EUROPEAN JOURNAL OF BIOMEDICAL AND PHARMACEUTICAL SCIENCES
http://www.ejbps.com

ISSN 2349-8870
Volume: 6
Issue: 4
120-128
Year: 2019

PROTEIN SEQUENCE DATA BANKS

Shireen Begum*[1] and Amtul Ameena[2]

[1]Assistant Professor, Deccan School of Pharmacy,
[2]Student, Deccan School of Pharmacy (Affiliated by OU),
Department of Pharmaceutics, Deccan School of Pharmacy, Dar-us-Salam, Aghapura, Hyderabad-500 001, Telangana India.

***Corresponding Author: Shireen Begum**
Department of Pharmaceutics, Deccan School of Pharmacy, Dar-us-Salam, Aghapura, Hyderabad-500 001, Telangana India.

Article Received on 09/02/2019 **Article Revised on 01/03/2019** **Article Accepted on 23/03/2019**

ABSTRACT
A variety of protein sequence databases exit, ranging from simple sequence repositories which store data with little or no manual intervention in the creation of the records, to expertly curate universal databases that cover all the species and in which the original sequence data are enhanced by the manual addition of further information in each sequence records. As the focus of the researches moves from the genome to the 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.[1]

KEYWORDS: Bioinformatics, Biological databases, Protein analysis, Protein modeling.

INTRODUCTION

DATABASE: a collection of information organized in such a way that a computer program can quickly select desired pieces of data. You can think data base as a electronic filing system.

To access information from a database, need a database management system (DBMS). This is a collection of a programs that enables you to enter organize, and select data in a data base.

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.[2]
[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] To analyze 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.

In the following, we present the current status of the leading protein sequence databases.

OBJECTIVES

Increasingly the term database is used as shorthand for a database management system.

1. **Representation of a data:** a database is used to abstract very specific sorts of information about and organize it in a way that will prove useful.
2. **Organizing the data:** it is important to realize that data can be filed away in several different forms depending on how it needs to be used and accessed. Perhaps the simplest method is flat file or spreadsheet
3. **Flat files and spread sheaths:** flat file or spread sheath is a simple method of storing data. All records in this database have the same number of fields.
4. **Hierarchal files:** hierarchical files store in more than one type of record. This method is usually described as a "parent child, one too many ", relationship.
5. **Relationship files:** relationship files connect different files or tables without using internal pointers or keys
6. **Object oriented databases:**. It is these attributes that are stored in the database. Object oriented database has the advantage of organizing information in ways that scores often find easier to use the database

7. **Importance of the biological databases:** Internet has radically affected the way data are provided, handled and analyzed. This powerful combination of data and tools which allow easy access and analysis has changed and will continue to change our approach to the design and practice of biological research.[3]

PROTEIN SEQUENCE DATABASES

Introduction

Protein sequences (also known as polypeptides) are organic compounds made of amino acids arranged in a linear chain and folded into a globular form. The amino acids in a polymer chain are joined together by the peptide bonds between the carboxyl and amino groups of adjacent amino acid residues. The sequence of amino acids in a protein is defined by the sequence of a gene, which is encoded in the genetic code.

Protein sequence classification plays an important role in protein sequence analysis.

The protein sequence databases are the most comprehensive source of information on protein.

In the following sections you will find a short description of protein information resource (PIR) the oldest protein sequence database, and more detailed description of SWISS-PROT annotated universal database, and of TrEMBLE, the supplement of SWISS-PROT which can be classified as computer annotated sequence repository.[4]

(i) SWISS-PROT PROTEIN SEQUENCE DATABASE

Introduction

SWISS-PROT (1) is an annotated protein sequence database established in 1986 and maintained collaboratively, since 1987, by the Department of Medical Biochemistry of the University of Geneva and the EMBL Data Library (now the EMBL Outstation-The European Bioinformatics Institute;[5]
2). The SWISS-PROT protein sequence data bank consists of sequence entries. Sequence entries are composed of different line types, each with their own format. For standardization purposes the format of SWISS-PROT
(3) Follows as closely as possible that of the EMBL nucleotide sequence database.

