© 2002 Oxford University Press Nucleic Acids Research, 2002, Vol. 30, No. 1 245–248

The Protein Data Bank: unifying the archive

John Westbrook, Zukang Feng, Shri Jain, T. N. Bhat[1], Narmada Thanki[1], Veerasamy Ravichandran[1], Gary L. Gilliland[1], Wolfgang F. Bluhm[2], Helge Weissig[2], Douglas S. Greer[2], Philip E. Bourne[2,3,4] and Helen M. Berman*

Rutgers, The State University of New Jersey, Department of Chemistry, 610 Taylor Road, Piscataway, NJ 08854-8087, USA, [1]National Institute of Standards and Technology, Route 270, Quince Orchard Road, Gaithersburg, MD 20899, USA, [2]San Diego Supercomputer Center, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0505, USA, [3]Department of Pharmacology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0500, USA and [4]The Burnham Institute, 10901 North Torrey Pines Road, La Jolla, CA 92037, USA

Received September 25, 2001; Accepted October 1, 2001

ABSTRACT

The Protein Data Bank (PDB; http://www.pdb.org/) is the single worldwide archive of structural data of biological macromolecules. This paper describes the progress that has been made in validating all data in the PDB archive and in releasing a uniform archive for the community. We have now produced a collection of mmCIF data files for the PDB archive (ftp://beta.rcsb.org/pub/pdb/uniformity/data/mmCIF/). A utility application that converts the mmCIF data files to the PDB format (called CIFTr) has also been released to provide support for existing software.

INTRODUCTION

The PDB is the single archive of biological macromolecular structures (1,2). From July 1, 2000 to June 30, 2001, a total of 3148 structures were deposited with the PDB. Full data processing of entries by the Research Collaboratory for Structural Bioinformatics (RCSB), including author revisions, averages <2 weeks. The number and complexity of the entries continues to increase. In 1999, 2670 structures were deposited containing 1 million residues; in 2000, 2995 structures were deposited containing 1.3 million residues. This is a 30% increase in residues per year.

The access and distribution of the archival data is through the primary Web site at UCSD and through mirrors located throughout the world (Table 1). The PDB receives an average of 115 000 hits per day on the primary Web site alone. As of September 19, 2001, there are more than 16 000 structures in the PDB. The demographics of the current holdings are shown at http://www.rcsb.org/pdb/holdings.html.

Although we have continued to improve our query capabilities, the lack of uniform data, a result of the evolution of the format and content of the PDB over a 30 year period, necessarily limits our ability to provide reliable searches and the community's ability to perform quantitative science. In order to improve the querying capabilities of the PDB, we first addressed the uniformity of key data records for all entries in the PDB (3). Several records were targeted for this type of remediation: macromolecular names and synonyms, source organism, R-factor, resolution, enzyme names and classification, and primary citation. The results of record-wise uniformity processing are stored in the PDB relational database. This information is available as database output, in PDB reports and in Structure Explorer pages.

The practical effect of this type of data processing is that it makes queries on these records more reliable. For example, it is now possible to perform complex queries on enzymes and enzyme classes in a way that was not possible with the archival data. However, while this type of uniformity processing improves PDB queries, it does not produce improved files.

Changing an existing archive such as the PDB to comply with evolving data and nomenclature standards and imposing consistency constraints in data representation presents a variety of problems. Such changes, no matter how well

Table 1. PDB mirror sites

RCSB partner sites	
SDSC, La Jolla, CA	http://www.pdb.org/ ftp://ftp.rcsb.org/
Rutgers University, Piscataway, NJ	http://rutgers.rcsb.org/
NIST, Gaithersburg, MD	http://nist.rcsb.org/
Other RCSB mirrors	
CCDC, UK	http://pdb.ccdc.cam.ac.uk/ ftp://pdb.ccdc.cam.ac.uk/rcsb/
National University of Singapore	http://pdb.bic.nus.edu.sg/ ftp://pdb.bic.nus.edu.sg/pub/pdb/
Osaka University, Japan	http://pdb.protein.osaka-u.ac.jp/ ftp://pdb.protein.osaka-u.ac.jp/
Universidade Federal de Minas Gerais, Brazil	http://www.pdb.ufmg.br/ ftp://vega.cenapad.ufmg.br/pub/pdb/

