occur, the introduction of small quantities of ligands, products, substrate, substrate analogs, monovalent or divalent cations, organic reagents, etc., to the crystallization mixtures may facilitate crystal growth. Also, analogous to the vapor diffusion experiments, the search may be expanded to finer increments of pH if results warrant.

The example above illustrates how the data in the BMCD can be used to develop a general procedure for soluble proteins. The BMCD can be used to develop analogous procedures for other classes of biological macromolecules.

[29] Protein Data Bank Archives of Three-Dimensional Macromolecular Structures

By ENRIQUE E. ABOLA, JOEL L. SUSSMAN, JAIME PRILUSKY, and NANCY O. MANNING

Introduction

The growing reliance on the availability of structural data on macromolecules to help one understand biological processes highlights the importance of information resources such as the Protein Data Bank (PDB). Established at Brookhaven National Laboratory (BNL, Upton, NY) in 1971, the PDB has a 25-year history of service to a global community of researchers, educators, and students in a variety of scientific disciplines.[1-3] The common interest shared by this community is a need to access information that can relate the biological functions of macromolecules to their three-dimensional structures.

The PDB was started at the urging of members of the scientific community who in the late 1960s anticipated the growth in structural biology. The scientific and commercial importance of the resource was recognized even then by this community, who urged the adoption of a policy to encourage deposition, archiving, and distribution of data. Furthermore, it was deemed important to seek ways of having these data available free of charge to the community.

[1] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, *J. Mol. Biol.* **112,** 535 (1977).

[2] E. E. Abola, F. C. Bernstein, S. H. Bryant, T. F. Koetzle, and J. Weng, Protein Data Bank. *In* "Crystallographic Databases—Information Content, Software Systems, Scientific Applications" (F. H. Allen, G. Bergerhoff, and R. Sievers, eds.), pp. 107–132. Data Commission of the International Union of Crystallography, Bonn, 1987.

[3] *Nature New Biol.* **233,** 223 (1971). [Crystallography, Protein Data Bank: Announcement].

0076-6879/97 $25.00

The PDB was established as an international collaboration between the groups headed by W. Hamilton at BNL and O. Kennard at the University of Cambridge (Cambridge, UK). This collaboration was expanded later to include centers at the Commonwealth Scientific and Industrial Research Organization (CSIRO, Clayton, Australia) and at the University of Osaka (Osaka, Japan). Today, this international collaboration is being strengthened by the establishment of data collection and distribution centers at the European Bioinformatics Institute (EBI; Hinxton Hall, UK), at the University of Osaka, and at the Weizmann Institute of Science (Rehovot, Israel). In addition to these centers, a number of official "PDB mirror" sites are operating Web and FTP sites, facilitating access to data within regional areas (see PDB Web home page for a list of current mirror sites). Funds to operate the PDB initially were provided by the U.S. National Science Foundation (NSF, Arlington, VA). More recently, the U.S. Department of Energy (Washington, DC), the U.S. National Institutes of Health (Bethesda, MD), and the U.S. National Library of Medicine (Bethesda, MD) along with the NSF provide funds to operate the resource.

Today, the challenge facing the PDB is to keep abreast of the increasing flow of data, to maintain the archive as error free as possible, and to organize and present the stored information in ways that facilitate data retrieval, knowledge exploration, and hypothesis testing, without interrupting current services. We discuss in this chapter various facets of our activities with the hope that this will help users understand the nature and scope of the data in the PDB archives and help them access these data. We also provide practical guidelines to depositors that will help them in depositing data with the PDB.

Contents of Protein Data Bank

Several pieces of information related to an entry are archived by the PDB (see Table I). In addition to the coordinate entry file, the PDB stores files related to the experiment such as structure factors, nuclear Overhaüser effect (NOE) restraints, and lists of chemical shifts. Also archived are

TABLE I
CONTENTS OF PROTEIN DATA BANK ARCHIVES

Coordinate entries (released and obsolete) with correction history
Raw coordinate data files
Structure factors
NMR–NOE and chemical shifts data files
Topology and parameter file used in refinement

auxiliary files used in structure analysis and refinement such as X-PLOR parameter and topology files. Currently, the archives are managed as a set of individual files, and each entry may have several associated files. The PDB is in the process of building a relational database, 3DBase, that will replace the current data management and access system. A description of 3DBase, including an outline of how users can access its contents is provided below.

