

Chapter 4

Biological 3D Structural Databases



Yasser Gaber, Boshra Rashad, and Eman Fathy

Contents

4.1	Introduction.....	48
4.1.1	X-Ray Crystallography.....	52
4.1.2	Crystal Formation.....	52
4.1.3	Structure Determination.....	53
4.2	Macromolecular Structural Databases.....	54
4.2.1	Protein Data Bank wwPDB.....	54
4.2.1.1	RCSB PDB.....	54
4.3	PDBsum: Structural Summaries of PDB Entries.....	60
4.4	sc-PDB: A 3D Database of Ligandable Binding Sites.....	62
4.5	PDBTM: Protein Data Bank of Transmembrane Proteins.....	63
4.6	CATH Database.....	64
4.7	SCOP (Structural Classification of Proteins) Database.....	66
4.8	Structure Comparison Servers.....	68
4.9	Conclusion.....	70
	References.....	70

Y. Gaber (✉)

Department of Pharmaceutics and Pharmaceutical Technology, Faculty of Pharmacy,
Mutah University, Al-Karak, Jordan

Microbiology Department, Faculty of Pharmacy, Beni-Suef University, Beni-Suef, Egypt

e-mail: Yasser.Gaber@mutah.edu.jo; Yasser.Gaber@pharm.bsu.edu.eg

B. Rashad

Biotechnology and Life Sciences Department, Faculty of Postgraduate Studies for Advanced
Sciences (PSAS), Beni-Suef University, Beni-Suef, Egypt

e-mail: boshramohamed@psas.bsu.edu.eg

E. Fathy

Biotechnology and Life Sciences Department, Faculty of Postgraduate Studies for Advanced
Sciences (PSAS), Beni-Suef University, Beni-Suef, Egypt

Microbiology Department, Directorate of Health Affairs at Ministry of Health,
Beni-Suef, Egypt

e-mail: emanfathy@psas.bsu.edu.eg

© Springer Nature Switzerland AG 2019

N. A. Shaik et al. (eds.), *Essentials of Bioinformatics, Volume I*,
https://doi.org/10.1007/978-3-030-02634-9_4

4.1 Introduction

Structural databases are storage platforms that are devoted to the three-dimensional (3D) structural information of macromolecules. The 3D structure determination of biomacromolecules is essential for understanding phenomena such as the mechanisms of disease development that can aid in the design of new drugs. Also, 3D structures of biomacromolecules help to find the structure-function relationship. For instance, a point mutation in an enzyme can lead to a serious disease; this is exemplified by the glucose-6-phosphate dehydrogenase mutant enzyme that has lower ability to bind NADP+cofactor, thus resulting in the hemolytic anemia syndrome (Wang et al. 2008). The availability of 3D structural information of macromolecules will unveil the mysterious protein-protein interaction. Also, the conserved amino acid analysis using 3D structural features of proteins facilitates understanding the structure activity relationships. Proteins are polymers of amino acid sequence; it is amazing that only 20 different amino acids account for all the diversities of proteins, which are mainly arranged into primary, secondary, tertiary, and quaternary structural forms. The primary structure refers to the linear attachment of amino acids that make up the polypeptide chain. Secondary structure denotes repeated and regular folding patterns of the main chain sequence either an alpha helix or beta sheets connected via coils, turns, or loops. Tertiary structure is the characteristic three-dimensional shape resulted from the secondary structure elements found in the protein. Quaternary structure refers to two or more protein subunits that are linked to each other via non-covalent interaction.

The start of original structure biology dates back to the 1950s, when DNA double helix, hemoglobin, and myoglobin structures were determined. In the following years, scientists paid great attention to the evaluation and study of protein structure in terms of the relation between protein sequence, structure, and function. In 1971, structure biologists held an important meeting to discuss the allowance of the public accessibility to structural data; as a result, the Brookhaven National Laboratory hosted the Protein Data Bank (Berman et al. 2012). The structural databases aim at keeping the information about the structures of each biomacromolecule, annotate its properties, and facilitate to the users finding relevant information and related structures. Table 4.1 lists the main 3D structural databases, tools, and servers that are essential for biologists, bioinformaticians, and even the public interested in structure biology. There are several structural databases that are available free of charge for public use and are responsible for archiving and organizing the 3D structural information of biological macromolecules and proteins such as RCSB PDB, PDBe, PDBj, BMRB, SCOP, and CATH. The 3D structural information can be seen as a primary source of data that requires effort for extraction and interpretation of the useful information. Therefore, other several types of databases and web servers are developed to add further levels of information such as comparison to other structures or focus on certain property, for example, the membrane protein databases (Bagchi 2012).

Table 4.1 List of important structural biological databases and related web resources for structure analysis

Database	Use/description	Link	References
<i>1. Primary structural data centers and other browsers</i>			
PDBj	Protein Data Bank Japan archives macromolecular structures and provides integrated tools	https://pdbj.org/	Kinjo et al. (2016)
BMRB	Biological Magnetic Resonance Data Bank (NMR), a repository for data from NMR spectroscopy on proteins, peptides, nucleic acids, and other biomolecules	http://www.bmrwisc.edu/	Markley et al. (2008)
PDBe	Protein Data Bank in Europe (PDBe) archives biological macromolecular structures	http://www.ebi.ac.uk/pdbe/	Velankar et al. (2010) and Velankar et al. (2015)
RCSB PDB	Research Collaboratory for Structural Bioinformatics Protein Data Bank archives information about the 3D shapes of proteins, nucleic acids, and complex assemblies	https://www.rcsb.org/	Berman et al. (2000)
PDBsum	Pictorial analysis of macromolecular structures	www.ebi.ac.uk/pdbsum	Laskowski (2007) and Laskowski et al. (2018)
<i>2. Structure classification databases</i>			
CATH	Domain classification of structures	http://www.cathdb.info/	Knudsen and Wiuf (2010)
SCOP	SCOP2, structural and evolutionary classification	http://scop2.mrc-lmb.cam.ac.uk/	Lo Conte et al. (2000)
<i>3. Nucleic acid databases</i>			
NDB	Nucleic acid database	http://ndbserver.rutgers.edu/	Coimbatore Narayanan et al. (2013)
RNA FRABASE	3D structure of RNA fragments	http://rnafrabase.cs.put.poznan.pl/	Popenda et al. (2010)
NPIDB	3D structures of nucleic acid-protein complexes	http://npidb.belozersky.msu.ru/	Zanegina et al. (2015)
<i>4. Membrane protein database</i>			
MemProtMD	MemProtMD, database of membrane protein	http://sbc.bioch.ox.ac.uk/memprotmd/	Stansfeld et al. (2015)
<i>5. Ligands and binding sites and metalloproteins</i>			
PeptiSite	Is a comprehensive and reliable database of biologically and structurally characterized peptide-binding sites that can be identified experimentally from co-crystal structures in the Protein Data Bank	http://peptisite.ucsd.edu/	Acharya et al. (2014)

(continued)

Table 4.1 (continued)

Database	Use/description	Link	References
ComSin	Database of protein structures inbound (complex) and unbound (single) states in relation to their intrinsic disorder	http://antares.protes.ru/comsin/	Lobanov et al. (2009)
MetalPDB	MetalPDB collects and allows easy access to the knowledge on metal sites in biological macromolecules	http://metalweb.cerm.unifi.it/	Putignano et al. (2017)
Pocketome	The Pocketome is an encyclopedia of conformational ensembles of druggable binding sites that can be identified experimentally from co-crystal structures in the wwPDB	http://www.pocketome.org/	An et al. (2005)
MIPS	A database of all the metal-containing proteins available in the Protein Data Bank	http://dicsoft2.physics.iisc.ernet.in/cgi-bin/mips/query.pl	Mewes et al. (2002)
<i>6. Structure comparison servers</i>			
DALI	The Dali server is a service used for comparing protein 3D structures	http://ekhidna2.biocenter.helsinki.fi/dali/	Holm and Rosenström (2010)
VAST+	V ector A lignment S earch T ool, web-based tool for comparing 3D structure against all structures in the Molecular Modelling Database (MMDB), NCBI	https://structure.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html	Madej et al. (2013)
CE	A method for comparing and aligning protein structures	http://source.rcsb.org/ceHome.jsp	Shindyalov and Bourne (1998)
<i>7. Other databases</i>			
PTM-SD	Posttranslational modification database	http://www.dsimb.inserm.fr/dsimb_tools/PTM-SD/	Craveur et al. (2014)
PED3	Protein Ensemble Database The database of conformational ensembles describing flexible proteins	http://pedb.vib.be/	Varadi and Tompa (2015)
GFDB	Glycan Fragment Database (GFDB), identifying PDB structures with biologically relevant carbohydrate moieties and classifying PDB glycan structures based on their primary sequence and glycosidic linkage	http://www.glycanstructure.org/	Jo and Im (2012)
ChEBI	C hemical E ntities of B iological I nterest (ChEBI), a database focused on “small” chemical compounds	https://www.ebi.ac.uk/chebi/	Hastings et al. (2015)
ChEMBL	ChEMBL is a database of bioactive drug-like small molecules	https://www.ebi.ac.uk/chembl/	Gaulton et al. (2016)

Table 4.2 Experimental methods used for determination of macromolecule 3D structures

