Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

The Protein Data Bank

Helen M. Berman,[a]* Tammy Battistuz,[b] T. N. Bhat,[c] Wolfgang F. Bluhm,[b] Philip E. Bourne,[b,d,e] Kyle Burkhardt,[a] Zukang Feng,[a] Gary L. Gilliland,[c] Lisa Iype,[a] Shri Jain,[a] Phoebe Fagan,[c] Jessica Marvin,[a] David Padilla,[b] Veerasamy Ravichandran,[c] Bohdan Schneider,[a] Narmada Thanki,[c] Helge Weissig,[b] John D. Westbrook[a] and Christine Zardecki[a]

[a]RCSB, Department of Chemistry, Rutgers, The State University of New Jersey, 610 Taylor Road, Piscataway, NJ, USA, [b]RCSB, San Diego Supercomputer Center, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0537, USA, [c]RCSB, National Institute of Standards and Technology, Biotechnology Division and Informatics, 100 Bureau Drive, Gaithersburg, MD 20899-8314, USA, [d]Department of Pharmacology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0500, USA, and [e]The Burnham Institute, 10901 North Torrey Pines Road, La Jolla, CA 92037, USA

Correspondence e-mail:
berman@rcsb.rutgers.edu

Received 1 December 2001
Accepted 21 February 2002

The Protein Data Bank [PDB; Berman, Westbrook *et al.* (2000), *Nucleic Acids Res.* **28**, 235–242; http://www.pdb.org/] is the single worldwide archive of primary structural data of biological macromolecules. Many secondary sources of information are derived from PDB data. It is the starting point for studies in structural bioinformatics. This article describes the goals of the PDB, the systems in place for data deposition and access, how to obtain further information and plans for the future development of the resource. The reader should come away with an understanding of the scope of the PDB and what is provided by the resource.

1. Introduction

The Protein Data Bank was established at Brookhaven National Laboratory (BNL; Bernstein *et al.*, 1977) in 1971 as an archive for biological macromolecular crystal structures. Nobel prizes have been awarded for the determination and analysis of some of the structures in the PDB. It represents one of the earliest community-driven molecular-biology data collections. In the beginning the archive held seven structures and with each passing year a handful more were deposited. In the 1980s, the number of deposited structures began to increase dramatically. This was a consequence of the improvements in technology for all aspects of the crystallographic process, the addition of structures determined by nuclear magnetic resonance (NMR) methods and changes in community views about data sharing. By the early 1990s, the majority of journals required a PDBid for publication and at least one funding agency, the National Institute of General Medical Sciences (NIGMS), adopted the guidelines published by the International Union of Crystallography requiring data deposition for all structures determined using NIGMS funds. As of this writing on 13 November 2001, there are over 16 500 entries in the archive.

Accompanying this rapid growth, the mode of access to PDB data has changed over the years as a result of improved technology. Data distribution is now primarily *via* the World Wide Web rather than *via* magnetic media. Further, the need to analyze diverse subsets of the data led to the development of modern data-management systems.

Initial use of the PDB had been limited to a small group of experts involved in structural biology research. Today, depositors to the PDB have expertise in the techniques of X-ray crystal structure determination, NMR, cryo-electron microscopy and theoretical modeling. PDB users are a very diverse group of researchers in biology and chemistry, as well as educators and students of all levels. The tremendous influx of data, soon to be fueled by the structural genomics initiative

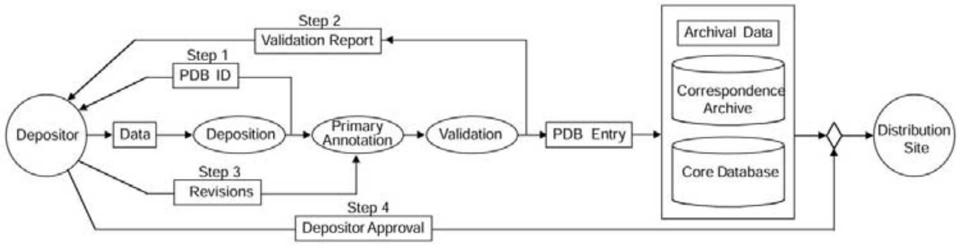

Figure 1
The steps involved in PDB data processing. Ellipses represent actions and rectangles define content. Figure reprinted from Berman, Westbrook *et al.* (2000) by permission of Oxford University Press.

and the increased recognition of the value of these data toward understanding biological function, continually demand new ways to collect, organize and distribute the data (Berman, Bhat *et al.*, 2000).

Since October 1998, the PDB has been managed by the three members of the Research Collaboratory for Structural Bioinformatics (RCSB) – Rutgers, The State University of New Jersey, the San Diego Supercomputer Center at the University of California, San Diego, and the National Institute of Standards and Technology. In this article, we describe the current procedures for collecting, validating, annotating and distributing PDB data. Finally, we describe the plans for further automating and improving the PDB so it can meet new emerging challenges posed by researchers and educators in the field of structural bioinformatics (Weissig & Bourne, 2002).

2. Data acquisition and processing

A key component of the PDB is the efficient capture and curation of the data: data processing. Data processing consists of data deposition, annotation and validation. These steps are part of a fully documented and integrated data-processing system shown in Fig. 1.