The SWISS-PROT database distinguishes itself from other protein sequence databases by three distinct criteria.

- **Annotation:** while the annotation consists of a description of the following items: (i) functions of the protein; (ii) post-translational modifications, for example carbohydrates, phosphorylation, acetylation, (iii) domains and sites, for example calcium binding regions, ATP binding sites, zinc fingers, (iv) secondary structure; (v) quaternary structure; (vi) similarities to other proteins; (vii) diseases associated with deficiency of the protein; (viii) sequence conflicts, variants, etc.[6]
- **Minimum redundancy:** we try as much as possible to merge all these data, so as to minimize the redundancy of the database. If conflicts exist between various sequencing reports they are indicated in the feature table of the corresponding entry.
- **Integration with other databases:** It is important to provide the users of biomolecular databases with a degree of 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.

(ii) PROTEIN INFORMATION RESOURCE

The protein information resource (PIR) in collaboration with Munich information center for protein sequence (MIPS) and japan information data base (JIPID) produces the PIR international protein sequence database (PIR-PSD).

The Protein Information Resource (PIR) has been providing the scientific community with databases and tools for the organization and analysis of protein sequence data (1,2). Together with MIPS and JIPID, we have undertaken a major restructuring to meet the challenges presented by the rapid growth of largely uncharacterized sequence data.

(III) THE PIR-INTERNATIONAL PROTEIN DATABASES

PIR, MIPS and JIPID constitute the PIR-International consortium that maintains the PIR-International Protein Sequence Database (PSD), the largest publicly distributed and freely available protein sequence database. The database has the following distinguishing features.

• It is a comprehensive, annotated, and non-redundant protein sequence database, containing over 142 000 sequences as of September 1999
• The PIR is the only sequence database to provide context cross-references between its own database entries. These cross-references assist searchers in exploring relationships such as subunit associations in molecular complexes, enzyme–substrate interactions, activation and regulation cascades, as well as in browsing entries with shared features and annotations.[7]

Superfamily and family classification

PIR-International has maintained the highest classification rate and provided the most comprehensive classification and alignments of proteins among all major public domain databases.

To deal efficiently with the many new sequences from genome sequencing projects, procedures for family and superfamily classification have been automated.

. There are currently >76 000 sequences in >8900 superfamilies, and 30 000 entries with 370 recognized homology domains in the PSD. Corresponding to the classification are 1500 superfamily, 2100 family and 400 domain alignments in the PIR-ALN database, and 15 000 family and 4500 superfamily alignments in the MIPS ProtFam database.[8]

THE PIR SEARCH AND ANALYSIS SYSTEM

The PIR search and analysis system provides search engines of three types (Table (Table2):2):

(i) interactive text-based search engines, which allow Boolean queries of text fields;
(ii) standard sequence similarity search engines, including Peptide Match, Pattern Match, BLAST, FASTA, Pairwise Alignment and Multiple Alignment
(iii) Advanced search engines that combine sequence similarity and annotation searches or evaluate gene family relationships, including Annotation-Sorted Similarity Search, Domain Search, Global and Domain Similarity Search and Gene FIND.

(IV) INTERPRO PROTEIN SEQUENCE DATABASE

Interpro is an integrated documentation resource for protein families: Domain and sites, developed initially as a mean of rationizing the complementary efforts of the PROSITE, PRINTS, PFAM and PRODOM and links are made back to the relevant member database, Databases with signatures diagnostic for protein families, domains or functional sites are important tools for the computational functional classification of newly determined sequences currently, the most commonly used signature and sequence cluster databases include PROSITE (1); Pfam (2); PRINTS (3); ProDom (4); and Blocks (5). Diagnostically, these resources have different areas of optimum application owing to the different strengths and weaknesses of their underlying analysis methods.

NEW FEATURES IN INTERPRO

Annotation

InterPro curators continue to integrate new signatures from member databases into entries. The entries are classified according to the type of signature they group together. Previously, the categories comprised family, domain, repeat, post-translational modification (PTM), active site and binding site. A new type has recently been introduced called ‘conserved.[9]

Protein matches and XML files

Matches of InterPro signatures to UniProtKB, UniParc and UniMES databases are continuously calculated. Each unique protein sequence is stored only once in UniParc and so, to minimize calculation overhead, searches are run cumulatively; only once per signature per unique sequence.

