International Tables for Crystallography (2006). Vol. F, Chapter 24.5, pp. 675–684.

24.5. The Protein Data Bank, 1999–

BY H. M. BERMAN, J. WESTBROOK, Z. FENG, G. GILLILAND, T. N. BHAT, H. WEISSIG, I. N. SHINDYALOV AND P. E. BOURNE

24.5.1. Introduction

The Protein Data Bank (PDB) was established at Brookhaven National Laboratory (BNL) (Bernstein *et al.*, 1977) in 1971 as an archive for biological macromolecular crystal structures. In the beginning there were seven structures, and each year a handful more were deposited. In the 1980s the number of deposited structures began to increase dramatically. This was due to the improved technology for all aspects of the crystallographic process, the addition of structures determined by nuclear magnetic resonance (NMR) methods and changes in the community views about data sharing. By the early 1990s the majority of journals required a PDB accession code and at least one funding agency (National Institute of General Medical Sciences) adopted the guidelines published by the IUCr requiring data deposition for all structures.

The mode of access to PDB data has changed over the years as a result of improved technology, notably the availability of the World Wide Web (WWW) replacing distribution solely *via* magnetic media. Further, the need to analyse diverse data sets required the development of modern data-management systems.

Initial use of the PDB had been limited to a small group of experts involved in structural research. Today depositors to the PDB have varying expertise in the techniques of X-ray crystal-structure determination, NMR, cryoelectron microscopy and theoretical modelling. Users are a very diverse group of researchers in biology and chemistry, community scientists, educators and students at all levels. The tremendous influx of data soon to be fuelled by the structural genomics initiative, and the increased recognition of the value of the data toward understanding biological function, demand new ways to collect, organize and distribute the data.

The vision of the Research Collaboratory for Structural Bioinformatics (RCSB)* is to create a resource based on the most modern technology that would facilitate the use and analysis of structural data and thus create an enabling resource for biological research. In October 1998, the management of the PDB became the responsibility of the RCSB.† In this chapter, we describe the current procedures for deposition, processing and distribution of PDB data by the RCSB. We conclude with some current developments of the PDB.

24.5.2. Data acquisition and processing

A key component of creating the public archive of information 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 the fully documented and integrated data-processing system shown in Fig. 24.5.2.1.

In the present system (Fig. 24.5.2.2), data (atomic coordinates, structure factors and NMR restraints) may be submitted *via* e-mail or *via* the *AutoDep Input Tool* [*ADIT*: http://pdb.rutgers.edu/adit (Westbrook *et al.*, 1998)] developed by the RCSB. *ADIT*, which is also used to process the entries, is built on top of the mmCIF dictionary, which is an ontology of 1700 terms that define the macromolecular structure and the crystallographic experiment (Bourne *et al.*, 1997), and a data-processing program called *MAXIT* (Macromolecular Exchange and Input Tool; Feng, Hsieh *et al.*, 1998). This integrated system helps to ensure that the data that are deposited for an entry are consistent and error-free after annotation.

After a structure has been deposited using *ADIT*, a PDB identifier is sent to the author automatically and immediately (Fig. 24.5.2.1, step 1). This is the first stage in which information about the structure is loaded into the internal core database (see Section 24.5.3). The entry is then annotated by PDB staff using *ADIT*; several validation reports about the structure are produced. The completely annotated entry as it will appear in the PDB resource, together with the validation information, is sent back to the depositor (step 2). After reviewing the processed file, the author sends any revisions (step 3). Depending on the nature of these revisions, steps 2 and 3 may be repeated. Once approval is received from the author (step 4), the entry and the tables in the internal core database are ready for distribution.

All aspects of data processing, including communications with the author, are recorded and stored in the correspondence archive. This makes it possible for the PDB staff to retrieve information about any aspect of the deposition process and to monitor the efficiency of PDB operations closely.

Current status information including a list of authors, title and release category 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 before 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).

24.5.2.1. *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 24.5.2.1 contains the general information that the PDB collects for all structures as well as the additional information collected for those structures determined by X-ray methods. The additional items listed for the NMR structures are derived from the International Union of Pure and Applied Chemistry recommendations (Markley *et al.*, 1998) and will be implemented in the near future.

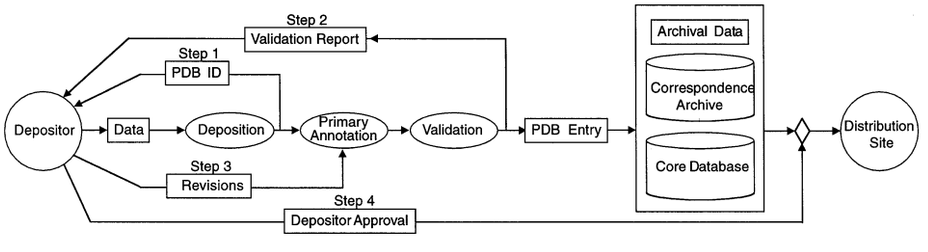

Fig. 24.5.2.1. The steps in PDB data processing. Ellipses represent actions and rectangles define content.

* The Research Collaboratory for Structural Bioinformatics (RCSB) is a consortium consisting of three institutions: Rutgers, The State University of New Jersey; San Diego Supercomputer Center, University of California, San Diego; and the National Institute of Standards and Technology.

† A call for proposals was issued by the National Science Foundation in 1998. The award was made to the RCSB after peer review of the proposals submitted.

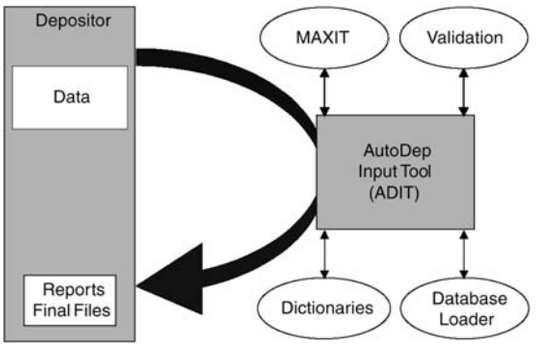

Fig. 24.5.2.2. The integrated tools of the PDB data-processing system.

The information content of data submitted by the depositor is likely to change as new methods for data collection, structure determination and refinement evolve and advance. In addition, the ways in which these data are captured is likely to change as the software for structure determination and refinement produce the necessary data items as part of their output. The data-input system for the PDB, *ADIT*, has been designed so as to incorporate these likely changes easily.