	X-Ray crystallography	Nuclear magnetic resonance	Cryo-EM
Experimental steps	<ol style="list-style-type: none"> 1. X-rays are scattered by electrons in the atoms of crystal. 2. Then recorded on a detector, e.g., CCDS. 3. Phase estimation and calculation of electron density map. 4. Fit primary sequence to electron density map (model). 5. Model refinement. 6. Deposition in PDB 	<ol style="list-style-type: none"> 1. Molecules absorb radiofrequency radiation held in a strong magnetic field. 2. Resonance frequency detection influenced by chemical environment. 3. Collection of conformational interatomic distance constraints. 4. Calculation of the 3D structure. 5. Deposition in PDB 	<ol style="list-style-type: none"> 1. Sample is vitrified at liquid nitrogen temp. 2. High-energy electron beam passes through it under high vacuum. 3. Image is produced when transmitted electrons are projected to a detector 4. Structure determination
Specimen	Crystals	Solution	Vitrified solution ^a
Protein size	Wide range	Below 40–50 KDa	>150 KDa
Contribution ^b	>89% of PDB entries	> 9% of PDB entries	>1% of PDB entries
Resolution	Higher resolution	High resolution	Significantly low >3.5 Å
Advantages	Well-developed Accurate, easy for model building	Provide dynamic information	Easy sample preparation Samples in its native environment
Disadvantages	Crystallization step Slow process	High purity sample is required Less precise than X-ray Intensive computational simulations	Cost Mainly for large molecules and assemblies

^aA vitrified solution is the solidification of a liquid into a noncrystalline or amorphous solid known as glass

The determination of the 3D structure for biological macromolecules is done by four fundamental techniques arranged in terms of familiarity and contribution as follows: X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, cryo-electron microscopy (Cryo-EM), and neutron diffraction. Table 4.2 summarizes the experimental steps adopted in the first three techniques and shows main advantages and disadvantages of these techniques. Although these techniques are viable and inestimable, they cannot build an atomic structure model from scratch without former knowledge of the proteins' chemical and physical properties and the proteins' primary sequences.

4.1.1 X-Ray Crystallography

X-ray protein crystallography is a branch of science that plays a vital role in many aspects including the determination of the 3D structure of proteins. Proteins' 3D structure determination enables us to perceive the relationships between the structure and the function of these molecules and characterizes drug targets such as G-coupled protein receptors (Rosenbaum et al. 2009), 3D structures of enzymes, DNA structure, and others. In 1895, Wilhelm Roentgen eternalized his name by discovering a new unknown type of rays that has a shorter wavelength than the UV rays; he named it X-rays. In 1912 Max von Laue demonstrated that X-rays can be diffracted upon interacting with a crystalline material. The following year, Bragg, the father, and his son, could solve a very challenging step in using X-rays for structure determination that was known as the phase problem; they succeeded in paving the way to use X-ray diffraction to know the 3D structure of a crystalline material. According to the current status, X-ray protein crystallography can be summarized in two main successive steps:

4.1.2 Crystal Formation

The X-ray crystallography experiment is based on shooting a protein crystal with X-rays. The process of getting crystals can be a cumbersome task since it is somehow a trial-and-error rather than systematic experiment. The process starts with obtaining a protein sample in high concentration. This step is done nowadays using different techniques of recombinant DNA technology. It is noteworthy to mention that the advancement in DNA synthesis has facilitated the process of gene cloning and expression. Advancement in genetics has not only facilitated the synthesis of genetic sequences at very reasonable cost but also assisted in controlling the gene expression by manipulating the molecular regulatory elements in the host cells (e.g., *Escherichia coli*, *Pichia*, or mammalian cells). Aided by different DNA techniques, the gene of interest can be overexpressed in suitable expression host to yield the target protein in a very good yield. Taking advantage of DNA recombinant technology, it is possible to add tags to the overexpressed proteins that will help in the purification steps (e.g., multi-histidine residues to the overexpressed protein to aid in metal affinity chromatography or SUMO tag that helps in an expression of the protein in a good yield) (Gaber et al. 2016). The overexpressed protein will undergo a process of purification until it is obtained in a high purity as judged by SDS-PAGE analysis. Afterward, a concentrated protein solution will be subjected to a crystal formation experiment. In such experiment, the concentrated protein solution will be exposed to different buffer solutions with different additives such as ethylene

glycol; the process is run in a miniaturized setting that allows testing hundreds of crystallization conditions in a short time and in an automated manner. The appearance of crystals in any of the tested conditions will be considered a positive hit that will lead to picking this specific condition and pursuing with the condition to reach a big crystal size of the protein. It is worth mentioning that membrane proteins are among the very difficult protein types to be crystallized. The difficulty comes from different reasons such as flexibility issues, instability, usage of detergent for extraction from cell membrane, purification, crystallization, data collection, and structure solution (Carpenter et al. 2008; Wlodawer et al. 2008).

4.1.3 *Structure Determination*

A special facility named synchrotron is used in the process of X-ray shooting. These facilities are located mostly in Europe, the USA, Japan, and Australia, for example, in Grenoble, France, and Lund, Sweden. The synchrotron is big laboratories that accelerate electrons to generate X-rays. The crystals obtained from the crystallization process are kept frozen in liquid nitrogen to protect them from destruction upon exposure to the high-energy rays. Special types of detectors collect the diffraction patterns obtained from the process of crystal exposure to the X-ray. These detectors have witnessed continuous development in order to facilitate the data collection process. The obtained data are then subjected to what is known data reduction in order to reduce the number of data obtained. Eventually, the data obtained will lead to what is known as electron density map which can be described as an in silico representation of a 3D shape of the protein revealed from the X-ray shooting experiment. The electron density map can be figured numerically by Fourier transformation (Wlodawer et al. 2008). The following step is to fit the protein primary amino acid sequence into the obtained electron density map providing the preliminary 3D model; this was a challenging task; however a plethora of programs are created to alleviate this issue; the most common one is COOT (Emsley et al. 2010). COOT is a widely used molecular graphics program for model building and biological molecule validation. It unveils atomic models and electron density maps and permits the manipulations of built models. Moreover, COOT supplies access to numerous validation and refinement tools. Validation of the preliminary model is vital before depositing final structure model into the PDB as a misinterpretation of data is liable. Many programs can help with this issue like PROCHECK. In addition, attempts to re-evaluate structures after deposition into PDB have been spotted; PDB-REDO is a good example of such efforts which can re-refine formerly deposited structures (Joosten et al. 2010).

4.2 Macromolecular Structural Databases

4.2.1 Protein Data Bank *wwPDB*

The Worldwide Protein Data Bank abbreviated as wwPDB (www.wwpdb.org) is the central organization that takes the responsibility to maintain and archive the 3D structural information of biomacromolecules. wwPDB stores 141,150 records of 3D structures (updated April 2018).

The wwPDB is composed of four partners:

- (i) Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) (Berman et al. 2000)
- (ii) Protein Data Bank in Europe (PDBe) (Velankar et al. 2010)
- (iii) Protein Data Bank Japan (PDBj) (Kinjo et al. 2016)
- (iv) Biological Magnetic Resonance Data Bank (BMRB) (Markley et al. 2008)

4.2.1.1 RCSB PDB

The RCSB Protein Data Bank (RCSB PDB, <http://www.rcsb.org>) is the US partner of wwPDB, and it presents the PDB archive in an organized and easy way to explore. The PDB archive is accessed by the public and serves diverse disciplines that encompass agricultural, pharmaceutical, and biotechnological applications. It is worth mentioning that a majority of PDB users are of limited expertise in structural biology. The design of the RCSB PDB webpage allows easy navigation and provides different options to find the structure of interest, facilitate in finding similar structures, and jump to related contents in different databases. Figure 4.1 shows a screenshot of RCSB PDB webpage viewing the accession code 2BH9. The page is organized into sections that include different types of information as indicated briefly as follows:

1. The front of the page shows the accession code 2BH9 and information about the authors and the deposition date.
2. The right corner contains a hyperlink to downloadable structural files of the 2BH9 as PDB file extension in addition to other types such as PDBx/mmCIF files. The typical form of storing 3D structure information is PDB file format. These files are typically opened with specific molecular visualization software such as PyMOL or YASARA (DeLano 2002; Krieger and Vriend 2014). However, the file can also be opened and edited – though is not advised for novice users – with text editor software programs such as Notepad or Microsoft Word. Figure 4.2 shows the PDB file for 2BH9 entry as an example; the file lists all the atoms present in the macromolecule (protein) and its coordinates as *X*, *Y*, and *Z*. A typical PDB file includes a header that gives a summary of the protein in terms of its source, author details, and the experimental techniques used. Since the size of

