Plant protein annotation in the UniProt Knowledgebase

Schneider, Michel; Bairoch, Amos Marc; Wu, Cathy H; Apweiler, Rolf

How to cite

SCHNEIDER, Michel et al. Plant protein annotation in the UniProt Knowledgebase. In: Plant physiology, 2005, vol. 138, n° 1, p. 59–66. doi: 10.1104/pp.104.058933

This publication URL: https://archive-ouverte.unige.ch/unige:38249
Publication DOI: 10.1104/pp.104.058933

Plant Protein Annotation in the UniProt Knowledgebase[1]

Michel Schneider*, Amos Bairoch, Cathy H. Wu, and Rolf Apweiler

Swiss Institute of Bioinformatics (M.S., A.B.), and Department of Structural Biology and Bioinformatics (A.B.), Centre Medical Universitaire, University of Geneva, 1211 Geneva 4, Switzerland; Department of Biochemistry and Molecular Biology, Georgetown University Medical Center, Washington, DC 20057–1414 (C.H.W.); and European Molecular Biology Laboratory Outstation, European Bioinformatics Institute, Hinxton, Cambridge CB10 1SD, United Kingdom (R.A.)

The Swiss-Prot, TrEMBL, Protein Information Resource (PIR), and DNA Data Bank of Japan (DDBJ) protein database activities have united to form the Universal Protein Resource (UniProt) Consortium. UniProt presents three database layers: the UniProt Archive, the UniProt Knowledgebase (UniProtKB), and the UniProt Reference Clusters. The UniProtKB consists of two sections: UniProtKB/Swiss-Prot (fully manually curated entries) and UniProtKB/TrEMBL (automated annotation, classification and extensive cross-references). New releases are published fortnightly. A specific Plant Proteome Annotation Program (http://www.expasy.org/sprot/ppap/) was initiated to cope with the increasing amount of data produced by the complete sequencing of plant genomes. Through UniProt, our aim is to provide the scientific community with a single, centralized, authoritative resource for protein sequences and functional information that will allow the plant community to fully explore and utilize the wealth of information available for both plant and nonplant model organisms.

BRIEF HISTORY

From the "Atlas" and PIR-PSD to Swiss-Prot

The history of protein sequence databases began when Margaret Dayhoff started to assemble all the information related to known protein sequences in a book called ''Atlas of Protein Sequence and Structure.'' The first edition, published in 1965 (Dayhoff et al., 1965), included 65 proteins. In 1984, the Protein Information Resource (PIR) of the National Biomedical Research Foundation extended this work by establishing the PIR-International Protein Sequence Database (PIR-PSD), the first computer protein sequence data bank ever created. Later, Amos Bairoch launched PIR+, an extended version based on the format of the European Molecular Biology Laboratory (EMBL) nucleotide sequence database with several advanced features. In 1986, this database started to be freely distributed under the name of Swiss-Prot. The first release contained roughly 3,900 annotated proteins.

Swiss-Prot and TrEMBL

In 1996, Swiss-Prot already contained 83,000 entries. However, the exponential data influx generated by genome sequencing projects resulted in a situation where most newly identified proteins were not readily available in the database. To cope with this problem, a complementary database, TrEMBL, was introduced. TrEMBL consists of computer-annotated entries derived from the translation of all coding sequences (CDS) proposed by authors in their sequence submission to EMBL/GenBank/DNA Data Bank of Japan (DDBJ), except for CDS already included in Swiss-Prot. Any entries redundant with Swiss-Prot/TrEMBL are merged and the remainder then progress into TrEMBL, awaiting manual annotation and subsequent transfer into Swiss-Prot.

Universal Protein Resource

Until 2002, Swiss-Prot/TrEMBL (Boeckmann et al., 2003) and PIR-PSD (Wu et al., 2003) still coexisted as two independent protein databases, although the contents of the databases and the priorities and their approaches to annotation were in fact complementary. Therefore, the Swiss Institute of Bioinformatics (SIB), the European Bioinformatics Institute (EBI), and the PIR group at the Georgetown University Medical Center and National Biomedical Research Foundation decided to join forces and form the Universal Protein Resource (UniProt) Consortium (Apweiler et al., 2004; http://www.uniprot.org). Its main goal is to provide a single, centralized, authoritative resource for protein sequences and functional information.

STRUCTURE OF UNIPROT

UniProt, described in detail in Apweiler et al. (2004) and in Bairoch et al. (2005), consists of three different sections, each optimized for a different use (Fig. 1).

UniProt Knowledgebase

Since the creation of UniProt, Swiss-Prot and TrEMBL ceased to exist as independent databases,

[1] This work was supported by the National Institutes of Health (grant no. U01 HG02712) and by the Swiss Federal Office of Education and Science and Genoplante (project no. Bi2001071).

