Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Automated and accurate deposition of structures solved by X-ray diffraction to the Protein Data Bank

Huanwang Yang, Vladimir Guranovic, Shuchismita Dutta, Zukang Feng, Helen M. Berman* and John D. Westbrook

Protein Data Bank, Research Collaboratory for Structural Bioinformatics, Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, 610 Taylor Road, Piscataway, NJ 08854-8087, USA

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

Received 14 June 2004
Accepted 6 August 2004

The RCSB Protein Data Bank (PDB) has a number of options for deposition of structural data and has developed software tools to facilitate the process. In addition to *ADIT* and the *PDB Validation Suite*, a new software application, *pdb_extract*, has been designed to promote automatic data deposition of structures solved by X-ray diffraction. The *pdb_extract* software can extract information about data reduction, phasing, molecular replacement, density modification and refinement from the output files produced by many X-ray crystallographic applications. The options, procedures and tools for accurate and automated PDB data deposition are described here.

1. Introduction

The number of structure determinations of biological molecules has increased dramatically during the last several years as a result of improved methods for protein production, crystallization, data collection, phase determination and refinement. An important focus of many current international initiatives in structural genomics is the creation of automated high-throughput pipelines for structure determination and analysis (Burley *et al.*, 1999). Thus, the number of structures deposited to the Protein Data Bank (Berman *et al.*, 2000, 2003; Bernstein *et al.*, 1977) will continue to grow rapidly.

In addition to the increased number of structures, information about the experiments that produced these structures is also increasing, although at a more modest rate (Fig. 1). Whereas a minority of depositions used to include structure factors, in 2003 almost 80% of crystal structure depositions included these data. The PDB Exchange Dictionary, which is an expanded form of the macromolecular Crystallographic Information File (mmCIF; Bourne *et al.*, 1997), now includes more than 4000 potential data items (http://deposit.pdb.org/mmcif). Even though all of these data items might not be appropriate for any one structure, only a small fraction of what is appropriate is currently represented in a typical PDB file.

For the continued improvement of structure-determination methods and for data-mining applications it is important that the protein-structure information deposited into the PDB is as complete and accurate as possible. Information about data collection, phasing and refinement is fully recorded in the output files produced by the software applications used in structure determination. Thus, it would be useful to harvest this information for deposition along with the coordinates and experimental data as described earlier (Henrick, 1998; Winn, 1999) and implemented in the *CCP*4 package (Collaborative Computational Package, Number 4, 1994). Web-based protein crystallography project-information systems have also been

© 2004 International Union of Crystallography
Printed in Denmark – all rights reserved

developed that allow users to track the progress of a crystal structure determination (Haebel *et al.*, 2001; Harris & Jones, 2002).

With the advent of high-throughput X-ray crystallography and the expected higher rate of data deposition, the RCSB PDB has developed various tools for automated and accurate structure deposition. *ADIT* is available both as a web-based tool and a standalone editor for assembling, editing, validating and depositing structural data. The application *pdb_extract* can extract information from the output of standard crystallographic programs at each step of the structure-determination process and merge the information into mmCIF files that are ready for validation and deposition. The *PDB Validation Suite* (Westbrook *et al.*, 2003) creates structure-validation reports and calculates derived information that could be used for assessing the quality of a structure or for monitoring progress during refinement. Since these tools utilize the PDB mmCIF Exchange Dictionary for data exchange, their use in structure deposition also facilitates annotating and processing the data.

2. Data deposition

In addition to the coordinates and structure-factor files, information regarding the source and sequence of the macromolecules in the structure, data-collection, data-processing, structure-solution, refinement and citation information are also required for data deposition. Thus, the deposition process consists of collecting, assembling and entering all this information and finally submitting it to the PDB. It is highly recommended that the files be validated before submission. We have developed tools for all of these steps so as to make the deposition process as automatic as possible while ensuring the accuracy and integrity of the data.

2.1. Data-deposition tools

Deposition tools include *ADIT*, *pdb_extract* and the *PDB Validation Suite*. These programs can be used either independently or in an integrated way, as shown in Fig. 2. Each of these tools is described in the sections below.