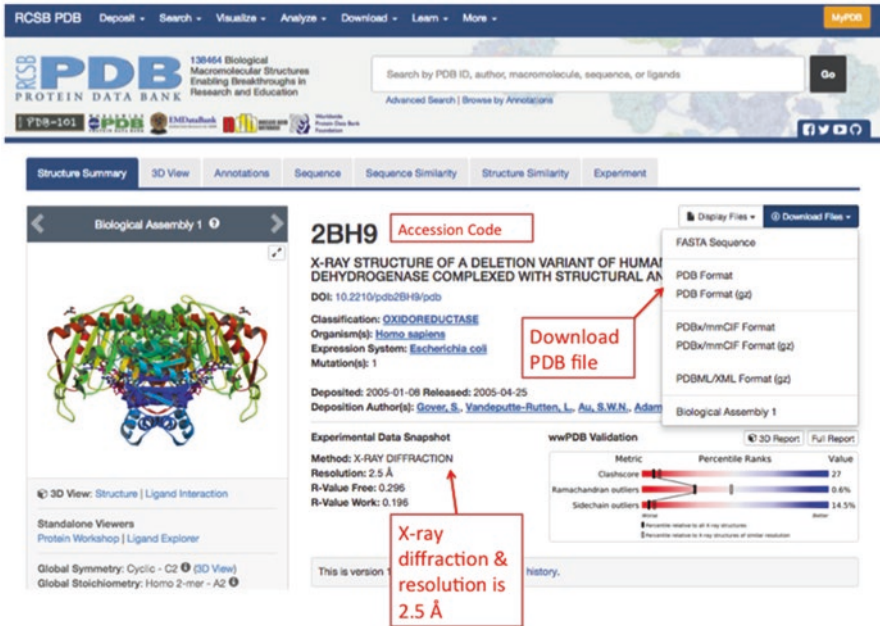


Fig. 4.1 A screenshot of PDB webpage interface for an oxidoreductase protein, deposited under the accession code (2BH9), structure determined by X-ray diffraction technique at a resolution of 2.5 Å. The source organism is *Homo sapiens* and overexpressed in *E. coli*. It also provides different types of downloadable file formats for the user to choose from, e.g., FASTA sequence, PDB, and mmCIF file formats

- 3D structure information is too big in few cases like virus capsid, a new file format – PDBx/mmCIF – is introduced to accommodate such large files.
- Information about the peer-reviewed publications linked to 2BH9 and the citation information.
 - Macromolecule section that shows the CATH classification of 2BH9 and the accession code of 2BH9 at UniProt database.
 - Experimental data snapshot: this section is devoted to the X-ray crystallography experiment and the statistical data revealing the resolution of the structure. In case of 2BH9, the structure was determined at a resolution of 2.5 Å, which is not a very good resolution. Resolution refers to the quality of the experimental data generated by X-ray crystallography. High-resolution structures will be determined at values of less than 1.5 Å or so; this level of accuracy of determining the atomic positions is high. Conversely, at a resolution of 3 Å or higher, the structure shape as global will be inferred; however the accurate positioning of the individual atoms is poor.

PDBe Protein Data Bank in Europe, (<http://www.ebi.ac.uk/pdbe/>) is the European equivalent to RCSB PDB. The PDBe home page provides an organized structure to

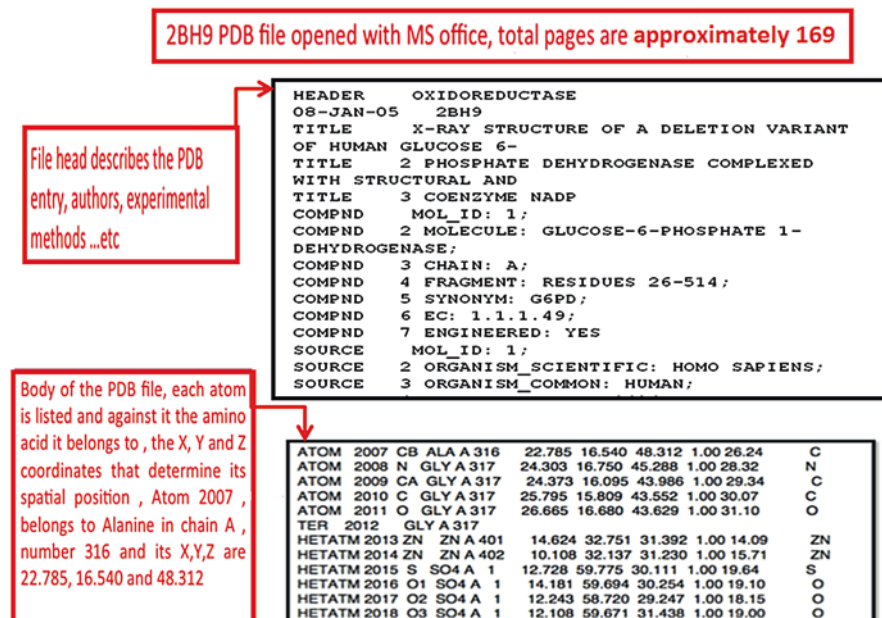


Fig. 4.2 PDB file format of the entry 2BH9 as an example, opened with Microsoft Word. A PDB file provides a full description of the entry such as a list of protein atoms and their 3D arrangement in space. For 2BH9 the header section provides information about the entry citation, authors, source of enzyme, and the experimental technique used in solving the structure. The file body provides information about the protein's atoms; each atom is listed opposite the amino acid it belongs to; moreover, it provides data about X, Y, and Z coordinates that determine its spatial position. For instance, atom number 2007 belongs to alanine in chain A number 316, and its X, Y, and Z coordinates are 22.785, 16.540, and 48.312, respectively

ease the browsing and exploration of the content. The tab (PDBe services) allows the access to categorize resources according to the user's interest and background; these categorized tabs are structural biologists, bioinformaticians, life scientists, students and teachers, medicinal chemists, journal editors and referees, and all services tab. For example, a good training and educational material are available under the PDBe training tab. Among the popular services that are provided by PDBe is FASTA protein sequence search that enables using protein sequence in the searching box. PDBeFold is a tool that finds similar structures starting from a PDB accession code as a query entry or via uploading a PDB file. To address the challenge of slow networking time, PDBe has developed a customized server named CoordinateServer that enables extraction of specific data for a given structure providing an advantage of high-speed exploration of PDB files and reduces the limitation of network file transfers. The server can provide several types of data extraction options such as finding residues interacting with a certain ligand and others. The server can be accessed via the link www.ebi.ac.uk/pdbe/coordinates/. PDBe has developed its own molecular visualization software LiteMole 3D viewer. The tool is

The screenshot displays the PDBE search results for entry 2BH9. The main title is "X-RAY STRUCTURE OF A DELETION VARIANT OF HUMAN GLUCOSE 6-PHOSPHATE DEHYDROGENASE COMPLEXED WITH STRUCTURAL AND COENZYME NADP". The source organism is *Homo sapiens* and the assembly composition is "protein only structure". Interacting compounds are listed as GOL and NAP. The experimental method is X-ray diffraction with a resolution of 2.5 Å. The release date is 25 Apr 2005. The page also shows a 3D molecular model of the protein structure.

Fig. 4.3 A screenshot of Protein Data Bank in Europe (PDBE) webpage interface of 2BH9 entry; the structure determined by X-ray diffraction technique at a resolution of 2.5 Å. In addition to that, binding ligands NAP (nicotinamide adenine dinucleotide phosphate) and GOL (glycerol); literature; the source organism – *Homo sapiens* – and protein assembly composition are also provided besides other buttons to access a plethora of information about the entry, e.g., binding ligands and protein family

compatible with many Internet browsers. The tool is WebGL-based viewer with too little memory footprint. PDBE has also developed a server that enables users to take part in developing their own search queries to meet their needs. The server is known as RESTful application programming interface (API) (Representational State Transfer) and is accessed via pdbe.org/api. Figure 4.3 shows a screenshot of PDBE webpage presenting search results for the entry 2BH9 including the experimental methods, the source organism, the assembly composition, and the interacting compounds (ligands). Additional details are available via other tabs such as macromolecules, compounds, and protein families.

PDBj The Protein Data Bank Japan (<https://pdbj.org/>), is one of the consortium members of wwPDB; the database is continuously updated to meet the user requirements with a focus on Asian and Middle Eastern users. The database offers a bunch of tools and services that assist the analysis and interpretation of structural data. These services include PDB deposition via an updated tool that supports X-ray, NMR, and EM structures. Group deposition is also available where a group ID is given to a set of structures that are related to each other and have been deposited at

the same time. PDBj also provides a tool for easy exploration of the PDB files via PDBj Mine, a tool for searching PDB using either accession codes, keywords, or via the advanced search function. Sequence-based structural alignment is also available via the tool known as SeSAW. The tool allows annotation of the conserved sequences and structural motifs found in the query proteins. eF-seek is a relatively new tool at PDBj that searches similar PDB files with a focus on the ligand binding sites. Omokage is another web-based tool for searching three-dimensional density maps and atomic models, with a focus on global shape similarities. ProMode Elastic database allows inspection of the PDB files regarding the dynamic rather than the static status. The database provides dynamic analysis for the PDB structures, and animations can be generated for PDB structure. PDBj has also developed its own molecular visualization graphic software known as Molmil that enables fast and enhanced graphics and is compatible with JavaScript and WebGL. Figure 4.4 shows a screenshot of PDBj showing summary for the entry 2BH9 including information about the related 3D structure 1QKI, functional keywords, and biological source; also other buttons are found for structural details, experimental details, functional details, sequence neighbor, history, and downloads. In the right side, download format options are available and structure view asymmetric unit.

The screenshot shows the PDBj website for entry 2BH9. At the top, there is a search bar and navigation links in multiple languages. The main navigation menu includes buttons for Summary, Structural details, Experimental details, Functional details, Sequence Neighbor, History, and Downloads. The entry title is 'X-RAY STRUCTURE OF A DELETION VARIANT OF HUMAN GLUCOSE 6-PHOSPHATE DEHYDROGENASE COMPLEXED WITH STRUCTURAL AND COENZYME NADP'. Below the title, there is a 'Summary for 2BH9' section with a table of related information.

