# Structural Databases of Biological Macromolecules

Advanced article

Online posting date: 17th September 2012

**Margaret J Gabanyi,** *Rutgers, The State University of New Jersey, Piscataway, New Jersey, USA*

**Helen M Berman,** *Rutgers, The State University of New Jersey, Piscataway, New Jersey, USA*

*Based in part on the previous version of this eLS article 'Structural Databases of Biological Macromolecules' (2005) by Helen M Berman.*

**A biological macromolecule's function is determined by the chemical and physical characteristics of its three-dimensional (3D) shape, or 'structure'. For this reason, knowing the structure of a biomolecule is very helpful if we want to be able to understand living systems and disease. The Protein Data Bank (PDB) began as an archive of the structural data available about biological macromolecules. The advances made in all technologies have been mirrored in further development of the PDB and in the structural speciality and structural characteristic databases that have also evolved. New resource portals such as the Protein Structure Initiative (PSI) Structural Biology Knowledgebase (SBKB) also collect all available genomic, structural, and functional information together to reduce the time needed to obtain the latest information on structurally determined proteins. This article will describe selected structural databases and resources available to the public today.**

## Historical Background

In 1957, the first structure of a biological macromolecule (myoglobin) was determined (Kendrew *et al*., 1958). This was followed by the determinations of several more key molecules, including haemoglobin (Perutz *et al*., 1960), lysozyme (Blake *et al*., 1965) and ribonuclease (Kartha *et al*., 1967; Wyckoff *et al*., 1967). In 1971, small-molecule and protein crystallographers from both sides of the Atlantic agreed to establish a data bank of the protein structures being determined. Its mission would be to collect, archive and disseminate data on the three-dimensional structures of biological macromolecules. Walter Hamilton of the Brookhaven National Laboratory and Olga Kennard of the Cambridge Structural Database (CSD) collaborated to manage the Protein Data Bank (PDB) resource (1971). Hamilton's interest was borne from his work on the high-resolution determination of amino acid crystal structures and from his visionary idea of setting up distributed computing resources whereby every crystallographer would have a graphics workstation on his/her desk with full network access to powerful high-speed computers. Kennard had founded the CSD in 1965 to create a database of organic and metal-organic compounds studied by X-ray and neutron diffraction, and was well experienced in managing structural data. **See also**: Crystallisation of Nucleic Acids; Crystallization of Proteins and Protein–Ligand Complexes

**eLS subject area:** Structural Biology

**How to cite:**
Gabanyi, Margaret J; and Berman, Helen M (September 2012)
Structural Databases of Biological Macromolecules. In: eLS. John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0005252.pub2

The PDB contained less than a dozen structures at its inception, with a few more structures added each year. The structures themselves were relatively small. The PDB file format was simple, and it was relatively easy to extract the structures from magnetic tape to find out what you wanted to know about any particular molecule.

In the 1980s, improvements in the technology required to study crystal structures began to evolve rapidly. Now, three decades later, modern molecular biology techniques have made it much more straightforward to obtain large quantities of proteins. Crystallisation methods have emerged that allow investigators to screen many different conditions using exceedingly small amounts of material. Data collection methods have improved at all levels. The lifetimes of crystals are routinely extended by flash freezing. The radiation sources are much more intense, especially with the emergence of powerful synchrotron beamlines. Detectors are much more sensitive and allow the very rapid collection of arrays of reflections. Methods for phase determination and refinement have improved. Indeed, crystallography is part of the armament of techniques that is readily accessible to biologists.