A summary of the contents of the November 1996 PDB release is given in Table II. More than 5000 coordinate entries are available, and another 1000 currently either are being processed or are not to be released for up to 1 year at the request of the depositors. In 1996, the PDB receives 5 new entries per day and the rate of deposition is growing exponentially as shown in Fig. 1. It is expected that by the year 2000, the PDB will contain between 20,000 and 30,000 entries.

Coordinate entries in the PDB are stored in separate files, each of which reports the results of an experiment or analysis that elucidates the structure of proteins, necleic acids, polysaccharides, and other biological macromolecules. Although most of the data are generated from single crystal X-ray diffraction studies, a growing number of PDB entries are from nuclear magnetic resonance (NMR) studies (see Table II).

Files are distributed using the PDB data interchange format that was introduced in 1976 and has been used since then without significant changes. Entries are distributed as flat files consisting of fixed length records. Each of these records is identified by a tag word such as HEADER, COMPND and ATOM. The PDB records are divided into fields, most of which are

TABLE II
PROTEIN DATA BANK HOLDINGS LIST:
NOVEMBER 15, 1996[a]

Molecule type
4500 proteins, peptides, and viruses
202 protein–nucleic acid complexes
361 nucleic acids
12 carbohydrates
Experimental technique
142 theoretical modeling
714 NMR
4219 diffraction and other
576 Structure Factor files
169 NMR Restraint files

[a] A total of 5075 released atomic coordinate entries, by molecule type, were available as of November 1996.

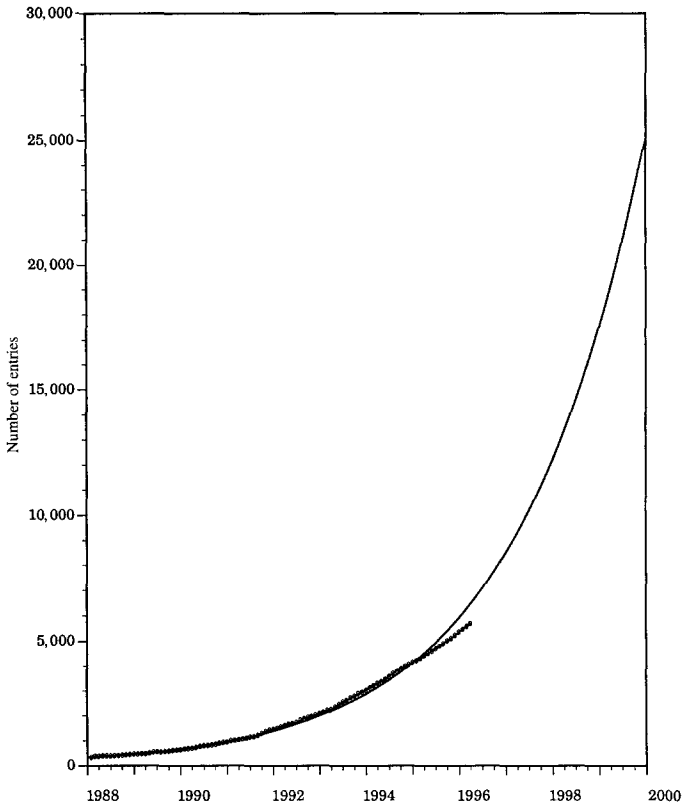

FIG. 1. Total number of atomic coordinate entries in the PDB extrapolated to year 2000 (exponential fit to 1988–1996 data).

fixed in length. More recently a number of records with varying field lengths and terminated by a character token (e.g., semicolon ";") have been introduced. The contents of the PDB file along with a full description of the format and general rules used by the PDB in representing biological and chemical data are given in the *Protein Data Bank Contents Guide.*[4] Table III lists all PDB record types in use in 1996. The REMARK section of PDB entries contains not only free text comments but also a number of standard tables such as those containing matrices needed to generate complete viral particles from the icosahedral asymmetric unit supplied by the depositor.

[4] "Protein Data Bank Contents Guide: Atomic Coordinate Entry Format Description," version 2.1, October 25, 1996. Available anonymous FTP: ftp.pdb.bnl.gov, and at URL http://www.pdb.bnl.gov.

