Open Access
Full Length Article

REVIEW ARTICLE

Structural Bioinformatics: Computational Software and Databases for the Evaluation of Protein Structure

Ayisha Amanullah* and Suad Naheed
Department of Biotechnology, Jinnah University for Women, Karachi, Pakistan

ABSTRACT

Databases are the computerized platform where information is stored and can be retrieved easily by public users. Biological databases are the repositories of biological data. These biological data libraries contain facts and figures related to various disciplines of research including genomics, proteomics, microarray technology, metabolomics and phylogenetics. By using biological databases, a broad collection of essential biological information can be exploited ranging from function, structure and localization of gene, clinical consequences of mutation to similarity index among biological sequences and structures. Nowadays, different kinds of biological databases are available on the web. The present write up focuses on biological databases and bioinformatics tools for protein structure analysis. This review also aims to elaborate the searching schemes, available in different structural databases. The wide variety of different levels and types of information content related to 3D protein structures are available on web-based databases. Regarding the biological functions and 3D structures of various proteins, these databases provide a huge range of useful links, schematic diagrams as well as strategies for detailed analysis of proteins and other macromolecules structures. 3D structural illustration of proteins stored in structural databases is determined and visualized by X-ray crystallography, electron microscopy and NMR spectroscopy. On regular basis, a large number of protein structures are submitted by structural biologists, updated and curated by subject experts. Most familiar biological databases that store 3D protein and other macromolecules structures include, PDB, 3D Genomics, CATH, & SCOP. These databases contain valuable information of overall protein structures, domains and motif structures, protein-protein complex systems and complex of protein with other biomolecules.

Keywords
Database, PDB, CATH, MMDB, SCOP, domain, protein-protein complex systems.

*Address of Correspondence
ayisha.aman24@gmail.com

Article info.
Received: April 02, 2018
Accepted: September 24, 2018

Cite this article: *Amanullah A, Naheed S. Structural Bioinformatics: Computational Software and Databases for the Evaluation of Protein Structure. RADS J. Biol. Res. Appl. Sci. 2018; 9(2): 94-101.*

Funding Source: *Nil*
Conflict of Interest: *Nil*

INTRODUCTION

A database is an organized and structured computational based repository that can be easily retrieved, analyzed, managed and updated. The biological database is an archival collection of biological data. Biological databases fall into two categories i.e. primary database and secondary database. In the primary database, the information submitted is basically in its original form which is obtained experimentally. While in secondary databases, the information is highly curated after the evaluation of data, from primary databases. Nowadays, gathering, processing and analyzing of data has become an important aspect of research, especially in the field of biological sciences. Bioinformatics, a new and foremost sub-discipline of biotechnology, mainly focuses on data

processing. Through the various strategies of bioinformatics, biological databases are designed after data gathering and processing. These biological databases are incorporated with specific features[1]. Among various databases, protein database is a collection of various sets of data about proteins. It is serving as a storehouse of information since 1971 and stores information about protein sequence, protein structure, its conformation and active sites. Moreover, it also describes the physical and chemical nature and biological role of proteins as proteins are solely responsible for all the biological functions in the body[2]. Protein structure databases elucidate the protein structure in normal as well as in diseased condition in the biological system. Through these structure databases, the 3D structure of proteins and cell dimension can be studied.

This review highlights the most common protein structure databases, that will be helpful to those individuals in the scientific community who would be interested to conduct research in bioinformatics specifically in protein structure studies.

PROTEIN DATA BANK (PDB)

Protein data bank is an online structural library of biological macromolecules, which is the only worldwide repository of macromolecular structure. The PDB was organized in 1971 at Brookhaven National Laboratories (BNL) as a platform of crystal structures of biomolecules. Over the years, the data submitted to PDB was modified and approaches to access the PDB have changed, as a result of advancements in technology.

In October 1998, Research Collaborator for Structural Bioinformatics (RCSB) has started to manage and maintain the activities of PDB. PDB data is curated and annotated by RCSB PDB. The major task of the RCSB is to generate such measures that allow the use and analysis of structural data[3]. PDB stores 3D structural information of biological molecules mainly nucleic acid and proteins. The structural information of biomolecules is commonly acquired experimentally by NMR spectroscopy, X-ray crystallography, electron microscopy etc.

