Annotation of post-translational modifications in the Swiss-Prot knowledge base

Nathalie Farriol-Mathis[1], John S. Garavelli[2], Brigitte Boeckmann[1], Séverine Duvaud[1], Elisabeth Gasteiger[1], Alain Gateau[1], Anne-Lise Veuthey[1] and Amos Bairoch[1]

[1]Swiss Institute of Bioinformatics, Centre Médical Universitaire (CMU), Geneva, Switzerland
[2]European Bioinformatics Institute, Genome Campus, Hinxton, Cambridge, UK

High-throughput proteomic studies produce a wealth of new information regarding post-translational modifications (PTMs). The Swiss-Prot knowledge base is faced with the challenge of including this information in a consistent and structured way, in order to facilitate easy retrieval and promote understanding by biologist expert users as well as computer programs. We are therefore standardizing the annotation of PTM features represented in Swiss-Prot. Indeed, a controlled vocabulary has been associated with every described PTM. In this paper, we present the major update of the feature annotation, and, by showing a few examples, explain how the annotation is implemented and what it means. Mod-Prot, a future companion database of Swiss-Prot, devoted to the biological aspects of PTMs (*i.e.*, general description of the process, identity of the modification enzyme(s), taxonomic range, mass modification) is briefly described. Finally we encourage once again the scientific community (*i.e.*, both individual researchers and database maintainers) to interact with us, so that we can continuously enhance the quality and swiftness of our services.

Received	22/10/03
Revised	27/1/04
Accepted	30/1/04

Keywords: Annotation / Bioinformatics / Database / Post-translational modifications / Prediction tools / Protein sequence

1 Introduction

With accelerating progress in the field of proteomics, biological knowledge bases such as Swiss-Prot [1] must cope with a huge wealth of information, in particular the protein modifications that play crucial structural and functional roles [2]. The challenge for curators is to include as much of this information as feasible in protein database entries, in a way that is consistent and logical, and allows easy retrieval by biologist expert users, as well as by computer programs. In order to approach this problem efficiently, one needs to catalogue the different types of modifications which occur in proteins. This requirement is fulfilled by the RESID Database of Protein Modifications [3]. The RESID database is a "comprehensive collection of annotations and structures for protein modifications including amino-terminal, carboxyl-terminal and peptide chain cross-link, pre-, co- and post-translational modifications". It is regularly updated with modified amino acids as they are discovered and their structure is determined. Release 35.00 on 30-Sep-2003 contained entries for 344 unmodified and modified amino acids.

Although the modification of amino acids does occur before (pre-), during (co-) and after (post-) the said amino acids are incorporated into proteins by ribosomes, they are usually referred to misleadingly as post-translational modifications, or PTMs.

PTMs arise from the cleaving or forming of covalent bonds and can be classified into three categories based on the following processes: cleavage (including pre- and propeptide processing, initiator methionine removal, *C*-terminal processing), linkage (including attachment of chemical groups from the simple such as acetyl, methyl, phosphoryl, or hydroxyl, to more complex entities such as glycans or lipids) and cross-linking (including disulphide, thioether, and thioester bonds). Many complex PTMs arise from both a cleavage and a linkage, including glycosylphosphatidylinositol GPI-anchor attachment, intein

Correspondence: Dr. Nathalie Farriol-Mathis, Swiss Institute of Bioinformatics, Centre Médical Universitaire (CMU), 1 rue Michel-Servet, CH-1211 Genève 4, Switzerland
E-mail: nathalie.farriol-mathis@isb-sib.ch
Fax: +41-22-379-58-58

Abbreviations: FT, feature table; **GPI**, glycosylphosphatidylinositol; **SRS**, sequence retrieval system

splicing and hedgehog processing. More than three hundred different modifications are currently known and new types are reported in the literature at a rate of roughly ten *per* year.

While some PTMs are absolutely necessary for the function of particular proteins, such as those which form or attach a cofactor, other PTMs can be regarded as optional when they target a protein to a cellular location (*e.g.* protein lipidation), or regulate protein-protein and cell-cell interactions (*e.g. N*-glycosylation), and activation state (*e.g.* phosphorylation). It is now clear that the activity state of a protein often depends on its modification state. Even if the expression of a gene is the same in two situations, the phosphorylation status can determine the activation or the inactivation of a given protein. For example, signal transduction pathways, the cell cycle, and many other crucial pathways in eukaryote development are clearly dependent on phosphorylation/dephosphorylation cycles. Recent studies on cytoplasmic glycosylation indicate that some sites that can be phosphorylated may alternatively be glycosylated [4]. A given serine or threonine residue can thus have one of three states: unmodified, glycosylated or phosphorylated. Efforts are made to characterize with high-throughput the protein modifications which occur in signal transduction cascades following cell activation or differentiation. This research field has been termed functional proteomics [5].

Taking into account alternative splicing along with alternative post-translational events, the number of different protein molecules produced by less than 25 000 human genes is generally estimated to be close to one million. This huge complexity of the human proteomes was estimated by counting, in 2-DE, the number of well-separated spots for different isoforms of the same protein, since all spots are marked by a unique monoclonal antibody. For example, ten different forms of the cellular tumor antigen p53 have been detected in human liver tissue [6]. Thanks to recent improvements in mass spectrometry technologies, not only can these proteins be identified but their modifications can also be characterized by taking into account the difference of mass caused by the addition of chemical groups such as phosphate, methyl, acetyl, *etc.* [7].

In Swiss-Prot release 42 of October 2003, the most represented PTMs are, in decreasing number of occurrences, signal peptide processing, disulphide bonds, *N*-glycosylation, phosphorylation, ligation of 4Fe-4S clusters, palmitoylation, *C*-terminal cleavage and amidation, ligation of 2Fe-2S clusters, and initiator methionine removal coupled to *N*-terminal acetylation. Concomitantly, the database documents more than 80 rare PTMs that have been found so far only in single protein families. For example: tyrosine iodination in thyroglobulins, and hypusine formation from lysine residues in translation initiation factor eIF-5A.

2 Links to the cited tools, databases and facilities

Table 1 lists resources cited in this paper that are accessible on web servers, and that are mostly available free of charge.

Table 1. Links to the cited tools, databases and facilities

ExPASy [9]	www.expasy.org/
SWISS-PROT [1]	www.expasy.org/sprot/
SWISS-PROT documents	www.expasy.org/sprot/sp-docu.html
Update requests	www.expasy.org/sprot/update.html
RESID [3]	www.ncifcrf.gov/RESID/
PDB [18]	www.rcsb.org/pdb/
GlycoSuiteDB [10]	www.glycosuite.com/
Phosphorylation site DB	vigen.biochem.vt.edu/xpd/xpd.htm
IMP Bioinformatics Group	mendel.imp.univie.ac.at/mendeljsp/index.jsp
SRS on ExPASy	www.expasy.org/srs/
SRS at the EBI [28]	srs.ebi.ac.uk
PubMed	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed
FindMod [13]	www.expasy.org/tools/findmod/
GlycoMod [15]	www.expasy.org/tools/glycomod/
PeptIdent	www.expasy.org/tools/peptident.html
FindPept [14]	www.expasy.org/tools/findpept.html
PeptideMass [16]	www.expasy.org/tools/peptide-mass.html
SignalP [22]	www.cbs.dtu.dk/services/SignalP/
TargetP [23]	www.cbs.dtu.dk/services/TargetP/
Predotar	www.inra.fr/predotar/
Big-PI predictor [19]	mendel.imp.univie.ac.at/sat/gpi/gpi_server.html
NMT [20]	mendel.imp.univie.ac.at/myristate/SUPLpredictor.htm
Sulfinator [21]	www.expasy.org/tools/sulfinator/

3 Swiss-Prot entry format

The format of Swiss-Prot entries is extensively described in various documents [8], including the Swiss-Prot user manual. Hence, we will not detail the format

here yet again but would instead advise the reader to have a look at the SWISS-PROT user manual provided on the internet.

