Protein sequence annotation in the genome era: the annotation concept of SWISS-PROT+TREMBL

Apweiler, Rolf; Gateau, Alain; Contrino, Sergio; Martin, Maria J; Junker, Viv; O'Donovan, Claire; Lang, Fiona; Mitaritonna, Nicoletta; Kappus, Stephanie; Bairoch, Amos Marc

Protein Sequence Annotation in the Genome Era: The Annotation Concept of SWISS-PROT + TREMBL

Rolf Apweiler, Alain Gateau(*), Sergio Contrino, Maria Jesus Martin, Vivien Junker, Claire O'Donovan, Fiona Lang, Nicoletta Mitaritonna, Stephanie Kappus, Amos Bairoch(*)

The EMBL Outstation - The European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton,
Cambridge CB10 1SD, UK
apweiler@ebi.ac.uk

(*) Department of Medical Biochemistry
University of Geneva
Geneva, Switzerland

Abstract

SWISS-PROT is a curated protein sequence database which strives to provide a high level of annotation, a minimal level of redundancy and high level of integration with other databases. Ongoing genome sequencing projects have dramatically increased the number of protein sequences to be incorporated into SWISS-PROT. Since we do not want to dilute the quality standards of SWISS-PROT by incorporating sequences without proper sequence analysis and annotation, we cannot speed up the incorporation of new incoming data indefinitely. However, as we also want to make the sequences available as fast as possible, we introduced TREMBL (TRanslation of EMBL nucleotide sequence database), a supplement to SWISS-PROT. TREMBL consists of computer-annotated entries in SWISS-PROT format derived from the translation of all coding sequences (CDS) in the EMBL nucleotide sequence database, except for CDS already included in SWISS-PROT. While TREMBL is already of immense value, its computer-generated annotation does not match the quality of SWISS-PROT's. The main difference is in the protein functional information attached to sequences. With this in mind, we are dedicating substantial effort to develop and apply computer methods to enhance the functional information attached to TREMBL entries.

Introduction[1]

SWISS-PROT (Bairoch and Apweiler 1997) is an annotated protein sequence database established in 1986 and maintained collaboratively, since 1987, by the Department of Medical Biochemistry of the University of Geneva and the EMBL Data Library (now the EMBL Outstation - The European Bioinformatics Institute (EBI) (Stoesser et al. 1997)).

The SWISS-PROT database distinguishes itself from other protein sequence databases by three distinct criteria:

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.

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 with 26 different databases (Table 1). 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.

Annotation

One of SWISS-PROT's leading concepts from the very beginning was to provide far more than a simple collection of protein sequences, but rather a critical view of what is known or postulated about each of these sequences.
In SWISS-PROT, as in most other 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), and the taxonomic data (description of the biological source of the protein), while

EMBL Database	EMBL Nucleotide Sequence Database
DICTYDB	Dictyostelium discoideum genome database
ECO2DBASE	Escherichia coli gene-protein database (2D gel spots)
ECOGENE	Escherichia coli K12 genome database (EcoGene)
ENZYME	ENZYME data bank
FLYBASE	Drosophila genome database (FlyBase)
GCRDB	G-protein--coupled receptor database (GCRDb)
HIV	HIV sequence database
HSC-2DPAGE	Harefield Hospital 2D gel protein databases
HSSP	Homology-derived secondary structure of proteins database (HSSP)
LISTA	Yeast (Saccharomyces cerevisiae) genome database
MAIZEDB	Maize genome database (MaizeDB)
MAIZE-2DPAGE	Maize genome 2D Electrophoresis database
MEDLINE	Medline from the National Library of Medicine (NLM)
MIM	Mendelian Inheritance in Man Database
PDB	Brookhaven Protein Data Bank
PIR	Protein sequence database of the Protein Information Resource
PROSITE	PROSITE dictionary of sites and patterns in proteins
REBASE	Restriction enzyme database
AARHUS/GHENT-DPAGE	Human keratinocyte 2D gel protein database from Aarhus and Ghent universities
SGD	Saccharomyces Genome Database
STYGENE	Salmonella typhimurium LT2 genome database (StyGene)
SUBTILIST	Bacillus subtilis 168 genome database (SubtiList)
SWISS-2DPAGE	Human 2D Gel Protein Database from the University of Geneva
TRANSFAC	Transcription factor database (Transfac)
WORMPEP	Caenorhabditis elegans genome sequencing project protein database (Wormpep)
YEPD	Yeast electrophoresis protein database

