

Swiss-Prot: Juggling between evolution and stability

Amos Bairoch

heads the Swiss-Prot group at the SIB and is a professor at the Department of Structural Biology and Bioinformatics of the University of Geneva.

Brigitte Boeckmann

has been working in the Swiss-Prot group for 16 years. She has been involved in annotation and tool development and is now coordinating automatic annotation in Swiss-Prot.

Serenella Ferro

has a background in biochemistry and chemistry and has worked as a Swiss-Prot head annotator for 15 years.

Elisabeth Gasteiger

coordinates software development in the SIB Swiss-Prot group and is in charge of the ExPASy server.

Keywords: *protein sequence, database, functional annotation, automatic annotation, sequence analysis, user feedback*

Amos Bairoch, Brigitte Boeckmann, Serenella Ferro and Elisabeth Gasteiger

Date received (in revised form): 22nd December 2003

Abstract

We describe some of the aspects of Swiss-Prot that make it unique, explain what are the developments we believe to be necessary for the database to continue to play its role as a focal point of protein knowledge, and provide advice pertinent to the development of high-quality knowledge resources on one aspect or the other of the life sciences.

INTRODUCTION

The goal of this article is not to depict the history of Swiss-Prot,¹ as this has already been done elsewhere,² but rather to explore some of the consequences of decisions taken about 20 years ago, to discuss how the database has constantly evolved and to describe the challenges that it currently faces. To say that the past 20 years have been exciting would be a major understatement. Most young scientists now starting a career in the life science fields are not aware of how much the combined technological revolutions that led to high-throughput sequencing and the WWW have quantitatively and qualitatively changed the universe of knowledge on proteins. Yet, while we now have to cater in the Swiss-Prot and TrEMBL sections of the UniProt knowledgebase³ for more than 1 million protein sequences, there is a continuously widening chasm between truly characterised proteins and those that have been solely predicted by genome-sequencing projects. For us, in Swiss-Prot, the ultimate in terms of a well-characterised protein is one for which not only the exact sequence, post-translational modifications, subcellular location, tissue specificity, interaction partners and 3D structure are known, but more crucially for which a functional role can be assigned.

What we hope to convey in this paper

are the particular aspects of Swiss-Prot that make it unique, and hopefully derive some advice that would be pertinent to someone embarking on the development of a high-quality knowledge resource on one aspect or the other of the life sciences. But before we do so, we want to enumerate six observations that we believe are important to communicate to any would-be developers of such databases:

- Your task will be much more complex and far bigger than you ever thought it could be.
- If your database is successful and useful to the user community, then you will have to dedicate all your efforts to develop it for a much longer period of time than you would have thought possible.
- You will always wonder why life scientists abhor complying with nomenclature guidelines or standardisation efforts that would simplify your and their life.
- You will have to continually fight to obtain a minimal amount of funding.
- As with any service efforts, you will be told far more what you do wrong rather than what you do right.
- But when you will see how useful your efforts are to your users, all the above drawbacks will lose their importance!

Amos Bairoch,
Swiss Institute of Bioinformatics,
Centre Médical Universitaire,
1 Rue Michel Servet,
1211 Geneva 4,
Switzerland

Tel: +41 22 379 50 50
Fax: +41 22 379 58 58
E-mail: swiss-prot@expasy.org

A SMALL BIT OF HISTORICAL INTROSPECTION

How Swiss-Prot started and how it institutionally evolved

In 1965, the late Margaret Dayhoff published the first edition of the 'Atlas of Protein Sequence and Structure'.⁴ It contained information on 65 protein sequences. In the introduction she expressed the mission of the Atlas as

locating all of the relevant publications; critically reviewing the data and resolving conflicting reports; transforming the data into a uniform format to reflect those aspects of the structure that have been experimentally determined and those that could reasonably be inferred by homology; identifying the material with regard to chemical function, biological source, genetic control, and evolutionary origin. . .

This ambitious and still highly pertinent mission statement is a tribute to the vision shown by Margaret Dayhoff. She pursued her task until her untimely death in 1983. At that time the Atlas had evolved into a protein sequence data bank known as the Protein Identification Resource (PIR) of the National Biomedical Research Foundation (NBRF). When in 1985, one of us (Amos Bairoch) was, in the context of a PhD thesis, developing a software package (PC/Gene⁵) to analyse protein sequences, he was faced with some deficiencies and omissions in the PIR database. As he did not receive satisfactory feedback from PIR, he resolved to develop a version of PIR in the format of the European Molecular Biology Laboratory (EMBL) nucleotide sequence database that would contain additional sequences and, more crucially, additional annotations on various aspects of the protein universe.

In mid-1986, the first release of Swiss-Prot came out. Almost immediately we approached the EMBL to see if they were interested in distributing and helping with the maintenance of the

database. With foresight they immediately accepted. The collaboration that grew from this early decision gave rise to the current situation: Swiss-Prot is a fully collaborative endeavour of what has become the Swiss-Prot group at the Swiss Institute of Bioinformatics (SIB) and the European Bioinformatics Institute (EBI), an outstation of EMBL. The last institutional development was the decision, in late 2003, of the NIH to award a major grant to a consortium composed of the EBI, the SIB and PIR to produce a universal resource on proteins, known as UniProt.

Today, in 2004, more than 120 people directly work on Swiss-Prot and TrEMBL (see below) or on resources that evolved out of Swiss-Prot. While the first reaction to this figure can be 'that's a lot of people', it pales when compared with the amount of work to be carried out. In fact this is a major issue shared by all life sciences information resources: long-term, high-quality curation of information is not cheap. It is not as glamorous as whole genome sequencing projects or any such well-defined scientific and technological efforts, yet it needs to be adequately and stably funded. Sadly, this is not yet widely recognised by funding bodies.

Why TrEMBL was developed

In the mid-1990s it was already clear that the increased data flow from genome projects was going to be a major challenge for Swiss-Prot. As will be explained further on, maintaining the high quality of the database requires careful sequence analysis and detailed annotation of every entry. This was, and still is, a major rate-limiting step. We did not wish to relax the editorial standards of Swiss-Prot and there was a limit to how much the annotation procedures could be accelerated. Yet it was vital to make new sequences available as quickly as possible. To address this concern, we introduced in 1996 TrEMBL (Translation of EMBL). TrEMBL consists of computer-annotated entries derived from the translation of all

Swiss-Prot contains mostly manual annotated entries

TrEMBL consists of computer-annotated entries, which are not yet in Swiss-Prot

coding sequences in the EMBL database, except for those already included in Swiss-Prot. TrEMBL is therefore a complement to Swiss-Prot and sequence entries only move out from TrEMBL and enter Swiss-Prot after having been manually curated by an annotator.

