# SWISS-PROT: connecting biomolecular knowledge via a protein database

Gasteiger, Elisabeth; Jung, Eva; Bairoch, Amos Marc

Curr. Issues Mol. Biol. (2001) 3(3): 47-55.

# SWISS-PROT: Connecting Biomolecular Knowledge Via a Protein Database

**Elisabeth Gasteiger*, Eva Jung, and Amos Bairoch**

SWISS-PROT group, Swiss Institute of Bioinformatics, CMU, 1 rue Michel-Servet, 1211 Genève 4, Switzerland

**Abstract**

**With the explosive growth of biological data, the development of new means of data storage was needed. More and more often biological information is no longer published in the conventional way via a publication in a scientific journal, but only deposited into a database. In the last two decades these databases have become essential tools for researchers in biological sciences. Biological databases can be classified according to the type of information they contain. There are basically three types of sequence-related databases (nucleic acid sequences, protein sequences and protein tertiary structures) as well as various specialized data collections. It is important to provide the users of biomolecular databases with a degree of integration between these databases as by nature all of these databases are connected in a scientific sense and each one of them is an important piece to biological complexity. In this review we will highlight our effort in connecting biological information as demonstrated in the SWISS-PROT protein database.**

## Data Integration Using Cross-References

The current situation of a research scientist has been described quite accurately by a quote from John Naisbitt, saying that "We are drowning in information, but starving for knowledge". The World Wide Web (Berners-Lee, 1999), which immensely facilitated information exchange between information providers and users, now offers the life science community a wealth of easily accessible knowledge and information. While clicking on hypertext links and thus navigating between databases maintained around the world seems to be a technically easy task, the challenge lies in extracting the complete and up-to-date information related to a research field from the hundreds of databases available. The user can be assisted in this task by the creators of information resources, who should attempt to provide a system that allows scientists to rapidly and efficiently consult all information pertinent to a given topic. This is usually done by establishing *cross-references* from each record in a database to related entries in other databases.

*For correspondence. Email Elisabeth.Gasteiger@isb-sib.ch; Tel. +41-22-7025050; Fax. +41-22-7025858.



## Cross-References in SWISS-PROT

SWISS-PROT (Bairoch *et al.*, 2000) is a curated protein sequence database, which strives to provide a high level of annotation (such as the description of the function of a protein, its domain structure, post-translational modifications (PTM), variants, etc.), a minimal level of redundancy and high level of integration with other databases.

### SWISS-PROT Entry Format, a Sample Entry and Line Types Implementing Cross-References

A SWISS-PROT entry is composed of different line types, and each line is introduced by a two-letter code indicating the type of data following on that line (see Figure 1 for a sample entry). The first section of every SWISS-PROT entry contains the entry name (ID), a unique primary accession number (AC), sometimes followed by several secondary accession numbers, and dates indicating when the entry was created and when its sequence and annotations were last updated (DT). The description line (DE) lists all names, including synonyms, under which the protein has been known, and the GN line contains the name(s) of the gene(s) coding for it. The following section contains taxonomic data about the organism from which the protein originates, in particular the organism name (OS), its classification in the taxonomic tree (OC) and a unique taxonomy identifier (OX). The reference section (RN, RP, RX, RA, RT and RL lines) contains all literature references consulted for the annotation of the protein. The list of references includes not only publications of the sequence itself, but also articles detailing post-translational modifications, 3-D structure, polymorphisms etc. The reference section is followed by the comment block (CC) containing textual information classified into different "topics" and describing the protein' s function, subcellular localisation, post-translational modifications, association with diseases etc.

Database cross-references are stored in the DR lines and allow the user to access related information in other databases. DR lines will be described in detail in the next section. The keyword section (KW line type) lists a number of terms from a controlled vocabulary, which can be used to retrieve subsets of the database. A very important part of a SWISS-PROT protein entry is the feature table (FT lines), which contains information about interesting sites or domains within the protein sequence, for which positional information is known. The feature table describes events such as post-translational modifications, sequence variants due to polymorphisms, domain structure, sequence conflicts, etc. Each feature line consists of a feature key, start and end positions of the described feature in the precursor sequence, and the feature description. Finally, there is the amino acid sequence itself.