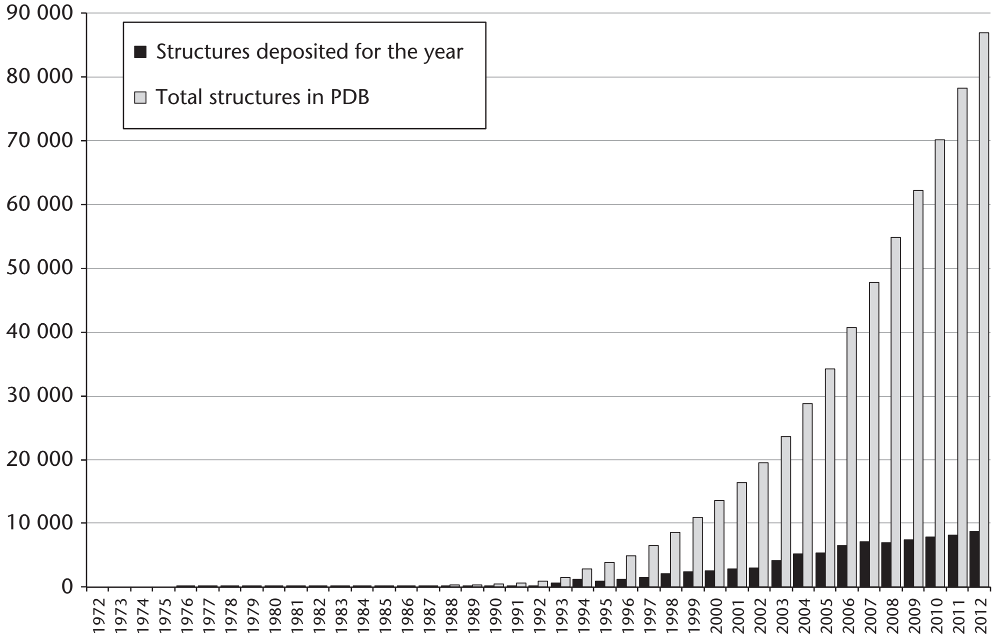


**Figure 1** Growth of the contents of the Protein Data Bank (as of May 2012). The number of structures deposited each year is shown in grey, the total number of structures available in black. 2012 values are projected deposited/total values based on deposition trends up to May 2012. This chart is regularly updated at http://www.rcsb.org

As crystallographic methods continue to improve, other structure determination methods have come of age. Nuclear magnetic resonance (NMR), which allows the determination of structures in solution, is currently responsible for approximately 12% of the structures released in the PDB. **See also**: Nuclear Magnetic Resonance (NMR) Spectroscopy: Structure Determination of Proteins and Nucleic Acids

Developments in electron cryomicroscopy (cryo-EM) have proven useful in the determination of very large assemblies such as membrane-bound receptors and pores, cellular enzyme complexes (e.g. ribosomes, chaperonins and synthases), and both polyhedral and complex viruses. As of June 2012, there are over 400 cryo-EM structures in the PDB. **See also**: Electron Cryomicroscopy and Three-dimensional Computer Reconstruction of Biological Molecules

The structural genomics initiative, which began in 2000, has determined over 10000 structures (14% of the PDB) in a high-throughput mode. Thus, the PDB holdings will continue to grow (**Figure 1**). **See also**: Structural Proteomics: Large-scale Studies

The level of activity in structural biology has made it essential that the PDB use the most modern technologies to collect, archive and disseminate data. The PDB is an *archival repository*, which contains coordinates and experimental data for biological macromolecules determined using public funds as well as many from the private sector. It is managed by the four members of the worldwide Protein Data Bank consortium (wwPDB), which work together to ensure data standardisation (Berman *et al*., 2003). The PDB archive also contains information about the methods and materials used to determine those structures. Other databases have emerged (**Table 1**) that extract some of the information contained in the PDB and organise that information in different ways so as to enable different types of query. These are *value-added databases*, which serve the needs of particular users. In this article we describe the PDB and some of these other structural databases. The web addresses for all data resources mentioned in this article can also be found in **Table 1**.

## The worldwide Protein Data Bank

The Protein Data Bank is the primary international repository for 3D coordinates of biological macromolecules such as proteins, nucleic acids, and their binding partners/ligands when present. It currently holds over 82000 entries and is managed by the four members of the worldwide Protein Data Bank (wwPDB) consortium: the Research Collaboratory for Structural Bioinformatics (RCSB, USA) (Berman *et al*., 2000), Protein Data Bank in Europe (PDBe, UK) (Velankar *et al*., 2012), Protein Data Bank in Japan (PDBj, Japan) (Kinjo *et al*., 2012), and the BioMagResBank (BMRB, USA) (Ulrich *et al*., 2008). Each week, the wwPDB releases data that have been fully checked, annotated, and approved by the depositors for release into the PDB FTP archive. The current distribution of structures in the PDB is shown in **Table 2**.