4 Annotations regarding PTMs

There are three main places in which PTMs are presented in Swiss-Prot entries: (a) The PTM comments (CC -!- PTM: . . .), where information is presented on PTMs that have not been located in the sequence. Additional data that we do not want to store in the feature lines, such as the effect of specific modifications on protein function, the conditions for the observation of PTMs, and the relationship between various modifications can be found; (b) The keywords (KW), where the "key" characteristics of the protein are listed. These are most useful when retrieving a selected set of proteins based on keywords associated with specific types of modifications. On the ExPASy server [9], we provide users with our definition of a keyword, as well as the list of all proteins linked to a particular keyword. This information is displayed by clicking once on the keyword directly in the protein entry; (c) The feature table (FT), where the data are separated into several classes, each with a specific feature key, *i.e.*, the first part of a feature line. The feature keys devoted to post-translational modifications are 'MOD_RES' (addition of small groups), 'CARBOHYD' (addition of carbohydrates), 'LIPID' (addition of lipids), 'DISULFID' (disulphide bonds) and 'CROSSLNK' (other cross-links). The 'METAL' feature key is used for both metal-coordinating sites, and the covalent binding of metal clusters. The 'BINDING' key is used for the covalent binding of larger molecular entities, such as hemes, or chromophores, in which case 'covalent' may appear in the feature description. The 'ACT_SITE' feature key is sometimes used to describe covalent modifications that are formed as intermediates during catalysis, in which case the description contains '. . . intermediate', such as 'phosphoserine intermediate'. Here also, the user can be provided with the definition of the feature key from the user manual by clicking on it once. The second part of the FT line locates the feature in the sequence, and the last part, the FT description, contains the name of the modification, except for the DISULFID FT lines. If the modification concerns specific splice isoforms or variants, its name, *e.g.* 'in isoform 3' or 'in variant B', is indicated. Quite recently, cross-references to GlycoSuiteDB [10], a database of glycoprotein glycan structures, have been integrated in feature lines. The licensed users can access directly the corresponding GlycoSuiteDB entry which contains additional information concerning the glycan bound at the given position (*e.g.* glycan structure, identification method, *etc.*), information which is outside the scope of a general database such as Swiss-Prot and is therefore not included as part of the annotation.

In addition to the CC, KW and FT lines, information concerning PTMs is also found in the reference position (RP) lines, where we indicate which data was annotated from the article (*e.g.*, 'methylation of Arg-x', 'phosphorylation' and 'carbohydrate structure'), the user can then access the desired PubMed abstract or the full-text paper by clicking on the appropriate link, and in database cross-reference (DR) lines, which point to relevant entries in databases, including specific PTM-oriented databases such as GlycoSuiteDB [10] and PhosSite (Phosphorylation Site DB).

5 Feature standardization

In order to increase the precision and accuracy of the feature annotations, the vocabulary used in the feature-description field is being standardized, revised and enhanced. To carry out this feature standardization, existing unique features have been identified, equivocal records are being resolved, and ambiguous records are being differentiated. The task is achieved by using the RESID database as a feature repository. Among other fields, it includes both Swiss-Prot and PIR feature-table annotations for each type of modification. The feature mapping facility in RESID not only allows us to revise the annotations but also facilitates the import of PIR experimental features into Swiss-Prot in the framework of the UniProt project [11]. This standardization work, done feature key by feature key, is gradually imposing a controlled vocabulary, thus allowing us to spot and immediately correct any nonstandard annotation. Standardization also ends up improving the efficiency of our proteomic tools that use the feature annotations (Table 2) [12–16]. Current controlled vocabularies for each of the standardized feature keys are provided in the Swiss-Prot user manual.

6 Sources of data

In the process of creating a Swiss-Prot entry, the corresponding computer-annotated TrEMBL entry [17] is edited and additional data from the literature and from various bioinformatic tools are integrated (see annbioch.txt at www.expasy.org/cgi-bin/lists?annbioch.txt). The resulting entry is therefore enriched with information, including the function, splice and mutation variants, PTMs, protein-protein interactions and subcellular location of the mature protein.

Table 2. ExPASy proteomic tools taking into account the annotated features [12]. All tools are accessible from http://www.expasy.org/tools/

PeptIdent	Identify proteins with peptide mass fingerprinting data, p*I* and Mw. Experimentally measured, user-specified peptide masses are compared with the theoretical peptides calculated for all proteins in Swiss-Prot, making extensive use of database annotations.
FindMod [13]	Predict potential protein PTMs and potential single amino acid substitutions in peptides. Experimentally measured peptide masses are compared with the theoretical peptides calculated from a specified Swiss-Prot entry or from a user-entered sequence, and mass differences are used to better characterize the protein of interest.
GlycoMod [15]	Predict possible oligosaccharide structures that occur on proteins from their experimentally determined masses (can be used for free or derivatized oligosaccharides and for glycopeptides).
FindPept [14]	Identify peptides that result from unspecific cleavage of proteins from their experimental masses, taking into account artefactual chemical modifications, PTMs and protease autolytic cleavages.
PeptideMass [16]	Calculate masses of peptides and their PTMs for a Swiss-Prot or TrEMBL entry or for a user sequence.

6.1 Experimental data

The paper(s) linked to a nucleotide sequence submission in the EMBL/GenBank/DDBJ databases are present in the TrEMBL entries and are generally the papers from which Swiss-Prot extracts information. However, they are often not the best source for data on PTM. When creating an entry, the annotator looks for additional information in pure characterization studies and review papers with the help of PubMed. Unfortunately, it is not an easy task to detect all papers which report information pertinent to PTMs. In addition, findings published after the creation of a Swiss-Prot entry are less likely to be integrated into the database, unless they are linked to a sequence submission, or submitted in the form of an user's update request and therefore given high priority. We highly encourage researchers to use this form, or to submit PTM data by email to swiss-prot@expasy.org, in order to help us to offer up-to-date data to all our users.

