;
CC         IsoId=Q15746-6; Sequence=VSP_004795;
CC       Name=5; Synonyms=Smooth-muscle isozyme;
CC         IsoId=Q15746-7; Sequence=VSP_018845;
CC         Note=Produced by alternative initiation at Met-923 of isoform 1;
CC       Name=6; Synonyms=Telokin;
CC         IsoId=Q15746-8; Sequence=VSP_018846;
CC         Note=Produced by alternative initiation at Met-1761 of isoform
CC         1. Has no catalytic activity;
...
FT   VAR_SEQ       1   1760       Missing (in isoform 6).
FT                                /FTId=VSP_018846.
FT   VAR_SEQ       1    922       Missing (in isoform 5).
FT                                /FTId=VSP_018845.
FT   VAR_SEQ     437    506       VSGIPKPEVAWFLEGTPVRRQEGSIEVYEDAGSHYLCLLKA
FT                                RTRDSGTYSCTASNAQGQVSCSWTLQVER -> G (in
FT                                isoform 2 and isoform 3B).
FT                                /FTId=VSP_004791.
FT   VAR_SEQ    1433   1439       DEVEVSD -> MKWRCQT (in isoform 3A,
FT                                isoform 3B and isoform 4).
FT                                /FTId=VSP_004792.
FT   VAR_SEQ    1473   1545       Missing (in isoform 4).
FT                                /FTId=VSP_004793.
FT   VAR_SEQ    1655   1705       Missing (in isoform 3A and isoform 3B).
FT                                /FTId=VSP_004794.
FT   VAR_SEQ    1790   1790       Missing (in isoform Del-1790).
FT                                /FTId=VSP_004795.


CC   -!- ALTERNATIVE PRODUCTS:
CC       Event=Alternative splicing, Alternative initiation; Named isoforms=3;
CC         Comment=Isoform 1 and isoform 2 arise due to the use of two
CC         alternative first exons joined to a common exon 2 at the same
CC         acceptor site but in different reading frames, resulting in two
CC         completely different isoforms;
CC       Name=1; Synonyms=p16INK4a;
CC         IsoId=O77617-1; Sequence=Displayed;
CC       Name=3;
CC         IsoId=O77617-2; Sequence=VSP_018701;
CC         Note=Produced by alternative initiation at Met-35 of isoform 1.
CC         No experimental confirmation available;
CC       Name=2; Synonyms=p19ARF;
CC         IsoId=O77618-1; Sequence=External;
..
FT   VAR_SEQ       1     34       Missing (in isoform 3).
FT                                /FTId=VSP_004099.
```

3.12.6. Syntax of the topic 'MASS SPECTROMETRY'

```
CC   -!- MASS SPECTROMETRY: Mass=mass(; Mass_error=error)?; Method=method; Range=ranges( (IsoformID))?(; Note=free_text)?; Source=refe
```

Where:

- 'Mass=XXX' is the determined molecular weight (MW);
- 'Mass_error=XX' (optional) is the accuracy or error range of the MW measurement;
- 'Method=XX' is the ionization method;
- 'Range=XX-XX[(Name)]' is used to indicate what part of the protein sequence entry corresponds to the molecular weight. In case of multiple products, the name of the relevant isoform is enclosed;
- 'Note={Free text}'. Comment in free text format;
- 'Source=PubMed:/Ref.n' indicates the relevant reference'.

3.12.7. DISRUPTION PHENOTYPE

Note that we only describe effects caused the complete absence of a gene and thus a protein in vivo (null mutants caused by random or target deletions, insertions of a transposable element etc.) To avoid description of phenotypes due to partial or dominant negative mutants, missense mutations are not described in this comment, but in FT MUTAGEN instead. Defects caused by transient inactivation by methods such as RNA interference (RNAi) or blockage by antibodies are not described in this comment due to the difficulty to interpret results, except for C. elegans RNAi studies, which are widely used and done in vivo.

3.12.8. Syntax of the topic 'SEQUENCE CAUTION'

The format of the SEQUENCE CAUTION topic is:

```
CC   -!- SEQUENCE CAUTION:
         Sequence=Sequence; Type=Type;[ Positions=Positions;][ Note=Note;]
```

Where:

- *Sequence* is the sequence which differs from the UniProtKB sequence. It is described by one of:
  - an EMBL protein identifier (with version number)
  - an EMBL accession number.
  - a literature reference (e.g. Ref.3).
- *Type* describes the cause for the sequence difference(s) and is one of:
  - Frameshift

  - Erroneous initiation
  - Erroneous termination
  - Erroneous gene model prediction
  - Erroneous translation
  - Miscellaneous discrepancy
- *Positions* describes the sequence position(s) or range(s) of the difference(s) compared to the displayed sequence where possible. Sometimes the term 'Several' is used to indicate that there are many differences.
- *Note* is an optional free text explanation.

These lines will not be wrapped and their length may therefore exceed 75 characters.

3.12.9. Syntax of the topic 'WEB RESOURCE'

```
CC   -!- WEB RESOURCE: Name=ResourceName[; Note=FreeText][; URL=WWWAddress].
```

Where:

- 'Name' is the name of the database;
- 'Note' (optional) is a free text note;
- 'URL' is the WWW address (URL) of the database;

The length of these lines may exceed 75 characters because long URL addresses are not wrapped into multiple lines.

3.13. The DR line

Table of contents

3.13.1. Definition

The DR (Database cross-Reference) lines are used as pointers to information in external data resources that is related to UniProtKB entries. The full list of all databases to which UniProtKB is cross-referenced can be found in the document dbxref.txt. It also contains references describing these resources and provides links to their web sites.

For example, if the X-ray crystallographic atomic coordinates of a sequence are stored in the Protein Data Bank (PDB) there will be one DR line pointing to each of the corresponding entries in PDB. For a sequence translated from a nucleotide sequence there will be DR line(s) pointing to the relevant entri(es) in the EMBL/GenBank/DDBJ database which correspond to the DNA or RNA sequence(s) from which it was translated.

The format of the DR line is:

```
DR   RESOURCE_ABBREVIATION; RESOURCE_IDENTIFIER; OPTIONAL_INFORMATION_1[; OPTIONAL_INFORMATION_2][; OPTIONAL_INFORMATION_3].
```

The cross-references to the EMBL/GenBank/DDBJ nucleotide sequence database and PROSITE are described in sections 3.25 and 3.25.126.

3.13.2. Resource abbreviation

The first field of the DR line, the 'RESOURCE_ABBREVIATION', is the abbreviated name of the referenced resource. The currently defined abbreviations are listed below.

Abbreviation	Description
EMBL	Nucleotide sequence database of EMBL/EBI (see 3.25)
Allergome	Allergome; a platform for allergen knowledge
ArachnoServer	ArachnoServer: Spider toxin database
Araport	Araport: Arabidopsis Information Portal
Bgee	Bgee dataBase for Gene Expression Evolution
BindingDB	The Binding Database
BioCyc	Collection of Pathway/Genome Databases
BioGrid	BioGrid, The Biological General Repository for Interaction Datasets
BioMuta	BioMuta curated single-nucleotide variation and disease association database
BRENDA	BRENDA Comprehensive Enzyme Information System
CarbonylDB	CarbonylDB database of protein carbonylation sites
CAZy	Carbohydrate-Active enZymes
CCDS	The Consensus CDS (CCDS) project
CDD	Conserved Domains Database
ChEMBL	A database of bioactive drug-like small molecules.
ChiTaRS	A database of human, mouse and fruit fly chimeric transcripts and RNA-sequencing data
CGD	Candida genome database
CleanEx	Public gene expression data via unique approved gene symbols
COMPLUYEAST-2DPAGE	2-D database at Universidad Complutense de Madrid
CollecTF	CollecTF database of bacterial transcription factor binding sites
ConoServer	ConoServer: Cone snail toxin database
CORUM	CORUM comprehensive resource of mammalian protein complexes
CTD	Comparative Toxicogenomics Database
dictyBase	Dictyostelium discoideum online informatics resource
DIP	Database of interacting proteins

DMDM	Domain mapping of disease mutations
DNASU	The DNASU plasmid repository
DOSAC-COBS-2DPAGE	2D-PAGE database from the dipartimento oncologico di III Livello
DisGeNET	DisGeNET discovery platform integrating information on gene-disease associations
DisProt	Database of protein disorders
DrugBank	The DrugBank database (DrugBank)
EchoBASE	The integrated post-genomic database for E. coli (EchoBASE)
EcoGene	Escherichia coli K12 genome database (EcoGene)
eggNOG	evolutionary genealogy of genes: Non-supervised Orthologous Groups
ELM	The Eukaryotic Linear Motif resource for functional sites in proteins
Ensembl	Database of automatically annotated sequences of large genomes (Ensembl database)
EnsemblBacteria	This databases is part of Ensembl Genomes, which has been created to complement the existing Ensembl site, the focus of which are vertebrate genomes.
EnsemblFungi	This databases is part of Ensembl Genomes, which has been created to complement the existing Ensembl site, the focus of which are vertebrate genomes.
EnsemblMetazoa	This databases is part of Ensembl Genomes, which has been created to complement the existing Ensembl site, the focus of which are vertebrate genomes.
EnsemblPlants	This databases is part of Ensembl Genomes, which has been created to complement the existing Ensembl site, the focus of which are vertebrate genomes.
EnsemblProtists	This databases is part of Ensembl Genomes, which has been created to complement the existing Ensembl site, the focus of which are vertebrate genomes.
EPD	The Encyclopedia of Proteome Dynamics is a resource that contains data from multiple, large-scale proteomics experiments aimed at characterising proteome dynamics in both human cells and model organisms.
ESTHER	The server ESTHER (ESTerases and alpha/beta-Hydrolase Enzymes and Relatives) is dedicated to the analysis of proteins or protein domains belonging to the superfamily of alpha/beta-hydrolases, exemplified by the cholinesterases.
euHCVdb	The European Hepatitis C Virus database
EuPathDB	Eukaryotic Pathogen Database Resources
EvolutionaryTrace	The Evolutionary Trace ranks amino acid residues in a protein sequence by their relative evolutionary importance.
ExpressionAtlas	Information on gene expression patterns under different biological conditions.
FlyBase	Drosophila genome database (FlyBase)
Gene3D	Database of structural assignments for genes (Gene3D)
GeneCards	GeneCards: human genes, protein and diseases
GeneDB	GeneDB pathogen genome database from Sanger Institute
GeneID	Database of genes from NCBI RefSeq genomes
GeneReviews	GeneReviews, a resource of expert-authored, peer-reviewed disease descriptions.
GeneWiki	GeneWiki, an initiative that aims to create seed articles for every notable human gene.
GenomeRNAi	Database of phenotypes from RNA interference screens in Drosophila and Homo sapiens.
GeneTree	The phylogenetic gene trees that are available at http://www.ensembl.org/ and http://ensemblgenomes.org/.
Genevisible	Genevisible is a free resource to explore public expression data for a gene of interest and find the top five tissues, cell lines, cancers or perturbations in which it has the highest expression or response.
GO	Gene Ontology (GO) database
Gramene	Comparative mapping resource for grains (Gramene)
GuidetoPHARMACOLOGY	An expert-driven guide to pharmacological targets and the substances that act on them
HGNC	Human gene nomenclature database (HGNC)
H-InvDB	Human gene database H-Invitational Database (H-InvDB)
HAMAP	Database of microbial protein families (HAMAP)
HOGENOM	Homologous genes from fully sequenced organisms database
HOVERGEN	Homologous vertebrate genes database
HPA	Human Protein Atlas
IMGT/GENE-DB	IMGT genome database for vertebrate immunoglobulin and T-cell receptor genes
InParanoid	InParanoid: Eukaryotic Ortholog Groups
IntAct	Protein interaction database and analysis system (IntAct)
InterPro	Integrated resource of protein families, domains and functional sites (InterPro)
IPI	International Protein Index
iPTMnet	iPTMnet integrated resource for PTMs in systems biology context
KEGG	Kyoto encyclopedia of genes and genomes
KO	KEGG Orthology
LegioList	Legionella pneumophila (strains Paris and Lens) genome database
Leproma	Mycobacterium leprae genome database (Leproma)
MaizeGDB	Maize Genetics/Genomics Database (MaizeGDB)
MalaCards	MalaCards human disease database
MaxQB	MaxQB - The MaxQuant DataBase

MEROPS	Peptidase database (MEROPS)
MGI	Mouse Genome Informatics Database (MGI)
MIM	Mendelian Inheritance in Man Database (MIM)
MINT	Molecular INTeraction database
mycoCLAP	mycoCLAP
neXtProt	neXtProt, the human protein knowledge platform
OGP	USC-OGP 2-DE database
OMA	Identification of Orthologs from Complete Genome Data
OpenTargets	Target Validation Platform
Orphanet	Orphanet; a database dedicated to information on rare diseases and orphan drugs
OrthoDB	Database of Orthologous Groups
PANTHER	Protein ANalysis THrough Evolutionary Relationships (PANTHER) Classification System
PATRIC	Pathosystems Resource Integration Center
PaxDb	A comprehensive absolute protein abundance database
PDB	3D-macromolecular structure Protein Data Bank (PDB)
PDBsum	PDB sum
PeptideAtlas	PeptideAtlas
PeroxiBase	Peroxidase superfamilies database
Pfam	Pfam protein domain database
PharmGKB	The Pharmacogenetics and Pharmacogenomics Knowledge Base
PhosphoSitePlus	Phosphorylation site database
PhylomeDB	Database for complete collections of gene phylogenies
PIR	Protein sequence database of the Protein Information Resource (PIR)
PIRSF	Protein classification system of PIR (PIRSF)
PMAP-CutDB	CutDB - Proteolytic event database
PomBase	Schizosaccharomyces pombe database
PRIDE	PRoteomics IDEntifications database
PRINTS	Protein Fingerprint database (PRINTS)
ProDom	ProDom protein domain database
PRO	PRO provides an ontological representation of protein-related entities.
ProMEX	Protein Mass spectra EXtraction database
PROSITE	PROSITE protein domain and family database (see 3.25.126)
ProteinModelPortal	Protein Model Portal, a module of the Protein Structure Initiative Knowledgebase (PSI KB) to unify the model data from the different sites.
PseudoCAP	Pseudomonas aeruginosa Community Annotation Project
Reactome	Curated resource of core pathways and reactions in human biology (Reactome)
REBASE	Restriction enzymes and methylases database (REBASE)
RefSeq	NCBI reference sequences
REPRODUCTION-2DPAGE	2D-PAGE database from the Lab of Reproductive Medicine at the Nanjing Medical University
RGD	Rat Genome Database (RGD)
SABIO-RK	Biochemical Reaction Kinetics Database
SFLD	Structure-Function Linkage Database
SGD	Saccharomyces Genome Database (SGD)
SignaLink	A signaling pathway resource with multi-layered regulatory networks
SIGNOR	A signaling network open resource (SIGNOR)
SMART	Simple Modular Architecture Research Tool (SMART)
SMR	The SWISS-MODEL Repository (SMR)
STRING	STRING: functional protein association networks
SUPFAM	Superfamily database of structural and functional annotation
SWISS-2DPAGE	2D-PAGE database from the Geneva University Hospital (SWISS-2DPAGE)
SwissLipids	SwissLipids knowledge resource for lipid biology
SwissPalm	SwissPalm database of S-palmitoylation events
TAIR	The Arabidopsis Information Resource (TAIR)
TCDB	Transport Classification Database
TIGRFAMs	TIGR protein family database (TIGRFAMs)
TopDownProteomics	TopDownProteomics is a resource from the Consortium for Top Down Proteomics that hosts top down proteomics data presenting validated proteoforms to the scientific community.
TreeFam	TreeFam database of animal gene trees
TuberculList	Mycobacterium tuberculosis H37Rv genome database (TubercuList)
UCD-2DPAGE	UCD-2DPAGE: University College Dublin 2-DE Proteome Database.
UniGene	UniGene database

UCSC	UCSC genome browser
UniCarbKB	UniCarbKB, a new collaborative approach to glycomics.
UniPathway	UniPathway: a resource for the exploration and annotation of metabolic pathways
VectorBase	Bioinformatics resource for invertebrate vectors of human pathogens
World-2DPAGE	The World-2DPAGE database
WormBase	A multi-species resource for nematode biology and genomics (WormBase)
WBParaSite	WormBase ParaSite (WBParaSite) resource for parasitic worms (helminths)
Xenbase	Xenopus laevis and tropicalis biology and genome database
ZFIN	Zebrafish Information Network genome database (ZFIN)

3.13.3. The resource identifier

The second field of the DR line, the 'RESOURCE_IDENTIFIER', is an unambiguous pointer to a record in the referenced resource.

- For Allergome, ArachnoServer, Bgee, BioCyc, BioGrid, CCDS, CDD, ChEMBL, CGD, CleanEx, CollecTF, ConoServer, CTD, DIP, DisGeNET, DisProt, DMDM, DNASU, DrugBank, EchoBASE, EcoGene, eggNOG, EvolutionaryTrace, FlyBase, Gene3D, GeneDB, GeneID, GeneReviews, GeneTree, GeneWiki, GenomeRNAi, GO, GuidetoPHARMACOLOGY, HOGENOM, HOVERGEN, HPA, InterPro, IPI, KEGG, KO, MEROPS, MGI, MIM, MINT, mycoCLAP, neXtProt, OpenTargets, Orphanet, OrthoDB, PANTHER, PATRIC, Pfam, PharmGKB, PIR, PRINTS, PRO, ProDom, Reactome, REBASE, RefSeq, REPRODUCTION-2DPAGE, RGD, SGD, SMART, SUPFAM, SwissLipids, TAIR, TCDB, TIGRFAMs, TreeFam, UCSC, UniGene, UniPathway, World-2DPAGE, Xenbase or ZFIN the resource identifier is the accession number (also called the Unique Identifier in some databases) of the referenced entry.
- For BindingDB, CarbonylDB, COMPLUYEAST-2DPAGE, CORUM, DOSAC-COBS-2DPAGE, ELM, EPD, ExpressionAtlas, InParanoid, IntAct, iPTMnet, MaxQB, OGP, PaxDb, PhosphoSitePlus, PhylomeDB, PMAP-CutDB, PeptideAtlas, PRIDE, ProMEX, ProteinModelPortal, SABIO-RK, SignaLink, SIGNOR, SMR, STRING, SWISS-2DPAGE, SwissPalm, TopDownProteomics, UCD-2DPAGE or UniCarbKB, the resource identifier is the UniProtKB accession number.
- For Araport, BioMuta, dictyBase, ESTHER, GeneCards, Genevisible, IMGT/GENE-DB, MalaCards, Peroxibase, PomBase, PseudoCAP the resource identifier is the official gene name.
- For BRENDA, the resource identifier is an EC number.
- For ChiTaRS, the resource identifier is a gene name.
- For CAZy, the resource identifier is the CAZy family number.
- For Ensembl, EnsemblBacteria, EnsemblFungi, EnsemblMetazoa, EnsemblPlants, EnsemblProtists, Gramene, VectorBase, WormBase and WBParaSite, the resource identifier is a transcript identifier.
- For EuPathDB, the primary identifier is a combination of the child database name and the accession number in this database. Both are concatenated by a ':'.
- For HGNC, the resource identifier is the unique identifier assigned by the HUGO Gene Nomenclature Committee.
- For H-InvDB, the resource identifier is the unique identifier of a cDNA cluster.
- For HAMAP, the resource identifier is the unique identifier of a HAMAP signature.
- For PDB, the resource identifier is the entry name.
- For PDBsum, the resource identifier is the PDB entry name.
- For PIRSF, the resource identifier is the protein family number.
- For OMA, the primary identifier consists of an OMA group fingerprint.
- For MaizeGDB, the resource identifier is the 'Gene-product' accession ID.
- For LegioList, Leproma or TubercuList, the resource identifier is the genome Open Reading Frame (ORF) code.
- For euHCVdb, the resource identifier is an EMBL accession number.

3.13.4. The optional information 1

The third field of the DR line, the 'OPTIONAL_INFORMATION_1', is used to provide optional information.

- For RefSeq this field is the nucleotide sequence identifier.
- For CDD, InterPro, PANTHER, Pfam, PIR, PRINTS, ProDom, REBASE, SFLD, SMART, SUPFAM or TIGRFAMs, this field is the entry name.
- For PDB, this field is the structure determination method, which is controlled vocabulary that currently includes: X-ray (for X-ray crystallography), NMR (for NMR spectroscopy), EM (for electron microscopy and cryo-electron diffraction), Fiber (for fiber diffraction), IR (for infrared spectroscopy), Model (for predicted models) and Neutron (for neutron diffraction).
- For dictyBase, EcoGene, FlyBase, CGD, HGNC, MGI, PomBase, RGD, SGD, Xenbase or ZFIN, this field is the gene designation. If the gene designation is not available, a dash ('-') is used.
- For GO, this field is a 1-letter abbreviation for one of the 3 ontology aspects, separated from the GO term by a column. If the term is longer than 46 characters, the first 43 characters are indicated followed by 3 dots ('...'). The abbreviations for the 3 distinct aspects of the ontology are P (biological Process), F (molecular Function), and C (cellular Component).
- For HAMAP, this field contains the HAMAP entry name for a protein family.
- For Ensembl, EnsemblBacteria, EnsemblFungi, EnsemblMetazoa, EnsemblPlants, EnsemblProtists, Gramene, VectorBase, WormBase and WBParaSite, this field is a protein identifier.
- For VectorBase, this field is the species of origin.
- For ESTHER and PIRSF, this field is the protein family name.
- For IntAct and BioGrid, this field indicates the number of interactors.
- For MIM, this field distinguishes between MIM "gene" and "phenotype" entries. Note that some MIM entries describe both a gene and a phenotype. In such a case, this field indicates "gene+phenotype".
- For Allergome, this field is the name of the allergen.
- For ArachnoServer, this field is the name of the toxin.
- For CAZy, this field is the CAZy family name.
- For ConoServer, this field is the name of the toxin.
- For Orphanet, this field is the name of the disease caused by defects in the protein.
- For ChiTaRS and UCSC, this field is the organism name.
- For PeroxiBase, this field is a name given by the database curators, based on organism and peroxidase classification.
- For Reactome, this field is the name of the pathway.
- For UniPathway, this field is the identifier of the reaction.
- For BRENDA and Genevisible, this field is an organism code.
- For DrugBank, this field is a drug generic name for which the protein is a target.
- For TCDB, this field is the transport classification family name.

- For ExpressionAtlas, this field is the expression patterns.
- For eggNOG, this field is the taxonomic scope.
- For TAIR, this field is the TAIR locus name (AGI number).
- For Bgee, BindingDB, BioCyc, CarbonylDB, CCDS, ChEMBL, CleanEx, COMPLUYEAST-2DPAGE, CollecTF, CORUM, CTD, DIP, DisGeNET, DisProt, DMDM, DNASU, DOSAC-COBS-2DPAGE, EchoBASE, ELM, EPD, EuPathDB, EvolutionaryTrace, Gene3D, GeneCards, GeneDB, GeneID, GeneReviews, GeneTree, GeneWiki, GenomeRNAi, GuidetoPHARMACOLOGY, HOGENOM, HOVERGEN, H-InvDB, HPA, InParanoid, IPI, iPTMnet, KEGG, LegioList, Leproma, MaizeGDB, MalaCards, MaxQB, MEROPS, MINT, mycoCLAP, nextProt, OMA, OGP, OpenTargets, OrthoDB, PATRIC, PaxDb, PDBsum, PeptideAtlas, PharmGKB, PhosphoSitePlus, PhylomeDB, PMAP-CutDB, PRIDE, PRO, ProMEX, ProteinModelPortal, PseudoCAP, REPRODUCTION-2DPAGE, SABIO-RK, SignaLink, SIGNOR, SMR, STRING, SWISS-2DPAGE, SwissLipids, SwissPalm, TAIR, TopDownProteomics, TreeFam, TubercuList, UCD-2DPAGE, UniCarbKB, UniGene and World-2DPAGE, this field is not used and a dash ('-') is displayed in that field.

3.13.5. The optional information 2

A number of DR lines possess a fourth field, the 'OPTIONAL_INFORMATION_2', which is used to provide further optional information.

- For the protein domain/family or structural databases CDD, Gene3D, HAMAP, PANTHER, Pfam, PIRSF, ProDom, SFLD, SMART, SUPFAM and TIGRFAMs, this field is the number of hits found in the sequence.
- For Ensembl, EnsemblBacteria, EnsemblFungi, EnsemblMetazoa, EnsemblPlants, EnsemblProtists, Gramene, VectorBase, WormBase and WBParaSite, this field is a gene identifier.
- For GO, this field is a 3-character GO evidence code. The GO evidence code is followed by the source database from which the cross-reference was obtained, separated by a colon. The definitions of the evidence codes are: IDA=inferred from direct assay, IMP=inferred from mutant phenotype, IGI=inferred from genetic interaction, IPI=inferred from physical interaction, IEP=inferred from expression pattern, TAS=traceable author statement, NAS=non-traceable author statement, IC=inferred by curator, ISS=inferred from sequence or structural similarity.
- For PDB, this field indicates the resolution of structures that were determined by X-ray crystallography or electron microscopy.

3.13.6. The optional information field 3

A number of DR lines possess a fifth field, the 'OPTIONAL_INFORMATION_3', which is used to provide further optional information.

- For PDB, this field indicates the chain(s) and the corresponding range, of which the structure has been determined. If the range is unknown, a dash is given rather than the range positions (e.g. 'A/B=-.'), if the chains and the range is unknown, a dash is used.
- For WormBase, this field indicates the gene designation.

3.13.7. The optional isoform sequence identifier field

Some of the resources to which we link contain information that is specific to an isoform sequence and where this is known, we indicate the corresponding UniProtKB isoform sequence identifier in our DR lines:

- Ensembl
- Gramene
- RefSeq
- UCSC
- EnsemblBacteria
- EnsemblFungi
- EnsemblMetazoa
- EnsemblPlants
- EnsemblProtists
- CCDS
- WBParaSite
- SwissLipids
- TopDownProteomics

Examples of complete DR lines are shown here:

```
DR   EMBL; U29082; AAA68403.1; -; Genomic_DNA.
```

```
DR   Allergome; 3541; Asc s 1.0101.
```

```
DR   ArachnoServer; AS000173; kappa-hexatoxin-Hv1b.
```

```
DR   Araport; AT4G08920; -.
```

```
DR   Bgee; ENSMUSG00000032315; -.
```

```
DR   BindingDB; P06709; -.
```

```
DR   BioCyc; EcoCyc:USHA-MONOMER; -.
```

```
DR   BioGrid; 69392; 1.
DR   BioMuta; TF; -.

DR   BRENDA; 3.5.99.5; 3804.

DR   CarbonylDB; Q14789; -.

DR   CAZy; GH109; Glycoside Hydrolase Family 109.

DR   CCDS; CCDS18166.1; -. [O89019-1]

DR   CDD; cd01948; EAL; 1.

DR   ChEMBL; CHEMBL4259; -.

DR   ChiTaRS; ATP6AP1; drosophila.

DR   CGD; CAL0001939; orf19.773.

DR   CleanEx; MM_KIFC2; -.

DR   CollecTF; EXPREG_00000150; -.

DR   COMPLUYEAST-2DPAGE; P41797; -.

DR   ConoServer; 2838; ArIA precursor.

DR   CORUM; P59998; -.

DR   CTD; 282646; -.

DR   dictyBase; DDB0191351; myoB.

DR   DIP; DIP:37N; -.

DR   DisGeNET; 348; -.

DR   DisProt; DP00239; -.

DR   DMDM; 44887889; -.

DR   DNASU; 1400; -.

DR   DOSAC-COBS-2DPAGE; P02774; -.

DR   DrugBank; APRD00096; Tegaserod.

DR   EchoBASE; EB4119; -.
DR   EcoGene; EG10054; araC.

DR   eggNOG; ENOG410IEUN; Eukaryota.

DR   ELM; P12931; -.

DR   Ensembl; ENST00000383775; ENSP00000373285; ENSG00000154813. [Q96FX2-2]

DR   EnsemblBacteria; EBSTAT00000032812; EBSTAP00000031682; EBSTAG00000032810.

DR   EnsemblFungi; YDR365W-B; YDR365W-B; YDR365W-B.

DR   EnsemblMetazoa; FBtr0071603; FBpp0071529; FBgn0020306.

DR   EnsemblPlants; AT1G66340.1; AT1G66340.1; AT1G66340.

DR   EnsemblProtists; DDB0305146; DDB0305146; DDB_G0286833.

DR   EPD; P00451; -.

DR   ESTHER; bacbr-grsb; Thioesterase.

DR   euHCVdb; AF271632; -.

DR   EuPathDB; GiardiaDB:GL50803_13101; -.

DR   EvolutionaryTrace; P06611; -.

DR   ExpressionAtlas; Q10Q48; baseline.

DR   FlyBase; FBgn0000055; Adh.

DR   Gene3D; 2.40.128.50; -; 2.

DR   GeneCards; LIPE; -.

DR   GeneDB; H25N7.01:pep; -.

DR   GeneID; 36674; -.

DR   UniCarbKB; P00750; -.

DR   GenomeRNAi; 37724; -.

DR   GeneReviews; MAGEL2; -.

DR   GeneTree; EMGT00050000006238; -.
DR   Genevisible; P31946; HS.

DR   GeneWiki; Dock7; -.

DR   GO; GO:0003677; F:DNA binding; IPI:SGD.

DR   Gramene; OS03T0727600-01; OS03T0727600-01; OS03G0727600.

DR   GuidetoPHARMACOLOGY; 2242; -.

DR   HGNC; HGNC:12849; YWHAB.

DR   H-InvDB; HIX0004037; -.

DR   HAMAP; MF_00120; GatA; 1.

DR   HOGENOM; HBG282443; -.

DR   HOVERGEN; HBG057182; -.

DR   HPA; CAB004311; -.

DR   IMGT_GENE-DB; IGHM; -.

DR   InParanoid; O04196; -.

DR   IntAct; P14653; 1.

DR   InterPro; IPR001650; Helicase_C.

DR   IPI; IPI00000005; -.

DR   iPTMnet; Q15796; -.

DR   KEGG; bsu:BG10490; -.

DR   KO; K09972; -.

DR   LegioList; lpp2301; -.

DR   Leproma; ML0485; -.

DR   MaizeGDB; 25342; -.

DR   MalaCards; LIPE; -.

DR   MaxQB; O94656; -.
DR   MEROPS; M41.009; -.

DR   MGI; MGI:87920; Adfp.

DR   MIM; 140050; gene.

DR   MINT; MINT-640764; -.

DR   mycoCLAP; MAN26A_PIRSP; -.

DR   neXtProt; NX_Q69383; -.

DR   OGP; P31946; -.

DR   OMA; GLCHYFS; -.

DR   OpenTargets; ENSG00000157764; -.

DR   Orphanet; 64; Alstrom syndrome.

DR   OrthoDB; EOG94QWM6; -.

DR   PANTHER; PTHR12103; HAD-IG-Ncltidse; 1.

DR   PATRIC; fig|511145.12.peg.1604; -.

DR   PaxDb; P85829; -.

DR   PDB; 1NB3; X-ray; 2.80 A; A/B/C/D=116-335, P/R/S/T=98-105.

DR   PDBsum; 1HF1; -.

DR   PeptideAtlas; P32494; -.

DR   PeroxiBase; 79; AtPrx03.

DR   Pfam; PF00017; SH2; 1.

DR   PharmGKB; PA134908332; -.

DR   PhosphoSitePlus; P01266; -.

DR   PhylomeDB; A4WFL4; -.

DR   PIR; A00682; KIPGA.