Related	1QKI
Descriptor	GLUCOSE-6-PHOSPHATE 1-DEHYDROGENASE, NADP NICOTINAMIDE-ADENINE-DINUCLEOTIDE PHOSPHATE, GLYCEROL, ... (4 entities in total)
Functional Keywords	oxidoreductase, oxidoreductase (cho[n](d)-nadp), carbohydrate metabolism, glucose metabolism
Biological source	HOMO SAPIENS (HUMAN)

On the right side, there is a 'Downloads' section with options for Sequence (fasta), PDB/mmCIF, PDBML format (no-atom), PDB format (full), and Validation report (PDF). Below this is a 'Structures' section with a 'View Asymmetric Unit' button. At the bottom right, there is a 3D molecular model of the protein structure.

Fig. 4.4 A screenshot of Protein Data Bank Japan (PDBj) webpage interface shows detailed informative data about the entry 2BH9, represented in the main navigation menu containing many buttons which provide information about the entry's summary, structural details, experimental details, and functional details. Moreover, it also provides different types of downloadable file formats such as FASTA sequence, PDB format, PDBx/mmCIF file formats, and others

BMRB Biological Magnetic Resonance Data Bank, aims to archive and annotate the nuclear magnetic resonance data obtained from macromolecules and their metabolites. The database is unique and provides an important repository for NMR data for peptides, proteins, and nucleic acids. The current content (May 2018) of BMRB archive includes 11,628 entries of proteins/peptides, 398 entries of DNA, and 345 entries of RNA (Fig. 4.5). BMRB can be accessed via the URL <http://www.bmrb.wisc.edu/>, which is sponsored by the University of Wisconsin-Madison, the National Library of Medicine, and National Institutes of Health. The website is organized into different tabs such as search archive, validation tools, deposit data, NMR statistics, programmers' corner, spectroscopists' corner, educational outreach, etc. (Ulrich et al. 2007).

NCBI Structure Resources The NCBI devotes one of its databases to the structure information. NCBI provides ENTREZ search function that allows searching keywords all over its databases including the structure database. The structure database is available in the link <https://www.ncbi.nlm.nih.gov/structure/>, accessed on March 2018. Figure 4.6 is a screenshot of structure summary MMDB webpage using the PDB ID 2BH9 (MMDB ID 33089) as an example. The page displays information about the experimental method, resolution, source organism, similar structures, and biological unit (molecular graphic, interactions) for 2BH9.

Biological Magnetic Resonance Data Bank
A Repository for Data from NMR Spectroscopy on Proteins, Peptides, Nucleic Acids, and other Biomolecules

Instant entry access: Searches all entries on many criteria: Title, Author, Entry, Organism, Database code, etc. Hover over a result for more information.

BMRB Query Grid Interface

Current Content of the BMRB Archive

BMRB entry list (12211)

Clicking on a link in one of the boxes in the above table will take you to a BMRB entry listings for the type of biopolymer and type of data represented by the location of the box in the grid. Values in parentheses indicate the number of entries for that category.

Data Type	Proteins/Peptides (11628)	DNA (398)	RNA (345)
All Chemical Shifts	7978455 (11328)	54065 (333)	75825 (293)
1H Chemical Shifts	4078255 (10995)	49568 (329)	47863 (292)
13C Chemical Shifts	2980117 (8251)	3419 (50)	23563 (181)
15N Chemical Shifts	823793 (8529)	121 (16)	3815 (119)
31P Chemical Shifts	-	1135 (73)	727 (55)
Other Chemical Shifts	-	-	-
Coupling Constants	28147 (363)	131 (5)	-
Dipolar Couplings	54191 (123)	-	-
T1 Values	37668 (248)	-	-
T2 Values	39226 (245)	-	-
Heteronuclear NOE Values	35789 (244)	-	-
S2 Values	15163 (93)	-	-
H-Exchange Rates	1561 (19)	-	-
H-Exchange Protection Factors	727 (19)	-	-
D/H-Fractionation Factors	-	-	-
pKa Values	-	-	-
3D Structure Entries	(1)	-	-

Contact bmrbhelp@bmrb.wisc.edu if you have any questions about this site
Copyright © The Board of Regents of the University of Wisconsin System.
Last Modified: Saturday, 05-May-2018 00:04:35 CDT

Funded by NIGMS

Fig. 4.5 A screenshot of Biological Magnetic Resonance Data Bank (BMRB) webpage interface shows the recent content of the three major classes of biomacromolecules' structures, determined by nuclear magnetic resonance spectroscopy, 11,628 protein/peptide entries, 398 DNA entries, and 345 RNA entries, and the derived information: coupling constants, chemical shifts, dipolar coupling, etc

NCBI
National Center for
Biotechnology Information

Structure Summary
MMDB

Enter PDB ID or MMDB ID

2BH9: X-Ray Structure Of A Deletion Variant Of Human Glucose 6- Phosphate Dehydrogenase Complexed With Structural And Coenzyme NADp

Citation:

Human glucose-6-phosphate dehydrogenase: the crystal structure reveals a structural NADP(+) molecule and provides insights into enzyme deficiency
Au SW, Gover S, Lam VM, Adams MJ
Structure (2000) 8 p.293-303

Abstract

BACKGROUND: Glucose-6-phosphate dehydrogenase (G6PD) catalyzes the first committed step in the pentose phosphate pathway; the generation of NADPH by this enzyme is essential for protection against oxidative stress. The human enzyme

PDB ID:	2BH9 <input type="button" value="Download"/>
MMDB ID:	33089 <input type="button" value=""/>
PDB Deposition Date:	2005/1/7 <input type="button" value=""/>
Updated in MMDB:	2007/10 <input type="button" value=""/>
Experimental Method:	x-ray diffraction <input type="button" value=""/>
Resolution:	2.5 Å <input type="button" value=""/>
Source Organism:	Homo sapiens <input type="button" value=""/>
Similar Structures:	VAST+ <input type="button" value=""/>
<input type="button" value="Download sequence data"/> <input type="button" value=""/>	

Biological Unit for 2BH9: dimeric; determined by author and by software (PQS)

Molecular Graphic

Interactions

Fig. 4.6 A screenshot of Molecular Modeling Database (MMDB) webpage interface of 2BH9, MMDB ID (33089). The structure is resolved by X-ray diffraction technique at a resolution of 2.5 Å, and the source organism is *Homo sapiens*. Besides, it provides a chemical graph, links to literature, and compact structures (3D structure domains) that help with identifying similar structures

4.3 PDBsum: Structural Summaries of PDB Entries

PDBsum available at <https://www.ebi.ac.uk/pdbsum> is an atlas of proteins and web server that helps to present the PDB entries in a visualized form. It was developed at the University College London (UCL) in 1995 and is aimed to provide a largely graphic compendium of the proteins and their complexes (Laskowski 2007; Babajan et al. 2011). The server can be accessed freely and is maintained by Laskowski and collaborators at the European Bioinformatics Institute (EBI) (Laskowski et al. 2018). PDBsum provides many different analytical tools for the content of the protein structure including the ligand interaction, protein-protein interaction, and CATH classification. The 3D structures are viewed interactively in PyMOL and RasMol, and users have the ability to upload their own PDB files – could be a homology model – and get them analyzed. Figure 4.7 illustrates some of the pictorial analyses presented by PDBsum. The example given is for PDB entry 2BH9 (G6PD-human) solved by X-ray crystallography at 2.5 Å resolution. The page shows different sections among which the 3D structures are presented interactively using molecular visualization JavaScript viewer called 3Dmol.js. This generated

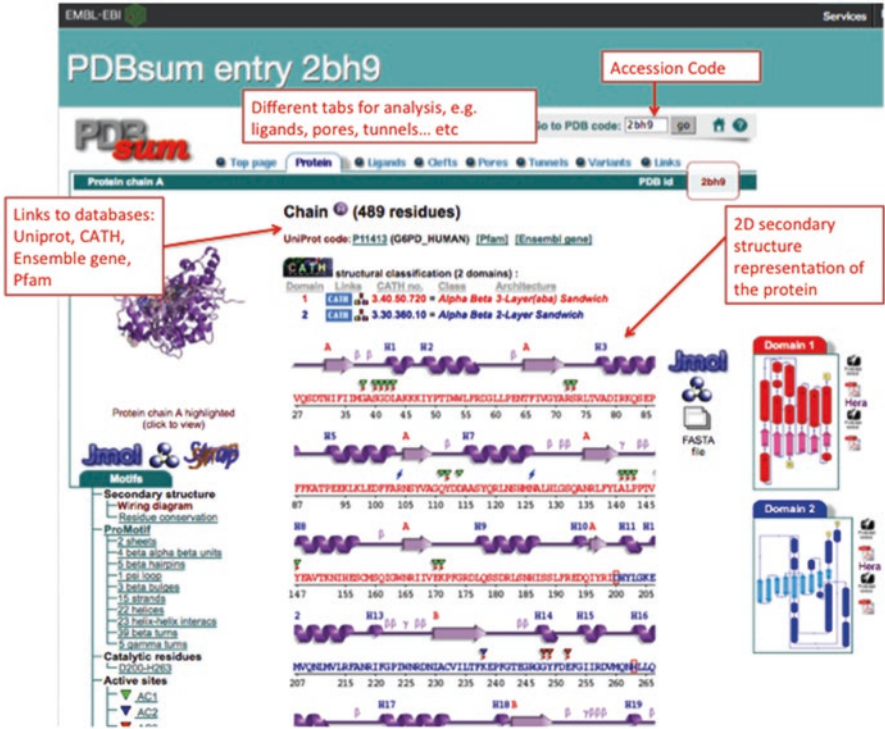


Fig. 4.7 A screenshot of PDBsum webpage interface of 2BH9 entry; a 2D secondary structure representation is shown in the figure. Tabs for protein-protein interactions, ligands, pores, tunnels, and others are seen in the figure. Hyperlinks to r databases like UniProt, Pfam, and Ensembl gene are also provided

image automatically gives only a rough idea of the sizes and locations of the clefts. Using the RasMol or Jmol options on the clefts tab, an idea about the clefts found in the structure can be obtained. PDBsum webpage also hosts useful links to databases and servers such as:

1. EC-PDB, Enzyme Structure Database, database includes approximately 73,000 PDB enzyme structures. The database classifies the entries according to the Enzyme Commission (EC) as EC1, EC2, EC3, EC4, EC5, and EC6. EC3 – the hydrolase family – is the highest represented family among others in this database including over 27,000 PDB structures.
2. Drug port is the second server which identifies all “drug targets” in the PDB and any drug fragments that exist as ligands in PDB structures. The server lists all the drugs in alphabetical order; therefore, for example, if you are looking for acetaminophen, you will find it under the alphabet A in the list, and visiting its page will show the information of the protein targets of this specific drug and hyperlinks to other related resources such as DrugBank and others.

3. ProFunc server: the server aims to help in the identification of protein of related biochemical function based on the 3D structure. The algorithm of ProFunc uses information such as the active site, fold matching, residue conservation, and surface analysis to do the task. The server allows to look for existing PDB file or to upload custom PDB file (Laskowski et al. 2018).
4. SAS, sequence annotated by the structure, is a tool by PDBsum; the tool allows multiple sequence alignment of a query protein that entered in different forms such as FASTA sequence, PDB accession code, PDB file, or UniProt accession code. The obtained multiple alignments can be color-coded according to different criteria, such as the secondary structure assignment, ligand binding site, and number of hydrogen bonds to ligands or residue similarity. The alignment can be adjusted according to the user needs using selection and sequence similarity filters.

4.4 sc-PDB: A 3D Database of Ligandable Binding Sites

The protein-ligand interaction is very important in determining the critical amino acids in the protein structure that interact with ligands, and based on this information, designing new ligands (drugs) is possible. The sc-PDB database archives and illustrates the ligandable binding sites found in protein structures that are listed in the PDB repository. The database was launched in 2004 and is accessible at <http://bioinfopharma.u-strasbg.fr/scPDB/>. The Sc-PDB provides specialized structure files that serve the need to do receptor-ligand docking studies. Currently, the sc-PDB stores 16,034 entries (binding sites) extracted from 4782 unique proteins and 6326 exclusive ligands. The sc-PDB database provides annotated druggable binding sites, the coordinates for protein-ligand complexes, and the physicochemical and geometrical properties of the ligands. It also provides a chemical description of ligands and functional explanation of the proteins. Metal ions are not included in sc-PDB, and the ligands included are classified into four main categories: (i) nucleotides of size <4 bases, (ii) peptides <9 amino acids, (iii) cofactors, and (iv) organic compounds. The binding site can be defined as the protein residues (including amino acids, cofactors, and important metal ions) that are in contact with one atom of the ligand within a distance of 6.5 Å. The sc-PDB is very useful in drug design tasks since it can predict receptors for any ligand and it can analyze different structural cavities and establish the interacting points between a ligand and the active site of the receptor (Desaphy et al. 2014; Kellenberger et al. 2006). Ligands can be searched using the chemical structure draw applet provided by ChemAxon. Figure 4.8 is a screenshot of the sc-PDB webpage showing the total number of entries (16034) including 4782 proteins and 6326 ligands. The database home page shows four buttons: ligand, protein, binding mode, and binding site. The database archive can be searched using the search anything box, PDB ID box, or protein UniProt accession code.

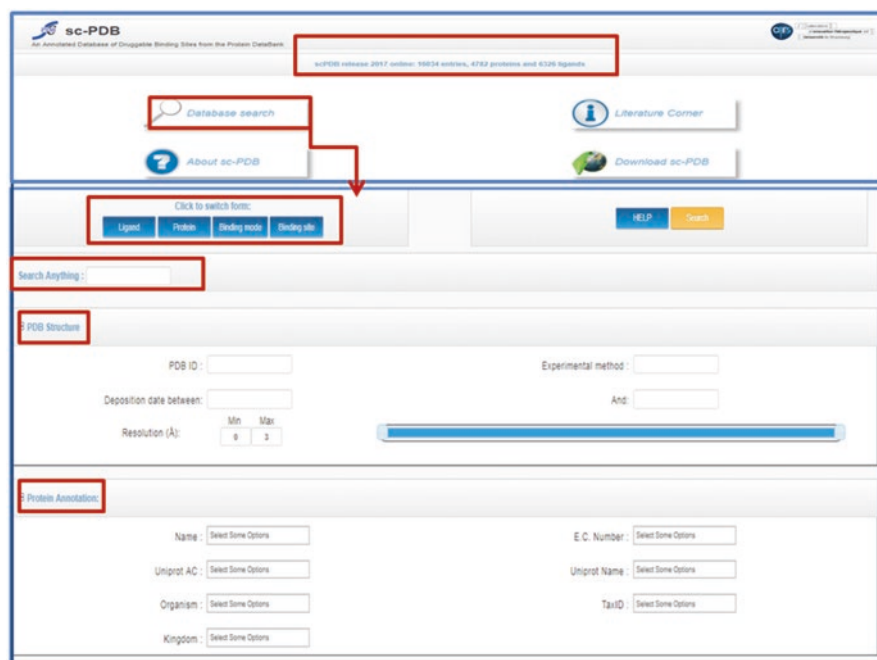


Fig. 4.8 A screenshot of sc-PDB webpage interface. It shows (16034) three-dimensional structures of binding sites found in the Protein Data Bank (PDB) and includes (4782) unique proteins and (6326) unique ligands. In addition, it provides the main navigation window for the user to navigate and switch views directly (ligand, protein, binding mode, and binding site)

4.5 PDBTM: Protein Data Bank of Transmembrane Proteins

Membrane proteins account for 20–30% of the all human proteins which participate in vital cellular processes and enzymatic reactions. Membrane proteins represent 60% of all druggable proteins in human (Yin and Flynn 2016). The experimental 3D structure determination of these proteins is difficult due to the complexity of obtaining soluble expressed proteins. Since the publication of the first membrane protein 3D structure in 1985, the number of membrane proteins in wwPDB is increasing slowly but steadily. Still, the current representation of the membrane proteins in PDB is low. There was a need to have specialized databases for membrane proteins. The PDBTM database is the first up-to-date and inclusive TM protein consisting of a list of PDB files of transmembrane proteins (Kozma et al. 2012). The database was launched in 2004 and is available at <http://pdbtm.enzim.hu>; PDBTM archives more than 3000 transmembrane proteins; most of them have the well-known alpha-helical structures. PDBTM is utilizing a special algorithm named TMDet to find transmembrane proteins found in the PDB based on the structural information. The algorithm is also able to determine the spatial arrangement of these proteins inside the lipid bilayer. PDBTM

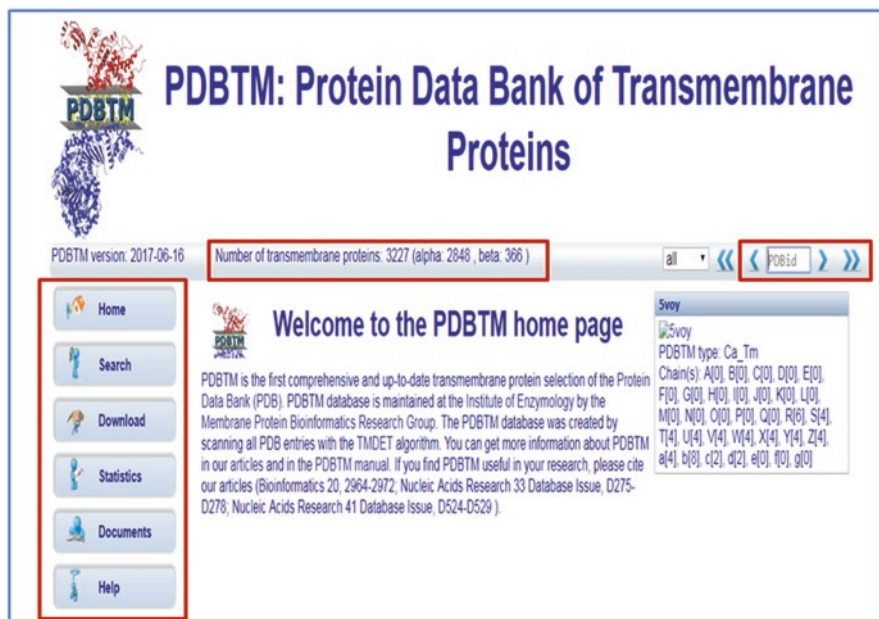


Fig. 4.9 A screenshot of Protein Data Bank of Transmembrane Protein (PDBTM). It shows a number of transmembrane proteins deposited in PDBTM; total number is 3227 entries: 2848 alpha structured and 366 beta structured

website allows to browse its content by the type of the membranes (alpha or beta structures), and it also permits to download datasets of TM protein structures. Figure 4.9 shows the home page of PDBTM and the number of transmembrane proteins that is archived until May 2018 (a total of 3227, including 2848 alpha structure and 366 beta structure). The search field using PDB ID exists in the right side, while the left side includes six vertical tabs (home, search, download, statistics, documents, and help).

