Nucleic Acids Research, 2004, Vol. 32, Database issue ***D223–D225***
DOI: 10.1093/nar/gkh096

The distribution and query systems of the RCSB Protein Data Bank

Philip E. Bourne[1,2,*], Kenneth J. Addess[1], Wolfgang F. Bluhm[1], Li Chen[3], Nita Deshpande[1], Zukang Feng[3], Ward Fleri[1], Rachel Green[3], Jeffrey C. Merino-Ott[1], Wayne Townsend-Merino[1], Helge Weissig, John Westbrook[3] and Helen M. Berman[3]

[1]San Diego Supercomputer Center, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0505, USA, [2]Department of Pharmacology, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0636, USA and [3]Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, 610 Taylor Road, Piscataway, NJ 08854-8087, USA

Received September 16, 2003; Revised and Accepted October 7, 2003

ABSTRACT

The Protein Data Bank (PDB; http://www.pdb.org) is the primary source of information on the 3D structure of biological macromolecules. The PDB's mandate is to disseminate this information in the most usable form and as widely as possible. The current query and distribution system is described and an alpha version of the future re-engineered system introduced.

INTRODUCTION

The Protein Data Bank (PDB) (1,2) is the single worldwide repository for data on the 3D structure of biological macromolecules, which is available from a website (http://www.pdb.org/), an FTP site (ftp://ftp.rcsb.org/) and a number of mirror websites worldwide (Table 1). The data are derived experimentally primarily from X-ray crystallography, nuclear magnetic resonance (NMR) and cryo-electron microscopy. Theoretical models are also accepted but not curated by the PDB, and are accessible only from the ftp archive to distinguish them from the experimental archive. In previous years of this volume, we have introduced the PDB as operated by the Research Collaboratory for Structural Bioinformatics (RCSB) (1), provided details of efforts to unify the data that have been collected over a 30 year period (3,4) and most recently described the PDB's response to the structural genomics initiatives worldwide (5).

Here another feature of the PDB's work is described—the distribution and query of PDB data. Query refers to access of PDB data from the RCSB PDB websites, and distribution refers to obtaining files of PDB data via the web, ftp and CD-ROM. Future distribution will also provide access to PDB data via CORBA and web services. While the current production system (1) has been continuously improved, the success of the data uniformity project (3,4) and new information technologies have enabled the development of a newly re-engineered system. At the time of this writing (September 2003), the PDB has an internal alpha release of this new system and it is anticipated that a public beta release will be available in the first quarter of 2004. Details will be made available on the PDB website.

CONTENT

When referring to the current and new query and distribution system in the sections that follow, it is important to associate each of them with the appropriate data content. The current web system is built from PDB files that have been available to the community for a number of years with the addition of data resulting from efforts in uniformity processing (3,4). Current PDB data distribution consists of the original PDB files, mmCIFs (6) that include all uniform data and XML files that are translations of these mmCIFs. All data collected since 1999 contain additional data items conforming to the mmCIF dictionary and have been subject to improved validation procedures. The re-engineered website is built from the mmCIFs and includes some derived data to facilitate browsing of the PDB. For example, the assignment of gene ontology terms (7) and any relationship to disease as taken from the Online Mendelian Inheritance in Man (OMIM) database (http://www.ncbi.nlm.nih.gov/omim).

THE CURRENT QUERY AND DISTRIBUTION SYSTEM

Architecture

The architecture of the current system has been described previously (1) and consists of four data sources derived from the original PDB files: a SYBASE relational database, a Property Object Model (POM) database (8), a Lucene (http://jakarta.apache.org) index and the PDB files themselves.

*To whom correspondence should be addressed. Tel: +1 858 534 8301; Fax: +1 858 822 0873; Email: bourne@sdsc.edu
Present address:
Helge Weissig, ActivX Biosciences Inc., 11025 North Torrey Pines Road, Suite 120, La Jolla, CA 92037, USA

Table 1. RCSB PDB mirror sites

Location	URL
San Diego Supercomputer Center, USA	http://www.pdb.org
Rutgers University, USA	http://rutgers.rcsb.org
National Institute of Standards and Technology, USA	http://nist.rcsb.org
Cambridge Crystallographic Data Center, UK	http://pdb.ccdc.cam.ac.uk
National University of Singapore, Singapore	http://pdb.bic.nus.edu.sg
Osaka University, Japan	http://pdb.protein.osaka-u.ac.jp
Universidade Federal de Minas Gerais, Brazil	http://www.pdb.ufmg.br
Max Delbrück Center for Molecular Medicine, Germany	http://www.pdb.mdc-berlin.de

Capabilities: query

Capabilities of the current query system were first described in Berman *et al*. (1). Briefly, a Perl/CGI web layer wraps around the databases described above and provides several query interfaces with keyword searches, searches of the relational database content, etc.

In the past 4 years, we have made many significant improvements to this system. Additional query functionalities include an enzyme classification browser and single-click searches for authors, EC numbers and ligands that were displayed for a structure from a previous search.

The ability to remove similar sequences both pre- and post-query has been added, based on weekly clustering of all protein sequences of >20 residues using the cd-hit program (9). For further details, see http://www.pdb.org/pdb/redundancy.html.

Netscape's LDAP keyword search has been replaced with the more efficient and robust Lucene keyword search. Most importantly, Lucene queries an index of the curated mmCIF files (rather than the original PDB files), which provides more accurate query results.