* Corresponding author; e-mail michel.schneider@isb-sib.ch; fax 41–22–379–58–58.

www.plantphysiol.org/cgi/doi/10.1104/pp.104.058933.

and they are now integral parts of the core section of UniProt, the UniProt Knowledgebase (UniProtKB). For continuity, the two names have been kept and what is now called UniProtKB/Swiss-Prot (Bairoch et al., 2004) contains all the nonredundant, fully manually annotated records, while UniProtKB/TrEMBL consists of all the computationally analyzed records awaiting full manual annotation. All suitable PIR-PSD sequences and annotations missing from the original Swiss-Prot or TrEMBL databases have been integrated into the UniProtKB. Taken together, UniProtKB/Swiss-Prot and UniProtKB/TrEMBL cover all proteins characterized or inferred from nucleotide sequences identified so far in any species, archaea, bacteria, or eukaryote.

UniProt Archive

The UniProt Archive (UniParc) is an archive that contains original protein sequences loaded from many sources such as UniProtKB/Swiss-Prot, UniProtKB/TrEMBL, PIR-PSD, the Ensembl database of animal genomes, the National Center for Biotechnology Information (NCBI) Reference Sequence collection, model organism databases such as FlyBase and WormBase, and protein sequences from the European, American, and Japanese patent offices. Sequence fragments are kept as separate entries. Every UniParc entry contains cross-references to the source databases from which the protein sequence was extracted.

UniProt Reference Clusters

UniProt Reference Clusters (UniRef) provides three nonredundant reference clusters of sequence data, UniRef100, UniRef90, and UniRef50. UniRef100 combines identical sequences across different species, including all splice isoforms and sequence fragments, into a single record. UniRef90 and UniRef 50 collapse all UniRef100 sequences that are at least 90% or 50% identical into a single cluster using the CD-HIT algorithm (Li et al., 2001), presenting only one representative sequence for each cluster. The three UniRef databases allow the user to choose between a truly comprehensive search and a fast one by reducing the size of the UniRef100 by approximately 40% in UniRef90 and by approximately 65% in UniRef50.

Prospective

DDBJ has joined the UniProt Consortium and will start in the near future to contribute to the project by both confirming sets of gene predictions for the Japanese cultivar of rice (*Oryza sativa*) and by assigning functional annotation, and secondary and tertiary structure to predicted proteins based on translations of a set of cDNA sequences deposited in public databases by Japanese consortiums.

TECHNICAL SPECIFICATIONS OF UNIPROT

Manual and Automatic Annotation

Proteins for which functional, biochemical, and/or structural data are published are the main targets for manual annotation. Curators add annotations such as protein functions, biologically relevant domains and sites, posttranslational modifications, subcellular location of the protein, developmental- or tissue-specific expression of the protein, splice isoforms, and the references used in the annotation process. Once manually annotated, entries are stored in the UniProtKB/Swiss-Prot section of the UniProtKB.

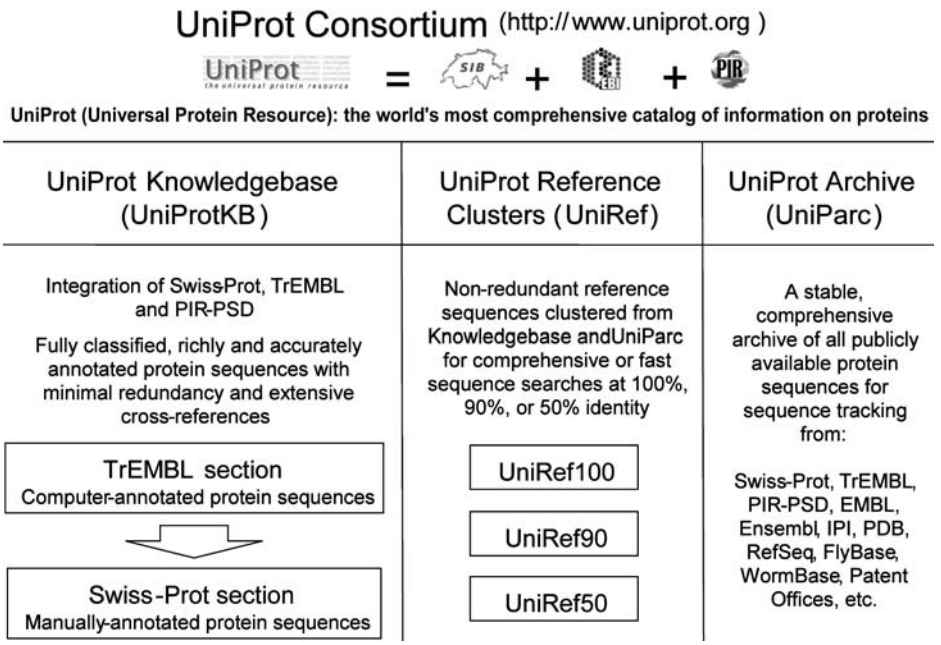

Figure 1. Structure and organization of UniProt.