The SWISS-PROT database was the first biomolecular

```
ID APE_HUMAN STANDARD; PRT; 317 AA.
AC P02649;
DT 21-JUL-1986 (Rel. 01, Created)
DT 21-JUL-1986 (Rel. 01, Last sequence update)
DT 01-OCT-2000 (Rel. 40, Last annotation update)
DE Apolipoprotein E precursor (Apo-E).
GN APOE.
OS Homo sapiens (Human).
OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX NCBI_TaxID=9606;
RN [1]
RP SEQUENCE FROM N.A. (VARIANT E3).
RX MEDLINE=84185684; PubMed=6325438;
RA Zannis V.I., McPherson J., Goldberger G., Karathanasis S.K.,
RA Breslow J.L.;
RT "Synthesis, intracellular processing, and signal peptide of human
RT apolipoprotein E.";
RL J. Biol. Chem. 259:5495-5499(1984).
RN [2]
RP SEQUENCE FROM N.A. (VARIANT E3).
RX MEDLINE=84212473; PubMed=6327682;
RA McLean J.W., Elshourbagy N.A., Chang D.J., Mahley R.W., Taylor J.M.;
RT "Human apolipoprotein E mRNA. cDNA cloning and nucleotide sequencing
RT of a new variant.";
RL J. Biol. Chem. 259:6498-6504(1984).

… 15 references omitted …

RN [18]
RP X-RAY CRYSTALLOGRAPHY (2.25 ANGSTROMS) OF 41-184.
RX MEDLINE=91289138; PubMed=2063194;
RA Wilson C., Wardell M.R., Weisgraber K.H., Mahley R.W., Agard D.A.;
RT "Three-dimensional structure of the LDL receptor-binding domain of
RT human apolipoprotein E.";
RL Science 252:1817-1822(1991).
RN [19]
RP X-RAY CRYSTALLOGRAPHY (2.0 ANGSTROMS) OF 41-181.
RX MEDLINE=96313129; PubMed=8756331;
RA Dong L.-M., Parkin S., Trakhanov S.D., Rupp B., Simmons T.,
RA Arnold K.S., Newhouse Y.M., Innerarity T.L., Weisgraber K.H.;
RT "Novel mechanism for defective receptor binding of apolipoprotein E2
RT in type III hyperlipoproteinemia.";
RL Nat. Struct. Biol. 3:718-722(1996).
RN [20]
RP X-RAY CRYSTALLOGRAPHY (1.85 ANGSTROMS) OF 22-165.
RX MEDLINE=20306971; PubMed=10850798;
RA Segelke B.W., Forstner M., Knapp M., Trakhanov S.D., Parkin S.,
RA Newhouse Y.M., Bellamy H.D., Weisgraber K.H., Rupp B.;
RT "Conformational flexibility in the apolipoprotein E amino-terminal
RT domain structure determined from three new crystal forms:
RT implications for lipid binding.";
RL Protein Sci. 9:886-897(2000).
CC -!- FUNCTION: APO-E MEDIATES BINDING, INTERNALIZATION, AND CATABOLISM
CC OF LIPOPROTEIN PARTICLES. IT CAN SERVE AS A LIGAND FOR THE LDL(APO
CC B/E) RECEPTOR AND FOR THE SPECIFIC APO-E RECEPTOR (CHYLOMICRON
CC REMNANT) OF HEPATIC TISSUES.
CC -!- SUBCELLULAR LOCATION: EXTRACELLULAR.
CC -!- TISSUE SPECIFICITY: OCCURS IN ALL LIPOPROTEIN FRACTIONS IN PLASMA.
CC IT CONSTITUTES 10-20% OF VERY LOW DENSITY LIPOPROTEINS (VLDL) AND
CC 1-2% OF HIGH DENSITY LIPOPROTEINS (HDL). APOE IS PRODUCED IN MOST
CC ORGANS. SIGNIFICANT QUANTITIES ARE PRODUCED IN LIVER, BRAIN,
CC SPLEEN, LUNG, ADRENAL, OVARY, KIDNEY, AND MUSCLE.
CC -!- PTM: SYNTHESIZED WITH THE SIALIC ACID ATTACHED BY O-GLYCOSIDIC
CC LINKAGE AND IS SUBSEQUENTLY DESIALATED IN PLASMA.
CC -!- POLYMORPHISM: THREE MAJOR ISOFORMS CAN BE RECOGNIZED, DESIGNATED
CC E2, E3, AND E4, ACCORDING TO THEIR RELATIVE POSITION AFTER
CC ISOELECTRIC FOCUSING. THE MOST COMMON VARIANT IS E3 AND IS PRESENT
CC IN 40-90% OF THE POPULATION.
CC -!- DISEASE: IN ADDITION TO THE INFLUENCE OF COMMON APOE VARIANTS ON
CC LIPOPROTEIN METABOLISM IN HEALTHY INDIVIDUALS, APOE VARIANTS ARE