**Table 1** Selected database resources for macromolecular structure mentioned in this review

| | | |
|---|---|---|
| *Repository for 3D biological macromolecule structures* | | |
| Protein Data Bank | macromolecular structures FTP archive | http://www.wwpdb.org |
| *Structural databases* | | |
| RCSB PDB (Berman *et al*., 2000) | macromolecular structure deposition, search, and analysis tools (USA) | http://www.rcsb.org |
| PDBe (Velankar *et al*., 2012) | macromolecular structure deposition, search, and analysis tools (Europe) | http://www.pdbe.org |
| PDBj (Kinjo *et al*., 2012) | macromolecular structure deposition, search, and analysis tools (Japan) | http://www.pdbj.org |
| BMRB (Ulrich *et al*., 2008) | NMR data set deposition, search, and validation tools | http://www.bmrb.wisc.edu |
| EMDataBank (Lawson *et al*., 2011) | Cryo-EM data deposition, search, and analysis tools | http://www.emdatabank.org |
| Nucleic Acid Database (Berman *et al*., 2002) | nucleic acid structure search and analysis tools | http://ndbserver.rutgers.edu |
| CSD (Allen *et al*., 1979) | small molecule structure database | http://www.ccdc.cam.ac.uk |
| Membrane Proteins of Known Structure (White, 2004) | membrane protein structure search and annotation | http://blanco.biomol.uci.edu/mpstruc/listAll/list |
| *Structural characteristic databases* | | |
| CATH (Cuff *et al*., 2009) | structural classification | http://www.cathdb.info |
| SCOP (Andreeva *et al*., 2004) | structural classification | http://scop.mrc-lmb.cam.ac.uk/scop/ |
| PDBeFold (SSM) (Krissinel and Henrick, 2004) | structural classification | http://pdbe.org/fold |
| PDBSum (Laskowski *et al*., 1997) | structural annotation and analysis | http://www.biochem.ucl.ac.uk/bsm/pdbsum/ |
| Gene3D (Lees *et al*., 2012) | structural alignment and annotation | http://gene3d.biochem.ucl.ac.uk/Gene3D/ |
| SUPERFAMILY (Wilson *et al*., 2007) | structural alignment and annotation | http://supfam.cs.bris.ac.uk/SUPERFAMILY/ |
| HSSP (Dodge *et al*., 1998) | structure comparison, search, and analysis | http://swift.cmbi.ru.nl/gv/hssp/ |
| *Specialty databases* | | |
| PDBePISA (Krissinel and Henrick, 2007) | intermolecular interactions and quaternary structure | http://pdbe.org/pisa |
| Dictionary of Interfaces in Proteins (DIP) (Salwinski *et al*., 2004) | protein–protein interactions | http://dip.doe-mbi.ucla.edu/ |
| DisProt (Sickmeier *et al*., 2007) | disordered protein database | http://www.disport.org |
| Protein Circular Dichroism Database (Whitmore *et al*., 2011) | data repository for circular dichroism experiments | http://pcddb.cryst.bbk.ac.uk/home.php |

## Structural Data Collected by the wwPDB

The PDB archive contains entries containing the three-dimensional Cartesian coordinates of resolved atoms in a biomolecule, along with the experimental details of its structure determination such as the crystal conditions or NMR solution. Related experimental data, which is now required upon deposition, include structure factors from X-ray experiments and chemical shifts and constraints derived from NMR experiments. Additional quantitative data related to NMR experiments are housed at the BMRB. The EM Data Bank (EMDB), the primary archive for experimentally determined maps obtained using three-dimensional electron microscopy methods, joined the PDB archive in March 2012 to provide EM maps and models from the same archive (Lawson *et al*., 2011).

**Table 2** Summary of different types of biological macromolecule structures in the Protein Data Bank (as of 15 May 2012)

| Experimental methods | Proteins, peptides and viruses | Nucleic acids | Protein–nucleic acid complexes | Other[a] | Total |
|---|---|---|---|---|---|
| X-ray | 67466 | 1365 | 3409 | 2 | 72242 |
| NMR | 8276 | 988 | 186 | 7 | 9457 |
| Electron Microscopy | 293 | 22 | 120 | 0 | 435 |
| Hybrid[b] | 44 | 3 | 2 | 1 | 50 |
| Other[c] | 141 | 4 | 5 | 13 | 163 |
| Total | 76220 | 2382 | 3722 | 23 | 82347 |

[a] Other biomolecules in the archive include legacy peptide-based polymers.
[b] Hybrid refers to structures that were solved using more than one determination method.
[c] Other methods include neutron diffraction, fiber diffraction and neutron scattering.

The deposited atomic data undergoes validation, which is the process of evaluating how well the fitted and refined model fits the experimental data. The covalent bond distances and angles, stereochemical validation, close contacts, ligand and atom nomenclature, sequence comparison, distant waters, and overall geometry are compared to accepted community standards, and authors are informed of any inconsistencies. The entries are then annotated with information related to the validation of the entry.

## Structural Databases

Although wwPDB members collaborate as data deposition and annotation centres, each site develops unique resources to access and analyse the data in the archive.

The RCSB PDB website can be used to perform simple and advanced searches based on annotations relating to sequence, structure and function, and to visualise, download, and analyse PDB data (Rose *et al.*, 2011). Data from external resources is incorporated with PDB data through links and clearly marked web page widgets. New features are added regularly; recent additions include the ability to tour the PDB archive by exploring the distributions of data across significant categories (organism, taxonomy, polymer type, *etc.*), and an improved top bar search mechanism to find entries by molecule name. The RCSB PDB offers educational resources through the ‘PDB-101’ online portal, such as Molecule of the Month columns that are aimed at promoting a structural view of biology for students of all ages. Users personalise their RCSB PDB usage by creating a free ‘myPDB’ account. This feature not only saves viewing preferences and personal notes, but also will automatically send emails if newly released structures match any saved searches.

