(*c*) AUDIT_LINK
- **_audit_link.block_code**
- **_audit_link.block_description**

The bullet (•) indicates a category key. Where multiple items within a category are marked with a bullet, they must be taken together to form a compound key. The arrow (→) is a reference to a parent data item. Items in italics have aliases in the core CIF dictionary formed by changing the full stop (.) to an underscore (_).

The sole data item in the category ENTRY, **_entry.id**, is a label that identifies the current data block. This label is used as the formal key in several categories that record information that is relevant to the entire data block (*e.g.* **_cell.entry_id**, **_geom.entry_id**), so care should be taken to select a label that is informative and unique.

Data items in the ENTRY_LINK category record the relationships between the current data block and other data blocks within the current file which may be referenced in the current data block. Since there are no formal constraints on the value of **_entry.id** assigned to each data block, authors must take care to ensure that an mmCIF comprised of several distinct data blocks uses a different value for **_entry.id** in each block.

As mentioned in the introductory paragraph of Section 3.6.9, the ENTRY_LINK category is used in mmCIF applications instead of the core category AUDIT_LINK. The latter is retained formally in the mmCIF dictionary for strict compatibility with the core dictionary, and the data items in this category, **_audit_link.blockcode** and **_audit_link.block_description**, are aliased to corresponding core data names (see Section 3.2.6.1). Their use is not recommended in mmCIF applications.

3.6.9.3. Other category classifications

The following categories, already described elsewhere in this chapter, are included in other formal category groups:

Compliance with earlier dictionaries

COMPLIANCE group
- DATABASE

Compatibility with PDB format files

PDB group
- DATABASE_PDB_CAVEAT
- DATABASE_PDB_MATRIX
- DATABASE_PDB_REMARK
- DATABASE_PDB_REV
- DATABASE_PDB_REV_RECORD
- DATABASE_PDB_TVECT

The COMPLIANCE group includes categories that appear in the mmCIF dictionary for the sole purpose of ensuring compliance with earlier dictionaries. They are not intended for use in the creation of new mmCIFs. As was discussed in Section 3.6.8.3, the DATABASE category of the core CIF is replaced in mmCIF by the more structured DATABASE_2 category. Thus the core CIF DATABASE category appears in the mmCIF COMPLIANCE group. At the time of writing (2005), DATABASE is the only category in the COMPLIANCE group.

The PDB group includes a number of categories that record unstructured information imported from various records in Protein Data Bank (PDB) format files. These categories are also part of the DATABASE group and were discussed in Section 3.6.8.3.2.

Appendix 3.6.1
Category structure of the mmCIF dictionary

Table A3.6.1.1 provides an overview of the structure of the mmCIF dictionary by category group and member categories.

Appendix 3.6.2
The Protein Data Bank exchange data dictionary

BY J. D. WESTBROOK, K. HENRICK, E. L. ULRICH AND H. M. BERMAN

In developing a data-management infrastructure, the Protein Data Bank (PDB; Berman *et al.*, 2000) has chosen the mmCIF dictionary technology for describing the data that it collects and disseminates. To accommodate the growth in the PDB's activities, data collection, processing and annotation now occur at three sites worldwide: the Research Collaboratory for Structural Bioinformatics (RCSB/PDB), the Macromolecular Structural Database (MSD) at the European Bioinformatics Institute (EBI) and the Protein Data Bank Japan (PDBj) at Osaka. Together these facilities form the Worldwide PDB (wwPDB) (Berman *et al.*, 2003). In order to maintain the fidelity of the single archive of three-dimensional macromolecular structure, a precise content description is required to support the accurate exchange of data among the different sites and the exchange of information between different file formats.

A key strength of the mmCIF technology is the extensibility afforded by a framework based on a software-accessible data dictionary. The PDB has exploited this functionality by using the mmCIF dictionary as a foundation and supplementing it with extensions in order to describe all aspects of data processing and database operations.

