ELSEVIER

Protein sequence databases

Rolf Apweiler[1,*], Amos Bairoch[2] and Cathy H Wu[3]

A variety of protein sequence databases exist, ranging from simple sequence repositories, which store data with little or no manual intervention in the creation of the records, to expertly curated universal databases that cover all species and in which the original sequence data are enhanced by the manual addition of further information in each sequence record. As the focus of researchers moves from the genome to the proteins encoded by it, these databases will play an even more important role as central comprehensive resources of protein information. Several the leading protein sequence databases are discussed here, with special emphasis on the databases now provided by the Universal Protein Knowledgebase (UniProt) consortium.

Addresses
[1]The EMBL Outstation — The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK
[2]Swiss Institute of Bioinformatics, Centre Medical Universitaire, 1 rue Michel Servet, 1211 Geneva 4, Switzerland
[3]Department of Biochemistry and Molecular Biology, Georgetown University Medical Center, 3900 Reservoir Road, NW, Box 571414, Washington, DC 20057-1414, USA
*e-mail: apweiler@ebi.ac.uk

Current Opinion in Chemical Biology 2004, **8**:76–80

This review comes from a themed issue on
Proteomics and genomics
Edited by Michael Snyder and John Yates III

1367-5931/$ – see front matter

DOI 10.1016/j.cbpa.2003.12.004

Abbreviations

DDBJ	DNA Data Bank of Japan
EMBL	European Molecular Biology Laboratory
GO	Gene Ontology
NCBI	National Center of Biotechnology Information
NREF	non-redundant reference databases
PDB	Protein Data Bank
PIR	Protein Information Resource
PIR-PSD	Protein Information Resource Protein Sequence Database
RefSeq	Reference Sequence
TrEMBL	Translation from EMBL
UniParc	UniProt Archive

Introduction

With the availability of over 165 completed genome sequences from both eukaryotic and prokaryotic organisms, efforts are now being focused on the identification and functional analysis of the proteins encoded by these genomes. The large-scale analysis of these proteins has started to generate huge amounts of data due to the new information provided by the genome projects and to a range of new technologies in protein science. For example, mass spectrometry approaches are being used in protein identification and in determining the nature of post-translational modifications [1]. These and other methods make it possible to quickly identify large numbers of proteins, to map their interactions, to determine their location within the cell [2••] and to analyse their biological activities. Protein sequence databases play a vital role as a central resource for storing the data generated by these and more conventional efforts, and making them available to the scientific community.

To exploit the various resources fully, it is essential to distinguish between them and to identify the types of data they contain. Universal protein databases cover proteins from all species whereas specialized data collections contain information about a particular protein family or group of proteins, or related to a specific organism. Universal protein sequence databases can be further subdivided into two categories: sequence repositories, in which data are stored with little or no manual intervention in the creation of the records; and expertly curated databases, in which the original data are enhanced by the addition of further information. In the following, we present the current status of the leading protein sequence databases.

Sequence repositories

Several protein sequence databases act as repositories of protein sequences. These databases add little or no additional information to the sequence records they contain and generally make no effort to provide a non-redundant collection of sequences to users.

GenPept

The most basic example of this type of database is the GenBank Gene Products Data Bank (GenPept), produced by the National Center of Biotechnology Information (NCBI) [3]. The entries in the database are derived from translations of the sequences contained in the nucleotide database maintained collaboratively by the DNA Data Bank of Japan (DDBJ) [4], the European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database [5] and GenBank [6], and contain minimal annotation, which has been extracted primarily from the corresponding nucleotide entry. The entries lack additional annotation and the database does not contain proteins derived from amino acid sequencing. Also, each protein may be represented by multiple records and no attempt is made to group these records into a single database entry.

NCBI's Entrez Protein

NCBI's Entrez Protein (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein) is another example of a sequence repository. The database contains sequence data translated from the nucleotide sequences of the DDBJ/EMBL/GenBank database as well as sequences from Swiss-Prot [7], the Protein Information Resource (PIR) [8], RefSeq [9] and the Protein Data Bank (PDB) [10]. The database differs from GenPept in that many of the entries contain additional information that has been extracted from curated databases such as Swiss-Prot and PIR. As with GenPept, the sequence collection is redundant.