Structural information of some chemical ligands and nucleotides are also available on PDB. PDB ID is a four-character identifier that is actually entitled as PDB entry. A user can access PDB at http://www.rcsb.org/pdb/ or http://www.pdb.org (Fig. **1**).

Fig. 1: A Homepage of PDB.

Searching through PDB is done by a vast range of search engines ranges from PDB ID and keywords to structural features of proteins and other biomolecules (Fig. **2**).

Fig. 2: Searching strategies available at PDB.

There are two formats that PDB uses to keep structural data: The PDB file format and macromolecular crystallographic information file format (mmCIF). PDB file design is more commonly used in protein community as compared to mmCIF.

PDB offers various molecular structural visualization soft wares including Jmol, PDB simple viewer, PDB protein workshop and RCSB-Kiosk. Structural confirmation of secondary structure is also provided by PDB[4].

The PDB depository is run by an association, named the Worldwide Protein Data Bank (wwPDB) which guarantees that the information is freely accessible to the public. Structures for huge numbers of the proteins and nucleic acids required in the central procedures of life are available on PDB[5].

PDB SUM

At present, more than 13000 3D structures of biomolecules have been exploited experimentally by using NMR spectroscopy and X-ray crystallography. Most of these 3D structures mainly include proteins, DNA and protein-ligand complexes (Fig. **3** & **4**). Along with the

sequence, functional and physiochemical properties give a bundle of information, quite sufficient for developing knowledge regarding biochemical processes. These structures are submitted to PDB which can be obtained from RCSB's PDB web page. There are other structural archives that collect further information on the particular type of molecules or particular features of the molecules.

PDBsum is one of the principal database. PDBsum was established in 1995 with the goal to make an available pictorial summary of biological macromolecules structures (DNA, RNA, protein, metal ions, water molecules and small molecule ligands), present in PDB, along with their major structural attributes. PDBsum keeps 3D images of the protein structure, protein secondary structure annotations, detailed structural statistics, developed from the PROMOTIF computer software, summary PROCHECK consequences and flow diagrams that indicate connections between protein, ligands and DNA molecules. RasMol focuses the crucial features of the protein structure including domains, protein-ligand correlations and PROSITE format for interactional 3D visualization.

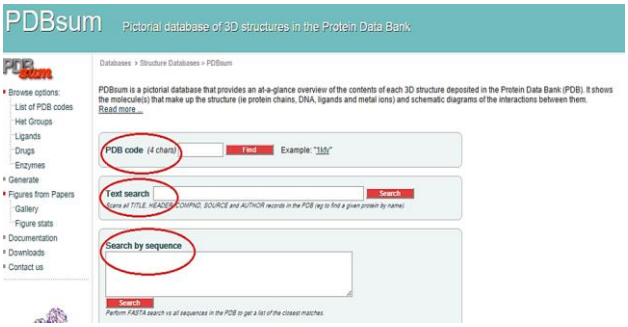

Fig. 3: Search engines available at PDBsum.

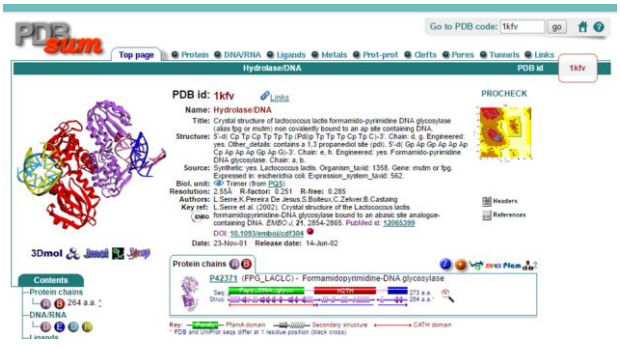

Fig. 4: Pictorial summary of Hydrolase/DNA with PDB id: 1kfv at PDBsum.

PDBsum is frequently upgraded, whenever any new structure is submitted to PDB. PDBsum is accessed freely through URL: http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum [6].