Table 1: List of the databases cross-referenced to SWISS-PROT

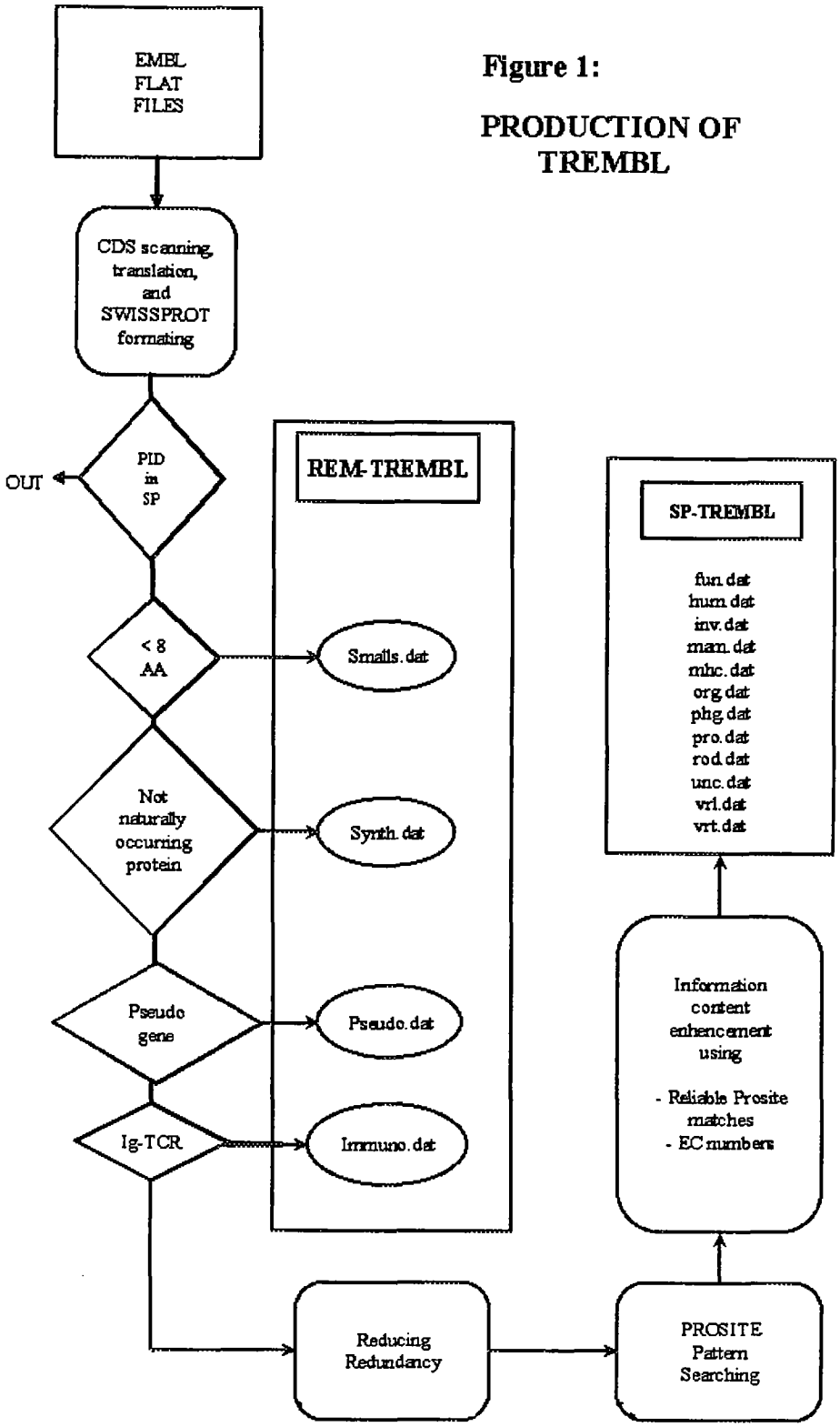
the annotation consists of the description of the following items:

o Function(s) of the protein
o Post-translational modification(s). For example carbohydrates, phosphorylation, acetylation, GPI-anchor, etc.
o Domains and sites. For example calcium binding regions, ATP-binding sites, zinc fingers, homeobox, kringle, etc.
o Secondary structure
o Quaternary structure
o Similarities to other proteins
o Disease(s) associated with deficiencie(s) in the protein
o Sequence conflicts, variants, etc.

In SWISS-PROT, annotation is mainly found in the comment lines (CC), in the feature table (FT) and in the keyword lines (KW). We use a controlled vocabulary whenever possible; this approach permits the easy retrieval of specific categories of data from the database.
We include as much annotation 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 our 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.
However, due to the increased data flow from genome projects to the sequence databases we face a number of challenges to our way of database curation. The attachment of biological knowledge abstracted from publications to the sequences is a skilled and labour-intensive task. Maintaining the high quality of sequence and annotation in SWISS-PROT requires careful sequence analysis and detailed curation of every entry. It is the rate-limiting step in the production of SWISS-PROT. The ever-increasing rate of determination of new sequences requires new approaches if SWISS-PROT is to keep up. While we do not wish to relax the high editorial standards of SWISS-PROT, it is clear that there is a limit to how much we can speed up such painstaking work. On the other hand, it is also vital that we make new sequences available as quickly as possible. To address this concern, we introduced, with SWISS-PROT release 33, TREMBL (TRanslation of EMBL nucleotide sequence database), a supplement to SWISS-PROT.