A new file (feature.xml) has been created which contains non-signature match data from the structural databases (PDB, MODBASE and SWISS-MODEL) for UniProtKB proteins. Proteins from UniProtKB that do not match any of the signatures in InterPro's member databases have been added to our match XML file.

Web services: New SOAP-based Web Services have been added to complement the existing InterProScan Web Service. These allow users to programmatically retrieve InterPro entry data such as the abstract, integrated signature lists or GO terms. Users can download a range of clients from http://www.ebi.ac.uk/Tools/webservices/clients/dbfetch, including PERL, C#.NET and Java clients, to access this data.

AVAILABILITY

The database and related software are freely available to be downloaded and distributed, so long as the appropriate Copyright notice is supplied.

DISCUSSION

In the early stages of InterPro's evolution, signature development between the member databases was not a coordinated effort and resulted in a high level of redundancy, with some InterPro entries eventually containing up to 10 signatures. Each database is cultivating its own niche in signature development, with the aim of expanding sub-families and building signatures representative of newly characterized families, rather than duplicating work. This trend is illustrated in Figure 1.

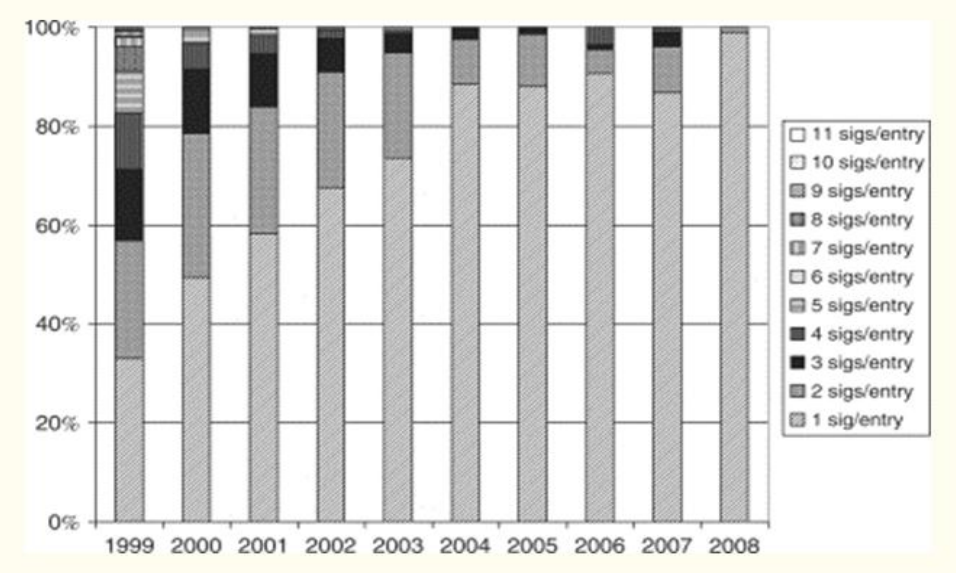


Figure 1: Trends in number of signatures integrated into a single entry, categorized by the year the entry was first created. Initially, these entries would have only contained signatures from the founding four consortium members. However, as other member databases joined, they also may have had signatures covering the same families and domains which consequently also became integrated into these entries, leading to the totals we see today. Note that the number of signatures integrated in a single year can vary (between 1000 and 5000 signatures) dependent on the member databases' release cycles.

CONTENTS OF CURRENT RELEASE

Individual InterPro entries contain a description of the protein family, domain, repeat or post-translational modification (e.g. *N*-glycosylation site); a list of member database signatures, Hidden Markov Models (HMMs), profiles or fingerprints associated with the entry; an abstract derived from merged annotation from the member databases; examples of representative sequences; literature references used to create the abstract; and links to tabular or graphical views of the matches to SWISS-PROT and TrEMBL. An example is shown in Figure 2.