In the present system, data (atomic coordinates, structure factors and NMR restraints) may be submitted *via* e-mail or *via* the web-based AutoDep Input Tool (*ADIT*; Westbrook *et al.*, 1998; http://deposit.pdb.org/adit/) developed by the RCSB PDB. *ADIT* is built on top of the mmCIF dictionary that contains 1700 terms that define the macromolecular structure and the crystallographic experiment (Bourne *et al.*, 1997; Westbrook & Bourne, 2000). The mmCIF dictionary has been further extended to form the PDB exchange dictionary, which includes terms needed for tracking and other information-management purposes. *ADIT* is complemented by *MAXIT* (MAcromolecular EXchange Input Tool; Feng, Hsieh *et al.*, 1998), a program that performs many of the data-processing tasks and checks. This integrated system helps to ensure that the data submitted are consistent with the mmCIF dictionary which defines data types, enumerates ranges of allowable values where possible and describes allowable relationships between data values.

Each deposition to the PDB is represented by the PDBid – a four-character code of the form *nxyz*, where *n* is an integer and *x*, *y* and *z* are alphanumeric characters, *e.g.* 4hhb. The PDBid is assigned arbitrarily and is an immutable reference to the structure and indeed is the only absolute way of retrieving a desired structure from the PDB, although this shortcoming is being addressed (refer to §3). PDBids are never reused and remain the link between the structure and the literature reference that describes that structure.

After a structure has been deposited using *ADIT*, a PDBid is automatically and immediately returned to the author (Fig. 1, Step 1). This is the first stage, in which information about the structure is loaded into the internal core database, validated and annotated (see also §§4.1 and 2.2). This step involves using *ADIT* to help diagnose errors or inconsistencies in the files. The completely annotated entry as it will appear in the PDB resource, together with the validation information, is sent back to the depositor (Fig. 1, Step 2). After reviewing the processed file, the author sends any revisions (Fig. 1, Step 3). Depending on the nature of these revisions, steps 2 and 3 may be repeated. Once approval is received from the author (Fig. 1, Step 4), the entry and the tables in the internal core database are ready for distribution. The schema of this core database is a subset of the conceptual schema specified by the mmCIF dictionary.

All aspects of data processing, including communications with the author, are recorded and stored in the electronic correspondence archive. This makes it possible for the PDB staff to retrieve information about any deposited entry. Current status information, including the entry's authors, title and release status, is stored for each entry in the core database and is made accessible for query *via* the WWW interface (http://www.rcsb.org/pdb/status.html). Entries prior to release are categorized as 'in processing' (PROC), 'in depositor review' (WAIT), 'to be held until publication' (HPUB) or 'on hold until a depositor-specified date' (HOLD).

2.1. Content of the data collected by the PDB

All the data collected from depositors by the PDB are considered primary data. Primary data contain, in addition to the coordinates, general information required for all deposited structures and information specific to the method of structure determination. Table 1 contains the general information that the PDB collects for all structures as well as the information specific to X-ray and NMR experiments.

Historically, NMR data have been placed in a format defined around crystallographic information. The PDB is currently working with an NMR Task Force and the BioMagResBank (BMRB: Ulrich *et al.*, 1989) to develop an NMR data dictionary as an extension to the mmCIF. This

Table 1
Content of data in the PDB (reprinted from Berman *et al.*, 2001).

- Content of all depositions (X-ray and NMR)
  - Source – specifications such as genus, species, strain or variant of gene (cloned or synthetic); expression vector and host or description of method of chemical synthesis
  - Sequence – full sequence of all macromolecular components
  - Chemical structure of cofactors and prosthetic groups
  - Names of all components of the structure
  - Qualitative description of the characteristics of the structure
  - Literature citations for the structure submitted
  - Three-dimensional coordinates
- Additional items for X-ray structure determinations
  - Temperature factors and occupancies assigned to each atom
  - Crystallization conditions, including pH, temperature, solvents, salts, methods
  - Crystal data, including the unit-cell dimensions and space group
  - Presence of non-crystallographic symmetry
  - Data-collection information describing the methods used to collect the diffraction data including instrument, wavelength, temperature and processing programs
  - Data-collection statistics including data coverage, $R_{\rm sym}$, data above 1, 2, 3σ levels and resolution limits
  - Refinement information including R factor, resolution limits, number of reflections, method of refinement, σ cutoff, geometry r.m.s.d., σ
  - Structure factors – h, k, l, $F_{\rm obs}$, $\sigma(F_{\rm obs})$
- Additional items for NMR structure determinations
  - For an ensemble, the model number for each coordinate set that is deposited and an indication if one should be designated as a representative
  - Data-collection information describing the types of methods used, instrumentation, magnetic field strength, console, probe head, sample tube
  - Sample conditions, including solvent, macromolecule concentration ranges, concentration ranges of buffers, salts, antibacterial agents, other components, isotopic composition
  - Experimental conditions, including temperature, pH, pressure and oxidation state of structure determination and estimates of uncertainties in these values
  - Non-covalent heterogeneity of sample, including self-aggregation, partial isotope exchange, conformational heterogeneity resulting in slow chemical exchange
  - Chemical heterogeneity of the sample (*e.g.* evidence for deamidation or minor covalent species)
  - A list of NMR experiments used to determine the structure, including those used to determine resonance assignments, NOE/ROE data, dynamical data, scalar coupling constants and those used to infer hydrogen bonds and bound ligands. The relationship of these experiments to the constraint files are given explicitly
  - Constraint files used to derive the structure as described in Task Force recommendations

dictionary includes descriptions of the solution components, the experimental conditions, enumerated lists of the instruments used and information about structure refinement. This dictionary will be used for deposition and validation tools specific to NMR structures. NMR coordinate data and constraints are currently deposited with the PDB and other NMR-specific experimental data are deposited with the BMRB. Plans are in place to have a single interface for the deposition of data to both the BMRB and the PDB.