Figure 1:

PRODUCTION OF TREMBL



Recent Developments: TREMBL

The Production of TREMBL

Translation and Entry Creation

TREMBL consists of computer-annotated entries in SWISS-PROT format derived from the translation of all coding sequences (CDS) in the EMBL nucleotide sequence database, except for CDS already included in SWISS-PROT. The production of TREMBL is illustrated in Figure 1. All the EMBL nucleotide sequence database divisions are scanned for CDS features and these are translated to produce TREMBL division files containing TREMBL entries in SWISS-PROT format.

The program to produce TREMBL is written in C and makes use of the srs4_02 library (Etzold and Argos 1993), which provides the basis for a first level parsing of EMBL database entries. This level allows text data to fit in structures such as ordered lists of features or bibliographic references, to assemble the coding sequences and to translate them. It is not possible to rely on the /translation qualifier in the EMBL database entries, since the TREMBL production program has to report extra features like conflicts and variants at the amino acid level. Each CDS leading to a correct translation results in one entry whose ID is the PID of the CDS. In the next step the structures are scanned to extract relevant data, to filter it and eventually to insert it properly formatted into the TREMBL entry. Only bibliographic references relevant to the given CDS are kept in the TREMBL entry. This is achieved by scanning the RP (Reference Position) lines of the EMBL entry and matching with the CDS position in the sequence. The RC (Reference Comment) line is built by assigning the SWISS-PROT equivalent of the following EMBL qualifiers:

"/plasmid","PLASMID=",
"/strain","STRAIN=",
"/isolate","STRAIN=", (2nd choice)
"/cultivar","STRAIN=CV. "
"/tissue_type","TISSUE=",
"/transposon","TRANSPOSON=",

The description line comes from the /product qualifier, when present, otherwise we make use of the EMBL DE line, the /gene and /note qualifiers. The EMBL DE line is only considered if the EMBL entry holds only one cds and is stripped of non-pertinent information such as the organism name, or things like 'complete cds'. The /gene qualifier is also used for the TREMBL GN line.

At the moment, because the EMBL and SWISS-PROT taxonomies are slightly different, we use equivalence tables to assign OS and OC lines in the entries. Where no equivalent is found, the EMBL OS and OC lines are kept. Fortunately, in the near future, Genbank (Benson et al. 1997), EMBL, DDBJ (Tateno and Gojobori 1997) and SWISS-PROT are going to adopt a new common taxonomic scheme.

The EMBL keywords are included in the TREMBL entry, but only when they match a subset of SWISS-PROT keywords which have the same meaning. This occurs only in cases where the EMBL entry has just one CDS so that no ambiguity is possible. Some extra keywords derived from the features and description lines are added.

A subset of SWISS-PROT features can be derived from the EMBL entry features. These are:

- SIGNAL from sig_peptide
- TRANSIT from transit_peptide
- CHAIN from mat_peptide
- VARIANT from allele, variation, misc_difference and mutation
- CONFLICT from conflict

Two examples of TREMBL entries, created in the way described before, are shown in Figure 2. In addition to this information parsed into TREMBL entries, data is put in the annotator's section of the entry, which is not visible to the public. This is used for further analysis both by programs and by biologists and consists of:

- The EMBL entry description lines
- EMBL CC lines
- Bibliographic reference titles
- Full CDS feature text
- Full text of other relevant features within the CDS range
- Number of CDS in the EMBL entry
- The date of the last entry update
- Information if the organism already exists in SWISS-PROT

Sorting the Entries

In the process of building TREMBL, different types of entries are put into different output files:

- CDS with a /dbxref="SWISS-PROT" or a /dbxref="SPTREMBL" are not translated (already in SWISS-PROT + TREMBL)
- CDS from mhc genes -> mhc.dat
- CDS from patent data -> patent.dat
- CDS from immunoglobulins and t-cell receptors -> immuno.dat
- CDS smaller than 8 amino acids -> smalls.dat
- CDS from artificial, synthetic or chimeric genes -> synthetic.dat
- CDS from pseudogenes -> pseudo.dat
- remaining CDS -> stay in their relative taxonomic TREMBL divisions

```
ID  G34313 PRELIMINARY;  PRT;  332 AA.
AC  X02152_1;
DT  23-DEC-1996 (EMBLREL. 49, CREATED)
DT  23-DEC-1996 (EMBLREL. 49, LAST SEQUENCE UPDATE)
DT  23-DEC-1996 (EMBLREL. 49, LAST ANNOTATION UPDATE)
DE  LACTATE DEHYDROGENASE.
OS  HOMO SAPIENS (HUMAN).
OC EUKARYOTA; METAZOA; CHORDATA; VERTEBRATA; TETRAPODA; MAMMALIA;
OC  EUTHERIA; PRIMATES.
RN  [1]
RP  SEQUENCE FROM N.A.
RX  MEDLINE; 85127030.
RA  TSUJIBO H., TIANO H.F., LI S.S.-L.;
RL  EUR. J. BIOCHEM. 147:9-15(1985).
DR  EMBL; X02152; G34313; -.
SQ  SEQUENCE  332 AA;  36689 MW;  FF7595E2 CRC32;
//

ID  G780261 PRELIMINARY;  PRT;  332 AA.
AC  X03077_1;
DT  23-DEC-1996 (EMBLREL. 49, CREATED)
DT  23-DEC-1996 (EMBLREL. 49, LAST SEQUENCE UPDATE)
DT  23-DEC-1996 (EMBLREL. 49, LAST ANNOTATION UPDATE)
DE  LACTATE DEHYDROGENASE.
OS  HOMO SAPIENS (HUMAN).
OC  EUKARYOTA; METAZOA; CHORDATA; VERTEBRATA; TETRAPODA; MAMMALIA;
OC  EUTHERIA; PRIMATES.
RN  [1]
RP  SEQUENCE FROM N.A.
RX  MEDLINE; 86076881.
RA  CHUNG F.Z., TSUJIBO H., BHATTACHARYYA U., SHARIEF F.S., LI S.S.-L.;
RL  BIOCHEM. J. 231:537-541(1985).
DR  EMBL; X03077; G780261; -.
DR  EMBL; X03078; G780261; JOINED.
DR  EMBL; X03079; G780261; JOINED.
DR  EMBL; X03080; G780261; JOINED.
DR  EMBL; X03081; G780261; JOINED.
DR  EMBL; X03082; G780261; JOINED.
DR  EMBL; X03083; G780261; JOINED.
SQ  SEQUENCE  332 AA;  36689 MW;  FF7595E2 CRC32;
//
```