The experimental results are annotated in the corresponding protein entry and in the highly homologous protein entries; in the latter, the 'By similarity' qualifier is added at the end of the feature line. With the aim of capturing recently published data for existing Swiss-Prot entries, we have been developing a text-mining tool to extract desired information automatically from PubMed abstracts (provided that the modified position is clearly indicated, *e.g.*, phosphorylation of Tyr-460), and from proteomic papers. In proteomic papers, the results are often presented as a table containing the names of the examined proteins and the sequences of the short peptides containing modified residues. These tables are analyzed automatically, *i.e.*, the short peptide is mapped and the modified residue is localised in the Swiss-Prot's precursor or mature protein sequence. The main problem encountered at this stage is due to the heterogeneity of these tables relative to the identity of the proteins and to the labelling of the modified residues. The protein identifiers can be as diverse as Swiss-Prot identifiers (ID), Swiss-Prot/TrEMBL accession numbers (AC), EMBL/GenBank identifiers, LocusLink identifiers, gene names, ORF numbers, *etc.* Regarding the label of the PTM, the modified residue can be underlined, in bold font, preceded by a special character, *e.g.* 'p' for phosphorylation, or even written with a special code, *e.g.*, 'DMA' for dimethylarginine. The data extraction from these tables would be greatly simplified if the authors used Swiss-Prot/TrEMBL accession numbers, and specific characters or inserted characters for, respectively, the identification of the protein, and the labelling of the modified residue. After manual checking of the output, the studied protein and its close homologues are annotated without and with the qualifier 'By similarity', respectively. This tool has been used successfully to integrate dozens of phosphorylation sites reported in traditional research papers as well as in proteomic papers. Experimental data are also extracted from the protein 3D-structure database PDB [18] and from PTM-specialized databases, such as GlycoSuiteDB [10] and Phosphorylation

Site DB, whereas direct data submission and update requests by authors often lead to the global update of a protein or a protein family.

6.2 PTM predictions and automated annotation

Protein sequence analysis tools for the prediction of post-translational modifications have been developed and made available to the scientific community by various groups including the Swiss Institute of Bioinformatics (SIB). For the annotation process we make use of only a limited number of tools (Table 3) [19–23], since many methods produce a high level of false positive hits. The predictions made by these programs are carefully checked by experienced annotators and introduced in the relevant entries tagged by the qualifier 'Potential'.

Before introducing a new PTM prediction program for Swiss-Prot annotation, it is evaluated. Once we have decided to make use of it, we apply it to all the protein sequences to which we expect the result to be relevant. Recently we undertook a major overhaul of the annotation of proteins that are attached to the membrane by a GPI-anchor, but for which the lipid-binding site had not been determined experimentally. The GPI-attachment sites predicted by the big-PI tool [19] were first checked by an experienced annotator and then integrated into the entry of the analyzed protein with the feature tagged by the nonexperimental qualifier 'Potential'.

Rule-based systems for the automated annotation are currently under development for further customization of the annotator's workbench (for example see HAMAP [24]). Bioinformatic tools are extremely helpful, once it is known that a certain modification exists or is likely to appear in a defined module or protein family, since we are looking for relevant PTM analysis results only in this context. In order to constrain the prediction tools, annotation rules must therefore be created. For the automated annotation we embed the PTM annotation into the module- or protein-specific annotation rules, which can then be applied to annotate related proteins. Another solution, more conducive to the maintenance of rules, is the creation of PTM-specific rules, which can then be called from within relevant protein or domain rules, the latter being essential for the automated annotation of proteins with complex domain architecture.

7 Annotation examples

7.1 Glycosylation

Protein glycosylation is the second most common type of post-translational modification. It plays a role in many cellular processes such as protein folding, localization and trafficking, protein solubility, antigenicity, biological half-life and cell-cell interactions [25]. The annotation of glycoproteins in Swiss-Prot was revised two years ago [26]. At that time it was decided to classify the types of glycosylation according to the atom to which the glycan is attached, *i.e.*, '*N*-linked' when bound to nitrogen, 'O-linked' when bound to oxygen and '*C*-linked' when bound to carbon, all described by using a 'CARBOHYD' FT key. The glycosyl or reducing terminal monosaccharide identity is provided if known, followed by three dots if the glycan is a polymer (Fig. 1). We draw the reader's attention to the fact that the figures which illustrate the annotation examples below are taken from Swiss-Prot release 43, and that the syntax of the feature (FT) lines may have changed since then.

Table 3. Prediction tools currently used in the Swiss-Prot annotation

SignalP [22][a)]	Signal peptide cleavage sites
TargetP [23][a)]	Chloroplast transit peptide (cTP), mitochondrial targeting peptide (mTP) or secretory pathway signal peptide (SP) cleavage sites
PROSITE [27] patterns[b)], including	*N*-glycosylation sites in eukaryotic proteins (PS00001) Asx hydroxylation site (PS00010) Phosphopantetheine attachment site (PS00012)
Big-PI [19][c)]	GPI-anchor modification sites
NMT [20][c)]	*N*-myristoylation sites in eukaryotic proteins
Sulfinator [21][b)]	Tyrosine sulfation sites

a) provided by the Center for Biological Sequence analysis, Denmark
b) provided by the Swiss Institute of Bioinformatics, Switzerland
c) provided by the Bioinformatics group at the Research Institute of Molecular Pathology, Austria


```
ID   TSP1_HUMAN     STANDARD;      PRT;  1170 AA.
AC   P07996; Q15667;
DT   01-AUG-1988 (Rel. 08, Created)
DT   01-AUG-1988 (Rel. 08, Last sequence update)
DT   15-SEP-2003 (Rel. 42, Last annotation update)
DE   Thrombospondin 1 precursor.
GN   THBS1 OR TSP1 OR TSP.
..
RN   [7]
RP   CARBOHYDRATE-LINKAGE SITES TRP-385; SER-394; TRP-438; TRP-441;
RP   THR-450; TRP-498 AND THR-507.
RC   TISSUE=Platelet;
RX   MEDLINE=21125860; PubMed=11067851;
RA   Hofsteenge J., Huwiler K.G., Macek B., Hess D., Lawler J.,
RA   Mosher D.F., Peter-Katalinic J.;
RT   "C-mannosylation and O-fucosylation of the thrombospondin type 1
RT   module.";
RL   J. Biol. Chem. 276:6485-6498(2001).
..
DR   GlycoSuiteDB; P07996; -.
..
KW   Glycoprotein; Cell adhesion; Calcium-binding; Heparin-binding; Repeat;
KW   EGF-like domain; Signal; 3D-structure.
FT   SIGNAL        1     18
FT   CHAIN        19   1170       THROMBOSPONDIN 1.
..
FT   CARBOHYD    248    248       N-LINKED (GLCNAC...).
FT   CARBOHYD    360    360       N-LINKED (GLCNAC...) (POTENTIAL).
FT   CARBOHYD    385    385       C-LINKED (MAN).
FT                                /FTId=CAR_000205.
FT   CARBOHYD    394    394       O-LINKED (FUC...).
FT                                /FTId=CAR_000206.
FT   CARBOHYD    438    438       C-LINKED (MAN).
FT                                /FTId=CAR_000207.
FT   CARBOHYD    441    441       C-LINKED (MAN).
FT                                /FTId=CAR_000208.
FT   CARBOHYD    450    450       O-LINKED (FUC...).
FT                                /FTId=CAR_000209.
FT   CARBOHYD    498    498       C-LINKED (MAN).
FT                                /FTId=CAR_000210.
FT   CARBOHYD    507    507       O-LINKED (FUC...).
FT                                /FTId=CAR_000211.
FT   CARBOHYD    708    708       N-LINKED (GLCNAC...) (POTENTIAL).
FT   CARBOHYD   1067   1067       N-LINKED (GLCNAC...) (POTENTIAL).
..
SQ   SEQUENCE   1170 AA;  129412 MW;  69B3EDE5AE3A395E CRC64;
..
//
```

Figure 1. Example of a Swiss-Prot entry containing *N*-, *C*- and O-linked glycans. Three *N*-linked glycosylation sites are predicted by matching with the PS00001 PROSITE pattern (note the qualifier 'Potential').