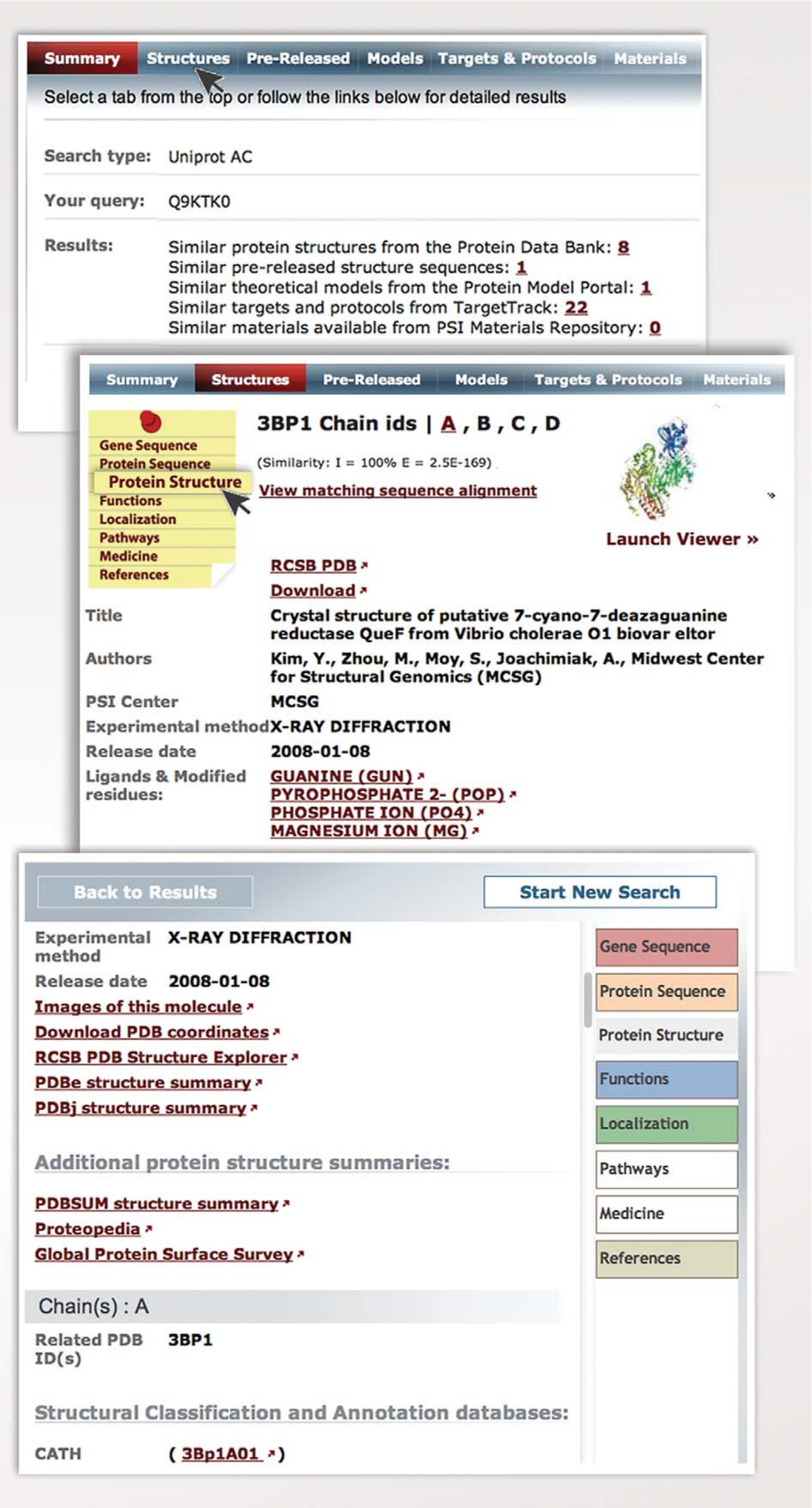
The PDBe site extends its search capabilities with many structural analysis and search tools. Upon entering a PDB ID of interest, it gives ‘one-click access’ to the structural entry in PDBeAtlas (Velankar *et al.*, 2012), the ability to download files, predict protein interfaces and quaternary assembly of the protein (PDBePISA) (Krissinel and Henrick, 2007), find similarly folded structures (PDBe-Fold/SSM) (Krissinel and Henrick, 2004), or find certain structural motifs or binding sites within each structural chain (PdbeMotif) (Golovin and Henrick, 2009). An educational section called ‘PDB Quips’ also highlights interesting protein structures.

PDBj has also developed unique tools for searching and visualising PDB entries (Kinjo *et al.*, 2012). Yorodumi is an interactive viewer that can visualise data from the PDB as well as EM Data Bank archives; EM Navigator help visualise EM volume data. It has also built tools for bioinformaticists and, in collaboration with the RCSB PDB group, has developed an eXtensible Markup Language (XML) version of the PDB archive (PDBML) (Westbrook *et al.*, 2005) and an extended version (PDBMLplus) with additional curated annotations which can be searched using the application PDBj Mine. It is also the only member site that supports browsing in additional languages such as Japanese, simplified/traditional Chinese, and Korean.