Figure 2: First level TREMBL entries (after translation and entry creation, sequence not shown)

At this stage the entries from the composite divisions of the EMBL database (STS, EST, and UNC) are added to their relative taxonomic TREMBL divisions.
Then all files are searched for entries that have recently been added to SWISS-PROT but which do not yet have a /dbxref="SWISS-PROT" qualifier in EMBL. These entries are removed and TREMBL is split into two different sections. SP-TREMBL (SWISS-PROT TREMBL) which contains entries that will be added, after complete annotation, to SWISS-PROT and REM-TREMBL (REMaining TREMBL) which contains entries not for inclusion in SWISS-PROT.

REM-TREMBL consists of 5 files (patent.dat, immuno.dat, smalls.dat, synthetic.dat, and pseudo.dat). SP-TREMBL consists of 12 files (fun.dat, inv.dat, hum.dat, mam.dat, mhc.dat, org.dat, phg.dat, pln.dat, pro.dat, rod.dat, vrl.dat, and vrt.dat) which will undergo further post-processing.

Post-processing the SP-TREMBL Entries
To post-process the SP-TREMBL entries, a collection of shell scripts and C programs are used.

```
ID  G34313 PRELIMINARY;  PRT;  332 AA.
AC  X02152_1;
DT  23-DEC-1996 (EMBLREL. 49, CREATED)
DT  23-DEC-1996 (EMBLREL. 49, LAST SEQUENCE UPDATE)
DT  23-DEC-1996 (EMBLREL. 49, LAST ANNOTATION UPDATE)
DE  LACTATE DEHYDROGENASE.
OS  HOMO SAPIENS (HUMAN).
OC  EUKARYOTA; METAZOA; CHORDATA; VERTEBRATA; TETRAPODA; MAMMALIA;
OC  EUTHERIA; PRIMATES.
RN  [1]
RP  SEQUENCE FROM N.A.
RX  MEDLINE; 85127030.
RA  TSUJIBO H., TIANO H.F., LI S.S.-L.;
RL  EUR. J. BIOCHEM. 147:9-15(1985).
RN  [2]
RP  SEQUENCE FROM N.A.
RX  MEDLINE; 86076881.
RA  CHUNG F.Z., TSUJIBO H., BHATTACHARYYA U., SHARIEF F.S., LI S.S.-L.;
RL  BIOCHEM. J. 231:537-541(1985).
DR  EMBL; X02152; G34313; -.
DR  EMBL; X03077; G780261; -.
DR  EMBL; X03078; G780261; JOINED.
DR  EMBL; X03079; G780261; JOINED.
DR  EMBL; X03080; G780261; JOINED.
DR  EMBL; X03081; G780261; JOINED.
DR  EMBL; X03082; G780261; JOINED.
DR  EMBL; X03083; G780261; JOINED.
SQ  SEQUENCE  332 AA;  36689 MW;  FF7595E2 CRC32;
//
```

Figure 3: Second level TREMBL entry (after merging, sequence not shown)

The first step is the reduction of redundancy. All full-length proteins in SP-TREMBL with the same sequence are merged into one entry. All fragment proteins with the same sequence from the same organism are merged provided they do not belong to a highly variable category of proteins like MHC proteins or viral proteins. For all SWISS-PROT entries, the CRC32 checksums of all the different annotated sequence reports are calculated and compared with the checksums of all SP-TREMBL entries. Identified matches are removed from SP-TREMBL and integrated into the corresponding SWISS-PROT entries. Figure 3 shows an example of an automatically merged TREMBL entry, created by merging of the two TREMBL entries shown in Figure 2.