These extensions include content required to support reversible format translation, noncrystallographic structure determination methods and the details of protein production. They also support recommendations by the International Union of Crystallography (IUCr) and the International Structural Genomics Organization (ISGO) as to which data should be deposited. In the following sections, the extensions to the mmCIF data dictionary developed by the PDB (http://mmcif.pdb.org/) are described.

A3.6.2.1. Data exchange and format translation

The majority of crystallographic and structural concepts embodied in the PDB are already well described in the mmCIF data dictionary. However, while there is a conceptual description of most crystallographic information in PDB-format files within the mmCIF dictionary, the precise representation of this information can differ subtly. To guarantee accurate data exchange and to facilitate reversible format translation between PDB and mmCIF formats, all such differences in representation must be resolved.

To accommodate content and semantic differences between formats, extensions to the dictionary have been created. These extensions take one of two forms: the addition of new definitions to existing categories or the creation of new categories. Where possible, extensions are added to existing categories. This is done when the new definition supplements the content of the category without changing the category definition or its fundamental organization. However, if a new definition cannot be added to an existing category, a new category is created to hold the extension. All new data items and categories include the prefix **pdbx_** in their names.

For example, the level of detail in the PDB description of the biological source exceeds the description provided by mmCIF. In this case, dictionary extensions have been added to the existing categories ENTITY_SRC_NAT and ENTITY_SRC_GEN (where 'nat' and 'gen' stand for naturally occurring and genetically engineered, respectively). The PDB description of atomic coordinates includes two items that are not described in mmCIF: the insertion code

Table A3.6.1.1. *Categories in the mmCIF dictionary*

Numbers in parentheses refer to the section of this chapter in which each category is described in detail.

ATOM group (§3.6.7.1)
- ATOM_SITE (§3.6.7.1.1(*a*))
- ATOM_SITE_ANISOTROP (§3.6.7.1.1(*b*))
- ATOM_SITES (§3.6.7.1.2(*a*))
- ATOM_SITES_ALT (§3.6.7.1.4(*a*))
- ATOM_SITES_ALT_ENS (§3.6.7.1.4(*b*))
- ATOM_SITES_ALT_GEN (§3.6.7.1.4(*c*))
- ATOM_SITES_FOOTNOTE (§3.6.7.1.2(*b*))
- ATOM_TYPE (§3.6.7.1.3)

AUDIT group (§3.6.9.1)
- AUDIT (§3.6.9.1(*a*))
- AUDIT_AUTHOR (§3.6.9.1(*b*))
- AUDIT_CONFORM (§3.6.9.1(*c*))
- AUDIT_CONTACT_AUTHOR (§3.6.9.1(*d*))
- AUDIT_LINK (§3.6.9.2(*c*))

CELL group (§3.6.5.1)
- CELL (§3.6.5.1(*a*))
- CELL_MEASUREMENT (§3.6.5.1(*b*))
- CELL_MEASUREMENT_REFLN (§3.6.5.1(*c*))

CHEM_COMP group (§3.6.7.2.2)
- CHEM_COMP (§3.6.7.2.2(*a*))
- CHEM_COMP_ANGLE (§3.6.7.2.2(*b*))
- CHEM_COMP_ATOM (§3.6.7.2.2(*c*))
- CHEM_COMP_BOND (§3.6.7.2.2(*d*))
- CHEM_COMP_CHIR (§3.6.7.2.2(*e*))
- CHEM_COMP_CHIR_ATOM (§3.6.7.2.2(*f*))
- *CHEM_COMP_LINK* (*see CHEM_LINK group*)
- CHEM_COMP_PLANE (§3.6.7.2.2(*h*))
- CHEM_COMP_PLANE_ATOM (§3.6.7.2.2(*i*))
- CHEM_COMP_TOR (§3.6.7.2.2(*j*))
- CHEM_COMP_TOR_VALUE (§3.6.7.2.2(*k*))