3D-GENOMICS

3D-GENOMICS is a structural database that gives information about the protein structure from the sequenced genome. The database stored information of 93 proteomes in August 2003. It stores information about homologous sequences from different sequence databases, patterns from Prosite, protein domain information from pfam and SCOP, and characteristics of other predicted sequences such as transmembrane regions and coiled coils. The structure of the protein and its biochemical function can be annotated after the genome has been sequenced by this database. Various searching policies are available on 3D-GENOMICS database that allow a user to,

- Retrieve data directly for individual protein sequence by accession numbers or keywords.
- Investigate the desired sequence selected from summarized annotations for a specific proteome.
- Or approach pre-calculated frequency based cross proteome comparative study[7].

Different methodologies are involved for retrieving the information such as the recognition of motifs which are responsible for structure and function, characterization of coiled and integral regions sequence identification and the identification of similar proteins whose function or the structure has already been reported[8].

CATH DATABASE

Protein evolution originates in the families of structurally closed protein, within which sequence identities can be very small. This can allow effective structure-based classification in distinguishing unpredicted associations in known structures and under ideal conditions, the function can also be allocated. The constant emerging variety of fully known protein structures is just so massive. Hence it is impossible to classify all proteins manually. In order to overcome the situation, automated techniques are needed for quick assessment of these biological molecule structures[9].

CATH is a protein structure classification bank which stores hierarchical ranking (taxonomic classification) of protein domains based on their folding mechanisms. CATH uses evolutionary information of proteins and classifies protein domains into superfamilies if they have emerged from same ancestors. CATH database obtains domains (Fig. **5**) from protein structures, stored in PDB[10].

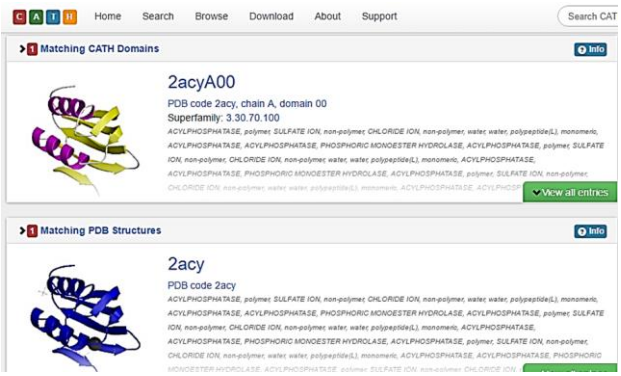

Fig. 5: Protein domains data, accessible through CATH.

Almost, 114,215 domains are present in the latest version of CATH, which was released in July 2008. These domains are graded in a hierarchical manner with four main divisions, named Class (C), Architecture (A), Topology (T) and Homologous superfamily (H), so the name coined "CATH". CATH can be browsed through URL www.cathdb.info. The CATH has four different visions when classifying protein domains,

1. **Class (C):** At this level, domains are categorized on the basis of their secondary structure into four structural types: including alpha, beta, mixed alpha beta and forth type consists of domains comprised of few secondary structures.
2. **Architecture (A):** This level classifies domains on the basis of the orientation of their secondary structures.
3. **Topology (T):** This level mainly focuses on the topological connectivity of the secondary structures.
4. **Homologous superfamily (H):** This level groups the domains according to their common evolutionary linkages; therefore, they can be demonstrated as homologous. The comparison among closely related domains is based on sequence and structural similarity, which is done by using SSAP- a dynamic programming algorithm[11].

Construction of Database

The information in CATH is received from PDB files submitted within the supermolecule information Bank. The database encloses structures resolved with 4A^{O} resolution or higher. The necessities of CATH mainly include that domain should consist of at least 40 residues and side chains should be resolved with 70% or more[12]. As described in the introduction, the current edition of CATH consists of 114,215 domains, treated from the proteins deposited in PDB.

New domains structures are incorporated in CATH on the basis of two fundamental principles,

1) Deposited protein chains are chopped to acquire domains of those fragments.
2) Afterward, these resulting domains are classified[11].

By analyzing structural families produced by CATH database shows the outstanding options of macromolecule structure area. A well-defined info relevant to the structural families of macromolecules like CATH can help the assigning of structure-function and phylogenetic relationships to each known and freshly determined structures of macromolecules.

