Nucleic Acids Research, 2001, Vol. 29, No. 1 221–222

# PDBsum: summaries and analyses of PDB structures

**Roman A. Laskowski***

Department of Crystallography, Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK

Received August 31, 2000; Accepted October 4, 2000

## ABSTRACT

**PDBsum is a web-based database providing a largely pictorial summary of the key information on each macromolecular structure deposited at the Protein Data Bank (PDB). It includes images of the structure, annotated plots of each protein chain's secondary structure, detailed structural analyses generated by the PROMOTIF program, summary PROCHECK results and schematic diagrams of protein–ligand and protein–DNA interactions. RasMol scripts highlight key aspects of the structure, such as the protein's domains, PROSITE patterns and protein–ligand interactions, for interactive viewing in 3D. Numerous links take the user to related sites. PDBsum is updated whenever any new structures are released by the PDB and is freely accessible via http://www.biochem.ucl.ac.uk/bsm/pdbsum.**

## INTRODUCTION

To date, the 3D structures of over 13 000 biological macromolecules have been determined experimentally, principally by X-ray crystallography and NMR spectroscopy. The majority of these are protein structures, including protein–DNA and protein–ligand complexes. Together with sequence, physicochemical and functional annotations they provide a wealth of information crucial for the understanding of biological processes.

Each new structure is deposited in the Protein Data Bank (PDB) (1), which is currently run by the Research Collaboratory in Structural Biology (RCSB) (2). The structures can be downloaded from the RCSB's PDB web server, which also provides additional information about each one. Further information, some of it focusing on specific types of molecules or specific aspects of the molecules, can be obtained from a large number of other structural databases (3) on the Web. One such database is PDBsum, which is the subject of this paper.

## DESCRIPTION

The PDBsum database at http://www.biochem.ucl.ac.uk/bsm/pdbsum was created in 1995 (4). Its aim was to provide an at-a-glance summary of the molecules contained in each PDB entry (i.e. protein and DNA/RNA chains, small-molecule ligands, metal ions and waters), together with annotations and analyses of their key structural features. Thus, for each PDB entry there is a corresponding summary web page in PDBsum, accessible by the four-character PDB identifier.

The original PDBsum paper (4) described the basic contents of each entry, namely a block of 'header' information, relating to the entry as a whole, followed by a list of the molecules making up the structure, together with any relevant structural analyses of each. The header details start with a thumbnail image of the molecule(s) in question plus buttons for viewing the whole structure in 3D using RasMol (5) or VRML (Virtual Reality Modelling Language). These are followed by information extracted directly from the header records of the PDB file, summary PROCHECK (6) analyses (including a Ramachandran plot) giving an indication of the stereochemical 'quality' of all the protein chains in the structure, and links to related databases. In the list of molecules that follows, each protein chain is shown schematically by a 'wiring diagram' depicting its secondary structural motifs, primary sequence, structural domains and highlighting active site residues and residues that interact with ligands, metals or DNA/RNA molecules. The secondary structural motifs are computed by the PROMOTIF (7) program, whose detailed outputs are available via hyperlinks, while the domain definitions come from the CATH protein structural classification database (8,9). For each ligand molecule a LIGPLOT (10) diagram gives a schematic depiction of the hydrogen bonds and non-bonded interactions between it and the residues of the protein with which it interacts.

In the time since the original paper was published, a number of new analyses, links and functions have been added, and these are described in the remainder of this paper.

## NEW FEATURES

The first of the additions relates only to protein–DNA and DNA–ligand complexes. The interactions between the DNA chains and any other molecules in the complex are shown schematically in a diagram generated by the NUCPLOT (11) program. Like the LIGPLOT diagrams of protein–ligand interactions, the NUCPLOT diagrams show all the hydrogen bonds and non-bonded interactions between the molecules, as calculated by HBPLUS (12). The diagrams are output in PostScript format (see, for example, the PDBsum entry for PDB code 2OR1).

Next, each protein chain now has a direct link to the SAS (Sequence Annotated by Structure) (13) database. Clicking on the link initiates a FASTA search that scans the given chain's sequence of amino acid residues against a database of all sequences in the PDB. The net result is a list of all other chains in the PDB that are similar at the sequence level to the one of interest. The SAS database provides a variety of different

*Tel: +44 20 7419 3890; Fax: +44 20 7380 7193; Email: roman@biochem.ucl.ac.uk

annotations of the resultant multiple-sequence alignment, as well as enabling the user to view the superposed structures in 3D in RasMol.