The acceptor site Asn-Xaa-Ser/Thr-Xaa (Xaa not Pro) is used to predict *N*-linked glycosylation in eukaryotic secreted or membrane-bound proteins. The precursor sequence must include a type I or type II signal peptide, and only the extracellular sites are annotated. In the Swiss-Prot annotation process, all suitable proteins are screened by matching with the PROSITE [27] pattern PS00001 for potential *N*-glycosylation sites, but only probable and certified glycoproteins get the keyword 'glycoprotein'.

7.2 Lipid addition

Modifications resulting from the covalent attachment of a lipid moiety are described in a 'LIPID' FT line and the corresponding entries contain the keyword 'Lipoprotein'. The lipid feature descriptions have been recently modified to contain the name of the modified amino acid and also, when necessary, the atom of attachment, rather than just that of the added group, *e.g.*, 'S-palmitoyl cysteine' instead of 'PALMITATE'.

An important protein lipid modification consists in the attachment of a GPI-anchor, composed of a tetraglycan and phosphatidylinositol, which anchors an otherwise secreted protein to the membrane of eukaryotic cells, on the extracellular side. Many proteins are experimentally shown to be GPI-anchored *via* their release from the cell surface by phosphatidylinositol-specific phospholipase (PI-PLC), without further characterization of the modification site. These proteins and their homologues are analyzed with the big-PI [19] prediction tool, and the sites predicted with a satisfactory score are annotated as puta-

tive attachment sites of a GPI-anchor. This modification not only consists of the attachment of a GPI-anchor but also of the cleavage of a *C*-terminal propeptide, annotated in a 'PROPEP' FT line. Being extracellular, the protein sequence must include a signal peptide, described in a 'SIGNAL' FT line. Thus, the entry of a GPI-anchored protein contains the keywords 'Signal', 'GPI-anchor', 'Lipoprotein' and 'Glycoprotein' (Fig. 2).

Besides GPI-anchor attachment, a number of lipid modifications of proteins which occur in the cytoplasm are important in protein cellular targeting and cellular regulation. These include myristoylation, palmitoylation, and prenylation.

N-terminal myristoylation consists in the attachment of myristic acid to the *N*-terminal glycine of eukaryotic proteins in the cytosol, after the removal of the initiator methionine or, more rarely, the cleavage of a propeptide. This modification increases the hydrophobicity of the modified protein, and potentially influences its interactions with intracellular substructures or proteins. When additionally palmitoylated on one or more nearby cysteine side chains, the protein becomes anchored to the cell membrane through insertion of the fatty acids in the lipid bilayer. In predicting this modification, the eukaryotic proteins whose sequences begin with the methionine-glycine dipeptide are analyzed with the NMT prediction tool [20] and the predicted sites, as well as the indication that the initiator methionine is potentially removed, are annotated in the entries (*i.e.*, *N*-myristoyl glycine) and the keyword 'Myristate' is added (Fig. 3).

Palmitoylation is the modification of an amino acid by palmitic acid. So far, it has been observed to occur on cysteine side chains (S-palmitoyl cysteine) and *N*-terminal cysteines (*N*-palmitoyl cysteine), on serine or threonine side chains (*O*-palmitoyl serine or threonine), and on lysine side chains (*N*(6)-palmitoyl lysine). The keyword 'Palmitate' is added (Fig. 3). In eukaryotes, S-palmitoyl

```
ID   PPBI_RAT       STANDARD;      PRT;   540 AA.
AC   P15693;
DT   01-APR-1990 (Rel. 14, Created)
DT   01-APR-1990 (Rel. 14, Last sequence update)
DT   15-SEP-2003 (Rel. 42, Last annotation update)
DE   Alkaline phosphatase, intestinal 1 precursor (EC 3.1.3.1) (IAP-I).
..
RN   [1]
RP   SEQUENCE FROM N.A., AND SEQUENCE OF 21-34 AND 287-300.
RX   MEDLINE=90167110; PubMed=2155025;
RA   Lowe M., Strauss A.W., Alpers R., Seetharam S., Alpers D.H.;
RT   "Molecular cloning and expression of a cDNA encoding the membrane-
RT   associated rat intestinal alkaline phosphatase.";
RL   Biochim. Biophys. Acta 1037:170-177(1990).
RN   [2]
RP   GPI-ANCHOR.
RX   MEDLINE=95263539; PubMed=7744844;
RA   Engle M.J., Mahmood A., Alpers D.H.;
RT   "Two rat intestinal alkaline phosphatase isoforms with different
RT   carboxyl-terminal peptides are both membrane-bound by a glycan
RT   phosphatidylinositol linkage.";
RL   J. Biol. Chem. 270:11935-11940(1995).
..
CC   -!- SUBCELLULAR LOCATION: Attached to the membrane by a GPI-anchor.
..
KW   Hydrolase; Zinc; Magnesium; Phosphorylation; Transmembrane;
KW   Multigene family; Glycoprotein; GPI-anchor; Signal; Lipoprotein.
FT   SIGNAL        1     20
FT   CHAIN        21    511       ALKALINE PHOSPHATASE, INTESTINAL 1.
FT   PROPEP      512    540       REMOVED IN MATURE FORM (POTENTIAL).
FT   DISULFID    141    203       BY SIMILARITY.
FT   DISULFID    487    494       BY SIMILARITY.
FT   ACT_SITE    112    112       PHOSPHOSERINE INTERMEDIATE
FT                                (BY SIMILARITY).
FT   LIPID       511    511       GPI-anchor amidated asparagine
FT                                (Potential).
FT   CARBOHYD    142    142       N-LINKED (GLCNAC...) (POTENTIAL).
FT   CARBOHYD    301    301       N-LINKED (GLCNAC...) (POTENTIAL).
FT   CARBOHYD    428    428       N-LINKED (GLCNAC...) (POTENTIAL).
SQ   SEQUENCE   540 AA;  58402 MW;  29AC2B543CBE6B52 CRC64;
..
//
```