DR   PIRSF; PIRSF006414; Ftr_formyl_trnsf; 1.
DR   PMAP-CutDB; P38398; -.
DR   PomBase; SPAC1565.08; cdc48.
DR   PRIDE; P10144; -.
DR   PRINTS; PR00237; GPCRRHODOPSN.
DR   PRO; PR:O42634; -.
DR   ProDom; PD000511; Aconitase_N; 1.
DR   ProMEX; P49200; -.
DR   PROSITE; PS00107; PROTEIN_KINASE_ATP; 2.
DR   ProteinModelPortal; P84155; -.
DR   PseudoCAP; PA0892; -.
DR   Reactome; REACT_1675.1; mRNA Processing.
DR   REBASE; 993; EcoRI.
DR   RefSeq; NP_611010.4; NM_137166.4.
DR   REPRODUCTION-2DPAGE; IPI00022774; -.
DR   RGD; 70968; Ddah1.
DR   SABIO-RK; P10172; -.
DR   SFLD; SFLDS00014; RuBisCO; 1.
DR   SGD; S000000170; AAR2.
DR   SignaLink; P41935; -.
DR   SIGNOR; P00533; -.
DR   SMART; SM00369; LRR_TYP; 2.
DR   SMR; P38479; -.
DR   STRING; P23946; -.
DR   SUPFAM; SSF48371; ARM-type_fold; 1.
DR   SWISS-2DPAGE; P10599; -.
```

```
DR   SwissLipids; SLP:000000316; -.
```

```
DR   SwissPalm; Q13530; -.
```

```
DR   TAIR; locus:2064097; AT2G38610.
```

```
DR   TCDB; 1.A.1.11.11; voltage-gated ion channel (VIC) superfamily.
```

```
DR   TIGRFAMs; TIGR00630; uvra; 1.
```

```
DR   TopDownProteomics; P10599-1; -. [P10599-1]
```

```
DR   TreeFam; TF324882; -.
```

```
DR   TubercuList; Rv0001; -.
```

```
DR   UCD-2DPAGE; P02648; -.
```

```
DR   UCSC; uc001olc.4; human. [A5YM72-3]
```

```
DR   UniCarbKB; O02197; -.
```

```
DR   UniGene; Hs.505267; -.
```

```
DR   UniPathway; UPA00842; UER00808.
```

```
DR   VectorBase; AAEL004386-RA; AAEL004386-PA; AAEL004386.
```

```
DR   World-2DPAGE; 0001:P77845; -.
```

```
DR   WormBase; F26C11.2; CE01560; WBGene00006744; unc-4.
```

```
DR   WBParaSite; Bm6838; Bm6838; WBGene00227099.
```

```
DR   Xenbase; XB-FEAT-481105; hdac3.
```

```
DR   ZFIN; ZDB-GENE-980526-290; hoxb1b.
```

3.14. Cross-references to the nucleotide sequence database

The specific format for cross-references to the EMBL/GenBank/DDBJ nucleotide sequence database is:

```
DR   EMBL; ACCESSION_NUMBER; PROTEIN_ID; STATUS_IDENTIFIER; MOLECULE_TYPE.
```

where 'PROTEIN_ID' stands for the 'Protein Sequence Identifier'. It is a string which is stored, in nucleotide sequence entries, in a qualifier called '/protein_id' which is tagged to every CDS in the nucleotide database. Example from EMBL:

```
FT   CDS 302..2674
FT   /protein_id="CAA03857.1"
FT   /db_xref="SWISS-PROT:P26345"
FT   /gene="recA"

FT   /product="RecA protein"
```

The Protein Sequence Identifier (Protein_ID) consists of a stable ID portion (8 characters: 3 letters followed by 5 numbers) plus a period and a version number. The version number only changes when the protein sequence coded by the CDS changes, while the stable part remains unchanged. The Protein_ID effectively replaces what was previously known as the 'PID'.

The 'STATUS_IDENTIFIER' provides information about the relationship between the sequence in the entry and the CDS in the corresponding EMBL entry:

a) In most cases the translation of the EMBL nucleotide sequence CDS results in the same sequence as shown in the corresponding entry or the differences are mentioned in the feature (FT) lines as CONFLICT, VARIANT or VAR_SEQ and in the RP lines. The status identifier is then a dash ('-').

Example:

```
DR   EMBL; AJ297977; CAC17465.1; -; Genomic_DNA.
```

```
DR   EMBL; X56491; CAA39846.1; ALT_FRAME; mRNA.
```

b) In some cases the translation of the EMBL nucleotide sequence CDS results in a sequence different from the sequence shown in the corresponding entry. When the differences are either not mentioned in the feature (FT) lines as CONFLICT, VARIANT or VAR_SEQ (see this section) and in the RP lines, or do simply not meet the criteria for such situations, the differences are indicated as follows:

1. If the difference is due to a different start of the sequence (i.e. Swiss-Prot believes that the start of the sequence is upstream or downstream of the site annotated as the start of the sequence in the EMBL database), the status identifier shows the comment 'ALT_INIT'. Example:

```
DR   EMBL; L29151; AAA99430.1; ALT_INIT; mRNA.
```

2. If the difference is due to a different termination of the sequence (i.e. Swiss-Prot believes that the termination of the sequence is upstream or downstream of the site annotated as the end of the sequence in the EMBL database), the status identifier shows the comment 'ALT_TERM'. Example:

```
DR   EMBL; L20562; AAA26884.1; ALT_TERM; Genomic_DNA.
```

3. If the difference is due to frameshifts in the EMBL sequence, the status identifier shows the comment 'ALT_FRAME'. Example:

```
DR   EMBL; X56420; CAA39814.1; ALT_FRAME; mRNA.
```

4. If the difference is not due to any of the cases mentioned above (e.g. wrong intron-exon boundaries given in the EMBL entry) or to a mixture of the cases mentioned above, the status identifier shows the comment 'ALT_SEQ'. Example:

```
DR   EMBL; M28482; AAA26378.1; ALT_SEQ; Genomic_DNA.
```

c) In some cases the nucleotide sequence of a complete CDS is divided into exons present in different EMBL entries. We point to the exon-containing EMBL entries by citing the Protein_ID as optional information field 1 and adding the comment 'JOINED' as the status identifier. These EMBL entries do not contain a CDS feature but contain exons joined to a CDS feature which is labeled with the given Protein_ID.

Example:

```
DR   EMBL; M63397; AAA51662.1; -; Genomic_DNA.

DR   EMBL; M63395; AAA51662.1; JOINED; Genomic_DNA.

DR   EMBL; M63396; AAA51662.1; JOINED; Genomic_DNA.
```

In the above example the Swiss-Prot sequence is derived from the CDS labeled with the Protein_ID AAA51662. This CDS feature can be found in the EMBL entry M63397. Exons belonging to this CDS are not only found in EMBL entry M63397, but also in the EMBL entries M63395 and M63396.

d) In some cases there is no CDS feature key annotating a protein translation in an EMBL entry and thus no Protein_ID for the CDS. Therefore it is not possible for us to point to a Protein_ID in the optional information field 1. In these cases we point to the relevant EMBL entries by including a dash ('-') in the position of the missing Protein_ID and 'NOT_ANNOTATED_CDS' into the status identifier.

Example:

```
DR   EMBL; AJ243418; -; NOT_ANNOTATED_CDS; mRNA.
```

The 'MOLECULE_TYPE' provides information about the biological source of the molecule. The molecule type is controlled vocabulary, which corresponds to that of The International Nucleotide Sequence Database Collaboration (INSD) comprised of DDBJ (Japan), the EMBL-Bank(UK) and GenBank (USA). Relevant to the UniProt Knowledgebase are the following values:

- Genomic_DNA: any genomic DNA; includes nuclear, organelle, plasmid ...
- Genomic_RNA: genomic RNA
- Transcribed_RNA: any transcribed RNA that is not mRNA; includes unmature RNA (pre-RNA)
- mRNA: mRNA; includes cDNA
- Viral_cRNA: positive cRNA molecule from a single stranded genomic RNA
- Unassigned_DNA: unknown in vivo DNA
- Unassigned_RNA: unknown in vivo RNA
- Other_DNA: synthetic DNA
- Other_RNA: synthetic RNA
- -: unknown

3.14.1. Cross-references to the PROSITE database

The specific format for cross-references to the PROSITE protein domain and family database is:

```
DR   PROSITE; ACCESSION_NUMBER; ENTRY_NAME; NUMBER_OF_MATCHES.
```

Where 'ACCESSION_NUMBER' stands for the accession number of the PROSITE pattern or profile entry; 'ENTRY_NAME' is the name of the entry and 'NUMBER_OF_MATCHES' is the number of matches of the pattern or profile in that particular protein sequence.

Examples of PROSITE cross-references:

```
DR   PROSITE; PS00107; PROTEIN_KINASE_ATP; 2.
```

```
DR   PROSITE; PS00028; ZINC_FINGER_C2H2_1; 4.
```

3.15. The PE line

Table of contents

The PE (Protein existence) line gives indication on the evidences that we currently have for the existence of a protein. Because most protein sequences are derived from translation of nucleotide sequences and are mere predictions, the PE line indicates what the evidences are of the existence of a protein.

Note that the 'PE' line does not give information on the accuracy or correctness of the sequence displayed. While it gives information on the existence of a protein, it may happen that the sequence slightly differ, especially for sequences derived from gene model predictions from genomic sequences.

The format of the PE line is:

```
PE   Level: Evidence;
```

With the following values:

- 1: Evidence at protein level
- 2: Evidence at transcript level
- 3: Inferred from homology
- 4: Predicted
- 5: Uncertain

Example:

```
PE   1: Evidence at protein level;
```

The status 'Evidence at protein level' indicates that there is clear experimental evidence (such as a characterization paper, partial to complete Edman sequencing, clear identification by mass spectrometry (MSI), X-ray or NMR structure, detection by antibodies etc.) for the existence of this protein.

The status 'Evidence at transcript level' is used to indicate that the existence of a protein has not been strictly proved but that expression data (presence of cDNAs, RT-PCR, Northern blots) indicates the existence of a transcript.

The status 'Inferred from homology' is used to indicate that the existence of a protein is probable because clear orthologs exist in closely related species.

The status 'Predicted' is used for entries without evidence at protein, transcript, or homology levels.

The status 'Uncertain' indicates that the existence of the protein is unsure.

Criteria used to assign a PE level to entries are described in the document file pe_criteria.txt

3.16. The KW line

Table of contents

The KW (KeyWord) lines provide information that can be used to generate indexes of the sequence entries based on functional, structural, or other categories. The keywords chosen for each entry serve as a subject reference for the sequence. The document keywlist.txt lists all the keywords and a definition of their usage in the database. Often several KW lines are necessary for a single entry. The list of keywords associated to one entry can be downloaded using the following URL: http://www.uniprot.org/uniprot/?query=reviewed%3ayes+keyword%3a*&force=yes&format=tab&columns=id,keywords.

The format of the KW line is:

```
KW   Keyword[; Keyword...].
```

More than one keyword may be listed on each KW line; semicolons separate the keywords, and the last keyword is followed by a period. An entry often contains several KW lines. Keywords may consist of more than one word (they may contain blanks), but are never split between lines. The keywords are stored by alphabetical order. An example of KW lines in an entry is:

```
KW   3D-structure; Alternative splicing; Alzheimer disease; Amyloid;

KW   Apoptosis; Cell adhesion; Coated pits; Copper;

KW   Direct protein sequencing; Disease mutation; Endocytosis;

KW   Glycoprotein; Heparin-binding; Iron; Membrane; Metal-binding;

KW   Notch signaling pathway; Phosphorylation; Polymorphism;

KW   Protease inhibitor; Proteoglycan; Serine protease inhibitor; Signal;

KW   Transmembrane; Zinc.
```

TrEMBL makes use of the same controlled list of keywords as is used in Swiss-Prot but, as most keywords in an entry are added during the annotation process, TrEMBL entries generally contain fewer keywords than Swiss-Prot entries. The main sources of TrEMBL keywords are:

- The underlying nucleotide entry. The nucleotide databases add keywords and any which are also found in the UniProtKB controlled list are added to the corresponding TrEMBL entry.
- The program which creates TrEMBL entries. This adds keywords based on information in the underlying nucleotide entry. For example, if a nucleotide entry contains the word "kinase" in the description, the program adds the keyword "Kinase" to the corresponding TrEMBL entry.
- Automatic annotation.

3.17. The FT line

Table of contents

The FT (Feature Table) lines provide a precise but simple means for the annotation of the sequence data. The table describes regions or sites of interest in the sequence. In general the feature table lists posttranslational modifications, binding sites, enzyme active sites, local secondary structure or other characteristics reported in the cited references. Sequence conflicts between references are also included in the feature table.

The FT lines have a fixed format. The column numbers allocated to each of the data items within each FT line are shown in the following table (column numbers not referred to in the table are always occupied by blanks).

Columns	Data item
1-2	FT
6-13	Key name
15-20	'From' endpoint
22-27	'To' endpoint
35-75	Description

The key name and the endpoints are always on a single line, but the description may require one or more additional lines. In this event, the following line contains blanks in the columns 3-34, and the description continues from column 35 onwards as in the line above. Thus a blank key always denotes a continuation of the previous description.

An example of a feature table is shown below:

```
FT   NON_TER       1      1


FT   SIGNAL       <1     10       {ECO:0000250}.