The information content of data submitted by the depositor is likely to change as new methods for data collection, structure determination and refinement evolve. A case in point is the need for structural genomics projects to collect all information that would be in the 'material and methods' section of a paper describing a structure. The ways in which these data are captured are also changing as the software for structure determination and refinement evolve to produce the necessary data items as part of their output. The ontology-driven approach to software development used by the PDB makes it simple to collect new items of data once they are described in the mmCIF or extension dictionary.

2.2. Validation and annotation

Validation refers to the procedure for assessing the quality of deposited atomic models (structure validation) and for assessing how well these models fit the experimental data (experimental validation). Annotation refers to the process of adding information to the entry that results from the validation process. The PDB validates structures using accepted community standards as part of *ADIT*'s integrated data-processing system. The following checks are run and are summarized in a letter that is communicated directly to the depositor.

2.2.1. Covalent bond distances and angles. Proteins are compared with standard values from Engh & Huber (1991); nucleic acid bases are compared with standard values from Clowney *et al.* (1996); sugar and phosphates are compared with standard values from Gelbin *et al.* (1996).

2.2.2. Stereochemical validation. All chiral centers of proteins and nucleic acids are checked for correct stereochemistry.

2.2.3. Atom nomenclature. The nomenclature of all atoms is checked for compliance with IUPAC standards (IUPAC–IUB Joint Commission on Biochemical Nomenclature, 1983) and is adjusted if necessary.

2.2.4. Close contacts. The distances between all atoms within the asymmetric unit of crystal structures and the unique molecule of NMR structures are calculated. For crystal structures, contacts between symmetry-related molecules are also checked.

2.2.5. Ligand and atom nomenclature. Residue and atom nomenclature is compared against a standard dictionary (ftp://ftp.rcsb.org/pub/pdb/data/monomers/het_dictionary.txt) for all ligands as well as standard residues and bases. Unrecognized ligand groups are flagged and any discrepancies in known ligands are listed as extra or missing atoms. New ligands are added to the dictionary as they are deposited.

2.2.6. Sequence comparison. The sequence provided by the depositor is compared with the sequence derived from the coordinate records. This information is displayed in a table where any differences or missing residues are annotated. During the annotation process the sequence database references provided by the author are checked for accuracy. If no reference is given, a *BLAST* (Altschul *et al.*, 1990) search is used to find the best match. Any conflict between the depositor's sequence and the sequence derived from the coordinate records is further resolved and annotated by comparison with other sequence databases as needed.

2.2.7. Distant waters. The distances between all water O atoms and all polar atoms (oxygen and nitrogen) of the macromolecules, ligands and solvent in the asymmetric unit are calculated. Distant solvent atoms are repositioned using crystallographic symmetry such that they fall within the solvation sphere of the macromolecule.

In almost all cases, serious errors detected by these checks have been corrected through annotation and correspondence with the authors. It is also possible for authors to run these validation checks against structures before they are deposited. A validation server (http://deposit.pdb.org/validate/) has been made available for this purpose. In addition to the summary report letter, the server also provides output from *PROCHECK* (Laskowski *et al.*, 1993), *NUCheck* (Feng, Westbrook *et al.*, 1998) and *SFCHECK* (Vaguine *et al.*, 1999). A summary atlas page and molecular-graphics images are also produced.

The PDB continuously reviews the validation methods used and will continue to integrate new procedures as they become available and are accepted as community standards.

2.3. Data-deposition sites

Data are deposited to the PDB *via* one of three sites. Because it is critical that the final archive is kept uniform, the content and format of the final files as well as the methods used to check them must be the same.

The RCSB–PDB deposition site (http://deposit.pdb.org/adit/) has developed software programs for data deposition, validation and processing, including *ADIT* and the Validation Server. The *ADIT* system, as described above, is also used to process the data deposited.

The Institute for Protein Research at Osaka University in Japan has collaborated with the PDB to establish another deposition center (http://pdbdep.protein.osaka-u.ac.jp/adit/). All data deposited at this center (primarily depositors in Asia) are also processed by this Osaka group using the *ADIT* system.

The Macromolecular Structure Database group at the European Bioinformatics Institute (MSD-EBI) processes data that are submitted to them *via AutoDep* (http://autodep.ebi.ac.uk/). After processing, the data are sent to the RCSB in PDB format for inclusion in the central archive. A common mmCIF exchange dictionary has been developed with this group, which will help ensure a higher degree of data uniformity in the archival data in the future.

The PDB has also ported its data-processing software to a stand-alone system that does not require Internet access. This system is soon to be released for use by authors who wish to check data in their home laboratories.

2.4. Data-processing statistics

Production processing of PDB entries by the RCSB began on 27 January 1999. The median time from deposition to the completion of data processing including author interactions is less than two weeks. The number of structures with a HOLD release status remains at about 16% of all submissions: 63% are hold until publication and 21% are released immediately after processing.

Fig. 2 shows the growth of PDB data since the archive began. Fig. 2(*a*) shows the total number of structures available in the archive per year. Fig. 2(*b*) shows the number of residues released in the PDB each year, indicating how the complexity of structures released into the archive has increased over time.

The current breakdown of the types of structures in the PDB can be found at http://www.rcsb.org/pdb/holdings.html.

Table 2
Demographics of data released in the PDB (as of 13 November 2001).