RefSeq

A more ambitious approach is taken by the Reference Sequence (RefSeq) collection produced by the NCBI (http://www.ncbi.nih.gov/RefSeq). The aim of the project is to provide a non-redundant collection of reference protein sequences. RefSeq sequences exist for a limited set of species, including approximately 1100 viruses and 150 bacteria, as well as a small number of higher organisms, such as human, mouse, rat, zebrafish, honeybee, sea urchin, cow, and several important plant species. The main features of the RefSeq collection include non-redundancy, explicitly linked nucleotide and protein sequences, updates to reflect current knowledge of sequence data and biology, data validation and format consistency, distinct accession series, and ongoing curation by NCBI staff and collaborators, with review status indicated on each record. However, the majority of the records are automatically generated with minimal manual intervention. In December 2003, the database contained 831 000 entries with approximately 44 000 manually reviewed entries so the database is closer to a sequence repository than to any of the curated databases discussed below.

Universal curated databases

Although repositories are an essential means of providing the user with sequences as quickly as possible, it is clear that, when additional information is added to a sequence, this greatly increases the value of the resource for users. The curated databases enrich the sequence data by adding additional information, which gets validated by expert biologists before being added to the databases to ensure that the data in these collections can be considered to be highly reliable. There is also a large effort invested in maintaining non-redundant datasets by compiling all reports for a given protein sequence into a single record.

PIR-PSD

The oldest universal curated protein sequence database is the Protein Information Resource Protein Sequence Database (PIR-PSD) (http://pir.georgetown.edu/). It was established in 1984 as a successor to the original National Biomedical Research Foundation Protein Sequence Database, developed over a 20 year period by the late Margaret O Dayhoff and published as the 'Atlas of Protein Sequence and Structure' from 1965 to 1978 [11].

It compiles comprehensive, non-redundant protein sequence data, organized by superfamily and family, and annotated with functional, structural, bibliographic and genetic data. In addition to the sequence data, the database contains the name and classification of the protein, the name of the organism in which it naturally occurs, references to the primary literature, function and general characteristics of the protein, and regions of biological interest within the sequence. The database is extensively cross-referenced with DDBJ/EMBL/GenBank nucleic acid and protein identifiers, PubMed and MEDLINE IDs, and unique identifiers from many other source databases. In October 2003, the database contained 273 339 annotated and classified entries, covering the entire taxonomic range and organized into 36 000 superfamilies and over 100 000 families.

Protein family classification is central to the organization and annotation of PIR. Automated procedures have allowed the placement of more than 99% sequences into families and more than 70% into superfamilies. The classification approach improves the sensitivity of protein identification, helps to detect and correct genome annotation errors systematically, and allows a more complete understanding of sequence–function–structure relationships.

Swiss-Prot

The leading universal curated protein sequence database is Swiss-Prot (http://www.ebi.ac.uk/swissprot/index.html), which contained as of November 2003 (release 42.6) 140 000 curated sequence entries from over 8300 different species. The database is non-redundant, which means that all reports for a given protein are merged into a single entry, and is highly integrated with other databases [12].

Each entry in Swiss-Prot is thoroughly analysed and annotated by biologists to ensure that the database is of a high quality. Literature-based curation is used to extract experimental data. This experimental knowledge is supplemented by manually confirmed results from various sequence analysis programs. Annotation includes the description of properties such as the function of a protein, post-translational modifications, domains and sites, secondary and quaternary structure, similarities to other proteins, diseases associated with deficiencies in a protein, developmental stages in which the protein is expressed, in which tissues the protein is found, pathways in which the protein is involved, and sequence conflicts and variants.

The annotation added is stored mainly in the description (DE) and gene (GN) lines, the comment (CC) lines, the feature table (FT) lines, and the keyword (KW) lines. There are over 480 000 comments and 739 307 sequence features in Swiss-Prot release 42.6. The addition of several qualifiers in the comment and feature table lines during the annotation process allows users to distinguish between experimentally verified data, data that has been propagated from a characterized protein based on sequence similarity, and data for which no experimental evidence currently exists [13].

To provide correct gene names, use is made of authoritative gene name sources such as Genew, the database of the Human Genome Organization (HUGO) gene nomenclature committee [14], FlyBase [15] and the Mouse Genome Database (MGD) [16].

TrEMBL

To produce a fully curated Swiss-Prot entry is a highly labor-intensive process and is the rate-limiting step in the growth of the database because more new sequences are submitted than can be efficiently annotated manually and integrated into the database. To address this, the TrEMBL (Translation from EMBL) database (http://www.ebi.ac.uk/trembl/) was introduced to make new sequences available as quickly as possible [7]. TrEMBL consists of computer-annotated entries derived from the translation of all coding sequences in the DDBJ/EMBL/GenBank nucleotide sequence database that are not yet included in Swiss-Prot. To ensure completeness, it also contains several protein sequences extracted from the literature or submitted directly by the user community. TrEMBL Release 25.6 of November 2003 contained 1 079 094 entries from more than 62 000 different species. TrEMBL follows the Swiss-Prot format and conventions described above as closely as possible.