CC ALSO ASSOCIATED WITH FAMILIAL DYSBETALIPOPROTEINEMIA (FD):
CC INDIVIDUALS WITH TYPE III HYPERLIPOPROTEINEMIA, ARE CLINICALLY
CC CHARACTERIZED BY XANTHOMAS, YELLOWISH LIPID DEPOSITS IN THE PALMAR
CC CREASE THAT ARE PATHOGNOMONIC FOR FD, OR LESS SPECIFIC ON TENDONS
CC AND ON ELBOWS. FD RARELY MANIFESTS BEFORE THE THIRD DECADE IN MEN.
CC IN WOMEN, IT IS USUALLY EXPRESSED ONLY AFTER THE MENOPAUSE. THE
CC VAST MAJORITY OF FD PATIENTS ARE HOMOZYGOUS FOR APOE2. FD HAS ALSO
CC BEEN OBSERVED IN INDIVIDUALS HETEROZYGOUS FOR RARE APOE VARIANTS,
CC IN THESE CASES FD IS MORE SEVERE. THE INFLUENCE OF APOE ON LIPID
CC LEVELS IS OFTEN SUGGESTED TO HAVE MAJOR IMPLICATIONS FOR THE RISK
```

```
CC OF CORONARY ARTERY DISEASE (CAD). INDIVIDUALS CARRYING THE COMMON
CC APOE4 VARIANT ARE AT HIGHER RISK OF CAD.
CC -!- SIMILARITY: BELONGS TO THE APOA1 / APOA4 / APOE FAMILY.
CC -!- DATABASE: NAME=HotMolecBase; NOTE=ApoE entry;
CC WWW="http://bioinformatics.weizmann.ac.il/hotmolecbase/entries/apoe.htm".
DR EMBL; M12529; AAB59518.1; -.
DR EMBL; K00396; AAB59546.1; -.
DR EMBL; M10065; AAB59397.1; -.
DR EMBL; AF050154; AAD02505.1; -.
DR PIR; A03093; LPHUE.
DR PIR; JS0084; JS0084.
DR PDB; 1LE2; 15-OCT-92.
DR PDB; 1LE4; 15-OCT-92.
DR PDB; 1LPE; 15-OCT-92.
DR PDB; 1NFN; 27-JAN-97.
DR PDB; 1NFO; 27-JAN-97.
DR PDB; 1OEF; 07-DEC-96.
DR PDB; 1OEG; 07-DEC-96.
DR PDB; 1BZ4; 11-NOV-98.
DR SWISS-2DPAGE; P02649; HUMAN.
DR MIM; 107741; -.
DR InterPro; IPR000074; -.
DR Pfam; PF01442; Apolipoprotein; 1.
KW Glycoprotein; Plasma; Lipid transport; HDL; VLDL; Chylomicron;
KW Heparin-binding; Repeat; Signal; 3D-structure; Disease mutation;
KW Polymorphism.
FT SIGNAL 1 18
FT CHAIN 19 317 APOLIPOPROTEIN E.
FT DOMAIN 158 168 LDL RECEPTOR BINDING (POTENTIAL).
FT DOMAIN 162 165 HEPARIN-BINDING.
FT DOMAIN 229 236 HEPARIN-BINDING.
FT DOMAIN 80 255 8 X 22 AA APPROXIMATE TANDEM REPEATS.
FT REPEAT 80 101 1.
FT REPEAT 102 123 2.
FT REPEAT 124 145 3.
FT REPEAT 146 167 4.
FT REPEAT 168 189 5.
FT REPEAT 190 211 6.
FT REPEAT 212 233 7.
FT REPEAT 234 255 8.
FT CARBOHYD 212 212 O-LINKED (GALNAC...).
FT VARIANT 21 21 E -> K (IN E5-TYPE).
FT /FTId=VAR_000645.
FT VARIANT 31 31 E -> K (IN E5-TYPE AND E4 PHILADELPHIA).
FT /FTId=VAR_000646.
FT VARIANT 46 46 L -> P (IN E4 FREIBURG).
FT /FTId=VAR_000647.
FT VARIANT 60 60 T -> A (IN E3 FREIBURG).
FT /FTId=VAR_000648.
FT VARIANT 99 99 Q -> K (IN E5 FRANKFURT).
FT /FTId=VAR_000649.
FT VARIANT 102 102 P -> R (IN E5-TYPE; NO HYPERLIPIDEMIA).
FT /FTId=VAR_000650.
FT VARIANT 117 117 A -> T (IN E3*).
FT /FTId=VAR_000651.

… several variants omitted …

FT VARIANT 292 292 R -> H (IN E4 P.D.).
FT /FTId=VAR_000671.
FT VARIANT 314 314 S -> R (IN E4 H.G.).
FT /FTId=VAR_000672.
FT HELIX 43 59
FT TURN 60 60
FT HELIX 63 70
FT HELIX 73 96
FT TURN 97 98
FT HELIX 105 141
FT TURN 142 145
FT HELIX 149 180
FT TURN 181 182
SQ SEQUENCE 317 AA; 36154 MW; 91AFC04210A30689 CRC64;
MKVLWAALLV TFLAGCQAKV EQAVETEPEP ELRQQTEWQS GQRWELALGR FWDYLRWVQT
LSEQVQEELL SSQVTQELRA LMDETMKELK AYKSELEEQL TPVAEETRAR LSKELQAAQA
RLGADMEDVC GRLVQYRGEV QAMLGQSTEE LRVRLASHLR KLRKRLLRDA DDLQKRLAVY
QAGAREGAER GLSAIRERLG PLVEQGRVRA ATVGSLAGQP LQERAQAWGE RLRARMEEMG
SRTRDRLDEV KEQVAEVRAK LEEQAQQIRL QAEAFQARLK SWFEPLVEDM QRQWAGLVEK
VQAAVGTSAA PVPSDNH
//
```

