Chapter 2

A Guide to UniProt for Protein Scientists

Claire O'Donovan and Rolf Apweiler

Abstract

One of the essential requirements of the proteomics community is a high quality annotated nonredundant protein sequence database with stable identifiers and an archival service to enable protein identification and characterization. The scope of this chapter is to illustrate how Universal Protein Resource (UniProt) (The UniProt Consortium, Nucleic Acids Res. 38:D142–D148, 2010) can be best utilized for proteomics purposes with a particular focus on exploiting the knowledge captured in the UniProt databases, the services provided and the availability of complete proteomes.

Key words: Protein sequence database, Annotation, Stable identifiers, Complete proteome, Archive, Nonredundant

1. Introduction

The Proteomics community has evolved intensively over the last decade but one constant is the need to identify the resulting proteins and their potential functions. This requires the availability of a nonredundant protein sequence database, with maximal coverage including splice isoforms, disease variant(s) and posttranslational modifications. Sequence archiving is an essential feature in order to be able to interpret and maintain the proteomic set results. Stable identifiers, consistent nomenclature and controlled vocabularies are highly beneficial for protein identification. The last but by no means least requirement is the provision of detailed information on protein function, biological processes, and molecular interactions and pathways cross-referenced to appropriate external sources. In this chapter, we will show how the Universal Protein Resource fulfils these criteria.

Cathy H. Wu and Chuming Chen (eds.), *Bioinformatics for Comparative Proteomics*, Methods in Molecular Biology, vol. 694,
DOI 10.1007/978-1-60761-977-2_2,

2. Materials

The mission of the Universal Protein Resource (UniProt) is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information, which is essential for modern biological research. UniProt is produced by the UniProt Consortium, which consists of groups from the European Bioinformatics Institute (EBI), the Protein Information Resource (PIR), and the Swiss Institute of Bioinformatics (SIB). Its activities are mainly supported by the National Institutes of Health (NIH) with additional funding from the European Commission and the Swiss Federal Government.

It has five components optimized for different uses. The UniProt Knowledgebase (UniProtKB) (1) is an expertly curated database, a central access point for integrated protein information with cross-references to multiple sources. The UniProt Archive (UniParc) (2) is a comprehensive sequence repository, reflecting the history of all protein sequences. UniProt Reference Clusters (UniRef) (3) merge closely related sequences based on sequence identity to speed up searches whereas the UniProt Metagenomic and Environmental Sequences database (UniMES) was created to respond to the expanding area of metagenomic data. UniProtKB Sequence/Annotation Version Archive (UniSave) is the UniProtKB protein entry archive, which contains all versions of each protein entry (Fig. 1).

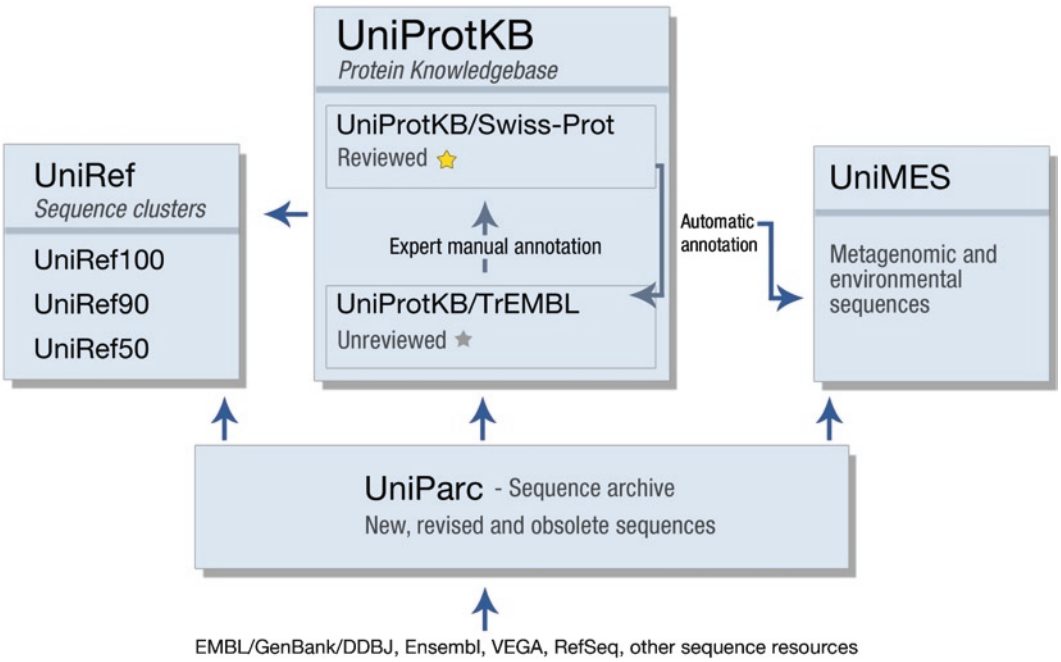


Fig. 1. UniProt databases.