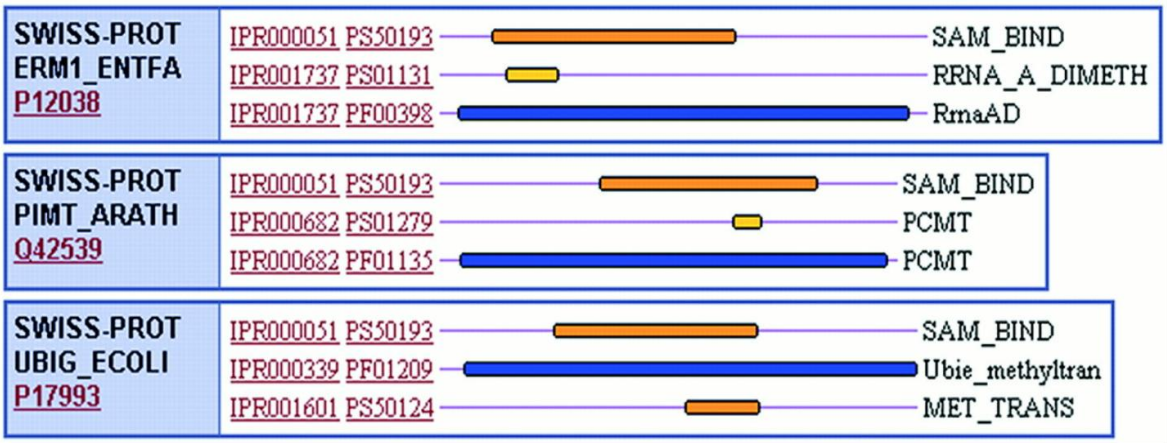


Figure 2:. An example of an InterPro entry. This is IPR000890, an entry containing signatures describing the acetate and butyrate kinase protein family. The 'i' information buttons have links to help files describing, for example, the 'Family' concept.[10]

APPLICATIONS OF INTERPRO

- InterPro is an international initiative that was conceived in an attempt to streamline the efforts of the signature database providers
- A primary application of InterPro's family, domain and functional site definitions will be in annotation and functional classification of uncharacterized sequences.
- The EBI is using InterPro for enhancing the automated annotation of TrEMBL
- InterPro has also proven its usefulness for whole proteome analysis in the comparative genome analysis of *Drosophila melanogaster, Caenorhabditis elegans* and *Saccharomyces cerevisiae*.
- Another major use of InterPro will be in identifying those families and domains for which the existing

discriminators are not optimal and could hence be usefully supplemented with an alternative pattern

- Alternatively, InterPro is likely to highlight key areas where none of the databases has yet made a contribution and hence where the development of a specific pattern might be useful[11]
- As it evolves, InterPro will streamline the analysis of newly determined sequences for the individual user and will make a significant contribution in the demanding task of automatic classification of predicted proteins from genome sequencing projects.

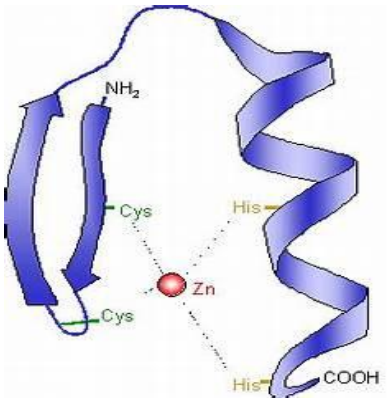
PROTEIN MOTIFS RELATED TP STRUCTURE /FUNCTION

Protein Motifs

Protein motifs may be defined by their primary sequence or by the arrangement of secondary structure elements.

Motifs

The term motif is used in two different ways in structural biology. The first refers to a particular amino-acid sequence that is characteristic of a specific biochemical function. An example is the so-called zinc finger motif.



Sequence motifs

Sequence motifs can often be recognized by simple inspection of the amino-acid sequence of a protein, and when detected provide strong evidence for biochemical function. The protease from the human immunodeficiency virus was first identified as an aspartyl protease because a characteristic sequence motif for such proteases was recognized in its primary structure. The second, equally common, use of the term motif refers to a set of contiguous secondary structure elements.

Structure motifs

A structure motifs is a super secondary structure, which appears in a variety of other molecules.