Figure 1. SWISS-PROT entry P02649, Human apolipoprotein E precursor (Apo-E).

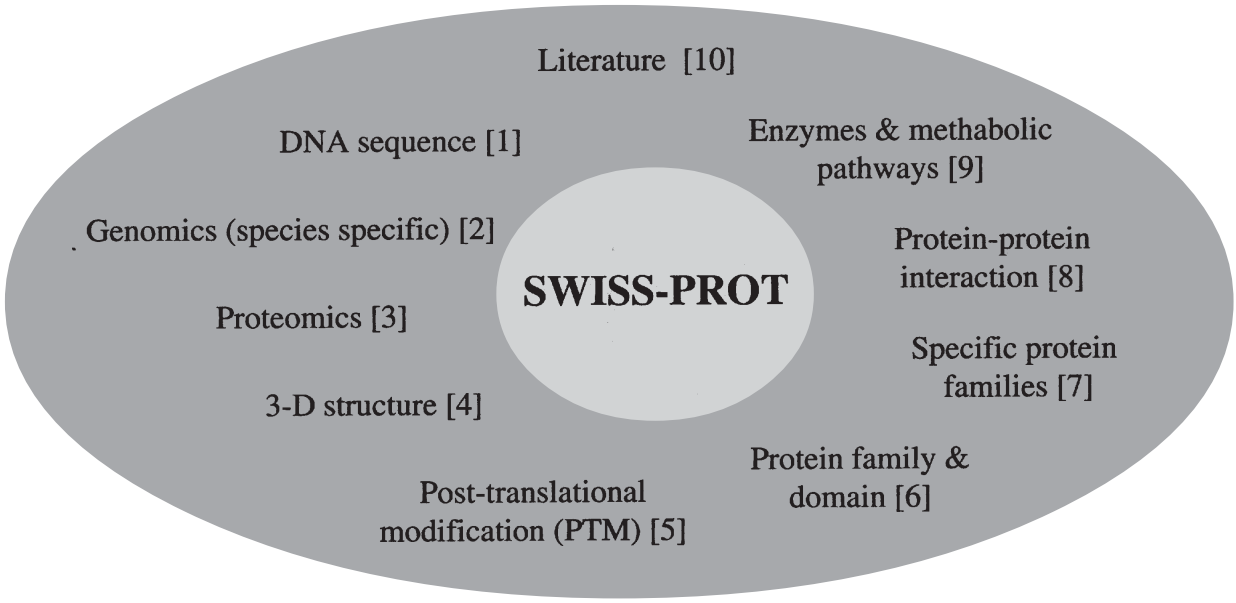


Figure 2. SWISS-PROT and cross-references to other databases. The different types of data repositories are shown to which SWISS-PROT has established links in the description (DE) line, reference cross-reference (RX) line or database cross-reference (DR) lines: [1] DNA sequence: DDBJ (Tateno *et al.*, 2000); EMBL (Stoesser *et al.*, 2001); GenBank (Benson *et al.*, 2000). [2] Genomics (species specific): DictyDb (Smith *et al.*, 1997); EcoGene (Rudd, 2000); FlyBase (The FlyBase consortium, 1999); GeneCards (Rebhan *et al.*, 1998); GeneCensus (Gerstein, 1998); HIV (Kuiken *et al.*, 1999); HUGE (Kikuno R. *et al.*, 2000); MaizeDB (Polacco *et al.*, 1999); Mendel (Price *et al.*, 2001); MGD (Blake *et al.*, 2001); Micado (Perriere *et al.*, 1999); NRSUB (Perriere *et al.*, 1998); MIM (Wheeler *et al.*, 2001); SGD (Ball *et al.*, 2001); StyGene (Sanderson *et al.*, 1995); SubtiList (Moszer, 1998); TIGR (Quackenbush *et al.*, 2001); TubercuList (Cole, 1999); WormPep (Sonnhammer *et al.*, 1997); YPD (Costanzo *et al.*, 2001); ZFIN (Sprague *et al.*, 2001). [3] Proteomics: ECO2DBASE (VanBogelen *et al.*, 1999); HSC-2DPAGE (Evans *et al.*, 1997); MAIZE-2DPAGE (Touzet *et al.*, 1996); SWISS-2DPAGE (Hoogland *et al.*, 2000); Aarhus/Ghent-2DPAGE (Celis *et al.*, 1998); YEPD (Latter *et al.*, 1995). [4] 3D structure: HSSP (Dodge *et al.*, 1998); PDB (Bhat *et al.*, 2001); PRESAGE (Brenner *et al.*, 1999) SWISS-3DIMAGE (Peitsch *et al.*, 1995). [5] Post-translational modification: CarbBank (Doubet *et al.*, 1989); GlycoSuite (Cooper *et al.*, 2001). [6] Protein family and domain: BLOCKS (Henikoff *et al.*, 2000); DOMO (Gracy *et al.*, 1998); InterPro (Apweiler *et al.*, 2000); Pfam (Bateman *et al.*, 2000); PRINTS (Attwood *et al.*, 2000); ProDom (Corpet *et al.*, 2000); PROSITE (Hofmann *et al.*, 1999); ProtoMap (Yona *et al.*, 2000). [7] Specific protein families: GCRDb (Kolakowski, 1994); GPCRDB (Horn *et al.*, 2001); IMGT (Lefranc, 2001); MEROPS (Rawlings *et al.*, 2000); NucleaRDB (Horn *et al.*, 2001); REBASE (Roberts *et al.*, 2001). [8] Protein-Protein Interaction: DIP (Xenarios *et al.*, 2001). [9] Enzymes and methabolic pathways: EcoCyc (Karp *et al.*, 2000); ENZYME (Bairoch, 2000); MEROPS (Rawlings *et al.*, 2000); REBASE (Roberts *et al.*, 2001). [10] Literature: PubMed, Medline (Wheeler *et al.*, 2001).

database to include cross-references in its entries – long before the advent of the World Wide Web, which made navigation between data resources distributed all over the planet become second nature to all its users. There are five different types of cross-references available in SWISS-PROT: *explicit* and *implicit* cross-references in the DR lines, URL addresses under the comment (CC) topic "DATABASE", and cross-references departing from certain key types in the feature table (FT). Finally, the Medline/PubMed (Wheeler *et al.*, 2001) identifiers of literature references are stored in RX (Reference Cross(X)reference) lines and thus allow direct access to these literature databases. There are a number of other annotation items in SWISS-PROT that might also be termed cross-references and that are, in the World Wide Web version, enhanced with active hypertext links, namely scientific journal references (RL lines), taxonomy identifier (OX lines) or enzyme classification numbers (DE lines). These different types of cross-references will be described in more detail in subsequent subsections.

In addition to cross-references provided by SWISS-PROT itself, SWISS-PROT also plays an important role for federated 2D-PAGE databases (Appel *et al.*, 1996), which achieve much of the integration of data located and maintained at different sites through SWISS-PROT as their main index. We will explain this concept in a later subsection.

**DR Lines**

The DR (Database cross-Reference) lines are used as pointers to information related to SWISS-PROT entries and found in data collections other than SWISS-PROT (see Figure 2). The full list of all databases to which SWISS-PROT is cross-referenced can be found in the document file dbxref.txt (http://www.expasy.ch/cgi-bin/lists?dbxref.txt).

For example, 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 (Stoesser et al., 2001; Benson *et al.*, 2000; Tateno *et al.*, 2000), which correspond to the DNA or RNA sequence(s) from which it was translated. If the X-ray crystallographic atomic coordinates of a sequence are stored in the Protein Data Bank (PDB) (Bhat *et al.*, 2001), there will be DR line(s) pointing to the corresponding entri(es) in that database.

*Explicit and Implicit Links*

The database cross-references in DR (Database cross-Reference) lines available from the Web version of SWISS-PROT on ExPASy (example: http://www.expasy.ch/cgi-bin/

Table 1. Databases explicitly referenced in SWISS-PROT DR lines. Abbreviations used: Database identifier: short identifier as used in SWISS-PROT DR lines; Nb. entries: total number of SWISS-PROT entries in release 39.14 with cross-references to this database; Nb. DR lines: Total number of DR lines linking to this database in SWISS-PROT release 39.14. The full names corresponding to the database identifiers used here can be extracted from the reference list at the end of this article, and are listed in the SWISS-PROT document http://www.expasy.ch/cgi-bin/lists?dbxref.txt. Databases cross-referenced in TrEMBL are highlighted in grey.