24.5.2.2. *Validation*

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). The PDB validates structures using accepted community standards as part of *ADIT*'s integrated data-processing system. All validation reports are communicated directly to the depositor. It is also possible to run these validation checks against structures that are not being deposited. A validation server (http://pdb.rutgers.edu/validate/) has been made available for this purpose.

Several types of checks are used in this process: *PROCHECK* (Laskowski *et al.*, 1993) is used for checking the structural features of proteins and *NUCheck* (Feng, Westbrook & Berman, 1998) is used for checking the structural features of nucleic acids. The information currently checked includes the following: bond lengths and bond angles, nomenclature, sequence, stereochemistry, torsion angles, ligand geometry, planarity of peptide bonds, intermolecular

Table 24.5.2.1. *Content of data in the PDB*

(*a*) 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 in structure
Qualitative description of characteristics of structure
Literature citations for the structure submitted
Three-dimensional coordinates

(*b*) 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 noncrystallographic 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_{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.
Structure factors – h, k, l, F_{obs}, $\sigma(F_{obs})$

(*c*) Additional items for NMR structure determinations

Model number for each coordinate set that is deposited and an indication if one should be designated as a representative, or an energy-minimized average model provided
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

Table 24.5.2.2. *Demographics of the released data in the PDB as of 14 September 1999*

Experimental technique	Molecule type				
	Proteins, peptides, and viruses	Protein–nucleic acid complexes	Nucleic acids	Carbohydrates and other	Total
X-ray diffraction and other	7946	390	439	14	8789
NMR	1365	53	270	4	1692
Theoretical modelling	202	16	15	0	233
Total	9513	459	724	18	10714

contacts, and positions of water molecules. In consultation with the community, other structure checks will be implemented over the next few years.

The experimental data are also checked. Currently, X-ray crystallographic data are validated and plans for checking NMR data are in progress. For X-ray crystallographic structures, the structure factors are validated using *SFCHECK* (Vaguine *et al.*, 1999). This program extracts the deposited *R* factor, resolution and model information, and then compares them with values calculated from coordinate and structure-factor files. It also calculates an overall *B* factor, coordinate errors, an effective resolution and completeness. The summary of the density correlation shift and *B* factor are reported for each residue. As specific procedures are developed for checking NMR structures against experimental data, they will be incorporated into the PDB validation procedures.

24.5.2.3. *NMR data*

The PDB staff recognize that NMR data need a special development effort. Historically these data have been retro-fitted into a PDB format defined around crystallographic information. As a first step towards improving this situation, the PDB carried out an extensive assessment of the current NMR holdings and presented the findings to a task force consisting of a cross section of NMR researchers. The PDB is working with this group, the BioMagResBank (BMRB; Ulrich *et al.*, 1989) and other members of the NMR community to develop an NMR data dictionary along with deposition and validation tools specific for NMR structures.

24.5.2.4. *Data-processing statistics*

Production processing of PDB entries by the RCSB began on 27 January 1999. As of 1 July 1999, when the RCSB became fully responsible for the PDB, approximately 80% of all structures submitted to the PDB are deposited *via ADIT* and processed by the RCSB. Another 20% are submitted *via AutoDep* to the European Bioinformatics Institute (EBI), who process these submissions and forward them to the PDB for archiving and distribution. The average time from deposition to the completion of data processing including author interactions is two weeks. The number of structures with a HOLD release status remains at about 20% of all submissions; 57% are held until publication (HPUB); and 23% are released immediately after processing.

Table 24.5.2.2 shows the breakdown of the types of structures in the PDB. As of 14 September 1999, the PDB contained 10 714 publicly accessible structures with another 1169 entries on hold (not shown). Of these, 8789 (82%) were determined by X-ray methods, 1692 (16%) were determined by NMR and 233 (2%) were theoretical models. Overall, 35% of the entries have deposited experimental data.

24.5.3. The PDB database resource

24.5.3.1. *The database architecture*

In recognition of the fact that no single architecture can fully express the information content of the PDB, an integrated system of heterogeneous databases and indices that store and organize the structural data has been created. At present there are five major components (Fig. 24.5.3.1):

(1) The core relational database managed by Sybase (Sybase Inc., 1995) provides the central physical storage for the primary experimental and coordinate data described in Table 24.5.2.1. The core PDB relational database contains all deposited information in a tabular form that can be accessed across any number of structures.

(2) The final curated data files (in PDB format) and data dictionaries are the archival data and are present as ASCII files in the ftp archive.

(3) The POM-based databases (Shindyalov & Bourne, 1997) consist of indexed objects containing native (*e.g.* atomic coordinates) and derived properties (*e.g.* calculated secondary-structure assignments and property profiles). Some properties require no derivation, for example, *B* factors; others must be derived, for example, exposure of each amino-acid residue (Lee & Richards, 1971) or Cα contact maps. Properties requiring significant computation time, such as structure neighbours (Shindyalov & Bourne, 1998), are pre-calculated when the database is incremented to save considerable user-access time.

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

(5) The Netscape LDAP server is used to index the textual content of the PDB in a structured format and provides support for keyword searches.

In the current implementation, communication among databases has been accomplished using the common gateway interface (CGI). An integrated web interface dispatches a query to the appropriate database(s), which then executes the query. Each database returns the PDB identifiers that satisfy the query, and the CGI program

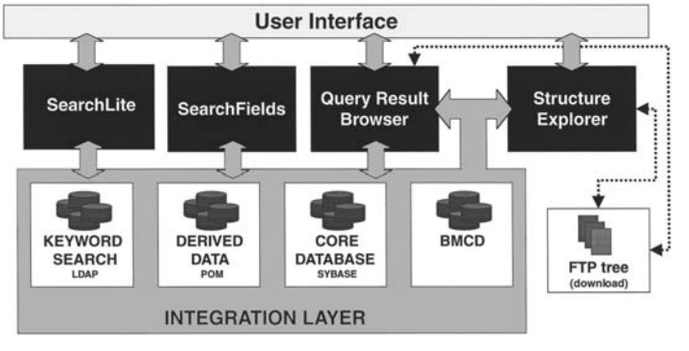

Fig. 24.5.3.1. The integrated query interface to the PDB.