TABLE III
PROTEIN DATA BANK RECORD TYPES[a]

Section	Description	Record type
Title	Summary descriptive remarks	HEADER, OBSLTE, TITLE, CAVEAT, COMPND, SOURCE, KEYWDS, EXPDTA, AUTHOR, REVDAT, SPRSDE, JRNL
Remark	Bibliography, refinement, annotations	REMARKs 1, 2, 3 and others
Primary structure	Peptide and/or nucleotide sequence and relationship between PDB sequence and that found in sequence database(s)	DBREF, SEQADV, SEQRES, MODRES
Heterogen	Description of nonstandard groups	HET, HETNAM, HETSYN, FORMUL
Secondary structure	Description of secondary structure	HELIX, SHEET, TURN
Connectivity annotation	Chemical connectivity	SSBOND, LINK, HYDBND, SLTBRG, CISPEP
Miscellaneous features	Features within macromolecule	SITE
Crystallographic	Description of crystallographic cell	CRYST1
Coordinate transformation	Coordinate transformation operators	ORIGXn, SCALEn, MTRIXn, TVECT
Coordinate	Atomic coordinate data	MODEL, ATOM, SIGATM, ANISOU, SIGUIJ, TER, HETATM, ENDMDL
Connectivity	Chemical connectivity	CONECT
Bookkeeping	Summary information, end-of-file marker	MASTER, END

[a] Various sections of a PDB coordinate entry and the records comprising them.

3DBase: Relational Database Management System for Protein Data Bank

In 1994, the PDB started work on building a new relational database for managing and accessing the contents of the data bank. The new database, 3DBase, is constructed with the SYBASE[5] Relational Database Management System (RDBMS), the Object-Protocol Model (OPM), and the OPM data management tools[6] developed by the Markowitz group at Lawrence Berkeley National Laboratory (Berkeley, CA). SYBASE provides a powerful and robust environment for data management, the OPM tools allow rapid development of SYBASE databases, and the object-oriented view

[5] "SYBASE SQL Server," Unix version 10.0. Sybase, Inc., Emeryville, California, 1994.
[6] I. A. Chen and V. M. Markowitz, *Inf. Syst.* **20**(5), 393 (1995); article and related information available at URL:http://gizmo.lbl.gov/DM_TOOLS/OPM/opm.html.

of the OPM provides a scientifically intuitive representation of data. For example, a purely relational view of the data requires the construction of several tables to store all the information related to the coordinates of a residue. Queries that require access to these data will then have to be constructed by naming and joining data found in each of these tables. In contrast, only one object definition is required to store and access the same data in an object-oriented view.

This development effort attempts to address the needs of the diverse user community served by the PDB. A conceptual view (also referred to as a *conceptual schema*) of the data was developed that supports queries by those interested in answering both crystallographic and molecular biology questions. The system is designed to federate 3DBase easily with other biological databases. It is expected that federation will permit complex queries to be submitted to the database, returning a composite answer built from a set of diverse databases. Interoperability is addressed through the use of a schema that shares with other OPM-based databases and supports a variety of data interchange formats in the query results. In addition to providing users with a powerful environment to do complex *ad hoc* queries, 3DBase also will facilitate management of the growing archive, which is expected to contain more than 20,000 structural reports by the year 2000.

This work is being done as a collaboration among the following groups: The Protein Data Bank (Brookhaven National Laboratory), Bioinformatics Unit (Weizmann Institute of Science), OPM Data Management Tools Project (Lawrence Berkeley National Laboratory), and Genome Database (GDB, Johns Hopkins University, Baltimore, MD).

Schema Development

The OPM is a semantic data model that includes constructs that are powerful enough to represent the diversity and complexity of data found in PDB entries. The OPM has constructs such as object class, object attribute, class hierarchy and inheritance, and derived attribute. In addition, multiple views of the data are constructed in the OPM by using derived classes. This mechanism allows others to develop their own conceptual view of the data without having to alter the underlying database. The schema for 3DBase has been developed using the OPM, and is available for perusal through the PDB WWW (World Wide Web) home page. Among its notable features is a description of the coordinate data set from two perspectives. The object class *Experiment* provides users with the classic view of a PDB entry, which is a report of crystallographic or NMR analysis. An alternative view is presented in the class *Aggregate,* which describes the macromolecule and its complexes with ligands, or other macromolecules. A

clear example that demonstrates the differences between these classes is the case of the hemoglobin molecule. The *Experiment* object contains the coordinates for the crystallographic asymmetric unit, which is usually a dimer of α and β subunits. The full tetramer generated by using a crystallographic twofold symmetry operator will, however, be presented in the *Aggregate* object. The latter case is normally what molecular biologists are interested in when accessing PDB entries. Those wanting to do crystal-packing studies or further crystallographic refinement will need access to the *Experiment* object to obtain the asymmetric unit.