BMRB joined the wwPDB in 2006. It is a repository that collects data measured from any NMR experiment, not only those related to structure determination. It contains experimental data such as the assigned chemical shifts, coupling constants; peak lists for a variety of biological macromolecules (even small ones excluded from the PDB), as well as derived data such as hydrogen exchange rates, pKa values, and relaxation parameters that explain real-time biochemical processes. BMRB also collects the NMR restraints for PDB entries, time domain spectral data, and NMR data on hundreds of metabolites and standard compounds (Ulrich *et al.*, 2008).

Additional structural databases exist that focus on specific subsets of the PDB archive. The Nucleic Acid Database

**Figure 2** Example of a structure query using the Structural Biology Knowledgebase. Users can search the SBKB by protein or DNA sequence, by Protein Data Bank (PDB) ID, UniProt Accession Code (AC) or by text. Red links direct users to the primary data resources. (Top) The summary of search results includes matching structures, theoretical models, structure determination targets, protocols, and available DNA clones from the PSI Materials Repository. (Middle) the Structures tab organises the links to primary data resources. (Bottom) The SBKB’s annotation notebook will provide available links to over 150 key biological databases; biological categories in the right-hand tabs that have no existing annotations are greyed out.



Select a tab from the top or follow the links below for detailed results

Search type: Uniprot AC

Your query: Q9KTK0

Results:
Similar protein structures from the Protein Data Bank: 8
Similar pre-released structure sequences: 1
Similar theoretical models from the Protein Model Portal: 1
Similar targets and protocols from TargetTrack: 22
Similar materials available from PSI Materials Repository: 0

3BP1 Chain ids | A , B , C , D

(Similarity: I = 100% E = 2.5E-169)

View matching sequence alignment

RCSB PDB

Download

| | |
|---|---|
| Title | Crystal structure of putative 7-cyano-7-deazaguanine reductase QueF from Vibrio cholerae O1 biovar eltor |
| Authors | Kim, Y., Zhou, M., Moy, S., Joachimiak, A., Midwest Center for Structural Genomics (MCSG) |
| PSI Center | MCSG |
| Experimental method | X-RAY DIFFRACTION |
| Release date | 2008-01-08 |
| Ligands & Modified residues: | GUANINE (GUN)<br>PYROPHOSPHATE 2- (POP)<br>PHOSPHATE ION (PO4)<br>MAGNESIUM ION (MG) |

| | |
|---|---|
| Experimental method | X-RAY DIFFRACTION |
| Release date | 2008-01-08 |

Images of this molecule

Download PDB coordinates

RCSB PDB Structure Explorer

PDBe structure summary

PDBj structure summary

Additional protein structure summaries:

PDBSUM structure summary

Proteopedia

Global Protein Surface Survey

Chain(s) : A

| | |
|---|---|
| Related PDB ID(s) | 3BP1 |

Structural Classification and Annotation databases:

| | |
|---|---|
| CATH | ( 3Bp1A01 ) |

focuses on the structures of nucleic acids polymers. In addition to search functionality, the site contains tools for visualisation, for predicting protein–DNA interactions, and motif-based searches for similar structures (Berman *et al.*, 2002). EMDataBank is a global deposition and retrieval network for cryoEM map, model and associated metadata, as well as a portal for software tools for standardized map format conversion, map, segmentation and model assessment, visualisation, and data integration (Lawson *et al.*, 2011). The Membrane Proteins of Known Structures database (MPStruc) annotates entries with tertiary structure information along with their orientation and assembly in the membrane (White, 2004).

## Structural Characteristic Databases

Although the focus of many structural databases begins with individual structures, some databases organise their data according to tertiary structural characteristics. SCOP (a Structural Classification of Proteins) classifies each structure in the PDB according to 'family', 'superfamily', 'common fold' and 'class' (Andreeva *et al.*, 2004). *Families* are classified according to their sequence similarities. Families with similar structure and function belong to the same superfamily. Families and superfamilies with the same arrangement of secondary structures, which are connected with one another in the same way, have the same common fold. *Class* refers to the types of secondary structures (all alpha helix, all beta sheet, alpha–beta, *etc*.). SCOP was one of the earliest databases that attempted to integrate sequence, structure and function information; it continues to be a major resource in structural biology.

CATH provides another classification scheme based on class (C), architecture (A), topology (T) and homologous superfamilies (H) (Cuff *et al.*, 2009). *Class* defines the secondary structure content as in SCOP. Architecture defines the description of the arrangement of these secondary structures without consideration of the connectivities. *Topology* is equivalent to fold in SCOP. Finally, homologous superfamilies contain all folds with a similar function. CATH has a systematic classification system for all structures analogous to the EC classification for enzyme function. The type of research possible with this database is exemplified by an analysis of all enzymes in which it was shown that the topology of enzymes is more related to the ligands that bind to them rather than the enzyme EC class (Martin *et al.*, 1998).