2.1.1. *ADIT*. *ADIT* (http://deposit.pdb.org/adit) is an integrated software system for assembling, editing, checking, validating and depositing structural data to the PDB. The functionality of this mmCIF editor has been described elsewhere (Berman *et al.*, 2000). In an *ADIT* session, three operations can be performed: a data-format pre-check, validation and actual deposition. Optimally, all these steps should be executed for a structure deposition. In the data-format pre-check step, the format of the coordinate data file is checked to ensure that it conforms to either PDB or mmCIF format. In the validation step, the data are checked for consistency with known standards and a report is created as described below. If any major errors or warnings are highlighted here, the structure should be corrected accordingly before proceeding further. During deposition, all categories in *ADIT* should be completed appropriately and checked before submitting the structure to the PDB. Upon successful completion of a

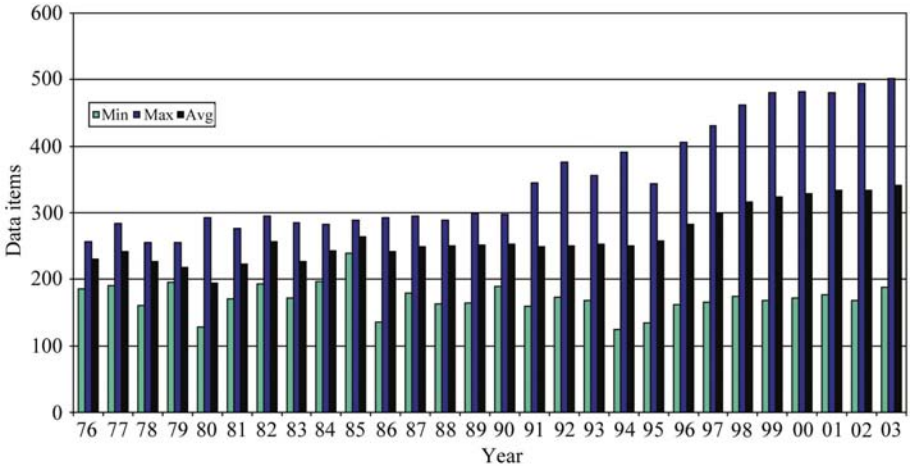

Figure 1
The number of unique data items deposited to the PDB by year. The turquoise, blue and black bars represent the minimum, maximum and average number of data items for each structure, respectively.

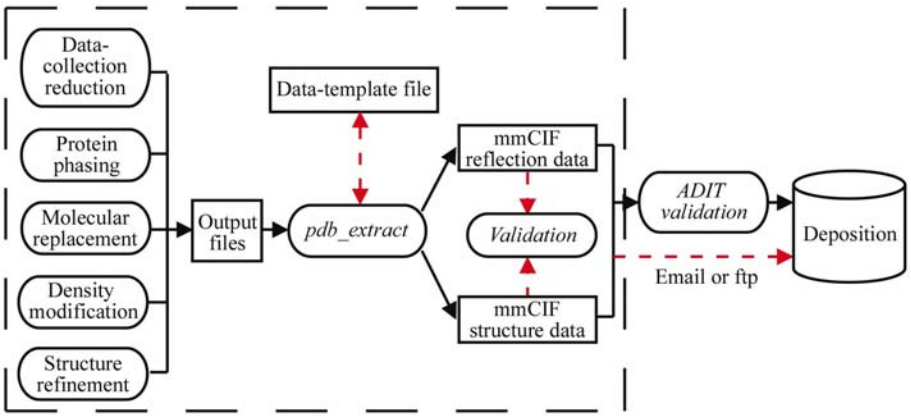

Figure 2
Ways to complete a PDB deposition using RCSB PDB tools. The flow chart in the dashed box is for the data extraction using *pdb_extract*. The black arrows (solid lines) show the recommended way for PDB deposition using *pdb_extract* and *ADIT*. The output files generated from each crystallographic step are extracted and merged into two mmCIF files which can be uploaded to *ADIT* for adding non-electronically captured information, validation and submission to the PDB. The red arrows (dotted lines) show an alternative method for PDB deposition using *pdb_extract*. The output mmCIF files generated by *pdb_extract* can either be validated by a command line or by the standalone version of *ADIT*. The validated mmCIF files can be directly sent to the PDB by ftp (ftp://pdb.rutgers.edu) or e-mail (deposit@rcsb.rutgers.edu).