Currently we are working on a further reduction of redundancy by establishing rules to automatically merge sub-fragments with full-length sequences and on the identification of sequence differences due to polymorphisms, strain variations and sequencing errors with the goal of establishing rules to automatically merge conflicting sequence reports about the same sequence into one entry.

This work is done in collaboration with Jean-Jacques Codani from INRIA, France. His group developed LASSAP (LArge Scale Sequence compArison Package), a programmable, high performance system designed to overcome current limitations of sequence comparison programs in order to fit the needs of large scale analysis (Glemet and Codani 1997. LASSAP provides an Application Programming Interface allowing the integration of any generic pairwise-based algorithm. Whichever pairwise algorithm is used in LASSAP, it shares numerous enhancements with all other algorithms such as:

- intra and inter data bank comparisons
- computational requests (selections and computations are achieved on the fly)
- frame translations on queries and data banks
- structured results allowing easy and powerful post-analysis
- performance improvements by parallelization and the driving of specialised hardware.

LASSAP allows the use of several sequence comparison methods: BLAST (Altschul et al. 1990), FASTA (Pearson and Lipman 1988), dynamic programming with local (Smith and Waterman 1981) and global (Needleman and Wunsch 1970) similarity searches, string matching with

```
ID  P00338     PRELIMINARY;  PRT;  332 AA.
AC  P00338;
DT  01-FEB-1997 (TREMBLREL. 02, CREATED)
DT  01-FEB-1997 (TREMBLREL. 02, LAST SEQUENCE UPDATE)
DT  01-FEB-1997 (TREMBLREL. 02, LAST ANNOTATION UPDATE)
DE  L-LACTATE DEHYDROGENASE (EC 1.1.1.27).
OS  HOMO SAPIENS (HUMAN).
OC  EUKARYOTA; METAZOA; CHORDATA; VERTEBRATA; TETRAPODA; MAMMALIA;
OC  EUTHERIA; PRIMATES.
RN  [1]
RP  SEQUENCE FROM N.A.
RX  MEDLINE; 85127030.
RA  TSUJIBO H., TIANO H.F., LI S.S.-L.;
RL  EUR. J. BIOCHEM. 147:9-15(1985).
RN  [2]
RP  SEQUENCE FROM N.A.
RX  MEDLINE; 86076881.
RA  CHUNG F.Z., TSUJIBO H., BHATTACHARYYA U., SHARIEF F.S., LI S.S.-L.;
RL  BIOCHEM. J. 231:537-541(1985).
CC  -!- CATALYTIC ACTIVITY: L-LACTATE + NAD(+) = PYRUVATE + NADH.
CC  -!- SUBUNIT: HOMOTETRAMER (BY SIMILARITY).
CC  -!- PATHWAY: FINAL STEP IN ANAEROBIC GLYCOLYSIS.
DR  EMBL; X02152; G34313; -.
DR  EMBL; X03077; G780261; -.
DR  EMBL; X03078; G780261; JOINED.
DR  EMBL; X03079; G780261; JOINED.
DR  EMBL; X03080; G780261; JOINED.
DR  EMBL; X03081; G780261; JOINED.
DR  EMBL; X03082; G780261; JOINED.
DR  EMBL; X03083; G780261; JOINED.
DR  PROSITE; PS00064; L_LDH.
KW  OXIDOREDUCTASE; NAD; GLYCOLYSIS.
FT  ACT_SITE   193   193     BY SIMILARITY.
SQ  SEQUENCE   332 AA;  36689 MW;  FF7595E2 CRC32;
//
```

Figure 4: Third level TREMBL entry (after complete post-processing, sequence not shown)

(Landau and Vishkin 1986, Hunt and Szymanski 1977) or without errors (Boyer and Moore 1977) and pattern matching with (Baeza-Yates and Gonnet 1989, Wu and Manber 1991) or without errors. We already use LASSAP to find sequences in SP-TREMBL which are sub-fragments of SWISS-PROT or other SP-TREMBL entries, and to find sequences in SP-TREMBL which are the full-length sequences of sequence fragments in SWISS-PROT. The comparisons are done with the Boyer-Moore algorithm. Identified matches are removed from SP-TREMBL and integrated into the corresponding SWISS-PROT or SP-TREMBL entries.

Although we have already identified thousands of SP-TREMBL entries with exact matches in SWISS-PROT or SP-TREMBL, there are tens of thousands of subtle redundancies due to polymorphisms, strain variations and sequencing errors which we would like to eliminate. Also, we need to establish further rules to automatically merge differing SP-TREMBL sequence reports about the same sequence into one entry.