	Molecule type				
Experimental technique	Proteins, peptides and viruses	Protein–nucleic acid complexes	Nucleic acids	Carbo-hydrates/ other	Total
X-ray diffraction/other	12448	601	591	14	13654
NMR	2056	78	407	4	2545
Theoretical modeling	301	23	23	0	347
Total	14805	702	1021	18	16546

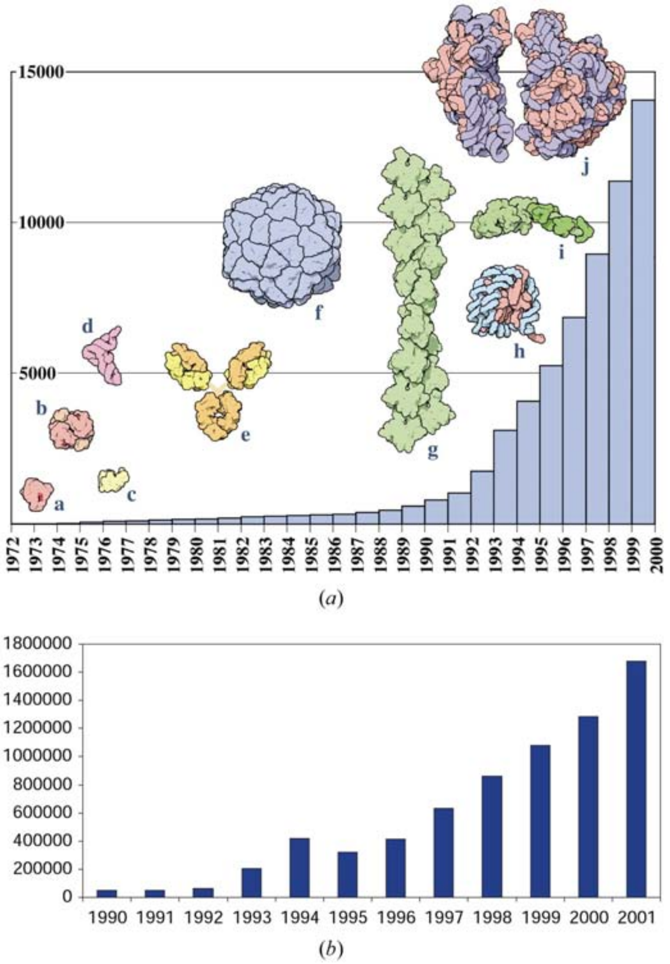

Figure 2
(*a*) Growth chart of the PDB showing the total number of structures available in the PDB archive per year and highlighting example structures from different time periods: a, myoglobin; b, hemoglobin; c, lysozyme; d, transfer RNA; e, antibodies; f, entire viruses; g, actin; h, the nucleosome; i, myosin; j, 30S ribosomal subunits. Images were created by Dr David Goodsell, who creates the PDB's Molecule of the Month series. Figure originally appeared in the International Union of Crystallography Newsletter (2001). Images, descriptions and the molecules and links to related information can be found at http://www.rcsb.org/pdb/molecules/molecule_list.html. (*b*) Number of residues released in the PDB per year.

Data at the time of writing (13 November 2001) are shown in Table 2. The PDB contained 16 546 publicly accessible structures. Of these, 13 654 (82%) were determined by X-ray methods, 2545 (15%) were determined by NMR and 347 (2%) were theoretical models. Overall, 44% of the entries include experimental data. Another 1959 entries are on hold.

Table 3
Query results on uniform *versus* non-uniform data (from 29 August 2000).

The attributes listed can be searched by using the SearchFields interface. The numbers given are the result of entering the query term in the desired field on both non-uniform and uniform data. These data are currently available as database tables, but not available in the individual PDB data files. They are available in the mmCIF data files described in §3. Information about the data-uniformity project is archived at http://www.rcsb.org/pdb/uniformity/. Table reprinted from Bhat *et al.* (2001) by permission of Oxford University Press.

Attribute	Query term	Non-uniform	Uniform
Resolution (Å)	2.1–2.5	3061	3492
Primary citation	*J. Mol. Biol.*	1953	2331
Journal name	*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

Table 4
Summary of released entries containing nomenclature and chemical representation errors.

	Incorrect sequence	Sequence-coordinate mismatch	Atom nomenclature errors	Stereochemical labeling errors
Legacy data (8368 entries)†	166	90	3311	294
1999 data (3150 entries)‡	0	5	162	3
2000 data (3569 entries)‡	0	0	31	3

† Pre-October 1998 entries (excluding nucleic acid-containing crystal structures). ‡ Structures processed and released by the RCSB.

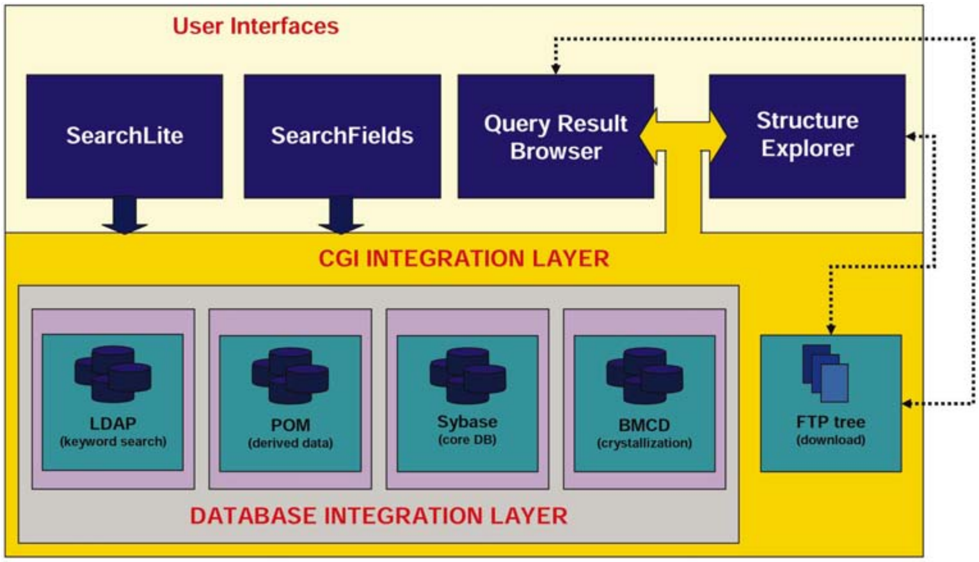

Figure 3
The integrated query interface to the PDB.