Table 1
The major information extracted by *pdb_extract* from each step of crystallographic structure determination.

Category	Extracted information	Software
Data collection and reduction	Intensities and standard deviations	*HKL*/*SCALEPACK*, version 1.3–1.96 (Otwinowski & Minor, 1997)
	Data completeness (overall, resolution shells)	*d*TREK*, version 7.0SSI (Pflugrath, 1999)
	Redundancy (overall, resolution shells)	*SAINT*, version 6.35A (Bruker Analytical X-ray Systems)
	$R_{\rm merge}$, $R_{\rm sym}$ (overall, resolution shells)	*SCALA*, version 3.1.4–3.2.3 (Evans, 1997)
	$\langle I\rangle/\langle\sigma(I)\rangle$ (overall, resolution shells)	
	Total unique reflections collected	
	Resolution range	
	Collection temperature	
	Wavelength, f', f''	
Molecular replacement	Low and high resolution used in rotation and translation	*CNS*, version 0.9–1.1 (Brünger *et al.*, 1998)
	Rotation and translation methods	*AMoRe*, version *CCP*4 (4.0–5.0) (Navaza, 1994)
	Reflection cutoff criteria, reflection completeness	*MolRep*, version 7.5.01 (Vagin & Teplyakov, 2000)
	Correlation coefficients for I (or F) between observed and calculated	*EPMR*, version 2.5 (Kissinger *et al.*, 1999)
	R factor, packing information and model details	
Protein phase determination	FOM (acentric, centric, overall, resolution shells)	*CNS*, version 0.9-1.1 (Brünger *et al.*, 1998)
	$R_{\rm Cullis}$ (acentric, centric, overall, resolution shells)	*SOLVE*, version 2.0-2.06 (Terwilliger & Berendzen, 1999)
	$R_{\rm Kraut}$ (acentric, centric, overall, resolution shells)	*MLPHARE*, version *CCP*4 (4.0–5.0) (Collaborative Computational Project, Number 4, 1994)
	Phasing power (acentric, centric, overall, resolution shells)	*SHARP*/*autoSHARP*, version 1.3.*x*–2.02 (de La Fortelle & Bricogne, 1997)
	No. heavy-atom sites, heavy-atom type	*SHELXD*/*SHELXS*, version 97 (Sheldrick & Schneider, 1997)
	Heavy-atom B factor, occupancies and xyz coordinates	*PHASES*, version 95 (Furey & Swaminathan, 1997)
	Heavy-atom location method	*SnB*, version 2.0–2.2 (Weeks & Miller, 1999)
	Phasing method and software for phasing	*BnP*, version 0.93–0.96 (Weeks *et al.*, 2002)
Density modification	Density-modification method	*CNS*, version 0.9–1.1 (Brünger *et al.*, 1998)
	FOM after density modification (overall, resolution shells)	*DM*, version 2.0–2.1 (Cowtan, 1994)
	Solvent-mask determination method	*SOLOMON*, version *CCP*4 (4.0–5.0) (Collaborative Computational Project, Number 4, 1994)
		RESOLVE, version 2.0–2.06 (Terwilliger, 2000)
		SHELXE, version 97 (Sheldrick & Schneider, 1997)
Structure refinement	R factor, $R_{\rm free}$ (overall, resolution shells), $R_{\rm all}$ and $R_{\rm obs}$	*CNS*, version 0.9–1.1 (Brünger *et al.*, 1998)
	Resolution range	*REFMAC*5, version 5.0–5.2 (Murshudov *et al.*, 1999)
	No. reflections used for R factor and $R_{\rm free}$	*RESTRAIN*, version 4.7.7 (Collaborative Computational Project, Number 4, 1994)
	No. atoms refined	*SHELXL*, version 97 (Sheldrick & Schneider, 1997)
	Cartesian coordinates of all the atoms	*TNT*, version 5F (Tronrud, 1997)
	R.m.s. bond distances, bond angles, torsion angles	
	Isotropic temperature-factor restraints	
	Non-crystallographic symmetry restraints	
	Solvent model used	
	Overall average isotropic B factor	
	Overall isotropic and anisotropic B factor	
	Unit-cell parameters and space group	
	Refinement method and software	

structure deposition, a PDB ID is automatically assigned to the entry and displayed in the deposition window.