The CATH database is effective for researcher and bioinformaticians. Web surfing is made simple and convenient even for a single domain via a manageable user-friendly net interface for researchers having a particular task, whereas bioinformaticians with insight on comprehensive research can explore entire datasets provided for downloading. The database is frequently updated and supplemented with numerous approaching extensions with horizontal layers corresponding to the hierarchical data structure; CATH is probably going to be an excellent useful protein taxonomic computational tool in the future. Thus, operating with CATH is remarkably uncomplicated[13].

PROTEIN FOLDING DATABASE

The Protein Folding Database (PFD) is freely available data house that is supplied with the data concerning to methodological, structural annotations, kinetic and thermodynamic folding patterns of a large number of proteins i.e. greater than 50 proteins belonging to 39 families[14]. This database collects overall protein folding

information into a solitary, effectively open asset. An easy to use web database has been produced that permits strong searching, browsing, information mining and data recovery while giving connections to other protein repositories. The protein folding database architecture represents the view of folding impressions in a helpful and innovative manner, with the everlasting goal that encourages data retrieval and bioinformatics procedures[15].

Structural Organization of the Database

A salient purpose of the database is to permit the examination of the exact and hypothetical connections between structural characteristics and folding frequencies of a protein, for instance, topology. Consequently, new tables were included into databases in order to assemble information like sequence, expression tags, construct length, PDB identifier and disordered regions. Additional tables also enable the accumulation of crude kinetic information and blunders for every single numerical data are presently recorded[14].

Folding Research and PFD

The PFD deposits a stockpile of folding pattern information that can be examined with various factors and the results show or appear in the structural form. The repository allows a full information, spreadsheet-like rundown of outcomes allowing rapid exploration of common or general fashion in data. The inquiry data outcomes can be looked for on any caption, which is valuable, e.g. while reviewing the inconstancy of folding frequencies among proteins inside a family. Information related to publications of each entry and URL to access NCBI PubMed writing database is recorded[15].

Futuristic Approaches

The PFD has been redesigned by the laws defined by the International Fold Omics Consortium. Latest approaches in this tool will motivate the further construction of the database, and new methods of presenting data information graphically will upgrade its utilization in protein science field. The main objective of future research will be on the advancement of further graphical portrayals of the folding data information. It is important to reveal here that this database is not only an information archive but rather turns into an intense analytical tool in folding research[14].

DATABASE OF MACROMOLECULAR MOTIONS

Macromolecular motion is crucial to understand the function. The macromolecular motion allows understanding the mode of catalysis, signaling and the mechanisms involved in the formation of complexes. Database of macromolecular motion is a platform, to which structures are submitted and this database creates putative motion trajectories, that play a vital role in structural biology[16]. The database of macromolecular motions is used generally as a structural community. This database is freely available at http://bioinfo.mbb.yale.edu/MolMovDB, it schematizes the protein and nucleic acid movement in order to obtain structural information. Evidence about experimental information, structural similarity is available in this database. In order to implement a database, we can use the design of standard relation. In addition, the heterogeneity, as well as complexity regarding information, is available in the database thus promoting its link to an object-oriented approach. However, in order to keep complex data, the database is equipped with imaginable depictions for motion pathways, obtained from 3D interpolation among the known conformation[17].

SCOP DATABASE

There is another protein structural repository that stores information of those proteins that are similar at the structural level and possess common ancestors. SCOP (Structural Classification of Proteins) was constructed for this purpose[18]. In 1994, SCOP was developed in the laboratory of molecular biology and center for protein. SCOP database consists of comprehensive details of evolutionary relatedness of known protein structures. The basic unit of classification is a protein domain, which is hierarchically classified into species, proteins, families, superfamilies, folds and classes.

Classification

SCOP classifies the protein based on hierarchical levels, which are discussed below,

- **Family**, proteins are grouped into families according to two strategies, based on common evolutionary relationships. 1st all proteins with 30% or greater

residue identities, 2nd proteins having small sequence identities but they are similar at functional and structural level.