In 3DBase, literature citation data are being loaded into the citation database (CitDB) of references that was developed by the GDB.[7] A pointer to the appropriate entry in the CitDB is loaded in the *Experiment* object of 3DBase. This is an example of the strategy that the PDB is following in linking to external databases. The CitDB will be managed as a federation run by a number of database centers that include the GDB and PDB. There are several advantages to this scenario. By sharing the schema and management of the citation databases, access to information stored in each of the databases via the bibliographic citation becomes straightforward. Duplication of effort is also minimized. Today, it is still common to have several public databases build and maintain their own bibliographic databases. This will no longer be economically feasible with the expected rapid growth in database size.

A notable feature of the 3DBase schema is the inclusion of an object class that allows users, depositors, and other editors to add their own annotation to objects. This can, for example, be used to attach to an *Experiment* object the output of a program that describes error estimates using a novel data-checking technique. The PDB and its advisers, along with the rest of the community, will have to arrive at a workable editorial policy before general use of this annotation mechanism is permitted.

Building Semantic Links to External Data Sources

Links to contents of sequence databases are provided in 3DBase via the *PrimarySeq* and *SeqAdv* classes. These classes form another set of objects that link 3DBase objects to external databases. Representing, building, and maintaining these links will be one of the primary tasks of the PDB in the coming years. There are several issues that must be addressed for this effort to succeed. Data representation issues are foremost. Each database uses different data models to represent and store information. Semantic contents are rarely the same; for example, the primary sequence data stored in

[7] K. H. Fasman, A. J. Cuticchia, and D. T. Kingsbury, *Nucleic Acids Res.* **22,** 3462 (1994).

sequence databases such as SWISS-PROT[8] (Medical Biochemistry Department, University of Geneva, Switzerland and European Bioinformatics Institute, Hinxton, England) or Protein Information Resource[9] (PIR, National Biomedical Research Foundation, Washington, DC) are presented using a view that differs significantly from that used by the PDB.

In general, PIR and SWISS-PROT entries contain information on the naturally occurring wild-type molecules. Each entry normally contains the sequence of one gene product, and some entries include the complete precursor sequence. Annotation is provided to describe residue modifications. In both databases, the residue names used are limited to the 20 standard amino acids.

In contrast, PDB entries contain multichain molecules with sequences that may be wild type, variant, or synthetic. Sequences also may have been modified through protein-engineering experiments. A number of PDB entries report structures of domains cleaved from larger molecules.

The *PrimarySeq* object class was designed to account for these differences by providing explicit correlations between contiguous segments of sequences as given in PDB ATOM records and PIR or SWISS-PROT entries. Several cases are easily represented using this class. Molecules containing heteropolymers may be linked to different sequence database entries. In some cases, such as those PDB entries containing immunoglobulin Fab fragments, each PDB chain may be linked to several different SWISS-PROT entries.

This facility is needed because these databases represent sequences for the various immunoglobulin domains as separate entries. *PrimarySeq* also should be able to represent molecules engineered by altering the gene (fusing genes, altering sequences, creating chimeras, or circularly permuting sequences). In addition, it will be possible to link segments of the structure to entries in motif databases (e.g., PROSITE,[10] BLOCKS[11]).

Initial building of these links is straightforward, and requires analysis of a few entries coming out of a sequence comparison search using the program FASTA[12] or BLAST[13] against the sequence databases. An issue

[8] A. Bairoch and B. Boeckmann, *Nucleic Acids Res.* **22,** 3578 (1994); available at URL: http://expasy.hcuge.ch/sprot/top.html.

[9] D. G. George, W. C. Barker, H. W. Mewes, F. Pfeiffer, and A. Tsugita, *Nucleic Acids Res.* **22,** 3569 (1994); available at URL: http://www-nbrf.georgetown.edu/pir.

[10] A. Bairoch and P. Bucher, *Nucleic Acids Res.* **22,** 3583 (1994); available at URL: http://expasy.hcuge.ch/.