Since the number of protein sequences in UniProtKB/TrEMBL continues to grow exponentially, an automatic procedure providing high-throughput annotation and functional characterization is required. InterPro (Mulder et al., 2003) combines several databases, for example PROSITE (Hulo et al., 2004), Pfam (Bateman et al., 2002), and TIGRFAMs (Haft et al., 2003), which use different methodologies to derive protein signatures. Based on such motifs and domains, protein sequences are grouped into families and superfamilies. The annotation associated with all functionally characterized UniProtKB/Swiss-Prot proteins belonging to the same family is then transferred to the UniProtKB/TrEMBL entries. This transfer is controlled by RuleBase (Apweiler, 2001), a database containing more than 500 annotation rules and conditions, and another rule set generated by decision trees (Kretschmann et al., 2001). At this step, all sequences derived from the same species that are strictly 100% identical are merged into a single entry.

Automatic annotation relies also on the PIRSF family classification concept (Wu et al., 2004a); proteins classified in the same family are both homologous (sharing common ancestry) and homeomorphic (sharing full-length sequence similarity with common domain architecture). Based on this system, rules are developed and manually curated for annotating and propagating position-specific features, such as active or binding sites, and for protein names and gene ontology (GO) terms.

Format of the Database

The main distribution format of UniProt is as an ASCII flat file. Since their creation, Swiss-Prot and TrEMBL have used a data format that followed as closely as possible that of the EMBL Nucleotide Sequence Database. For integration of the two databases into the UniProtKB, this original format was maintained and both UniProtKB/Swiss-Prot and UniProtKB/TrEMBL entries are structured so as to be usable by human readers as well as by computer programs. Each line begins with a two-character line code that indicates the type of data contained in the line. The current line types and line codes and the order in which they appear in an entry are described in the UniProt user manual (http://www.expasy.org/sprot/userman.html).

Availability and Distribution

The UniProt databases are released biweekly. The UniProtKB is distributed as two gzip-compressed files (uniprot_sprot.dat.gz and uniprot_trembl.dat.gz) that, when decompressed, produce ASCII files in a flat-file format. The same data are also available in FASTA and XML formats. This core data set is further supplemented by two files available under the FASTA format exclusively, containing the sequences of all additional splice isoforms annotated in UniProtKB/Swiss-Prot and UniProtKB/TrEMBL sections (uniprot_sprot_varsplic.fasta and uniprot_trembl_varsplic.fasta). The program VARSPLIC (Kersey et al., 2000) is used to generate those additional records.

In addition to the complete databases, we will soon provide the data in taxonomic divisions for archaea, bacteria, fungi, human, invertebrates, mammals, plants, rodents, vertebrates, viruses, and unclassified.

The three UniRef databases are downloadable as FASTA or XML files. The FASTA files, containing representative sequences of the UniRef clusters, are useful for FASTA, BLAST, and other sequence similarity searches. However, the sequence files do not contain annotations, which can be generated from the UniProtKB if needed. As a sequence archive containing both active and obsolete sequences and no annotation information, UniParc is unsuitable for large-scale parsing or manipulation, and is therefore not available for download.

The UniProt distribution files can be downloaded from the anonymous FTP servers maintained by the UniProt Consortium at SIB (ftp.expasy.org), EBI (ftp.ebi.ac.uk), and PIR (ftp.uniprot.org). The files are located in the same directory (/databases/uniprot/knowledgebase/) at all three FTP sites, and are easily accessible from the download center at the UniProt Web site (http://www.uniprot.org/database/download.shtml).

PLANTS IN UNIPROT

The Plant Proteome Annotation Program

Shortly after the publication of the complete genome of Arabidopsis (*Arabidopsis thaliana*) and the prediction of 25,498 protein-encoding genes (Arabidopsis Genome Initiative, 2000), the Swiss-Prot group initiated the Plant Proteome Annotation Program (PPAP). This program is focused on the annotation of plant-specific proteins and protein families (i.e. originating from viridiplantae (or green plants) as defined by the Tree of Life project (http://tolweb.org), and each protein is annotated according to our usual standards (Boeckmann et al., 2003). Our major effort is directed toward Arabidopsis, without neglecting annotation of proteins from other plant species (rice, maize [*Zea mays*], wheat [*Triticum aestivum*], poplar [*Populus* spp.], soybean [*Glycine max*], *Medicago sativa*, etc.). At the end of February 2005 (UniProt release 4.2), 11,970 and 174,229 plant sequence entries are present in the UniProtKB/Swiss-Prot and UniProtKB/TrEMBL sections, respectively. Fifteen plant species are represented in UniProtKB/Swiss-Prot with 100 or more proteins annotated (Table I), while more than 2,700 different plant species are present in the manually annotated section of the UniProtKB.