3. Data uniformity

A key goal of the PDB is to make the archive as consistent and error-free as possible. As indicated above, all new depositions are reviewed carefully by annotators before release. Errors found subsequent to release by authors and PDB users are addressed as rapidly as possible. Minor errors result in revisions to the entry which are annotated within the entry; major errors lead to a superceding entry or entry withdrawal. Corrections and updates to entries are sent to deposit@rcsb.rutgers.edu.

'Legacy data', that is, data submitted prior to October 1998, comply with several different PDB formats, and variation exists in how the same features are described for different structures within each format. The inconsistency of formats and nomenclature conventions make it difficult to consistently parse these data and query across the archive. As an immediate solution to the query problem, particular records across all entries in the archive were corrected; these included citation, R factor and resolution (Bhat *et al.*, 2001). These corrections were loaded into the database and thus it was possible to query on these features and obtain accurate results (Table 3). However, these data were not available in the PDB files. To provide uniform data for each structure we used the software that was developed and tested for primary processing and revalidated all the data in the archive. Corrections were made to nomenclature and special attention was paid to consistency of the chemical description of the macromolecule and the ligand. Examples of the types of errors that were found and corrected are shown in Table 4.

The corrected files were released in mmCIF format and can be found at ftp://beta.rcsb.org/pub/pdb/uniformity/data/mmCIF/ (Westbrook *et al.*, 2002). The original PDB files will continue to be available as they are a historical record and have been the basis of many research projects. Software is available from the PDB to transform the mmCIF files to PDB-formatted files. In the future, these files will form the basis of the PDB databases accessible *via* the WWW.

4. Data access

The PDB is presently incremented once per week, with new data becoming available on Wednesday mornings in most parts of the world through a number of mirror sites. The following describes the database

Table 5
Current query capabilities of the PDB.

Table reprinted from Berman, Westbrook *et al.* (2000) by permission of Oxford University Press.

Query options	
SearchLite	Any word or combination of words in the PDB
SearchFields	General information: PDB identifier, citation author, chain type (protein, DNA *etc*.), PDB HEADER, experimental technique, deposition/release date, citation, compound information, EC number, text search Sequence and secondary structure: chain length, FASTA search, short sequence pattern, secondary-structure content Crystallographic experimental information: resolution, space group, unit-cell dimensions, parameters
Status	PDB identifier, deposition author, title, holding status, deposition date, release date, prereleased sequences
Result analysis	
Single Structure: Structure Explorer	
Summary	Compound name, authors, experimental method, classification, source, primary citation, deposition date, release date, resolution, *R* value, space group, unit-cell parameters, polymer chain identifiers, number of residues, HET groups, number of atoms
View Structure	VRML, *RasMol*, QuickPDB (Java Applet), *Chime*, still images
Download/Display File	HTML and text formats for display; PDB and mmCIF formats with different compression options for download
Structural Neighbors	List of sites for finding structural homologues
Geometry	Unusual dihedral angles, bond angles and bond lengths
Other Sources	Links to other sources of information
Sequence Details	Chain Ids, number of residues per chain, molecular weight, chain type, secondary-structure assignment; download sequence only in FASTA format
Crystallization information	Conditions under which the crystals were obtained
Previous Versions	Versions of the structure replaced by the current version if applicable
Nucleic Acid Database Atlas Entry	Detailed information from the NDB (if applicable)
Quick Report	Nucleic acid geometry if applicable
Structure Factors	Experimental data if available
Multiple Structure: Results Browser	
Summary List	Deposition date, resolution, experimental method, classification, compound name
Download Structures or Sequences	mmCIF and PDB compressed files (gzip, tar, compressed); sequences in FASTA format
Query Refinement	Iterative query over result set using OR, AND or NOT Boolean logic
Tabular Report	Cell dimensions, primary citation, structure identifiers, sequence, experimental details, refinement details
Query Review	Summary of queries submitted thus far with the option to return to one of them

architecture used by the PDB, how users access these databases *via* the web and how data files can be accessed *via* the web and *via* ftp.

4.1. Database architecture

The current PDB data-management system consists of several heterogeneous data sources that are integrated through Perl CGI scripts (Fig. 3). While this leads to some redundancy, since parts of the data are stored multiple times, it allows efficient access. The complete system is currently being re-engineered to maintain this efficiency while providing more manageability with less redundancy. The new system will be based upon a new relational database-management system. We consider here the five core components of the current system.

First, the core relational database (Sybase SQL server release 11.0, Emeryville, CA, USA) stores the primary experimental and coordinate data as described in Table 1. These data are retrieved by the reporting options available through the web interface. Second, the ftp archive (ftp://ftp.rcsb.org/pdb/) provides the data files in PDB and mmCIF formats as well as the data dictionaries to which they correspond. Third, the Property Object Model (POM) data-management system (Shindyalov & Bourne, 1997) is used for more efficient access to certain structural features, such as sequence. POM consists of indexed objects containing native data (*e.g.* atomic coordinates) and derived properties [*e.g.* secondary structure calculated according to Kabsch & Sander (1983)]. Fourth, the Netscape LDAP server is used to index the textual content of the PDB and provides support for keyword searches. Fifth, the *Molecular Information Agent* (*MIA*; http://mia.sdsc.edu) is used to collect and store hyperlinks and limited other information for approximately 75 external data resources in a separate Sybase database (see http://www.rcsb.org/pdb/mia.html). *MIA* formulates a query to each of these data sources based on the PDBid and parses the results of the query to provide the information viewable through the 'Other Sources' option of an entry's Structure Explorer page. *MIA* includes housekeeping software, for example, to coordinate the simultaneous access to these data sources and to timeout if a particular site is down. These five components, associated software and web pages constitute the system which is mirrored to a number of sites worldwide (see below).