The second post-processing step is the information enhancing process. For SP-TREMBL to act as a computer-annotated supplement to SWISS-PROT, new procedures have been introduced whereby valuable annotation can be added automatically. Firstly, all SP-TREMBL entries are scanned for PROSITE (Bairoch, Bucher, and Hofmann 1997) patterns compatible with their taxonomic range. The results are added to the annotator's section of the SP-TREMBL entry which is not visible to the public. Some of the patterns are known to be very reliable (i.e. no known false positive). These are used to enhance the information content of the DE, CC, DR, and KW fields by adding information about the potential function of the protein, metabolic pathways, active sites, cofactors, binding sites, domains, subcellular location, and

```
ID  LDHM_HUMAN    STANDARD;     PRT;  331 AA.
AC  P00338;
DE  L-LACTATE DEHYDROGENASE M CHAIN (EC 1.1.1.27) (LDH-A).
GN  LDHA.
OS  HOMO SAPIENS (HUMAN).
RN  [1]
RP  SEQUENCE FROM N.A.
RX  MEDLINE; 85127030.
RA  TSUJIBO H., TIANO H.F., LI S.S.-L.;
RL  EUR. J. BIOCHEM. 147:9-15(1985).
RN  [2]
RP  SEQUENCE FROM N.A.
RX  MEDLINE; 86076881.
RA  CHUNG F.Z., TSUJIBO H., BHATTACHARYYA U., SHARIEF F.S., LI S.S.-L.;
RL  BIOCHEM. J. 231:537-541(1985).
RN  [3]
RP  VARIANT CYS-314.
RX  MEDLINE; 93075246.
RA  SUDO K., MAEKAWA M., SHIOYA M., IKEDA K., TAKAHASHI N., ISOGAI Y.,
RA  LI S.S.-L., KANNO T., MACHIDA K., TORIUMI J.;
RL  BIOCHEM. INT. 27:1051-1057(1992).
RN  [4]
RP  VARIANT GLU-221.
RX  MEDLINE; 94199831.
RA  MAEKAWA M., SUDO K., KOBAYASHI A., SUGIYAMA E., LI S.S.-L., KANNO T.;
RL  CLIN. CHEM. 40:665-668(1994).
CC  -!- CATALYTIC ACTIVITY: L-LACTATE + NAD(+) = PYRUVATE + NADH.
CC  -!- SUBUNIT: HOMOTETRAMER.
CC  -!- PATHWAY: FINAL STEP IN ANAEROBIC GLYCOLYSIS.
CC  -!- THERE ARE THREE TYPES OF LDH CHAINS: M (LDH-A) FOUND PREDOMINANTLY
CC      IN MUSCLE TISSUES, H (LDH-B) FOUND IN HEART MUSCLE AND X (LDH-C)
CC      WHICH IS PRESENT IN THE SPERMATOZOA OF MAMMALS, IN THE COLUMBIDAE
CC      OF BIRDS AND IN ACTINOPTERYGIAN FISH.
CC  -!- DISEASE: EXERTIONAL MYOGLOBINURIA IS DUE TO A DEFECT IN LDH-A.
DR  EMBL; X02152; G34313; -.
DR  EMBL; X03077; G780261; -.
DR  EMBL; X03078; G780261; JOINED.
DR  EMBL; X03079; G780261; JOINED.
DR  EMBL; X03080; G780261; JOINED.
DR  EMBL; X03081; G780261; JOINED.
DR  EMBL; X03082; G780261; JOINED.
DR  EMBL; X03083; G780261; JOINED.
DR  PIR; A00347; DEHULM.
DR  HSSP; P00344; 1LDB.
DR  AARHUS/GHENT-2DPAGE; 2207; NEPHGE.
DR  MIM; 150000; -.
DR  PROSITE; PS00064; L_LDH.
KW  OXIDOREDUCTASE; NAD; GLYCOLYSIS; MULTIGENE FAMILY; DISEASE MUTATION;
KW  POLYMORPHISM.
FT  INIT_MET    0    0
FT  ACT_SITE    192  192     ACCEPTS A PROTON DURING CATALYSIS.
FT  VARIANT     221  221     K -> E.
FT  VARIANT     314  314     R -> C (IN LDHA DEFICIENCY).
SQ  SEQUENCE  331 AA;  36557 MW;  DF367487 CRC32;
//
```

Figure 5: Fully annotated SWISS-PROT entry (DT lines, OC lines and sequence not shown)

other annotation to the entry wherever appropriate. We also use the ENZYME database (Bairoch 1996), using the EC number as a reference point, to generate standardised description lines for enzyme entries and to allow information such as catalytic activity, cofactors and relevant keywords to be taken from ENZYME and to be added automatically to SP-TREMBL entries. Furthermore we use specialised databases like Flybase (The FlyBase Consortium 1997), SGD, GDB (Fasman et al. 1997), and MGD (Blake et al. 1997) to parse information like the correct gene nomenclature and cross-references to these databases into SP-TREMBL entries.

The now fully post-processed TREMBL entry, already used as an example before, is shown in Figure 4. Although this computer-generated annotation is already enhancing the information about the sequence drastically, it is still a long way to the quality of the corresponding SWISS-PROT entry (shown in Figure 5), fully annotated by biologists. Also, this is not a representative example of a TREMBL entry. A (fortunately also not representative) example of a TREMBL entry with low annotation quality is shown in Figure 6. In this case, a badly annotated nucleotide sequence resulted in a badly annotated TREMBL entry, which we could not improve with our current, limited range of automated annotation tools.