[11] S. Henikoff and J. G. Henikoff, *Genomics* **19,** 97 (1994); available at URL: http://blocks.fhcrc.org/.

[12] W. R. Pearson and D. J. Lipman, *Proc. Natl. Acad. Sci. U.S.A.* **85,** 2444 (1988).

[13] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, *J. Mol. Biol.* **215,** 403 (1990).

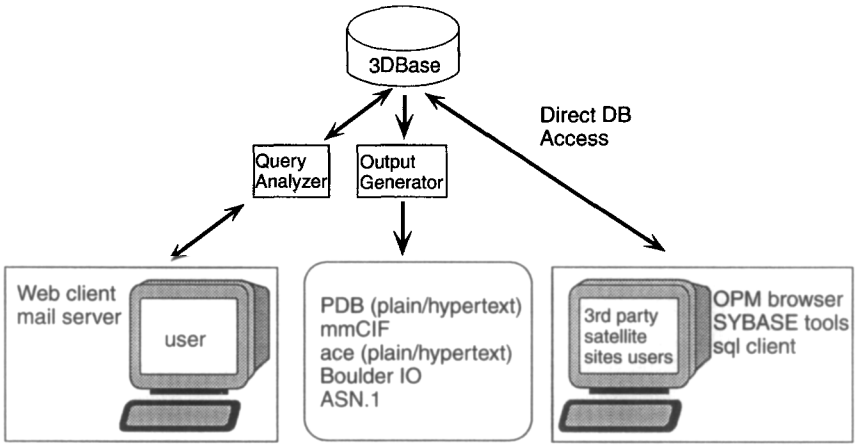

FIG. 2. Access to 3DBase.

that must be resolved in the long run will be the updating of these links as new experimental evidence is encountered, leading to a correction in either database. Both PIR and SWISS-PROT have similar problems as they build pointers to PDB entries. To help obviate these difficulties the PDB has agreed to establish a closer interaction among the databases. A protocol is being set up that will broadcast to each database changes that occur that could, in turn, affect specific entries.

Accessing Data in 3DBase

User queries to 3DBase will be via the Internet, using general-purpose graphical user interfaces such as Mosaic and Netscape. Access also will be possible through the use of software developed by third parties (commercial developers). As diagrammed in Fig. 2, user queries will be addressed to the Query Analyzer (PDB-QA), a program module running at the server site that will parse queries and pass them on to 3DBase. Query results will be returned through the Output Generator (PDB-OG) in the format requested by the user. Queries placed over the network generally will be in the form of uniform resource locators (URLs), which are easily generated from HyperText links, HyperText markup language (HTML)-based forms, or by use of programs or scripts employing the National Center for Supercomputing Applications libraries[14] for more sophisticated applications. As part of the query, the user may specify the format of the response, as we

[14] National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Illinois. Available at URL: http://www.ncsa.uiuc.edu.

do at the present time in the PDB 3DB Browser. The response will be frequently in the form of an HTML document, but it can also be a PDB- or Crystallographic Information File (CIF)-formatted file.[15] The information returned may be either a complete or partial entry, and may include information from linked databases or external programs.

A 3DBase browser has been built using the Letovky Genera system.[16] Users specify search criteria by filling out an HTML form. Software at BNL processes this form and generates the required structured query language (SQL). System performance is improved by using stored SYBASE SQL procedures that access each predefined object. The fields available are similar to those in our PDBBrowse program, and answer most of the questions that users have been asking.

For those familiar with (or willing to learn about) the OPM protocol, access to the object layer will be provided using a high-level OPM-based query language. As part of the PDB open database policy, direct access to the underlying RDBMS will be allowed and actively supported. These queries are not parsed by the PDB-QA module, so better response time can be expected. This provides third-party developers with the opportunity either to incorporate SQL clients in their products or to learn more of the OPM protocol and, thereby, gain access to all of the benefits that the Object model affords (e.g., active external links, programs). As depicted in Fig. 2, the output generator will return query results using a variety of data interchange formats. The PDB will continue to support its current format for the foreseeable future. We plan also to extend this format to allow us to represent objects being stored in 3DBase. In addition, a "raw format" is being provided that returns an attribute/value pair. This form is easily parsed and is more compact than the PDB format.