| Database identifier | No. entries | No. DR lines |
|---|---|---|
| EMBL | 87502 | 158541 |
| InterPro | 66712 | 97619 |
| Pfam | 64205 | 76620 |
| PROSITE | 49062 | 75151 |
| HSSP | 24705 | 24705 |
| PRINTS | 24141 | 30638 |
| TIGR | 6281 | 6301 |
| MIM | 5424 | 6046 |
| SGD | 4799 | 4847 |
| EcoGene | 4046 | 4048 |
| MGD | 3917 | 3928 |
| PDB | 2971 | 10073 |
| Mendel | 2826 | 2915 |
| SubtiList | 2207 | 2208 |
| MEROPS | 2110 | 2199 |
| WormPep | 2005 | 2039 |
| FlyBase | 1785 | 1833 |
| TubercuList | 1235 | 1260 |
| GCRDb | 972 | 1661 |
| TRANSFAC | 970 | 1052 |
| StyGene | 754 | 755 |
| SWISS-2DPAGE | 733 | 734 |
| MaizeDB | 397 | 401 |
| HIV | 354 | 370 |
| REBASE | 306 | 308 |
| DictyDb | 303 | 306 |
| ECO2DBASE | 299 | 351 |
| GlycoSuiteDB | 198 | 198 |
| ZFIN | 138 | 138 |
| YEPD | 120 | 129 |
| Aarhus/Ghent 2DPAGE | 98 | 128 |
| HSC-2DPAGE | 85 | 85 |
| MAIZE-2DPAGE | 39 | 39 |
| CarbBank | 21 | 41 |

niceprot.pl?P00750 or http://www.expasy.ch/cgi-bin/get-sprot-entry?P00750) include both *explicit* and *implicit* cross-references.

*Explicit Links*

Typically, a SWISS-PROT entry will have cross-references to its parent DNA sequence(s), to a genomic database (MIM (Wheeler *et al.*, 2001), MGD (Blake *et al.*, 2001), FlyBase (The FlyBase consortium, 1999), SGD (Ball *et al.*, 2001), etc.), to information detailing its three-dimensional (3D) structure (PDB), etc. All these cross-references are stored in the SWISS-PROT flat file, generally in the form "DR database name; primary identifier; secondary identifier." and are termed "explicit". The primary identifier is an unambiguous pointer to the information entry in the database to which reference is made; for most databases, this corresponds to the (unique) accession number of the remote entry. The secondary identifier is generally used to complement the information given by the first identifier. Examples for secondary identifiers can be entry names or release numbers. The SWISS-PROT user manual (http://www.expasy.ch/txt/userman.txt) provides a detailed explanation of primary and secondary identifiers for each of the referenced databases. Databases cross-referenced via explicit links have their own system of unique identifiers, which distinguishes them from the resources referenced via implicit links, as explained in the following subsection.

SWISS-PROT release 39.14 of 21/02/2001 consists of 93,407 protein sequence entries, which contain 564,764 explicit DR lines. Table 1 lists the 34 databases referenced in this manner, sorted by decreasing total number of SWISS-PROT entries linked to each of these databases. The absolute numbers shown in this table will of course be already obsolete at time of publication as table 1 is merely a snapshot of release 39.14 – however, it is important to note that each SWISS-PROT entry contains an average of 6 explicit cross-references. It is further noteworthy that those entries with cross-references to EMBL, *i.e.* those derived from nucleic acid sequences, contain an average of 1.81 DR EMBL lines. This illustrates SWISS-PROT's high emphasis to reduce redundancy and to merge entries describing the same protein. TrEMBL (Bairoch *et al.*, 2000) is a computer-annotated supplement to SWISS-PROT which, by definition, lacks much of the high quality annotation. However, as far as cross-references are concerned, many of the above-mentioned principles for SWISS-PROT also apply to TrEMBL. Numerous types of explicit cross-references can indeed be built automatically (often in collaboration with the database to which the link is established), and a large number of DR lines are systematically added and kept up to date in TrEMBL. These databases are highlighted in grey in Table 1.

*Implicit Links*

The ExPASy server further enhances the database interoperability offered by the explicit links by automatically adding, where appropriate, some so-called "implicit" links. As opposed to databases referenced via explicit links, the data collections linked in this manner usually do not have their own system of unique identifiers; however, they can be referenced via identifiers such as SWISS-PROT accession numbers, gene names, EMBL accession numbers, etc. Two broad categories of implicit links exist:

a) Various databases have been developed in the last ten years that are completely based on SWISS-PROT (and sometimes also TrEMBL) and offer a specific analytical view of the database. For example, the ProDom (Corpet *et al.*, 2000) and DOMO (Gracy *et al.*, 1998) databases provide automatically derived domain views of each protein in SWISS-PROT; the ProtoMap (Yona *et al.*, 2000) database is a hierarchical classification of all SWISS-PROT proteins. In such cases, it is straightforward to add implicit links. There should be, for each SWISS-PROT entry, a corresponding entry in these derived databases, and such an entry can be accessed using the primary accession number of the SWISS-PROT entry.

b) There are specialized databases, which are supposed to share some form of "identifiers" with SWISS-PROT. A typical example is GeneCards (Rebhan *et al.*, 1998), a database containing information on human genes. It can be accessed using the HUGO (Human Genome Organization) approved symbol for a relevant gene. Because SWISS-PROT also uses, as the first name listed on the GN line, the HUGO approved gene symbol, it is

possible to automatically generate links between SWISS-PROT and GeneCards.

In both cases, no extra DR line for such data collections is added to the SWISS-PROT flat file. Indeed, such DR would take up "space" without really adding information: Knowing that information related to every human SWISS-PROT entry can be found in GeneCards, and that this information is accessible by searching GeneCards for the SWISS-PROT accession number, an explicit line "DR GeneCards; P00001." is redundant and will not provide the user with pertinent new information. In the Web version however, a hypertext link is useful and allows the user to navigate directly to the related information provided by the remote data collection.

Whilst the distinction between explicit and implicit links may not seem very important to a user querying SWISS-PROT through the World Wide Web, as both types of links are clickable, there are two noteworthy issues:

Firstly, if you look or print an entry from the Web server, it will contain lines that do not exist in the version distributed with and accessible through various software packages or from other Web servers; ExPASy and the EBI servers are the only ones to provide most of the implicit links described above. Secondly, such automatically generated links can fail in some cases. For example, a new SWISS-PROT entry may not yet have a corresponding entry in a derived database. Or, to take the example of GeneCards, it could happen that the gene symbol has not yet been "synchronized" (either GeneCards has updated a gene name before SWISS-PROT or the reverse).