- **Superfamily**, this level occupies the families of proteins that hold least sequence identities, but their structural and in many conditions, functional characteristics share a same evolutionary origin.
- **Common fold**, it groups the proteins of families and superfamilies, have major similar secondary structures with the same order and topological connectivity.
- **Class**, this level groups different folds into classes. These protein folds are categorized into five classes,
    1. All alpha, whose structure is composed of alpha helices.
    2. All beta, whose structure consists of beta sheets.
    3. α/β, whose structure is formed from alpha helices and beta sheets.
    4. α+β, in these proteins, alpha helices and beta strands are massively sequestered.
    5. Multidomain, protein domains with different folds and for which homologs are unknown.

Searching Facilities at SCOP

For the browsing and searching of particular families of protein at SCOP, SCOP is provided with the variety of search engines and techniques for navigations including,

- Surfing through SCOP hierarchy
- Browsing by using amino acid sequence
- Search through a key
- Search through PDB identifier
- Browsing using history[19].

SCOP is freely available at URL link http://scop.mrc-lmb.cam.ac.uk/scop/.

SCOP 2

There was a closure of SCOP in 2010 and in January the prototype for a new SCOP2 database has become easily accessible, which reflects a novel approach for the classification of proteins that differs quietly from the previous version of SCOP, but best features have been introduced.

In SCOP2 (Fig. **6**), proteins are also arranged according to the evolutionary and structural relationship but not like old version of SCOP which provides hierarchy in simple tress like structures, this database describes the classes

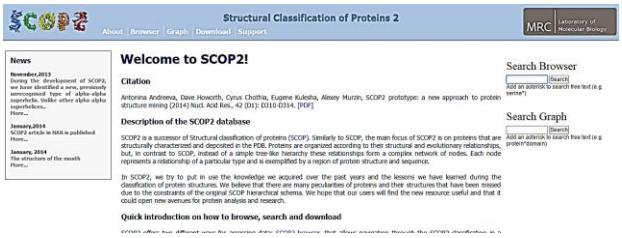

Fig. 6: A Homepage of SCOP2.

of proteins in a form of graph which is actually acyclic in shape with a network of complex nodes, which defines a relationship of protein of particular type with others[20].

MMDB (MOLECULAR MODELING DATABASE)

MMDB is the 3D structural database of biological macromolecules and their molecular interaction, developed by NCBI. MMDB can be accessed through NCBI Entrez's search engine[21]. The main aim of MMDB is to circulate the information of the 3D structure of macromolecules mainly protein and functional annotations among molecular biologists [22]. MMDB filters the contents of PDB. It links protein 3D structure information along with sequence information, sequence classification resources and PubChem- an archive that keeps the chemical structural data of small molecule and their biological roles, thus making available various approaches to 3D structure data for molecular biologists, structural biologists, and chemists. MMDB gives a comprehensive detail on structural alignments and software for 3D structure visualization by graphical viewer Cn3D.

MMDB also focuses the quaternary structures and the molecular interactions between its components. MMDB is accessible through the World Wide Web at URL http:// www.ncbi.nlm.nih.gov/structure [21].

CONCLUSION

There are numbers of web-based protein structure databases that contain different extents of information on biological macromolecules structures on a different basis. Generalized databases contain general information about protein structures; include those protein structures experimentally investigated by NMR Spectroscopy and X-ray Crystallography. They also give useful web resources and schematic diagrams related to the protein 3D

structure and their biological functions. There is another category of protein structure database which systemizes 3D structures by their folds as they display evolutionary relatedness which may be difficult to determine from sequence alignment alone. In addition, there are diverse databases and servers that contrast folds of protein structures are beneficial for newly determined protein structures, and especially for those proteins or protein structures which are still unknown. Beyond these, there is a wide range of databases for the most specialized users that deal with particular families, different structural qualities, diseases and so on.

PDB is the most widely used and popular among protein structure communities that provides excellent details of protein structure in 3D, allowing the user to study and visualize the 3D structures by different computational based software (Jmol, NGL view etc.). PDB interprets the 3D structures to the user in an understandable way. On the other hand, SCOP and CATH categorize protein 3D structures on the basis of their folds and evolutionary origin. MMDB is also structural archive, headed by NCBI that also stores an extensive number of macromolecular 3D structures.

Although the number of these structural databases is increasing day by day with great advancements and facilities but the databases that have been discussed in this article are the milestones in the field of bioinformatics[1].