Finally, there is a close integration to three external resources (*i.e.* not mirrored as part of the PDB).

(i) The Biological Macromolecule Crystallization Database (BMCD; Gilliland, 1988; Gilliland *et al.*, 2002) is organized as a relational database within Sybase and contains three general categories of literature-derived information: macromolecular, crystal and summary data.

(ii) The Nucleic Acid Database (NDB; Berman *et al.*, 1992, 2002), which contains information specifically pertaining to DNA and RNA.

(iii) The CE database (Shindyalov & Bourne, 1998) of three-dimensional protein structure alignments.

The latter raises an important point of PDB policy. The alignment of structures depends to some degree on the assumptions of the method being used. Since there is no agreement in the community at present as to a *de facto* standard method for protein structure alignment, the PDB's policy is to provide access to a variety of alignments and classification schemes (Murzin *et al.*, 1995; Gibrat *et al.*, 1996; Orengo *et al.*, 1997; Holm & Sander, 1998; Shindyalov & Bourne, 1998). In short, the PDB's policy is to provide a portal (entry point) to relevant information, but not impose judgment on which methodology should be used.

In the current implementation, communication among the five PDB components and these databases has been accomplished using the Common Gateway Interface (CGI) in such a way as to hide the intricacies of the underlying databases from the user. An integrated web interface dispatches a query to the appropriate database(s), which then execute the query. Each database returns the PDBids that satisfy the query and the CGI program integrates the results. Complex queries are performed by repeating the process and having the interface program perform the appropriate Boolean operation(s) on the collection of query results. A variety of output options are then available for use with the final list of selected structures.

The newly created and uniform mmCIFs will enable the PDB to substantially improve its underlying database architecture. The mmCIFs are loaded into a new relational database with a schema that conceptually conforms closely to the mmCIF dictionary. The results will provide access to data not currently available and do so in a way that is easier to maintain.

4.2. User web access

Currently, three distinct query interfaces are available for the query of data within PDB: Status Query (http://www.rcsb.org/pdb/status.html), SearchLite (http://www.rcsb.org/pdb/searchlite.html) and SearchFields (http://www.rcsb.org/pdb/cgi/queryForm.cgi). Table 5 summarizes the current query and analysis capabilities of the PDB. Fig. 4 illustrates how the various query options are organized.

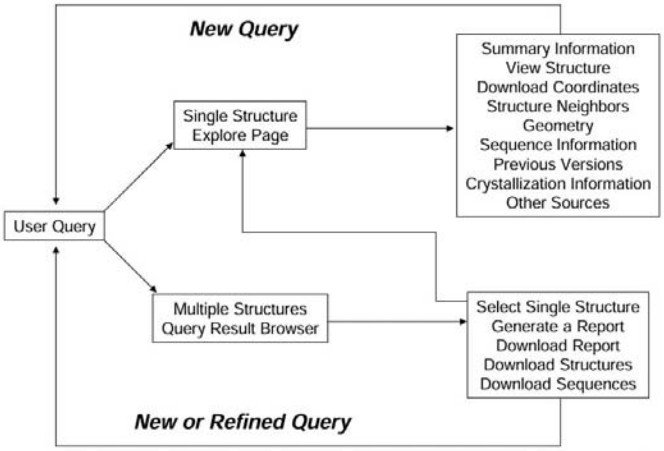

Figure 4
The layout of the PDB query system. Figure reprinted from Berman *et al.* (2001).

The Status query allows the user to review information on structures deposited but not yet released. In addition to an author list, title and release status, the depositor may opt to display sequence information for the unreleased entry. This provides a set of useful targets for structure-prediction studies.

SearchLite provides a single form field for keyword searches. Textual information within the PDB file, such as dates and some experimental data, are searched. Boolean searching and restriction of keywords can be used to conform to specific attributes. For example, 'green' can be attributed to an author name or a common name for a protein.

SearchFields is for more advanced searches and presents a customizable query form that can be used to search different data items, including macromolecule type, citation authors, sequence (*via* a FASTA search; Pearson & Lipman, 1988) and release dates. For enzymes, it is possible to browse these structures using the Enzyme Commission hierarchy. The numbers of entries at each level are reported as the user traverses the hierarchy.

Search results are displayed in the 'Query Result Browser', which can be used to generate reports, download data and perform further searches. The 'Structure Explorer' interface provides detailed information on a single structure. On-line tutorials for accessing PDB data *via* the WWW are available at http://www.rcsb.org/pdb/info.html#PDB_Users_Guides.

4.3. Application web access

Interfaces to both single and multiple structures are accessible to other web resources and applications through the simple CGI application programmer interface (API) described at http://www.rcsb.org/pdb/linking.html. Stated another way, a URL can be constructed with either single or multiple embedded PDBids and used to return results on those structures. Many web sites worldwide use this mechanism to reference PDB structures.

The PDB web site is maintained on redundant load balanced servers and receives in excess of 100 000 page hits per day. On average, a structure is downloaded every second 24 h per day, seven days per week.

4.4. Ftp access

All structures, in PDB and mmCIF formats, are available for download from the PDB ftp site. Dictionaries, documentation and PDB-provided software are also available. Instructions and software for mirroring the PDB ftp archive as a local copy are available at http://www.rcsb.org/pdb/ftpproc.final.html.