Arabidopsis

Arabidopsis is considered as a common model organism for plants and, as such, is the focus of our

main effort in plant protein annotation. Currently, 3,124 proteins have been manually annotated in the UniProtKB/Swiss-Prot section. A detailed list, including the chromosome locus, the UniProt accession number, entry name, description, and gene name(s), can be retrieved from the PPAP Web site (http://www.expasy.org/cgi-bin/lists?arath.txt). All 67 proteins encoded in the chloroplast and 114 proteins encoded in the mitochondrion are present in UniProtKB/Swiss-Prot.

Alternative Splicing

A list of more than 2,500 Arabidopsis genes with alternatively spliced gene models was published by The Institute for Genomic Research in 2003 (http://www.tigr.org/tdb/e2k1/ath1/altsplicing/splicing_variations.shtml). For several of these genes, the alternative splicing occurs in the 5′- or 3′-untranslated regions of the gene and has no effect on the sequence of the encoded protein. For the remaining proteins that are present in UniProtKB/Swiss-Prot, the various isoforms are annotated and the feature table of the entry allows the recreation of the different splice variants.

Since splice isoforms may differ considerably from one to another, with potentially less than 50% sequence similarity between isoforms, it may be of interest to extend similarity searches to all the various isoforms. As indicated previously, the VARSPLIC program can recreate all annotated splice variants from the feature table of a UniProtKB/Swiss-Prot entry. Most sequence analysis and proteomic tools (BLAST or PeptIdent, for example) present on the ExPASy Web server (http://www.expasy.org) have been adapted to take into account all annotated splice isoforms. As the splice isoforms have also been incorporated into UniRef100, they are also directly searchable and retrievable from the UniProt Web site.

Table I. *The fifteen most-represented plants species in the UniProtKB/Swiss-Prot section (Release 4.2)*

Rank	No. of UniProtKB/Swiss-Prot Entries	Plant Species
1	3,124	Arabidopsis
2	513	Maize
3	448	Rice
4	368	Tobacco (*Nicotiana tabacum*)
5	302	Pea (*Pisum sativum*)
6	283	Wheat
7	272	Barley (*Hordeum vulgare*)
8	264	Soybean
9	262	Tomato (*Lycopersicon esculentum*)
10	256	Potato (*Solanum tuberosum*)
11	238	Spinach (*Spinacia oleracea*)
12	212	Chlamydomonas (*Chlamydomonas reinhardtii*)
13	153	Liverwort (*Marchantia polymorpha*)
14	135	Rape (*Brassica napus*)
15	106	Mesostigma (*M. viride*)

AGI Gene Codes

In 1999, the Arabidopsis Genome Initiative (AGI) proposed a uniform gene nomenclature system (http://mips.gsf.de/proj/thal/db/about/agicodes.html). Although those codes identify specific chromosome loci and not proteins, they are useful for labeling and discriminating between highly similar genes. However, users should note that AGI numbers are not completely stable and are subject to change when gene models are split or merged, for example. Since every existing Arabidopsis entry in the UniProtKB/Swiss-Prot section is correlated with genomic data, the AGI number (e.g. At5g39190) is indicated in the gene name line. The few exceptions represent sequences that are either in yet-unsequenced regions of the Arabidopsis genome or have major problems. When available, the bacterial artificial chromosome name or names (e.g. F26P21.150) are also indicated.

RNA Editing

RNA editing that leads to one or more amino acid changes is common in chloroplasts and plant mitochondria. The protein sequence shown in UniProtKB/Swiss-Prot entries is always the translation of the edited RNA. If the CDS proposed in the EMBL entry does not correspond to this edited protein sequence, then the corresponding cross-reference in the UniProt entry is flagged with an alternative sequence (ALT_SEQ) statement. Editing is sometimes partial at the RNA level. In plants, we do not take this into account, as it seems that only the protein resulting from the fully edited RNA is functional (A. Brennicke, personal communication).

Genoplante

Since 2003, Swiss-Prot has participated in Genoplante, an initiative of the French joint program in plant genomics. This project is based on a network of scientists working in different fields of research. One of its goals is to obtain extensive, homogeneous, reliable, documented, and traceable annotations for Arabidopsis nuclear genes and gene products. Working in a family-oriented manner, all expert-curated annotations of paralogous genes are gathered into an added-value database named GeneFarm (Aubourg et al., 2005). When available, cross-links between UniProtKB/Swiss-Prot and GeneFarm entries are provided.