Submitting Data to Protein Data Bank

The PDB is evolving to operate as a direct-deposition archive, providing mechanisms allowing depositors to load data with minimal staff intervention. This strategy is essential if the PDB is to meet present projections of exponential growth in depositions against a fixed staff size. This is particularly challenging owing to the complexity of the data being handled, the need for a common viewpoint of the entry description, and the community requirement that these data be accessible immediately on receipt.

[15] P. M. D. Fitzgerald, H. M. Berman, P. E. Bourne, and K. Watenpaugh, *American Crystallographic Association Annual Meeting [program and abstracts]*, Ser.2, **21,** 33 (1993).

[16] S. I. Letovsky, "Genera" (computer program, 1994). URL: http://gdbdoc.gdb.org/letovsky/genera.

With direct deposition, there will be a concomitant need to increase the power of data validation procedures. These procedures must reflect current models for identifying errors and must be as complete as possible. Quality control issues assume a more central and difficult role in direct deposition strategies. Distributed data must be of the highest quality; otherwise users will lose their trust in the archived data and will have to revalidate data received from the PDB before using them, clearly an unproductive scenario.

Current Data Deposition Procedures

Since its inception in 1971, the method followed by the PDB for entering and distributing information has paralleled the review-and-edit mode used by scientific journals. The author submits information that is converted into a PDB entry and is run against PDB validation programs by a PDB processor. The entry and the output of the validation suite are then evaluated by a PDB scientific staff member, who completes the annotations and returns the entry to the author for comment and approval. Table IV summarizes checks included in the current data validation suite. Corrections from the author are incorporated into the entry, which is reanalyzed and validated before being archived and released.

Originally data flow was a manual system, designed for a staff of 1–2 scientists, and a deposition rate of about 25–50 entries per year. One person processed an entry from submission through its release. By the late 1980s, when the first steps at automation were being introduced, running the

TABLE IV
DATA VALIDATION WITH CURRENT SYSTEM

Class	What is checked
Stereochemistry	Bond distances and angles, Ramachandran plot (dihedral angles), planarity of groups, chirality
Bonded/nonbonded interactions	Crystal packing, unspecified inter- and intraresidue links
Crystallographic information	Matthews coefficient, Z value, cell transformation matrices
Noncrystallographic transformation	Validity of noncrystallographic symmetry
Primary sequence data	Correlations with sequence databases
Secondary structure	Generated automatically or visually checked
Heterogen groups	Identification, geometry, and nomenclature
Miscellaneous checks	Solvent molecules outside hydration sphere, syntax checks, internal data consistency checks

validation programs took about 4 hr per entry. Today, the same step, which includes a vastly improved set of validation programs, takes about 1 min.

The current deposition load of ~100 entries a month is handled by about 10 staff members who annotate and validate entries. The process is a production line in which checking is repeated at various steps to ensure that errors and inconsistencies in data representation are minimized. Prior to June 1994, a significant number of depositions required that administrative staff manually input information provided in a deposition form. Introduction of the current Electronic Deposition Form, together with a new parsing program, has greatly reduced hand entry of information.

Today, most of the processing time is spent resolving data representation issues and ensuring that outliers are identified and annotated. The most troublesome areas are consistently those involving handling of heterogens, resolving crystal-packing issues, representing molecules with noncrystallographic symmetry, and resolving conflicts between the submitted amino acid sequence and that found in the sequence databases. Publications and other references are sometimes consulted to verify factual information such as crystal data, biological details, and reference information. Although much improved over those used in 1991, processing programs still allow errors to pass undetected through the system, requiring a visual check of all entries. The PDB continually improves these programs, and also acquires software from collaborators to address deficiencies that both we and our users have identified. In addition, the PDB now has formed a quality control group that will be identifying sources of errors and recommending steps to improve data quality.

Development of Automatic Deposition and Validation

The PDB must overcome many challenges for direct deposition to work. In a workshop held to assess the needs of PDB users, crystallographers and NMR spectroscopists were unanimous in their desire for a system that did not require additional work on their part when depositing data. On the other hand, consumers (who included these same depositors) were vocal in their desire for entries to contain more information than is currently available within the PDB. The PDB is striving to develop a suite of deposition and validation programs that accommodates these somewhat conflicting desires while ensuring that the archives maintain the highest standard of accuracy.