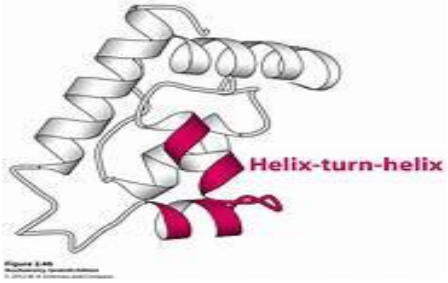
Structure motifs are short segments of protein 3D structure, which are spatially close but not necessarily adjacent in the sequence. Examples of structural motifs that represent a large part of a stably folded domain include the four-helix bundle.

Functional motifs

A structure motif that perform the biological function

- Short continuous stretch of primary sequence
- Define in terms of protein architecture

Bioinformatics era



Any primary sequence pattern that is associated with the biological function.

Along with the functional sequence motifs, the former are known generally as functional motifs. An example is the helix-turn-helix motif found in many DNA-binding proteins.[12]

(i) THE PROSITE DATABASE

Introduction

The PROSITE database uses two kinds of signatures or descriptors to identify conserved regions, i.e. patterns and generalized profiles, which both have their own strengths and weaknesses defining their area of optimum application. Each PROSITE signature is linked to an annotation document where the user can find information on the protein family or domain detected by the signature: origin of its name, taxonomic occurrence, domain architecture, function, 3D structure, main characteristics of the sequence, domain size and some references.[13]

NEW FUNCTIONAL PREDICTION TOOL

PROSITE has a long experience in documentation and detailed annotation of domains, families and functional sites.

Two types of information are stored in ProRule:

1. General information: the occurrence of a match with a profile is enough to trigger this annotation. Usually, it is restricted to the name of the domain and the position of its boundaries.
2. Conditional information: this is dependent on the presence of given amino acids at precise positions, on the occurrence of other domains or on taxonomic specificity. This information is only transferred if the conditions are fulfilled. For example, an enzymatic active site is annotated only if the correct amino acid is found at the required position.

WEB PAGE

The PROSITE website was redesigned and new predictive tools were implemented to assign more detailed functional information to the scanned proteins.

Users who want to scan their own proteins against all PROSITE entries or to scan a PROSITE entry against a protein database will find a new version of the ScanProsite web page.[14]

The documentation page has also been reorganized. It now contains three main sections:

1. The description part that exposes the main characteristics of the domain or the family and a representative list of proteins that contain the domain or belong to the family.
2. A technical section that refers to the descriptors used to identify the domain or family.
3. The third section is the reference block where, for each reference, we added the PubMed ID and a direct link to the article.

HOW TO OBTAIN PROSITE

PROSITE and proRule are freely available to academic users

- Weekly updates of PROSITE are available on our FTP server: ftp://ftp.expasy.org/databases/prosite/release_with_updates/.
- PROSITE is also accessible from the Hits page: http://hits.isb-sib.ch/.
- Frame-tolerant scans can be performed at the following address: http://www.isrec.isb-sib.ch/software/PFRAMESCAN_form.html.

(ii) THE PFAM DATABASE

Introduction

PFAM is a collection of protein families and domains. PFAM contains multiple protein alignments and profile-HMM of these families. PFAM is a semi-automatic protein family database, which aims to be comprehensive as well as accurate.

PFAM is a database of multiple alignment of protein domains or conserved protein regions.

PFAM is composed of two parts; the first part, PFAM-A, contains curated families each with an associated (profile-HMM) that can be used for alignments and database searching.

The second part of PFAM is PFAM-B in which sequence segments that are not included in PFAM-A are clustered automatically, allowing PFAM to be comprehensive.