Figure 2. Example of a Swiss-Prot entry containing a GPI-anchor and three predicted *N*-linked glycosylation sites. The position of the GPI-anchor attachment is a prediction made by big-PI (note the qualifier 'Potential'). The signal peptide length is known by amino acid sequencing. The active site and disulphide bond positions are inferred by similarity with the murine orthologue (AC P24822) and a human paralogue (AC P05187), respectively (note the qualifier 'By similarity').


```
ID   FYN_HUMAN      STANDARD;      PRT;   536 AA.
AC   P06241;
DT   01-JAN-1988 (Rel. 06, Created)
DT   01-FEB-1994 (Rel. 28, Last sequence update)
DT   15-SEP-2003 (Rel. 42, Last annotation update)
DE   Proto-oncogene tyrosine-protein kinase FYN (EC 2.7.1.112) (P59-FYN)
DE   (SYN) (SLK).
GN   FYN.
..
RN   [3]
RP   MYRISTOYLATION, AND PHOSPHORYLATION OF TYR-530.
RX   MEDLINE=91016431; PubMed=1699196;
RA   Peters D.J., McGrew B.R., Perron D.C., Liptak L.M., Laudano A.P.;
RT   "In vivo phosphorylation and membrane association of the fyn proto-
RT   oncogene product in IM-9 human lymphoblasts.";
RL   Oncogene 5:1313-1319(1990).
..
KW   Proto-oncogene; Transferase; Tyrosine-protein kinase; Phosphorylation;
KW   ATP-binding; Myristate; SH3 domain; SH2 domain; Palmitate;
KW   Lipoprotein; 3D-structure; Polymorphism.
FT   INIT_MET      0      0
FT   LIPID         1      1       N-myristoyl glycine.
FT   LIPID         2      2       S-palmitoyl cysteine (By similarity).
FT   LIPID         5      5       S-palmitoyl cysteine (By similarity).
..
FT   MOD_RES      11     11       PHOSPHORYLATION (BY PKC) (BY SIMILARITY).
..
FT   MOD_RES     419    419       PHOSPHORYLATION (AUTO-) (BY SIMILARITY).
FT   MOD_RES     530    530       PHOSPHORYLATION.
FT   VARIANT     444    444       I -> F (in dbSNP:1801121).
FT                                /FTId=VAR_014661.
..
SQ   SEQUENCE   536 AA;  60630 MW;  57436625365A0977 CRC64;
..
//
```

Figure 3. Example of a Swiss-Prot entry for a myristoylated and palmitoylated phosphoprotein. *N*-terminal myristoylation and one phosphorylation site are experimental results. Two palmitoylation sites on cysteine side chains and two additional phosphorylation sites are inferred by similarity with homologous proteins (note the qualifier 'By similarity'). For one phosphorylation site, the kinase is known (note the comment '(by PKC)'). The initiatior methionine is removed prior to *N*-myristoylation (note the 'INIT_MET 0 0' line).

cysteine is often found in association with *N*-myristoyl glycine in cytosolic proteins, where it strengthens membrane anchorage. It is also frequently observed alone in cytoplasmic parts of integral membrane proteins, where it probably serves to keep the cytoplasmic segment close to the membrane. For the convenience of our users, a controlled vocabulary list for the lipid modifications is provided in the Swiss-Prot user manual.

7.3 Addition of small groups

Modifications of amino acids by addition of hydroxyl, methyl or acetyl groups and by phosphate or sulphate, to cite only a few, are annotated in a 'MOD_RES' FT line. In this category, the annotation is also being standardized in the same way as has been done in the 'LIPID' features, *i.e.*, to show the product name instead of the process name (note: the revised features are in mixed case, whereas the ones awaiting standardization remain in upper case).

The only tool we currently use to predict modifications of this type is the Sulfinator [21]. This tool is used to predict sulphated tyrosines in extracellular parts of proteins. The keyword 'Sulfation' is assigned to sulphated proteins (Fig. 4). Extensively studied, because of their involvement in cellular signalling, the serine/threonine and tyrosine phosphorylations are catalyzed by protein kinases in the cytoplasm or the nucleus. In the feature line, we optionally indicate the name of the kinase found to phosphorylate the protein at a given position. The difficulty in predicting the position of a phosphorylation site resides in the variety of existing narrow- and broad-specificity kinases. We are therefore not using any prediction tool to analyze phosphoproteins. However, we annotate data resulting from *in vitro* assays, provided that two other pieces of evidence are present: (i) the protein is effectively phosphorylated *in vivo*, and (ii) the kinase used in the study exists in the corresponding organism and cellular compartment – in which case the feature is tagged with a comment '*in vitro*' (Fig. 5).

7.4 Cross-linking of two amino acids

Cross-linking is defined as the binding of two amino acids and results in the tight attachment of two parts of the protein or of two protein molecules, which itself results in the formation of a dimer. The disulphide bonds are described in 'DISULFID' FT lines, whereas the other cross-link features are described in 'CROSSLNK' FT lines. A controlled vocabulary list of cross-links is provided as part of the Swiss-Prot user manual. When


```
ID   FA10_BOVIN     STANDARD;      PRT;   492 AA.
AC   P00743;
DT   21-JUL-1986 (Rel. 01, Created)
DT   13-AUG-1987 (Rel. 05, Last sequence update)
DT   15-SEP-2003 (Rel. 42, Last annotation update)
DE   Coagulation factor X precursor (EC 3.4.21.6) (Stuart factor).
GN   F10.
..
RN   [2]
RP   SEQUENCE OF 41-180.
RX   MEDLINE=80130563; PubMed=6766735;
RA   Enfield D.L., Ericsson L.H., Fujikawa K., Walsh K.A., Neurath H.,
RA   Titani K.;
RT   "Amino acid sequence of the light chain of bovine factor X1 (Stuart
RT   factor).";
RL   Biochemistry 19:659-667(1980).
..
RN   [5]
RP   SEQUENCE OF 183-233, AND CARBOHYDRATE-LINKAGE SITES.
RX   MEDLINE=94062825; PubMed=8243461;
RA   Inoue K., Morita T.;
RT   "Identification of O-linked oligosaccharide chains in the activation
RT   peptides of blood coagulation factor X. The role of the carbohydrate
RT   moieties in the activation of factor X.";
RL   Eur. J. Biochem. 218:153-163(1993)...
..
RN   [7]
RP   PROCESSING.
RX   MEDLINE=76053121; PubMed=1059122;
RA   Fujikawa K., Titani K., Davie E.W.;
RT   "Activation of bovine factor X (Stuart factor): conversion of factor
RT   Xa-alpha to factor Xa-beta.";
RL   Proc. Natl. Acad. Sci. U.S.A. 72:3359-3363(1975).
..
RN   [9]
RP   SULFATION.
RX   MEDLINE=86140210; PubMed=3949800;
RA   Morita T., Jackson C.M.;
RT   "Localization of the structural difference between bovine blood
RT   coagulation factors X1 and X2 to tyrosine 18 in the activation
RT   peptide.";
RL   J. Biol. Chem. 261:4008-4014(1986).
..
CC   -!- PTM: The vitamin K-dependent, enzymatic carboxylation of some
CC       glutamate residues allows the modified protein to bind calcium.
CC   -!- PTM: N- and O-glycosylated.
CC   -!- PTM: The activation peptide is cleaved by factor IXA (in the
CC       intrinsic pathway), or by factor VIIA (in the extrinsic pathway).
..
DR   GlycoSuiteDB; P00743; -.
..
KW   Glycoprotein; Hydrolase; Serine protease; Plasma; Blood coagulation;
KW   Gamma-carboxyglutamic acid; Hydroxylation; Calcium-binding; Vitamin K;
KW   Signal; Zymogen; EGF-like domain; Repeat; Sulfation; 3D-structure.
FT   SIGNAL        1     23       POTENTIAL.
FT   PROPEP       24     40
FT   CHAIN        41    180       FACTOR X LIGHT CHAIN.
FT   CHAIN       183    492       FACTOR X HEAVY CHAIN.
FT   PROPEP      183    233       ACTIVATION PEPTIDE.
FT   CHAIN       234    492       ACTIVATED FACTOR XA, HEAVY CHAIN.
..
FT   MOD_RES     200    200       SULFATION (partial).
FT   CARBOHYD    208    208       O-LINKED (GALNAC...).
FT   CARBOHYD    218    218       N-LINKED (GLCNAC...).
FT                                /FTId=CAR_000011.
..
SQ   SEQUENCE   492 AA;  54510 MW;  D5BD911FB72F1D30 CRC64;
..
//
```