From 1996 to the end of 2003, Swiss-Prot grew by 83,000 sequences to reach a total of 140,000 entries. In this period of time, TrEMBL grew from the 86,000 entries in its first release to about 1.1 million entries!

WHAT MAKES SWISS-PROT SPECIAL

Aiming for the perfect sequence

Even if it may be obvious to many of its users, it is important to restate that Swiss-Prot is a corpus of knowledge centred on protein sequences. As will become apparent in the following sections of this paper, we add many layers of information around the sequence data, yet most of that information is in one way or another dependent on the sequence. It is therefore important to capture and to represent the most correct sequence. This is an important aspect of the work of Swiss-Prot that escapes the notice of most of its users.

The overwhelming majority (>99 per cent) of the sequence data represented in Swiss-Prot originates from the translation of nucleotide sequences submitted to the EMBL/Genbank/DDBJ database. Only a very small proportion of the sequences are obtained directly at the amino acid level using Edman degradation or mass spectrometry. This situation already existed in 1986. What has happened since was obviously an enormous quantitative increase in the amount of nucleotide sequence data, but also, more relevant to our quest toward quality, a significant increase in nucleotide sequence quality and a sociological change in the breakdown of the originators of sequence data. The increase in sequence quality is mainly due to the growing use of very sophisticated automated sequencing

machines. In 1986, most nucleotide sequences submitted to the DNA databases originated from individual laboratories that were sequencing a single gene or a small region of a genome. Today, the biggest (in terms of quantity) contributors are major sequencing centres that either provide complete genomic sequences or massive amounts of data from full-length cDNAs.

As we depend on primary sequence data that have been submitted to the nucleotide sequence databases, it would seem at first glance that there is not really anything we can do to improve the quality of the derived protein sequences. This is far from being true, and in fact there are many things we can do by comparing sequences. Sequence comparison is essential to the process of creating or updating a Swiss-Prot entry. One needs to remember that Swiss-Prot is a non-redundant database. What this means is that we took the decision from the very beginning to merge the protein sequences from the same organism originating from the same gene. Thus we are often faced with many complete or partial sequences that need to be merged and whose discrepancies have to be taken into account. Sequence discrepancies are annotated with the feature (FT) keys CONFLICT, VARIANT, MUTAGEN or VARSPLIC. The FT key VARIANT is used to describe polymorphisms and disease mutations, MUTAGEN for experimentally altered sites and CONFLICT for sequence differences of any other reason. Insertions or gaps within alignments of otherwise identical sequences are usually due to alternative splicing events, which are annotated using the FT key VARSPLIC.

Thus sequence comparisons can already help us in determining what is the most correct sequence. This is especially true in organisms that are the focus of many sequencing efforts. For example, we currently have an average of 3.7 independent sequence reports (cDNA or genomic DNA) for each human protein. Such a redundancy in the nucleotide

The correct protein sequence is the basis for high-quality annotation

Redundancy removal: Merging entries point out sequence discrepancies

Splice isoforms

sequence database helps flagging potential sequencing errors. Further errors can be found when comparing orthologous and paralogous sequences across species. The relevance of such approaches is increasing as more and more full genome sequences are becoming available.

Frameshifts

One of the advantages of comparing many sequences is the detection of probable frameshift errors. They stand up in multiple protein sequence alignments as locally divergent regions. If the divergence can be explained at the nucleotide level by the insertion or deletion of a single nucleotide, it is likely (but not certain) that it is due to a sequencing error. The total number of potential frameshift errors that were corrected by Swiss-Prot annotators is difficult to estimate as it often happens that incorrect DNA sequences are later resubmitted by the original authors, correcting sequencing errors, generally by taking into account the correction made in the corresponding Swiss-Prot entries. In the current release we have 1 per cent of the entries that are flagged with at least one potential frameshift error in one of the cross-referenced nucleotide sequence entries.

Initiation sites and exon boundaries

In many cases, the N-terminal initiation sites of bacterial or archaeal genes or the exon/intron boundaries of eukaryotic genes are incorrectly predicted. It is important to note that these predictions are of a very heterogeneous quality and to recognise that not all sequencing centres produce the same level of quality in terms of both sequences and of protein-coding gene predictions. Swiss-Prot annotators are aware of this heterogeneity and know what data can be more or less trusted. We currently observe that in 7.1 per cent of our entries we disagree with the translation provided by the submitter.

Access to published information before the internet era

Annotation of CDs not annotated in the nucleotide sequence databases

It often happens that annotators have to translate, from a nucleotide entry, protein sequences that have been overlooked by the original submitters. Currently we have 2.5 per cent of our entries that contain such translations.

Finally, the work of the Swiss-Prot annotators is also to reject putative protein sequences that are obviously bogus, either because they originate from a pseudogene or because they were incorrectly predicted either from non-coding DNA or a wrong open reading frame (ORF).

If you take all the above factors and tasks into consideration, you can see why we believe that the correction of amino acid sequences is an important part of the annotation process, and that it is far from trivial to achieve. This is not necessarily apparent to the user, but it is one of the reasons why Swiss-Prot has always been considered as the reference database for protein sequences. Of course the drawback of such an approach is that it is time consuming and can be applied only to manually annotated entries. Such an approach can consequently not be applied to TrEMBL, where the represented protein sequences are those that have been indicated by the submitters of the original nucleotide sequence entry. It would therefore be important to develop semi-automatic systems that allow some aspects of sequence correction to be applied to TrEMBL.

Extracting information from the literature

Fifteen years ago, Swiss-Prot annotators typically went through the following process: they photocopied all relevant papers from the reference list of the entry they were annotating. The publications were read and important information was marked in the paper copy. Information was then added to the entry in either free text (comments lines) or structured feature lines. Access to reference databases and computing tools considerably facilitated the above procedures, but also brought along a higher level of complexity. Being an annotator in the early 1990s was already not a trivial job, but it has since become a much more demanding task.

When Medline became available at the workplace first on CD-ROMs, and later via the internet, most journal abstracts

The primary source of protein knowledge are journal articles

Text mining tools will guide annotators through the wealth of publications

could immediately be read – or discarded if not relevant – and information was retrieved directly from here, which was particularly helpful when the journal was not available from local libraries. But it is online access to full text articles that has completely changed the life of annotators. They can look at many more relevant papers than they used to do when they needed to go to the library. This is particularly useful nowadays as information on a given protein is generally spread between many different reports in a wide variety of journals. Such a trend is exemplified by the journal citation statistics of Swiss-Prot: in 1993, 461 different journals were cited in the database, while today the number has risen to about 1,400. Although some journals (such as *J. Biol. Chem.* and *Proc. Natl Acad. Sci USA*) were and still are major sources of articles useful for the annotation process, there has been a clear trend toward a ‘decentralisation’ of the sources of protein-related publications. Of course, journal articles are not the only source of information, and we also make use of electronic journals, book articles, theses, patent applications and external information resources, but overwhelmingly the primary source of experimental information remains published journal articles.