In addition to the web-based version (http://deposit.pdb.org/adit), there is a standalone Linux workstation version of *ADIT* that allows the user to prepare, check and validate structures on a local computer before actually submitting the file to the PDB. Except for actual deposition, this version contains all the functionalities of the web-based *ADIT* and can also be used to record information during different stages of structure solution or monitor the progress of a refinement. Once a file has been prepared and saved using the standalone version of *ADIT*, it can be uploaded and submitted *via* the web version.

2.1.2. *pdb_extract*. The *pdb_extract* application automatically extracts information from output and log files generated by standard software used in X-ray crystallographic structure determination for data collection, data reduction, protein phasing, molecular replacement, density modification and refinement. Since multiple software packages may be used for each step of structure determination, *pdb_extract* has been designed to accommodate the researcher's preferences in software applications. Thus, this program can extract information from the output files from a number of commonly used applications at all stages of structure determination. Table 1 lists the software applications that are supported by *pdb_extract* and the information that is currently extracted.

The PDB Exchange Dictionary contains definitions of all the data items that are extracted. Once relevant information is extracted from these files, they are merged to create two mmCIF data files: one with structure factors and the other with details of the structure including its coordinates.

There are three versions of *pdb_extract*: a web interface (http://pdb-extract.rutgers.edu), a standalone application (available from RCSB PDB) and part of the *CCP*4 package (version 5.00 and above). In all cases, the program can extract the sequence information of all polymers (protein or nucleic acid) present in the structure from the coordinate file. The sequence should be examined and any residues which were present in the crystallized molecule(s) but not modeled owing to missing electron density should be inserted here. Also, the sequence of any residue modeled as Gly or Ala owing to missing side-chain density should also be corrected.

In the web version, the sequence information automatically populates the data item corresponding to the sequence of macromolecular components. These data eventually form the `entity_poly_seq` category in mmCIF format and the SEQRES record in PDB format files. When using the standalone version or *pdb_extract* as part of the *CCP*4 package, the program creates two text files while extracting the sequence information: a data-template file (called `data_template.text`) and a script-input file (called `log_script.inp`). The data-template file contains the sequence information and also has fields for adding non-electronically produced information such as author name, citation, release status, structure title, related entries, protein source, protein-expression details, molecular names, crystallization conditions, crystal properties, radiation source, temperature and data-collection protocols. These fields may either be completed here or later when using the *ADIT* editor for deposition. The advantage of completing these fields in the data-template file is obvious when preparing multiple related structures for deposition. In this case, many of these fields are identical. Thus, instead of manually typing of all this information in *ADIT* for each deposition, the information in these fields can be copied when editing the respective data-template files. It should be noted here that the web interface of *pdb_extract* does not provide fields for including the non-electronically produced information. Thus, it should be used in conjunction with *ADIT* to produce fully populated data files that can be deposited to the PDB.

In addition to the coordinates, structure-factor files and sequence information, *pdb_extract* requires the names of the programs used for structure determination, along with their appropriate output and log files representing the final or best trial for that step of structure determination. In the *CCP4i* and web interface of *pdb_extract* the program names can be selected from lists provided and the appropriate output and log files can be uploaded. The standalone version and *pdb_extract* as part of the *CCP*4 package can both be executed either from the command line or using a script. When using the script method, the script-input file (`log_script.inp`, generated along with the data-template file) is used to list the names of the applications used for structure solution along with their output and log files, while in the command-line method all this information is directly provided at the command line using specific arguments described in the documentation for this program.

In future, when information about protein production and crystallization is produced using computer-controlled equipment, *pdb_extract* will be extended to automatically harvest this information too.