Each PFAM-A family consists of four elements:

(i) Annotations
(ii) A seed alignments
(iii) A profile HMM
(iv) A full alignment

Pfam data are available in a variety of formats, which include flat files. The Pfam website (available at http://pfam.sanger.ac.uk/andhttp://pfam.janelia.org/) provides different ways to access the database content, providing both graphical representations of and interactive access to the data.[15]

RECENT CHANGES TO THE DATABASE CONTENT

Removing dubious sequences from the underlying database

Each Pfam release is calculated against a fixed sequence database, called pfam seq, which is derived from UniProtKB. At the beginning of a release cycle, we take a copy of the current version of UniProtKB and process it in two ways, the second of which is a novel addition for release 27.0. First, we remove sequences that contain non-consecutive regions. The linear sequence-information in these proteins will be inaccurate, as adjacent residues in the sequence can flank an intervening number of unsequenced residues. There are currently <1000 UniProt entries that contain non-consecutive sequence regions. The second, new processing step is the removal of sequences derived from spurious open reading frames, which are identified by searching AntiFam (11) models against the sequence database.

Family full alignments and trees

When building a Pfam release, we aim to ensure that the same set of post-processing operations are performed on all families regardless of size, thereby providing consistency both to the database and to the website. One of the distinguishing features of Pfam compared with most e protein family databases is our provision of full alignments.

USING PFAM

The PFAM web sites allow the database to be queried in one of three ways:

First, a user may have a sequence for which they know nothing, in such a case, this sequence can be searched against the current collection of PFAM profile HMMs to locate regions of the sequence that belongs to known domain families.

Second, if the user already has a SISS-PROT or SP TrEMBL identifier for the sequence they can access pre-calculated matches using the SWISSPFAM resource.

CHANGES TO PFAM

PFAM-B is an automatically generated supplement to PFAM that provides completeness in terms of coverage. It has also provided a useful resource for new PFAM-A families. PFAM-B in principle is made from parts of PRODOM not covered by PFAM-A.

SEARCHING PFAM

The US and UK PFAM servers provide users the ability to search query protein sequence against one, all or a few PFAM HMMs.[16]

(iii) PROFILE SCAN

Description

Profile Scan uses the method of Gribskov et al. (CABIOS **4(1)**; 61-66 (1988)) to find structural and sequence motifs in protein sequences. These motifs are represented as profiles in a library. Profile Scan aligns each profile motif to the sequence, and displays all alignments between the profile and sequence that have a normalized score above a set threshold. Because more than one alignment between a sequence and a particular motif can be found, each repeat of a duplicated structure (such as the zinc finger motif) can be presented.[17]

Functions

Profile Scan uses a database of profiles to find structural and sequence motifs in protein sequences.

THE PROTEIN DATA BANK

The Protein Data Bank (PDB) was established at Brookhaven National Laboratories (BNL) (1) in 1971 as an archive for biological macromolecular crystal structures.

The Protein Data Bank is the single worldwide archive of structural data of biological macromolecules. This paper describes the goals of the PDB, the systems in place for data deposition and access, how to obtain further information, and near-term plans for the future development of the resource. The primary goals of the resource are,

- To enable you to locate structures of interest
- To perform simple analysis on one or more structures.
- To act as a portal to additional information available on a structure notably the Cartesian atomic coordinates for further analysis

Search method

The search tools can be accessed from the PDB home page. The types of possible searches are:
By providing a PDB identification code (PDB) id.
By searching the text found in PDB files
By searching against specific fields of information –for example, deposition data or author
By searching on the status of entry, on Hold or released (status)
By iterating on a previous search.

File format

The file format initially used by the PDB was called the PDB file format.

The structure files can be downloaded in any of these three formats. In fact, individual files are easily downloaded into graphics packages using web addresses:

- For PDB format files, use, e.g., http://www.pdb.org/pdb/files/4hhb.pdb.gz or http://pdbe.org/download/4hhb
- For PDBML (XML) files, use, e.g., http://www.pdb.org/pdb/files/4hhb.xml.gz or http://pdbe.org/pdbml/4hhb

Content of the data collected by the PDB

All the data collected from depositors by the PDB are considered primary data. Primary data contain, in addition to the coordinates, general information required for all deposited structures and information specific to the method of structure determination.

The information content of data submitted by the depositor is likely to change as new methods for data collection, structure determination and refinement evolve and advance. In addition, the ways in which these data are captured are likely to change as the software for structure determination.

CURRENT DEVELOPMENTS

In the coming months, the PDB plans to continue to improve and develop all aspects of data processing. Deposition will be made easier, and annotation will be more automated. In addition, software for data deposition and validation will be made available for in-laboratory use.

