Published online 16 November 2006 *Nucleic Acids Research, 2007, Vol. 35, Database issue* ***D301–D303***
doi:10.1093/nar/gkl971

The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data

Helen Berman*, Kim Henrick[1], Haruki Nakamura[2] and John L. Markley[3]

RCSB Protein Data Bank, Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, 610 Taylor Road, Piscataway, NJ 08854-8087, USA, [1]MSD-EBI, EMBL Outstation-Hinxton, Cambridge CB10 1SD, UK, [2]PDBj, Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan and [3]BioMagResBank, University of Wisconsin-Madison, Department of Biochemistry, 433 Babcock Drive, Madison, WI 53706, USA

Received September 14, 2006; Accepted October 20, 2006

ABSTRACT

The worldwide Protein Data Bank (wwPDB) is the international collaboration that manages the deposition, processing and distribution of the PDB archive. The online PDB archive is a repository for the coordinates and related information for more than 38 000 structures, including proteins, nucleic acids and large macromolecular complexes that have been determined using X-ray crystallography, NMR and electron microscopy techniques. The founding members of the wwPDB are RCSB PDB (USA), MSD-EBI (Europe) and PDBj (Japan) [H.M. Berman, K. Henrick and H. Nakamura (2003) *Nature Struct. Biol.*, 10, 980]. The BMRB group (USA) joined the wwPDB in 2006. The mission of the wwPDB is to maintain a single archive of macromolecular structural data that are freely and publicly available to the global community. Additionally, the wwPDB provides a variety of services to a broad community of users. The wwPDB website at http://www.wwpdb.org/ provides information about services provided by the individual member organizations and about projects undertaken by the wwPDB.

HISTORY AND BACKGROUND

The Protein Data Bank (PDB) was founded in 1971 to provide a repository for three-dimensional (3D) structure data of experimentally determined biological macromolecules (1–3). The PDB archive contains 3D coordinate data, information about the chemical content such as polymer sequence and ligand chemistry, information about the experiment used to derive the structure and some qualitative descriptions of the structure. When the PDB was in its infancy, the archive contained seven structures composed of loosely structured free text. Today, the PDB archive contains close to 40 000 structures and relies upon strict ontologies that define the content of these entries.

The data contained in the PDB are generated and submitted by scientists from around the globe to sites in the United States, Europe and Asia. The worldwide PDB (wwPDB) was established in 2003 to formally recognize the international nature of the PDB archive (2,4) and to ensure that the data files remain uniform in content and format. The founding members are the RCSB PDB (USA) (1), the Macromolecular Structure Database at the European Bioinformatics Institute (MSD-EBI) (5) and the Protein Data Bank Japan (PDBj) at Osaka University. These wwPDB sites share responsibilities in data deposition, processing and distribution of the PDB archive, and agree to support a single, standardized archive of structural data (Table 1). The BioMagResBank (BMRB) at the University of Wisconsin-Madison (USA) (6) became a member in 2006 and will be a deposition site for primary experimental data and PDB data.

A wwPDB Advisory Committee (wwPDBAC) consists of representatives appointed by each member site as well as representatives of the international X-ray, NMR and electron microscopy (EM) communities. wwPDBAC meets yearly and provides advice about policies governing the content, format and distribution of the PDB data files.

The website (http://www.wwpdb.org/) contains the formal agreement for the operation of the wwPDB organization, links to the deposition and access sites, and news and announcements about policies and projects related to the wwPDB.

MEMBER DEPOSITION SITES

The advances in protein cloning, expression, labeling, purification through to structure determination has resulted in a rapid increase in the rate at which new protein structures are determined. Progress is also being made in structure determinations of nucleic acids, particularly RNA molecules. A key component of the wwPDB is that its tools are able to efficiently capture and curate data as the amount deposited

*To whom correspondence should be addressed. Tel: +1 732 445 4667; Fax: +1 732 445 4320; Email: berman@rcsb.rutgers.edu

Table 1. wwPDB Data deposition and access sites

	Access PDB FTP	Deposit data	Main website
RCSB PDB	ftp://ftp.rcsb.org/pub/pdb	http://deposit.rcsb.org	http://www.pdb.org
MSD-EBI	ftp://ftp.ebi.ac.uk/pub/databases/rcsb/pdb	http://www.ebi.ac.uk/msd-srv/autodep4	http://www.ebi.ac.uk/msd
PDBj	ftp://pdb.protein.osaka-u.ac.jp/pub/pdb	http://www.pdbj.org/deposit.html	http://www.pdbj.org
BMRB		http://batfish.bmrb.wisc.edu/bmrb-adit/YES	http://www.bmrb.wisc.edu

Table 2. PDB structures deposited and processed by year and site (as of August 28, 2006)

Year	Total depositions	Deposited to			Processed by		
		RCSB PDB	PDBj	EBI	RCSB PDB	PDBj	EBI
2000	2983	2445	10	528	2294	161	528
2001	3286	2673	118	495	2407	384	495
2002	3563	2769	289	505	2401	657	505
2003	4830	3488	673	669	3135	1026	669
2004	5508	3796	900	812	3083	1613	812
2005	6677	4506	1166	1005	3562	2110	1005
2006	4728	3239	725	764	2659	1305	764
Total	31 575	22 916	3881	4778	19 545	7252	4778