Table 24.5.3.1. *Current query capabilities of the PDB*

(*a*) Query – single or iterative

Free text – any word in the PDB
Specific data items – compound name, author, description, deposition date, resolution, source, citation, cell dimensions, experimental method, data-collection method, refinement method, broad structure type, ligand (using the PDB HET records)
Property pattern – sequence, secondary structure
Structure similarity – 3D comparison

(*b*) Results analysis – single structure

Synopsis/Snapshot/Atlas – compound name, sequence, chemical components, citation, space group, cell constants, crystallization conditions, refinement details, structure views
Quick report – compound name, author, description, deposition date, resolution, source, citation, cell dimensions, experimental method, data-collection method, refinement method, geometry features
Full report – Quick report results plus secondary structure, chemical components, solvent
Property profiles – sequence, secondary structure
Links – see Table 24.5.3.2
Render – *RasMol*, *Chime*, *QuickPDB* (Java applet), VRML, *Protein Explorer*
Geometry – bond lengths, bond angles, dihedrals, close contacts, summary visual inspection

(*c*) Results analysis – multiple structure

Quick report – as above, but collated over multiple structures
Full report – as above, but collated over multiple structures
Structure neighbours – pairwise structure comparison

(*d*) Other query output options

mmCIF and PDB data files
Compressed files (gzip, tar, compressed)

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 CGI approach (and in the future a CORBA-based approach) will permit other databases to be integrated into this system, for example, those containing extended data on different protein families. The same approach could also be applied to include NMR data found in the BMRB or data found in other community databases.

24.5.3.2. *Database queries*

Three distinct query interfaces are available for querying data within the 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 24.5.3.1 summarizes the current query and analysis capabilities of the PDB. Fig. 24.5.3.2 illustrates how the various query options are organized.

SearchLite, which provides a single form field for keyword searches, was introduced in February 1999. All textual information within the PDB files as well as dates and some experimental data are accessible *via* simple or structured queries. *SearchFields*, accessible since May 1999, is a customizable query form that allows searching over many different data items, including compound, citation authors, sequence (*via* a *FASTA* search; Pearson & Lipman, 1988) and release or deposition dates.

Two user interfaces provide extensive information for results sets from *SearchLite* or *SearchFields* queries. The 'Query result browser' interface allows access to some general information, access to more detailed information in tabular format and the possibility of downloading whole sets of data files for result sets consisting of multiple PDB entries. The 'Structure explorer' interface provides information about individual structures as well as cross-links to many external resources for macromolecular structure data (Table 24.5.3.2). Both interfaces are accessible to other data resources through the simple CGI application programmer interface (API) described at http://www.rcsb.org/pdb/linking.html.

Table 24.5.3.3 indicates that usage has climbed dramatically since the system was first introduced in February 1999. Currently the PDB receives approximately 90000 web hits per day, or, on average, one query every second, seven days a week, 24 hours a day.

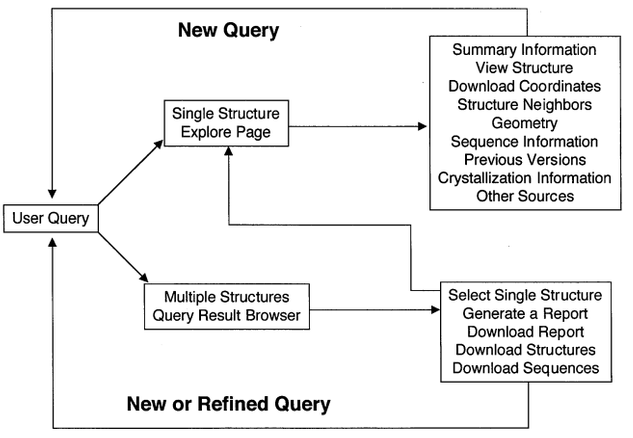

Fig. 24.5.3.2. The various query options that are available for the PDB.

Table 24.5.3.2. *Static cross-links to other data resources currently provided by the PDB*

Resource	Information content
3Dee (Siddiqui & Barton, 1996)	Structural domain definitions
BMCD (Gilliland, 1988)	Crystallization information about biomacromolecules
CATH (Orengo *et al.*, 1997)	Protein fold classification
CE (Shindyalov & Bourne, 1998)	Complete PDB and representative structure comparison and alignments
DSSP (Kabsch & Sander, 1983)	Secondary-structure classification
Enzyme Structures Database (Laskowski & Wallace, 1998)	Enzyme classifications and nomenclature
FSSP (Holm & Sander, 1998)	Structurally similar families
GRASS (Nayal *et al.*, 1999)	Graphical representation and analysis
HSSP (Dodge *et al.*, 1998)	Homology-derived secondary structures
Image (Sühnel, 1996)	Image library of biological macromolecules
MMDB (Hogue *et al.*, 1996)	Database of three-dimensional structures
MEDLINE (National Library of Medicine, 1989)	Direct access to MEDLINE at NCBI
NDB (Berman *et al.*, 1992)	Database of three-dimensional nucleic acid structures
PDBObs (Weissig *et al.*, 1998)	Obsolete structures database
PDBSum (Laskowski *et al.*, 1997)	Summary information about protein structures
SCOP (Murzin *et al.*, 1995)	Structure classifications
STING (Neshich *et al.*, 1998)	Simultaneous display of structural and sequence information
Tops (Westhead *et al.*, 1998)	Protein structure motif comparisons topological diagrams
VAST (Gibrat *et al.*, 1996)	Vector Alignment Search Tool (NCBI)
Whatcheck (Hooft *et al.*, 1996)	Protein structure checks

24.5.4. Data distribution

Data are distributed to the community in the following ways:

(1) From primary PDB web and ftp sites at UCSD, Rutgers and NIST that are updated weekly.

(2) From complete web-based mirror sites that contain all databases, data files, documentation and query interfaces, updated weekly.

(3) From ftp-only mirror sites that contain a complete or subset copy of data files, updated at intervals defined by the mirror site. The steps necessary to create an ftp-only mirror site are described at http://www.rcsb.org/pdb/ftpproc.final.html.

(4) Quarterly CD-ROM.

Data available for distribution include PDB files, mmCIF files, derived information, structure factors, NMR restraints, documentation, data dictionaries and software.

The RCSB has been responsible for distribution of PDB data since 3 February 1999. Data are distributed once a week. New data officially become available at 1 a.m. Pacific Standard Time each Wednesday. This follows the tradition developed by BNL and has minimized the impact of the transition on existing mirror sites. Since May 1999, two ftp archives have been provided: ftp://ftp.rcsb.org, a reorganized and more logical organization of all PDB data, software and documentation; and ftp://bnlarchive.rcsb.org, a near-identical copy of the original BNL archive which is maintained for purposes of backward compatibility. RCSB-style PDB mirrors have been established in Japan (Osaka University), Singapore (National University Hospital), Brazil (Universidade Federal de Minas Gerais Brazil) and in the UK (the Cambridge Crystallographic Data Centre). Plans call for operating mirrors in Australia, Canada, Germany and possibly India.