4.6 CATH Database

CATH (Class, Architecture, Topology, Homology) database classifies the protein domains according to the amino acid sequence and the structural and the functional properties. CATH provides a big deal of help for researchers with proteins that have insignificant similarity in sequences yet can be functionally and structurally related. CATH is also a valuable destination for both bioinformatician and biologists. Inexperienced users benefit from the user-friendly web interface; on the other hand, bioinformaticians seeking for analysis of a huge number of domains can find complete downloadable datasets. Therefore CATH has the potentials to be a really valuable and promising recourse. In CATH, domains are classified hierarchically into

four levels named as class (C), architecture (A), topology (T), and homologous superfamily (H), hence giving the acronym CATH (Knudsen and Wiuf 2010):

- (i) C level: categorize domains into four main groups according to secondary structures as alpha mainly, beta mainly, α - β mixed, and finally category group domains with few alpha and beta structures.
- (ii) A level: categorize domains by the general orientation of the secondary structures.
- (iii) T level: categorization depends upon the connectivity of secondary structures.
- (iv) H level: categorization depends upon a combination of sequence similarity and structural similarity.

Exploration of the contents of the databases can also be done via different links given in the web server, for example, (1) searching by domain ID or keywords, (2) searching by the sequence in FASTA format, and (3) exploring the database from the hierarchy top and download datasets. A list encompasses the names of all domains in CATH – along with their individual groupings – which is likewise accessible, and the amino acid sequences of all domains ordered in CATH are open for download in the FASTA file format (Knudsen and Wiuf 2010). Figure 4.10 is the search results for the PDB ID (2BH9); the figure shows the matching CATH superfamilies and the matching CATH domains.

The screenshot shows the CATH/Gene3D search interface. At the top, there is a navigation bar with 'Home', 'Search', 'Browse', 'Download', 'About', and 'Support'. A search bar contains '2BH9' and a 'Search' button. Below the search bar, there are three tabs: 'Search by Text or ID' (selected), 'Search by Sequence', and 'Search by Structure'. The 'Results' section is divided into two main parts: 'Matching CATH Superfamilies' and 'Matching CATH Domains'. The 'Matching CATH Superfamilies' section shows a 3D ribbon diagram of a protein structure and the following text: '3.30.360.10 Dihydrodipicolinate Reductase: domain 2'. The 'Matching CATH Domains' section shows a 3D ribbon diagram of a protein structure and the following text: '2bh9A01 PDB code 2bh9, chain A, domain 01'. On the right side, there is a 'Current Search Filters' section with '2BH9*' and a 'Filter by Keyword / CATH ID' section with a search input field. Below that, there is a 'Top Keywords' section with the text: '(nadp(+)), +/- 1-dehydrogenase 1.1.1.363 1.1.1.388, 6-phospho-6-phospho-d-glucono-1,5-lactone = a activity and binding biosynthetic branch catalyzes cellular of cholesterol cytoplasm cytoplasmic cytosol d-

Fig. 4.10 A screenshot of CATH/Gene3D webpage interface of the entry 2BH9. The websites provide different ways of search: text or ID, search by sequence, or search by the structure. In the current example, the screenshot shows the matching CATH superfamilies and domains related to 2BH9

CATH/Gene3D database is complementary to the original CATH database; it is available at <http://www.cathdb.info/>; it classifies 95 million protein domains into 6119 superfamilies (Dawson et al. 2016). CATH/Gene3D scans the protein sequence information found in UniProt database; it also classifies the structural domains found in the structural files in wwPDB. Annotation of the structure is created using hidden Markov models making use of the domain families deposited in CATH. Moreover, all information is downloadable in an XML file format, enabling users to perform a complex search at their computers (Yeats et al. 2006). Furthermore, Gene3D exploits the data in CATH to predict the position of structural domains on a host of protein sequences available at wwPDB which allows inclusion of informative annotations such as information, function, and residues of the active site. It also provides a broad prediction of globular domains in proteins (Dawson et al. 2016; Dawson et al. 2017).

4.7 SCOP (Structural Classification of Proteins) Database

Structural Classification of Protein (SCOP), available at <http://scop.mrc-lmb.cam.ac.uk/scop/>, is a database with a focus on structure and evolutionary classifications of proteins. SCOP adopts the following hierarchical scheme to classify protein structures:

- A. Family: similar protein structures are assembled into families based on two criteria that suggest a common evolutionary source; the first criterion is a similarity in protein sequence, and the second criterion is a similarity in structure and function.
- B. Superfamily: families whose proteins have little sequence similarity yet their function and structure imply typical evolutionary origin are clustered together in superfamilies.
- C. Common fold: protein families and subfamilies that have similar secondary structures and same topological associations are assigned to have a common fold.
- D. Class: the distinctive folds have been gathered into classes.

The majority of the folds are grouped into one of the five structural classes:

1. All $-\alpha$: structures that are basically formed of α -helices.
2. All $-\beta$: structures that are basically formed of β -sheets.
3. α/β : structures formed of α -helices and β -strands.
4. $\alpha + \beta$: structures formed of α -helices and β -strands are to a great extent segregated.
5. Multi-domain: structures with domains of various classes and for which no homologs are yet known.

SCOP is updated into the new version SCOP2, where improvements in the classification criteria were done. SCOP2 classification is based on four criteria, i.e., the protein type, the evolutionary analysis, the structure class, and the protein relation-

ships. The protein types indicate four possible types of proteins, i.e., membrane, soluble, fibrous, and intrinsically disordered proteins. The evolutionary analysis considers the classification of proteins according to the major evolutionary events that had have happened to certain protein class. The third criterion is the secondary structure arrangement of the protein as an efficient way in the classification of protein structures. The protein relationships are unique to SCOP2 compared to SCOP. The database is accessible via the link <http://scop2.mrc-lmb.cam.ac.uk/>. SCOP2 can be explored in two different ways: SCOP2-graph and SCOP2-browser. SCOP2-graph shows graphical representation for the database entries, while SCOP2-browser allows the exploration of the SCOP2 contents according to the four classification criteria mentioned above in addition to a possibility of keyword search. The SCOP2 additionally provides hyperlinks whenever possible to each entry archived to the external databases such as UniProt and PDB and the original SCOP record (Andreeva et al. 2007; Hubbard et al. 1997). Figure 4.11 is a screenshot of SCOP2-graph database webpage interface. It illustrates a hierarchal classification of protein domains by the structure and evolutionary relevance.

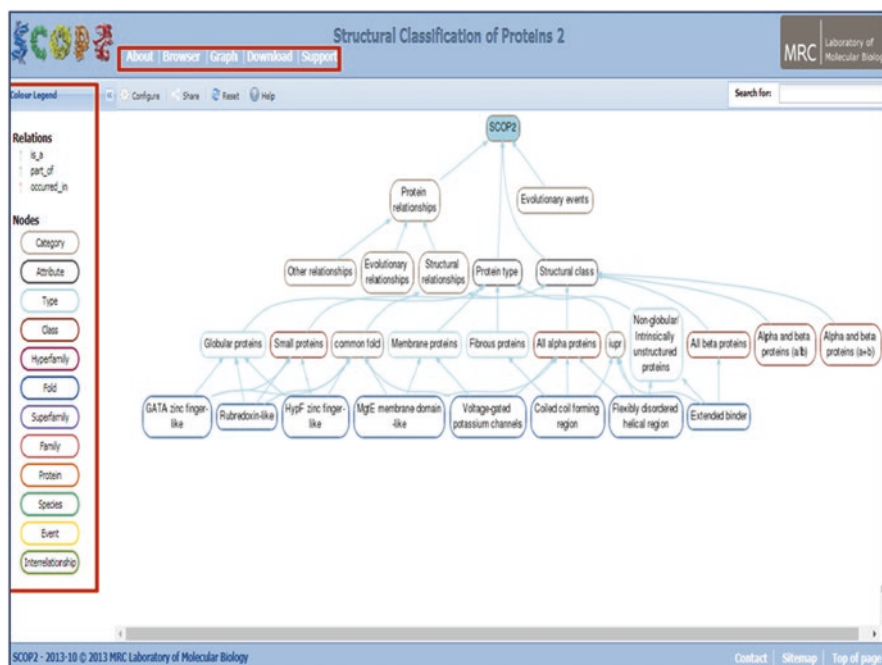


Fig. 4.11 A screenshot of SCOP2-graph database webpage interface. It illustrates a hierarchal classification of protein domains in accordance with the structure and evolutionary relevance; these relationships appear as compound node networks; also, it provides accessible links to the SCOP entries and hence provides a possibility for the users to compare both databases