Also new is the identification of any PROSITE (14) patterns present in each protein chain. These are patterns of residues that are found in regions that are highly conserved across all members of a given protein family and consequently characterise both the family itself and the biologically significant sites in its member proteins. In PDBsum the matching residues are coloured according to their conservation (and hence importance): from red for highly conserved, to blue for highly variable. Not all matching PROSITE patterns are shown; only those that appear to be true positives are included (15). The residues matching the PROSITE pattern can be viewed in RasMol to see where they lie in relation to the rest of the protein structure. A RasMol script renders the residues as thick sticks, coloured as on the PDBsum page, while showing the rest of the protein as a white backbone trace and any nearby ligands in spacefill. This often gives a clear indication of the structural and functional significance of the PROSITE pattern residues. See, for example, the entry for 1AAW, an aspartate aminotransferase, which contains the PROSITE pattern AA_TRANSFER_CLASS_1 corresponding to the Class 1 aminotransferases.

The RasMol scripts that display the PROSITE residues are generated on the fly by a program called RomLas (the name being a carefully chosen anagram of RasMol). The program is used throughout PDBsum to generate RasMol scripts for highlighting specific structural features. For example, below each LIGPLOT diagram there is a button for generating a RasMol script that displays the given ligand in the 3D context of the protein residues with which it interacts; the ligand is shown in thick sticks, while the protein residues are shown in wireframe and are labelled with the residue name and number.

Other new features include a simple text search facility on the home page and full listings of all the ligands and hetero groups found in the database. Links to a number of useful new databases have been added.

## ACKNOWLEDGEMENTS

PDBsum is maintained at University College, London. The authors of the programs used in generating and running the PDBsum database include David Smith, Gail Hutchinson, Alex Michie, Andrew Martin, Ian McDonald, Andrew Wallace, Nick Luscombe, Duncan Milburn and Atsushi Kasuya. I would like to thank Martin Jones and John Bouquiere for their contribution to the database's development and running. Thanks also to Frances Pearl, Malcolm MacArthur, Edith Chan and, most of all, Janet Thornton.

## REFERENCES

1. Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F.,Jr, Brice,M.D., Rogers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.
2. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242. Updated article in this issue: *Nucleic Acids Res.* (2001), **29**, 214–218.
3 Berman,H.M. (1999) The past and future of structure databases. *Curr. Opin. Struct. Biol.*, **10**, 76–80.
4. Laskowski,R.A., Hutchinson,E.G., Michie,A.D., Wallace,A.C., Jones,M.L. and Thornton,J.M. (1997). PDBsum: a Web-based database of summaries and analyses of all PDB structures. *Trends Biochem. Sci.*, **22**, 488–490.
5. Sayle,R.A. and Milner-White,E.J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374–376.
6. Laskowski,R.A., MacArthur,M.W., Moss,D.S. and Thornton,J.M. (1993) PROCHECK - a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.*, **26**, 283–291.
7. Hutchinson,E.G. and Thornton,J.M. (1996) PROMOTIF – a program to identify and analyze structural motifs in proteins. *Protein Sci.*, **5**, 212–220.
8. Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) CATH: a hierarchic classification of protein domain structures, *Structure*, **5**, 1093–1108.
9. Pearl,F.M.G., Lee,D., Bray,J.E., Sillitoe,I., Todd,A.E., Harrison,A.P., Thornton,J.M. and Orengo,C.A. (2000) Assigning genomic sequences to CATH. *Nucleic Acids Res.*, **28**, 277–282. Updated article in this issue: *Nucleic Acids Res.* (2001), **29**, 223–227.
10. Wallace,A.C., Laskowski,R.A. and Thornton,J.M. (1995) LIGPLOT: A program to generate schematic diagrams of protein–ligand interactions. *Protein Eng.*, **8**, 127–134.
11. Luscombe,N.M., Laskowski,R.A. and Thornton,J.M. (1997) NUCPLOT: a program to generate schematic diagrams of protein–nucleic acid interactions. *Nucleic Acids Res.*, **25**, 4940–4945.
12. McDonald,I.K. and Thornton,J.M. (1994) Satisfying hydrogen-bonding potential in proteins. *J. Mol. Biol.*, **238**, 777–793.
13. Milburn,D., Laskowski,R.A. and Thornton,J.M. (1998) Sequences annotated by structure: a tool to facilitate the use of structural information in sequence analysis. *Protein Eng.*, **11**, 855–859.
14. Hofmann,K., Bucher,P., Falquet,L. and Bairoch,A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.*, **27**, 215–219.
15. Kasuya,A. and Thornton,J.M. (1999) Three-dimensional structure analysis of PROSITE patterns. *J. Mol. Biol.*, **286**, 1673–1691.