The first PDB CD-ROM distribution by the RCSB contained the coordinate files, experimental data, software and documentation as found in the PDB on 30 June 1999. Data are currently 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 presently no charge for this service.

24.5.5. Data archiving

The PDB is establishing a central master archiving facility. The master archive plan is based on five goals: reconstruction of the current archive in the case of a major disaster; duplication of the

Table 24.5.3.3. *Web query statistics for the primary RCSB site (www.rcsb.org)*

	Daily average		Monthly totals			
Month	Hits	Files	Sites	Kbytes	Files	Hits
August 1999	63768	47675	34928	31781561	1477927	1976818
July 1999	75693	54427	38698	35652864	1687265	2346495
June 1999	33256	27054	11586	11164410	622264	764894
May 1999	26890	22085	12405	12463441	684650	833597
April 1999	21140	17099	12261	9925351	512990	634224
March 1999	8406	6911	6292	3560629	214255	260610
February 1999	2944	2433	2246	844536	68133	82453
January 1999	1563	1353	1153	92014	35202	40641

Table 24.5.9.1. *PDB information sources*

Source	Information content
http://www.rcsb.org/pdb/ and http://www.pdb.org/	Main PDB web site
http://rutgers.rcsb.edu/pdb/ (Rutgers)	RCSB member institution PDB web sites
http://nist.rcsb.org/pdb/ (NIST)	
http://www.rcsb.org/pdb/mirrors.html	List of all RCSB PDB mirrors
http://pdb.rutgers.edu/adit/	*ADIT* web site (Rutgers)
http://pdbdep.protein.osaka-u.jp/adit/	*ADIT* web site (Osaka University, Japan)
http://pdb.rutgers.edu/validate/	*ADIT* validation server
http://www.rcsb.org/pdb/newsletter.html	RCSB PDB newsletter
http://www.rcsb.org/pdb/linking.html	Enzyme classifications and nomenclature
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
deposit@rcsb.rutgers.edu	Data processing correspondence

contents of the PDB as it existed on a specific date; preservation of software, derived data, ancillary data and all other computerized and printed information; automatic archiving of all depositions and the PDB production resource; and maintenance of the PDB correspondence archive that documents all aspects of deposition. During the transition period, all physical materials including electronic media and hard-copy materials were inventoried and stored, and are being catalogued.

24.5.6. Maintenance of the legacy of the BNL system

One of the goals of the PDB has been to provide a smooth transition from the system at BNL to the new system. Accordingly *AutoDep*, which was developed by BNL (Brookhaven National Laboratory, 1998) for data deposition, has been ported to the RCSB site and enables depositors to complete partial depositions as well as to make new depositions. In addition, the EBI accepts data using *AutoDep*. Similarly, the programs developed at BNL for data query and distribution (*PDBLite*, 3*DB Browser etc.*) are being maintained by the remaining BNL-style mirrors. The RCSB provides data in a form usable by these mirrors. Finally, the style and format of the BNL ftp archive is being maintained at ftp://bnlarchive.rcsb.org.

Links to the PDB at BNL were automatically redirected to the RCSB after BNL closed operations on 30 June 1999 using a network redirect implemented jointly by RCSB and BNL staff. External resources linking to the PDB are advised to change any URLs from http://www.pdb.bnl.gov to http://www.rcsb.org.

24.5.7. Current developments

An important role of the PDB is to foster new standards and technologies important to researchers and educators using macromolecular structure data. To this end, the following are under development at the PDB.

The RCSB is leading the Object Management Group Life Sciences Initiative's efforts to define a CORBA interface definition for the representation of macromolecular structure data. This is a standard developed under a strict procedure to ensure maximum input by members of various academic and industrial research communities. At this stage, proposals for the interface definition, including a working prototype that uses the standard, are being accepted. For further details refer to http://www.omg.org/cgi-bin/doc?lifesci/99-08-15. The finalized standard interface will facilitate the query and exchange of structural information not just at the level of complete structures, but at finer levels of detail.

As multimedia become more common, the opportunity exists to use them to deliver information on structure and function to a broad PDB user community *via* the web. To date we have developed prototype protein documentaries (Quinn, Taylor *et al.*, 1999) that explore these new media in describing structure–function relationships in proteins. It is also possible to develop educational materials that will run using a recent web browser (Quinn, Wang *et al.*, 1999).

Finally, it is recognized that structures exist both in the public and private domains. To this end we are planning on providing a subset of database tools for local use. Users will be able to load both public and proprietary data and use the same search and exploratory tools used at the PDB resources.

24.5.8. PDB advisory boards

The PDB has several advisory boards. Each member institution of the RCSB has its own local PDB Advisory Committee. Each institution is responsible for implementing the recommendations of those committees, as well as the recommendations of an international advisory board. Initially, the RCSB presented a report to the advisory board previously convened by BNL. At their recommendation, a new board has been assembled which contains previous members and new members. The goal was to have the board accurately reflect the depositor and user communities and thus include experts from many disciplines.

Serious issues of policy are referred to the major scientific societies, notably the International Union of Crystallography (IUCr). The goal is to make decisions based on input from a broad international community of experts. The IUCr maintains the mmCIF dictionary as the data standard upon which the PDB is built.

24.5.9. Further information

The PDB seeks to keep the community informed of new developments *via* weekly news updates to the web site, quarterly newsletters and an annual report. Users can request information at any time by sending an e-mail to info@rcsb.org. Finally, the pdb-l @rcsb.org listserver provides a community forum for the discussion of PDB-related issues. Changes to PDB operations that may affect the community, for example data-format changes, are posted here and users have 60 days to discuss the issue before changes are made

according to major consensus. Table 24.5.9.1 indicates how to access these resources.

24.5.10. Conclusion

These are exciting and challenging times to be responsible for the collection, curation and distribution of macromolecular structure data. Since the RCSB assumed responsibility for data deposition in February 1999, the number of depositions has averaged approximately 50 a week. However, with the advent of a number of structure genomics initiatives worldwide, this number is likely to increase. We estimate that the PDB, which at writing contains approximately 10 500 structures, could triple or quadruple in size over the next five years. This presents a challenge of timely distribution while maintaining high quality. The PDB's approach of using modern data-management practices should permit us to accommodate a large data influx.