4.8 Structure Comparison Servers

Finding homologous protein structure is very important in the area of structural bioinformatics. Therefore early efforts were carried out to devise algorithms for structural alignment; in 1960, Perutz et al. described the structure similarity of hemoglobin and myoglobin (Perutz et al. 1960). It is known that protein structures are more conserved compared to protein sequences; this is the base of evolutionary analysis of related protein structures. It is important to differentiate between two terms, i.e., structure superposition and structure alignment. Structure superposition refers to the spatial fitting of two structures that already have similar starting points – usually in the C-alpha backbone – which work as guiding points in the process of fitting these two structures over each other. The aim is to find the best match between the two structures as judged by the root-mean-square deviation (RMSD) value. RMSD is a measure of the average distance between atoms of two or more superimposed protein structures and is measured in angstrom. Structure alignment does not require prior information of equivalent spatial positions of two structures. However, the alignment algorithm tries to find structures between two 3D structures or more based on the 3D information. There are few clear reasons behind the effort for finding similar protein structures:

1. To help in structure classification and fold assignment
2. To aid the process of function identification, since similar protein structures can provide a wealth of information about the function of an unknown protein
3. To aid, in the process of homology or comparative modeling, the process of predicting protein 3D structure based on similarity to already known 3D structure
4. To aid in the tasks of protein engineering (Gaber 2016; Pavelka et al. 2009)

CATH and SCOP databases were used in the endeavors of finding similar structures based on detection of similar structural domains. Currently, some online servers and tools are used in finding homologous 3D structures of proteins; among these servers are:

1. Combinatorial Extension (CE) is a tool for aligning and comparing protein structures deposited into RCSB PDB (Shindyalov and Bourne 1998). CE is an indispensable part of identifying and annotating protein structures with unknown function. The comparison can be performed on a complete PDB or on structurally representative subsets of proteins. Also, it can be performed in two ways either using a structural representative subset of protein or on the full PDB records. The most direct task is to locate every single similar structure to a starting protein that exceeds 30 residues long and exists in the wwPDB. The superimposed structures can be visualized with programs such as RasMol and Protein Explorer (utilizing Chime) or in an exceptionally outlined Java applet Compare3D. The applet enables the user to investigate the two similarities and differences between the aligned structures both from a sequence and structure viewpoint. It is worth mentioning that the site is always subjected to modification and editing by the Bourne Laboratory staff to keep it up to date (Shindyalov and Bourne 2001).

- PDBeFold is an online server that is provided by EMBL-EBI (European Molecular Biology Laboratory-European Bioinformatics Institute). PDBeFold can be accessed from the PDBe webpage and is considered a structure alignment server that allows both pairwise or multiple 3D alignments. Searching homologous structures can be initiated by providing the PDB accession code.
- VAST+ is online server hosted by NCBI and is devoted to finding similar 3D structures; the server does not rely on sequence comparison; hence it can find 3D structures of too low sequence similarity. Figure 4.12 shows the interface of VAST+ using the PDB entry 2BH9.
- DALI web server was established in 2000 at Helsinki Lab; the server aims to compare 3D structures of proteins to those found in the Protein Data Bank. A new version of DALI known as DALI Lite has been released to do pairwise structural superimposition. Figure 4.13 shows a screenshot of DALI structure comparison server exemplified by a search using the entry 2BH9. DALI website displays nine horizontal tabs as follows: about, PDB search, PDB25, pairwise, all against all, gallery, references, statistics, and tutorial.

2BH9: X-Ray Structure Of A Deletion Variant Of Human Glucose 6- Phosphate Dehydrogenase Complexed With Structural And Coenzyme Nadp

Biological unit 1: dimeric
 Source organism: Homo sapiens
 Number of proteins: 2 (GLUCOSE-6-PHOSPHATE 1-DEHYDROGENASE ▼)
 Number of chemicals: 8 (Glycerol (4),Nadp Nicotinamide-adenine-dinucleo... ▼)

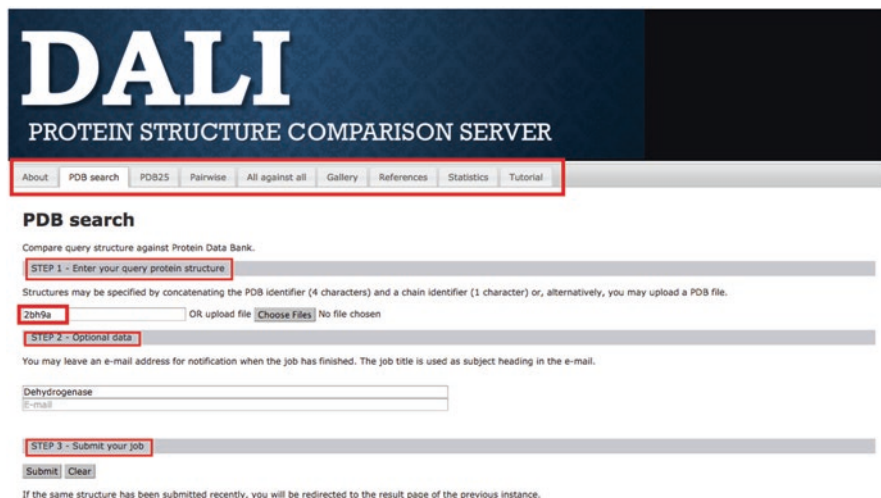
Similar Structures (2308)

Display filters

Showing 1 to 10 out of 2308 selected structures

PDB ID	Description	Taxonomy	Aligned Protein	RMSD	Aligned Residues	Sequence Identity
1QKI	X-Ray Structure Of Human Glucose 6-Phosphate Dehydrogenase (Variant Canton R459I) Complexed With Structural Nadp+	Homo sapiens	2	0.67Å	584	98%
SUKW	Crystal Structure Of Human Glucose 6-phosphate Dehydrogenase Mutant (a277c) Complexed With G6p	Homo sapiens	2	0.75Å	952	99%
2BHL	X-ray Structure Of Human Glucose-6-phosphate Dehydrogenase (deletion Variant) Complexed With Glucose-6-phosphate	Homo sapiens	2	1.03Å	603	100%
4LGV	X-ray Crystal Structure Of Glucose-6-phosphate 1-dehydrogenase From Mycobacterium Avium	Mycobacterium avium 104	2	1.15Å	357	34%

Fig. 4.12 A screenshot of VAST+ webpage interface of the entry 2BH9. It provides information about macromolecules that share similar three-dimensional structures. Concerning 2BH9, there is 2308 structure similar to it. It is worth noting that filters can be used to limit the number of matching molecules at will. The RMSD values shown indicate the structural similarity between the query 2BH9 and the retrieved hits; lower RMSD values indicate high structural similarity



DALI
PROTEIN STRUCTURE COMPARISON SERVER

About PDB search PDB25 Pairwise All against all Gallery References Statistics Tutorial

PDB search

Compare query structure against Protein Data Bank.

STEP 1 - Enter your query protein structure

Structures may be specified by concatenating the PDB Identifier (4 characters) and a chain identifier (1 character) or, alternatively, you may upload a PDB file.

2bh9a OR upload file No file chosen

STEP 2 - Optional data

You may leave an e-mail address for notification when the job has finished. The job title is used as subject heading in the e-mail.

Dehydrogenase
E-mail

STEP 3 - Submit your job

If the same structure has been submitted recently, you will be redirected to the result page of the previous instance.

Fig. 4.13 A screenshot of DALI server webpage interface and example input of the entry 2BH9 is shown. The website provides three different types of searches: PDB search, pairwise comparison, and all-against-all comparison which performs a database search comparing a query structure supplied by the user against the database of known structures (PDB) and returns the list of structural neighbors using the e-mail

4.9 Conclusion

Structural databases are providing essential information not only to the scientific community but also to the public. The content of such databases is a really precious information; precious is not just a metaphor; to explain, solving 1000 protein structures costs 150 million USD and the effort of 180 scientists (Ledford 2010). Fortunately, the advancement in the computational sciences allowed structural databases to be explored by both experts and novice users to navigate and easily extract the required information from its content. It is also very feasible to find related contents in the different database based on the interconnectedness between the different databases. The availability of such data allowed new generations of databases to evolve and to provide new layers of information that help in solving serious problems such as designing new drugs or engineering new proteins for different purposes.