The PDB will also continue to develop ways of exchanging information between databases.

Finally it is recognized that structures exist both in the public and private domains. Users will be able to load both public and proprietary data and use the same search and exploratory tools used at PDB resources.[18]

DATA BASE FOR PROTEIN STRUCTURE CLASSIFCATION

(i) SCOP: a Structural Classification of Proteins database

The Structural Classification of Proteins (SCOP) database provides a detailed and comprehensive description of the relationships of known protein structures. The classification is on hierarchical levels: the first two levels, family and superfamily, describe near and distant evolutionary relationships; the third, fold, describes geometrical relationships.

The sequences of proteins in SCOP provide the basis of the ASTRAL sequence libraries that can be used as a source of data to calibrate sequence search algorithms and for the generation of statistics on, or selections of, protein structures.

CLASSIFICATION

The levels of SCOP are as follows.

1. Class: Types of folds, e.g., beta sheets.
2. Fold: The different shapes of domains within a class.
3. Superfamily: The domains in a fold are grouped into superfamilies, which have at least a distant common ancestor.

4. Family: The domains in a superfamily are grouped into families, which have a more recent common ancestor.
5. Protein domain: The domains in families are grouped into protein domains, which are essentially the same protein.
6. Species: The domains in "protein domains" are grouped according to species.
7. Domain: part of a protein. For simple proteins, it can be the entire protein.

The classification of the proteins in SCOP is on hierarchical levels as follows:

- **Class**. The different folds have been grouped into classes. Most of the folds are assigned to one of the five structural classes:

1. All-α, those whose structure is essentially formed by α-helices;
2. all-β, those whose structure is essentially formed by β-sheets;
3. α/β, those with α-helices and β-strands;
4. α+β, those in which α-helices and β-strands are largely segregated;
5. Multi-domain, those with domains of different fold and for which no homologues are known at present.

- **Folds:** Each class contains a number of distinct folds. This classification level indicates similar tertiary structure, but not necessarily evolutionary relatedness. For example, the "All-α proteins" class contains >280 distinct folds, including: Globin-like (core: 6 helices; folded leaf, partly opened), long alpha-hairpin (2 helices; antiparallel hairpin, left-handed twist).
- **Family.** Proteins are clustered together into families on the basis of one of two criteria that imply their having a common evolutionary origin: first, all proteins that have residue identities of 30% and greater; second, proteins with lower sequence identities but whose functions and structures are very similar; for example, globins with sequence identities of 15%.
- **Superfamily.** Families whose proteins have low sequence identities but whose structures and, in many cases, functional features suggest that a common evolutionary origin is probable, are placed together in superfamilies; for example, the variable and constant domains of immunoglobulins.
- **Common fold**. Superfamilies and families are defined as having a common fold if their proteins have the same major secondary structures in the same arrangement and with the same topological connections Other classes have been assigned for peptides, small proteins, theoretical models, nucleic acids and carbohydrates.

PDB entry domains

A "TaxId" is the taxonomy ID number and links to the NCBI taxonomy browser, which provides more information about the species to which the protein belongs. Clicking on a species or isoform brings up a list of domains. For example, the "Hemoglobin, alpha-chain from Human (Homo sapiens)".

Clicking on the PDB numbers is supposed to display the structure of the molecule.

There are now a number of other databases which classify protein structures, such as CATH (4), FSSP (5), Entrez (6) and DDBASE (7),[19]

(i) Class, architecture, topology and homologous superfamily

Cath: protein structure classification

The CATH database provides hierarchical classification of protein domains based on their folding patterns.

The CATH database is a classification of protein domains based not only on sequence information, but also on structural and functional properties. CATH offers an important tool to researchers.

The CATH hierarchy

The domains are then classified within the CATH structural hierarchy:

- At the Class (C) level, domains are assigned according to their secondary structure content, i.e. all alpha, all beta, a mixture of alpha and beta, or little secondary structure;
- At the Architecture (A) level, information on the secondary structure arrangement in three-dimensional space is used for assignment;
- At the Topology/fold (T) level, information on how the secondary structure elements are connected and arranged is used
- Assignments are made to the Homologous superfamily (H) level if there is good evidence that the domains are related by evolution[2] i.e. they are homologous.