We are often asked whether annotators are ‘really sitting there and reading publications’. Yes, they are. Knowledge extracted from the articles is mostly added to the appropriate topics of the comment (CC) lines, and to the feature table (FT), whenever a description concerns a defined region or site within the sequence. But we also add new synonyms for protein names (DE line), gene names (GN line), compare or complete author names with the ones given in a reference block (RA line), annotate a reference block (RP and RC lines), add additional relevant references to an entry, and much more. All experimental findings and authors’ conclusions are compared with the knowledge available on related proteins and the results from various

protein sequence analysis tools. When contradictory results have been published and there is not enough information to prefer one hypothesis over the others, the annotation is performed in a way that draws the user’s attention to the contradictory conclusions. Finally the content of an entry is summarised in form of a list of keywords (KW line) from a controlled vocabulary.

Both abstracts and full text articles are the target of text-mining tools, which will soon become an indispensable help for annotators to quickly find the publications of interest from the wealth of information available. We believe that efforts to build efficient software tools allowing the semi-automated extraction of information from repositories of full text articles will be essential to anyone trying to build comprehensive information resources for life scientists. The fact that we will rely on such tools to hunt and extract information is paradoxical. Anyone outside the life sciences field would believe that such important information would be immediately made available in a structured way by the experimentalists to the relevant databases. As we will see in the next section, this is unfortunately not the case.

User submissions and updates

We have always strongly encouraged user feedback, as well as the submission of updates and corrections, initially by asking people to contact us by e-mail. Also, very early on, a list of ‘on-line experts’ was compiled, ie a list of email addresses of scientists working with specific protein families or domains, who agreed to review protein sequences in Swiss-Prot relevant to their field of research. This list is regularly updated and the ~150 experts’ e-mail addresses, grouped by fields of expertise, are listed in the document.⁶

However, it does not seem clear to most users – who have grown accustomed to the repository nature of the nucleotide sequence databases, where only the original authors are allowed to correct and update existing entries – that

User-submitted updates are highly valued

Swiss-Prot is extremely different in that respect, and that we do have an ongoing editorial policy. We do indeed highly value our users' expertise, and we believe that it is only with the assistance of our user community that we can do our job of being comprehensive and up to date. We are therefore actively seeking any type of updates and/or corrections, whether they have been published or not, and would like to be notified about annotations to be updated, eg if the function of a protein has been clarified, or if new post-translational modification information has become available. In order to increase the visibility of these aspects, and to encourage our users to let us know about outdated protein entries or errors, we have implemented update forms on the ExPASy server (see the section below, 'Making Swiss-Prot available to the users'). The forms, accessible from the bottom of every Swiss-Prot entry, prompt users to provide their corrections and updates in any format. Update requests are treated with a very high priority by annotators. We currently receive about 300 update requests for Swiss-Prot entries per year, a number that we would very much like to see growing in the future!

Web submission forms for updates

On the other hand, annotators send newly annotated entries to the original authors of reports cited in these entries so as to check the validity of the annotations. We generally get useful feedback, but not as much as we would like!

Direct protein sequence submission

Another point of interaction with users is sequence submission directly to Swiss-Prot and TrEMBL. We accept submission of sequences that have been obtained only as amino acid sequence. A web submission tool (SPIN) has just been made available, which guides the submitter through the process, and prompts for all required pieces of information. There are about 300 such sequence submissions per year. It is interesting to note that 10 per cent of the proteins originate from venomous animals. This is explained by the fact that toxins can easily be purified in large

quantity from venom and are generally quite small, thus they are easily sequenced at protein level.

We have to admit that we are disappointed by the low level of input from users in the updating of the database. We may have been insufficiently efficient in publicising our willingness and eagerness to welcome any type of help. Yet, after years of discussions with researchers, we believe that the root of the project is of a sociological nature. The career of life scientists is driven by the famous 'publish or perish' injunction and submitting data to a database does not get any credit points on a CV. So we have to rely on the altruism of some individuals. We are indeed indebted to those persons who take the time to make sure that we adequately represent the results of their research in our database. However, we believe it is time that the community as a whole addresses this issue and initiates a process of responsibility toward the biomolecular databases.

Tools for annotation***The basic data organisation, the editor and the syntax checker***

The working copy of Swiss-Prot is arranged in flat files, grouping proteins by family or other functional criteria. Although it was apparent from the beginning that the complexity of protein relationships could not be simulated simply by grouping entries one-dimensionally into separate files, this system allows curators to immediately find orthologues, which can all be updated when new findings become available for at least one protein, or when a review article summarises relevant knowledge on a protein family or subfamily and comes to new conclusions. The quick availability of all related entries (all in the same file) also ensures consistent annotation of all relevant entries. The ~140,000 entries in the current release are thus split into ~3,000 files.

Most of the annotation is done manually with the help of a continuously growing number of tools. We currently

From a single text editor to an adapted annotation platform

use a text editor, Crisp (from Vital, Inc.), that is easy to use and comes with a powerful C-like macro language that we use extensively both for literature-driven textual annotation and as a platform to launch sequence analysis programs (see next section). An extensive series of macro-commands have been developed to reformat references, comment lines, feature lines or sequences, to check controlled vocabulary or syntax, and to retrieve entries from other databases. Analysis tools are also run directly from the editor with the help of macro-commands that send the sequence and other relevant information to the analysis program, and then retrieve the result and format it in the annotation platform. All commands are available both from keyboard shortcuts (which are preferred by experienced annotators) and from menus and dialogue boxes that are fully integrated in the editor's graphical user interface (GUI) environment.

Only well-structured data is easily accessible

Swiss-Prot annotation has always been subjected to very strict rules and guidelines. All entries are reviewed before they enter the database, which guarantees the homogeneity of the annotation. We developed a 'syntax checker' so as to make sure that our annotation and format rules are enforced. This syntax checker, implemented in Perl, is much more than a program that verifies the basic syntax of a Swiss-Prot entry. It also enforces the use of controlled vocabularies (see section below, 'Standardisation and controlled vocabularies') and checks for dependencies and consistencies between different portions of an entry. In December 2003, the syntax checker contained almost 1,100 different rules, each of which can lead to the detection of errors or inconsistencies.