A considerable variety of information is archived about each structure, which must be supplied by the authors. The new PDB Web-based data deposition program, AutoDep, simplifies the process (see Table V). It includes a convenient and interactive electronic deposition protocol that

TABLE V
IMPORTANT HIGHLIGHTS OF AutoDep

Program allows author to fill in form automatically from existing PDR entry or from previous deposition. AutoDep enters data from designated file to appropriate fields in new form. Author need only update fields to reflect new structure
X-Ray structural refinement software is available to write PDB records that can be merged automatically into the deposition form. For example, new releases of X-PLOR and SHELXL write refinement details as PDB records that will be read by program and entered in relevant sections. PDB is continuing to work with authors of various programs and anticipates that increasing numbers of programs will be integrated with PDB
Each session has Help files, examples, and links to related documentation and useful URLs to support author during AutoDep session
At any time during AutoDep session, Deposition Form or resultant header portion of PDB file can be viewed to check progress
AutoDep session can be interrupted at any time and resumed later. Session ID number and password must be recorded to continue with same deposition
When author is satisfied with completed Deposition Form, Submit button is provided to initiate following: 1. Coordinates are passed through syntax checker 2. If they fail, depositor is asked to correct problem and resubmit coordinates 3. If they pass, depositor immediately is sent acknowledgment letter containing PDB ID code 4. Entry enters PDB processing flow

guides the author in providing information. It also contains tools for data verification and validation, and is able to flag errors in syntax or spelling. The form requests approximately 1000 items, including a description of the experiment and the molecule under study. Steps are being taken to help ease the burden of filling out this form. For example, the program can fill in fields using data from an existing PDB entry or from any file containing PDB format-compliant records. These data then can be modified to reflect the contents of the new deposition. Checks against other databases are an important and evolving part of this process. Thus, names of organisms are checked against the taxonomy database of the National Center for Biotechnology Information (NCBI, Bethesda, MD),[17] chemical names against IUPAC (International Union of Pure and Applied Chemistry[18])

[17] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland (producer). Available URL: http://www.ncbi.nih.gov. Available anonymous FTP: ncbi.nlm.nih.gov. Directory: /repository/taxonomies.

[18] C. Liebecq (ed.), "Biochemical Nomenclature and Related Documents: A compendium," 2nd Ed. International Union of Biochemistry and Molecular Biology, Portland Press, London, 1992.

nomenclature tables, and author names and citations against MEDLINE[19] (CitDB when it becomes available). FASTA/BLAST programs are run against the SWISS-PROT and PIR databases to verify protein sequences. Links between the PDB entry and these databases are established in the process. To handle the increasing number of entries with nonstandard residues (*heterogens*), a standard residue and heterogen dictionary has been developed for use in the data entry and checking process. The PDB is also adopting programs developed by the Cambridge Crystallographic Data Center[20] (CCDC, Cambridge, UK), and elsewhere, to handle heterogens automatically for use in AutoDep.

In addition to the deposition form that is filled out by AutoDep, authors are requested to submit the coordinate data entry and other experimental data files for processing and archiving. Facilities are provided by AutoDep that simplify this process. An FTP script is provided that uploads author-specified local filenames to the PDB server site.

The completed form is then converted automatically into a file in PDB format and, along with the coordinate data, is submitted to a set of validation programs for checking and further annotation. These programs are designed to check (1) the quality, consistency, and completeness of the experimental data; (2) possible violations of physical or stereochemical constraints (e.g., no two atoms in the same place, appropriate bond angles); (3) compliance with our data dictionary (syntax checks); and (4) in the near future, the correspondence of the experimental data to the derived structure. Development of the validation suite will continue to evolve with advice from the community and encompass programs currently in use, written both within and outside the PDB.

The validation software automatically generates, and includes in the entry, measures of data quality and consistency, as well as annotations giving details of apparent inconsistencies and outliers from normal values. This output is returned to the depositor for review. Entries whose data quality and consistency meet appropriate standards may then be sent by the depositor directly for final review by the PDB staff and entry into the database. Entries that do not pass the quality and consistency checks may be revised by the depositor to correct inadvertent errors; alternatively, more experimental work may be needed to resolve problems uncovered.

Apparent inconsistencies or outliers may remain in a submitted entry, provided these are explained by the depositor in an annotation. In the most

[19] MEDLINE (on-line and CD-ROM). National Library of Medicine, National Institutes of Health, Bethesda, Maryland (producer). Available: NLM, DIALOG, BRS, SilverPlatter.