Rice

In April 2002, a draft sequence of the rice genome became available (Goff et al., 2002; Yu et al., 2002). Although a large part of the sequence is still not annotated or freely available, we decided to start a new rice annotation effort. We will work in close

collaboration with the DDBJ group, a new UniProt Consortium member, to manually annotate proteins for which experimental data are available. In this way, we will provide the scientific community with high quality data and facilitate comparisons between monocot and dicotyledonous plants. Since the genome assembly is incomplete and not yet stable, we will concentrate on the annotation of well-characterized proteins.

By choice, the taxonomical classification used in the UniProtKB/Swiss-Prot section stops at the species level (Phan et al., 2003); we do not distinguish between subspecies or cultivars, although we indicate them, when available, in the ''Reference Comment'' lines of the entry. Consequently, all the rice entries are grouped under a single TaxID (TaxID = 4530) and information concerning the same protein extracted from an Indica cultivar group or a Japonica cultivar group is merged in a single entry, with eventual differences in the sequence annotated as variants. However, due to the data submission procedure to the nucleotide sequence databases, rice entries in the UniProtKB/TrEMBL section are stored under three different TaxIDs (TaxID = 39946 for the Indica cultivars, TaxID = 39947 for the Japonica cultivars, and TaxID = 4530 when the subspecies is not specified). We are currently addressing this problem to resolve this inconsistency between the two UniProtKB sections.

iProPlants

Plant-Containing and Plant-Specific Protein Families

The PIR group has started a new project called iProPlants, extended from the iProClass database of integrated protein family, structure, and function (Wu et al., 2004b). iProPlants will be a plant-centric, integrated protein resource coupling PIRSF family classification and data integration to facilitate functional annotation of plant genomes. Plant-specific and plant-containing protein families will be catalogued and curated in the PIRSF classification framework for comparative analyses of protein functions. This systematic, classification-driven approach will allow standardized and rich annotation across plant genomes.

Currently, there are over 5,000 curated PIRSF families and 36,000 not-yet-curated preliminary clusters covering over two-thirds of UniProt sequences. Among them are about 3,900 plant-containing families (1,300 curated, 2,600 preliminary clusters), consisting of protein members from plants and other taxonomic division, as well as over 2,300 plant-specific families (over 300 curated, 2,000 preliminary clusters), consisting of plant proteins only. Over 10,000 Arabidopsis proteins have been classified, with about 40% coverage of the Arabidopsis proteome.

Community Annotation of Plant Protein Families

To provide a research infrastructure for community annotation of plant protein families, the PIRSF family curation interface will be made available to collaborating plant researchers. The family curation platform is implemented in the n-tier J2EE software framework with a JavaWebStart client, which allows worldwide curators to access the PIRSF classification system from a Web browser at any time, always using the most current versions of software tools and data. The curation interface can launch several analysis and visualization tools and a directed acyclic graph (DAG) editor. The tools include iterative BLASTClust with a tree view, multiple sequence alignment with phylogenetic tree and protein annotation table, a taxonomy tree browser, and the SEED program for genome context and phylogenetic profile analysis (Osterman and Overbeek, 2003). Interested researchers can register (pirmail@georgetown.edu) to access this installation-free Web-based family curation system via a sign-on mechanism with role-based identification and password authentication.

WHAT DIFFERENTIATES UNIPROT FROM OTHER DATABASES?

Manual Annotation and Minimal Redundancy

Currently, the UniProt project has over 100 staff members, with five plant biologists assigned full time to the curation of plant proteins and protein families. Plant-specific proteins or proteins involved in metabolic or signaling pathways important in plants get a high priority for manual annotation. Most information is extracted from journal articles and occasionally patents. We also rely on direct submission to the database and on experts who help with nomenclature or provide first-hand experimental data.

Curators minimize database redundancy by merging all data from different literature reports into a single entry. When several protein sequences are available, they are compared through multiple alignments. If the sequences differ by only a few dispersed amino acids, the annotator checks if the same cultivar has been used as starting material and if she/he is really dealing with a single-copy gene. If that is the case, the differences are annotated as conflicts. If several contiguous amino acids differ in only one sequence, then the curator goes back to the DNA level and compares the sequences of the genes. That allows discrimination between simple frameshifts or alternative splicing, events that are annotated accordingly. Moreover, if a protein is defined only by a single-gene model prediction created by a computer program running on a genomic sequence, then multiple alignments with other members of the family to which the protein belongs (paralogs) and/or with proteins having the same function in other related species (orthologs) allow checking and often correction of the initial gene model prediction by confirming the presence and position of the end points of the various exons.

The sequence shown in each UniProtKB/Swiss-Prot entry is the most correct sequence version according to annotator judgment, and differences attributed to splice variants, polymorphisms, experimental sequence modifications, or sequencing errors are indicated in the feature table. Consequently, only a single entry usually relates to one given protein.