The number of prediction methods used in Swiss-Prot steadily increases

Many people are surprised to hear that Swiss-Prot annotation is done from within a text editor. However, those same people are usually even more surprised once they see how powerful the annotation platform developed around that text editor is, and that almost every command can be launched, and its results

treated, from within the editor, in a remarkable speed. One major disadvantage of this environment is that it relies heavily on the flat file format. We are now developing a Swiss-Prot specific editor, which will work with the version of the databases formatted by extensible mark-up language (XML), and will include many consistency checks and context-specific menus. The new annotation platform will also include many graphical features, eg visualisation of domain and site predictions along the sequence. We believe that such a development is highly desirable, as it will allow the implementation of consistency checks directly at the level of the annotation platform while we now have to rely on a regular post-processing check of the data, using the syntax checker to enforce consistency.

Sequence analysis tools

The task of annotating Swiss-Prot entries has always relied on the use of the most appropriate sequence analysis programs so as to predict important sequence features. Over the years we have implemented many different methods and programs in our annotation platform. We have also spent a considerable amount of time testing new methods and selecting the most appropriate ones. In some cases, when no existing program could satisfy our needs, we have developed our own set of predictive methods.^{7,8} All these activities are carried out by a small research component within the Swiss-Prot group whose missions are to carry out technological watch and to develop new methodologies for protein sequence analysis.

Currently we use software tools (a full list with references is available in the Swiss-Prot document `annbioch.txt`) to predict the following sequence features:

- signal sequences of type 1, type 2 (lipoprotein) and type 3;
- mitochondrial and plastid targeting sequences;
- transmembrane domains;
- coiled coil domains;

- specific repeats (leucine-rich repeat (LRR), tetratricopeptide repeat (TRR), WD (Trp-Asp) repeat, etc);
- statistically significant runs of amino acids and regions enriched in particular amino acids;
- N-glycosylation sites;
- glycosylphosphatidylinositol (GPI) anchors;
- sulphation sites;
- N-terminal myristoylation sites.

PROSITE was created for the annotation of conserved domains and functional sites

In addition to the above list, we make extensive use of domain/family databases to annotate specific domains. In fact the development of the PROSITE⁹ database, which was first released in 1990, was specifically driven by the need to detect and annotate protein domains. The combined usage of profiles and patterns allows the detection of domains (profile) and the functional sites within domains (pattern). As mentioned in the section on cross-references below, there are now many other protein domain databases and we occasionally make use of most of them to annotate specific domains not yet covered by PROSITE. The reasons of our preference for PROSITE over other similar databases are pragmatic: PROSITE domain descriptors are specifically tailored for their use in the context of protein sequence annotation in order not to predict overlapping domains. Cut-off values are selected conservatively to minimise the number of false positives: we prefer to miss the occurrence of a domain rather than to over-predict its existence.

HAMAP proved that automated annotation is not necessarily accompanied by a decrease in quality

We believe that the use of the most up-to-date sequence analysis tools is essential to any protein sequence annotation effort. In addition anyone considering applying such methods on a large scale needs to develop internal benchmarks so as to objectively judge the validity and the scope of the methods. In many instances we have observed that the claims of developers of sequence analysis methods are slightly overblown and that one obtains unexpected results when using such methods on large and highly heterogeneous sets of sequences.

Automation: Trying to simulate the expertise of annotators

Thanks to genome sequencing efforts, there has been a tremendous rise in the number of available protein sequences. Yet clearly this is only the beginning and what exists now will represent only a drop in an ocean of uncharacterised sequences. And there lies both the problem and a possible solution: on one hand the overwhelming majority of genome-derived sequences are currently not the target of experimental characterisation and are probably not going to be so in the next decade. On the other hand, we have encapsulated in Swiss-Prot a tremendous amount of knowledge, some of which is specific to a given protein, while the majority can be carefully propagated to well-defined orthologous sequences. Automatic annotation is far from being a novel concept. But what we want to achieve in Swiss-Prot differs from what others expect from such systems. Their aim is to analyse new genomic sequences and predict a maximum of potential information items so as to be able to infer hypotheses on the potential biological processes present in the organism. Our aim is to make sure that we produce high-quality annotation with a minimal amount of incorrect inferences.

Our first automatic annotation project is called HAMAP,¹⁰ which stands for High-quality Automated and Manual Annotation of microbial Proteomes. In the context of this project, proteins from complete bacterial and archaeal proteomes, together with the related plastid proteins, are automatically annotated based on manually created family rules for complete protein annotation, with template-based feature propagation. Proteins with no similarity to other proteins in Swiss-Prot, which we call ORFans, undergo an automated protein sequence analysis procedure that looks for many of the sequence features described in the preceding section. These features are then automatically annotated according to rules of consistency and dependency. A paper with further

Anabelle extends the scope of automated annotation to proteins with a complex domain architecture

statistics on HAMAP is currently under review by another journal.

We have just developed a second system called Anabelle that strives to annotate not only ORFans and well-defined proteins, but also any protein with one or more conserved or functional domains or sites detected by one of the methods carefully selected for their accuracy by the Swiss-Prot team. The information retrieved from all results is logically combined according to selection rules and logical rules, thus coming to more trustworthy conclusions than possible when just looking at one result at a time. Anabelle is integrated in the annotator's workbench: the automatically pre-selected analysis results are visualised in a graphical system, from which the annotator can choose the true positive results and easily generate annotation based on sequence similarity and sequence analysis. Not only does this speed up annotation, but it also promotes the consistent transfer of entire information blocks that logically group together, ensuring the usage of standardised vocabulary and minimising the probability of errors and typos.

Controlled vocabularies demand continuous attention

We believe that careful application of rules to produce automatically or semi-automatically annotated protein entries brings about many advantages for users of Swiss-Prot. We know that many are apprehensive of the word 'automation' and are afraid that we will drown high-quality manually annotated entries with lower-quality 'automated' entries. We are very aware of this danger and are almost paranoid in our effort to ensure that automatic annotation will produce data of a quality up to that of manual curation. Finally it must be noted that one of the important changes planned in the Swiss-Prot format (see section on 'Evolution of entry structure and format' below) is very pertinent to this issue: the introduction of 'evidence tags', should allow us to unambiguously flag whether an information item has been derived manually or automatically.

Standardisation and controlled vocabularies

A long tradition of using controlled vocabularies in Swiss-Prot

To allow effective and precise database retrieval and searches, the same concepts need to be described with the same terms everywhere in the database. Controlled vocabularies or indexing terms can serve this purpose. A controlled vocabulary is defined as 'an organized list of words and phrases, or notation system, that is used to initially tag content, and then to find it through navigation or search'.¹¹

Since its creation, Swiss-Prot has stored information under specific line types, many of which are structured in such a way as to facilitate text searches in the database. Even the fields that appear to contain unstructured text are often written according to strict guidelines to ensure consistency. In some cases, lists are made where 'preferred' terms are associated with synonyms, spelling differences, abbreviations, or yet other terms considered as equivalents.