The production of TrEMBL starts with the translation of coding sequences in the DDBJ/EMBL/GenBank nucleotide sequence database. At this stage, all annotation in a TrEMBL entry derives from the corresponding nucleotide entry. The next steps involve redundancy removal through merging of multiple records [17] and the automated enhancement of the information content in TrEMBL [18]. The process is based on a system of standardized transfer of annotation from well-characterized proteins in Swiss-Prot to unannotated TrEMBL entries belonging to defined groups [19]. To assign entries to these groups, InterPro [20], an integrated resource of protein families, domains and functional sites, is used. This amalgamates the efforts of the member databases, which are currently PROSITE [21], PRINTS [22], Pfam [23], ProDom [24] SMART [25], TIGRFAMS [26], PIR SuperFamilies [27] and SUPERFAMILY [28]. This process of adding accurate, high-quality information to TrEMBL entries brings the standard of annotation in TrEMBL closer to that found in Swiss-Prot.

UniProt: the next generation of protein sequence databases

One of the most significant developments with regard to protein sequence databases is the recent decision by the National Institutes of Health to award a grant [29] to combine the Swiss-Prot, TrEMBL and PIR-PSD databases into a single resource, UniProt (http://www.uniprot.org) [30••]. UniProt was launched on 15 December 2003 and comprises three components: first, the UniProt Knowledgebase which will continue the work of Swiss-Prot, TrEMBL and PIR by providing an expertly curated database; second, the UniProt Archive (UniParc) into which new and updated sequences are loaded on a daily basis; third, the UniProt non-redundant reference databases (UniProt NREF), which provide non-redundant views on top of the UniProt Knowledgebase and UniParc.

The UniProt Archive (UniParc)

UniParc is the most comprehensive publicly accessible non-redundant protein sequence collection available. It contains publicly available protein sequences from Swiss-Prot, TrEMBL, PIR-PSD, EMBL, Ensembl [31], International Protein Index (IPI) (http://www.ebi.ac.uk/IPI), PDB, RefSeq, FlyBase, WormBase [32] and the patent offices in Europe, the United States and Japan, making it the most comprehensive protein sequence database available. While a protein sequence may exist in multiple databases and more than once in a given database, UniParc stores every unique sequence only once and assigns a unique UniParc identifier. Furthermore, UniParc provides cross-references to the source databases (accession numbers), sequence versions, and status (active or obsolete). A UniParc sequence version is also provided, and incremented each time the underlying sequence changes, thus making it possible to observe sequence changes in all source databases.

The UniProt knowledgebase (UNIPROT)

Swiss-Prot, TrEMBL and PIR-PSD have been merged to form the UniProt knowledgebase. All suitable PIR-PSD sequences that are missing from Swiss-Prot + TrEMBL were incorporated into UniProt. Bi-directional cross-references between Swiss-Prot + TrEMBL and PIR-PSD were created to allow the easy tracking of the PIR-PSD entries. The transfer into UniProt of references and experimentally verified data present in PIR but missing from Swiss-Prot + TrEMBL is ongoing.

The UniProt knowledgebase consists of two parts: a section containing fully manually annotated records resulting from literature information extraction and curator-evaluated computational analysis, and a section with computationally analysed records awaiting full manual

annotation. For the sake of continuity and name recognition, the two sections are referred to as 'Swiss-Prot' and 'TrEMBL'.

The main principles of the UniProt knowledgebase follow the established procedures used to annotate Swiss-Prot, TrEMBL and PIR, and this has already been explained in some detail. However, some additional advances have been made during the creation of UniProt.

Gene Ontology annotation

To provide the high level of annotation described above, the UniProt curators read a large amount of scientific literature related to each protein. This enables them to contribute to the work of the Gene Ontology (GO) consortium [33] by assigning GO terms during the annotation process as they extract information related to each of the gene ontologies (i.e. the function of a protein, what processes it is involved in and where in the cell it is located) [34•].

Integration with other databases

UniProt provides cross-references to external data collections such as the underlying DNA sequence entries in the DDBJ/EMBL/GenBank nucleotide sequence databases, 2D PAGE and 3D protein structure databases, various protein domain and family characterization databases, post-translational modification databases, species-specific data collections, variant databases and disease databases. As a result of this, UniProt acts as a central hub for biomolecular information archived in more than 50 cross-referenced databases.