2.1. The UniProt Archive

UniParc is the main sequence storehouse and is a comprehensive repository that reflects the history of all protein sequences. UniParc contains all new and revised protein sequences from all publicly available sources (http://www.uniprot.org/help/uniparc) to ensure that complete coverage is available at a single site. To avoid redundancy, all sequences 100% identical over the entire length are merged, regardless of source organism. New and updated sequences are loaded on a daily basis, cross-referenced to the source database accession number, and provided with a sequence version that increments on changes to the underlying sequence. The basic information stored within each UniParc entry is the identifier, the sequence, cyclic redundancy check number, source database(s) with accession and version numbers, and a time stamp. If a UniParc entry lacks a cross-reference to a UniProtKB entry, the reason for its exclusion from UniProtKB is provided (e.g., pseudogene). In addition, each source database accession number is tagged with its status in that database, indicating if the sequence still exists or has been deleted in the source database and cross-references to NCBI GI and TaxId if appropriate.

2.2. The UniProt Knowledgebase

UniProtKB consists of two sections, UniProtKB/Swiss-Prot and UniProtKB/TrEMBL. The former contains manually annotated records with information extracted from literature and curator-evaluated computational analysis. Annotation is done by biologists with specific expertise to achieve accuracy. In UniProtKB/Swiss-Prot, annotation consists of the description of the following: function(s), enzyme-specific information, biologically relevant domains and sites, post-translational modifications, subcellular location(s), tissue specificity, developmental specific expression, structure, interactions, splice isoform(s), associated diseases or deficiencies, or abnormalities etc. The UniProt Knowledgebase aims to describe, in a single record, all protein products derived from a certain gene from a certain species. After an inspection of the sequences, the curator selects the reference sequence, does the corresponding merging, and lists the splice and genetic variants along with disease information when available. This results in not only the whole record having an accession number but also unique identifiers for each protein form derived by alternative splicing, proteolytic cleavage, and posttranslational modification. The freely available tool VARSPLIC (4) enables the recreation of all annotated splice variants from the feature table of a UniProt Knowledgebase entry, or for the complete database. A FASTA-formatted file containing all splice variants annotated in the UniProt Knowledgebase can be downloaded for use with similarity search programs.

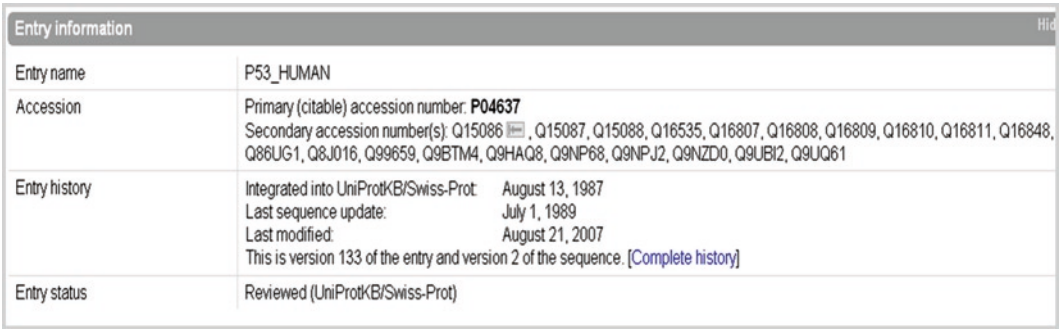
UniProtKB/TrEMBL contains high quality computationally analyzed records enriched with automatic annotation and classification. The computer-assisted annotation is created using both automatically generated rules as well as manually curated rules

(UniRule) based on protein families (5–8). UniProtKB/TrEMBL contains the translations of all coding sequences (CDS) present in the EMBL/GenBank/DDBJ Nucleotide Sequence Databases and, with some defined exclusions, *Arabidopsis thaliana* sequences from The Arabidopsis Information Resource (TAIR) (9), yeast sequences from the Saccharomyces Genome Database (SGD) (10) and *Homo sapiens* sequences from the Ensembl database (11). It will soon be extended to include other Ensembl organism sets and RefSeq records. Records are selected for full manual annotation and integration into UniProtKB/Swiss-Prot according to defined annotation priorities.