2.1.3. PDB validation. The *PDB Validation Suite* (Westbrook *et al.*, 2003; http://deposit.pdb.org/validate) creates reports based upon the following information: close contacts between all atoms both within the asymmetric unit and between symmetry-related molecules, covalent bond length and angle deviations (Clowney *et al.*, 1996; Gelbin *et al.*, 1996; Engh & Huber, 1991), chirality errors with respect to IUBMB and IUPAC conventions (Li'ebecq, 1992; Markley *et al.*, 1998), ligand and atom nomenclature according to the chemical component dictionary (ftp://ftp.rcsb.org/pub/pdb/data/monomers/het_dictionary.txt), sequence comparison and water distances. The reports produced by the *PDB Validation Suite* are output in a plain text file and in PostScript files showing both the asymmetric unit and crystal packing. Presently, validation reports from *SFCHECK* (Vaguine *et al.*, 1999), *PROCHECK* (Laskowski *et al.*, 1993) and *NUCHECK* (Feng *et al.*, 1998) are also produced. Reports based on other validation programs like *WHAT_CHECK* (Hooft *et al.*, 1996) and *MolProbity* (Lovell *et al.*, 2003) may be included here in future versions.

2.2. Data-deposition procedure

Using the tools described here can make deposition of structural data produced by X-ray diffraction experiments quick, easy, automated, complete and error-free. Fig. 2 illustrates the different ways to complete a PDB deposition using these tools developed by the RCSB.

The method used until now has been to upload coordinate and structure-factor files into *ADIT*. After (optional) validation of the files, any information not available in the uploaded files is manually typed into *ADIT* by the depositor; the data file is then ready for deposition. An improved method is to use *pdb_extract* to automatically retrieve data from the output and log files of structure-determination programs for deposition into the PDB. Data extraction can be performed either using the *pdb_extract* web interface (shown by the black arrows in Fig. 2) or using the standalone version (shown by the red arrows in Fig. 2). Additionally, *pdb_extract* may also be used as part of the *CCP*4 package. The two mmCIF files produced by all these methods can be imported to the *ADIT* web interface for online validation and submission. Alternatively, after validation the user can send the two mmCIF files produced by *pdb_extract* to the PDB using ftp (ftp://pdb.rutgers.edu) or e-mail (deposit@rcsb.rutgers.edu). The PDB ID for such an e-mail or ftp deposition is usually assigned within the next working day.

The advantage of using *pdb_extract* is that it reduces manual editing and the data are less likely to contain errors and inconsistencies. It also allows the depositor to easily capture detailed information regarding the structure determination, which leads to a more complete deposition. Optimal use of the data-template file is helpful in efficiently preparing multiple related depositions. Since the files deposited are created using software based upon the PDB Exchange Dictionary, the annotation process is also easier and takes much less time.

3. Examples of data extraction, validation and deposition

Here, we use an example to discuss a few ways of using *pdb_extract* to deposit a set of coordinates and structure-factor data into the PDB. In the example, a single crystal was used to collect data at three wavelengths (*e.g.* inflection, peak, remote edge) for a multiple anomalous diffraction (MAD; Hendrickson, 1991) experiment. The data were indexed and scaled using *HKL*2000 (Otwinowski & Minor, 1997). All three reflection-data files were used for phase determination and phase refinement using *SOLVE* (Terwilliger & Berendzen, 1999) followed by density modification using *RESOLVE* (Terwilliger, 2000). The final structure refinement was performed using the reflection data collected at the inflection edge (infl.cv) by *CNS* (Brünger *et al.*, 1998).

The relevant output and log files generated by each of the programs used in this example include three reflection data files (`scalepack1.sca`, `scalepack2.sca`, `scalepack3.sca`) and three log files (`scalepack1.log`, `scalepack2.log`, `scalepack3.log`) generated by *HKL*2000, one log file (`solve.prt`) containing phasing statistics and a PDB file (`ha.pdb`) containing heavy-atom coordinates (Se in this case) generated by *SOLVE*, one log file (`resolve.log`) containing statistics after density modification by *RESOLVE* and one mmCIF file (`cns.cif`) containing atomic coordinates and refinement statistics generated by *CNS*.