References

- Acharya C, Kufareva I, Ilatovskiy AV, Abagyan R (2014) PeptiSite: a structural database of peptide binding sites in 4D. *Biochem Biophys Res Commun* 445(4):717–723
- An J, Totrov M, Abagyan R (2005) Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol Cell Proteomics* 4(6):752–761

- Andreeva A, Howorth D, Chandonia J-M, Brenner SE, Hubbard TJ, Chothia C, Murzin AG (2007) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res* 36(suppl_1):D419–D425
- Babajan B, Chaitanya M, Rajsekhar C, Gowsia D, Madhusudhana P, Naveen M et al (2011) Comprehensive structural and functional characterization of *Mycobacterium tuberculosis* UDP-NAG enolpyruvyl transferase (Mtb-MurA) and prediction of its accurate binding affinities with inhibitors. *Interdiscip Sci* 3(3):204–216. <https://doi.org/10.1007/s12539-011-0100-y>
- Bagchi A (2012) A brief overview of a few popular and important protein databases. *Computat Mol Biosci* 2(04):115
- Berman HM, Bhat TN, Bourne PE, Feng Z, Gilliland G, Weissig H, Westbrook J (2000) The Protein Data Bank and the challenge of structural genomics. *Nat Struct Mol Biol* 7:957–959
- Berman HM, Kleywegt GJ, Nakamura H, Markley JL (2012) The Protein Data Bank at 40: reflecting on the past to prepare for the future. *Structure* 20(3):391–396
- Carpenter EP, Beis K, Cameron AD, Iwata S (2008) Overcoming the challenges of membrane protein crystallography. *Curr Opin Struct Biol* 18(5):581–586
- Coimbatore Narayanan B, Westbrook J, Ghosh S, Sweeney B, Zirbel CL, Leontis NB, Berman HM (2013) The nucleic acid database: new features and capabilities. *Nucleic Acids Res* 42(D1):D114–D122
- Craveur P, Rebehmed J, de Brevern AG (2014) PTM-SD: a database of structurally resolved and annotated posttranslational modifications in proteins. *Database*:2014
- Dawson NL, Lewis TE, Das S, Lees JG, Lee D, Ashford P et al (2016) CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Res* 45(D1):D289–D295
- Dawson NL, Sillitoe I, Lees JG, Lam SD, Orengo CA (2017) CATH-Gene3d: generation of the resource and its use in obtaining structural and functional annotations for protein sequences. *Protein Bioinforma* 1558:79–110
- DeLano WL (2002) The PyMOL molecular graphics system. <http://pymol.org>
- Desaphy J, Bret G, Rognan D, Kellenberger E (2014) sc-PDB: a 3D-database of ligandable binding sites—10 years on. *Nucleic Acids Res* 43(D1):D399–D404
- Emsley P, Lohkamp B, Scott WG, Cowtan K (2010) Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* 66(4):486–501
- Gaber Y (2016) In-silico smart library design to engineer a xylosetolerant hexokinase variant. *Afr J Biotechnol* 15(21):910–916
- Gaber Y, Mekasha S, Vaaje-Kolstad G, Eijsink VG, Fraaije MW (2016) Characterization of a chitinase from the cellulolytic actinomycete *Thermobifida fusca*. *Biochim Biophys Acta* 1864(9):1253–1259
- Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, Motow P, Atkinson F, Bellis LJ, Cibrián-Uhalte E (2016) The ChEMBL database in 2017. *Nucleic Acids Res* 45(D1):D945–D954
- Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, Swainston N, Mendes P, Steinbeck C (2015) ChEBI in 2016: improved services and an expanding collection of metabolites. *Nucleic Acids Res* 44(D1):D1214–D1219
- Holm L, Rosenström P (2010) Dali server: conservation mapping in 3D. *Nucleic Acids Res* 38(suppl_2):W545–W549
- Hubbard TJ, Murzin AG, Brenner SE, Chothia C (1997) SCOP: a structural classification of proteins database. *Nucleic Acids Res* 25(1):236–239
- Jo S, Im W (2012) Glycan fragment database: a database of PDB-based glycan 3D structures. *Nucleic Acids Res* 41(D1):D470–D474
- Joosten RP, Te Beek TA, Krieger E, Hekkelman ML, Hooft RW, Schneider R et al (2010) A series of PDB related databases for everyday needs. *Nucleic Acids Res* 39(suppl_1):D411–D419
- Kellenberger E, Muller P, Schalon C, Bret G, Foata N, Rognan D (2006) sc-PDB: an annotated database of druggable binding sites from the Protein Data Bank. *J Chem Inf Model* 46(2):717–727
- Kinjo AR, Bekker G-J, Suzuki H, Tsuchiya Y, Kawabata T, Ikegawa Y, Nakamura H (2016) Protein Data Bank Japan (PDBj): updated user interfaces, resource description framework, analysis

- tools for large structures. *Nucleic Acids Res* 45:D282–D288. <https://doi.org/10.1093/nar/gkw962>
- Knudsen M, Wiuf C (2010) The CATH database. *Hum Genomics* 4(3):207
- Kozma D, Simon I, Tusnady GE (2012) PDBTM: Protein Data Bank of transmembrane proteins after 8 years. *Nucleic Acids Res* 41(D1):D524–D529
- Krieger E, Vriend G (2014) YASARA View—molecular graphics for all devices—from smartphones to workstations. *Bioinformatics* 30(20):2981–2982
- Laskowski RA (2007) Enhancing the functional annotation of PDB structures in PDBsum using key figures extracted from the literature. *Bioinformatics* 23(14):1824–1827
- Laskowski RA, Jabłońska J, Pravda L, Vařeková RS, Thornton JM (2018) PDBsum: structural summaries of PDB entries. *Protein Sci* 27(1):129–134
- Ledford H (2010) Big science: the cancer genome challenge. *Nat News* 464(7291):972–974
- Lo Conte L, Ailey B, Hubbard TJ, Brenner SE, Murzin AG, Chothia C (2000) SCOP: a structural classification of proteins database. *Nucleic Acids Res* 28(1):257–259
- Lobanov MY, Shoemaker BA, Garbuzynskiy SO, Fong JH, Panchenko AR, Galzitskaya OV (2009) ComSin: database of protein structures in bound (complex) and unbound (single) states in relation to their intrinsic disorder. *Nucleic Acids Res* 38(suppl_1):D283–D287
- Madej T, Lanczycki CJ, Zhang D, Thiessen PA, Geer RC, Marchler-Bauer A, Bryant SH (2013) MMDB and VAST+: tracking structural similarities between macromolecular complexes. *Nucleic Acids Res* 42(D1):D297–D303
- Markley JL, Ulrich EL, Berman HM, Henrick K, Nakamura H, Akutsu H (2008) BioMagResBank (BMRB) as a partner in the Worldwide Protein Data Bank (wwPDB): new policies affecting biomolecular NMR depositions. *J Biomol NMR* 40(3):153–155
- Mewes H-W, Frishman D, Güldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Münsterkötter M, Rudd S, Weil B (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* 30(1):31–34
- Pavelka A, Chovancova E, Damborsky J (2009) HotSpot Wizard: a web server for identification of hot spots in protein engineering. *Nucleic Acids Res* 37(suppl_2):W376–W383
- Perutz MF, Rossmann MG, Cullis AF, Muirhead H, Will G, North A (1960) Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-Å resolution obtained by X-ray analysis. *Nature* 185(4711):416
- Popenda M, Szachniuk M, Blazewicz M, Wasik S, Burke EK, Blazewicz J, Adamiak RW (2010) RNA FRABASE 2.0: an advanced web-accessible database with the capacity to search the three-dimensional fragments within RNA structures. *BMC Bioinformatics* 11(1):231
- Putignano V, Rosato A, Banci L, Andreini C (2017) MetalPDB in 2018: a database of metal sites in biological macromolecular structures. *Nucleic Acids Res* 46(D1):D459–D464
- Rosenbaum DM, Rasmussen SG, Kobilka BK (2009) The structure and function of G-protein-coupled receptors. *Nature* 459(7245):356
- Shindyalov IN, Bourne PE (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng* 11(9):739–747
- Shindyalov IN, Bourne PE (2001) A database and tools for 3-D protein structure comparison and alignment using the Combinatorial Extension (CE) algorithm. *Nucleic Acids Res* 29(1):228–229
- Stansfeld PJ, Goose JE, Caffrey M, Carpenter EP, Parker JL, Newstead S, Sansom MS (2015) MemProtMD: automated insertion of membrane protein structures into explicit lipid membranes. *Structure* 23(7):1350–1361
- Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J et al (2007) BioMagResBank. *Nucleic Acids Res* 36(suppl_1):D402–D408
- Varadi M, Tompa P (2015) The protein ensemble database. Intrinsically disordered proteins studied by NMR spectroscopy. Springer, pp 335–349
- Velankar S, Alhroub Y, Alili A, Best C, Boutselakis HC, Caboche S et al (2010) PDBe: protein data bank in Europe. *Nucleic Acids Res* 39(suppl_1):D402–D410

- Velankar S, van Ginkel G, Alhroub Y, Battle GM, Berrisford JM, Conroy MJ, Dana JM, Gore SP, Gutmanas A, Haslam P (2015) PDBe: improved accessibility of macromolecular structure data from PDB and EMDB. *Nucleic Acids Res* 44(D1):D385–D395
- Wang XT, Chan TF, Lam V, Engel PC (2008) What is the role of the second “structural” NADP⁺-binding site in human glucose 6-phosphate dehydrogenase? *Protein Sci* 17(8):1403–1411
- Wlodawer A, Minor W, Dauter Z, Jaskolski M (2008) Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. *FEBS J* 275(1):1–21
- Yeats C, Maibaum M, Marsden R, Dibley M, Lee D, Addou S, Orengo CA (2006) Gene3D: modelling protein structure, function and evolution. *Nucleic Acids Res* 34(suppl_1):D281–D284
- Yin H, Flynn AD (2016) Drugging membrane protein interactions. *Annu Rev Biomed Eng* 18:51
- Zanegina O, Kirsanov D, Baulin E, Karyagina A, Alexeevski A, Spirin S (2015) An updated version of NPIDB includes new classifications of DNA–protein complexes and their families. *Nucleic Acids Res* 44(D1):D144–D153