Isoform and feature identifiers

Unique and stable feature identifiers (FTId) allow reference to a position-specific annotation item in the feature table. Currently, these are systematically attributed to FT VARIANT lines of human sequence entries, to alternative splicing events (VARSPLIC), and to certain glycosylation sites (CARBOHYD), but will ultimately be assigned to all types of FT lines.

Isoform identifiers have been introduced for splice isoforms, which may differ considerably from one another, with potentially less than 50% sequence similarity between isoforms. The tool VARSPLIC [35], which is freely available, enables the recreation of all annotated splice variants from the feature table of a UniProt entry, or for the complete database. A FASTA-formatted file containing all splice variants annotated in UniProt can be downloaded for use with similarity search programs.

The UniProt NREF (UniRef) databases

The UniProt NREF (UniRef) databases, NREF100, NREF90 and NREF50, provide a complete coverage of sequence space while hiding redundant sequences from view.

NREF100 provides a comprehensive non-redundant sequence collection clustered by sequence identity and taxonomy with source attribution. Identical sequences and subfragments from the same source organism (species) are presented as a single NREF entry with accession numbers of all the merged UniProt entries, the protein sequence, taxonomy, bibliography, links to the corresponding UniProt knowledgebase and archive records, as well as close sequence neighbors (with at least 95% sequence identity) from the same source organism.

NREF90 and NREF50 are built from NREF100 to provide non-redundant sequence collections for the scientific user community to perform faster homology searches. All records from all source organisms with mutual sequence identity of >90% or >50%, respectively, are merged into a single record that links to the corresponding UniProt knowledgebase records. NREF90 and NREF50 yield a size reduction of approximately 40% and 65%, respectively.

Conclusions

Complete and up-to-date databases of biological knowledge are vital for information-dependent biological and biotechnological research. With the rapid accumulation of genome sequences for many organisms, attention is turning to the identification and function of proteins encoded by these genomes. The recent joining of forces by the major protein databases Swiss-Prot, TrEMBL and PIR in the UniProt consortium to handle the increasing volume and variety of protein sequences and functional information will provide a cornerstone for scientists active in modern biological research.

Acknowledgements

UniProt is mainly supported by the National Institutes of Health (NIH) grant 1 U01 HG02712-01. Minor support for the EBI's involvement in UniProt comes from the two European Union contracts BioBabel (QLRT-2000-00981) and TEMBLOR (QLRI-2001-00015) and from the NIH grant 1R01HGO2273-01. Swiss-Prot activities at the SIB are supported by the Swiss Federal Government through the Federal Office of Education and Science. PIR activities are also supported by the National Science Foundation (NSF) grants DBI-0138188 and ITR-0205470.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Sickmann A, Mreyen M, Meyer HE: **Mass spectrometry – a key technology in proteome research**. *Adv Biochem Eng Biotechnol* 2003, **83**:141-176.

2. •• Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK: **Global analysis of protein expression in yeast**. *Nature* 2003, **425**:686-691.

This article describes the construction and analysis of yeast strains expressing full-length, chromosomally tagged green fluorescent protein fusion proteins, representing 75% of the yeast proteome. The authors provide localization information for 70% of previously unlocalized proteins. Analysis of this dataset will help to define the functions of proteins in the context of the cellular environment.

3. Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Tatusova TA, Wagner L: **Database resources of the National Center for Biotechnology**. *Nucleic Acids Res* 2003, **31**:28-33.

4. Miyazaki S, Sugawara H, Gojobori T, Tateno Y: **DNA Data Bank of Japan in XML**. *Nucleic Acids Res* 2003, **31**:13-16.

5. Stoesser G, Baker W, van den Broek A, Garcia-Pastor M, Kanz C, Kulikova T, Leinonen R, Lin Q, Lombard V, Lopez R *et al.*: **The EMBL Nucleotide Sequence Database: major new developments**. *Nucleic Acids Res* 2003, **31**:17-22.

6. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank**. *Nucleic Acids Res* 2003, **31**:23-27.

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

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

9. Pruitt KD, Tatusova T, Maglott DR: **NCBI Reference Sequence Project: update and current status**. *Nucleic Acids Res* 2003, **31**:34-37.

10. Westbrook J, Feng Z, Chen L, Yang H, Berman HM: **The Protein Data Bank and structural genomics**. *Nucleic Acids Res* 2003, **31**:489-491.

11. Dayhoff MO: *Atlas of Protein Sequence and Structure*, vol 5, suppl. 3. Washington, DC: National Biomedical Research Foundation; 1978.

12. Gasteiger E, Jung E, Bairoch A: **SWISS-PROT: connecting biomolecular knowledge via a protein database**. *Curr Issues Mol Biol* 2001, **3**:47-55.