We have established collaboration with The Arabidopsis Information Resource (TAIR; Rhee et al., 2003) to provide more up-to-date Arabidopsis sequences in the UniProtKB/TrEMBL section. Presently, many UniProtKB/TrEMBL sequences are genomic bacterial artificial chromosome sequences annotated with predicted genes. Some of these will fail to match any of the currently annotated Arabidopsis proteins because of faulty gene predictions in the initial round of annotation. Others may match to two current Arabidopsis proteins or partially overlap with a current protein. There is an ongoing effort to map UniProt sequences to TAIR and to identify potential inconsistencies for correction. Reciprocally, when newly identified Arabidopsis proteins that are not yet correlated with specific chromosome loci are manually added to UniProtKB/Swiss-Prot, TAIR is asked to assign new AGI numbers.

Cross-References

UniProt serves as a central hub for biomolecular information with access to more than 60 other resources. It provides cross-references to external data collections such as the underlying DNA sequence entries in the DDBJ/EMBL/GenBank nucleotide sequence databases, two-dimensional (2D) PAGE and three-dimensional protein structure databases, various protein domain and family characterization databases, posttranslational modification databases, species-specific data collections, variant databases, and disease databases. A document listing all databases cross-referenced in UniProt (http://www.uniprot.org/support/docs/dbxref.shtml) contains, for each database, a short description and the server URL. This interconnectivity is achieved via database cross-reference lines. In addition, the manually annotated UniProtKB/Swiss-Prot section has 30 other implicit links that are automatically generated when entries are accessed through the ExPASy Web server or one of its mirror sites.

Of special interest for plant biologists are the links to TAIR; to Gramene, a comparative mapping resource for grains (Ware et al., 2002); to MaizeDB, the Maize genome database (Lawrence et al., 2004); to MAIZE-2DPAGE, the Maize genome 2D electrophoresis database (Touzet et al., 1996); to GeneFarm (Aubourg et al., 2005); and to SWISS-2DPAGE, a database of proteins identified on 2D PAGE, including one of Arabidopsis origin, maintained by the Geneva University Hospital (Hoogland et al., 2000).

The link to TAIR is currently implicit and, therefore, is only implemented on the ExPASy server and its mirror sites. A mapping of UniProt sequences to TAIR is ongoing and explicit links will be incorporated in the near future. For the time being, each time a link is followed, a ''Quick search'' form is automatically filled on the TAIR Web page with the AGI number displayed for the UniProtKB/Swiss-Prot entry.

Note that all the plant proteins whose structures have been determined are cross-linked to the corresponding Protein Data Bank entries (Bhat et al., 2001) and 70% (324 out of 466) have been fully manually annotated and incorporated into UniProtKB/Swiss-Prot.

Controlled Vocabularies and GO Terms

To facilitate text searches and database interoperability, controlled vocabularies are used for several annotation fields, such as species names, strains, tissues, keywords, or description of the posttranslational modifications. A number of lists of controlled vocabularies (strains.txt, keywlist.txt, tisslist.txt, etc.) can be found in the UniProt documentation.

Whenever available, we use the official gene or protein names provided by international nomenclature committee while still providing all the published synonyms. Collaborations and regular data exchange with other databases and organizations allow the implementation of community-specific nomenclatures. The UniProtKB uses a unified keyword list based on the previous Swiss-Prot keywords augmented by selected PIR keywords that represent new concepts or new parent/child nodes of preexisting Swiss-Prot keywords. In an attempt to address the need for consistent descriptions of gene products, several databases joined forces to form the GO Consortium (Gene Ontology Consortium, 2000). Three ontologies that describe gene products in terms of their associated biological processes, cellular components, and molecular functions in a species-independent manner have been created. In the UniProtKB/TrEMBL section, GO terms are assigned to gene products by a combination of electronic and manual annotation, while UniProtKB/Swiss-Prot entries contain only GO terms with experimental evidence codes.

Evidence Attribution

Since 2001, UniProt is introducing evidence tags in UniProtKB/TrEMBL entries. All relevant data will ultimately be linked to one or several pieces of evidence that support the information. In addition, the classification of the evidence tags in three categories allows the user to discriminate between the various sources of data. ''Evidence category = 'program''' indicates that a computer program has created and added this piece of information. ''Evidence category = 'import''' flags data imported from other sources, while information manually added by the annotator are labeled as ''evidence category = 'curator.''' Evidence tags are available from the XML

distribution file. The introduction of evidence tags to the UniProtKB/Swiss-Prot section is scheduled for the second part of 2005.

This implementation of evidence tags will allow the user to easily identify particular classes of data of interest such as experimentally proven protein annotation. The current amount of such experimentally verified data being rather limited, we invite direct submission of any new or missing information from the scientific community. Consequently, a bibliography submission system has been developed by UniProt (http://www.expasy.uniprot.org/bibliography/bibliosubmit.shtml). So far, a few thousand experimental features have been incorporated into the UniProtKB after being associated with publications and cross-referenced to the corresponding PubMed identifiers.