Table 1 provides a partial description of where and how Swiss-Prot either makes use of existing controlled vocabularies or has developed such corpora. This list, even if incomplete, is impressive; yet it does not capture the whole complexity of issues surrounding the use of nomenclature and controlled vocabularies in the life sciences. We need to state here that if physicists or chemists behaved like biologists, we would probably live in a world without computers or plastic (this may sound like an attractive proposition to some!). Life scientists do not receive, during their training, the perception of the importance of following nomenclature rules. Yet, they are the first to complain when they look for specific information across one or many databases and fail to obtain a comprehensive answer because that information is heterogeneously described. Therefore we always felt that Swiss-Prot had a mission to fulfil in enforcing existing rules and more and more, as time passed by, to actively participate in the development of

Table 1: Standardisation efforts and use of existing or in-house controlled vocabularies in Swiss-Prot, listed by line type. (Note: Refer to the Swiss-Prot user manual for further information on all the information present in a Swiss-Prot record)

Protein names (DE line)	We use as primary name the one that seems to be the most appropriate according to the function of a protein, to the nomenclature adopted by the specialists in that field or to the gene name, etc. We keep all synonyms used in publications and authors' submissions except if they are misleading. Furthermore we transfer the same name to the orthologues of related organisms.
Gene names (GN line)	Whenever a nomenclature committee (HUGO, FlyBase, etc.) provides 'official' gene names for a given organism, we try to enforce their choice of gene names, yet keeping what authors originally provided as synonyms.
Species names (OS line)	The species names used in Swiss-Prot are listed in a document (speclist.txt). From the very beginning, care has been taken to store not only the official (scientific) name, but also the most useful common names and synonyms.
Species taxonomy (OC and OX lines)	We make use of the taxonomy compiled by NCBI which is used by most major biomolecular sequence databases.
Organelle (OG line)	We standardise plasmid name usage and list them in a Swiss-Prot document (plasmid.txt).
Reference comments (RC line)	Among other uses, the RC line allows us to indicate the tissue from which a protein originates (TISSUE), or the strain (STRAIN). The tissues are reported in the file tisslist.txt and the strains in strains.txt. Both lists contain indications on synonyms.
Reference authors (RA line)	As far as possible, the names of authors are stored according to consistent rules. For example the German umlaut is replaced by an 'e' following the vowel on which the umlaut was perched, the hyphen is retained between two initials (which is removed in Medline/PubMed), we keep all the initials (even where PubMed only keeps two) and we often correct misspelling in author names!
Reference location (RL line)	Journal abbreviations in Swiss-Prot follow whenever possible those used by the National Library of Medicine (NLM). We provide a journal list (journalist.txt) that, in addition to the journal names and abbreviations, also provides ISSN (International Standard Serial Number), CODEN number, publishers and journal home page web addresses.
Comments (CC line)	The CC lines mainly contain free text comments classified under 24 different topics. If a piece of information cannot be classified under a specific topic, it is put under 'MISCELLANEOUS'. However, with time, the information in the CC lines is becoming less 'free' so to speak, and more and more CC line topics are subjected to controlled vocabularies. For example, this is the case of the 'CATALYTIC ACTIVITY' topic whose text is taken from the ENZYME database ¹² for all known enzymes, referred to by their EC (Enzyme Classification) numbers in the DE lines. We are currently standardizing the use of the 'COFACTOR', 'PATHWAY' and 'SUBCELLULAR LOCATION' topics.
Keywords (KW line)	Keywords were one of the first sets of controlled vocabulary in Swiss-Prot. They were introduced to summarise the content of an entry and to group entries according to different aspects related to biological processes, molecular function, subcellular location, domains, ligands, sequence modifications and diseases. We provide a keyword list (keywlist.txt) that is being superseded by a dictionary that provides the precise definition of the usage of a keyword in the context of Swiss-Prot. The dictionary also includes synonyms, groups keywords into categories and provides a mapping between Swiss-Prot keywords and GO terms (see section 'Going ahead with GO in Swiss-Prot').
Feature table (FT line)	We are currently establishing a controlled vocabulary for the features describing post-translational modifications (PTMs). ¹³ We are also building a PTM database to store, for each type of modification, information such as the general description, target(s), chemical formula, subcellular localisation of modified site, enzyme(s) carrying out the PTM, etc. Domain-type (DOMAIN, REPEAT, DNA_BIND, ZN_FING, etc.) feature descriptions are also standardised across all of Swiss-Prot.
Sequence	The sequences are stored in the one-letter code adopted by the commission on Biochemical Nomenclature of the IUPAC-IUBMB.

new nomenclature and controlled vocabularies. Anecdotally such an active role can have some unexpected consequence: we were once threatened with a lawsuit because we did not accept to use as a valid gene symbol the one proposed by an author.

All of this leads us to give the following advice to would-be developers of databases:

- Try to follow as much as possible existing controlled vocabularies and nomenclatures.
- Do not hesitate to contact the groups maintaining these resources and to

point out inconsistencies and/or errors.

- Do not be afraid to take a firm stand toward your users when they request the representation in your database of terms that do not follow a specific guideline. You can always (and you should!) store this information as a synonym.

Going ahead with GO in Swiss-Prot

If we assume, as mentioned above, that 'users and database should agree on the meaning of the term being used', given the large number of biomolecular databases available, this indirectly implies

Swiss-Prot introduces Gene Ontology terms very carefully

that **all** databases should agree on the meaning of a term! In an attempt to achieve this ambitious goal, maintainers of FlyBase, MGD and SGD joined forces and formed the GeneOntology (GO) Consortium.¹⁴ They established three ontologies, gathering key terms for cellular components, biological process and molecular function, thus catering for a large need for standardisation that could be observed all across the scientific community.

From the beginning of the GO activities, we were repeatedly approached by users wondering when we would introduce GO terms to Swiss-Prot and TrEMBL. However, while clearly welcoming the effort made by the GO Consortium, we were reluctant to add links to GO at that time: given the initially small scope (GO specialised in three major organism groups, whereas Swiss-Prot has to deal with thousand of different species), and the fact that many mappings had been created automatically and were thus likely to assign GO terms to unrelated proteins, we considered it dangerous to mislead users into incorrect assumptions. We did not want to risk the situation where someone would happily accept a GO assignment indicating a function for an otherwise uncharacterised protein, without further questioning the assignment because they trust the judgment of Swiss-Prot annotators and the high quality of the manual annotations.

It was only in 2003 that we felt it had become 'safe' to start introducing GO terms in Swiss-Prot. We felt that GO had indeed considerably matured and had increased its coverage. What is more, several species-specific databases have established manually curated mappings between GO terms and their gene catalogues. The EBI GO team has mapped Swiss-Prot keywords to GO terms. Evidence tags are available in GO to indicate whether an assignment has been done automatically or by manual curation. The time had come to follow the demands, and to introduce cross-

references (see section on 'Cross-references in Swiss-Prot' below) from Swiss-Prot to GO. We added them in all cases where they originated from manual annotation efforts. We also are in the process of introducing GO terms for all members of microbial protein families that fall under the scope of the HAMAP annotation project.