CHEM_LINK group (§3.6.7.2.3)
- CHEM_COMP_LINK (§3.6.7.2.2(*g*))
- CHEM_LINK (§3.6.7.2.3(*a*))
- CHEM_LINK_ANGLE (§3.6.7.2.3(*b*))
- CHEM_LINK_BOND (§3.6.7.2.3(*c*))
- CHEM_LINK_CHIR (§3.6.7.2.3(*d*))
- CHEM_LINK_CHIR_ATOM (§3.6.7.2.3(*e*))
- CHEM_LINK_PLANE (§3.6.7.2.3(*f*))
- CHEM_LINK_PLANE_ATOM (§3.6.7.2.3(*g*))
- CHEM_LINK_TOR (§3.6.7.2.3(*h*))
- CHEM_LINK_TOR_VALUE (§3.6.7.2.3(*i*))
- ENTITY_LINK (§3.6.7.2.3(*j*))

CHEMICAL group (§3.6.7.2)
- CHEMICAL (§3.6.7.2.1(*a*))
- CHEMICAL_CONN_ATOM (§3.6.7.2.1(*b*))
- CHEMICAL_CONN_BOND (§3.6.7.2.1(*c*))
- CHEMICAL_FORMULA (§3.6.7.2.1(*d*))

CITATION group (§3.6.8.1)
- CITATION (§3.6.8.1(*a*))
- CITATION_AUTHOR (§3.6.8.1(*b*))
- CITATION_EDITOR (§3.6.8.1(*c*))

COMPUTING group (§3.6.8.2)
- COMPUTING (§3.6.8.2(*a*))
- SOFTWARE (§3.6.8.2(*b*))

DATABASE group (§§3.6.8.3, 3.6.9.3)
- DATABASE (§3.6.8.3.1(*a*))
- DATABASE_2 (§3.6.8.3.1(*b*))
- *The following also belong to the PDB group*
- DATABASE_PDB_CAVEAT (§3.6.8.3.2(*e*))
- DATABASE_PDB_MATRIX (§3.6.8.3.2(*c*))
- DATABASE_PDB_REMARK (§3.6.8.3.2(*f*))
- DATABASE_PDB_REV (§3.6.8.3.2(*a*))
- DATABASE_PDB_REV_RECORD (§3.6.8.3.2(*b*))
- DATABASE_PDB_TVECT (§3.6.8.3.2(*d*))

DIFFRN group (§3.6.5.2)
- DIFFRN (§3.6.5.2(*a*))
- DIFFRN_ATTENUATOR (§3.6.5.2(*b*))
- DIFFRN_DETECTOR (§3.6.5.2(*c*))
- DIFFRN_MEASUREMENT (§3.6.5.2(*d*))
- DIFFRN_ORIENT_MATRIX (§3.6.5.2(*e*))
- DIFFRN_ORIENT_REFLN (§3.6.5.2(*f*))
- DIFFRN_RADIATION (§3.6.5.2(*g*))
- DIFFRN_RADIATION_WAVELENGTH (§3.6.5.2(*h*))
- DIFFRN_REFLN (§3.6.5.2(*i*))
- DIFFRN_REFLNS (§3.6.5.2(*j*))
- DIFFRN_REFLNS_CLASS (§3.6.5.2(*k*))
- DIFFRN_SCALE_GROUP (§3.6.5.2(*l*))
- DIFFRN_SOURCE (§3.6.5.2(*m*))
- DIFFRN_STANDARD_REFLN (§3.6.5.2(*n*))
- DIFFRN_STANDARDS (§3.6.5.2(*o*))

ENTITY group (§3.6.7.3)
- ENTITY (§3.6.7.3.1(*a*))
- ENTITY_KEYWORDS (§3.6.7.3.1(*b*))
- *ENTITY_LINK* (*see CHEM_LINK group*)
- ENTITY_NAME_COM (§3.6.7.3.1(*c*))
- ENTITY_NAME_SYS (§3.6.7.3.1(*d*))
- ENTITY_POLY (§3.6.7.3.2(*a*))
- ENTITY_POLY_SEQ (§3.6.7.3.2(*b*))
- ENTITY_SRC_GEN (§3.6.7.3.1(*e*))
- ENTITY_SRC_NAT (§3.6.7.3.1(*f*))