Information such as those organised in CATH and SCOP are being used to close the sequence-structure gap by looking for relationships between comparable structures. HSSP (Homology-derived structures of proteins) provides a list of sequence homologues for each entry in the PDB, with the sequences aligned to the PDB protein (Dodge *et al.*, 1998). SUPERFAMILY, part of the SCOP family of databases, organises structures into evolutionarily related groups to promote the annotation of under-characterized proteins (Wilson *et al.*, 2007). Gene3D, part of the CATH family of databases, plays a similar role in assigning CATH domains to gene products of unknown structure (Lees *et al.*, 2012).

## Speciality Databases

Knowing the structure of a protein alone is not enough; one must also know how it might interact with its environment. The PDBePISA tool can be used to determine the quaternary assembly of macromolecules derived from calculations made upon the molecule's surface or observed interfaces. The Dictionary of Interfaces in Proteins (DIP) is a data bank of complementary molecular surface patches and is meant to enable molecular recognition research (Salwinski *et al.*, 2004). The Global Protein Surface Survey (GPSS) analyses surfaces to create 3D surface libraries and for protein surface comparison (Binkowski, 2009).

A large number of databases have also been created which allow for a structural view of specific biological processes. A comprehensive list of speciality structural databases is maintained by *Nucleic Acids Research* for their annual 'Databases' special issue. The open access catalogue can be found at http://www.oxfordjournals.org/nar/database/cat/4.

A protein's shape, stability or solubility might not be conducive to current structure determination methods; so several databases collect additional biophysical or calculated structure evidence. Protein Circular Dichroism (CD) Data Bank holds CD spectra with secondary structure information for nearly 300 proteins, both structurally determined (for reference purposes) and those not structurally determined (Whitmore *et al.*, 2011). DisProt is the key data resource that collects provides information about proteins that lack fixed 3D structure in their putatively native states (Sickmeier *et al.*, 2007).

## A Portal to Combined Structural Information

As the number of publicly available structural resources grows, the need to be able to combine this information effectively becomes more of a challenge for researchers. For this reason, the Structural Biology Knowledgebase (SBKB) was designed as a portal to integrate information about structurally determined proteins (or those targeted for structure determination) from many resources in order to enable new knowledge (Gabanyi *et al.*, 2011). When searching the SBKB by a protein's amino acid sequence, the SBKB runs a BLAST search and returns matching and homologous structures from the PDB, relevant theoretical models from the Protein Model Portal, and biological descriptions (annotations) from over 150 genomics, structural, and function-related databases that hold information about the protein of interest. Since the SBKB is created as a part of the Protein Structure Initiative programme, it will also return progress on similar proteins targeted for structural determination, including histories and protocols for protein production and structure determination from TargetTrack, and deoxyribonucleic acid (DNA) plasmid availability from the PSI Materials Repository (Cormier *et al.*, 2012). Searches by plain text will find related structures and annotations, related research and technical highlights from the Nature Publishing Group, and reports and articles within the PSI Technology (Gifford *et al.*, 2012) and PSI Publication Portals.

The SBKB has been designed so that both the novice user and the advanced structural biologist can gain insight from structural data and analysis tools. Searches are started by entering a query into the search box on the SBKB homepage (**Figure 2**). SBKB uses simple-to-use molecular visualisation tools (Martz, 2009) with the list of returned structure matches which can graphically render the biomolecule in several biologically relevant styles. Continuing with the structure results, a yellow sticky note lists the categories for which additional biological annotations exists; clicking on the Protein Structure heading will provide details of the structure, as well as links to the structure summary pages of all 3 wwPDB sites (4 if NMR structure), additional derived protein summary resources, structure classification databases, and experimental details for the structure determination.

A unique feature of the SBKB is its inclusion of theoretical comparative models which can be useful in the absence of an experimental structure. The Protein Model Portal (Bordoli and Schwede, 2012) provides centralized access to over 22 million theoretical models for nearly 4 million UniProt sequences from Swiss-Model Repository (Kiefer *et al.*, 2009), MODBASE (Pieper *et al.*, 2009), and 6 Protein Structure Initiative centres. It also provides real-time calculation of new comparative models using the services of 5 modelling groups: MODELLER (Eswar *et al.*, 2008), M4T (Rykunov *et al.*, 2009), I-Tasser (Roy *et al.*, 2010), Swiss-Model, and HHPred (Hildebrand *et al.*, 2009).

## Challenges for the Structural Community

The PDB is now much more than a repository of coordinate data. To make this resource even more useful, all the files need to be represented consistently and accurately across the archive. The same technologies that have contributed to the dramatic increases of data in the PDB also push the ways in which PDB data are represented. Regular reviews across the data help to identify areas where data needs to be remediated for improved representation in the archive (Henrick *et al.*, 2008; Lawson *et al.*, 2008). It is also important for the wwPDB to be able to review the quality of the data in the archive. Method-specific task forces have been convened by the wwPDB to develop consensus on what types of validation and validation software should be used to review PDB data. Recommendations have been published by the X-ray Validation Task Force (Read *et al.*, 2011) along with a first report from the Electron Microscopy Task Force (Henderson *et al.*, 2012). Additional task force meetings and reports will be published.