Evolution of entry structure and format

Since its creation in 1986, the basic structure of a Swiss-Prot entry has not changed significantly. The distinct line types defined by a two-letter code are generally relevant to all entries and cover the core data, while the actual protein information is given in the comment (CC) lines and in the feature table (FT). While the general framework has been very stable, we have carried out many changes over the years. New line types were introduced, the structure of existing line types was constantly refined and new sub-fields (comments topics, feature keys) were added. Such changes are always documented (in release notes and other documents) and users are warned in advance of pending changes so that they can adapt their software tools. While the general stability of the Swiss-Prot flat file format may be seen as a proof of foresight, careful planning and experience, one can also say that in some respect Swiss-Prot had become a victim of its own success: even the smallest modification to the flat file format, or the introduction of new fields, needs to be considered carefully, and it happens that ideas are discarded for the sole reason that 'this will cause the crash of thousands of programs out there. . .'.

Swiss-Prot and TrEMBL have traditionally been maintained and distributed as flat files. An inherent problem of flat file databanks is that their maintenance becomes increasingly difficult when they grow in size and many people are involved in the production of the data. Since 2002, Swiss-Prot and TrEMBL have also been distributed in

XML format

XML,¹⁵ the extensible mark-up language that makes it possible to define the content of a document separately from its formatting, making it easy to reuse that content in other applications or for other presentation environments. XML allows, in contrast to HTML, the authors of a document to create their own mark-up tags suiting their needs and allowing the best structure for the data. But what is more, XML allows implementing rules that are not limited to formatting, but can be used to formulate dependencies. We are also in the process of porting the production of Swiss-Prot and TrEMBL to a relational database management system. In order to develop the relational and XML schema, we have designed conceptual data models, using the Unified Modelling Language (UML) notation, to represent the structure and constraints present in the data.

Relational database

In the meantime, until the production copy of Swiss-Prot is managed in a relational database management system, we still need to introduce certain format changes to the flat file in order to accommodate more complex concepts. Such changes can be quite substantial and time-consuming, as they are always introduced in a way that not only new annotation is performed according to the new format, but all existing entries need to be converted. As a consequence, this can involve, in addition to the creation of conversion software, and to the modification of documentation and annotation tools, a lot of manual cleaning. That we need to embark on such manual cleaning steps is not due to the structure or the format of the database, but rather to our pathological urge to make sure that all aspects of Swiss-Prot are self-consistent. Therefore, whenever we introduce a new type of data, we try as much as possible to update all the entries where such data have some relevance.

Evidence tags

There are many changes we plan to make to the flat file format. For example, in the near future, we plan to overhaul the format of the GN (gene) line so that it will allow a more structured

representation of the information concerning gene names. The new format will allow distinguishing official gene name, synonyms, ordered locus name and ORF names. This change allows a better representation of the complexity of gene and locus naming schemes.

As we described in the section on automatic annotation, it is important to provide users with a means to track down the origin of all information items in a Swiss-Prot entry. Such a need was not apparent in the early days of Swiss-Prot as most information was derived from a single paper that both reported the sequence and its characterisation. This is no longer true and some entries contain information originating from up to 110 references as well as the results of many sequence analysis tools. It is therefore necessary to provide 'evidence tags'. These are links between an information item and its source, whether a reference, the judgment of annotator or the result of a program. Such evidence tags already exist in TrEMBL. We have been very slow in the process of providing them in Swiss-Prot, partly because they are difficult to implement in the current annotation platform and because they are very cumbersome in the current flat file format. Evidence tags are therefore going to be implemented in the XML and relational versions of Swiss-Prot and will probably not be available in the flat file distribution.

Cross-references***Cross-references in Swiss-Prot***

Cross-references as a way to access related information in other databases have been an integral part of Swiss-Prot almost since the beginning (they were introduced in release 4 of April 1987). Navigating between databases is much less of a challenge now, thanks to the web, than it was back in the late 1980s. The early presence of DR (Database cross-Reference) lines in Swiss-Prot shows how anticipatory we were in conceiving the database in a way that facilitates data integration. One of the first important

The Sequence Retrieval System (SRS)

software applications that made use of Swiss-Prot cross-references was the Sequence Retrieval System (SRS),¹⁶ developed by Thure Etzold at EMBL, from 1990 on. In addition to providing a search interface for multiple databases with a single query, an important feature of SRS is its ability to combine all indexed databanks into a network, where new ways of linking information from different sources can be explored. One of the main reasons why this became possible was the fact that Swiss-Prot, one of the first databases indexed under SRS, was so highly cross-referenced. SRS documentation contained in 1990, and still contains in 2003, an image showing biological databases linked to each other in form of a network, the centre of which is Swiss-Prot, connected with practically all the other databases indexed under SRS.

Link statistics

The first databases cross-referenced in Swiss-Prot were the primary DNA and protein sequence databases EMBL and PIR, and the PDB protein structure database. New links were regularly added at each of the major Swiss-Prot releases. Currently Swiss-Prot is linked to 55 different databases and each entry contains an average of 9.1 links. One would naively assume that an entry does not contain more than a single cross-reference to a given external database. This is not always true, for a variety of reasons that generally depend on the structure of the external database. For example, there is an average of 1.92 cross-references to the EMBL DNA sequence database per Swiss-Prot entry. This reflects the redundant archival nature of the nucleotide databases. However, this overall average does not convey the true nature of the situation: 58 per cent of all Swiss-Prot entries contain only one cross-reference to EMBL, while 6.2 per cent contain more than five such cross-references.

A special emphasis should be given to the cross-references to family/domain databases. PROSITE was the first of these databases to be created and accordingly the first to be cross-referenced in Swiss-

Prot. When cross-references to PROSITE were introduced in 1990, there was an average of 0.42 per Swiss-Prot entry. In 2003, this number was more than twice as high, an increase that can be explained by improved methods to detect domains, but also by the fact that PROSITE increasingly reacts to the demands from Swiss-Prot annotators: Whenever a newly annotated protein family carries a particular domain that is not yet present in PROSITE, the PROSITE staff creates a discriminator (pattern or profile) for that domain. Many other family/domain databases were created in the past ten years, most of which are cross-referenced in Swiss-Prot and also incorporated in the InterPro¹⁷ resource which unites these databases 'under one roof'. Today a Swiss-Prot entry contains an average of 5.2 links to family/domain databases. These cross-references can also be seen as a pointer to the existence of a specific domain in a given protein sequence.