Figure 4. Example of a Swiss-Prot entry for a sulphated protein with additional *N*- and *O*-linked carbohydrates, all experimental. The protein precursor is a preproprotein, which becomes a mature protein after proteolytic processing (note the 'SIGNAL' and 'PROPEP' lines). Note also the comment 'partial' indicating that not all protein molecules are sulphated.


```
ID   GELS_HUMAN     STANDARD;      PRT;   782 AA.
AC   P06396; Q8WVV7;
DT   01-JAN-1988 (Rel. 06, Created)
DT   01-JAN-1988 (Rel. 06, Last sequence update)
DT   15-SEP-2003 (Rel. 42, Last annotation update)
DE   Gelsolin precursor, plasma (Actin-depolymerizing factor) (ADF)
DE   (Brevin) (AGEL).
GN   GSN.
..
RN   [9]
RP   PHOSPHORYLATION OF TYR-86; TYR-409; TYR-465; TYR-603 AND TYR-651.
RX   MEDLINE=99224907; PubMed=10210201;
RA   De Corte V., Demol H., Goethals M., Van Damme J., Gettemans J.,
RA   Vandekerckhove J.;
RT   "Identification of Tyr438 as the major in vitro c-Src phosphorylation
RT   site in human gelsolin: a mass spectrometric approach.";
RL   Protein Sci. 8:234-241(1999).
..
CC   -!- FUNCTION: Calcium-regulated, actin-modulating protein that binds
CC       to the plus (or barbed) ends of actin monomers or filaments,
CC       preventing monomer exchange (end-blocking or capping). It can
CC       promote the assembly of monomers into filaments (nucleation) as
CC       well as sever filaments already formed.
..
CC   -!- SUBCELLULAR LOCATION: Cytoplasmic (isoform 2); secreted (isoform
CC       1).
CC   -!- ALTERNATIVE PRODUCTS:
CC       Event=Alternative initiation;
CC         Comment=2 isoforms, 1/Secreted/Plasma (shown here) and
CC         2/Cytoplasmic, are produced by alternative initiation;
..
CC   -!- PTM: Phosphorylation on Tyr-86, Tyr-409, Tyr-465, Tyr-603 and Tyr-
CC       651 in vitro is induced in presence of phospholipids.
..
KW   Cytoskeleton; Actin-binding; Repeat; Calcium; Alternative initiation;
KW   Signal; Amyloid; Disease mutation; 3D-structure; Phosphorylation;
KW   Actin capping.
FT   SIGNAL        1     27
FT   CHAIN        28    782       GELSOLIN, ISOFORM 1.
FT   CHAIN        53    782       GELSOLIN, ISOFORM 2.
FT   INIT_MET     52     52       FOR ISOFORM 2.
..
FT   DISULFID    215    228       IN ISOFORM 1.
FT   MOD_RES      86     86       PHOSPHORYLATION (BY SRC) (IN ISOFORM 2)
FT                                (IN VITRO).
FT   MOD_RES     409    409       PHOSPHORYLATION (BY SRC) (IN ISOFORM 2)
FT                                (IN VITRO).
FT   MOD_RES     465    465       PHOSPHORYLATION (BY SRC) (IN ISOFORM 2)
FT                                (MAJOR) (IN VITRO).
FT   MOD_RES     603    603       PHOSPHORYLATION (BY SRC) (IN ISOFORM 2)
FT                                (IN VITRO).
FT   MOD_RES     651    651       PHOSPHORYLATION (BY SRC) (IN ISOFORM 2)
FT                                (IN VITRO).
..
SQ   SEQUENCE   782 AA;  85697 MW;  8CEBC52257A160F7 CRC64;
..
//
```

Figure 5. Example of a Swiss-Prot entry for a phosphoprotein with *in vitro* phosphorylation sites annotated (note the comment *in vitro* in the corresponding feature line). Note also that two isoforms are produced by the use of two alternative initiation methionines, leading to their targeting to two different locations. Isoform 1 contains a signal peptide and is secreted whereas isoform 2 is cytoplasmic and becomes phosphorylated.

a cross-link occurs between two protein molecules, it is simply indicated by the word 'interchain' in the feature description (Fig. 6).