Integration between the three types of sequence-related databases (nucleic acid sequences, protein sequences, and protein tertiary structures) as well as with specialized data collections is important for the UniProt users. UniProtKB is currently cross-referenced with more than ten million links to 114 different databases with regular update cycles. This extensive network of cross-references allows UniProt to act as a focal point of biomolecular database interconnectivity. All cross-referenced databases are documented at http://www.uniprot.org/docs/dbxref and if appropriate are included in the UniProt ID mapping tool at http://www.uniprot.org/help/mapping with the file for download at ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping.

2.3. The UniProt Reference Clusters

UniRef provides clustered sets of all sequences from the UniProt Knowledgebase (including splice forms as separate entries) and selected records from the UniProt Archive to achieve complete coverage of sequence space at identity levels of 100, 90, and 50% while hiding redundant sequences (3). The UniRef clusters are generated in a hierarchical manner; the UniRef100 database combines identical sequences and sub-fragments into a single UniRef entry, UniRef90 is built from UniRef100 clusters and UniRef50 is built from UniRef90 clusters. Each individual member sequence can exist in only one UniRef cluster at each identity level and have only one parent or child cluster at another identity level. UniRef100, UniRef90, and UniRef50 yield a database size reduction of ~10, 40, and 70%, respectively. Each cluster record contains source database, protein name, and taxonomy information on each member sequence but is represented by a single selected representative protein sequence and name; the number of members and lowest common taxonomy node for the membership is also included. The representative protein sequence or cluster representative is automatically selected using an algorithm that accounts for (1) Quality of entry annotation: order of preference is a member from UniProtKB/Swiss-Prot, UniProtKB/TrEMBL, then UniParc; (2) Meaningful name: members with protein names that do not contain words such as "hypothetical" or "probable"



Entry information	Hid
Entry name	P53_HUMAN
Accession	Primary (citable) accession number: **P04637** Secondary accession number(s): Q15086, Q15087, Q15088, Q16535, Q16807, Q16808, Q16809, Q16810, Q16811, Q16848, Q86UG1, Q8J016, Q99659, Q9BTM4, Q9HAQ8, Q9NP68, Q9NPJ2, Q9NZD0, Q9UBI2, Q9UQ61
Entry history	Integrated into UniProtKB/Swiss-Prot: August 13, 1987 Last sequence update: July 1, 1989 Last modified: August 21, 2007 This is version 133 of the entry and version 2 of the sequence. [Complete history]
Entry status	Reviewed (UniProtKB/Swiss-Prot)

Fig. 2. UniSave link.

are preferred; (3) Organism: members from model organisms are preferred; (4) Sequence length: longest sequence is preferred. UniRef100 is one of the most comprehensive and nonredundant protein sequence dataset available. The reduced size of the UniRef90 and UniRef50 datasets provide faster sequence similarity searches and reduce the research bias in similarity searches by providing a more even sampling of sequence space.

2.4. The UniProt Metagenomic and Environmental Sequences

The UniProt Knowledgebase contains entries with a known taxonomic source. However, the expanding area of metagenomic data has necessitated the creation of a separate database, the UniProt Metagenomic and Environmental Sequences database (UniMES). UniMES currently contains data from the Global Ocean Sampling Expedition (GOS), which predicts nearly six million proteins, primarily from oceanic microbes. By combining the predicted protein sequences with automatic classification by InterPro, the integrated resource for protein families, domains and functional sites, UniMES uniquely provides free access to the array of genomic information gathered.