The current status of TREMBL

The TREMBL release created from the EMBL Nucleotide Sequence Database release 49 contains (February 1997) 116,379 sequence entries, comprising 31,293,053 amino acids. This TREMBL release was distributed with SWISS-PROT release 34 (containing 59,021 sequence entries with 21,210,389 amino acids).

Most of the 96,757 sequence entries currently in SP-TREMBL are additional sequence reports of entries already in SWISS-PROT and will lead to updates of those SWISS-PROT entries. However, some 30,000 to 40,000 entries now in SP-TREMBL will eventually be included as new sequence entries in SWISS-PROT. Approximately 20% of the SP-TREMBL entries have been post-processed.

The majority of REM-TREMBL entries (currently approximately 13,000 of 19,620) are immunoglobulins and T-cell receptors. In SWISS-PROT, we have translations of the germ line genes for immunoglobulins and T-cell receptors but we do not wish to add all known somatic recombined variant sequences as this would bias database-wide searches. Such entries will be placed in IMGT-TREMBL (ImMunoGeneTics-TREMBL). We will, in collaboration with IMGT (Guidicelli et al. 1997), develop IMGT-TREMBL as a specialist protein database for immunoglobulins and T-cell receptors. This supplement to SWISS-PROT will be presented in SWISS-PROT format and cross-referenced to SWISS-PROT, the EMBL Nucleotide Sequence Database, and IMGT.

Another category of data which will not be included in SWISS-PROT is synthetic sequences (SWISS-PROT represents only naturally occurring sequences). Again, we do not wish to leave these entries in TREMBL. Ideally one should build a specialized database for artificial sequences as a further supplement to SWISS-PROT. The remainder of the REM-TREMBL entries are patents, pseudogenes (SWISS-PROT does not represent genes known not to be expressed), and sequence fragments of 8 amino acid residues or less.

The EMBL Nucleotide Sequence Database release 50 contained 20,000 new CDS, leading to the creation of a corresponding number of new TREMBL entries in only 3 months (EMBL releases are created every 3 months). However, in the same period we integrated 10,000 SP-TREMBL sequence reports into SWISS-PROT or merged them with existing SWISS-PROT + TREMBL entries. In 1997, we expect approximately 80,000 new CDS in the EMBL Nucleotide Sequence Database. Although we hope to increase the number of sequence entries integrated into SWISS-PROT from 59,000 at the end of 1996 to 75,000 at the end of 1997, the number of entries in TREMBL will probably increase (after removal of redundancy) from 105,000 at the end of 1996 to 150,000 at the end of 1997. This prediction underlines the fact that the ever-increasing automation of SP-TREMBL annotation methods is the only long-term viable approach to the constantly increasing data flow.

The future of annotation in TREMBL

Most of the sequence data nowadays is coming from genome projects and lacks biochemical evidence to provide hard data on the function of the protein. The prediction of functional information from primary sequence information is a comparative problem based on a set of general rules and relationships derived from the current set of known proteins. Sequence similarity searches, pattern and profile searches, and clustering of sequences are currently helping us to take in the annotation process advantage of the relationship between primary sequence and function.

Modern sensitive database search algorithms find already characterised sequences similar to new sequences and enable us to annotate new sequences by analogy to old sequences. Secondary pattern and profile databases are used to enhance SP-TREMBL entries by adding information about the potential functions of proteins, metabolic pathways, active sites, cofactors, binding sites, domains, subcellular location, and other annotation. We are automating the similarity and motif searches to accelerate the upgrading of SP-TREMBL entries to SWISS-PROT standard. The annotation task, whether automated or carried out by database curators, can proceed far more quickly if large groups of related proteins, such as families of sequences sharing a similar

```
ID  P77135     PRELIMINARY;  PRT;  127 AA.
AC  P77135;
DT  01-FEB-1997 (TREMBLREL. 02, CREATED)
DT  01-FEB-1997 (TREMBLREL. 02, LAST SEQUENCE UPDATE)
DT  01-FEB-1997 (TREMBLREL. 02, LAST ANNOTATION UPDATE)
DE  O94 AND HYPOTHETICAL PROTEIN GENES, PARTIAL CDS
DE  (FRAGMENT).
GN  O94.
OS  ESCHERICHIA COLI.
OC  PROKARYOTA; GRACILICUTES; SCOTOBACTERIA; FACULTATIVELY ANAEROBIC RODS;
OC  ENTEROBACTERIACEAE.
RN  [1]
RP  SEQUENCE FROM N.A.
RC  STRAIN=MG1655;
RA  ROBERTS D., ALLEN E., ARAUJO R., APARICIO A., CHUNG E., DAVIS K.,
RA  DUNCAN M., FEDERSPIEL N., HYMAN R., KALMAN S., KOMP C., KURDI O.,
RA  LEW H., LIN D., NAMATH A., OEFNER P., SCHRAMM S., DAVIS R.W.;
RL  SUBMITTED (JAN-1997) TO EMBL/GENBANK/DDBJ DATA BANKS.
DR  EMBL; U83187; G1773217; -.
FT  NON_TER     1     1
SQ  SEQUENCE   127 AA;  14752 MW;  883882DD CRC32;
//
```

Figure 6: TREMBL entry with low annotation quality (sequence not shown)

motif, can be annotated together. We are attempting to do this by creation of alignments of all proteins and building of clusters of proteins with high percentage identities, indicating which sequence reports belong to the same group of proteins or which proteins share similar domains. These clusters will also be used for automatic motif detection and as a starting point to develop profile related algorithms.