Table 3. PDB structures released per year (experimentally solved structures only, as of August 28, 2006)

Year	Total
2000	2632
2001	2840
2002	3018
2003	4185
2004	5230
2005	5421
2006	4154
Total	27 480

grows exponentially (Table 1). Although the sites are physically dispersed and use three different tools for data capture and processing (ADIT, ADIT-NMR and AutoDep), all the data are annotated and processed using common standards. To ensure that the core data are represented uniformly, the wwPDB sites actively collaborate to exchange core reference information (e.g. the dictionary description for ligands) and to ensure that standard practices are followed. The annotators at all sites maintain daily communication via video teleconferencing, exchange visits and email; they are currently extending and updating the annotation manuals that will be made publicly available.

Every week, the data processed at each site are forwarded to the RCSB PDB for inclusion in the archive. At present, the RCSB PDB is the archive keeper and as such has sole write access to the PDB archive.

Statistics about the PDB structures deposited and processed by the wwPDB are available from http://www.wwpdb.org/stats.html (Tables 2 and 3).

DATA ACCESS: MEMBER FTP AND WEBSITES

The 'PDB archive' is the collection of flat files that are maintained in three different formats: the legacy PDB file format; the PDB exchange format that follows the mmCIF syntax (http://deposit.pdb.org/mmcif/); and the PDBML/XML format (7) that is a direct translation of the PDB exchange format. Each wwPDB site distributes the same PDB archive via FTP. The archive is updated weekly.

Time-stamped snapshots of the PDB archive are added each year to ftp://snapshots.rcsb.org. They provide a frozen copy of the archive as it appeared at that time for research and historical purposes. The most recent snapshot was added in January 2006. It includes the 34 421 experimentally determined coordinate files that were current (i.e. not obsolete) as of January 3, 2006, and the directory containing the frozen content as of January 6, 2005. Scripts are available to download all, or part, of a snapshot automatically.

In addition to providing access to the PDB archive, each wwPDB site provides databases and websites that provide different views and analyses of the structural data contained within the PDB archive (8–14).

DATA UNIFORMITY

wwPDB members collaborate to ensure the uniformity of the PDB archive. The PDB Exchange Dictionary consolidates content from a variety of dictionaries and includes extensions to describe NMR, EM and protein production data (15). wwPDB data processing, exchange and annotation depend upon this dictionary and the mmCIF format (16) to help make the data more consistent across the archive.

In the past, query across the complete PDB archive has been limited by missing, erroneous and inconsistently reported data, nomenclature and functional annotation. The evolution of experimental methods, functional knowledge of proteins and methods used to process these data has introduced various inconsistencies into the PDB archive and has inspired different versions of the PDB format.

Over the years, the MSD-EBI, PDBj and the RCSB PDB have been working individually on correcting errors in the archive. Under the wwPDB banner, these groups are now working to integrate all remediation efforts into a single consistent collection of data files. This work includes improving the representation of PDB small molecule data, assessing the required chemical definitions and their correspondences in PDB entries, resolving any remaining differences in the macromolecular sequences assigned by each group and resolving differences in primary citation assignments. The BMRB has been collaborating with MSD-EBI and RCSB PDB on standardizing restraint data associated with PDB depositions (17,18).

The remediated data (PDB V.2) will be made available for public review in 2007 and will form the basis of the wwPDB websites. The data released before remediation (PDB V.1) will continue to be available for the historical record.

PHASING OUT THEORETICAL MODEL DEPOSITIONS TO THE PDB ARCHIVE

Effective October 15, 2006, PDB depositions were restricted to atomic coordinates that are substantially determined by experimental measurements on specimens containing biological macromolecules. This policy was recommended and endorsed by a working group composed of structural and computational biologists and endorsed by the wwPDB Advisory Committee. Thus, theoretical model depositions (such as models determined purely *in silico* using, for example, homology or *ab initio* methods) will no longer be accepted.

NEWS AND ANNOUNCEMENTS

The News sections of the wwPDB website gives information about the outcome of the wwPDBAC meetings and policy statements affecting the PDB data files. A recent example is the announcement of the policy for the archiving of *in silico* models (19).

ACKNOWLEDGEMENTS

The RCSB PDB is operated by Rutgers, The State University of New Jersey and the San Diego Supercomputer Center and the Skaggs School of Pharmacy and Pharmaceutical Sciences at the University of California, San Diego. It is supported by funds from the National Science Foundation, the National Institute of General Medical Sciences, the Office of Science, Department of Energy, the National Library of Medicine, the National Cancer Institute, the National Center for Research Resources, the National Institute of Biomedical Imaging and Bioengineering, the National Institute of Neurological Disorders and Stroke and the National Institute of Diabetes and Digestive and Kidney Diseases. E-MSD gratefully acknowledges the support of the Wellcome Trust (GR062025MA), the EU (TEMBLOR, NMRQUAL and IIMS), CCP4, the BBSRC, the MRC and EMBL. PDBj is supported by grant-in-aid from the Institute for Bioinformatics Research and Development, Japan Science and Technology Agency (BIRD-JST), and the Ministry of Education, Culture, Sports, Science and Technology (MEXT). The BMRB is supported by NIH grant LM05799 from the National Library of Medicine. Funding to pay the Open Access publication charges for this article was provided by the agencies supporting the RCSB PDB.

