

Protein Sequence Database Resources

The availability of completed genome sequences of several eukaryotic and prokaryotic species has shifted the focus of scientists from genomics towards the identification and characterization of the gene products, the proteins. Proteomics represents a milestone in biological research as proteins return to the central stage. However, researchers face a huge challenge in determining the structure and function of proteins – a task that is considerably more daunting than the sequencing and mapping of genes. Protein sequence databases are the most comprehensive source of information on proteins available to scientists so it is timely to review their current status.



Rolf Apweiler



Claire O'Donovan, Maria Jesus Martin

Protein Sequence Databases

There are a number of different protein sequence databases available with different aims. It is important to distinguish between universal databases covering proteins from all species and specialised data collections storing information about specific families or groups of proteins, or about proteins of a specific organism. Furthermore, two categories of universal protein sequence databases can be discerned: expertly manually curated sequence databases and sequence repositories.

The expertly manually curated sequence databases have faced a serious challenge in responding to the protein data explosion generated by the genome sequencing efforts.

Swiss-Prot is a protein knowledgebase established in 1986 and maintained collaboratively by the Swiss Institute of Bioinformatics (SIB) and the European Bioinformatics Institute (EBI). It strives to provide a high level of annotation, a minimal level of redundancy, a high level of integration with other biomolecular databases, and extensive external documentation. Each entry in Swiss-Prot is thoroughly analysed and annotated by biologists to ensure a high standard of annotation and to maintain the quality of the database. In June 2003, the database release 41.12 contained 128,586 annotated sequences entries from more than 8'000 different species.

The TrEMBL protein sequence database was created in 1996 as a complement to Swiss-Prot in response to the need to

make new sequences available as quickly as possible. This was not achievable through integration into Swiss-Prot as maintaining the quality of Swiss-Prot is a time-consuming process that involves extensive sequence analysis and detailed curation by expert annotators. TrEMBL (Translation of EMBL nucleotide sequence database) initially consisted of computer annotated entries derived from the translation of all coding sequences (CDS) in the DDBJ/EMBL-Bank/GenBank nucleotide sequence database, except for those already included in Swiss-Prot. It now additionally contains protein sequences that are extracted from the literature or submitted to Swiss-Prot. TrEMBL has grown from 105'288 entries in release 1 in November 1996 to over a million entries in release 24 in June 2004, thoroughly justifying its creation.



The other universal protein sequence database is the Protein Information Resource. PIR is a joint effort between Georgetown University Medical Centre and the National Biomedical Research Foundation in Washington, D.C. It was established in 1984 and resulted from the work of Dr. Margaret Dayhoff. Her *Atlas of Protein Sequence and Structure*, published from 1965–1978, was the first comprehensive collection of protein sequences. In 1974, Dr. Dayhoff devised the concept of the protein family and superfamily, defined by sequence similarity, as a means of organising and classifying proteins. In recent years, this concept has been exploited by the PIR Protein Sequence Database (PIR-PSD) to enable them to computer-annotate their entries with functional and structural data. This has facilitated an increase in the number

of sequences in the database. In May 2003, it contained over 283,000 sequences organised into 36,000 superfamilies and over 100,000 families. Two other protein databases are provided by PIR in response to the flood of protein sequence data: the IProClass integrated resource of family relationships and structural and functional features of proteins (over 320,000 non-redundant PIR and Swiss-Prot proteins with links to over 45 biological databases) and the comprehensive PIR-NREF database of 1,235,044 non-redundant protein sequences from PIR-PSD, Swiss-Prot, TrEMBL, RefSeq, GenPept, and PDB.

While Swiss-Prot and PIR-PSD are examples of expertly manually curated protein sequence database, TrEMBL and its American counterparts GenPept and RefSeq exemplify sequence repositories. The most basic example is the GenBank Gene Products Data Bank (GenPept) produced by the National Center of Biotechnology Information (NCBI). The entries are derived from GenBank nucleotide sequence databank entries and contain minimal annotation, primarily extracted from the corresponding GenBank entries. Release 135.0 from April 2003 contains 1,387,534 entries.