FT   CHAIN        19     87       A-agglutinin.


FT   PROPEP       22     43       Removed by a dipeptidylpeptidase.


FT   MOD_RES      41     41       Arginine amide. {ECO:0000250}.


FT   DISULFID    110    115


FT   CARBOHYD    251    251       N-linked (GlcNAc...) asparagine.


FT   CONFLICT    327    327       E -> R (in Ref. 1; AAA52173).


FT   CONFLICT     78     78       Missing (in Ref. 1; AAD48408).
```

The first item on each FT line is the key name, which is a fixed abbreviation (of up to 8 characters) with a defined meaning. A list of the currently defined key names can be found in this section of this document.

Following the key name are the 'FROM' and 'TO' endpoint specifications. These fields designate (inclusively) the endpoints of the feature named in the key field. In general, these fields simply contain residue numbers which indicate positions in the sequence as listed. Note that these positions are always specified assuming a numbering of the listed sequence from 1 to n; this numbering is not necessarily the same as that used in the original reference(s). The following should be noted:

- If the 'FROM' and 'TO' specifications are identical, the feature involves one single amino acid;
- When a feature is known to extend beyond the position that is given in the feature table, the endpoint specification will be preceded by '<' for features which continue to the left end (N-terminal direction) or by '>' for features which continue to the right end (C- terminal direction);
- Unknown endpoints are denoted by '?'. Uncertain endpoints are denoted by a '?' before the position, e.g. '?42'.

See also the notes concerning each of the key names in this section

The remaining portion of the FT line is a description that contains additional information about the feature. For example, for a posttranslationally modified residue (key MOD_RES) the chemical nature of the modified residue is given, while for a sequence variation (key VARIANT) the nature of the variation is indicated. This portion of the line is generally in free form, and may be continued on additional lines when necessary.

3.17.1. Feature identifiers

Some features are associated with a unique and stable feature identifier (FTId), which allows to construct links directly from position-specific annotation in the feature table to specialized protein-related databases. The FTId is always the last component of a feature in the description field. The format of a feature with a feature identifier is:

```
FT   KEY_NAME     x      x        [Description.]
FT                                /FTId=XXX_number.
```

where XXX is the 3-letter code for the specific feature key, separated by an understore from a 6 to 10-digit number.

Feature identifiers currently exist for the feature keys CARBOHYD, CHAIN, PEPTIDE, PROPEP, VARIANT and VAR_SEQ . The format of the corresponding FTIds is shown in the following table:

Key name	Format of the FTId	Availability
CARBOHYD	CAR_number	Currently only for residues attached to an oligosaccharide structure annotated in the UniCarbKB database
CHAIN, PEPTIDE	PRO_number	Any mature polypeptide
PROPEP	PRO_number	Any processed propeptide
VARIANT	VAR_number	Currently only for protein sequence variants of Hominidae (great apes and humans)
VAR_SEQ	VSP_number	Any sequence with a VAR_SEQ feature

Examples of features with FTIds are given below:

```
FT   CARBOHYD    251    251       N-linked (GlcNAc...) asparagine.
FT                                /FTId=CAR_000070.


FT   CHAIN        23    611       Halfway protein.
FT                                /FTId=PRO_0000021413.


FT   PEPTIDE      20     57       Histatin 1.
FT                                /FTId=PRO_0000021416.


FT   PROPEP       25     48
FT                                /FTId=PRO_0000021449.


FT   VARIANT     214    214       V -> I.
FT                                /FTId=VAR_009122.


FT   VAR_SEQ      33     83       TPDINPAWYTGRGIRPVGRFGRRRATPRDVTGLGQLSCLPL
FT                                DGRTKFSQRG -> SECLTYGKQPLTSFHPFTSQMPP (in
FT                                isoform 2).
FT                                /FTId=VSP_004370.
```

INIT_MET - Initiator methionine.

This feature key is associated with a '1' value in the 'FROM' and 'TO' fields to indicate that the initiator methionine has been cleaved off:

```
FT   INIT_MET      1      1       Removed.
```

It is not used when the initiator methionine is not cleaved off

SIGNAL - Extent of a signal sequence (prepeptide).

```
FT   SIGNAL        1     26
```

```
FT   SIGNAL        1     23       {ECO:0000255}.
```

PROPEP - Extent of a propeptide.

Examples of PROPEP key feature lines:

```
FT   PROPEP       27     28       Activation peptide.
```

```
FT   PROPEP      550    574       Removed in mature form.
```

TRANSIT - Extent of a transit peptide (mitochondrion, chloroplast, thylakoid, cyanelle, peroxisome etc.).

Examples of TRANSIT key feature lines:

```
FT   TRANSIT       1     42       Chloroplast.
```

```
FT   TRANSIT       1     65       Cyanelle.
```

```
FT   TRANSIT       1     25       Mitochondrion. {ECO:0000250}.
```

```
FT   TRANSIT       1     34       Glyoxysome. {ECO:0000255}.
```

```
FT   TRANSIT       1      ?       Chloroplast. {ECO:0000255}.
FT   TRANSIT       ?     77       Thylakoid. {ECO:0000269|PubMed:11719511,
FT                                ECO:0000269|PubMed:11826309}.
```

CHAIN - Extent of a polypeptide chain in the mature protein.

- In Swiss-Prot, it is present:
  1. For proteins that are not processed (those for which the mature protein sequence corresponds to the translated cDNA sequence) the CHAIN covers the whole protein sequence:

     ```
     FT   CHAIN         1    333      COP9 signalosome complex subunit 5.
     ```

  2. For processed proteins, it describes the mature part of the protein once preprotein parts such as propeptides and signal sequences have been removed:

     ```
     FT   CHAIN        21    119       Beta-2-microglobulin.
     ```

     ```
     FT   CHAIN        41    180       Factor X light chain.
     ```

- For TrEMBL, it is present only in those entries where the submitters to the nucleotide sequence databases have provided this information for processed proteins.

PEPTIDE - Extent of a released active peptide.

Examples of PEPTIDE key feature lines:

```
FT   PEPTIDE       1      9       Arg-vasopressin.
```

```
FT   PEPTIDE     235    239       Met-enkephalin.
```

TOPO_DOM - Topological domain.

Examples:

```
FT   TOPO_DOM    337    371       Cytoplasmic.
```

```
FT   TOPO_DOM     17    471       Mitochondrial matrix. {ECO:0000255}.
```

TRANSMEM - Extent of a transmembrane region.

Examples:

```
FT   TRANSMEM    241    264       Helical.
```

```
FT   TRANSMEM    184    204       Helical. {ECO:0000255}.
```

```
FT   TRANSMEM    171    191       Helical; Name=3. {ECO:0000255}.
```

```
FT   TRANSMEM   1663   1685       Helical; Name=S5 of repeat IV.
FT                                {ECO:0000255}.
```

```
FT   TRANSMEM    490    516       Helical; Anchor for type IV membrane
FT                                protein. {ECO:0000250}.
```

```
FT   TRANSMEM     60     69       Beta stranded.
```

INTRAMEM - Extent of a region located in a membrane without crossing it.

Examples:

```
FT   INTRAMEM    306    320
```

```
FT   INTRAMEM   1145   1165       {ECO:0000255}.
```

```
FT   TRANSMEM    534    551       Helical. {ECO:0000250}.
```

```
FT   INTRAMEM    260    282       Pore-forming; Name=P region.
FT                                {ECO:0000255}.
```

```
FT   INTRAMEM    500    514       Helical. {ECO:0000250}.
FT   INTRAMEM    515    517       Note=Loop between two helices.
FT                                {ECO:0000250}.
FT   INTRAMEM    518    529       Helical. {ECO:0000250}.
FT   INTRAMEM    530    534       Note=Loop between two helices.
FT                                {ECO:0000250}.
```

```
FT   INTRAMEM    412    432       Pore-forming; Name=Segment H5.
FT                                {ECO:0000255}.
```

DOMAIN - Extent of a domain, which is defined as a specific combination of secondary structures organized into a characteristic three-dimensional structure or fold.

The domain type is given in the description field. If there exist several copies of a domain, the domains are numbered. Examples of DOMAIN key feature lines:

```
FT   DOMAIN       27    128       Cadherin 1.


FT   DOMAIN      745    939       Ras-GAP.


FT   DOMAIN      746    805       SH3.
```

REPEAT - Extent of an internal sequence repetition.

Examples of REPEAT key feature lines:

```
FT   REPEAT      225    307       1.
FT   REPEAT      341    423       2.
FT   REPEAT      455    537       3; approximate.
```

CA_BIND - Extent of a calcium-binding region.

Example:

```
FT   CA_BIND      20     31       1.
```

ZN_FING - Extent of a zinc finger region.

The zinc finger 'category' is indicated in the description field. Examples of ZN_FING key feature lines:

```
FT   ZN_FING     803    827       GATA-type.


FT   ZN_FING     559    579       NR C4-type.
```

DNA_BIND - Extent of a DNA-binding region.

The nature of the DNA-binding region is given in the description field. Examples of DNA_BIND key feature lines:

```
FT   DNA_BIND    335    415       ETS.


FT   DNA_BIND     69    128       Homeobox.


FT   DNA_BIND     16     67       Myb 2.


FT   DNA_BIND    135    200       TEA.
```

NP_BIND - Extent of a nucleotide phosphate-binding region.

The nature of the nucleotide phosphate is indicated in the description field. Examples of NP_BIND key feature lines:

```
FT   NP_BIND     182    189       ATP.


FT   NP_BIND      38     45       GTP. {ECO:0000255}.