4.5. Distribution

As stated, the PDB distributes coordinate data in PDB and CIF formats, structure-factor files and NMR constraint files. In addition it provides derived data, documentation and software. The PDB files enjoy widespread usage by individual researchers and by databases of structural properties (Weissig & Bourne, 2002).

New data officially become available at 2:00 AM PST each Wednesday. PDB mirrors have been established in Japan (Osaka University), Singapore (National University Hospital), Brazil (Universidade Federal de Minas Gerais) and in the UK (Cambridge Crystallographic Data Centre). Additionally other sources of PDB data exist, but are provided through different interfaces. For a partial list see http://www.rcsb.org/pdb/mirrors.html. The PDB also distributes a quarterly CD-ROM set that is essentially a copy of the ftp site. Data are distributed as compressed files using the compression utility program gzip. Refer to http://www.rcsb.org/pdb/cdrom.html for details of how to order CD-ROM sets. There is no charge for this service.

Table 6
Web links.

(*a*) PDB mirror sites

RCSB partner sites	
SDSC, La Jolla, CA (US)	http://www.pdb.org/, ftp://ftp.rcsb.org/
Rutgers University, Piscataway, NJ (US)	http://rutgers.rcsb.org/
NIST, Gaithersburg, MD (US)	http://nist.rcsb.org/
Other RCSB mirrors	
CCDC, United Kingdom	http://pdb.ccdc.cam.ac.uk/, ftp://pdb.ccdc.cam.ac.uk/rcsb/
National University of Singapore, 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/

(*b*) PDB sites of interest.

Source	Information content
Deposition	
http://deposit.pdb.org/adit/ (RCSB-Rutgers)	*ADIT* web site (deposit@rcsb.rutgers.edu)
http://pdbdep.protein.osaka-u.ac.jp/adit/ (Osaka University)	*ADIT* web site (adit@adit.protein.osaka-u.ac.jp)
http://autodep.ebi.ac.uk/ (MSD-EBI)	*AutoDep* (pdbhelp@ebi.ac.uk)
http://deposit.pdb.org/validate/	*ADIT* validation server
http://deposit.pdb.org/	Deposition, format and *ADIT* FAQs
Query	
http://www.rcsb.org/pdb/status.html	PDB status search
http://www.rcsb.org/pdb/searchlite.html	SearchLite
http://www.rcsb.org/pdb/cgi/queryForm.cgi	SearchFields
http://www.rcsb.org/pdb/linking.html	Information on linking to the PDB
http://mia.sdsc.edu	Molecular information agent
http://www.rcsb.org/pdb/mia.html	MIA at the PDB FAQ
http://www.rcsb.org/pdb/ftpproc.final.html	RCSB PDB mirror protocol
PDB features	
http://www.rcsb.org/pdb/strucgen.html	Structural genomics resources
http://www.rcsb.org/pdb/uniformity	Information about the PDB Data Uniformity Project
http://www.rcsb.org/pdb/lists/pdb-l/	PDB Listserv for community announcements
http://www.rcsb.org/pdb/cdrom.html	CD-ROM information
http://www.rcsb.org/pdb/info.html#PDB_Users_Guides	Tutorials for deposition and query
http://www.rcsb.org/pdb/newsletter/	PDB Newsletters
http://www.rcsb.org/pdb/holdings.html	Statistics on the PDB archive
http://www.rcsb.org/pdb/ftpproc.final.html	FTP mirroring information
http://www.rcsb.org/pdb/cdrom.html	CD-ROM ordering information
info@rcsb.org	General help desk

5. Outreach

Active outreach ensures that the community of PDB users is fully informed about our capabilities and activities and that the PDB receives feedback that allows it to improve its services. Outlined below are some of the key outreach activities.

5.1. Help desk

The electronic help desk (info@rcsb.org) addresses questions about all aspects of the PDB and about general structural biology. Questions are generally addressed within one or two working days. The list receives an average of 130 inquiries per month. The PDB also maintains two other addresses: deposit@rcsb.rutgers.edu, for questions concerning data deposition, and help@rcsb.rutgers.edu, for questions concerning *ADIT*.

5.2. pdb-l@rcsb.org

A list server at pdb-l@rcsb.org is maintained for use by the community to make announcements and conduct discussions on activities relating to the PDB. It is PDB policy that this list be reserved for open discussion by the community and not for use by the PDB itself. An archive of the discussions that take place on this list can be found at http://www.rcsb.org/pdb/lists/pdb-l/.

5.3. PDB web site

The web site is updated weekly with news, recent developments, and improvements to existing documents. The site includes tutorials and user guides for query, deposition and file formats.

5.4. PDB publications

The PDB publishes a quarterly newsletter available *via* e-mail or postal mail (see http://www.rcsb.org/pdb/newsletter.html). PDF versions of the PDB newsletter dating back to September 1974 are available at this site. Flyers, tutorials and an Annual Report are accessible from the PDB web site and by sending mail to info@rcsb.org.

5.5. Scientific meetings

PDB members attend a wide variety of meetings, presenting posters, talks and exhibit booths. Among the meetings attended are the American Crystallographic Association's Annual Meeting, Protein Society's Symposium, the Intelligent Systems in Molecular Biology's annual meeting and the

International Union of Crystallography's Congress and General Assembly.

6. Future

Structural biology is a fast-evolving field that poses challenges to the collection, curation and distribution of macromolecular structure data. Over the past three years, the number of depositions has averaged approximately 50 per week. However, with the advent of a number of structural genomics initiatives worldwide this number is likely to increase. We estimate that the PDB could contain 35 000 structures by 2005. This presents a challenge to timely distribution while maintaining high quality. We believe our approach to information management should permit us to scale to accommodate the anticipated large data influx. We are endeavoring to work closely with all structural genomics projects to automatically collect more data and redesigning our database systems to provide a more scalable system. In terms of access, we have worked with the Object Management Group to define a CORBA (Common Object Request Broker) standard for macromolecular structure which is closely aligned with the mmCIF dictionary. Eventually, this will provide a fine-grained access to items of PDB data by users and their applications. This will be achieved by providing a CORBA server that is currently under development.