A more ambitious approach is taken by the Reference Sequence (RefSeq) collection, which is also produced by the NCBI but only for selected organisms. NCBI provides RefSeqs for over 1000 viruses and 100 bacteria and is in the process of producing collections for numerous higher organisms, such as human, mouse, rat, zebrafish, honeybee, sea urchin, cow and several important plant species. The aims 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 most of the entries are automatically generated without any manual intervention or annotation so this database should still be viewed mainly as

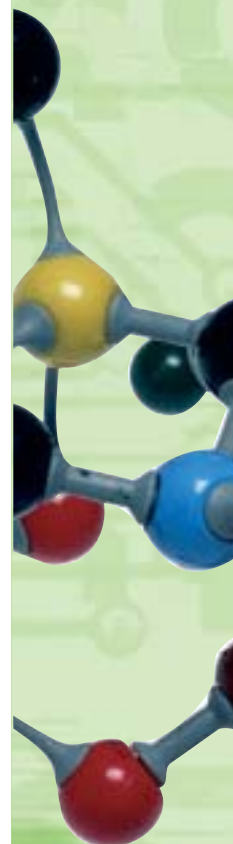
a sequence repository. In June 2003, RefSeq contained 39,130 entries with 2017 manually reviewed entries. The TrEMBL database also aspires to provide more than just the sequences to the user. It has limited redundancy, cross-references to over 45 biomolecular databases and a system of automatic annotation. This novel system involves standardised transfer of annotation from well-characterized proteins in Swiss-Prot to non-annotated TrEMBL entries. RuleBase manages and stores more than 600 annotation rules, which are applied to defined protein groups in TrEMBL. To assign TrEMBL entries into protein groups, the highly diagnostic protein family signature database InterPro is used. This system has been used to improve the annotation in 32% of all TrEMBL entries.

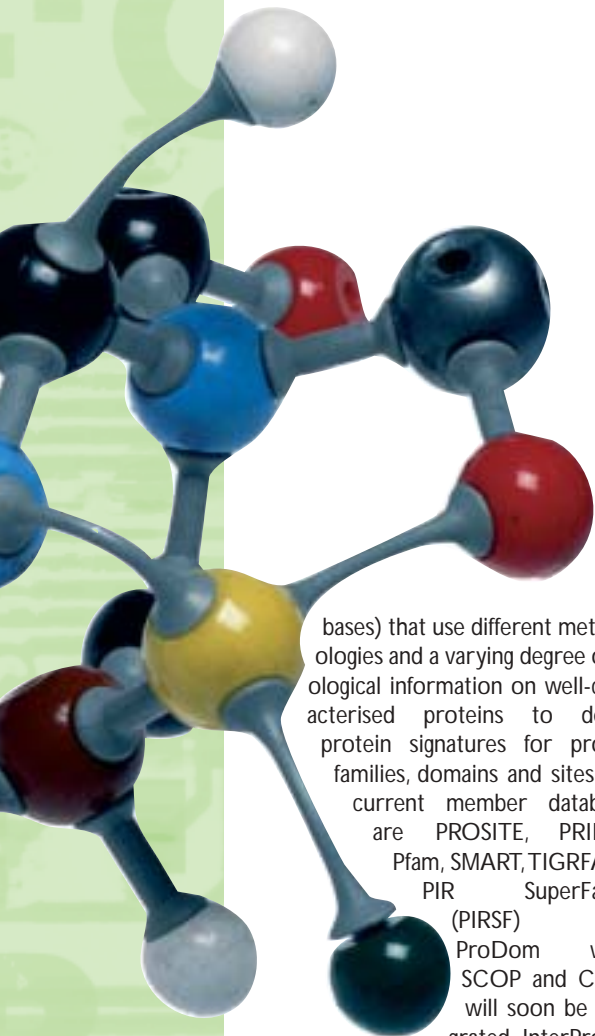
All of this demonstrates that although sequence depositories are essential to provide the sequences to the user as quickly as possible, it is clearly recognised that as much relevant information as possible should be attached to those sequences as soon as possible.