ENTRY group (§3.6.9.2)
- ENTRY (§3.6.9.2(*a*))
- ENTRY_LINK (§3.6.9.2(*b*))

EXPTL group (§3.6.5.3)
- EXPTL (§3.6.5.3.1(*a*))
- EXPTL_CRYSTAL (§3.6.5.3.1(*b*))
- EXPTL_CRYSTAL_FACE (§3.6.5.3.1(*c*))
- EXPTL_CRYSTAL_GROW (§3.6.5.3.2(*a*))
- EXPTL_CRYSTAL_GROW_COMP (§3.6.5.3.2(*b*))

GEOM group (§3.6.7.4)
- GEOM (§3.6.7.4(*a*))
- GEOM_ANGLE (§3.6.7.4(*b*))
- GEOM_BOND (§3.6.7.4(*c*))
- GEOM_CONTACT (§3.6.7.4(*d*))
- GEOM_HBOND (§3.6.7.4(*e*))
- GEOM_TORSION (§3.6.7.4(*f*))

IUCR group (§3.6.8.4)
- JOURNAL (§3.6.8.4.1(*a*))
- JOURNAL_INDEX (§3.6.8.4.1(*b*))
- PUBL (§3.6.8.4.2(*a*))
- PUBL_AUTHOR (§3.6.8.4.2(*b*))
- PUBL_BODY (§3.6.8.4.2(*c*))
- PUBL_MANUSCRIPT_INCL (§3.6.8.4.2(*d*))

PHASING group (§3.6.6.1)
- PHASING (§3.6.6.1.1)
- PHASING_AVERAGING (§3.6.6.1.2)
- PHASING_ISOMORPHOUS (§3.6.6.1.3)
- PHASING_MAD (§3.6.6.1.4(*a*))
- PHASING_MAD_CLUST (§3.6.6.1.4(*b*))
- PHASING_MAD_EXPT (§3.6.6.1.4(*c*))
- PHASING_MAD_RATIO (§3.6.6.1.4(*d*))
- PHASING_MAD_SET (§3.6.6.1.4(*e*))
- PHASING_MIR (§3.6.6.1.5(*a*))
- PHASING_MIR_DER (§3.6.6.1.5(*c*))
- PHASING_MIR_DER_REFLN (§3.6.6.1.5(*d*))
- PHASING_MIR_DER_SHELL (§3.6.6.1.5(*e*))
- PHASING_MIR_DER_SITE (§3.6.6.1.5(*f*))
- PHASING_MIR_SHELL (§3.6.6.1.5(*b*))
- PHASING_SET (§3.6.6.1.6(*a*))
- PHASING_SET_REFLN (§3.6.6.1.6(*b*))
- PUBL (*see IUCR group*)
- PUBL_AUTHOR (*see IUCR group*)
- PUBL_BODY (*see IUCR group*)
- PUBL_MANUSCRIPT_INCL (*see IUCR group*)

REFINE group (§3.6.6.2)
- REFINE (§3.6.6.2.1(*a*))
- REFINE_ANALYZE (§3.6.6.2.2)
- REFINE_B_ISO (§3.6.6.2.4(*a*))
- REFINE_FUNCT_MINIMIZED (§3.6.6.2.1(*b*))
- REFINE_HIST (§3.6.6.2.5)
- REFINE_LS_RESTR (§3.6.6.2.3(*a*))
- REFINE_LS_RESTR_NCS (§3.6.6.2.3(*b*))
- REFINE_LS_CLASS (§3.6.2.2.3(*e*))
- REFINE_LS_RESTR_TYPE (§3.6.6.2.3(*c*))
- REFINE_LS_SHELL (§3.6.6.2.3(*d*))
- REFINE_OCCUPANCY (§3.6.6.2.4(*b*))