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. Some of these efforts are acknowledged below. New input is constantly being sought and the PDB invites comments at any time by e-mail to info@rcsb.org.

Acknowledgements

The continuing support of Ken Breslauer (Rutgers), John Rumble (NIST) and Sid Karin (SDSC) is gratefully acknowledged. Current collaborators contributing to the future development of the PDB are the BioMagResBank, the Cambridge Crystallographic Data Centre, the HIV Protease Database Group, The Institute for Protein Research, Osaka University, The National Center for Biotechnology Information, the ReLiBase developers, the Swiss Institute for Bioinformatics/Glaxo and the European Bioinformatics Institute.

The cooperation of the BNL PDB staff is also gratefully acknowledged.

Parts of this chapter have appeared in *Nucleic Acids Research* (Berman *et al.*, 2000) and are reproduced here with permission of Oxford University Press.

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

24.1

Abola, E. E. (1994). *PDB-SHELL*. Available at ftp://pdb.bmc.uu.se/pub/databases/pdb/pdb_software/pdbshell/.

Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987). *Protein Data Bank*. In *Crystallographic databases – information content, software systems, scientific applications*, edited by F. H. Allen, G. Bergerhoff & R. Sievers, pp. 107–132. Bonn: International Union of Crystallography.

Abola, E. E., Sussman, J. L., Prilusky, J. & Manning, N. O. (1997). *Protein Data Bank archives of three-dimensional macromolecular structures*. *Methods Enzymol*. **277**, 556–571.

Bairoch, A. (1994). *The ENZYME data bank*. *Nucleic Acids Res*. **22**, 3626–3627.

Bairoch, A. & Boeckmann, B. (1994). *The SWISS-PROT protein sequence data bank: current status*. *Nucleic Acids Res.* **22**, 3578–3580.

Baker, E. N., Blundell, T. L., Vijayan, M., Dodson, E., Dodson, G., Gilliland, G. L. & Sussman, J. L. (1996). *Crystallographic data deposition*. *Nature (London)*, **379**, 202.

Bloom, F. E. (1998). *Policy change*. *Science*, **281**, 175.

Cambell, P. (1998). *New policy for structure data*. *Nature (London)*, **394**, 105.

Commission on Biological Macromolecules (2000). *Guidelines for the deposition and release of macromolecular coordinate and experimental data*. *Acta Cryst*. D**56**, 2.

Editorial Board (1998). *New policy on release of structural coordinates*. *Proc. Natl Acad. Sci. USA*, **95**, iii.

Jiang, J., Abola, E. & Sussman, J. L. (1999). *Deposition of structure factors at the Protein Data Bank*. *Acta Cryst.* D**55**, 4.

Kwong, P. D., Wyatt, R., Robinson, J., Sweet, R. W., Sodroski, J. & Hendrickson, W. A. (1998). *Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody*. *Nature (London)*, **393**, 648–659.

Lin, D., Manning, N. O., Jiang, J., Abola, E. E., Stampf, D., Prilusky, J. & Sussman, J. L. (2000). *AutoDep: a web-based system for deposition and validation of macromolecular structural information*. *Acta Cryst.* D**56**, 828–841.

Madden, D. R., Garboczi, D. N. & Wiley, D. C. (1993). *The antigenic identity of peptide–MHC complexes: a comparison of the conformations of five viral peptides presented by HLA-A2*. *Cell*, **75**, 693–708.

Peitsch, M. C., Stampf, D. R., Wells, T. N. C. & Sussman, J. L. (1995). *The Swiss 3D-image collection and Brookhaven Protein Data Bank browser on the World-Wide Web*. *Trends Biochem. Sci.* **20**, 82–84.

Phillips, D. C. (1971). *Protein crystallography 1971: coming of age*. *Cold Spring Harbor Symp. Quant. Biol.* pp. 589–592.

Rizzuto, C. D., Wyatt, R., Hernandez-Ramos, N., Sun, Y., Kwong, P. D., Hendrickson, W. A. & Sodroski, J. (1998). *A conserved HIV gp120 glycoprotein structure involved in chemokine receptor binding*. *Science*, **280**, 1949–1953.

Sayle, R. A. & Milner-White, E. J. (1995). *RASMOL: biomolecular graphics for all*. *Trends Biochem. Sci.* **20**, 374–376.

Schultz, S. C., Shields, G. C. & Steitz, T. A. (1991). *Crystal structure of a CAP–DNA complex: the DNA is bent by 90 degrees*. *Science*, **253**, 1001–1007.

Seavey, B. R., Farr, E. A., Westler, W. M. & Markley, J. L. (1991). *A relational database for sequence-specific protein NMR data*. *J. Biomol. Nucl. Magn. Reson.* **1**, 217–236.

Stampf, D. R., Felder, C. E. & Sussman, J. L. (1995). *PDBBrowser – a graphics interface to the Brookhaven Protein Data Bank*. *Nature (London)*, **374**, 572–574.

Sussman, J. L. (1997). *Bridging the gap*. *Nature Struct. Biol.* **4**, 517.

Sussman, J. L. (1998). *Protein Data Bank deposits*. *Science*, **282**, 1991.

Sussman, J. L., Lin, D., Jiang, J., Manning, N. O., Prilusky, J., Ritter, O. & Abola, E. E. (1998). *Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules*. *Acta Cryst.* D**54**, 1078–1084.

24.2

Allen, F. H., Bellard, S., Brice, M. D., Cartwright, B. A., Doubleday, A., Higgs, H., Hummelink, T., Hummelink-Peters, B. G., Kennard, O., Motherwell, W. D. S., Rodgers, J. R. & Watson, D. G. (1979). *The Cambridge Crystallographic Data Centre: computer-based search, retrieval, analysis and display of information*. *Acta Cryst.* B**35**, 2331–2339.

Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S. H., Srinivasan, A. R. & Schneider, B. (1992). *The Nucleic Acid Database – a comprehensive relational database of three-dimensional structures of nucleic acids*. *Biophys. J.* **63**, 751–759.

24.2 (*cont.*)

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

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. E., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *Protein Data Bank: a computer-based archival file for macromolecular structures*. *J. Mol. Biol.* **112**, 535–542.

Bourne, P., Berman, H. M., Watenpaugh, K., Westbrook, J. D. & Fitzgerald, P. M. D. (1997). *The macromolecular Crystallographic Information File (mmCIF)*. *Methods Enzymol.* **277**, 571–590.

Brünger, A. T. (1992). *X-PLOR*. Version 3.1. *A system for X-ray crystallography and NMR*. Yale University Press, New Haven, CT, USA.