Specialised databases

In addition to the universal protein sequence databases, there are numerous specialised databases available to the life science community. Some are devoted to one particular aspect of proteins while others seek to consolidate and exploit already existing resources to their full potential. One example of the former type of database is PDB. Over the history of the Protein Data Bank (PDB) this archive of three-dimensional structural data has grown from 7 files in 1971 to a database containing over 21,390 structures as of June 2003. The archive's growth has been accompanied by increases in both data content and the structural complexity of individual entries. A further acceleration is expected due to developments in high-throughput structural determination methodologies and worldwide structural genomics efforts with an estimated tripling or quadrupling in size over the next five years. This has led to PDB completely overhauling their submission and browsing facilities in order to be able to respond appropriately.

On the other hand, InterPro and SPTP are examples of integrated protein resources. InterPro combines a number of databases (referred to as member data-





bases) that use different methodologies and a varying degree of biological information on well-characterised proteins to derive protein signatures for protein families, domains and sites. The current member databases are PROSITE, PRINTS, Pfam, SMART, TIGRFAMS, PIR SuperFamily (PIRSF) and ProDom while SCOP and CATH will soon be integrated. InterPro release 6.2 from June 2003 contains 8423 entries, representing 1815 domains, 6383 families, 160 repeats and 45 sites. By uniting the member databases, InterPro capitalises on their individual strengths, producing a powerful integrated diagnostic tool. The creation of TrEMBL enabled the EBI to develop SP_TR_NRDB (or abbreviated SPTR or SWALL) at the end of 1997 to provide a

comprehensive, non-redundant, well-annotated and up-to-date single protein sequence database as frequently requested by the user community. The components of SPTR are the weekly updated Swiss-Prot workrelease, the weekly updated TrEMBL workrelease and TrEMBLnew, the weekly updated new data to be incorporated in TrEMBL at release time.

Discussion

All the protein sequence databases are in a period of change as they seek to respond to the flood of protein data that is now available and increasing exponentially in both volume and complexity. These databases will play an essential role in the development and exploitation of the proteomics era but only if they are able to provide the data in a timely and informative manner to the user community. The different databases are responding in different ways to this challenge but one of the most exciting is the recent announcement that the National Institutes of Health have awarded a three-year, \$15-million grant for the creation of the United Protein Database or UniProt. UniProt will combine the resources of Swiss-Prot, TrEMBL and PIR to create the three-layer approach of the UniProt protein knowledgebase, UniProt Archive and UniProt non-redundant reference databases. The commitment of PIR-PSD, Swiss-Prot and TrEMBL to expertly curated annotation will continue with the UniProt protein knowledgebase while the UniProt Archive (UniParc) will be the most comprehensive non-redundant protein sequence database available.

New and updated protein sequences will be loaded daily into UniParc from the databases Swiss-Prot, TrEMBL, PIR-PSD, EMBL, Ensembl, IPI, PDB, RefSeq, FlyBase, WormBase, European Patent Office, United States Patent and Trademark Office and Japan Patent Office. As a result, performing a sequence search against UniParc will be equivalent to performing the same search against all databases crossreferenced by UniParc providing a major advantage for the user. The UniProt NREF (Non-redundant REFERENCE) databases will be created to achieve non-redundancy to facilitate sequence merging in the UniProt knowledgebase and to allow faster and more informative sequence similarity searches. With the increasing volume and variety of protein sequences and functional information, UniProt, as the central database of protein sequence, will function as a cornerstone for a wide range of scientists active in modern biological research, especially in the field of proteomics. A new website will be created for UniProt and the web address will be <http://www.uniprot.org>.

Dr Rolf Apweiler
(Head of Sequence Database Group)
Claire O'Donovan
Maria Jesus Martin
EMBL Outstation – The European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton, Cambridge, CB10 1SD, UK.
Phone +44 1223 494435
Fax + 44 1223 494468
apweiler@ebi.ac.uk
odonovan@ebi.ac.uk
martin@ebi.ac.uk