REFLN group (§3.6.6.3)
- REFLN (§3.6.6.3.1(*a*))
- REFLN_SYS_ABS (§3.6.6.3.1(*b*))
- REFLNS (§3.6.6.3.2(*a*))
- REFLNS_CLASS (§3.6.6.3.2(*d*))
- REFLNS_SCALE (§3.6.6.3.2(*b*))
- REFLNS_SHELL (§3.6.6.3.2(*c*))

- *SOFTWARE* (*see COMPUTING group*)
- *SPACE_GROUP* (*see SYMMETRY group*)
- *SPACE_GROUP_SYMOP* (*see SYMMETRY group*)

STRUCT group (§3.6.7.5)
- STRUCT (§3.6.7.5.1(*a*))
- STRUCT_ASYM (§3.6.7.5.1(*b*))
- STRUCT_BIOL (§3.6.7.5.1(*c*))
- STRUCT_BIOL_GEN (§3.6.7.5.1(*d*))
- STRUCT_BIOL_KEYWORDS (§3.6.7.5.1(*e*))
- STRUCT_BIOL_VIEW (§3.6.7.5.1(*f*))
- STRUCT_CONF (§3.6.7.5.2(*b*))
- STRUCT_CONF_TYPE (§3.6.7.5.2(*a*))
- STRUCT_CONN (§3.6.7.5.3(*b*))
- STRUCT_CONN_TYPE (§3.6.7.5.3(*a*))
- STRUCT_KEYWORDS (§3.6.7.5.1(*g*))
- STRUCT_MON_DETAILS (§3.6.7.5.4(*a*))
- STRUCT_MON_NUCL (§3.6.7.5.4(*b*))
- STRUCT_MON_PROT (§3.6.7.5.4(*c*))
- STRUCT_MON_PROT_CIS (§3.6.7.5.4(*d*))
- STRUCT_NCS_DOM (§3.6.7.5.5(*c*))
- STRUCT_NCS_DOM_LIM (§3.6.7.5.5(*d*))
- STRUCT_NCS_ENS (§3.6.7.5.5(*a*))
- STRUCT_NCS_ENS_GEN (§3.6.7.5.5(*b*))
- STRUCT_NCS_OPER (§3.6.7.5.5(*e*))
- STRUCT_REF (§3.6.7.5.6(*a*))
- STRUCT_REF_SEQ (§3.6.7.5.6(*b*))
- STRUCT_REF_SEQ_DIF (§3.6.7.5.6(*c*))
- STRUCT_SHEET (§3.6.7.5.7(*a*))
- STRUCT_SHEET_HBOND (§3.6.7.5.7(*e*))
- STRUCT_SHEET_ORDER (§3.6.7.5.7(*d*))
- STRUCT_SHEET_RANGE (§3.6.7.5.7(*c*))
- STRUCT_SHEET_TOPOLOGY (§3.6.7.5.7(*b*))
- STRUCT_SITE (§3.6.7.5.8(*a*))
- STRUCT_SITE_GEN (§3.6.7.5.8(*c*))
- STRUCT_SITE_KEYWORDS (§3.6.7.5.8(*b*))
- STRUCT_SITE_VIEW (§3.6.7.5.8(*d*))

SYMMETRY group (§3.6.7.6)
- SPACE_GROUP (§3.6.7.6(*c*))
- SPACE_GROUP_SYMOP (§3.6.7.6(*d*))
- SYMMETRY (§3.6.7.6(*a*))
- SYMMETRY_EQUIV (§3.6.7.6(*b*))

VALENCE group (§3.6.7.7)
- VALENCE_PARAM group (§3.6.7.7(*a*))
- VALENCE_REF group (§3.6.7.7(*b*))

and the model number. These have been added to the mmCIF category ATOM_SITE (as `_atom_site.pdbx_PDB_ins_code` and `_atom_site.pdbx_PDB_model_num`) and to all related categories that include atom nomenclature.

