The PDB data uniformity project

T. N. Bhat, Philip Bourne[1,2], Zukang Feng[3], Gary Gilliland, Shri Jain[3], Veerasamy Ravichandran, Bohdan Schneider[3], Kata Schneider[3], Narmada Thanki, Helge Weissig[1], John Westbrook[3] and Helen M. Berman[3,*]

Biotechnology Division, National Institute of Standards and Technology, Gaithersburg, MD 20899-8310, USA, [1]San Diego Supercomputer Center, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0537, USA, [2]Department of Pharmacology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0500, USA and [3]Department of Chemistry, Rutgers University, 610 Taylor Road, Piscataway, NJ 08854-8087, USA

Received October 3, 2000; Revised and Accepted October 17, 2000

ABSTRACT

The Protein Data Bank (PDB; http://www.rcsb.org/pdb/) is the single worldwide archive of structural data of biological macromolecules. This paper describes the data uniformity project that is underway to address the inconsistency in PDB data.

INTRODUCTION

The Protein Data Bank (PDB) is the single archive of biological macromolecular structures (1). The Research Collaboratory for Structural Bioinformatics (RCSB; http://www.rcsb.org/) has been fully responsible for its management since July 1, 1999. The architecture and functionality of the systems used to collect, archive, distribute and query the data were described previously (2). In the past year the deposition rate has increased, with 2693 structures deposited to the PDB in the period from June 1999 to July 2000. Full data processing of entries by the RCSB, including author revisions, averages less than 2 weeks. The complexity of structures has also increased substantially. Several ribosomal units have been released and the structure of the large subunit of the ribosome, which includes 2833 RNA nucleotides and 27 proteins, was released in August 2000 (3). As of September 26, 2000, there were 13 270 structures in the PDB. The demographics of the current holdings are shown at http://www.rcsb.org/pdb/holdings.html.

The access and distribution of the archival data is through the primary Web site at UCSD and through mirrors located at Rutgers University, NIST and other locations throughout the world. The PDB receives an average of 90 000 hits per day on the primary Web site alone. The PDB Web sites provide users with direct query and reporting capabilities using the underlying databases. The query capabilities are quite extensive, and have been substantially improved with the introduction of the Molecular Information Agent (MIA; http://mia.sdsc.edu/), which provides frequently updated links to a growing number of databases. Query across the complete PDB has nevertheless been limited by missing, erroneous and inconsistently reported experimental data, nomenclature and functional annotation. Inconsistency, in particular, reflects the evolution of experimental methods, functional knowledge of proteins, and methods used to process these data over the years. The result is that only searches by PDB ID can provide completely reliable results. This paper describes the data uniformity project that is underway to address the non-uniformity in PDB data and the benefits this will bring.

DATA UNIFORMITY

In discussing the uniformity of the PDB archive, it is useful to divide the content of the archival entries into records containing coordinate information and records describing chemical features, experimental details and derivative structural features. The coordinate data are by far the most widely used data in the archive. For the most part, errors in these records are confined to the labeling of the residues and atoms and to their correspondences with records specifying chemical information, such as SEQRES, SHEET, HELIX, FORMUL, etc., (see http://www.rcsb.org/pdb/docs/format/pdbguide2.2/guide2.2_frame.html for a description of these record types). At this stage, the uniformity project does not address the intrinsic validity of the model.

Conversely, the content, level of detail and representation of other PDB data records have undergone a variety of changes during the life of the archive. In the early history of the archive, much of the chemical and experimental description was provided as free text. In later years the representation of these data has become more structured. For instance, between 1992 and 1996 the format for the description of the source of the macromolecule changed from a single text record, to a structured form with over 30 record types. During this period the level of detail in specifying refinement and data collection information was also dramatically increased. Also, many new data records and remarks were added to describe things such as related entries in sequence databases and the coordinate transformations required to produce a biologically relevant molecule. These

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

Research Collaboratory for Structural Bioinformatics (RCSB) is a consortium consisting of three institutions: Rutgers University; San Diego Supercomputer Center, University of California, San Diego; and the National Institute of Standards and Technology. The current RCSB PDB staff include the authors indicated and Kyle Burkhardt, Victoria Colflesh, Lisa Iype, Rachel Kramer, Jessica Marvin, Christine Zardecki (Rutgers); Haiyan Cheng, Phoebe Fagan, Diane Hancock, Michael Tung (NIST); Peter Arzberger, Tammy Battistuz, Wolfgang Bluhm, Nita Deshpande, Ward Fleri, Douglas Greer, Michael Gribskov, Brad Kroeger, David Padilla, Lynn Ten Eyck (UCSD).