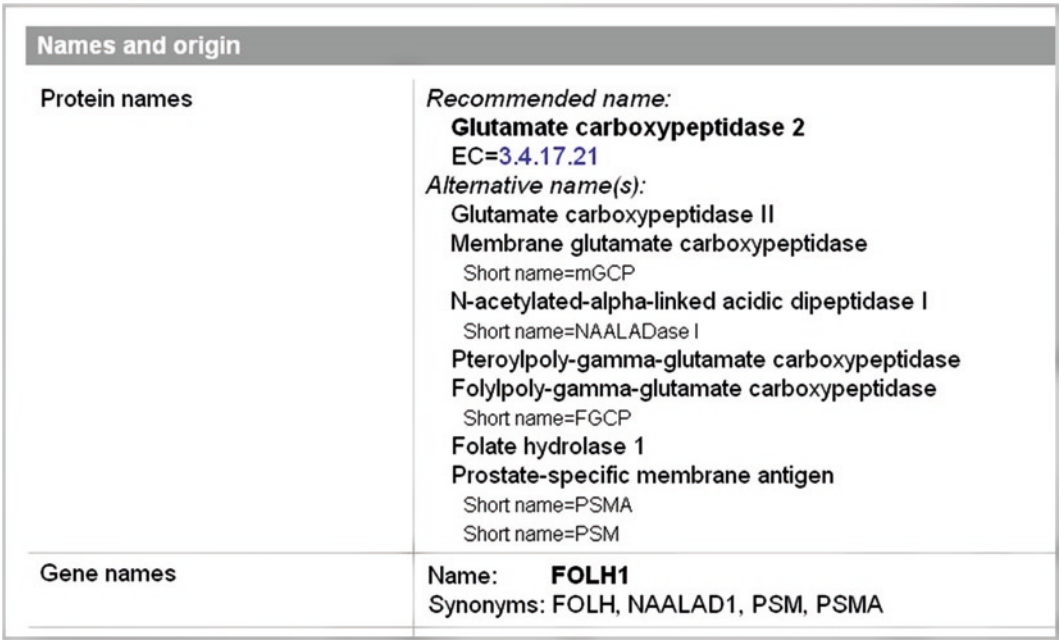
2.5. The UniProtKB Sequence/Annotation Version Archive

UniSave is a repository of UniProtKB/Swiss-Prot and UniProtKB/TrEMBL entry versions and provides the backend to the UniProtKB entry history service (Fig. 2) and is also provided as a standalone service at http://www.ebi.ac.uk/uniprot/unisave.

These descriptions of our databases should illustrate that UniProt does provide a high quality annotated nonredundant database with maximal coverage and sequence archiving.

3. Methods

This section will describe particular features of the UniProt activities, which fulfill the proteomics community requirements of detailed information on protein function, biological processes, molecular



Names and origin	
Protein names	*Recommended name:* **Glutamate carboxypeptidase 2** EC=3.4.17.21 *Alternative name(s):* Glutamate carboxypeptidase II Membrane glutamate carboxypeptidase Short name=mGCP N-acetylated-alpha-linked acidic dipeptidase I Short name=NAALADase I Pteroylpoly-gamma-glutamate carboxypeptidase Folylpoly-gamma-glutamate carboxypeptidase Short name=FGCP Folate hydrolase 1 Prostate-specific membrane antigen Short name=PSMA Short name=PSM
Gene names	Name: **FOLH1** Synonyms: FOLH, NAALAD1, PSM, PSMA

Fig. 3. UniProt nomenclature.

interactions and pathways cross-referenced to appropriate external sources and stable identifiers, consistent nomenclature and controlled vocabularies.

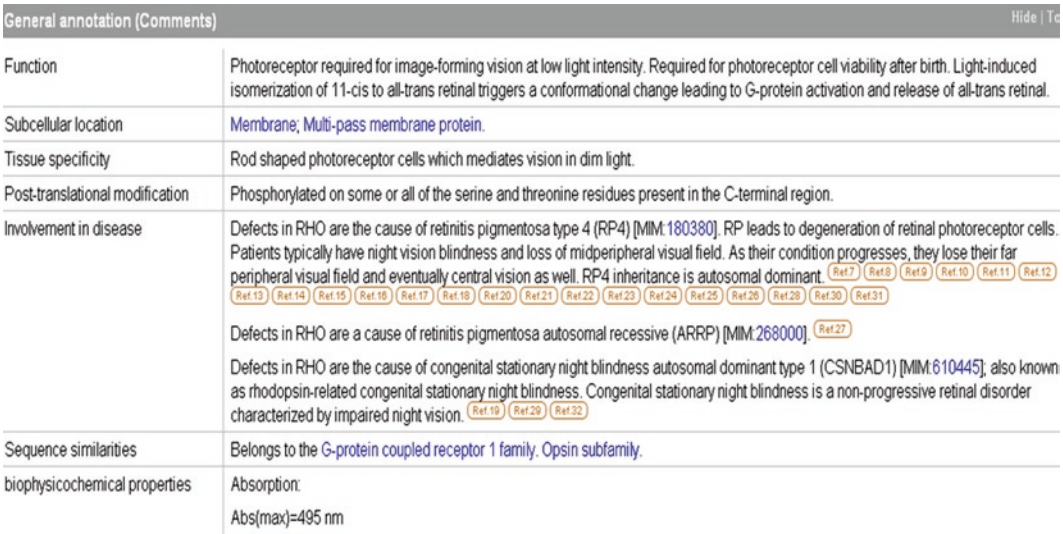
3.1. Protein Annotation

UniProtKB consists of two sections, Swiss-Prot and TrEMBL.