As mentioned above, in 2003, we have added cross-references to the three GO ontologies. These cross-references have a dual purpose: they allow navigation toward an external resource (here GO), and they also serve as information items. This may be better explained by the following example:

```
DR GO; GO:0012501; P:programmed
cell death; TAS.
```

In the above line, the GO accession number 'GO:0012501' provides a handle to access the GO database (navigation), the 'P:programmed cell death' indicates that the protein is involved in the biological process ('P') of programmed cell death and the 'TAS' stands for 'Traceable Author Statement'.

Cross-referencing versus integrating

Over the years, it became clear that our strategy to 'delegate' specialist tasks to the specialists (and establish reciprocal links), while concentrating on the more 'generalist' annotation was satisfactory.

Explicit and implicit links

This was facilitated and influenced by the appearance of more and more databases: the WWW made it a lot easier to publish expert knowledge. Existing and well-established databases (eg FlyBase) took advantage of the increased visibility offered by the web, and many additional new information resources burgeoned. A number of these databases were constructed around the primary sequence or organism-specific gene nomenclature databases, and used the accession numbers of the sequence databases (or the primary gene names) as their set of unique identifiers. An example is GeneCards, a database of 'information cards' on every human protein in Swiss-Prot and TrEMBL. Such databases are usually cross-referenced to Swiss-Prot via 'implicit' links, created on the fly by the NiceProt tool (see section below, 'Making Swiss-Prot available to the users') that displays a Swiss-Prot entry on ExPASy. In addition to the explicit cross-references 'hard-coded' in the Swiss-Prot DR lines, the concept of implicit links enforces the role of Swiss-Prot as a central hub for molecular biology information.¹⁸

There may seem to be certain drawbacks related to the strategy of establishing extensive cross-links *v.* the idea of integration of all data locally:

- 'loss of control';
- cross-references create a certain dependency (when free public access to the Yeast Proteome Database (YPD) was discontinued, expectations grew again for Swiss-Prot to provide more extensive annotation for *Saccharomyces cerevisiae*);
- necessity to rely on the willingness to collaborate of providers of the specialised cross-referenced databases (eg use of standard nomenclature and common identifiers, provide or at least help with mappings between Swiss-Prot accession numbers and their database);
- some foresight and knowledge of the related field is necessary, in order not to make the effort of adding links to a

resource that will not be updated or that is likely to lose funding – with the consequence of being forced to remove those links after a short while.

However, these disadvantages are easily outweighed by a gain in time and the relief not to 'have to be an expert in every field', as well as the reward of fruitful collaborations and exchanges. Procedures have been established to obtain mappings between Swiss-Prot sequences on one side, and relatively heterogeneous information on the other: nucleotide sequences, gene names, modification sites, domain descriptors, ontologies, etc. Many cross-references, in particular those that are based on sequence searches, ie domain and family classification, are now already applied to TrEMBL. This means that an entry comes with a certain number of DR lines before manual annotation even starts. Some other DR lines, however, require careful checking by an annotator, and yet others have to be added completely 'manually' as they can only be established after perusal of literature and other sources (eg MIM). While the list of cross-referenced databases keeps growing, it does happen that we are obliged to remove links to certain databases. This can have several different reasons, the most frequent ones being a lack of funding and subsequent discontinuation of a database, or the decision of a database maintainer to commercialise a resource and discontinue free web access even for academic users.

Some thoughts on unique and stable identifiers

There are some important observations to make about cross-referencing in general. To implement cross-referencing to a database, that database needs to provide unique and stable identifiers (USI) for each of their entries. These USI are often known as accession numbers. Such a requirement may seem obvious, but it is still often the case that databases do not see the need for stable identifiers. For example, a species-specific database may

use gene names as their unique identifiers. The problem is that such identifiers may be unique but are certainly not stable as it is most probable that some of the gene names will change over time. Far more important for future developments is our belief that major objects in a database require their own independent sets of USI. We became aware of this when we saw the need to add USI to a number of objects in Swiss-Prot, thus allowing external databases to seamlessly implement cross-references to a specific object in Swiss-Prot rather than at the level of the entire entry. A good example of such developments is the creation of feature identifiers (FTId) for all human protein sequence variants in Swiss-Prot. These identifiers allow specialised databases that report mutations concerning a specific set of genes to make a cross-reference to the representation of that mutation in Swiss-Prot.

The NiceProt view

MAKING SWISS-PROT AVAILABLE TO THE USERS

In prehistoric times – ie before the Web – Swiss-Prot reached its users by a variety of means. It was sent on computer tapes by the EMBL, it was distributed on floppy disks by companies selling sequence analysis software and, in 1989, it became the first major biomolecular database to be distributed on CD-ROM. In parallel to the physical distribution of Swiss-Prot, the database was made available by anonymous ftp and was searchable from a number of on-line resources such as BIONET and the NCBI IRX database retrieval software.

When the World-Wide Web began in 1993, Swiss-Prot became available on the ExpASy¹⁹ server,²⁰ which was born on 1st August, 1993. At that date there were fewer than 150 web servers worldwide. To the best of our knowledge ExpASy was the first web server for the life science community. We were very pleased to see that it was accessed 7,295 times during its first month of activity. We never imagined that a few years later it would be accessed at a rate of 8–10 million hits per

month. It has now been accessed more than 300 million times by a total of more than three million computer hosts from 200 countries. Seven mirror sites, ie exact copies of the main site in Switzerland, have been established in Australia, Bolivia, Canada, China, Korea, Taiwan and the USA. It is also noteworthy to mention that ExpASy and the EBI server²¹ are far from being the only web servers that redistribute Swiss-Prot and TrEMBL, we estimate that there are about 50 such sites worldwide.

ExpASy has constantly evolved in its ten years of existence. It is outside the scope of this paper to describe all of what is available on the server, yet we want to point out two significant developments that reflect our response to the needs of users.

In autumn 1998, we initiated 'NiceProt', with the intention to provide scientists with a more user-friendly way of looking at Swiss-Prot and TrEMBL entries. Instead of showing the raw Swiss-Prot data format (with its two-letter line types), we decided to make use of HTML tables to group certain fields under common headings, to replace the line type by a more explicit key (eg 'Cross-references' instead of 'DR'). This was initially targeted at users who are not familiar with the Swiss-Prot data format, but rapidly caught on in the scientific community. Gradually, more and more functionalities were added, including many implicit cross-references, and links to context-specific documentation. During the first eight months of 2003, ExpASy treated about 1 million requests for individual Swiss-Prot or TrEMBL entries on average per month. An overwhelming majority of these hits (85 per cent) are for NiceProt, whereas the remaining 15 per cent account for accesses to the raw text version, or the 'htmlised' view that was prevalent prior to September 1998.