The convention for defining the hydrogen bonding in β-sheets differs between the PDB and mmCIF representations. Because the PDB model is fundamentally different from that found in mmCIF, a new category was created to hold the PDB data: PDBX_STRUCT_SHEET_HBOND. The correspondence between the PDB and mmCIF formats is tabulated at http://deposit.pdb.org/mmCIF/dictionaries/pdb-correspondence/pdb2mmcif.html.

A3.6.2.2. Extensions for structural genomics

An International Task Force on Deposition, Archiving, and Curation of Primary Information for Structural Genomics was formed under the auspices of the International Structural Genomics Organization (ISGO) in 2001 (Berman, 2001) and was asked to develop specifications for data from structural genomics projects to be deposited with the PDB. The recommendations from this working group are summarized at http://deposit.pdb.org/mmcif/sg-data/xstal.html and http://deposit.pdb.org/mmcif/sg-data/nmr.html. For data from crystallography-based projects, the content extensions are largely focused on a more detailed description of phasing, tracing and density modification. All of the ISGO recommendations have been incorporated into the PDB exchange dictionary.

A3.6.2.3. Noncrystallographic methods

The IUCr-sponsored development of data dictionaries has been focused exclusively on crystallographic methods. As the repository for all three-dimensional macromolecular structure data, the PDB accepts structures determined using noncrystallographic techniques such as NMR and cryo-electron microscopy. The description of noncrystallographic methods is beyond the remit of the IUCr, so the PDB has worked with the NMR and cryo-electron microscopy communities to develop data dictionaries that describe these techniques within the mmCIF framework.

A3.6.2.3.1. *NMR*

The PDB exchange dictionary includes a description of NMR sample preparation, structure solution methodology, refinement and refinement metrics. These extensions were developed in collaboration with the BioMagResBank (BMRB; Ulrich *et al.*, 1989). The BMRB is the archive for experimental NMR data for biological macromolecules and has played an active role in the development of the mmCIF data dictionary. In selecting a format for archiving NMR data, the BMRB opted to use the STAR syntax (Hall, 1991) rather than the more restrictive CIF syntax. Despite this difference in syntax, the conceptual representation of macromolecular structure in the NMR dictionary (NMRStar) has remained semantically very close to the mmCIF representation. This has facilitated the exchange of data and dictionaries between the BMRB and the PDB, the sharing of software tools, and the development of a common platform for depositing data.

A3.6.2.3.2. *Cryo-electron microscopy*

Cryo-electron microscopy (as a technique for the determination of the structure of large molecular assemblies) is also described in the PDB exchange dictionary. The data extensions for cryo-electron microscopy include a description of the sample preparation, raw volume data (Henrick *et al.*, 2003), structure solution and refinement. These extensions have a prefix of `em_` (http://mmcif.pdb.org/dictionaries/mmcif_iims.dic/Index/).

A3.6.2.3.3. *Protein production*

The International Task Force on Deposition, Archiving, and Curation of Primary Information for Structural Genomics (Section A3.6.2.2) has also provided recommendations for the deposition of information about protein production. These recommendations are summarized at http://deposit.pdb.org/mmcif/sg-data/protprod.html. These data extensions have been used as the foundation for the Protein Expression Purification and Crystallization database (PEPCdb, http://pepcdb.pdb.org/) and for the protein production process model developed to support the Structural Proteomics in Europe initiative (SPINE; http://www.spineurope.org/).

A3.6.2.4. Supporting software

The RCSB/PDB has developed a set of software tools which support the PDB exchange dictionary framework (Chapter 5.5). These include *PDB_EXTRACT*, a tool to extract data from the output files of structure determination applications; *ADIT*, a web-based editor for data files based on the PDB exchange dictionary; and *CIFTr*, a translator from mmCIF to PDB format. These applications and other supporting utilities can be downloaded from http://sw-tools.pdb.org/.