UniProtKB/Swiss-Prot contains manually annotated records with information extracted from literature and curator-evaluated computational analysis. Manual annotation consists of a critical review of experimentally proven or computer-predicted data about each protein. An essential aspect of the annotation protocol is the use of official nomenclatures and controlled vocabularies that facilitate consistent and accurate identification (Fig. 3).

Annotation consists of the description of the following: functions(s), enzyme-specific information, biologically relevant domains and sites, posttranslation modifications, subcellular location(s), tissue specificity, developmental specific expression, structure, interactions, splice isoforms(s), associated diseases or deficiencies, or abnormalities etc (Fig. 4).

Another important part of the annotation process involves merging of different reports for a single protein. After an inspection of the sequences the curator selects the reference sequence, does the corresponding merging and lists the splice and genetic variants along with disease information when available (Fig. 5). Data are continuously updated by an expert team of biologists.



General annotation (Comments)	Hide | To
Function	Photoreceptor required for image-forming vision at low light intensity. Required for photoreceptor cell viability after birth. Light-induced isomerization of 11-cis to all-trans retinal triggers a conformational change leading to G-protein activation and release of all-trans retinal.
Subcellular location	Membrane; Multi-pass membrane protein.
Tissue specificity	Rod shaped photoreceptor cells which mediates vision in dim light.
Post-translational modification	Phosphorylated on some or all of the serine and threonine residues present in the C-terminal region.
Involvement in disease	Defects in RHO are the cause of retinitis pigmentosa type 4 (RP4) [MIM:180380]. RP leads to degeneration of retinal photoreceptor cells. Patients typically have night vision blindness and loss of midperipheral visual field. As their condition progresses, they lose their far peripheral visual field and eventually central vision as well. RP4 inheritance is autosomal dominant. Ref.7 Ref.8 Ref.9 Ref.10 Ref.11 Ref.12 Ref.13 Ref.14 Ref.15 Ref.16 Ref.17 Ref.18 Ref.20 Ref.21 Ref.22 Ref.23 Ref.24 Ref.25 Ref.26 Ref.28 Ref.30 Ref.31 Defects in RHO are a cause of retinitis pigmentosa autosomal recessive (ARRP) [MIM:268000]. Ref.27 Defects in RHO are the cause of congenital stationary night blindness autosomal dominant type 1 (CSNBAD1) [MIM:610445]; also known as rhodopsin-related congenital stationary night blindness. Congenital stationary night blindness is a non-progressive retinal disorder characterized by impaired night vision. Ref.19 Ref.29 Ref.32
Sequence similarities	Belongs to the G-protein coupled receptor 1 family. Opsin subfamily.
biophysicochemical properties	Absorption: Abs(max)=495 nm