ABBREVIATIONS

PDB: Protein Data Bank
CATH: Class Architecture Topology Homologous superfamily
SCOP: Structural Classification of Proteins
RCSB: Research Collaboratory for Structural Bioinformatics
MMDB: Molecular Modeling Database
PFD: Protein Folding Database

REFERENCES

1. Bagchi A. A brief overview of a few popular and important protein databases. Computational Molecular Bioscience. 2012; 2:115-2.
2. Zhang Y, Zhu Y, He F. An overview of human protein databases and their application to functional proteomics in health and disease. Sci China Life Sci. 2011; 54(11): 988-98.
3. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, *et al*. The Protein Data Bank. Nucleic Acids Res. 2000; 28(1): 235-42.
4. Xu D, Xu Y. Protein databases on the internet. Curr Protoc Mol Biol. 2004, ch. 19, Unit 19.4.
5. About RCSB PDB: Enabling Breakthroughs in Scientific and Biomedical Research and Education. RCSB PDB; [cited 2018 March 19]. Available from: http://www.rcsb.org/pages/about-us/index
6. Laskowski RA. PDBsum: summaries and analyses of PDB structures. Nucleic Acids Res. 2001; 29(1):221–2.
7. Fleming K, Muller A, MacCallum RM, Sternberg MJ. 3D GENOMICS: a database to compare structural and functional annotations of proteins between sequenced genomes. Nucleic Acids Res. 2004; 32: D245–50.
8. Li C, Dong X, Fan H., Wang C, Ding G, Li Y. The 3DGD: a database of genome 3D structure. Bioinformatics. 2014:*30*(11): 1640-2.
9. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH: A hierarchic classification of protein domain structures. Structure. 1997;5(8):1093-108.
10. Sillitoe I, Lewis TE, Cuff A, Das S, Ashford P, Dawson N.L, *et al*. CATH: comprehensive structural and functional annotations for genome sequences. Nucleic Acids Res.2015; 43: D376–81.
11. Kundsen M, Wiuf C. The CATH Database. Human Genomics. 2010; 4(3):207-12.
12. Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, Dibley M, *et al*. The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Res. 2007;35: D291-7.
13. Cuff A, Redfern OC, Greene L, Sillitoe I. TheCATH hierarchy revisited – Structural divergence in domain superfamilies and the continuity of fold space. Structure. 2009;17(8):1051-62.
14. Fulton K., Bate M, Faux N, Mahmood K., Betts C, Buckle A. Protein Folding 8. Database (PFD 2.0): an online environment for the International Foldeomics Consortium. Nucleic Acids Res. 2007; 35(Database issue): D304-7.
15. Fulton KF, Devlin GL, Jodun RA, Silvestri L, Bottomley SP, Fersht AR, BuckleAM. PFD: a database for the investigation of protein folding kinetics and stability. Nucleic Acids Res. 2005; 33(Database issue): D279-83.
16. Flores S, Echols N, Milburn D, Hespenheide B, Keating K, Lu J, *et al*. The database of macromolecular motions: new features added at the decade mark. Nucleic Acids Res. 2006; 34: D296-301.
17. Gerstein M, Krebs W. A database of macromolecular motions. Nucleic Acids Res.1998; 26(18):4280-90.
18. Murzin AG, Brenner SE, Hubbard T. and Chothia C. SCOP: a structural classification of proteins database

for the investigation of sequences and structures. J. Mol. Biol.1995; 247:536-40.

19. Lo Conte L, Ailey B, Hubbard TJ, Brenner SE, Murzin AG. And Chothia C. SCOP: a structural classification of proteins database. Nucleic Acids Res.2000; 28: 257-9.
20. Andreeva A, Howorth D, Chothia C, Kulesha E & Murzin AG. *SCOP2 prototype: a new approach to protein structure mining.* Nucleic Acids Res.2014; 42: D310-14.
21. Madej T, Addess KJ, Fong JH, Geer LY, Geer RC, Lanczycki CJ, *et al*. MMDB: 3D structures and macromolecular interactions. Nucleic Acids Res. 2012; 40: D461-4.
22. Wang Y, Anderson JB, Chen J, Geer LY, He S, Hurwitz DI, *et al.* MMDB: Entrez's 3D-structure database. Nucleic Acids Res.2002; 30:249-52.