However, currently most of these results are not included in the entries as seen by the public. These results are only accessible internally by the curators. We are developing a rule-based system which considers all the results obtained by the different methods to add annotation to SP-TREMBL entries. This system consists of growing numbers of sequence analysis methods, rules, and hierarchical classifications of the annotation content of SWISS-PROT entries, where all nodes in these hierarchical trees are linked to certain annotation. The rules consider the sequence analysis results to decide which node(s) in the classification tree(s) are sufficiently similar to the query sequence and lead subsequently to the incorporation of the appropriate annotation (linked to the node) in the SP-TREMBL entry. The incorporated annotation is flagged as annotation based on sequence analysis methods. These results internally accessible by the curators only are used by the SWISS-PROT team to establish additional rules to enhance the annotation of SP-TREMBL entries. We only add information based on our automatic analysis to SP-TREMBL entries, if we are convinced that the computer-generation creates correct annotation in more than 90% of the cases.

With this annotation concept of SWISS-PROT + TREMBL we try to combine the strengths of annotation carefully done by human experts with biological knowledge and after consultation of the relevant literature and thorough sequence analysis with the power of automation of sequence analysis and computer-generation of annotation. Since predicted annotation assignments and assignments based on hard experimental evidence are clearly distinguishable, we will present in TREMBL highly reliable although putative functional predictions, without lowering the high editorial standards of the standard SWISS-PROT entries.

References

Altschul S.F., Gish W., Miller W., Myers E.W., and Lipman D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403-410.

Baeza-Yates R.A., and Gonnet G.H. 1989. Efficient text searching of regular expressions. *Proceedings ICALP* 16:46-62.

Bairoch A. 1996. The ENZYME data bank in 1995. *Nucleic Acid Res.* 24: 221-222.

Bairoch A., Bucher P., and Hofmann K. 1997. The PROSITE database, its status in 1997. *Nucleic Acid Res.* 25:217-221.

Bairoch A., and Apweiler R. 1997. The SWISS-PROT protein sequence data bank and its supplement TREMBL. *Nucleic Acids Res.* 25:31-36.

Benson D.A., Boguski M., Lipman D.J., and Ostell J. 1997. GenBank. *Nucleic Acid Res.* 25:1-6.

Blake J.A., Richardson J.E., Davisson M.T., Eppig J.T., and the Mouse Genome Informatics Group 1997. The Mouse Genome Database (MGD). A comprehensive public resource of genetic, phenotypic and genomic data. *Nucleic Acid Res.* 25:85-91.

Boyer R., and Moore S. 1977. A fast string searching algorithm. *C.A.C.M.* 20:762-772.

Etzold T., and Argos P. 1993. SRS, an indexing and retrieval tool for flat file data libraries. *Comput. Appl. Biosci.* 9:49-57.

Fasman K.H., Letovsky S.I., Li P., Cottingham R.W., Kingsbury D.T. 1997. The GDB Human Genome Database Anno 1997. *Nucleic Acid Res.* 25:72-80.

FlyBase Consortium 1997/ FlyBase: a Drosophila database. *Nucleic Acid Res.* 25:63-66.

Glemet E., and Codani, J.-J. 1997. LASSAP, a Large Scale Sequence compArison Package. *CABIOS*: In Press.

Guidicelli V., Chaume D., Bodmer J., Mueller W., Busin C., Marsh S., Bontrop R., Marc L., Malik A., and Lefranc M.-P. 1997. IMGT, the international ImMunoGeneTics database. *Nucleic Acid Res.* 25:206-211.

Hunt J., and Szymanski T.G. 1977. A fast algorithm for computing longest common subsequences. *Theoretical Computer Science* 20:350-353.

Landau G.M., and Vishkin U. 1986. Efficient string matching with k mismatches. *Theoretical Computer Science* 43:239-249.

Needleman S.B., and Wunsch C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48:443-453.

Pearson W.R., and Lipman D.J. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* 85:2444-2448.

Smith T.F., and Waterman M.S. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* 147:195-197.

Stoesser G., Sterk P., Tuli M.A., Stoehr P.J., and Cameron G.N. 1997. The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.* 25:7-13.

Tateno Y., and Gojobori T. 1997. DNA Data Bank of Japan in the age of information biology. *Nucleic Acid Res.* 25:14-17.

Wu S., and Manber U. 1991. Fast text searching with errors. Technical Report TR 91-11, University of Arizona.