The development of the mmCIF dictionary and DDL2 has been an enormous task, and any list of contributors to the effort will certainly be incomplete. Still, we must try. We have so appreciated the people that have taken the time to think carefully and constructively about all of this, and we would like to recognize their efforts. We begin by recognizing Syd Hall, David Brown and Frank Allen, who began the entire CIF effort and who recruited us to do the extensions for macromolecular structure.

Chapter 1.1 describes the formation of the original mmCIF working group, chaired by Paula Fitzgerald and including Enrique Abola, Helen Berman, Phil Bourne, Eleanor Dodson, Art Olson, Wolfgang Steigemann, Lynn Ten Eyck and Keith Watenpaugh. However, the number of people who contributed to the original design of the mmCIF data structure is much larger. We would like to thank Steve Bryant, Vivian Stojanoff, Jean Richelle, Eldon Ulrich and Brian Toby.

There are also the people who realized the shortcomings of the original DDL and worked hard to convince us that a more rigorous underpinning for the dictionary would be needed. Among them are Michael Scharf, Peter Grey, Peter Murray-Rust, Dave Stampf and Jan Zelinka.

Writing the dictionary and developing the new DDL were just the starting points for evaluation and critique, and this effort has been greatly aided by the input from COMCIFS, the IUCr committee with oversight over this process (David Brown, Chair). But the real process of review, after the dictionary was released to the public for comment in August 1995, has involved a much larger number of people. We cannot say enough about the valuable input we have received from Frances Bernstein, Herbert Bernstein, Dale Tronrud and Peter Keller.

Our efforts have been greatly enabled by the staff of the Nucleic Acid Database at Rutgers University, who have dealt with many of the technical issues of the implementation of mmCIF with real data. So we would also like to thank Anke Gelbin, Shu-Hsin Hsieh and Christine Zardecki.

Without the three CIF workshops described in Chapter 1.1, this effort would never have taken the shape and focus it now has, and we are eternally gratefully to Eleanor Dodson (York), Phil Bourne (Tarrytown) and Shoshana Wodak (Brussels), who organized the workshops, and also to Helen Berman and John Westbrook for hosting the subsequent workshop at Rutgers following the publication of the mmCIF dictionary. We thank the European Science Foundation (ESF), the European Union (EU), the National Science Foundation (NSF) and the US Department of Energy (DOE), who provided the funding.

The RCSB/PDB is operated by Rutgers, The State University of New Jersey; the San Diego Supercomputer Center at the University of California, San Diego; and the Center for Advanced Research in Biotechnology/UMBI/NIST. RCSB/PDB is supported by funds

from the National Science Foundation (NSF), the National Institute of General Medical Sciences (NIGMS), the Office of Science, Department of Energy (DOE), the National Library of Medicine (NLM), the National Cancer Institute (NCI), the National Center for Research Resources (NCRR), the National Institute of Biomedical Imaging and Bioengineering (NIBIB) and the National Institute of Neurological Disorders and Stroke (NINDS).

References

Altona, C. & Sundaralingam, M. (1972). *Conformational analysis of the sugar ring in nucleosides and nucleotides. New description using the concept of pseudorotation. J. Am. Chem. Soc.* **94**, 8205–8212.

Berman, H. M. (2001). Chair. *Report of the task force on the deposition, archiving, and curation of the primary information.* Task Force Reports from the Second International Structural Genomics Meeting, Airlie, Virginia, USA. http://www.nigms.nih.gov/news/reports/airlie_tasks.html.

Berman, H. M., Henrick, K. & Nakamura, H. (2003). *Announcing the worldwide Protein Data Bank. Nature Struct. Biol.* **10**, 980.

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.

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

Brändén C.-I. & Jones, T. A. (1990). *Between objectivity and subjectivity. Nature (London)*, **343**, 687–689.