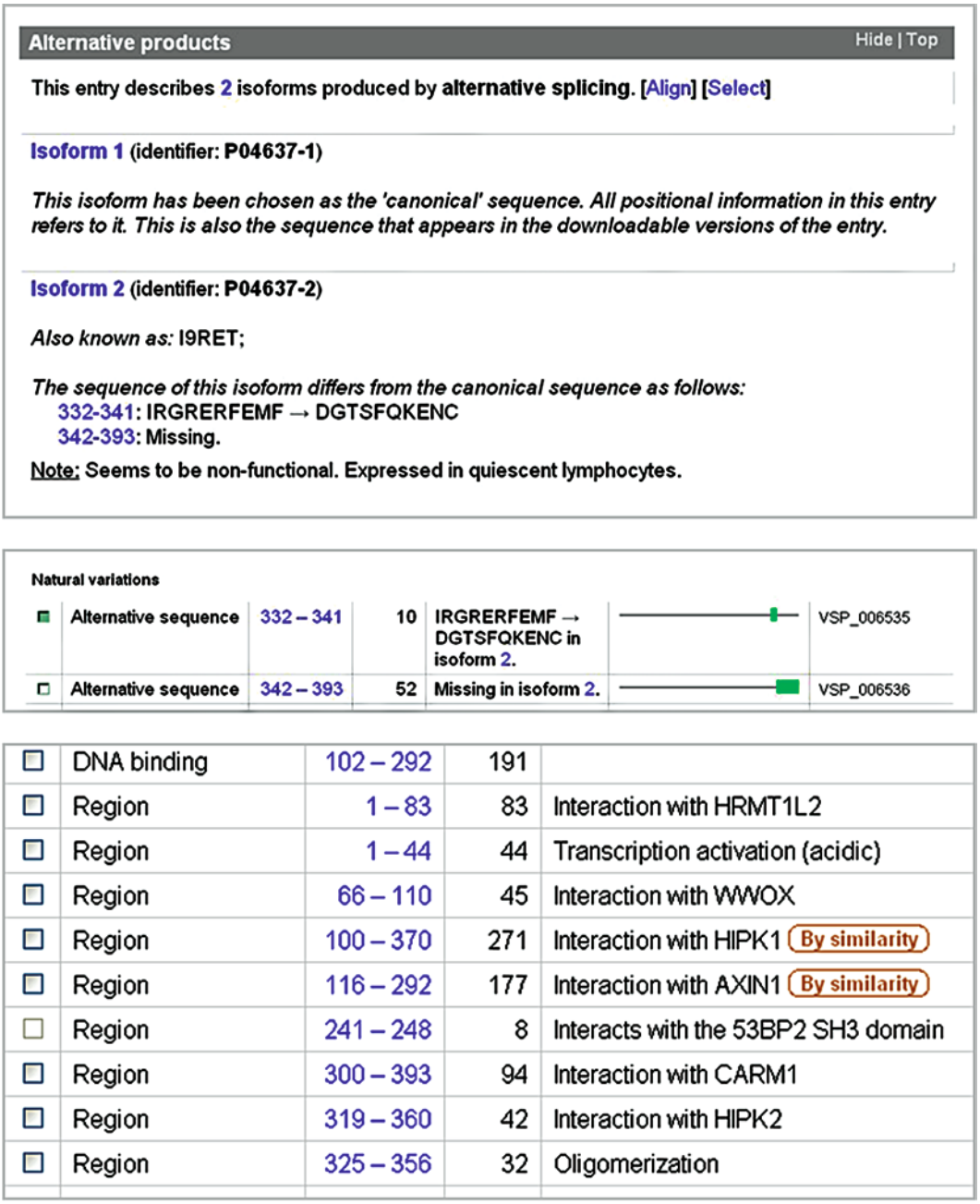
Fig. 4. Protein annotation.

3.2. The Gene Ontology Consortium and UniProt

To promote database interoperability and provide consistent annotation, the UniProt Consortium is a key member of the Gene Ontology Consortium (12) and benefits from the presence of the GO editorial office at the EBI. UniProt curators will continue to assign Gene Ontology (GO terms) to the gene products in UniProtKB during the UniProt manual curation process. UniProtKB also profits from GO annotation carried out by other GO Consortium members. Currently we include manual GO annotations from 19 GO Consortium annotation groups, and we further supplement this with high-quality annotations from other manual annotation sources (including the Human Protein Atlas and LIFEdb). In addition to this manually curated GO annotation, automatic GO annotation pipelines exist and will be further developed to ensure that the functional knowledge supplied by various UniProtKB ontologies, Ensembl orthology data, and InterPro matches are fully exploited to provide high-quality, comprehensive set of GO annotation predictions for all UniProtKB entries.

3.3. Cross-references to External Sources

One challenge in life sciences research is the ability to integrate and exchange data coming from multiple research groups. The UniProt Consortium is committed to fostering interaction and exchange with the scientific community, ensuring wide access to UniProt resources, and promoting interoperability between resources. An essential component of this interoperability is the provision of cross-references to these resources in UniProt entries (Fig. 6).



Alternative products Hide | Top

This entry describes 2 isoforms produced by **alternative splicing**. [Align] [Select]

Isoform 1 (identifier: **P04637-1**)

This isoform has been chosen as the 'canonical' sequence. All positional information in this entry refers to it. This is also the sequence that appears in the downloadable versions of the entry.

Isoform 2 (identifier: **P04637-2**)

Also known as: I9RET;

The sequence of this isoform differs from the canonical sequence as follows:
332-341: IRGRERFEMF → DGTSFQKENC
342-393: Missing.

Note: Seems to be non-functional. Expressed in quiescent lymphocytes.

Natural variations

Alternative sequence	332 – 341	10	IRGRERFEMF → DGTSFQKENC in isoform 2.		VSP_006535
Alternative sequence	342 – 393	52	Missing in isoform 2.		VSP_006536

DNA binding	102 – 292	191	
Region	1 – 83	83	Interaction with HRMT1L2
Region	1 – 44	44	Transcription activation (acidic)
Region	66 – 110	45	Interaction with WWOX
Region	100 – 370	271	Interaction with HIPK1 By similarity
Region	116 – 292	177	Interaction with AXIN1 By similarity
Region	241 – 248	8	Interacts with the 53BP2 SH3 domain
Region	300 – 393	94	Interaction with CARM1
Region	319 – 360	42	Interaction with HIPK2
Region	325 – 356	32	Oligomerization

Fig. 5. Sequence annotation.