The NEWT²² taxonomy browser²³ is a service introduced in 2002 that serves as an entry point into Swiss-Prot and TrEMBL using taxonomic search criteria.

The NEWT Taxonomy browser

The core of NEWT consists in the integration of Swiss-Prot specific taxonomy information with the NCBI taxonomy data in a relational database. Taxonomic nodes are stored in a hierarchical tree; this allows easy navigation through the taxonomy lineage from every taxon. The web interface to NEWT allows users to search and browse the daily updated taxonomy data. Users can navigate through the taxonomy tree and access corresponding Swiss-Prot and TrEMBL protein entries. Additionally, a manually curated selection of over 24,000 external links (including more than 13,000 photographs) provides specific information on selected species.

Both UniProt and NEWT are representatives of the trend toward a 'customisation' of the representation of knowledge. We believe that this trend will not abate; there are many specific communities of life scientists that require information on proteins, yet want them to be represented in a style or perspective specific to their field of research. We are in the process of developing new types of views.

Mining the server log files

We also believe that the ExpASY server access log files are a valuable source of information as to the most frequently consulted TrEMBL entries (ie unannotated entries that will greatly benefit from manual annotation) scientists' use of search engines, the context in which certain entries are consulted etc. We therefore plan to mine the ExpASY log files and expect to be able to draw enlightening conclusions!

CONCLUSIONS

Being a well-established database, we can say that the tireless effort of juggling between evolution and stability has been an exhausting but suitable strategy for the development of the Swiss-Prot protein knowledgebase. Early design features of the database such as the detailed structuring of the entry format, the standardisation of nomenclature, the regular review of the annotation of protein families have been shown to be

indispensable. The explosive growth in uncharacterised sequence data has led us to the implementation of automatic and semi-automatic processes. They are designed to ensure the same high-quality standards that have always been the hallmark of Swiss-Prot. Automation has to go in parallel with the introduction of evidence tags that will allow distinguishing data sources and inferences. We strongly believe that the future of Swiss-Prot and of any similar curated information resource relies on the active participation of the life sciences community. This will require an increased educational effort on our part. It is also dependent on the commitment of scientific societies, publishers and funding agencies to provide a framework to facilitate community efforts and give due credit to the participating scientists.

As a closing remark, we would like to thank all the persons involved in the development of Swiss-Prot at the SIB and EBI as well as all the funding agencies and companies that have financially contributed to the continuous evolution of the Swiss-Prot knowledgebase.

Acknowledgments

The work described in this paper covers activities funded by various sources including NIH:1 U01 HG02712-01, EU:BioMinT; QLRT-2001-02770, EU:Temblor; QLRT-2001-00015, EU:BioBabel; QLRI-CT-2001-00981, SNF:3100-063879.

References

1. Boeckmann, B., Bairoch, A., Apweiler, R. *et al.* (2003), 'The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003', *Nucleic Acids Res.*, Vol. 31, pp. 354-370.
2. Bairoch, A. (2000), 'Serendipity in bioinformatics, the tribulations of a Swiss bioinformatician through exciting times!', *Bioinformatics*, Vol. 16, pp. 48-64.
3. Apweiler, R., Bairoch, A., Wu, C. H. *et al.* (2004), 'UniProt: The universal protein knowledgebase', *Nucleic Acids Res.*, Vol. 32, pp. D115-D119.
4. Dayhoff, M. O., Eck, R. V., Chang, M. A. and Sochard, M. R. (1965), 'Atlas of Protein Sequence and Structure', Vol. 1, National

- Biomedical Research Foundation, Silver Spring, MD.
5. Moore, J., Engelberg, A. and Bairoch, A. (1988), 'Using PC/GENE for protein and nucleic acid analysis', *Biotechniques*, Vol. 6, pp. 566–572.
 6. URL: <http://www.expasy.org/cgi-bin/experts>
 7. Monigatti, F., Gasteiger, E., Bairoch, A. *et al.* (2002), 'The Sulfinator: Predicting tyrosine sulfation sites in protein sequences', *Bioinformatics*, Vol. 18, pp. 769–770.
 8. Bologna, G., Veuthey, A.-L., Yvon, C. *et al.* (2004), 'N-terminal myristoylation predictions by ensembles of neural networks', *Proteomics*, in press.
 9. Hulo, N., Sigrist, C., LeSaux, V. *et al.* (2004), 'Recent improvements to the PROSITE database', *Nucleic Acids Res.*, Vol. 32, pp. D134–D137.
 10. Gattiker, A., Michoud, K., Rivoire, C. *et al.* (2003), 'Automated annotation of microbial proteomes in Swiss-Prot', *Comput. Biol. Chem.*, Vol. 27, pp. 49–58.
 11. Warner, A. (URL: <http://www.lexonomy.com/publications/aTaxonomyPrimer.html>).
 12. Bairoch, A. (2000), 'The ENZYME database in 2000', *Nucleic Acids Res.*, Vol. 28, pp. 304–305.
 13. Farriol-Mathis, N., Garavelli, J. S., Boeckmann B. *et al.* (2004), 'Annotation of post-translational modifications in the Swiss-Prot knowledgebase', *Proteomics*, in press.
 14. Ashburner, M., Ball, C. A., Blake, J. A. *et al.* (2000), 'Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium', *Nature Genet.*, Vol. 25, pp. 25–29.
 15. URL: <http://www.ebi.ac.uk/swissprot/SP-ML>
 16. Etzold, T. and Argos, P. (1993), 'SRS – an indexing and retrieval tool for flat file data libraries', *Comput. Appl. Biosci.*, Vol. 9, pp. 49–57.
 17. Mulder, N. J., Apweiler, R., Attwood, T. K. *et al.* (2003), 'The InterPro Database, 2003 brings increased coverage and new features', *Nucleic Acids Res.*, Vol. 31, pp. 315–318.
 18. Gasteiger, E., Jung, E. and Bairoch, A. (2001), 'SWISS-PROT: Connecting biological knowledge via a protein database', *Curr. Issues Mol. Biol.*, Vol. 3, pp. 47–55.
 19. Gasteiger, E., Gattiker, A., Hoogland, C. *et al.* (2003), 'ExPASy – the proteomics server for in-depth protein knowledge and analysis', *Nucleic Acids Res.*, Vol. 31, pp. 3784–3788.
 20. URL: <http://www.expasy.org>
 21. URL: <http://www.ebi.ac.uk>
 22. Phan, I. Q., Pilbout, S. F., Fleischmann, W. and Bairoch, A. (2003) 'NEWt, a new taxonomy portal'. *Nucleic Acids Res.*, Vol. 31, pp. 3822–3823.
 23. URL: <http://www.ebi.ac.uk/newt/>