Several graphical viewers, such as MICE (10), STING (11) and the Swiss-PDB Viewer (12), have been added for displaying the crystallographic asymmetric unit. More importantly, graphical views and the Cartesian coordinates of the complete biological molecule are now provided. This is particularly relevant for virus structures where the application of both non-crystallographic and crystallographic symmetry is required to generate the protein capsid. For entries released since 1999, this information has been verified by the depositor of the data. For entries released prior to 1999, this information is generated with reference to the Protein Quaternary Structure (PQS) server (13) and Swiss-Prot (14).

Capabilities: distribution

The RCSB PDB is responsible for the distribution of data files in PDB, mmCIF and XML formats. These data are contributed by the RCSB, the MSD at the European Bioinformatics Institute (EBI; http://www.ebi.ac.uk/msd/) and PDBj (http://www.pdbj.org/) at Osaka University, Japan. These organizations are committed to the maintenance of a single PDB archive, and access to this data is provided by a variety of worldwide resources. Table 1 describes the RCSB resource and associated mirrors. The data distributed through the RCSB's websites and ftp archives include sequences and complete structural descriptions in PDB, mmCIF and XML formats. These data are available via various compression formats. The structure of the ftp archive is given at http://www.pdb.org/pdb/ftp_plan.html. Users may also mirror the complete ftp archive at their local sites. Several software solutions for this purpose are provided at ftp://ftp.rcsb.org/pub/pdb/software/.

FUTURE PLANS

The RCSB PDB is now in the process of re-engineering its site and database, using feedback derived from the PDB help desk, conference attendance, focus groups and other personal interactions between the users of the PDB and RCSB staff. This site is expected to be available for public testing (beta) in the first quarter of 2004. The new system has been designed using an Enterprise Java framework and is based on a three-tier model—underlying database, presentation layer and middle tier connecting them. The current query system and associated schemas cannot take full advantage of the successful efforts to unify the data across the entire PDB archive (3,4). Therefore, extensive efforts have gone into redesigning a relational database built entirely from the curated mmCIF files, which will allow improved query access to the unified data. Future distribution of PDB data will use both CORBA and web services. Users wishing to establish a CORBA server may do so now using C++ (http://deposit.pdb.org/mmcif/FILM/) or the Java OpenMMS software (http://openmms.sdsc.edu) (15). The complete application program interface (API) based on mmCIF and details of how to access web services will be available with the beta release of the new system.

CONCLUSION

The PDB's mandate extends to providing accurate and timely structural information to a worldwide community of users regardless of local hardware and software and geographic location. The current and future query and distribution system strives to achieve this mandate based on input from a wide variety of users. Comments and suggestions are always welcome by sending email to info@rcsb.org.

ACKNOWLEDGEMENTS

The PDB is operated by Rutgers, The State University of New Jersey; the San Diego Supercomputer Center at the University of California, San Diego; and the Center for Advanced Research in Biotechnology (CARB) at the National Institute of Standards and Technology (NIST)—three members of the Research Collaboratory for Structural Bioinformatics (RCSB). This work is supported by grants from the National Science Foundation, the Department of Energy and the National Institutes of Health.

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. 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.
3. Bhat,T.N., Bourne,P., Feng,Z., Gilliland,G., Jain,S., Ravichandran,V., Schneider,B., Schneider,K., Thanki,N., Weissig,H. *et al.* (2001) The PDB data uniformity project. *Nucleic Acids Res.*, **29**, 214–218.
4. Westbrook,J., Feng,Z., Jain,S., Bhat,T.N., Thanki,N., Ravichandran,V., Gilliland,G.L., Bluhm,W., Weissig,H., Greer,D.S. *et al.* (2002) The Protein Data Bank: unifying the archive. *Nucleic Acids Res.*, **30**, 245–248.
5. Westbrook,J., Feng,Z., Chen,L., Yang,H. and Berman,H.M. (2003) The Protein Data Bank and structural genomics. *Nucleic Acids Res.*, **31**, 489–491.
6. Bourne,P.E., Berman,H.M., Watenpaugh,K., Westbrook,J.D. and Fitzgerald,P.M.D. (1997) The macromolecular Crystallographic Information File (mmCIF). *Methods Enzymol.*, **277**, 571–590.
7. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.
8. Shindyalov,I.N. and Bourne,P.E. (1997) Protein data representation and query using optimized data decomposition. *Comput. Appl. Biosci.*, **13**, 487–496.
9. Li,W., Jaroszewski,L. and Godzik,A. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, **17**, 282–283.
10. Tate,J.G., Moreland,J.L. and Bourne,P.E. (2001) Design and implementation of a collaborative molecular graphics environment. *J. Mol. Graph. Model.*, **19**, 280–287, 369–273.
11. Neshich,G., Togawa,R.C., Mancini,A.L., Kuser,P.R., Yamagishi,M.E., Pappas,G.,Jr, Torres,W.V., Fonseca e Campos,T., Ferreira,L.L., Luna,F.M. *et al.* (2003) STING Millennium: a web-based suite of programs for comprehensive and simultaneous analysis of protein structure and sequence. *Nucleic Acids Res.*, **31**, 3386–3392.
12. Guex,N., Diemand,A. and Peitsch,M.C. (1999) Protein modelling for all. *Trends Biochem. Sci.*, **24**, 364–367.
13. Henrick,K. and Thornton,J.M. (1998) PQS: a protein quaternary structure file server. *Trends Biochem. Sci.*, **23**, 358–361.
14. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
15. Greer,D.S., Westbrook,J.D. and Bourne,P.E. (2002) An ontology driven architecture for derived representations of macromolecular structure. *Bioinformatics*, **18**, 1280–1281.