A growing number of databases (currently 19) can be accessed via implicit links in the SWISS-PROT version displayed on the ExPASy server. The documentation file dbxref.txt (http://www.expasy.ch/cgi-bin/lists?dbxref.txt) lists URLs for all of these databases and details the rules and criteria used for the construction of the implicit links.

**CC Lines**

While the DR lines provide access to general databases relating to a large number of SWISS-PROT entries, there are hundreds of resources, databases and data collections available on the World Wide Web that specialize in one specific protein or protein family, or provide detailed information about mutations or polymorphisms. A comprehensive and regularly updated list of such resources is available at http://www.expasy.ch/alinks.html. These data collections are usually extremely heterogeneous, and not necessarily very structured, and are generally accessed via one central URL rather than by accession number. Cross-references to such resources are provided in SWISS-PROT CC (comments) lines, under the topic 'DATABASE'. The syntax of the 'DATABASE' topic is:

"CC -!- DATABASE: NAME=Text[; NOTE=Text][; WWW="Address"][; FTP="Address"]."

where 'NAME' is the name of the database; 'NOTE' (optional) is a free text note; 'WWW' (optional) is the WWW address (URL) of the database; 'FTP' (optional) is the anonymous FTP address (including the directory name) where the database file(s) are stored. An example for such a cross-reference can be seen in Figure 1.

In SWISS-PROT release 39.14, 131 different resources were referenced in this way, from 440 CC DATABASE topics in 411 entries.

**FT Lines**

Certain specialized protein-related databases have entries that do not correspond to the complete protein sequence described by a SWISS-PROT entry, but rather to a particular sub-sequence or even just one amino acid. Examples are databases specializing in certain types of post-translational modifications of proteins, or in mutations. The section of SWISS-PROT that contains position-specific annotation being the feature table (FT), it makes sense to construct links directly from relevant feature table lines to the corresponding information in other databases. The feature identifiers (FTId) in FT VARIANT lines of human sequence entries for example allow to refer to a sequence variation in a unique and stable manner, and serve as anchors for specifically directed links. A federated single human mutation database (HmutDB; http://www2.ebi.ac.uk/mutaions/central/proposal.html) has been proposed, and the complete set of all FT VARIANT lines has been indexed for SRS at EBI (http://srs.ebi.ac.uk/), under the name SWISSCHANGE. The database SWISSCHANGE can be queried by SWISS-PROT FTIds.

In the future, the same principle will be used to further enhance the links to GlycoSuiteDB (Cooper *et al*., 2001). GlycoSuiteDB is an annotated database of glycan structures. For human Alpha-2-HS-glycoprotein (P02765) for example, GlycoSuiteDB provides detailed information about two different oligosaccharide structures attached to five different sites within the sequence. In addition to providing a global link to GlycoSuiteDB by creating an explicit DR line in SWISS-PROT entry P02765, we will create unique feature identifiers for each of the 5 FT CARBOHYD lines, which will allow direct access to the corresponding glycan structures.

**RX Lines**

For each published reference cited in SWISS-PROT, for which an entry in the literature databases Medline/PubMed exists, the reference block of the SWISS-PROT entry contains an RX line providing the Medline and PubMed identifiers. This allows quick and direct access to the abstract of the publication.

**Other Cross-References**

When looking at a SWISS-PROT entry on ExPASy (e.g. http://www.expasy.ch/cgi-bin/get-sprot-entry?P00750), one immediately notices the large number of hypertext links (clickable text portions displayed in blue and underlined, unless the Web browser is configured to display them differently). Many of them are links to resources local to ExPASy, but there are a number of annotation items that might, even if they do not belong to any of the cross-reference types described earlier, also be termed cross-references. Two examples are the taxonomy identifier (Tax_ID), the unique identifier of each organism in the NCBI taxonomy classification (Wheeler *et al*., 2001) (in the OX line of each SWISS-PROT entry), and the Enzyme Classification (EC) numbers (Bairoch, 2000) which can be

found in SWISS-PROT description (DE) lines of relevant protein sequence entries.

Another example are the coordinates (journal name, volume and page numbers and year of publication) for scientific journal references, which can be categorized as both explicit and implicit links: Explicit, because the RL line of the SWISS-PROT flat file allows the user to find the published article, either the paper copy at a library, or the electronic version through the journal's web site. And implicit, because the ExPASy Web version of SWISS-PROT adds some extra value to this hard-coded reference, by providing a hypertext link to the publisher site, using some additional information such as the date or volume number of the first issues available on-line. We plan to further enhance these links in the future, by encouraging publishers to link back to SWISS-PROT (and thus establish bi-directional links), and by creating more stable links using ISSN numbers and Digital Object Identifiers (DOI; see http://www.doi.org/).

**SWISS-PROT as a Common Index for Federated Databases**

With data exchange becoming more and more convenient, scientists can easily collaborate via the Internet, which can often result in very powerful projects combining the expertise of all participants. Instead of creating one central database for a specific topic, several independent, separately and heterogeneously maintained databases can be joined using the concept of *federated databases*. This principle has been successfully applied to the field of proteomics, where currently 15 federated 2D-PAGE databases exist (http://www.expasy.ch/ch2d/2d-index.html). These databases agreed to comply with a number of rules (Appel *et al.*, 1996), which are mainly based on cross-references: Among other requirements, the database must be linked to other databases through active hypertext cross-references, which link together all related databases and combine them into one large virtual database. In addition to these hypertext links between federated databases, a main index has to be supplied that provides a means of querying all databases through one unique entry point. Bi-directional cross-references must exist between the main index and the other databases. SWISS-PROT currently acts as this main index.

This concept of database federation could also be applied to other fields of specialisation. One might imagine federated databases for protein post-translational modifications (where GlycoSuiteDB could be considered as the first such federated database), or for polymorphisms and mutations. Both these subjects are extremely complex, and require expertise that is probably best shared between centers of competence around the planet rather than centralized in one single database. The possibility of creating links from such specific entities as SWISS-PROT feature table lines opens a large potential of database interoperability where SWISS-PROT can serve as a common index. We highly encourage any interested database provider, in particular those specializing in post-translational modifications, to collaborate with us in order to provide users with an even more comprehensive view of all data available for their protein of interest.