13. Junker V, Apweiler R, Bairoch A: **Representation of functional information in the Swiss-Prot data bank**. *Bioinformatics* 1999, **15**:1066-1067.

14. Wain HM, Lush M, Ducluzeau F, Povey S: **Genew: the human gene nomenclature database**. *Nucleic Acids Res* 2002, **30**:169-171.

15. FlyBase consortium: **The FlyBase database of the *Drosophila* genome projects and community literature.** *Nucleic Acids Res* 2003, **31**:172-175.

16. Blake JA, Richardson JE, Bult CJ, Kadin JA, Eppig JT: **MGD: the Mouse Genome Database**. *Nucleic Acids Res* 2003, **31**:193-195.

17. O'Donovan C, Martin MJ, Glemet E, Codani J, Apweiler R: **Removing redundancy in Swiss-Prot and TrEMBL**. *Bioinformatics* 1999, **15**:258-259.

18. Apweiler R: **Functional information in Swiss-Prot: the basis for large-scale characterisation of protein sequences**. *Brief Bioinform* 2001, **2**:9-18.

19. Fleischmann W, Moeller S, Gateau A, Apweiler R: **A novel method for automatic and reliable functional annotation**. *Bioinformatics* 1999, **15**:228-233.

20. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P *et al.*: **The InterPro database, 2003 brings increased coverage and new features**. *Nucleic Acids Res* 2003, **31**:315-318.

21. Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJA, Hofmann K, Bairoch A: **The PROSITE database, its status in 2002**. *Nucleic Acids Res* 2002, **30**:235-238.

22. Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, Moulton G, Nordle A, Paine K, Taylor P *et al.*: **PRINTS and its automatic supplement, prePRINTS**. *Nucleic Acids Res* 2003, **31**:400-402.

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

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

25. Letunic I, Goodstadt L, Dickens NJ, Doerks T, Schultz J, Mott R, Ciccarelli F, Copley RR, Ponting CP, Bork P: **Recent improvements to the SMART domain-based sequence annotation resource**. *Nucleic Acids Res* 2002, **30**:242-244.

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

27. Huang H, Barker WC, Chen Y, Wu CH: **iProClass: an integrated database of protein family, function and structure information**. *Nucleic Acids Res* 2003, **31**:390-392.

28. Gough J, Karplus K, Hughey R, Chothia C: **Assignment of homology to genome sequences using a library of Hidden Markov models that represent all proteins of known structure**. *J Mol Biol* 2001, **313**:903-919.

29. Butler D: **NIH pledges cash for global protein database**. *Nature* 2002, **419**:101.

30. •• Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro F, Gasteiger E, Huang H, Lopez R, Magrane M *et al.*: **UniProt: the Universal Protein Knowledgebase.** *Nucleic Acids Res* 2004, **32**:in press.

To provide the scientific community with a single, centralized, authoritative resource for protein sequences and functional information, the Swiss-Prot, TrEMBL and PIR protein database activities have united to form the Universal Protein Knowledgebase (UniProt) consortium. This paper describes the broad goals of the UniProt Consortium, and gives the basic practical information about UniParc, UniProt and UniRef, the three databases components provided by the consortium.

31. Clamp M, Andrews D, Barker D, Bevan P, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V *et al.*: **Ensembl 2002: accommodating comparative genomics**. *Nucleic Acids Res* 2003, **31**:38-42.

32. Harris TW, Lee R, Schwarz E, Bradnam K, Lawson D, Chen W, Blasier D, Kenny E, Cunningham F, Kishore R *et al.*: **WormBase: a cross-species database for comparative genomics**. *Nucleic Acids Res* 2003, **31**:133-137.

33. The Gene Ontology Consortium: **Creating the gene ontology resource: design and implementation.** *Genome Res* 2001, **11**:1425-1433.

34. • Camon E, Magrane M, Barrell D, Binns D, Fleischmann W, Kersey P, Mulder N, Oinn T, Maslen J, Cox A, Apweiler R: **The Gene Ontology Annotation (GOA) project: implementation of GO in Swiss-Prot, TrEMBL and InterPro**. *Genome Res* 2003, **13**:662-672.

Gene Ontology Annotation (GOA) is a project that aims to provide assignments of GO terms to gene products in Swiss-Prot and TrEMBL. This paper explains in detail the manual and automatic annotation procedures used to assign millions of GO terms to hundreds of thousands of proteins.

35. Kersey P, Hermjakob H, Apweiler R: **VARSPLIC: alternatively-spliced protein sequences derived from Swiss-Prot and TrEMBL**. *Bioinformatics* 2000, **16**:1048-1049.