Conflict of interest statement. None declared.

REFERENCES

1. 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.
2. Berman,H.M., Henrick,K. and Nakamura,H. (2003) Announcing the worldwide Protein Data Bank. *Nature Struct. Biol*., **10**, 980.
3. Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F.,Jr, Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol*., **112**, 535–542.
4. Henrick,K., Berman,H.M. and Nakamura,H. (2005) The Protein Data Bank and the wwPDB. In Jorde,L.B., Little,P.F.R., Dunn,M.J. and Subramaniam,S. (eds), *Encyclopedia of Genomics, Proteomics, and Bioinformatics*. John Wiley & Sons Ltd, Chichester, Vol. 7, pp. 3335–3339.
5. Golovin,A., Oldfield,T.J., Tate,J.G., Velankar,S., Barton,G.J., Boutselakis,H., Dimitropoulos,D., Fillon,J., Hussain,A., Ionides,J.M. *et al*. (2004) E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res*., **32**, D211–D216.
6. Ulrich,E.L., Markley,J.L. and Kyogoku,Y. (1989) Creation of a nuclear magnetic resonance data repository and literature database. *Protein Seq. Data Anal*., **2**, 23–37.
7. Westbrook,J., Ito,N., Nakamura,H., Henrick,K. and Berman,H.M. (2005) PDBML: the representation of archival macromolecular structure data in XML. *Bioinformatics*, **21**, 988–992.
8. Deshpande,N., Addess,K.J., Bluhm,W.F., Merino-Ott,J.C., Townsend-Merino,W., Zhang,Q., Knezevich,C., Xie,L., Chen,L., Feng,Z. *et al*. (2005) The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res*., **33**, D233–D237.
9. Kouranov,A., Xie,L., de la Cruz,J., Chen,L., Westbrook,J., Bourne,P.E. and Berman,H.M. (2006) The RCSB PDB information portal for structural genomics. *Nucleic Acids Res*., **34**, D302–D305.
10. Tagari,M., Tate,J., Swaminathan,G.J., Newman,R., Naim,A., Vranken,W., Kapopoulou,A., Hussain,A., Fillon,J., Henrick,K. *et al*. (2006) E-MSD: improving data deposition and structure quality. *Nucleic Acids Res*., **34**, D287–D290.
11. Henrick,K. and Thornton,J.M. (1998) PQS: a protein quarternary file server. *Trends Biochem. Sci*., **23**, 358–361.
12. Kinoshita,K. and Nakamura,H. (2004) eF-site and PDBjViewer: database and viewer for protein functional sites. *Bioinformatics*, **20**, 1329–1330.
13. Standley,D.M., Toh,H. and Nakamura,H. (2005) GASH: an improved algorithm for maximizing the number of equivalent residues between two protein structures. *BMC Bioinformatics*, **6**, 221.
14. Wako,H., Kato,M. and Endo,S. (2004) ProMode: a database of normal mode analyses on protein molecules with a full-atom model. *Bioinformatics*, **20**, 2035–2043.
15. Westbrook,J., Yang,H., Feng,Z. and Berman,H.M. (2005) The use of mmCIF architecture for PDB data management. In Hall,S.R. and McMahon,B. (eds), *International Tables for Crystallography*. Springer, Dordrecht, The Netherlands, Vol. G, pp. 539–543.
16. Fitzgerald,P.M.D., Westbrook,J.D., Bourne,P.E., McMahon,B., Watenpaugh,K.D. and Berman,H.M. (2005) Macromolecular dictionary (mmCIF). In Hall,S.R. and McMahon,B. (eds), *International Tables for Crystallography*. Springer, Dordrecht, The Netherlands, Vol. G, pp. 295–443.
17. Doreleijers,J.F., Mading,S., Maziuk,D., Sojourner,K., Yin,L., Zhu,J., Markley,J.L. and Ulrich,E.L. (2003) BioMagResBank database with sets of experimental NMR constraints corresponding to the structures of over 1400 biomolecules deposited in the Protein Data Bank. *J. Biomol. NMR*, **26**, 139–146.
18. Doreleijers,J.F., Nederveen,A.J., Vranken,W., Lin,J., Bonvin,A.M., Kaptein,R., Markley,J.L. and Ulrich,E.L. (2005) BioMagResBank databases DOCR and FRED containing converted and filtered sets of experimental NMR restraints and coordinates from over 500 protein PDB structures. *J. Biomol. NMR*, **32**, 1–12.
19. Berman,H.M., Burley,S.K., Chiu,W., Sali,A., Adzhubei,A., Bourne,P.E., Bryant,S.H., Dunbrack,J.R.L., Fidelis,K., Frank,J. *et al*. (2006) Outcome of a workshop on archiving structural models of biological macromolecules. *Structure*, **14**, 1211–1217.