The four main levels of the CATH hierarchy:

#	Level	Description
1	**C**lass	The overall secondary-structure content of the domain. (Equivalent to the SCOP Class)
2	**A**rchitecture	High structural similarity but no evidence of homology. (Equivalent to the 'fold' level in SCOP)
3	**T**opology/fold	a large-scale grouping of topologies which share particular structural features
4	**H**omologous superfamily	Indicative of a demonstrable evolutionary relationship. (Equivalent to SCOP superfamily)

CONCLUSION

Since the origin of the protein sequence database the classification of protein sequence into super families has provided a biological meaningful organization of the data. As discussed above the classification provides a systematic scheme for verification of the information in the database and for interfering additional information by homology in the controlled way.

Information generated from large scale sequencing projects is incomplete and not well understood. The major task of computational biology is to assign biological meaning to this data. Homology is the major opening principle employed in these analyses.

The superfamily classification provides a useful architecture for self-consistent and objective examination of sequence data by homology.[20]

REFERENCES

1. Text book of bioinformatics concept, skills and application (2nd edition) by SC rastogi, namita mendiratta, parag rastogi.
2. Bairoch and boeckmann, B (1991) the SWISS-PROT protein sequence data banks nucleic acid res., 19: 2247-22479.
3. Berman HM, Westbrook J, Feng Z, Gilliland G. et al. The Protein Data Bank.
4. Dayhoff M.O., Eck, R.V., Chang, M.A. and Sochard, M.R. (1965) *Atlas of Protein Sequence and Structure*, Vol. 1. National Biomedical Research Foundation, Silver Spring, MD.
5. Dayhoff M.O. (1979) *Atlas of Protein Sequence and Structure*, Vol. 5, Supplement 3. National Biomedical Research Foundation, Washington, DC.
6. Barker W.C., George, D.G., Mewes, H.-W., Pfeiffer, F. and Tsugita, A. (1993) Nucleic Acids Res., 21: 3089–3092. [PMC free article] [PubMed]
7. Westbrook J. and Bourne, P.E. (2000) Bioinformatics, in press.
8. Orengo CA, Michie AD, Jones DT, Swindells MB. et al. CATH: A hierarchic classification of protein domain structures. Structure, 1997.
9. Bairoch A. and Apweiler, R. (1999) Nucleic Acids Res., 27: 49–54. [PMC free article] [PubMed]
10. Stoesser G., Tuli, M.A., Lopez, R. and Sterk, P. (1999) Nucleic Acids Res., 27: 18–24. Updated article in this issue: *Nucleic Acids Res.*, 2000; 28: 19–23.
11. Westerfield M., Doerry, E., Kirkpatrick, A.E. and Douglas, S.A. (1999) Methods Cell Biol., 60: 339–355. [PubMed]
12. Bernstein F.C., Koetzle, T.F., Williams, G.J., Meyer, E.E., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) J. Mol. Biol., 112: 535–542. [PubMed]
13. Bourne P., Berman, H.M., Watenpaugh, K., Westbrook, J.D. and Fitzgerald, P.M.D. (1997) Methods Enzymol, 277: 571–590. [PubMed]
14. Corpet F, Gouzy J, Kahn D. recent improvement of proDom database of protein domain families. Nucleic acid res., 1999; 263-267.
15. Bairoch A. and Apweiler, R (2000). THE SWISS-PROT protein sequence database and in supplement TreMBL in 2000, nucleic acid res., 28(!): 45-48.
16. Arnold k, kiefer F, Koppy J, battery JN, Podinec M, Westbrook JD, Berman HM, Bordoli L, Schwede T. the protein model portal. J structure function genomics, 2009.
17. https://www.ebi.ac.uk/swiss-prot/and http://www.expasy.org/sprot/
18. https://www.ncbi.nlm .nih.gov/PMC/articles/PMC102476
19. https://www.ncbi.nlm.nih.gov/PMC/articles/PM352 5972
20. https./www.expasy.ch/sprot http:/www.ebi.ac.uk/swissprot