A

```
REMARK   3
REMARK   3 REFINEMENT. MOLECULAR DYNAMICS REFINEMENT BY THE METHOD OF
REMARK   3  A. BRUNGER, J. KURIYAN, AND M. KARPLUS (PROGRAM *XPLOR*).
REMARK   3  THE R VALUE IS 0.186 FOR ALL 42851 REFLECTIONS IN THE
REMARK   3  RESOLUTION RANGE 10 TO 1.8 ANGSTROMS.
```

B

```
REMARK   3
REMARK   3  DATA USED IN REFINEMENT.
REMARK   3   RESOLUTION RANGE HIGH (ANGSTROMS) : 1.8
REMARK   3   RESOLUTION RANGE LOW  (ANGSTROMS) : 10.0
REMARK   3   DATA CUTOFF             (SIGMA(F)) : 0.0
REMARK   3   DATA CUTOFF HIGH          (ABS(F)) : NULL
REMARK   3   DATA CUTOFF LOW           (ABS(F)) : NULL
REMARK   3   COMPLETENESS (WORKING+TEST)    (%) : NULL
REMARK   3   NUMBER OF REFLECTIONS              : 42851
```

Figure 1. Example data formatted as a PDB REMARK using the format documented in 1992 (**A**) and the format documented in 1996 (**B**).

changes reflect both the improvements and standardization in the methods for structure determination, and the greater demand for detailed structure data in biological research. Content changes that track changes in technology, scientific understanding and the needs of the users will always be required in order for the archive to remain most useful.

The first step in unifying the archive is establishing a robust data specification that not only describes all of the items of data currently in the archive, but can also adapt to changes and extensions that will be required in the future. To meet these needs we have chosen the macromolecular Crystallographic Information File (mmCIF) (4) as our data specification. Given this specification we have approached the data uniformity project in two ways. The first is a file-by-file approach that involves the evaluation of the complete entry for every structure in the archive. The second is a record-by-record approach that concentrates on specific PDB records across the complete archive.

Data specification

Although the PDB format has served the community for more than two decades, many of the uniformity problems that exist within the current archive result from the lack of extensibility and specificity in the existing PDB format. This format consists of fixed format records informally described in a document, the latest being the PDB Guide for Authors Version 2.1 (http://www.rcsb.org/pdb/docs/format/pdbguide2.2/guide2.2_frame.html; a FAQ is available at http://pdb.rutgers.edu/format-faq-v1.html). Each item of data is specified in a range of character positions in one of many PDB record types (HEADER, SOURCE, REMARK, etc.). The PDB format has evolved such that entries currently exist in several different so-called PDB formats. Figure 1 demonstrates how early PDB entries can differ from more recent counterparts.

Another problem is that the number of character fields reserved for the atom serial number and the chain ID is limited. This has not affected legacy data (1971–1998) but is having an impact on large structure files recently deposited with the PDB.

Clearly a consistent data specification cannot be automatically derived for this diverse representation. Our approach to solving this problem for all new structures deposited with the PDB and as part of the uniformity project is to use mmCIF, a comprehensive and consistent means of specifying data that is recognized by the International Union of Crystallography (IUCr).

```
_refine.ls_number_reflns_obs   42851
_refine.ls_d_res_low            10.0
_refine.ls_d_res_high            1.8
_refine.ls_percent_reflns_obs  100.0
_refine.ls_R_Factor_obs        0.186
_refine.ls_R_Factor_all        0.186
```

Figure 2. Example data formatted as mmCIF keyword value pairs.

mmCIF data are organized as a collection of name-value pairs in which each name is precisely described in a machine readable data dictionary. Extensions to the dictionary (i.e., additional content) do not change the syntax of the file format. These changes only add new names. This permits new science and technology to be added to data files with less disruption to existing users and software applications. In creating this dictionary, great care was exercised to ensure that there would be correspondences to the current PDB format; this guarantees that mmCIFs can be transformed to PDB format. Figure 2 shows an example of a fragment of an mmCIF data file describing the same refinement data as Figure 1, which shows prior and current versions of the PDB format.