*To whom correspondence should be addressed. Tel: +1 732 445 4667; Fax: +1 732 445 4320; Email: berman@rcsb.rutgers.edu intended, may corrupt references to a wide variety of published and Web-accessible work. However, from the perspective of a new PDB user who is not connected with the archive's rich history, the state of the PDB with respect to uniformity is difficult to reconcile. To participate in current and future scientific challenges, the PDB must advance to a level of data quality that facilitates systematic archive-wide analyses and integration with other biological and structural databases. The path to attain this level of data quality and maintain historical continuity is a difficult one with many trade-offs. We describe here how we have addressed this difficult issue and what we have done to create a new set of uniform PDB entries.

DATA PROCESSING OF NEW ENTRIES

In order to provide the community with high quality data, the RCSB has developed a number of tools that support the deposition and processing of X-ray and NMR structures. For depositing structures, the integrated web-based user interface the AutoDep Input Tool (ADIT, http://deposit.pdb.org/adit/) takes data from an uploaded file and presents a Web-based editor to modify and make additions to an entry. Data processing involves checking various aspects of the structure and the data collected through ADIT. Deposited information is converted to mmCIF representation and is subsequently processed by PDB validation programs. Over the years, many programs and procedures have been developed to diagnose the errors in PDB files (4–7). These programs have allowed authors to detect errors and correct them prior to deposition in the PDB. The PDB has incorporated many of these methods and has developed a series of procedures to review and validate structures.

A skilled annotator reviews the output of these validation checks. A distilled summary report of the validation diagnostics is forwarded to depositors. Since the author is the most knowledgeable about his/her own structure, the PDB collaborates with the author to help ensure that the structure that is ultimately released to the public is the best possible representation of the results of the experiment.

The PDB validation report summarizes results of the following checks: stereochemistry, close contacts in asymmetric unit and unit cell, occupancy, sequence in PDB SEQRES records and coordinates, distant waters, experimental data [SFCHECK (8)], comparison with standard values (9–11).

This validation software allows the user to check the format of coordinate and structure factor files and to perform a variety of validation tests on the structure prior to deposition in the PDB. These checks can be done independently by the user via the Validation Server on the Web at http://deposit.pdb.org/validate/ or by downloading the software from http://deposit.pdb.org/software/. The format precheck and validation steps are also optional steps of the ADIT deposition process.

REPROCESSING OF LEGACY FILES

In addressing uniformity issues with the legacy data, we have focused on formatting, nomenclature and sequence structure consistency. The decision to concentrate on these aspects is the result of many discussions with the community of PDB users. These modifications have been made very conservatively and do not change the coordinates of the structural model. To do this we have applied the software developed for primary data processing for the validation and standardization of the 8368 data files released into the archive prior to October 1998. The gain in efficiency in data processing has been largely transferable to the task of reprocessing legacy files. The combination of improved software and the experience gained from 3 years of primary processing has made it possible to attempt a more automated remediation of the legacy data in a batch mode. The classes of errors that have now been remediated are described in the following sections.

Sequence representation

The complete polymer sequence for the macromolecule under study is encoded in PDB SEQRES records as a list of three-letter residue codes. These records are intended to describe the full polymer sequence for the macromolecule or domain for which coordinates are deposited.

In comparing the legacy sequence data with data from sequence databases (12,13), cases were found in which the legacy sequence was incorrect. In most cases, these sequence errors reflect gaps in the model sequence or incompletely modeled residues where residues or side-chains were not experimentally observed. In all of these cases, the sequences were updated with the correct or missing residues. In some instances, two PDB chains were used to represent a single polymer with a residue gap. These sequences were consolidated into single PDB chains.