Clowney, L., Jain, S. C., Srinivasan, A. R., Westbrook, J., Olson, W. K. & Berman, H. M. (1996). *Geometric parameters in nucleic acids: nitrogenous bases*. *J. Am. Chem. Soc.* **118**, 509–518.

Feng, Z., Hsieh, S.-H., Gelbin, A. & Westbrook, J. (1998). *MAXIT: macromolecular exchange and input tool*. NDB-120. Rutgers University, New Brunswick, NJ, USA.

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

Gelbin, A., Schneider, B., Clowney, L., Hsieh, S.-H., Olson, W. K. & Berman, H. M. (1996). *Geometric parameters in nucleic acids: sugar and phosphate constituents*. *J. Am. Chem. Soc.* **118**, 519–528.

Grzeskowiak, K., Yanagi, K., Privé, G. G. & Dickerson, R. E. (1991). *The structure of B-helical C-G-A-T-C-G-A-T-C-G and comparison with C-C-A-A-C-G-T-T-G-G: the effect of base pair reversal*. *J. Biol. Chem.* **266**, 8861–8883.

Lavery, R. & Sklenar, H. (1989). *Defining the structure of irregular nucleic acids: conventions and principles*. *J. Biomol. Struct. Dyn.* **6**, 655–667.

Parkinson, G., Vojtechovsky, J., Clowney, L., Brünger, A. T. & Berman, H. M. (1996). *New parameters for the refinement of nucleic acid-containing structures*. *Acta Cryst.* D**52**, 57–64.

Sayle, R. & Milner-White, E. J. (1995). *RasMol: biomolecular graphics for all*. *Trends Biochem. Sci.* **20**, 374.

Schneider, B., Neidle, S. & Berman, H. M. (1997). *Conformations of the sugar–phosphate backbone in helical DNA crystal structures*. *Biopolymers*, **42**, 113–124.

Scott, W. G., Finch, J. T. & Klug, A. (1995). *The crystal structure of an all-RNA hammerhead ribozyme: a proposed mechanism for RNA catalytic cleavage*. *Cell*, **81**, 991–1002.

Vaguine, A. A., Richelle, J. & Wodak, S. J. (1999). *SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model*. *Acta Cryst.* D**55**, 191–205.

Westbrook, J. (1998). *AutoDep input tool*. NDB-406. Rutgers University, New Brunswick, NJ, USA.

24.3

Abola, E. E., Sussman, J. L., Prilusky, J. & Manning, N. O. (1997). *Protein Data Bank archives of three-dimensional macromolecular structures. Methods Enzymol.* **277**, 556–571.

Allen, F. H., Davies, J. E., Galloy, J. J., Johnson, O., Kennard, O., Macrae, C. F., Mitchell, E. M., Mitchell, G. F., Smith, J. M. & Watson, D. G. (1991). *The development of versions 3 and 4 of the Cambridge Structural Database system. J. Chem. Inf. Comput. Sci.* **31**, 187–204.

Allen, F. H., Kennard, O., Watson, D. G., Brammer, L., Orpen, A. G. & Taylor, R. (1987). *Tables of bond lengths determined by X-ray and neutron diffraction. Part 1. Bond lengths in organic compounds. J. Chem. Soc. Perkin Trans.* 2, pp. S1–S19.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1972). *The Protein Data Bank: a computer-based archival file for macromolecular structures. J. Mol. Biol.* **112**, 535–547.

Bruno, I. J., Cole, J. C., Lommerse, J. P. M., Rowland, R. S., Taylor, R. & Verdonk, M. L. (1997). *IsoStar: a library of information about nonbonded interactions. J. Comput.-Aided Mol. Des.* **11**, 525–537.

Hall, S. R., Allen, F. H. & Brown, I. D. (1991). *The crystallographic information file (CIF): a new standard archive file for crystallography. Acta Cryst.* A**47**, 655–685.

Hayes, I. C. & Stone, A. J. (1984). *An intermolecular perturbation theory for the region of moderate overlap. J. Mol. Phys.* **53**, 83–105.

Kennard, O. & Allen, F. H. (1993). *3D search and research using the Cambridge Structural Database. Chem. Des. Autom. News*, **8**, 1, 31–37.

RCSB (2000). The Protein Data Bank. Research Collaboratory for Structural Bioinformatics, Department of Chemistry, Rutgers University, Piscataway, NJ, USA (http://www.rcsb.org).

Sayle, R. (1996). *The RASMOL visualiser*. Glaxo Wellcome Research, Stevenage, Hertfordshire, England.

24.4

Ariyoshi, M., Vassylyev, D. G., Iwasaki, H., Fujishima, A., Shinagawa, H. & Morikawa, K. (1994). *Preliminary crystallographic study of Escherichia coli RuvC protein. An endonuclease specific for Holliday junctions. J. Mol. Biol.* **241**, 281–282.

Athanasiadis, A. & Kokkinidis, M. (1991). *Purification, crystallization and preliminary X-ray diffraction studies of the PvuII endonuclease. J. Mol. Biol.* **222**, 451–453.

Balendiran, K., Bonventre, J., Knott, R., Jack, W., Benner, J., Schildkraut, I. & Anderson, J. E. (1994). *Expression, purification, and crystallization of restriction endonuclease PvuII with DNA containing its recognition site. Proteins*, **19**, 77–79.

Bannikova, G. E., Blagova, E. V., Dementiev, A. A., Morgunova, E. Yu., Mikchailov, A. M., Shlyapnikov, S. V., Varlamov, V. P. & Vainshtein, B. K. (1991). *Two isoforms of Serratia marcescens nuclease. Crystallization and preliminary X-ray investigation of the enzyme. Biochem. Int.* **24**, 813–822.

Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S.-H., Srinivasan, A. R. & Schneider, B. (1992). *The Nucleic Acid Database: a comprehensive relational database of three-dimensional structures of nucleic acids. Biophys. J.* **63**, 751–759.

Blundell, T. L. & Johnson, L. N. (1976). *Protein crystallography*. New York: Academic Press.

Bozic, D., Grazulis, S., Siksnys, V. & Huber, R. (1996). *Crystal structure of Citrobacter freundii restriction endonuclease Cfr10I at 2.15 Å resolution. J. Mol. Biol.* **255**, 176–186.

Carter, C. W. Jr & Carter, C. W. (1979). *Protein crystallization using incomplete factorial experiments. J. Biol. Chem.* **254**, 12219–12223.