File-by-file uniformity processing

For file-by-file processing, the data items within each PDB record (i.e. HEADER, SOURCE, COMPND, REMARK, etc.) are examined collectively for a group of structures, typically from the same protein family. Changes in a group of structures to achieve a uniform representation in these data records may be required for three reasons. First, depositors and other users have supplied corrections, some of which were never applied to the archival files. Secondly, as shown in Figure 1A, files released prior to 1992 used a simpler record syntax, and in these files important experimental and structural details are described as unstructured text in remark records. It is often possible to extract and explicitly specify these key details from the free text of the remark records. Thirdly, information may be absent in a subset of the legacy files, yet can be extracted from other data resources. For example, useful information from other Internet-accessible data resources can be found by using sequence information. Other missing information may only be available by returning to the primary literature reference, a time-consuming process.

The uniformity in an archival file requires that all of the related data records within a file are consistent and properly integrated across the group of files. For instance, within an entry the list of residues defining the chemical sequence (SEQRES) must be consistent with the residue sequence in the coordinate records and with references to residues in describing the secondary structure features. Similar consistency issues among data records arise for references to polymer chains and individual atoms. Resolving the consistency of all atom and residue labels throughout each entry is the greatest challenge in file-by-file processing.

Atom nomenclature is checked for compliance with existing PDB conventions for standard amino acid and nucleotide residues as well as for hydrogen atoms. Missing atoms are noted. For ligands, atom nomenclature is standardized to the chemical descriptions in the PDB ligand dictionary (see below). The final step in file-by-file processing is geometrical validation, which includes checks of covalent geometry, torsion angles and intermolecular contacts. The result is a consistent, validated and complete mmCIF file that can be converted to a PDB file as needed. Files processed in this way to date include all crystal determinations of nucleic acid-containing entries in the Nucleic Acid Database (NDB) (5), globins and proteases. The nucleic acid containing files were made available in December of 2000.

It is important to emphasize that despite the care that is taken in processing and checking each entry, some errors will certainly remain. From our extensive experience with the nucleic acid containing entries we have found that uniformity is not achieved in a single step. Rather, uniformity is improved incrementally where the experience and improvements at each stage make further improvements possible.

Ligand chemical descriptions

The presence of small molecules, for example metals, ligands, potential drugs, and counter ions either covalently or non-covalently bound to the protein, DNA and RNA, plays a critical role in obtaining a full biological understanding of the structure. These so-called heterogens (HET groups) must also be described in a uniform manner, and this has been undertaken. Historically, the PDB had only recorded heavy atom connectivities for small molecules. This information was provided as a set of PDB CONECT records in a HET group dictionary. The entries in this dictionary are organized using a three-letter code identifier assigned to each unique chemical component. As part of the uniformity project, the information in this dictionary has been supplemented with hydrogen atom connectivity and bonding type. Systematic chemical names have been checked and common names have been added as synonyms. In the past, it was difficult to search this dictionary for molecules with common structures. This led to some redundancy in the assignment of different three-letter code identifiers to the same molecule. These redundant groups have been noted so that they can be standardized.

The current small molecule dictionary including full chemical descriptions is encoded as an mmCIF reference file (ftp://ftp.rcsb.org/pub/pdb/data/monomers/components.cif.Z). This dictionary forms the basis of classification for new structures and is incremented as new ligands are found in new structure depositions.

Record-by-record uniformity processing

File-by-file uniformity processing is very labor intensive and although progress has been made it will require several years to complete this work for the full archive. In the meantime there is a need to improve the uniformity of certain key data items to facilitate reliable queries on these items. This requires addressing the uniformity of these data for all entries in the PDB rather than for specific families of structures. We refer to this as record-by-record uniformity processing. This process is automated as much as possible. Revised data are available in PDB reports and Structure Explorer pages, but are not yet available within the PDB files. These revised data serve as a source of information for the file-by-file uniformity processing and can also be integrated into a legacy PDB file automatically. The values are read from the PDB file, examined, parsed and extracted automatically. Uniformity is imposed where possible and relational database tables are created that are then examined for missing data as well as outliers. Problems are resolved by visual examination of the original PDB file and returning to the publication if needed. The approach is described in more detail for each record type examined thus far:

R-factor. Values for R-factors were abstracted automatically wherever possible from the REMARK 3 records. If several R-factors were reported, the lowest was selected. Approximately 500 data files did not report values, while several had values that were outliers. Values for these were obtained from the literature. R-factors reported for NMR structures were deleted.