Sequence/coordinate mismatches

Sequence information can also be derived from PDB coordinate records. Since coordinate data may not be deposited for all of the residues in structure, the PDB SEQRES records are provided to define the full chemical sequence. Even though the coordinate records may not provide complete sequence information, the sequence information in the SEQRES and coordinate records should be consistent. We found 90 cases in the legacy data in which SEQRES and coordinate records were non-corresponding. The majority of these inconsistencies result from labeling residues in the coordinate records missing side-chains as alanines. Only four of the 90 cases could not be reconciled on the basis of a missing side chain.

Atom and ligand nomenclature

The most common problems found in the legacy data are related to the labeling of atoms and ligands. Atom nomenclature problems were found in 3311 (40%) of the legacy files. The labeling of terminal atoms was found to be the most common nomenclature error. Atoms adjacent to a gap of unobserved residues in continuous sequence were most commonly mislabeled as terminal atoms. All errors of this type were automatically corrected.

Labeling of ligand atoms and residues was the second most common nomenclature problem. Ligand atom names were standardized in software to the nomenclature used in the PDB ligand dictionary. This was accomplished by topology matching against the chemical descriptions in the dictionary. New ligand descriptions were created and added to the dictionary where necessary.

Another common nomenclature problem arises from the duplication of atom labels. Redundant atom labels were found in 636 legacy data files. This was most commonly the result of the mislabeling of alternate conformations. In a small number of cases, identical coordinate records were duplicated. All instances of duplicated atom records were resolved.

Stereochemical labeling

Perhaps the most serious class of errors in atom nomenclature is that related to stereochemistry. Errors in chirality were found in 549 legacy files. Only 255 of these cases could be resolved as errors in atom labeling; the remainder represents exceptions to current stereochemical conventions.

REVISITING RECENTLY PROCESSED ENTRIES

We also reviewed all the data that had been processed by the RCSB since October 1998. The re-validation of the 3150 files that we processed prior to January 2000 showed five entries with conflicts in sequence information between SEQRES and coordinate records, 162 errors in atom and ligand nomenclature, 19 duplicated atom labels, three errors in stereochemical labeling and 30 terminal atom labeling errors. The largest number of errors was related to ligand atom nomenclature. These resulted from changes in our ligand dictionary, which underwent significant correction and development during 1999. Any remaining errors were undetected by our software or were omissions in our annotation procedures during this period.

The results for files processed after January 2000 show further improvement. In this group of 3569 files, we found 31 errors in atom and ligand nomenclature, one duplicated atom label, three errors in stereochemical labeling and two terminal atom labeling errors. No sequence inconsistencies were detected. All of these errors were corrected and these corrections are described within each entry in the PDB revision records.

INTEGRATION AND DELIVERY OF UNIFIED DATA

The final step in this process was the integration of the results of record by record processing with the batch processing of all the entries in the archive. The results of the uniformity and data integration project are being delivered as a collection of mmCIF data files. The data items within these files are described in the PDB exchange data dictionary. This dictionary, which includes the data items in the standard mmCIF dictionary along with PDB extensions, is available at the PDB mmCIF Resource site (http://deposit.pdb.org/mmcif/).

The mmCIF data files are available on the PDB beta-ftp site for all legacy data and for current files deposited and processed with the PDB (ftp://beta.rcsb.org/pub/pdb/uniformity/data/mmCIF/). The beta release of these data files is to allow users to evaluate and comment. PDB will continue to correct and improve the uniformity of these data in response to user input.

SUPPORT FOR THE PDB AND OTHER FORMATS

In recognition that many software applications require the PDB format, we have provided a software tool (CIFTr) that translates the mmCIF into PDB format. The tool provides options that permit users to select their particular nomenclature preference. For instance, it is possible to select between the nomenclature used when the file was originally released and the nomenclature resulting from uniformity processing. In the future, CIFTr will provide translation to other file formats such as XML.