FT   NP_BIND      35     42       FAD.
```

REGION - Extent of a region of interest in the sequence.

Examples:

```
FT   REGION      620    631       Zymogen activation region.
FT   REGION       52     84       Hydrophobic.
```

COILED - Extent of a coiled-coil region.

Example:

```
FT   COILED      258    553       {ECO:0000255}.
```

MOTIF - Short (up to 20 amino acids) sequence motif of biological interest.

Examples:

```
FT   MOTIF        94    101       Nuclear localization signal.
```

```
FT   MOTIF       648    650       Microbody targeting signal.
FT                                {ECO:0000255}.
```

COMPBIAS - Extent of a compositionally biased region.

Examples:

```
FT   COMPBIAS    219    222       Poly-Ser.
```

```
FT   COMPBIAS     22    124       Glu-rich (acidic).
```

ACT_SITE - Amino acid(s) involved in the activity of an enzyme.

Examples of ACT_SITE key feature lines:

```
FT   ACT_SITE    238    238       Proton acceptor.
```

```
FT   ACT_SITE   1083   1083       Charge relay system; for serine protease
FT                                NS3 activity. {ECO:0000250}.
```

METAL - Binding site for a metal ion.

The description field indicates the nature of the metal. Examples of METAL key feature lines:

```
FT   METAL        52     52       Iron (heme axial ligand).
```

```
FT   METAL        28     28       Copper. {ECO:0000255}.
```

BINDING - Binding site for any chemical group (co-enzyme, prosthetic group, etc.).

The chemical nature of the group is given in the description field. Examples of BINDING key feature lines:

```
FT   BINDING      14     14       Heme (covalent).
```

```
FT   BINDING     132    132       Chloride.
```

SITE - Any interesting single amino-acid site on the sequence, that is not defined by another feature key. It can also apply to an amino acid bond which is represented by the positions of the two flanking amino acids.

Example:

```
FT   SITE        391    392       Cleavage; by thrombin.
```

NON_STD - Non-standard amino acid

This key describes the occurrence of non-standard amino acids selenocysteine and pyrrolysine. Note that we also display them in the sequence data section and use the IUPAC/IUBMB recommended one-letter codes 'U' for selenocysteine and 'O' for pyrrolysine

```
FT   NON_STD      52     52       Selenocysteine.
```

```
FT   NON_STD     356    356       Pyrrolysine. {ECO:0000250}.
```

MOD_RES - Posttranslational modification of a residue.

The chemical nature of the modified residue is given in the description. The general format of the MOD_RES description field is:

```
FT   MOD_RES     xxx    xxx       Name of the modified amino acid (comment).
```

The nature of the posttranslationally formed amino acid is annotated by using a controlled vocabulary; the currently defined list of controlled vocabulary, as well as other information such as the target amino acid, the related keyword, the species where the modification is annotated and the location of the mature protein are available in ptmlist.txt document. Links to example proteins and to the RESID database are also provided to help gain a better insight into every modification.

Modification	Description
ACETYLATION	N-terminal of some residues and side chain of lysine
AMIDATION	Generally at the C-terminal of a mature active peptide after oxidative cleavage of last glycine
BLOCKED	Unidentified N- or C-terminal blocking group
FORMYLATION	Generally of the N-terminal methionine
GAMMA-CARBOXYGLUTAMIC ACID	Of glutamate
HYDROXYLATION	Generally of asparagine, aspartate, proline or lysine
METHYLATION	Generally of N-terminal phenylalanine, side chain of lysine, arginine, histidine, asparagine or glutamate, and C-terminal cysteine
PHOSPHORYLATION	Of serine, threonine, tyrosine, aspartate, histidine or cysteine, and, more rarely, of arginine
PYRROLIDONE CARBOXYLIC ACID	N-terminal glutamine which has formed an internal cyclic lactam. This is also called 'pyro-Glu'. Very rarely, pyro-Glu can be produced by modification of a N-terminal glutamate
SULFATION	Of tyrosine, serine or threonine

Examples of MOD_RES key feature lines:

```
FT   MOD_RES       1      1       N-acetylalanine.
```

```
FT   MOD_RES      58     58       Methionine amide.
```

```
FT   MOD_RES      52     52       N6-methyllysine. {ECO:0000255}.
```

```
FT   MOD_RES     198    198       Phosphoserine; by CK2.
```

```
FT   MOD_RES     367    367       Sulfotyrosine. {ECO:0000250}.
```

Example of a feature where the identity of the amino acid is unknown (an X is shown at this position in the sequence):

```
FT   MOD_RES       1      1       Blocked amino end (Xaa).
```

LIPID - Covalent binding of a lipid moiety

The chemical nature of the bound lipid moiety is given in the description. The general format of the LIPID description field is:

```
FT   LIPID       xxx    xxx       Name of the modified amino acid.
```

The attached groups that are currently defined are listed below.

Attached group	Description
myristate	Myristate group attached through an amide bond to the N-terminal glycine residue of the mature form of a protein [1,2] or to an internal lysine residue. The myristate can also be attached through a thioester bond to an internal cysteine
palmitate	Palmitate group attached through an amide bond to the N-terminal cysteine of the mature form of a protein, or to an internal lysine. The palmitate can also be attached through a thioester bond to an internal cysteine or through an ester bond to a serine or threonine residue [1,2]
farnesyl	Farnesyl group attached through a thioether bond to a cysteine residue [3,4]
geranyl-	Geranyl-geranyl group attached through a thioether bond to a cysteine residue [3,4]

geranyl	
GPI-anchor	Glycosyl-phosphatidylinositol (GPI) group linked to the alpha-carboxyl group of the C-terminal residue of the mature form of a protein [5,6]
diacylglyceride	Glyceryl group bearing two ester-linked fatty acids attached through a thioether bond to the N-terminal cysteine of the mature form of a prokaryotic lipoprotein [7]
archaeol	Archaeol (2,3-di-O-phytanyl-sn-glycerol) lipid group attached through an thioether bond to the N-terminal cysteine of the mature form of a archaeal lipoprotein [8]
n-octanoate	n-octanoate group linked through an ester bond to a serine residue
cholesterol	Cholesterol group attached through an ester bond to the C-terminal glycine of the mature form of a protein

References:

1. Grand R.J.A.
   Biochem. J. 258:626-638(1989).
2. McIlhinney R.A.J.
   Trends Biochem. Sci. 15:387-391(1990).
3. Glomset J.A., Gelb M.H., Farnsworth C.C.
   Trends Biochem. Sci. 15:139-142(1990).
4. Sinensky M., Lutz R.J.
   BioEssays 14:25-31(1992).
5. Low M.G.
   FASEB J. 3:1600-1608(1989).
6. Low M.G.
   Biochim. Biophys. Acta 988:427-454(1989).
7. Hayashi S., Wu H.C.
   J. Bioenerg. Biomembr. 22:451-471(1990).
8. Nishihara M., Utagawa M., Akutsu H., Koga Y.
   J. Biol. Chem. 267:12432-12435(1992).

Examples of LIPID key feature lines:

```
FT   LIPID         2      2       N-myristoyl glycine; by host.
```

```
FT   LIPID         5      5       S-palmitoyl cysteine.
```

```
FT   LIPID       231    231       GPI-anchor amidated serine.
```

```
FT   LIPID        21     21       S-diacylglycerol cysteine.
```

The nature of the posttranslationally formed amino acid is annotated by using a controlled vocabulary; the currently defined list of controlled vocabulary, as well as other information such as the target amino acid, the related keyword, the species where the modification is annotated and the location of the mature protein are available in ptmlist.txt document. Links to example proteins and to the RESID database are also provided to help gain a better insight into every modification.

CARBOHYD - Glycosylation site.

This key describes the occurrence of the attachment of a glycan (mono- or polysaccharide) to a residue of the protein:

- The type of linkage (C-, N-, O- or S-linked) to the protein is indicated.
- If the nature of the reducing terminal sugar is known, its abbreviation is shown between parenthesis. If three dots '...' follow the abbreviation, this indicates an extension of the carbohydrate chain. Conversely, the absence of dots means that a monosaccharide is linked.

Examples of CARBOHYD key feature lines:

```
FT   CARBOHYD     53     53       N-linked (GlcNAc...) asparagine.
FT                                {ECO:0000269|PubMed:19159218}.
```

```
FT   CARBOHYD   1486   1486       O-linked (GlcNAc) threonine.
FT                                {ECO:0000269|PubMed:1629228}.
FT                                /FTId=CAR_000155.
```

```
FT   CARBOHYD    195    195       O-linked (Glc...) tyrosine.
FT                                {ECO:0000250|UniProtKB:P13280}.
```

```
FT   CARBOHYD     41     41       S-linked (Glc) cysteine.
FT                                {ECO:0000269|PubMed:21196935,
FT                                ECO:0000269|PubMed:21251913}.
```

```
FT   CARBOHYD     29     29       C-linked (Man) tryptophan.
FT                                {ECO:0000269|PubMed:10551839}.
FT   CARBOHYD     32     32       C-linked (Man) tryptophan; partial.
FT                                {ECO:0000269|PubMed:10551839}.
```

```
FT   CARBOHYD    152    152       O-linked (Ara...) hydroxyproline.
FT                                {ECO:0000305|PubMed:666730}.
```

The nature of the glycosylated amino acid is annotated by using a controlled vocabulary, which also includes information such as the target amino acid, the related keyword, the species where the modification is annotated and the location of the mature protein and is available in ptmlist.txt document. Links to example proteins and to the RESID database are also provided to help gain a better insight into every modification.

DISULFID - Disulfide bond.

The 'FROM' and 'TO' endpoints designate the two residues which are linked by an intra-chain disulfide bond. If the 'FROM' and 'TO' endpoints are identical, the disulfide bond is an interchain one and the description field indicates the nature of the cross-link. Examples of DISULFID key feature lines:

```
FT   DISULFID     23     84       {ECO:0000305}.
```

```
FT   DISULFID     29     29       Interchain (with C-8 in small chain).
```

CROSSLNK - Posttranslationally formed amino acid bonds.

The 'FROM' and 'TO' endpoints designate the two residues, which are linked by an intrachain bond, and the description field indicates the nature of the cross-link. If the 'FROM' and 'TO' endpoints are identical, the amino acid bond is an interchain one. The name of the linked peptide is indicated in the description field. Examples of 'CROSSLNK' key feature lines:

```
FT   CROSSLNK   1010   1013       Isoglutamyl cysteine thioester (Cys-Gln).
```

```
FT   CROSSLNK     60     77       Beta-methyllanthionine (Cys-Thr).
```

```
FT   CROSSLNK     63     73       Lanthionine (Ser-Cys).
```

```
FT   CROSSLNK     64     70       Beta-methyllanthionine (Cys-Thr).
```

```
FT   CROSSLNK     65     78       Lysinoalanine (Ser-Lys).
```

The nature of the posttranslationally formed amino acid bond is annotated by using a controlled vocabulary; the currently defined list of controlled vocabulary, as well as other information such as the target amino acid, the related keyword, the species where the modification is annotated and the location of the mature protein are available in ptmlist.txt document. Links to example proteins and to the RESID database are also provided to help gain a better insight into every modification.

VAR_SEQ - Description of sequence variants produced by alternative splicing, alternative promoter usage, alternative initiation and ribosomal frameshifting.

Examples of VAR_SEQ key feature lines:

```
FT   VAR_SEQ     653    672       VATSNPGKCLSFTNSTFTFT -> ALVSHHCPVEAVRAVHP
FT                                TRL (in isoform 2).
FT                                /FTId=VSP_003786.
FT   VAR_SEQ     673    913       Missing (in isoform 2).
FT                                /FTId=VSP_003787.
```

VARIANT - Authors report that sequence variants exist.

Examples of VARIANT key feature lines:

```
FT   VARIANT     214    214       V -> I.
FT                                /FTId=VAR_009122.


FT   VARIANT     237    237       I -> L (in strain: B293, Z3524, Z3910,
FT                                Z3915 and Z3918).