The goal of being able to relate structure to function will be facilitated by different types of database efforts. Databases that assemble information about particular protein families will be one avenue that will provide this information. In these databases the coverage is very narrow and deep, so that a truly full understanding of a single class of proteins with known function is possible. The lessons learned from these types of resources will perhaps allow us to develop some general principles about the relationships of structure and function.

## Acknowledgements

The RCSB PDB is supported by the National Science Foundation (NSF DBI 0829586), the Department of Energy, the National Library of Medicine, the National Cancer Institute, National Institute of Neurological Disorders and Stroke, and the National Institute of Diabetes and Digestive and Kidney Diseases. The SBKB is supported by the National Institute of General Medical Sciences (U01 GM093324).

## References

Allen FH, Bellard S, Brice MD *et al.* (1979) The Cambridge Crystallographic Data Centre: computer-based search, retrieval, analysis and display of information. *Acta Crystallographica Section B: Structural Science* **35**: 2331–2339.

Andreeva A, Howorth D, Brenner SE *et al.* (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Research* **32**(Database issue): D226–D229.

Berman HM, Henrick K and Nakamura H (2003) Announcing the worldwide Protein Data Bank. *Nature Structural Biology* **10**(12): 980.

Berman HM, Westbrook J, Feng Z *et al.* (2002) The nucleic acid database. *Acta Crystallographica Section D* **58**: 899–907.

Berman HM, Westbrook JD, Feng Z *et al.* (2000) The protein data bank. *Nucleic Acids Research* **28**: 235–242.

Binkowski A (2009) *Global Protein Surface Survey*. http://gpss.mcsg.anl.gov/.

Blake CCF, Koenig DF, Mair GA *et al.* (1965) Structure of hen egg-white lysozyme. A three dimensional Fourier synthesis at 2 Å resolution. *Nature* **206**: 757–761.

Bordoli L and Schwede T (2012) Automated protein structure modeling with SWISS-MODEL workspace and the protein model portal. *Methods in Molecular Biology* **857**: 107–136.

Cormier CY, Park JG, Fiacco M *et al.* (2012) PSI:Biology-materials repository: a biologist's resource for protein expression plasmids. *Journal of Structural and Functional Genomics* **12**(2): 55–62.

Cuff AL, Sillitoe I, Lewis T *et al.* (2009) The CATH classification revisited--architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Research* **37**(Database issue): D310–D314.

Dodge C, Schneider R and Sander C (1998) The HSSP database of protein structure-sequence alignments and family profiles. *Nucleic Acids Research* **26**: 313–315.

Eswar N, Eramian D, Webb B *et al.* (2008) Protein structure modeling with MODELLER. *Methods in Molecular Biology* **426**: 145–159.

Gabanyi MJ, Adams PD, Arnold K *et al.* (2011) The Structural Biology Knowledgebase: a portal to protein structures, sequences, functions, and methods. *Journal of Structural and Functional Genomics* **12**: 45–54.

Gifford LK, Carter LG, Gabanyi MJ *et al.* (2012) The protein structure initiative structural biology knowledgebase technology portal: a structural biology web resource. *Journal of Structural and Functional Genomics* **13**(2): 57–62.

Golovin A and Henrick K (2009) Chemical substructure search in SQL. *Journal of Chemical Information and Modeling* **49**(1): 22–27.

Henderson R, Sali A, Baker ML *et al.* (2012) Outcome of the first electron microscopy validation task force meeting. *Structure* **20**(2): 205–214.

Henrick K, Feng Z, Bluhm W *et al.* (2008) Remediation of the protein data bank archive. *Nucleic Acids Research* **36**(Database issue): D426–D433.

Hildebrand A, Remmert M, Biegert A *et al.* (2009) Fast and accurate automatic structure prediction with HHpred. *Proteins* **77**(suppl. 9): 128–132.

Kartha G, Bello J and Harker D (1967) Tertiary structure of ribonuclease. *Nature* **213**: 862–865.

Kendrew JC, Bodo G, Dintzis HM *et al.* (1958) A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature* **181**: 662–666.

Kiefer F, Arnold K, Kunzli M *et al.* (2009) The SWISS-MODEL repository and associated resources. *Nucleic Acids Research* **37**(Database issue): D387–D392.

Kinjo AR, Suzuki H, Yamashita R *et al.* (2012) Protein Data Bank Japan (PDBj): maintaining a structural data archive and resource description framework format. *Nucleic Acids Research* **40**(Database Issue): D453–D460.

Krissinel E and Henrick K (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica Section D: Biological Crystallography* **60**: 2256–2268.

Krissinel E and Henrick K (2007) Inference of macromolecular assemblies from crystalline state. *Journal of Molecular Biology* **372**(3): 774–797.