**Conclusions**

In this paper we illustrated our efforts to integrate biomolecular knowledge in our protein database SWISS-PROT. While aiming at providing as much annotation information on the protein as possible, we place a strong emphasis on integration with other biomolecular data repositories. Needless to say that in the era of proteomics we are already flooded with an increasing amount of data resulting from the analysis of the complex readout of genomes. Therefore we predict that there will be a lot more spezialised data repositories established in the future, and we hope to collaborate with these information resources. Cross-references in SWISS-PROT allow users with a specific interest in a particular biomolecular field to access information gathered by experts in their domain, whilst individual SWISS-PROT entries are not overloaded with spezialised information and remain easily comprehensible to the general interest user.

**References**

Appel, R.D., Bairoch, A., Sanchez, J.-C., Vargas, J.R., Golaz, O., Pasquali, C. and Hochstrasser, D.F. 1996. Federated 2-DE database: a simple means of publishing 2-DE data. Electrophoresis 17: 540-546.

Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N.J., Oinn, T.M., Pagni, M., Servant, F., Sigrist, C.J. and Zdobnov, E.M. 2000. InterPro - an integrated documentation resource for protein families, domains and functional sites. Bioinformatics 16: 1145-1150.

Attwood, T.K., Croning, M.D.R., Flower, D.R., Lewis, A.P., Mabey, J.E., Scordis, P., Selley, J.N. and Wright W. 2000. PRINTS-S: the database formerly known as PRINTS. Nucleic Acids Res. 28: 225-227.

Bairoch, A. 2000. The ENZYME database in 2000. Nucleic Acids Res. 28: 304-305.

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

Ball, C.A., Jin, H., Sherlock, G., Weng, S, Matese, J.C., Andrada, R., Binkley, G., Dolinski, K., Dwight, S.S., Harris, M.A., Issel-Tarver, L., Schroeder, M., Botstein, D. and Cherry, J.M. 2001. Saccharomyces Genome Database provides tools to survey gene expression and functional analysis data. Nucleic Acids Res. 29: 80-81.

Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L. and Sonnhammer, E.L. 2000. The Pfam protein families database. Nucleic Acids Res. 28: 263-266.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A. and Wheeler, D.L. 2000. GenBank. Nucleic Acids Res. 28: 15-18.

Berners-Lee, T. 1999. Weaving the Web. Harper, San Francisco.

Bhat, T.N., Bourne, P., Feng, Z., Gilliland, G., Jain, S., Ravichandran, V., Schneider, B., Schneider, K., Thanki, N., Weissig, H., Westbrook, J.and Berman, H.M. 2001.

The PDB data uniformity project. Nucleic Acids Res. 29: 214-218.

Blake, J.A., Eppig, J.T., Richardson, J.E., Bult, C.J. and Kadin, J.A. 2001. The Mouse Genome Database (MGD): integration nexus for the laboratory mouse. Nucleic Acids Res. 29: 91-94.

Brenner, S.E., Barken, D. and Levitt, M. 1999. The PRESAGE database for structural genomics. Nucleic Acids Res. 27: 251-253.

Celis, J.E., Ostergaard, M., Jensen, N.A., Gromova, I., Rasmussen, H.H. and Gromov, P. 1998. Human and mouse proteomic databases: novel resources in the protein universe. FEBS Lett. 430: 64-72.

Cole, S.T. 1999. Learning from the genome sequence of *Mycobacterium tuberculosis* H37Rv. FEBS Lett. 452: 7-10.

Cooper, C.A., Harrison, M.J., Wilkins, M.R. and Packer, N.H. 2001. GlycoSuiteDB: a new curated relational database of glycoprotein glycan structures and their biological sources. Nucleic Acids Res. 29: 332-335.

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

Costanzo, M.C., Crawford, M.E., Hirschman, J.E., Kranz, J.E., Olsen, P., Robertson, L.S., Skrzypek, M.S., Braun, B.R., Hopkins, K.L., Kondu, P., Lengieza, C., Lew-Smith, J.E., Tillberg, M. and Garrels, J.I. 2001. YPD, PombePD and WormPD: model organism volumes of the BioKnowledge library, an integrated resource for protein information. Nucleic Acids Res. 29: 75-79.

Dodge, C., Schneider, R. and Sander, C. 1998. The HSSP database of protein structure-sequence alignments and family profiles. Nucleic Acids Res. 26: 313-315.

Doubet, S., Bock, K., Smith, D., Darvill, A. and Albersheim, P. 1989. The Complex Carbohydrate Structure Database. Trends Biochem. Sci. 14: 475-477.

Evans, G., Wheeler, C.H., Corbett, J.M. and Dunn, M.J. 1997. Construction of HSC-2DPAGE: a two-dimensional gel electrophoresis database of heart proteins. Electrophoresis. 18: 471-479.

Gerstein, M. 1998. How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. Fold Des. 3: 497-512.

Gracy, J. and Argos, P. 1998. DOMO: a new database of aligned protein domains. Trends Biochem. Sci. 23: 495-497.

Henikoff, J.G., Greene, E.A., Pietrokovski, S. and Henikoff, S. 2000. Increased coverage of protein families with the blocks database servers. Nucleic Acids Res. 28: 228-230.

Hofmann, K., Bucher, P., Falquet, L. and Bairoch, A. 1999. The PROSITE database, its status in 1999. Nucleic Acids Res. 27: 215-219.

Hoogland, C., Sanchez, J.-C., Tonella, L., Binz, P.-A., Bairoch, A., Hochstrasser, D.F. and Appel, R.D. 2000. The 1999 SWISS-2DPAGE database update. Nucleic Acids Res. 28: 286-288.

Horn, F., Vriend, G. and Cohen, F.E. 2001. Collecting and harvesting biological data: the GPCRDB and NucleaRDB information systems. Nucleic Acids Res. 29: 346-349.

Karp, P.D., Riley, M., Saier, M., Paulsen, I.T., Paley, S.M. and Pellegrini-Toole, A. 2000. The EcoCyc and MetaCyc databases. Nucleic Acids Res. 28: 56-59.