Brünger, A. T. (1997). *Free R value: cross-validation in crystallography. Methods Enzymol.* **277**, 366–396.

Cruickshank, D. W. J. (1999). *Remarks about protein structure precision. Acta Cryst.* D**55**, 583–601.

Driessen, H., Haneef, M. I. J., Harris, G. W., Howlin, B., Khan, G. & Moss, D. S. (1989). *RESTRAIN: restrained structure-factor least-squares refinement program for macromolecular structures. J. Appl. Cryst.* **22**, 510–516.

Engh, R. A. & Huber, R. (1991). *Accurate bond and angle parameters for X-ray protein structure refinement. Acta Cryst.* A**47**, 392–400.

Fitzgerald, P. M. D., Berman, H., Bourne, P., McMahon, B., Watenpaugh, K. & Westbrook, J. (1996). *The mmCIF dictionary: community review and final approval. Acta Cryst.* A**52** (*Suppl.*), C575.

Fitzgerald, P. M. D., McKeever, B. M., VanMiddlesworth, J. F., Springer, J. P., Heimbach, J. C., Leu, C.-T., Kerber, W. K., Dixon, R. A. F. & Darke, P. L. (1990). *Crystallographic analysis of a complex between human immunodeficiency virus type 1 protease and acetyl-pepstatin at 2.0-A resolution. J. Biol. Chem.* **265**, 14209–14219.

Hall, S. R. (1991). *The STAR file: a new format for electronic data transfer and archiving. J. Chem. Inf. Comput. Sci.* **31**, 326–333.

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.

Hamilton, W. C. (1965). *Significance tests on the crystallographic R factor. Acta Cryst.* **18**, 502–510.

Hendrickson, W. A. & Konnert, J. H. (1979). *Stereochemically restrained crystallographic least-squares refinement of macromolecule structures.* In *Biomolecular structure, conformation, function and evolution*, edited by R. Srinavisan, Vol. I, pp. 43–57. New York: Pergamon Press.

Hendrickson, W. A. & Lattman, E. E. (1970). *Representation of phase probability distributions for simplified combination of independent phase information. Acta Cryst.* B**26**, 136–143.

Henrick, K., Newman, R., Tagari, M. & Chagoyen, M. (2003). *EMDep: a web-based system for the deposition and validation of high-resolution electron microscopy macromolecular structural information. J. Struct. Biol.* **144**, 228–237.

Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Cryst.* A**47**, 110–119.

Leonard, G. A., Hambley, T. W., McAuley-Hecht, K., Brown, T. & Hunter, W. N. (1993). *Anthracycline–DNA interactions at unfavourable base-pair triplet-binding sites: structures of d(CGGCCG)/daunomycin and d(TGGCCA)/adriamycin complexes. Acta Cryst.* D**49**, 458–467.

Luzzati, V. (1952). *Traitement statistique des erreurs dans la determination des structures cristallines. Acta Cryst.* **5**, 802–810.

Narayana, N., Ginell, S. L., Russu, I. M. & Berman, H. M. (1991). *Crystal and molecular structure of a DNA fragment: d(CGTGAATTCACG). Biochemistry*, **30**, 4449–4455.

Shapiro, L., Fannon, A. M., Kwong, P. D., Thompson, A., Lehmann, M. S., Grubel, G., Legrand, J. F., Als-Nielsen, J., Colman, D. R. & Hendrickson, W. A. (1995). *Structural basis of cell–cell adhesion by cadherins. Nature (London)*, **374**, 327–337.

Tickle, I. J., Laskowski, R. A. & Moss, D. S. (1998). *R_{free} and the R_{free} ratio. I. Derivation of expected values of cross-validation residuals used in macromolecular least-squares refinement. Acta Cryst.* D**54**, 547–557.

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.

Zanotti, G., Berni, R. & Monaco, H. L. (1993). *Crystal structure of liganded and unliganded forms of bovine plasma retinol-binding protein. J. Biol. Chem.* **268**, 10728–10738.