3.4. Nonredundant Complete UniProt Proteome Sets

UniProt constructs complete nonredundant proteome sets. Each set and its analysis is made available shortly after the appearance of a new complete genome sequence in the nucleotide sequence databases. A standard procedure is used to create, from the UniProtKB, proteome sets for bacterial, archaeal and some eukaryotic genomes. Proteome sets for certain metazoan genomes are

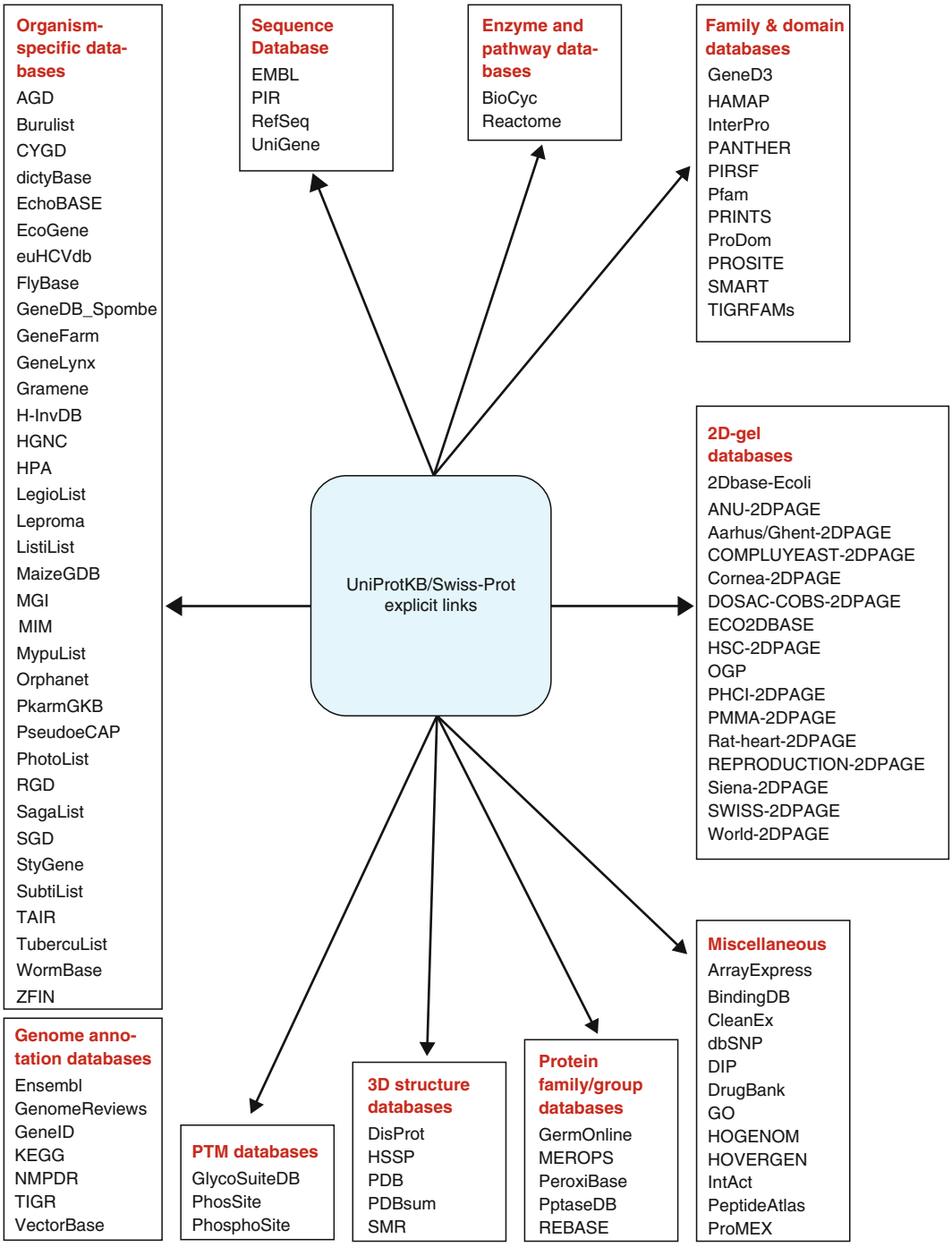


Fig. 6. UniProt cross-references.

produced by a separate procedure because a full and accurate set of coding sequence predictions are not yet available in the nucleotide sequence databases. Currently, the International Protein Index (IPI) (13) derived from the UniProt Knowledgebase, Ensembl,

and the NCBI's RefSeq project (and cross-referenced to from UniProtKB) provides this data but development is currently underway to extend the UniProtKB production pipeline to replace IPI's functionality and for UniProt to provide these sets directly in collaboration with Ensembl and RefSeq. It is envisaged that this development should be complete by mid 2010. It is a core goal for UniProt to provide meaningful annotation for these complete proteomes with a combination of our manual and automatic annotation protocols.