Resolution. Values for resolution were obtained from REMARK 2 records. Some X-ray structures had missing values, and these were obtained from the highest resolution shell in refinement statistics (present in REMARK 3). If both were missing, then values were obtained from the literature. NMR structures that had values for resolution had these values deleted.

Primary citation. Approximately one-third of the PDB entry primary citations were absent. This was because coordinates are generally deposited prior to publication and no mechanism existed for subsequent updates. Using existing author data within the entries, it was possible to automatically retrieve full citations as well as the abstracts from MEDLINE (http://www.nlm.nih.gov/databases/). Approximately 50% of missing citations were found in this way and manually checked. The remaining citations were sought manually using additional resources such as the Science Citation Index (http://www.isinet.com). Over 90% of citations have been found in this way. The remaining 10% appear to belong to structures that might not have been published.

Enzyme names and classification. EC numbers were extracted either from PDB COMPND records or by a text search of the PDB file. A relational table associating EC number with compound name was constructed for all entries with EC number assignments. This table was then used to search by compound name and when found, to assign EC numbers where they were missing. A manual verification of all the data was used for the final EC assignments. After all EC numbers had been assigned, all enzyme names in the legacy data were checked against the nomenclature of the Enzyme Commission. This final check provided a list of systematic and associated common names, both of which can be searched upon (see below).

Source, source synonyms and common names. Source information (genus and species as found in the PDB SOURCE record) was extracted for each entry and compared to the source taxonomy provided by the National Center for Biotechnology Information (NCBI) (6,7). This provided systematic and common names for the source for each structure (or component of a structure). Thus queries can now be conducted by specifying common or systematic names for source.

FUTURE WORK

Work is well underway on developing a table of synonyms for the compound names (from the PDB COMPND record) found in the PDB archive. The nomenclature of compound names has not been consistently maintained over the years, making querying very difficult. For example, as of July 2000 there were 101 HIV-1 protease structures in the PDB. There were at least 17 different ways in which the protein name has been specified in these entries: HIV-1 protease, HIV 1 protease, HIV-I protease, HIV I protease, human immunodeficiency virus type 1 protease, etc. A synonym table permits a query with any one of these names to give a consistent result. Further work is also in progress to provide systematic classifications of PDB contents similar to that developed for Macromolecular Structures (8).

The PDB ligand dictionary (with associated entries in PDB HET records) is being supplemented with additional synonyms providing better retrieval capability based on ligands. A synonym table is also being created that includes the several ways the same ligands may have been defined in different PDB data files, and these related to the corresponding PDB files.

Some larger ligands have been represented in PDB entries as multiple non-polymer groups, or as short polymer sequences. Although the subcomponents of these ligands may be well described, the identities of the compound ligand may be missing. Examples of such ligands are polymeric or non-polymeric inhibitors, prosthetic groups and co-factors.

INTEGRATION/DELIVERY OF UNIFIED DATA

The manner in which files resulting from uniformity processing are released by the PDB must address many community concerns. The interest in providing the most consistent data possible must be balanced against the need to provide continuity and stability in the archive. The latter is required in order to maintain connections with published literature and with other databases that use and analyze PDB data.

To this end, data are being entered in the PDB and made accessible via the Web based on release status. Specific files resulting from uniformity processing will be released in mmCIF format. By using mmCIF, it is possible to include both standardized and prior nomenclature in a single archival file. The contents of these files, unlike their PDB counterparts where author approval was required to make significant changes, are more dynamic. However, the content of mmCIF can be easily extended in a manner that is not disruptive to existing software applications and appropriate versioning exists.

mmCIF data files resulting from uniformity processing are placed in a special ftp area (ftp://beta.rcsb.org/pub/pdb/uniformity/data/mmCIF) for community evaluation. Recognizing that mmCIF is new to many PDB users, the mmCIF uniformity files are accompanied by a software tool that translates 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. Further details about these entries and the associated mmCIF to PDB translation software can be obtained from the PDB Web site (http://www.rcsb.org/pdb/uniformity/).

IMPACT

On query

The impact of the integration of curated data into the PDB databases is readily demonstrated by comparison of the query results using data available before and after the incorporation of uniform data. The PDB of August 29, 2000 was used to compile results given in Table 1.