Kikuno, R., Nagase, T., Suyama, M., Waki, M., Hirosawa, M., Ohara, O. 2000. HUGE: a database for human large proteins identified in the Kazusa cDNA sequencing project. Nucleic Acids Res 28: 331-332.

Kolakowski, L.F. 1994. GCRDb: a G-protein-coupled receptor database. Receptors Channels. 2: 1-7.

Kuiken, C.L., Foley, B., Hahn, B., Korber, B., McCutchan, F., Marx, P.A., Mellors, J.W., Mullins, J.I., Sodroski, J. and Wolinksy, S. 1999. Human Retroviruses and AIDS 1999: A Compilation and Analysis of Nucleic Acid and Amino Acid Sequences. In: Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM.

Latter, G.I., Boutell, T., Monardo, P.J., Kobayashi, R., Futcher, B., Mclaughlin, C.S.and Garrels, J.I. 1995. A *Saccharomyces cerevisiae* Internet protein resource now available. Electrophoresis. 16: 1170-1174.

Lefranc, M.P. 2001. IMGT, the international ImMunoGeneTics database. Nucleic Acids Res. 29: 207-209.

Moszer, I. 1998. The complete genome of Bacillus subtilis: from sequence annotation to data management and analysis. FEBS Lett. 430: 28-36.

Perriere, G., Bessieres, P. and Labedan, B. 1999. The Enhanced Microbial Genomes Library. Nucleic Acids Res. 27: 63-65.

Perriere, G., Gouy, M. and Gojobori, T. 1998. The non-redundant Bacillus subtilis (NRSub) database: update 1998. Nucleic Acids Res. 26: 60-62.

Peitsch, M.C., Wells, T.N., Stampf, D.R. and Sussman, J.L. 1995. The Swiss-3DImage collection and PDB-Browser on the World-Wide Web. Trends Biochem. Sci. 20: 82-84.

Polacco, M., Chen, S., Coe, E., Hancock, D.C., Kross, H., Schroeder, S. and Vargo, C. 1999. MaizeDB: integrated maize genome resource. Community curation, database interoperability, and comparative map displays. Maize Genetics Conference Abstracts 41.

Price, C.A. and Reardon, E.M. 2001. Mendel, a database of nomenclature for sequenced plant genes. Nucleic Acids Res. 29: 118-119.

Quackenbush, J., Cho, J., Lee, D., Liang, F., Holt, I., Karamycheva, S., Parvizi, B., Pertea, G., Sultana, R. and White, J. 2001. The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. Nucleic Acids Res. 29: 159-164.

Rawlings, N.D. and Barrett, A.J. 2000. MEROPS: the peptidase database. Nucleic Acids Res. 28: 323-325.

Rebhan, M., Chalifa-Caspi, V., Prilusky, J. and Lancet, D. 1998. GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. Bioinformatics 14: 656-664.

Roberts, R.J. and Macelis, D. 2001. REBASE—restriction enzymes and methylases. Nucleic Acids Res. 29: 268-269.

Rudd, K.E. 2000. EcoGene: a genome sequence database for *Escherichia coli* K-12. Nucleic Acids Res. 28: 60-64.

Sanderson, K.E., Hessel, A. and Rudd, K.E. 1995. Genetic map of *Salmonella typhimurium*, edition VIII. Microbiol.

Rev. 59: 241-303.
Smith, D.W. and Loomis, W.F. 1997. DictyDB - A Genomic Database for *Dictyostelium discoideum*. In: *Dictyostelium*. A Model System for Cell and Developmental Biology. Y. Maeda, K. Inouyea and I. Takeuchi, eds. Universal Academic Press, Inc. - Tokyo, Japan. p. 471-477.
Sonnhammer, E.L. and Durbin, R. 1997. Analysis of protein domain families in *Caenorhabditis elegans*. Genomics 46: 200-216.
Sprague, J., Doerry, E., Douglas, S. and Westerfield, M. 2001. The Zebrafish Information Network (ZFIN): a resource for genetic, genomic and developmental research. Nucleic Acids Res. 29: 87-90.
Stoesser, G., Baker, W., van den Broek, A., Camon, E., Garcia-Pastor, M., Kanz, C., Kulikova, T., Lombard, V., Lopez, R., Parkinson, H., Redaschi, N., Sterk, P., Stoehr, P. and Tuli, M.A. 2001. The EMBL nucleotide sequence database. Nucleic Acids Res. 29: 17-21.
Tateno, Y., Miyazaki, S., Ota, M., Sugawara, H. and Gojobori, T. 2000. DNA Data Bank of Japan (DDBJ) in collaboration with mass sequencing teams. Nucleic Acids Res. 28: 24-26.
The FlyBase Consortium. 1999. The FlyBase database of the Drosophila Genome Projects and community literature. Nucleic Acids Res. 27: 85-88.
Touzet, P., Riccardi, F., Morin, C., Damerval, C., Huet, J.-C., Pernollet, J.-C., Zivy, M. and de Vienne, D. 1996. The maize two dimensional gel protein database: towards an integrated genome analysis program. Theor. Appl. Genet. 93: 997-1005.
VanBogelen, R.A., Schiller, E.E., Thomas, J.D. and Neidhardt, F.C. 1999. Diagnosis of cellular states of microbial organisms using proteomics. Electrophoresis 20: 2149-2159.
Wheeler, D.L., Church, D.M., Lash, A.E., Leipe, D.D., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Tatusova, T.A., Wagner, L. and Rapp, B.A. 2001. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 29: 11-16.
Xenarios, I., Fernandez, E., Salwinski, L., Duan, X.J., Thompson, M.J., Marcotte, E.M. and Eisenberg, D. 2001. DIP: The Database of Interacting Proteins: 2001 update. Nucleic Acids Res. 29: 239-241.
Yona, G., Linial, N. and Linial, M. 2000. ProtoMap: automatic classification of protein sequences and hierarchy of protein families. Nucleic Acids Res. 28: 49-55.