8 Usage of Swiss-Prot annotations

8.1 Retrieval of particular sets of entries

The sequence retrieval system (SRS) [28] can be used to search for entries or features relevant to a given modification in Swiss-Prot. Every field of a Swiss-Prot entry is indexed in SRS, thus allowing the users to search for special strings. We want to draw the reader's attention to the concept of subentries used by SRS to permit more specific querying. A subentry is the coupling of two independent entities such as the elements of a feature line, *i.e.*, the feature key and the feature description. The search is thus performed on individual feature lines, instead of the whole feature table. Depending on which version of SRS is used (*i.e.*, SRS5 or SRS7), it is possible to define one subentry (*e.g.*, SRS5 where the features are considered as subentries), or several subentries (*e.g.*,


```
ID   QADG_PSEPK     STANDARD;      PRT;    78 AA.
ID   QADG_PSEPU
AC   Q88HA0; Q8VW83;
DT   15-SEP-2003 (Rel. 42, Created)
DT   15-SEP-2003 (Rel. 42, Last sequence update)
DT   15-SEP-2003 (Rel. 42, Last annotation update)
DE   Quinohemoprotein amine dehydrogenase gamma subunit (EC 1.4.99.-)
DE   (Quinohemoprotein amine dehydrogenase 9 kDa subunit) (Quinohemoprotein
DE   amine dehydrogenase catalytic subunit) (QH-AmDH).
GN   QHNDH OR PP3460.
..
RN   [2]
RP   SEQUENCE FROM N.A., SEQUENCE, AND MASS SPECTROMETRY.
RC   STRAIN=ATCC 12633, and IFO 15633;
RX   MEDLINE=21560975; PubMed=11555656;
RA   Vandenberghe I., Kim J.-K., Devreese B., Hacisalihoglu A., Iwabuki H.,
RA   Okajima T., Kuroda S., Adachi O., Jongejan J.A., Duine J.A.,
RA   Tanizawa K., van Beeumen J.;
RT   "The covalent structure of the small subunit from Pseudomonas putida
RT   amine dehydrogenase reveals the presence of three novel types of
RT   internal cross-linkages, all involving cysteine in a thioether
RT   bond.";
RL   J. Biol. Chem. 276:42923-42931(2001).
..
CC   -!- COFACTOR: CTQ.
..
CC   -!- PTM: The cysteine tryptophylquinone (CTQ) is generated by
CC       oxidation of the indole ring of a tryptophan residue to form
CC       tryptophylquinone, followed by covalent cross-linking with a
CC       cysteine residue.
CC   -!- PTM: Met-29 is oxidized to methionine sulfoxide, probably during
CC       sample preparation.
CC   -!- MASS SPECTROMETRY: MW=8486.5; METHOD=Electrospray; RANGE=1-78.
CC   -!- MASS SPECTROMETRY: MW=8504.0; METHOD=Electrospray; RANGE=Isoform
CC       with methionine sulfoxide.
CC   -!- MISCELLANEOUS: Is propably co-translocated into the periplasm when
CC       associated with the alpha and/or beta subunit, which contain both
CC       a signal peptide.
..
KW   Oxidoreductase; Electron transport; Periplasmic; CTQ; 3D-structure;
KW   Complete proteome.
FT   INIT_MET      0      0
FT   ACT_SITE     32     32       BASE.
FT   CROSSLNK      6     15       4-cysteinyl-glutamic acid (Cys-Glu).
FT   CROSSLNK     26     32       3-cysteinyl-aspartic acid (Cys-Asp).
FT   CROSSLNK     40     48       3-cysteinyl-aspartic acid (Cys-Asp).
FT   CROSSLNK     36     42       4'-cysteinyl-tryptophylquinone (Cys-Trp).
FT   MOD_RES      42     42       Tryptophylquinone.
..
SQ   SEQUENCE   78 AA;  8466 MW;  1A30663894AB2DFE CRC64;
..
//
```

Figure 6. Example of a Swiss-Prot entry containing three different cross-links. The name of the chemical compound is indicated ('4-cysteinyl-glutamic acid'), and the identity of the two amino acids is given ('Cys-Glu') for quality control and clarity purposes.

SRS7 where the publication references, comments, database cross-references, features and counter are considered as subentries). This option is activated by selecting 'feature' in the 'retrieve set of' or 'Get results of type' field on the ExPASy server or the EBI's server, respectively, instead of the default 'entry' specification. To illustrate this, let us say you are looking for a protein with a lipid bound to a predicted position (*i.e.*, with the qualifier 'Potential'). If the 'retrieve set of feature' is selected, the output contains 'LIPID' feature lines with the 'Potential' qualifier, whereas if the 'retrieve set of entries' is chosen, the output contains entries with a 'LIPID' FT key and a 'Potential' anywhere in the feature table.

8.2 Datasets for PTM prediction tool development

The prediction of post-translational modification sites is an essential bioinformatic task that helps quite significantly the functional characterization of a protein [29]. A prediction tool is obtained through the definition of a sequence motif that sets the probabilities for given amino acids to be present at the modification site and in the surrounding sequence. Most of the recent programs developed to predict PTM occurrence and localization use machine learning or statistical methods such as Artificial Neural Networks or Hidden Markov Models [30].

These methods operate through a learning process with positive and negative datasets, meaning sets of protein sequences or subsequences that respectively carry or do not carry the modification. For training purposes, it is really essential to construct clean datasets. Ideally, the positive set should only consist of protein sites where experimental proof of their modification has been found. The users can retrieve the concerned sequences with SRS, by searching Swiss-Prot entries that contain the specific feature line without any qualifier. In practice, these entries are selected by using the 'feature subentry' option as described in Section 8.2 (*e.g.*, the query '[swiss_prot-FtDescription: sulfotyrosine*]![swiss_prot-FtDescription: potential*]![swiss_prot-FtDescription: probable*]![swiss_prot-FtDescription: similarity*]' retrieves all proteins with an experimentally-proved sulphated tyrosine). In the future we plan to indicate positive information to features, such as 'experimental', to simplify this procedure for users.

Unfortunately, for many PTMs, the experimental results are too sparse to build a sufficient set for program training, requiring the consideration of PTM inferred by similarity. These can be included on the condition that the surrounding sequence diverges within a certain degree from that of its experimentally studied homologue (in order to avoid bias due to the presence of identical sequences). Methods that filter sequences according to the similarity level have been developed [31, 32]. On the other hand, PTM features qualified by a 'Potential' tag must not be included in the positive set as they already result from a prediction. We also want to draw the reader's attention to the fact that the dataset – whether it is created from Swiss-Prot features or not – must be carefully checked in order to be used in further steps of tool designing, so as not to introduce incorrect sequence motifs. With the aim of decreasing the number of false positives, programs should also be trained against a collection of sequences that do not carry the modification: the negative dataset.

Creating a negative dataset is a difficult task, since experimental negative results are rarely described in scientific papers. However, candidate proteins that have been sequenced at amino acid level or whose mass has been measured precisely by mass spectroscopy can be good negative set elements, as protein modifications would have been detected in these experiments. Indeed, the correspondence of the experimentally measured mass with the theoretical molecular weight of the peptide is strong evidence for the absence of modification. There remains yet another requirement, which the protein must absolutely fulfil in order to be included in the negative set: it must have the potentiality to be modified. That is to say, it must be located in the same cellular compartment as the modification enzyme, in order to make sure that the sequence motif is not recognized by the enzyme. In the future, we plan to add a negative indication to features which were indeed expected but turned out to be absent, from an experimental point of view.

8.3 Usage of feature annotations by proteomic tools

It is now widely acknowledged that most human proteins are post-translationally modified. Effectively, maybe all extracellular proteins bear *N*-linked glycans, probably all cytosolic proteins lack the intitator methionine, and nuclear proteins are often modified (*e.g.*, methylation of ribonucleoproteins, oxidation of transcription factors). Yet only 75% of human entries have PTMs annotated in Swiss-Prot. Even though there is a discrepancy between the number of PTMs which occur in nature, the number of known and studied PTMs, and the number of PTMs annotated in Swiss-Prot, the amount of relevant information in the database has increased considerably over the last few years. Common protein identification and characterization techniques include methods based on peptide mass fingerprinting [33]. Identification tools used in this context might miss peptides carrying PTMs because most of the tools work with theoretical masses of unmodified peptides. In the worst case, this might even prevent a highly modified protein from being picked up by an identification algorithm.

The tools provided by the ExPASy team [12–16] try to use a maximum amount of information from Swiss-Prot. In particular, they read the Swiss-Prot feature tables and parse out information on processing into mature chains and peptides, and on the addition of simple groups such as methyl, acetyl, palmitate and thirty odd other groups. The mass difference induced by each of these modifications is known and the tools can therefore model as accurately as possible the active processed and modified protein, instead of just using its precursor sequence translated from the gene.