CIFTr is available for download from http://deposit.pdb.org/software/ for SGI, Linux, Alpha and SUN platforms.

THE FUTURE

With the mmCIFs from the data uniformity project as a base, we will now examine the data items that were not included in our initial uniformity project. In particular we will examine the details of experimental data collection and refinement that are currently embedded in unstructured REMARK records in the older PDB files. As much as possible we will attempt to extract information from the text of these remarks and populate the corresponding mmCIF data items.

We are now redesigning the underlying PDB core relational database to take advantage of the new uniform and self-consistent data files. Owing to the greater internal consistency within the mmCIF datasets, the new database implementation will provide the ability to construct queries that span the range of structural detail from biological assembly to individual atoms.

Questions and comments about the PDB should be sent to info@rcsb.org.

ACKNOWLEDGEMENTS

The PDB team consists of the authors listed and Peter Arzberger, Bryan Banister, Tammy Battistuz, Kyle Burkhardt, Li Chen, Victoria Colflesh, Nita Deshpande, Phoebe Fagan, Ward Fleri, Michael Gribskov, Diane Hancock, Lisa Iype, Brad Kroeger, Jessica Marvin, David Padilla, Gnanesh Patel, Bohdan Schneider, Thomas Solomon, Lynn Ten Eyck, Michael Tung, Rosalina Valera and Christine Zardecki. We acknowledge the work of Dietmar Schomburg (Institute of Biochemistry at the University of Cologne), project leader of the BRENDA database, for enzyme classification remediation, the comments and corrections that have been contributed by colleagues at the NCBI, the EBI, the CCDC and the many users of the PDB. We also acknowledge the support of Angela Loh and Compaq Computer Corporation for the generous gift of hardware that has help to support the computational requirements of this project. This work is supported by grants from the National Science Foundation, the Office of Biological and Environmental Research at the Department of Energy, and two units of the National Institutes of Health: the National Institute of General Medical Sciences and the National Library of Medicine.

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., Meyer,E.E., 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., Westbrook,J. and Berman,H.M. (2001) The PDB data uniformity project. *Nucleic Acids Res.*, **29**, 214–218.

4. Laskowski,R.A., McArthur,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.
5. Laskowski,R.A., Rullmann,J.A., MacArthur,M.W., Kaptein,R. and Thornton,J.M. (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR*, **8**, 477–486.
6. Hooft,R.W., Sander,C. and Vriend,G. (1996) Verification of protein structures: side-chain planarity. *J. Appl. Crystallogr.*, **29**, 714–716.
7. Hooft,R.W., Vriend,G., Sander,C. and Abola,E.E. (1996) Errors in protein structures. *Nature*, **381**, 272.
8. Vaguine,A.A., Richelle,J. and Wodak,S.J. (1999) SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure–factor data and their agreement with the atomic model. *Acta Crystallogr.*, **D55**, 191–205.
9. Engh,R.A. and Huber,R. (1991) Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallogr.*, **A47**, 392–400.
10. Gelbin,A., Schneider,B., Clowney,L., Hsieh,S.-H., Olson,W.K. and Berman,H.M. (1996) Geometric parameters in nucleic acids: sugar and phosphate constituents. *J. Am. Chem. Soc.*, **118**, 519–528.
11. Clowney,L., Jain,S.C., Srinivasan,A.R., Westbrook,J., Olson,W.K. and Berman,H.M. (1996) Geometric parameters in nucleic acids: nitrogenous bases. *J. Am. Chem. Soc.*, **118**, 509–518.
12. Bairoch,A. and Boeckmann,B. (1994) The SWISS-PROT protein sequence databank: current status. *Nucleic Acids Res.*, **22**, 3578–3580.
13. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,D.L. (2000) GenBank. *Nucleic Acids Res.*, **28**, 15–18. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 17–20.