CONCLUSION

The creation of the UniProt Consortium established the basis for the building of a single, centralized, authoritative resource for protein sequences and functional information. With its two sections, UniProtKB/Swiss-Prot and UniProtKB/TrEMBL, the UniProtKB contains all known proteins, without species restriction.

In addition, UniProtKB/Swiss-Prot is a manually curated section, which means that biologists are looking at each entry, merging all redundant information, and validating the data included.

Our strategy at the UniProt Consortium has been to concentrate on generalist annotation by adding information gathered mainly from scientific publications and by establishing links to other specialized sources of annotation. Therefore, instead of trying to integrate all possible items or data into our entries, we decided to establish explicit cross-references. Of course, by doing that we relinquish control over part of the information that might be important for complete annotation of a protein. For example, when the Mendel database ceased regular updates, most of its information about the official nomenclature for plant genes became obsolete and links to this database had to be removed.

Our approach strongly supports the share-and-survive model advocated recently by Rhee (2004). Indeed, the launch of the UniProt Consortium combining forces from Europe, the United States, and Japan is a strong signal that we are convinced that we all belong to the same virtual community. The only way to avoid being overwhelmed by the huge amounts of biological data being produced daily is to work together and to share our forces and expertise.

Consistent with this philosophy, we urge that in addition to publishing their results, scientists should also always deposit their raw data in a public repository. Unfortunately, an increasing number of scientists are showing results about genes or PCR products without submitting the corresponding sequences to public nucleic acids databases. Statements like ''We isolated a clone X basically similar to the Z sequence'' are more and more frequent, and without access to the original DNA sequence, there is no way to discriminate between basic similarity due to an alternative splicing, gene duplication, sequencing error, or a problem of PCR fidelity. Potentially important information is not made available for the whole scientific community and is therefore lost.

Although every plant protein currently known is present in the UniProtKB, a specific effort was put on manual plant protein annotation. During the three years of existence of the PPAP, the number of plant entries manually annotated in UniProt by trained biologists has increased by almost 50%, while the number of Arabidopsis entries has more than doubled. With our newly established rice project, we will also increase the coverage of the detailed annotation of proteins of monocot origin.

Every new genome that is completely sequenced and for which a gene prediction has been performed is automatically added to the UniPortKB/TrEMBL section as soon as it is deposited in the public nucleotide databases (EMBL/GenBank/DDBJ). However, if the CDS annotation is missing or if the DNA sequences are deposited in a private database or in one with a restricted access, then the data are absent from the UniProtKB.

Classification into families, indication of plant-specific domains or features (e.g. targeting sequences directing the protein toward the chloroplast or the thylakoid) or of features highly represented in plant (like RNA editing for example), use of a clear and well established taxonomy, and the existence of a powerful retrieval system enable users to create numerous organism- or feature-specific datasets.

Due to the limited resources specifically allocated to curation of plant proteins, we are seeking active participation from the broad plant-scientific community by giving us feedback, informing us of high-priority proteins for annotation, and contributing to collaborative expert annotation. To maintain a high-quality database and continually improve the quality of our annotation, we actively solicit user feedback and input via our Web site. UniProt accepts submissions of new sequences, entry updates and corrections, and annotated bibliographic information for protein entries. Directions for submission are available at http://www.uniprot.org/support/submissions.shtml. When accessing UniProt through the ExPASy Web sites, links to specific pages for contacting us are added in the NiceProt view of any entry: ''Report form for updates or corrections of an existing (publicly available) UniProtKB/Swiss-Prot entry'' (http://www.expasy.org/sprot/update.html) and ''Request for priority annotation of UniProtKB/TrEMBL entry xxxxxx'' (http://www.expasy.org/cgi-bin/tr_annot_req.pl?xxxxxx).

Update requests get the highest priority for manual curation and, before the newly created or corrected

entry is released to the public, the submitter is contacted for approval. Unfortunately, we had less than 40 update requests or notification of errors concerning plant entries in 2004, although more than one million connections to the ExPASy Web sites, one of the three entry points to UniProt, are counted every month.

Through UniProt, we aim to provide a single, centralized, authoritative resource for protein sequences and functional information that allows the plant community to fully explore and utilize the wealth of information available in plant and nonplant model organisms.

ACKNOWLEDGMENTS

We would like to thank Alan Bridge, Michael Tognolli, and Sylvain Poux for critical reading of the manuscript.

Received December 23, 2004; returned for revision March 9, 2005; accepted March 21, 2005.