Cudney, B., Patel, S., Weisgraber, K., Newhouse, Y. & McPherson, A. (1994). *Screening and optimization strategies for macromolecular crystal growth. Acta Cryst.* D**50**, 414–423.

D'Arcy, A., Brown, R. S., Zabeau, M., van Resandt, R. W. & Winkler, F. K. (1985). *Purification and crystallization of the EcoRV restriction endonuclease. J. Biol. Chem.* **260**, 1987–1990.

Gilliland, G. L. (1988). *A biological macromolecule crystallization database: a basis for a crystallization strategy. J. Cryst. Growth*, **90**, 51–59.

Gilliland, G. L. & Bickham, D. (1990). *The Biological Macromolecule Crystallization Database: a tool to assist the development of crystallization strategies. Methods Companion Methods Enzymol.* **1**, 6–11.

Gilliland, G. L. & Davies, D. R. (1984). *Protein crystallization: the growth of large-scale single crystals. Methods Enzymol.* **104**, 370–381.

Gilliland, G. L., Tung, M., Blakeslee, D. M. & Ladner, J. E. (1994). *Biological Macromolecule Crystallization Database, version 3.0: new features, data and the NASA archive for protein crystal growth data. Acta Cryst.* D**50**, 408–413.

Gilliland, G. L., Tung, M. & Ladner, J. (1996). *The Biological Macromolecule Crystallization Database and NASA Protein*

24.4 (*cont.*)

Crystal Growth Archive. J. Res. Natl Inst. Stand. Technol. **101**, 309–320.

Hirsch, J. A., Wah, D. A., Dorner, L. F., Schildkraut, I. & Aggarwal, A. K. (1997). *Crystallization and preliminary X-ray analysis of restriction endonuclease FokI bound to DNA. FEBS Lett.* **403**, 136–138.

Jancarik, J. & Kim, S.-H. (1991). *Sparse matrix sampling: a screening method for crystallization of proteins. J. Appl. Cryst.* **24**, 409–411.

Ji, X., Johnson, W. W., Sesay, M. A., Dickert, L., Prasad, S. M., Ammon, H. L., Armstrong, R. N. & Gilliland, G. L. (1994). *Structure of the xenobiotic substrate binding site of a glutathione S-transferase as revealed by X-ray crystallographic analysis of product complexes with the diastereomers of 9-(S-glutathionyl)-10-hydroxy-9,10-dihydrophenanthrene. Biochemistry*, **33**, 1043–1052.

Kostrewa, D. & Winkler, F. K. (1995). *Mg^{2+} binding to the active site of EcoRV endonuclease: a crystallographic study of complexes with substrate and product DNA at 2 angstroms resolution. Biochemistry*, **34**, 683–696.

Kuo, C. F., McRee, D. E., Cunningham, R. P. & Tainer, J. A. (1992). *Crystallization and crystallographic characterization of the iron–sulfur-containing DNA-repair enzyme endonuclease III from Escherichia coli. J. Mol. Biol.* **227**, 347–351.

Kuo, C. F., McRee, D. E., Fisher, C. L., O'Handley, S. F., Cunningham, R. P. & Tainer, J. A. (1992). *Atomic structure of the DNA repair [4Fe–4S] enzyme endonuclease III. Science*, **258**, 434–440.

McPherson, A. (1982). *Preparation and analysis of protein crystals*. New York: Wiley.

McPherson, A. (1999). *Crystallization of biological macromolecules*. New York: Cold Spring Harbor Laboratory Press.

McPherson, A. Jr (1976). *The growth and preliminary investigation of protein and nucleic acid crystals for X-ray diffraction analysis. Methods Biochem. Anal.* **23**, 249–345.

Miller, M. D., Benedik, M. J., Sullivan, M. C., Shipley, N. S. & Krause, K. L. (1991). *Crystallization and preliminary crystallographic analysis of a novel nuclease from Serratia marcescens. J. Mol. Biol.* **222**, 27–30.

Morikawa, K., Ariyoshi, M., Vassylyev, D. G., Matsumoto, O., Katayanagi, K. & Ohtsuka, E. (1995). *Crystal structure of a pyrimidine dimer-specific excision repair enzyme from bacteriophage T4: refinement at 1.45 Å and X-ray analysis of the three active site mutants. J. Mol. Biol.* **249**, 360–375.

Morikawa, K., Matsumoto, O., Tsujimoto, M., Katayanagi, K., Ariyoshi, M., Doi, T., Ikehara, M., Inaoka, T. & Ohtsuka, E. (1992). *X-ray structure of T4 endonuclease V: an excision repair enzyme specific for a pyrimidine dimer. Science*, **256**, 523–526.

Morikawa, K., Tsujimoto, M., Ikehara, M., Inaoka, T. & Ohtsuka, E. (1988). *Preliminary crystallographic study of pyrimidine dimer-specific excision-repair enzyme from bacteriophage T4. J. Mol. Biol.* **202**, 683–684.

Newman, M., Strzelecka, T., Dorner, L. F., Schildkraut, I. & Aggarwal, A. K. (1994). *Structure of restriction endonuclease BamhI phased at 1.95 Å resolution by MAD analysis. Structure*, **2**, 439–452.

Scott, W. G., Finch, J. T., Grenfell, R., Fogg, J., Smith, T., Gait, M. J. & Klug, A. (1995). *Rapid crystallization of chemically synthesized hammerhead RNAs using a double screening procedure. J. Mol. Biol.* **250**, 327–332.

Sesay, M. A., Ammon, H. L. & Armstrong, R. N. (1987). *Crystallization and a preliminary X-ray diffraction study of isozyme 3–3 of glutathione S-transferase from rat liver. J. Mol. Biol.* **197**, 377–378.

Strzelecka, T., Newman, M., Dorner, L. F., Knott, R., Schildkraut, I. & Aggarwal, A. K. (1994). *Crystallization and preliminary X-ray analysis of restriction endonuclease BamHI–DNA complex. J. Mol. Biol.* **239**, 430–432.

Wah, D. A., Hirsch, J. A., Dorner, L. F., Schildkraut, I. & Aggarwal, A. K. (1997). *Structure of the multimodular endonuclease FokI bound to DNA. Nature (London)*, **388**, 97–100.

Winkler, F. K., Banner, D. W., Oefner, C., Tsernoglou, D., Brown, R. S., Heathman, S. P., Bryan, R. K., Martin, P. D., Petratos, K. & Wilson, K. S. (1993). *The crystal structure of EcoRV endonuclease and of its complexes with cognate and non-cognate DNA fragments. EMBO J.* **12**, 1781–1795.