Laskowski RA, Hutchinson EG, Michie AD *et al.* (1997) PDBSum: a Web-based database of summaries and analyses of all PDB structures. *Trends in Biochemical Sciences* **22**: 488–490.

Lawson CL, Baker ML, Best C *et al.* (2011) EMDataBank.org: unified data resource for CryoEM. *Nucleic Acids Research* **39**(Database issue): D456–D464.

Lawson CL, Dutta S, Westbrook J *et al.* (2008) Representation of viruses in the remediated PDB archive. *Acta Crystallographica Section D* **64**: 874–882.

Lees J, Yeats C, Perkins J *et al.* (2012) Gene3D: a domain-based resource for comparative genomics, functional annotation and protein network analysis. *Nucleic Acids Research* **40**(Database issue): D465–D471.

Martin ACR, Orengo CA, Hutchinson EG *et al.* (1998) Protein folds and functions. *Structure* **6**: 875–884.

Martz E (2009) *FirstGlance in Jmol*. http://firstglance.jmol.org.

Perutz MF, Rossmann MG, Cullis AF *et al.* (1960) Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5 Å resolution, obtained by X-ray analysis. *Nature* **185**: 416–422.

Pieper U, Eswar N, Webb BM *et al.* (2009) MODBASE, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Research* **37**(Database issue): D347–D354.

Protein Data Bank (1971) Protein Data Bank. *Nature New Biology* **233**: 223.

Read RJ, Adams PD, Arendall WB III *et al.* (2011) A new generation of crystallographic validation tools for the Protein Data Bank. *Structure* **19**(10): 1395–1412.

Rose PW, Beran B, Bi C *et al.* (2011) The RCSB protein data bank: redesigned web site and web services. *Nucleic Acids Research* **39**(Database issue): D392–D401.

Roy A, Kucukural A  and Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols* **5**(4): 725–738.

Rykunov D, Steinberger E, Madrid-Aliste CJ *et al.* (2009) Improved scoring function for comparative modeling using the M4T method. *Journal of Structural and Functional Genomics* **10**(1): 95–99.

Salwinski L, Miller CS, Smith AJ *et al.* (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Research* **32**(Database issue): D449–D451.

Sickmeier M, Hamilton JA, LeGall T *et al.* (2007) DisProt: the Database of Disordered Proteins. *Nucleic Acids Research* **35**(Database issue): D786–D793.

Ulrich EL, Akutsu H, Doreleijers JF *et al.* (2008) BioMag ResBank. *Nucleic Acids Research* **36**(Database issue): D402–D408.

Velankar S, Alhroub Y, Best C *et al.* (2012) PDBe: protein data bank in Europe. *Nucleic Acids Research* **40**(Database issue): D445–D452.

Westbrook JD, Ito N, Nakamura H *et al.* (2005) PDBML: the representation of archival macromolecular structure data in XML. *Bioinformatics* **21**: 988–992.

White SH (2004) The progress of membrane protein structure determination. *Protein Science: A Publication of the Protein Society* **13**(7): 1948–1949.

Whitmore L, Woollett B, Miles AJ *et al.* (2011) PCDDB: the protein circular dichroism data bank, a repository for circular dichroism spectral and metadata. *Nucleic Acids Research* **39**(Database issue): D480–D486.

Wilson D, Madera M, Vogel C *et al.* (2007) The SUPERFAMILY database in 2007: families and functions. *Nucleic Acids Research* **35**(Database issue): D308–D313.

Wyckoff HW, Hardman KD, Allewell N *et al.* (1967) The structure of ribonuclease-S at 6 Å resolution. *Journal of Biological Chemistry* **242**: 3749–3753.

## Further Reading

Arnold E, Himmel DM and Rossmann MG (eds) (2012) *International Tables for Crystallography, vol. F: Crystallization of Biological Macromolecules* [Chapters 21 and 24]. West Sussex, UK: John Wiley & Sons, Ltd.

Hall SR and McMahon B (eds) (2006) *International Tables for Crystallography, vol. G: Definition and exchange of crystallographic data*. West Sussex, UK: John Wiley & Sons, Ltd.

Anonymous (2007) Structural Genomics Supplement. *Structure* **16**(1): 1–160.

Anonymous (2012) Nucleic Acids Databases Issue. *Nucleic Acids Research* **40**: D1–D1317.

## Web Links

Canadian Bioinformatics.ca Links Directory. http://bioinformatics.ca/links_directory/category/protein

Structural Biology Knowledgebase Portal to information on structurally targeted proteins, structures, theoretical models, methods, materials, technologies, and information on the Protein Structure Initiative. http://sbkb.org

Structural data resources directories: ExPASy Bioinformatics Resource Portal. http://expasy.org

Worldwide Protein Data Bank Member organizations serve as data deposition, processing, and distribution centers for PDB data http://www.wwpdb.org