Table 1. Query results on uniform versus non-uniform data (from August 29, 2000)

Attribute	Query term	Non-uniform	Uniform
Resolution	2.1–2.5 Å	3061	3492
Primary citation/journal name	J. Mol. Biol.	1953	2331
	Biochemistry	1919	2522
	To be published	2856	760
EC number	3.2.1.17	264	570
Source (organism)	*E.coli*	5	1278
	Escherichia coli	1103	1278
	Mouse	451	477
	Mus musculus	444	477
	Human	1988	2388
	Homo sapiens	2010	2388

The attributes listed are what can be searched by using SearchFields on the PDB beta query site (http://beta.rcsb.org/pdb/cgi/queryForm.cgi). The features on this beta site will be available on the production PDB server by the end of the year. The numbers given are the result of entering in the query term in the field provided on both non-uniform and uniform data. The data are available as database tables, and are not available in the individual PDB data files. Queries based on source must be performed using the 'exact word match' option provided on the query form.

The results show improved return on query results across key categories of data. Categories that can improve the ability to locate structures according to experimental criteria and to biological relevance is a key step forward. The effects of synonym tables are quite dramatic, returning complete lists of structures for the first time. An important impact of uniformity is that it accurately defines the scope of data in the PDB to which users can be made aware. As an example, an option now exists to browse EC numbers with their associated names and select categories from the list based on the number of structures that exist at each level of the hierarchy. This is updated dynamically as new structures are added.

On deposition

The manner in which data enter the archive and are processed and annotated also benefits from uniformity processing. Standardization of nomenclature and the development of the associated controlled vocabularies simplify the deposition process and define rules about the data that can be implemented in software. The automation afforded by software speeds data processing and annotation and results in the production of entries, which share greater consistency with the archive as a whole. A further benefit of the development of controlled vocabularies is that these can be presented to the depositor as a menu of choices. This simplifies the deposition process and subsequent annotation.

CONCLUSIONS

Uniformity provides a number of features, with perhaps the most important being a clear specification of the systematic name and source of all components of a macromolecular structure. Previously it has only been possible to reference structures by their unique PDB identifier. Now it is possible to reliably reference all structures that belong to a specific functional class of biological molecule as well as systematically reference the ligands bound to these molecules. In the long term this will make the PDB much more useful to a wide spectrum of users. At present the scientific literature provides the main browser for locating a structure. A specific paper mentions an associated PDB ID, which can be located in the database. In the future other applications and data resources will be able to reference PDB structures by their functional specification, which is the natural way a structural biologist thinks about a problem. The functional specification may involve protein–protein interactions, or specific ligands binding, or the role of a particular protein in a complex biochemical pathway. The jump to permit the association of structure to biological function, which is the hope of structural genomics, is non-trivial. Systematic functional assignments for those structures of known function are vital in determining the identity of new structures of unknown function that will derive from structural genomics. Correct and systematic naming of the components of a macromolecular structure and functional classification of those components are the first steps towards a more detailed annotation of structure that the future of biology demands.

ACKNOWLEDGEMENTS

The PDB would like to acknowledge the work of Dietmar Schomburg (Institute of Biochemistry at the University of Cologne), Project Leader of the BRENDA database for enzyme classification remediation. We also gratefully acknowledge the comments and corrections that have been contributed by colleagues at the National Center for Biotechnology Information, the Cambridge Crystallographic Data Centre, the European Bioinformatics Institute and the many users of the PDB. This work is supported by grants from the National Science Foundation, the Office of Biology 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. 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.
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.
3. Nissen,P., Hansen,J., Ban,N., Moore,P.B. and Steitz,T.A. (2000) The structural basis of ribosome activity in peptide bond synthesis. *Science*, **289**, 920–930.
4. Bourne,P., 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.
5. Berman,H.M., Olson,W.K., Beveridge,D.L., Westbrook,J., Gelbin,A., Demeny,T., Hsieh,S.H., Srinivasan,A.R. and Schneider,B. (1992) The Nucleic Acid Database – a comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J.*, **63**, 751–759.
6. 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.
7. Wheeler,D.L., Chappey,C., Lash,A.E., Leipe,D.D., Madden,T.L., Schuler,G.D., Tatusova,T.A. and Rapp,B.A. (2000) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **28**, 10–14. Updated article in this issue: *Nucleic Acids Res.* (2001), **29**, 11–16.
8. Hendrickson,W.A. and Wüthrich,K. (eds) (1999) *Macromolecular Structures*. Elsevier Science Ltd., London.