Winkler, F. K., D'Arcy, A., Blocker, H., Frank, R. & van Boom, J. H. (1991). *Crystallization of complexes of EcoRv endonuclease with cognate and non-cognate DNA fragments. J. Mol. Biol.* **217**, 235–238.

24.5

Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S. H., Srinivasan, A. R. & Schneider, B. (1992). *The Nucleic Acid Database – a comprehensive relational database of three-dimensional structures of nucleic acids. 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). *The Protein Data Bank. Nucleic Acids Res.* **28**, 235–242.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. E., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *Protein Data Bank: a computer-based archival file for macromolecular structures. J. Mol. Biol.* **112**, 535–542.

Bourne, P., Berman, H. M., Watenpaugh, K., Westbrook, J. D. & Fitzgerald, P. M. D. (1997). *The macromolecular Crystallographic Information File (mmCIF). Methods Enzymol.* **277**, 571–590.

Brookhaven National Laboratory (1998). *AutoDep.* Version 2.1. Brookhaven National Laboratory, Upton, NY, USA.

Dodge, C., Schneider, R. & Sander, C. (1998). *The HSSP database of protein structure-sequence alignments and family profiles. Nucleic Acids Res.* **26**, 313–315.

Feng, Z., Hsieh, S.-H., Gelbin, A. & Westbrook, J. (1998). *MAXIT: macromolecular exchange and input tool.* NDB-120. Rutgers University, New Brunswick, NJ, USA.

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

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

Gilliland, G. L. (1988). *A Biological Macromolecule Crystallization Database: a basis for a crystallization strategy. J. Cryst. Growth*, **90**, 51–59.

Hogue, C. W., Ohkawa, H. & Bryant, S. H. (1996). *A dynamic look at structures: WWW-Entrez and the Molecular Modeling Database. Trends Biochem. Sci.* **21**, 226–229.

Holm, L. & Sander, C. (1998). *Touring protein fold space with Dali/FSSP. Nucleic Acids Res.* **26**, 316–319.

Hooft, R. W. W., Sander, C. & Vriend, G. (1996). *Verification of protein structures: side-chain planarity. J. Appl. Cryst.* **29**, 714–716.

Kabsch, W. & Sander, C. (1983). *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers*, **22**, 2577–2637.

Laskowski, R. A., Hutchinson, E. G., Michie, A. D., Wallace, A. C., Jones, M. L. & Thornton, J. M. (1997). *PDBsum: a web-based database of summaries and analyses of all PDB structures. Trends Biochem. Sci.* **22**, 488–490.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Cryst.* **26**, 283–291.

Laskowski, R. A. & Wallace, A. C. (1998). *Enzyme Structures Database.* http://www.biochem.ucl.ac.uk/bsm/enzymes/.

Lee, B. & Richards, F. M. (1971). *The interpretation of protein structures: estimation of static accessibility. J. Mol. Biol.* **55**, 379–400.

Markley, J. L., Bax, A., Arata, Y., Hilbers, C. W., Kaptein, R., Sykes, B. D., Wright, P. E. & Wüthrich, K. (1998). *Recommendations for the presentation of NMR structures of proteins and nucleic acids.*

24.5 (*cont.*)

IUPAC–IUBMB–IUPAB Inter-Union Task Group on the standardization of data bases of protein and nucleic acid structures determined by NMR spectroscopy. J. Biomol. Nucl. Magn. Reson. **12**, 1–23.

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). *SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol.* **247**, 536–540.

National Library of Medicine (1989). *MEDLINE* [database online]. Bethesday, MD, USA. Updated weekly. Available from: National Library of Medicine; OVID, Murray, UT; The Dialog Corporation, Palo Alto, CA.

Nayal, M., Hitz, B. C. & Honig, B. (1999). *GRASS: a server for the graphical representation and analysis of structures. Protein Sci.* **8**, 676–679.

Neshich, G., Togawa, R., Vilella, W. & Honig, B. (1998). *STING (Sequence To and withIN Graphics) PDB_Viewer. Protein Data Bank Q. Newsl.* **85**, 6–7.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). *CATH – a hierarchic classification of protein domain structures. Structure*, **5**, 1093–1108.

Pearson, W. R. & Lipman, D. J. (1988). *Improved tools for biological sequence comparison. Proc. Natl Acad. Sci. USA*, **24**, 2444–2448.

Quinn, G., Taylor, A., Wang, H.-P. & Bourne, P. E. (1999). *Development of internet-based multimedia applications. Trends Biochem. Sci.* **24**, 321–324.

Quinn, G., Wang, H.-P., Martinez, D. & Bourne, P. E. (1999). *Developing protein documentaries and other multimedia presentations for molecular biology*. In *Pacific symposium on biocomputing*, edited by R. Altman, K. Dunker, L. Hunter, T. Klein & K. Lauderdale, pp. 380–391. Singapore.

Shindyalov, I. N. & Bourne, P. E. (1997). *Protein data representation and query using optimized data decomposition*. *Comput. Appl. Biosci.* **13**, 487–496.

Shindyalov, I. N. & Bourne, P. E. (1998). *Protein structure alignment by incremental combinatorial extension of the optimum path. Protein Eng.* **11**, 739–747.

Siddiqui, A. & Barton, G. (1996). *Perspectives on protein engineering 1996*, Vol. 2, CD-ROM edition, edited by M. J. Geisow. BIODIGM Ltd (UK). ISBN 0-9529015-0-1.

Sühnel, J. (1996). *Image library of biological macromolecules. Comput. Appl. Biosci.* **12**, 227–229.

Sybase Inc. (1995). 70202-01-1100-01 SYBASE SQL server release 11.0. Emeryville, CA, USA.

Ulrich, E. L., Markley, J. L. & Kyogoku, Y. (1989). *Creation of a nuclear magnetic resonance data repository and literature database. Protein Seq. Data Anal.* **2**, 23–37.

Vaguine, A. A., Richelle, J. & Wodak, S. J. (1999). *SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. Acta Cryst.* D**55**, 191–205.

Weissig, H., Shindyalov, I. N. & Bourne, P. E. (1998). *Macromolecular structure databases: past progress and future challenges. Acta Cryst.* D**54**, 1085–1094.

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

Westhead, D., Slidel, T., Flores, T. & Thornton, J. (1998). *Protein structural topology: automated analysis and diagrammatic representation. Protein Sci.* **8**, 897–904.