LITERATURE CITED

Apweiler R (2001) Functional information in SWISS-PROT: the basis for large-scale characterisation of protein sequences. Brief Bioinform **2:** 9–18

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, et al (2004) UniProt: the Universal Protein knowledgebase. Nucleic Acids Res **32:** D115–D119

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408:** 796–815

Aubourg S, Brunaud V, Bruyère C, Cock M, Cooke R, Cottet A, Couloux A, Déhais P, Deléage G, Duclert A, et al (2005) GeneFarm, structural and functional annotation of Arabidopsis gene and protein families by a network of experts. Nucleic Acids Res **33:** D641–D646

Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, et al (2005) The Universal Protein Resource (UniProt). Nucleic Acids Res **33:** D154–D159

Bairoch A, Boeckmann B, Ferro S, Gasteiger E (2004) Swiss-Prot: juggling between evolution and stability. Brief Bioinform **5:** 39–55

Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer ELL (2002) The Pfam protein families database. Nucleic Acids Res **30:** 276–280

Bhat TN, Bourne PE, Feng Z, Gilliland G, Jain S, Ravichandran V, Schneider B, Schneider K, Thanki N, Weissig H, et al (2001) The 3D macromolecular structure Protein Data Bank (PDB) prepared by Research Collaboratory for Structural Bioinformatics (RCSB). Nucleic Acids Res **29:** 214–218

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, et al (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res **31:** 365–370

Dayhoff MO, Eck RV, Chang MA, Sochard MR (1965) Atlas of Protein Sequence and Structure, Vol 1. National Biomedical Research Foundation, Silver Spring, MD

Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. Nat Genet **25:** 25–29

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). Science **296:** 92–100

Haft DH, Selengut JD, White O (2003) The TIGRFAMs database of protein families. Nucleic Acids Res **31:** 371–373

Hoogland C, Sanchez JC, Tonella L, Binz PA, Bairoch A, Hochstrasser DF, Appel RD (2000) The 1999 SWISS-2DPAGE database update. Nucleic Acids Res **28:** 286–288

Hulo N, Sigrist CJA, Le Saux V, Langendijk-Genevaux PS, Bordoli L, Gattiker A, De Castro E, Bucher P, Bairoch A (2004) Recent improvements to the PROSITE database. Nucleic Acids Res **32:** D134–D137

Kersey P, Hermjakob H, Apweiler R (2000) VARSPLIC: alternatively-spliced protein sequences derived from SWISS-PROT and TrEMBL. Bioinformatics **11:** 1048–1049

Kretschmann E, Fleischmann W, Apweiler R (2001) Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on Swiss-Prot. Bioinformatics **17:** 920–926

Lawrence CJ, Dong Q, Polacco ML, Seigfried TE, Brendel V (2004) MaizeGDB, the community database for maize genetics and genomics. Nucleic Acids Res **32:** D393–D397

Li W, Jaroszewski L, Godzik A (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics **17:** 282–283

Mulder N, Apweiler R, Attwood T, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, et al (2003) The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Res **31:** 315–318

Osterman A, Overbeek R (2003) Missing genes in metabolic pathways: a comparative genomics approach. Curr Opin Chem Biol **7:** 238–251

Phan IQ, Pilbout SF, Fleischmann W, Bairoch A (2003) NEWT, a new taxonomy portal. Nucleic Acids Res **31:** 3822–3823

Rhee SY (2004) Carpe diem: retooling the ''publish or perish'' model into the ''share and survive'' model. Plant Physiol **134:** 543–547

Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M, et al (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. Nucleic Acids Res **31:** 224–228

Touzet P, Riccardi F, Morin C, Damerval C, Huet JC, Pernollet JC, Zivy M, De Vienne D (1996) The maize two-dimensional gel protein database: towards an integrated genome analysis program. Theor Appl Genet **93:** 997–1005

Ware DH, Jaiswal P, Ni J, Yap IV, Pan X, Clark KY, Teytelman L, Schmidt SC, Zhao W, Chang K, et al (2002) Gramene, a tool for grass genomics. Plant Physiol **130:** 1606–1613

Wu CH, Huang H, Nikolskaya A, Hu Z, Barker WC (2004b) The iProClass integrated database for protein functional analysis. Comput Biol Chem **28:** 87–96

Wu CH, Nikolskaya A, Huang H, Yeh LS, Natale DA, Vinayaka CR, Hu ZZ, Mazumder R, Kumar S, Kourtesis P, et al (2004a) PIRSF: family classification system at the Protein Information Resource. Nucleic Acids Res **32:** D112–D114

Wu CH, Yeh LS, Huang H, Arminski L, Castro-Alvear J, Chen Y, Hu Z, Kourtesis P, Ledley RS, Suzek BE, et al (2003) The Protein Information Resource. Nucleic Acids Res **31:** 345–347

Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). Science **296:** 79–92