9 Future developments

As the amount of data on protein modifications will increase, a relational database of protein modifications will be built to provide users and annotators with highly curated information on every biologically relevant aspect of a given protein modification. This database will be filled and updated *via* an internally developed data capture web interface (with controlled vocabulary and data consistency check, whenever possible). The stored information will be displayed on the ExPASy web server *via* HTML pages called from Swiss-Prot entries, proteomic tools

Table 4. Classified information of Mod-Prot entries

General description	Name(s) of modification, category, reversibility, *etc.*
Mass variation	Chemical composition modification (this information will mostly be imported from the RESID database), number of added units, *etc.*
Target	Amino acid(s), chemical group, environment
Occurrence rules	Rules governing the relationships between coexisting modification(s)
Taxonomic range	Information about the taxons in which the modification has been shown to exist, or not
Subcellular location of modified protein	Cellular location of the mature protein
Description of chemical reaction	Type of reaction, description of enzymes involved, subcellular location of reaction, *etc.*
Function	Function of the modification.
Associated disease	Short description of a disease due to the modification of the extent of a given modification.
Swiss-Prot annotation	Description of the annotations relative to the modification
Examples of modified proteins	Examples of paradigm annotation
Cross-references	General or specialized protein-modification databases
Prediction program	Neural network, hidden Markov model, pattern, *etc.*
References	Reviews, research papers giving information about the modification

or an upcoming protein modification browser. According to Swiss-Prot quality criteria, information on a protein modification is classified in several fields (Table 4). The so-called 'Mod-Prot' protein modification server will be available on ExPASy in the first semester of 2004.

10 Concluding remarks

We have described many aspects of the annotation of post-translational modifications in Swiss-Prot, *i.e.*, their meaning, the way they can be used, the work that has been done and the future developments we plan to do in this field. We place a strong emphasis on integration with other biomolecular data repositories. We therefore strongly encourage any interested database, 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. We also count on direct submissions of results or update requests from the authors themselves, not exactly to help us update entries but rather to allow us determine which proteins are to be updated with the highest priority.

The authors gratefully acknowledge the NIH (U01 HG02712-01) and the Swiss Confederation (grant 2000-2003) for supporting our work, Ursula Hinz (Swiss Institute of Bioinformatics) for useful discussions and Vivienne Baillie Gerritsen (Swiss Institute of Bioinformatics) for proof-reading of the manuscript.

11 References

[1] Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C. *et al.*, *Nucleic Acids Res.* 2003, *31*, 365–370.

[2] Han, K. K., Martinage, A., *Int. J. Biochem.* 1992, *24*, 1928.

[3] Garavelli, J. S., *Nucleic Acids Res.* 2003, *31*, 499–501.

[4] Kamemura, K., Hart, G. W., *Prog. Nucleic Acid Res. Mol. Biol.* 2003, *73*, 107–136.

[5] Godovac-Zimmermann, J., Brown, L. R., *Mass Spectrom. Rev.* 2001, *20*, 1–57.

[6] Sanchez, J.-C., Wirth, P., Jaccoud, S., Appel, R. D. *et al.*, *Electrophoresis* 1997, *18*, 638–641.

[7] Schweppe, R. E., Haydon, C. E., Lewis, T. S., Resing, K. A., Ahn, N. G., *Acc. Chem. Res.* 2003, *36*, 453–461.

[8] Gasteiger, E., Jung, E., Bairoch, A., *Curr. Issues Mol. Biol.* 2001, *3*, 47–55.

[9] Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I. *et al.*, *Nucleic Acids Res.* 2003, *31*, 3784–3788.

[10] Cooper, C. A., Joshi, H. J., Harrison, M. J., Wilkins, M. R., Packer, N. H., *Nucleic Acids Res.* 2003, *31*, 511–513.

[11] Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., *Nucleic Acids Res.* 2004, *32*, D115–D119.

[12] Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S. *et al.*, *Proteomics Handbook*, Humana Press, Totowa 2004, in press.

[13] Wilkins, M. R., Gasteiger, E., Gooley, A., Herbert, B. *et al.*, *J. Mol. Biol.* 1999, *289*, 645–657.

[14] Gattiker, A., Bienvenut, W. V., Bairoch, A., Gasteiger, E., *Proteomics* 2002, *2*, 1435–1444.

[15] Cooper, C. A., Gasteiger, E., Packer, N. H., *Proteomics* 2001, *1*, 340–349.

[16] Wilkins, M. R., Lindskog, I., Gasteiger, E., Bairoch, A. *et al.*, *Electrophoresis* 1997, *18*, 403–408.

[17] O'Donovan, C., Martin, M. J., Gattiker, A., Gasteiger, E. *et al.*, *Brief. Bioinform.* 2002, *3*, 275–284.

[18] Westbrook, J., Feng, Z., Chen, L., Yang, H., Berman, H. M., *Nucleic Acids Res.* 2003, *31*, 489–491.

[19] Eisenhaber, B., Bork, P., Eisenhaber, F., *J. Mol. Biol.* 1999, *292*, 741–758.

[20] Maurer-Stroh, S., Eisenhaber, B., Eisenhaber, F., *J. Mol. Biol.* 2002, *317*, 541–557.

[21] Monigatti, F., Gasteiger, E., Bairoch, A., Jung, E., *Bioinformatics* 2002, *18*, 769–770.

[22] Nielsen, H., Engelbrecht, J., Brunak, S., von Heijne, G., *Int. J. Neural Sys.* 1997, *8*, 581–599.

[23] Emanuelsson, O., Nielsen, H., Brunak, S., von Heijne, G., *J. Mol. Biol.* 2000, *300*, 1005–1016.

[24] Gattiker, A., Michoud, K., Rivoire, C., Auchincloss, A. H. *et al.*, *Comput. Biol. Chem.* 2003, *1*, 49–58.

[25] Varki, A., *Glycobiology* 1993, *3*, 97–130.

[26] Jung, E., Veuthey, A.-L., Gasteiger, E., Bairoch, A., *Proteomics* 2001, *1*, 262–268.

[27] Hulo, N., Sigrist, C. J. A., Le Saux, V., Langendijk-Genevaux, P. *et al.*, *Nucleic Acids Res.* 2004, *32*, D134–D137.

[28] Zdobnov, E. M., Lopez, R., Apweiler, R., Etzold, T., *Bioinformatics* 2002, *18*, 1149–1150.

[29] Jensen, L. J., Gupta, R., Blom, N., Devos, D. *et al.*, *J. Mol. Biol.* 2002, *319*, 1257–1265.

[30] Nielsen, H., Brunak, S., von Heijne, G., *Protein Eng.* 1999, *12*, 3–9.

[31] Sunyaev, S. R., Eisenhaber, F., Rodchenkov, I. V., Eisenhaber, B. *et al.*, *Protein Eng.* 1999, *12*, 387–394.

[32] Hansen, J. E., Lund, O., Tolstrup, N., Gooley, A. A. *et al.*, *Glycoconj. J.* 1998, *15*, 115–130.

[33] Jung, E., Hoogland, C., Chiappe, D., Sanchez, J.-C., Hochstrasser, D. F., *Electrophoresis* 2000, *21*, 3483–3487.