[20] F. H. Allen, J. E. Davies, J. J. Galloy, O. Johnson, O. Kennard, C. F. Macrae, E. M. Mitchell, G. F. Mitchell, J. M. Smith, and D. G. Watson, *J. Chem. Inf. Comput. Sci.* **31,** 187 (1991).

interesting cases, unusual features are a valid and important part of the structure. However, all such entries will be reviewed for possible errors by PDB staff, who may discuss any important issues with the depositor. The PDB staff will then forward acceptable entries to the database.

To make automatic deposition as easy as possible, the PDB is working with developers of software commonly used by our depositors. By modifying these programs to produce compliant data files and performing validation and consistency checks before submission, it may be possible to bypass most of the tedious steps in deposition. We are already working with A. Brünger (Yale University, New Haven, CT) to use procedures available through X-PLOR[21] to replace part of the validation suite for structures produced by X-ray crystallography and NMR. Diagnostic output will be included automatically as annotations in the entry. A limited version of X-PLOR will be available from BNL to all depositors for validation purposes only.

Validation of coordinate data against experimental X-ray crystallographic data requires access to structure-factor data, which are requested by the PDB, the International Union of Crystallography (IUCr), and some journals, but are not always supplied by the depositor. We are working toward building consensus in the community that structure-factor data are a necessary component of deposits of structures derived by X-ray crystallography. Statistics such as number of F and R values vs sin θ/λ will be calculated and included in the PDB entry as annotation for the experiment.

To make it easier for depositors to submit structure factors (as well as to exchange these data between laboratories), the PDB, in close collaboration with a number of macromolecular crystallographers, has developed a standard interchange format for these data. This standard is in CIF (Crystallographic Information File)[15,22] and was chosen both for simplicity of design and for being clearly self-defining, i.e., the file contains sufficient information for the file to be read and understood by either a program or a person. Details of this format are available through the PDB WWW server.

A consensus is still developing in the NMR community as to what types of experimental data should be deposited and what kinds of validation and consistency checks should be performed. Structural data produced by other methods also may have special features that should be archived or checked, for example, the sequence alignment used for modeling studies. Require-

[21] A. T. Brünger, "X-PLOR, Version 3.1: A System for X-Ray Crystallography and NMR." Yale University Press, New Haven, Connecticut, 1992.

[22] S. R. Hall, F. H. Allen, and I. D. Brown, *Acta Crystallogr.* **A47,** 655 (1991); related information available at URL: http://www.iucr.ac.uk/cif/home.html.

ments for the types of data to be deposited and proper ways of checking the validity and consistency of the data will be developed in cooperation with the experimental community for each category of structure data archived by the PDB.

[30] Macromolecular Crystallographic Information File

By PHILIP E. BOURNE, HELEN M. BERMAN, BRIAN MCMAHON, KEITH D. WATENPAUGH, JOHN D. WESTBROOK, and PAULA M. D. FITZGERALD

Introduction

The Protein Data Bank (PDB) format provides a standard representation for macromolecular structure data derived from X-ray diffraction and nuclear magnetic resonance (NMR) studies. This representation has served the community well since its inception in the 1970s[1] and a large amount of software that uses this representation has been written. However, it is widely recognized that the current PDB format cannot express adequately the large amount of data (content) associated with a single macromolecular structure and the experiment from which it was derived in a way (context) that is consistent and permits direct comparison with other structure entries. Structure comparison, for such purposes as better understanding biological function, assisting in the solution of new structures, drug design, and structure prediction, becomes increasingly valuable as the number of macromolecular structures continues to grow at a near exponential rate. It could be argued that the description of the required content of a structure submission could be met by additional PDB record types. However, this format does not permit the maintenance of the automated level of consistency, accuracy, and reproducibility required for such a large body of data.

A variety of approaches for improved scientific data representation is being explored.[2] The approach described here, which has been developed under the auspices of the International Union of Crystallography (IUCr), is to extend the Crystallographic Information File (CIF) data representation used for describing small-molecule structures and associated diffraction experiments. This extension is referred to as the macromolecular Crystallo-

[1] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rogers, O. Kennard, T. Shimanouchi, and M. Tasumi, *J. Mol. Biol.* **112,** 535 (1977).

[2] IEEE Metadata: *http://www.llnl.gov/liv_comp/metadata/* (1996).

0076-6879/97 $25.00