The maintenance and further development of the PDB are community efforts. The willingness of others to share ideas, software and data provides a depth to the resource not obtainable otherwise. It is important to acknowledge the contribution of scientists and staff at the BNL, who maintained the archive for many years. New input is constantly being sought and the PDB invites you to make comments at any time by sending e-mail to info@rcsb.org. A summary of all the URLs specified in this article is given in Table 6 for easy reference.

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. The efforts of the entire PDB staff (listed at http://www.rcsb.org/pdb/rcsb-group.html) are gratefully acknowledged. The Protein Data Bank does not recommend or endorse any specific commercial software products. This article is also appearing in the forthcoming book tentatively titled *Structural Bioinformatics* to be published by John Wiley & Sons, Inc. (P. E. Bourne & H. Weissig, Editors).

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). *J. Mol. Biol.*, **215**, 403–410.

Berman, H. M., Bhat, T. N., Bourne, P. E., Feng, Z., Gilliland, G., Weissig, H. & Westbrook, J. (2000). *Nature Struct. Biol.* **7**, 957–959.

Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S. H., Srinivasan, A. R. & Schneider, B. (1992). *Biophys. J.* **63**, 751–759.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2001). *International Tables for Crystallography*, Vol. F, *Crystallography of Biological Macromolecules*, edited by M. G. Rossmann & E. Arnold, pp. 675–662. Dordrecht: Kluwer Academic Publishers.

Berman, H., Westbrook, J., Feng, Z., Iype, L., Schneider, B. & Zardecki, C. (2002). *Acta Cryst.* D**58**, 889–898.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Bhat, T. N., Bourne, P., Feng, Z., Gilliland, G., Jain, S., Ravichandran, V., Schneider, B., Schneider, K., Thanki, N., Weissig, H., Westbrook, J. & Berman, H. M. (2001). *Nucleic Acids Res.* **29**, 214–218.

Bourne, P. E., Berman, H. M., Watenpaugh, K., Westbrook, J. D. & Fitzgerald, P. M. D. (1997). *Methods Enzymol.* **277**, 571–590.

Clowney, L., Jain, S. C., Srinivasan, A. R., Westbrook, J., Olson, W. K. & Berman, H. M. (1996). *J. Am. Chem. Soc.* **118**, 509–518.

Engh, R. A. & Huber, R. (1991). *Acta Cryst.* A**47**, 392–400.

Feng, Z., Hsieh, S.-H., Gelbin, A. & Westbrook, J. (1998*a*). NDB 120 *MAXIT: Macromolecular Exchange and Input Tool.* Rutgers University, New Brunswick, NJ, USA.

Feng, Z., Westbrook, J. & Berman, H. M. (1998*b*). NDB-407 *NUCheck.* Rutgers University, New Brunswick, NJ, USA.

Gelbin, A., Schneider, B., Clowney, L., Hsieh, S.-H., Olson, W. K. & Berman, H. M. (1996). *J. Am. Chem. Soc.* **118**, 519–528.

Gibrat, J.-F., Madej, T. & Bryant, S. H. (1996). *Curr. Opin. Struct. Biol.* **6**, 377–385.

Gilliland, G. L. (1988). *J. Cryst. Growth*, **90**, 51–59.

Gilliland, G. L., Tung, M. & Ladner, J. E. (2002). *Acta Cryst.* D**58**, 916–920.

Holm, L. & Sander, C. (1998). *Nucleic Acids Res.* **26**, 316–319.

International Union of Crystallography Newsletter (2001). Vol. 1. Issue 9. Buffalo, NY: IUCr.

IUPAC–IUB Joint Commission on Biochemical Nomenclature (1983). *Eur. J. Biochem.* **131**, 9–15.

Kabsch, W. & Sander, C. (1983). *Biopolymers*, **22**, 2577–2637.

Laskowski, R. A., McArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). *J. Mol. Biol.* **247**, 536–540.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). *Structure*, **5**, 1093–1108.

Pearson, W. R. & Lipman, D. J. (1988). *Proc. Natl Acad. Sci. USA*, **24**, 2444–2448.

Shindyalov, I. N. & Bourne, P. E. (1997). *CABIOS*, **13**, 487–496.

Shindyalov, I. N. & Bourne, P. E. (1998). *Protein Eng.* **11**, 739–747.

Ulrich, E. L., Markley, J. L. & Kyogoku, Y. (1989). *Protein Seq. Data Anal.* **2**, 23–37.

Vaguine, A. A., Richelle, J. & Wodak, S. J. (1999). *Acta Cryst.* D**55**, 191–205.

Weissig, H. & Bourne, P. E. (2002). *Acta Cryst.* D**58**, 908–915.

Westbrook, J. & Bourne, P. E. (2000). *Bioinformatics*, **16**, 159–168.

Westbrook, J., Feng, Z. & Berman, H. M. (1998). RCSB-99 *ADIT: The AutoDep Input Tool.* Department of Chemistry, Rutgers, the State University of New Jersey, USA.

Westbrook, J., Feng, Z., Jain, S., Bhat, T. N., Thanki, N., Ravichandran, V., Gilliland, G. L., Bluhm, W., Weissig, H., Greer, D. S., Bourne, P. E. & Berman, H. M. (2002). *Nucleic Acids Res.* **30**, 245–248.