FT   VARIANT      21     22       Missing (in 35% of the chains).
FT                                /FTId=VAR_006353.


FT   VARIANT     265    344       Missing (in allele D4.2).
FT                                /FTId=VAR_003465.


FT   VARIANT     156    156       I -> V (in RNA edited version).
FT                                /FTId=VAR_010166.
```

Explicit links are present in protein sequence entries of Hominidae to the Single Nucleotide Polymorphism database (dbSNP) (Nucleic Acids Res. 29:308-311(2001); PMID: 11125122). The dbSNP identifiers in human FT VARIANT lines are prefixed by "rs". NCBI/dbSNP has rs and ss numbers, but we only refer to SNPs with rs numbers. The format of such links is:

```
FT   VARIANT    from     to       Description (in dbSNP:rsaccession_number).
FT                                /FTId=VAR_number.
```

Example of a feature with a link to dbSNP:

```
FT   VARIANT     246    246       S -> A (in dbSNP:rs2228541).
FT                                /FTId=VAR_024350.
```

MUTAGEN - Site which has been experimentally altered by mutagenesis.

Examples of MUTAGEN key feature lines:

```
FT   MUTAGEN     119    119       C->R,E,A: Loss of cADPr hydrolase and
FT                                ADP-ribosyl cyclase activity.


FT   MUTAGEN     169    177       Missing: Abolishes ATP-binding.
```

UNSURE - Uncertainties in the sequence

Used to describe region(s) of a sequence for which the authors are unsure about the sequence assignment.

```
FT   UNSURE       12     12       V or Y.
```

CONFLICT - Different sources report differing sequences.

Examples of CONFLICT key feature lines:

```
FT   CONFLICT    484    484       Missing (in Ref. 2).


FT   CONFLICT    802    802       K -> Q (in Ref. 4, 5 and 10).


FT   CONFLICT      6     10       GSDSE -> RIRLR (in Ref. 2).


FT   CONFLICT    405    405       V -> A (in Ref. 6; AAF71067).
```

NON_CONS - Non-consecutive residues.

Indicates that two residues in a sequence are not consecutive and that there are a number of unsequenced residues between them. Example of a NON_CONS key feature line:

```
FT   NON_CONS   1683   1684
```

NON_TER - The residue at an extremity of the sequence is not the terminal residue.

If applied to position 1, this means that the first position is not the N-terminus of the complete molecule. If applied to the last position, it means that this position is not the C-terminus of the complete molecule. There is no description field for this key. Examples of NON_TER key feature lines:

```
FT   NON_TER       1      1
```

```
FT   NON_TER      29     29
```

Secondary structure (HELIX, STRAND, TURN) - The feature table of sequence entries of proteins whose tertiary structure is known experimentally contains the secondary structure information corresponding to that protein. The secondary structure assignment is made according to DSSP (see Kabsch W., Sander C.; Biopolymers, 22:2577-2637(1983)) and the information is extracted from the coordinate data sets of the Protein Data Bank (PDB).

In the feature table only three types of secondary structure are specified: helices (key HELIX), beta-strands (key STRAND) and turns (key TURN). Residues not specified in one of these classes are in a 'loop' or 'random-coil' structure. Because the DSSP assignment has more than the three common secondary structure classes, we have converted the different DSSP assignments to HELIX, STRAND and TURN as shown in the table below.

DSSP code	DSSP definition	Swiss-Prot assignment
H	Alpha-helix	HELIX
G	3(10) helix	HELIX
I	Pi-helix	HELIX
E	Hydrogen-bonded beta-strand (extended strand)	STRAND
B	Residue in an isolated beta-bridge	STRAND
T	H-bonded turn (3-turn, 4-turn or 5-turn)	TURN
S	Bend (five-residue bend centered at residue i)	Not specified

One should be aware of the following facts:

a. Segment length. For helices (alpha and 3-10), the residue just before and just after the helix as given by DSSP participates in the helical hydrogen-bonding pattern with a single H-bond. For practical purposes, one can extend the HELIX range by one residue on each side, e.g. HELIX 25-35 instead of HELIX 26-34. Also, the ends of secondary structure segments are less well defined for lower-resolution structures. A fluctuation of one residue is common.
b. Missing segments. In low-resolution structures, badly formed helices or strands may be omitted in the DSSP definition.
c. Special helices and strands. Helices of length three are 3-10 helices, those of length four and longer are either alpha-helices or 3-10 helices (pi helices are extremely rare). A strand of one residue corresponds to a residue in an isolated beta-bridge. Such bridges can be structurally important.
d. Missing secondary structure. No secondary structure is currently given in the feature table in the following cases:
    - No sequence data in the PDB entry;
    - Structure for which only C-alpha coordinates are in PDB;
    - NMR structure with more than one coordinate data set;
    - Model (i.e. theoretical) structure.

Examples:

```
FT   HELIX         2     11
FT   HELIX        12     14
FT   HELIX        21     35
FT   TURN         36     36
FT   TURN         46     47
FT   HELIX        49     59
FT   TURN         60     63
FT   HELIX        66     84
```

3.18. The SQ line

Table of contents

The SQ (SeQuence header) line marks the beginning of the sequence data and gives a quick summary of its content.

The format of the SQ line is:

```
SQ   SEQUENCE XXXX AA; XXXXX MW; XXXXXXXXXXXXXXXX CRC64;
```

The line contains the length of the sequence in amino acids ('AA') followed by the molecular weight ('MW') rounded to the nearest mass unit (Dalton) and the sequence 64-bit CRC (Cyclic Redundancy Check) value ('CRC64').

The molecular weight (MW) indicated on the SQ line is not meant to be that of the mature processed and posttranslationally modified protein. The MW of the mature protein can be predicted by the program Compute pI/MW tool on the ExPASy server.

The algorithm to compute the CRC64 is described in the ISO 3309 standard. The generator polynomial is x64 + x4 + x3 + x + 1.
Reference: Press W.H., Flannery B.P., Teukolsky S.A. and Vetterling W.T. "Numerical recipes in C", 2nd ed., pp896-902, Cambridge University Press (1993). (See http://library.lanl.gov/numerical/)

It should be noted that while, in theory, two different sequences could have the same CRC64 value, the likelihood that this would happen is extremely low.

An example of an SQ line is shown here:

```
SQ   SEQUENCE   486 AA;  55639 MW;  D7862E867AD74383 CRC64;
```

The information in the SQ line can be used as a check on accuracy or for statistical purposes. The word 'SEQUENCE' is present solely for readability.

3.19. The sequence data line

Table of contents

The sequence data line has a line code consisting of two blanks rather than the two-letter codes used until now. The sequence counts 60 amino acids per line, in groups of 10 amino acids, beginning in position 6 of the line.

The characters used for the amino acids are the standard IUPAC one letter codes (see Amino-acid codes section).

An example of sequence data preceded by the corresponding SQ line is shown here:

```
SQ   SEQUENCE   97 AA;  9110 MW;  E3C20C259858B830 CRC64;
     MTILASICKL GNTKSTSSSI GSSYSSAVSF GSNSVSCGEC GGDGPSFPNA SPRTGVKAGV
     NVDGLLGAIG KTVNGMLISP NGGGGGMGMG GGSCGCI
```

3.20. The // line

Table of contents

The // (terminator) line contains no data or comments and designates the end of an entry.

Appendix A : Amino-acid codes

Table of contents

The one-letter and three-letter codes for amino acids used in the knowledgebase are those adopted by the commission on Biochemical Nomenclature of the IUPAC-IUB (see the reference listed below).

One-letter code	Three-letter code	Amino-acid name
A	Ala	Alanine
R	Arg	Arginine
N	Asn	Asparagine
D	Asp	Aspartic acid
C	Cys	Cysteine
Q	Gln	Glutamine
E	Glu	Glutamic acid
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
L	Leu	Leucine
K	Lys	Lysine
M	Met	Methionine
F	Phe	Phenylalanine
P	Pro	Proline
S	Ser	Serine
T	Thr	Threonine
W	Trp	Tryptophan
Y	Tyr	Tyrosine
V	Val	Valine
O	Pyl	Pyrrolysine
U	Sec	Selenocysteine
B	Asx	Aspartic acid or Asparagine
Z	Glx	Glutamic acid or Glutamine
X	Xaa	Any amino acid

Reference

IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN).
Nomenclature and Symbolism for Amino Acids and Peptides. Recommendations 1983.

Eur. J. Biochem. 138:9-37(1984).

See also: http://www.chem.qmw.ac.uk/iupac/AminoAcid/

Appendix B : Format differences between the Swiss-Prot and EMBL databases

Table of contents

B.1 Generalities

The format of Swiss-Prot follows as closely as possible that of the EMBL database. The general structure of an entry is identical in both databases. The status are the same except though Swiss-Prot does not make use of the data classes and uses the 'reviewed' status instead. One line type used in Swiss-Prot does not exist in the EMBL database (see this section); conversely Swiss-Prot does not currently make use of every EMBL line type (see this section).

B.2 Differences in line types present in both databases

B.2.1 The ID line (IDentification)

Differences with the EMBL database ID line format are:

- The entry name can be up to 10 characters long (instead of 9 in EMBL) and can begin with a numerical character;
- EMBL entry ID lines have an additional three-letter taxonomic division 'token' inserted between the status and the molecule type;
- The molecule type is not listed;
- The length of the molecule is followed by 'AA' (Amino Acid) instead of 'BP' (Base Pairs).

B.2.2 The AC line (ACcession number)

The format of this line type is identical to that defined by the EMBL database. Swiss-Prot and TrEMBL accession numbers do not overlap those used in the EMBL/GenBank/DDBJ nucleotide sequence database. However, it should be noted that there are differences in the format of the accession numbers themselves. In Swiss-Prot and TrEMBL accession numbers consist of 6 alphanumerical characters in the following format:

1	2	3	4	5	6
[A-N,R-Z]	[0-9]	[A-Z]	[A-Z, 0-9]	[A-Z, 0-9]	[0-9]
[O,P,Q]	[0-9]	[A-Z, 0-9]	[A-Z, 0-9]	[A-Z, 0-9]	[0-9]

Examples: P01234; Q1AA12.

In EMBL, two different types of accession numbers co-exist:

1. Accession numbers with 6 alphanumerical characters, where the first character is any letter with the exception of O,P or Q and the five other characters are numbers (example: M23765);
2. Accession numbers with 8 alphanumerical characters, where the first two characters are letters and the following six characters are numbers (example: AB001084).

B.2.3 The DT line (DaTe)

Differences with the EMBL database DT line format are:

- In EMBL there are two DT lines per entry whereas there are three in Swiss-Prot;
- In EMBL the format of the DT line that indicates when an entry was created is identical to that defined in Swiss-Prot; but the two DT lines that convey information relevant to the updating of an entry are replaced by a single line in EMBL. This is shown in the example below.

DT lines in a Swiss-Prot entry:

```
DT   01-OCT-1989, integrated into UniProtKB/Swiss-Prot.
DT   01-OCT-1989, sequence version 1.
DT   07-FEB-2006, entry version 76.
```

DT lines in an EMBL database entry:

```
DT   10-MAR-1990 (Rel. 22, Created)
DT   12-APR-1990 (Rel. 23, Last updated, Version 3)
```

B.2.4 The DE line (DEscription)

- In the UniProt Knowledgebase the species of origin is not included in the description;
- In EMBL the last DE line is not terminated by a period.

B.2.5 The OS line (Organism Species)

- In EMBL the last OS line is not terminated by a period.

B.2.6 The OG line (OrGanelle)

- EMBL makes a distinction between 'Mitochondrion', and 'Kinetoplast', while Swiss-Prot does not use the latter designation;
- EMBL makes a distinction between 'Chloroplast' and 'Plastid', while Swiss-Prot does not use the latter designation;
- In EMBL the OG line is not terminated by a period.

B.2.7 The RP and RC lines

- In EMBL, unlike Swiss-Prot, the RC line precedes the RP line;
- In EMBL the RC line is in free format and is generally not used.

B.2.8 The RT line (Reference Title)

In EMBL the reference title is not terminated by a period, a question mark or an exclamation mark.

B.2.9 The FT line (Feature Table)

The format of this line is totally different from that currently defined for the EMBL database. The format used in Swiss-Prot is similar to that which was used in older versions of the EMBL database, prior to the introduction of the common EMBL/GenBank/DDBJ feature table.

B.2.10 The CC line (Comment)

The comment lines, which are free text and can appear anywhere in an EMBL entry, are grouped together in the Swiss-Prot database. They are always listed below the last reference line, and follow a precise syntax (see section 3.22).

B.2.11 The SQ line (SeQuence header)

Although the rough format and purpose of this line type is conserved, its exact content differs from that of the EMBL database. The numerical length of the sequence is listed, followed by 'AA' (Amino Acid) instead of 'BP' (Base Pairs). Rather then indicating the sequence composition which, for protein sequences, would not fit in a single line, the molecular weight and the 64-bit CRC (Cyclic Redundancy Check) value of the sequence are indicated.

B.3 Line types defined by Swiss-Prot but currently not used by EMBL

Presently, there are two line types in Swiss-Prot, which are not used in the EMBL database: the GN and OX lines.

B.4 Line types defined by EMBL but currently not used by Swiss-Prot

There are three line types in the EMBL database, which are not used in Swiss-Prot:

- FH and XX. The FH and XX lines contain no data and are present in EMBL only to improve readability of an entry when it is printed or displayed on a terminal screen. These lines are not included in Swiss-Prot so as to keep it as compact as possible and thereby facilitate its use on small computer systems.
- SV. The SV (Sequence Version) line contains an identifier specific to nucleic acid sequences. It has no meaning in the context of Swiss-Prot.

Appendix C : Documentation files

Table of contents

The knowledgebase is distributed with a large number of documentation files, which are available on the web and ftp sites.

Appendix D : UniProt Knowledgebase

Table of contents

The UniProt Knowledgebase is a non-redundant and complete protein sequence database consisting of two components:

1. Swiss-Prot
2. TrEMBL

Every four weeks two files are completely rebuilt and made available on our FTP server. These files are named: uniprot_sprot.dat.gz and uniprot_trembl.dat.gz. As indicated by their '.gz' extension, these are gzip-compressed files which, when decompressed, produce ASCII files in Swiss-Prot format.

The same data is available in fasta and xml format. The fasta files are useful for building the databases used by FASTA, BLAST and other sequence similarity search programs. Please do not use these files for any other purpose, as you will lose all annotations by using this stripped-down format.

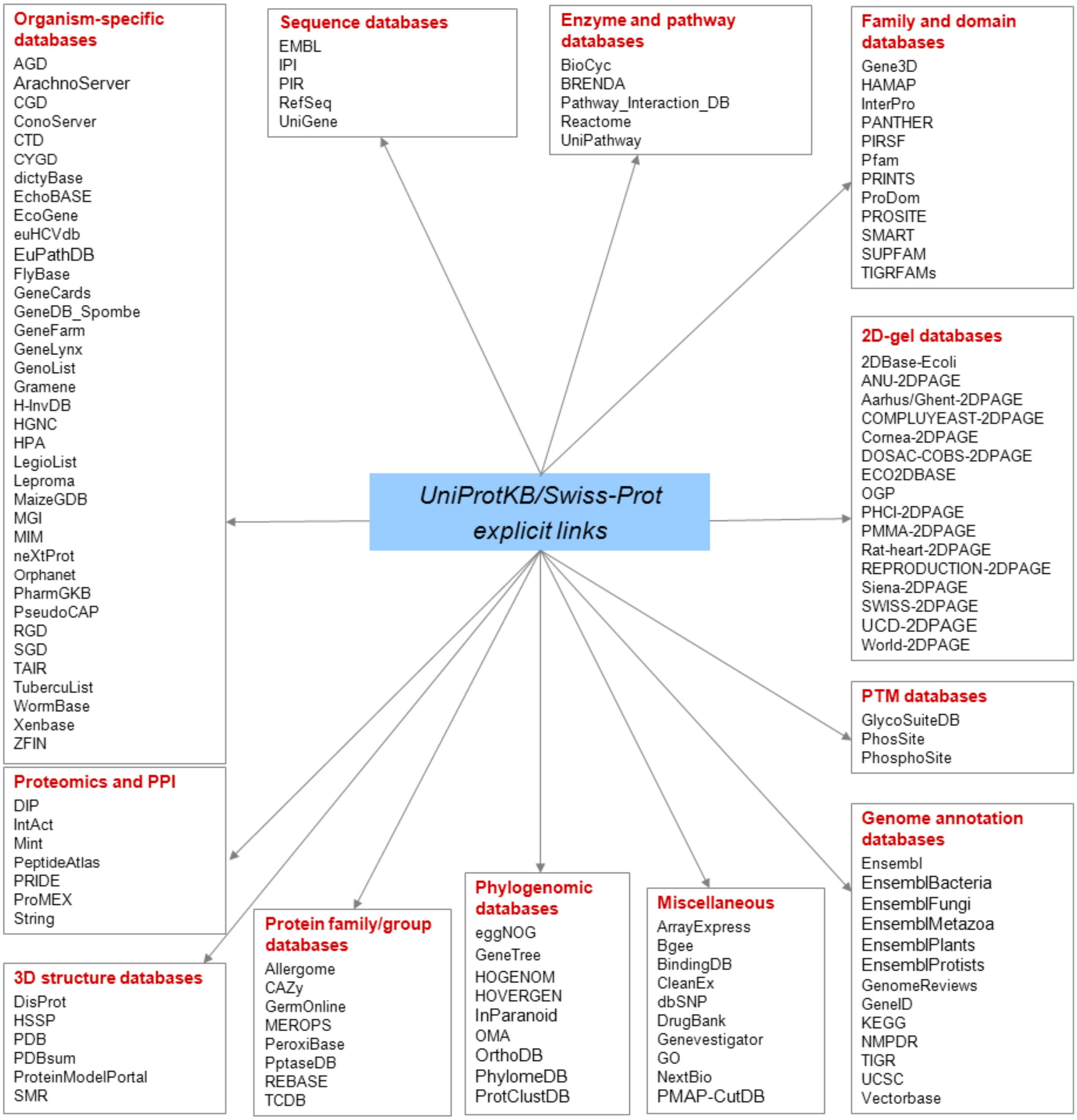
Additional notes:

- The Swiss-Prot file continuously grows as new annotated sequences are added.
- The TrEMBL file slightly decreases in size as sequences are moved out of TrEMBL after being annotated and moved into Swiss-Prot. However, at the same time, TrEMBL increases on a much larger scale by the new coding sequences submitted to the nucleotide sequence databases.
- Swiss-Prot and TrEMBL share the same system of accession numbers. Therefore you will not find any primary accession number duplicated between the two sections. A TrEMBL entry (and its associated accession number(s)) can either move to Swiss-Prot as a new entry or be merged with an existing Swiss-Prot entry. In the latter case, the accession number(s) of that TrEMBL entry are added to that of the Swiss-Prot entry.
- While these two files allow you to build what we call a 'non-redundant' database, the UniProt Knowledgebase, it must be noted that this is not completely a true statement. Without going into a long explanation we can say that this is currently the best attempt in providing a complete selection of protein sequence entries while trying to eliminate redundancies (2 or more entries for the same gene of a species). While Swiss-Prot is non-redundant, TrEMBL is far from being non-redundant and the addition of Swiss-Prot + TrEMBL is even less so.
- To describe to your users the version of the UniProt Knowledgebase that you are providing them with, you should use a statement of the form "UniProtKB release 2011_07 of Jun 28, 2011".

Appendix E : Relationships between Swiss-Prot and some biomolecular databases

Table of contents

The current status of the relationships (cross-references) between Swiss-Prot and some biomolecular databases is shown in the following schema:



UniProtKB/Swiss-Prot explicit links

Organism-specific databases

AGD
ArachnoServer
CGD
ConoServer
CTD
CYGD
dictyBase
EchoBASE
EcoGene
euHCVdb
EuPathDB
FlyBase
GeneCards
GeneDB_Spombe
GeneFarm
GeneLynx
GenoList
Gramene
H-InvDB
HGNC
HPA
LegioList
Leproma
MaizeGDB
MGI
MIM
neXtProt
Orphanet
PharmGKB
PseudoCAP
RGD
SGD
TAIR
TubercuList
WormBase
Xenbase
ZFIN

Sequence databases

EMBL
IPI
PIR
RefSeq
UniGene

Enzyme and pathway databases

BioCyc
BRENDA
Pathway_Interaction_DB
Reactome
UniPathway

Family and domain databases

Gene3D
HAMAP
InterPro
PANTHER
PIRSF
Pfam
PRINTS
ProDom
PROSITE
SMART
SUPFAM
TIGRFAMs

2D-gel databases

2DBase-Ecoli
ANU-2DPAGE
Aarhus/Ghent-2DPAGE
COMPLUYEAST-2DPAGE
Cornea-2DPAGE
DOSAC-COBS-2DPAGE
ECO2DBASE
OGP
PHCI-2DPAGE
PMMA-2DPAGE
Rat-heart-2DPAGE
REPRODUCTION-2DPAGE
Siena-2DPAGE
SWISS-2DPAGE
UCD-2DPAGE
World-2DPAGE

PTM databases

GlycoSuiteDB
PhosSite
PhosphoSite

Proteomics and PPI

DIP
IntAct
Mint
PeptideAtlas
PRIDE
ProMEX
String

3D structure databases

DisProt
HSSP
PDB
PDBsum
ProteinModelPortal
SMR

Protein family/group databases

Allergome
CAZy
GermOnline
MEROPS
PeroxiBase
PptaseDB
REBASE
TCDB

Phylogenomic databases

eggNOG
GeneTree
HOGENOM
HOVERGEN
InParanoid
OMA
OrthoDB
PhylomeDB
ProtClustDB

Miscellaneous

ArrayExpress
Bgee
BindingDB
CleanEx
dbSNP
DrugBank
Genevestigator
GO
NextBio
PMAP-CutDB

Genome annotation databases

Ensembl
EnsemblBacteria
EnsemblFungi
EnsemblMetazoa
EnsemblPlants
EnsemblProtists
GenomeReviews
GeneID
KEGG
NMPDR
TIGR
UCSC
Vectorbase