3.5. Using the UniProt Website

The UniProt consortium released its new improved unified website in 2009: a new interface, a new search engine, and many new options to serve its user community better. User feedback and the analysis of the use of our previous sites have led us to put more emphasis on supporting the most frequently used functionalities: database searches with simple (and sometimes less simple) queries that often consist of only a few terms have been enhanced by a good scoring system and a suggestion mechanism. Searching with ontology terms is assisted by auto-completion, and we also provide the possibility of using ontologies to browse search results. The viewing of database entries is improved with configurable views, a simplified terminology and a better integration of documentation. Medium-to-large sized result sets can now be retrieved directly on the site, so people no longer need to be referred to commercial, third party services. Access to the following most common bioinformatics tools have been simplified: sequence similarity searches, multiple sequence alignments, batch retrieval, and a database identifier mapping tool can now be launched directly from any page, and the output of these tools can be combined, filtered, and browsed like normal database searches. Programatic access to all data and results is possible via simple HTTP (REST) requests (http://www.uniprot.org/help/technical). In addition to the existing formats that support the different data sets (e.g., plain text, FASTA, and XML for UniProtKB), now it also provides (configurable) tab-delimited, RSS and GFF downloads where possible, and all data is available in RDF (http://www.w3.org/RDF/), a W3C standard for publishing data on the Semantic Web. Extensive documentation on how to best use this resource is available at: http://www.uniprot.org/help/.

References

1. The UniProt Consortium. (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.* **38**, D142–D148.
2. Leinonen R, Diez FG, Binns D, Fleischmann W, Lopez R, Apweiler R. (2004) UniProt archive. *Bioinformatics* **20**, 3236–3237.
3. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**, 1282–1288.
4. Kersey P, Hermjakob H, Apweiler R. (2000) VARSPLIC: alternatively-spliced protein

sequences derived from Swiss-Prot and TrEMBL. *Bioinformatics* **16**, 1048–1049.

5. Gattiker A, Michoud K, Rivoire C, Auchincloss AH, Coudert E, Lima T, Kersey P, Pagni M, Sigrist CJ, Lachaize C, et al. (2003) Automated annotation of microbial proteomes in SWISS-PROT. *Comput. Biol. Chem.* **27**, 49–58.
6. Fleischmann W, Moller S, Gateau A, Apweiler R. (1999) A novel method for automatic functional annotation of proteins. *Bioinformatics* **15**, 228–233.
7. Wu CH, Nikolskaya A, Huang H, Yeh L-S, Natale DA, Vinayaka CR, Hu ZZ, Mazumder R, Kumar S, Kourtesis P, et al. (2004) PIRSF: family classification system at the Protein Information Resource. *Nucleic Acids Res.* **32**, D112–D114.
8. Natale DA, Vinayaka CR, Wu CH. (2004) Large-scale, classification-driven, rule-based functional annotation of proteins. In: *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics* – Subramaniam S, ed. Bioinformatics John Wiley West Sussex, England.
9. Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, Li D, Meyer T, Muller R, Ploetz L, et al. (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.* **36**, D1009–D1014.
10. Hong EL, Balakrishnan R, Dong Q, Christie KR, Park J, Binkley G, Costanzo MC, Dwight SS, Engel SR, Fisk DG, et al. (2008) The Ontology annotations at SGD: new data sources and annotation methods. *Nucleic Acids Res.* **36**, D577–D581.
11. Flicek P, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, et al. (2008) Ensembl 2008. *Nucleic Acids Res.* **36**, D707–D714.
12. The Gene Ontology Consortium. (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29.
13. Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, Birney E, Apweiler R. (2004) The International Protein Index: an integrated database for proteomics experiments. *Proteomics* **4**, 1985–1988.

http://www.springer.com/978-1-60761-976-5

Bioinformatics for Comparative Proteomics

Wu, C.H.; Chen, C. (Eds.)

2011, XIII, 387 p., Hardcover
ISBN: 978-1-60761-976-5
A product of Humana Press