3.1. Using the *pdb_extract* web interface

The first step is to upload the coordinate file (`cns.cif`) into the *pdb_extract* web interface. This automatically extracts the sequences of all macromolecules present in the structure. Here, the web-interface window is split into two frames. The top frame is used for collecting information about the structure-factor file(s) and statistics related to data processing, while the bottom frame is for the experimental details and coordinates. Names of the applications are selected and the appropriate output and log files are added to both the frames. Thus, the structure factors for final refinement (infl.cv) and the reflection data files used for phasing (`scalepack1.sca`, `scalepack2.sca`, `scalepack3.sca`) along with their corresponding log files (`scalepack1.log`, `scalepack2.log`, `scalepack3.log`) are uploaded in the top frame. The program names and files uploaded in the bottom frame include *SOLVE* (with `solve.prt`, `ha.pdb`), *RESOLVE* (with `resolve.log`) and *CNS* (with `cns.cif`). The sequence information displayed should be corrected or completed as

Extracted mmCIF items	PDB file
_refine.ls_d_res_high 2.10	REMARK 3 RESOLUTION RANGE HIGH (ANGSTROMS) : 2.10
_refine.ls_d_res_low 79.06	REMARK 3 RESOLUTION RANGE LOW (ANGSTROMS) : 79.06
_refine.ls_percent_reflns_obs 99.400	REMARK 3 COMPLETENESS FOR RANGE (%) : 99.40
_refine.ls_number_reflns_obs 13837	REMARK 3 NUMBER OF REFLECTIONS : 13837
_refine.pdbx_ls_cross_valid_method THROUGHOUT	REMARK 3 CROSS-VALIDATION METHOD : THROUGHOUT
_refine.pdbx_R_Free_selection_details RANDOM	REMARK 3 FREE R VALUE TEST SET SELECTION : RANDOM
_refine.ls_R_factor_all 0.188	REMARK 3 R VALUE (WORKING + TEST SET) : 0.18831
_refine.ls_R_factor_R_work 0.186	REMARK 3 R VALUE (WORKING SET) : 0.18561
_refine.ls_R_factor_R_free 0.239	REMARK 3 FREE R VALUE : 0.23893
_refine.ls_percent_reflns_R_free 5.000	REMARK 3 FREE R VALUE TEST SET SIZE (%) : 5.0
_refine.ls_number_reflns_R_free 733	REMARK 3 FREE R VALUE TEST SET COUNT : 733
_refine.B_iso_mean 24.456	REMARK 3 MEAN B VALUE (OVERALL, A**2) : 24.456
_refine.aniso_B[1][1] -0.420	REMARK 3 B11 (A**2) : -0.42
_refine.aniso_B[2][2] 1.180	REMARK 3 B22 (A**2) : 1.18
_refine.aniso_B[3][3] -0.770	REMARK 3 B33 (A**2) : -0.77
_refine.aniso_B[1][2] 0.000	REMARK 3 B12 (A**2) : 0.00
_refine.aniso_B[1][3] 0.330	REMARK 3 B13 (A**2) : 0.33
_refine.aniso_B[2][3] 0.000	REMARK 3 B23 (A**2) : 0.00
_refine.correlation_coeff_Fo_to_Fc 0.939	REMARK 3 CORRELATION COEFFICIENT FO-FC : 0.939
_refine.correlation_coeff_Fo_to_Fc_free 0.913	REMARK 3 CORRELATION COEFFICIENT FO-FC FREE : 0.913
_refine.pdbx_overall_ESU_R 0.234	REMARK 3 ESU BASED ON R VALUE (A): 0.234
_refine.pdbx_overall_ESU_R_Free 0.195	REMARK 3 ESU BASED ON FREE R VALUE (A): 0.195
_refine.pdbx_overall_ESU_ML 0.114	REMARK 3 ESU BASED ON MAXIMUM LIKELIHOOD (A): 0.114
_refine.pdbx_overall_ESU_B 4.027	REMARK 3 ESU FOR B VALUES BASED ON MAXIMUM LIKELIHOOD (A**2): 4.027
_refine.pdbx_solvent_vdw_probe_radii 1.40	REMARK 3 VDW PROBE RADIUS : 1.40
_refine.pdbx_solvent_ion_probe_radii 0.80	REMARK 3 ION PROBE RADIUS : 0.80
_refine.pdbx_solvent_shrinkage_radii 0.80	REMARK 3 SHRINKAGE RADIUS : 0.80
loop_	REMARK 3 FIT IN THE HIGHEST RESOLUTION BIN.
_refine_ls_shell.d_res_low	REMARK 3 TOTAL NUMBER OF BINS USED : 20
_refine_ls_shell.d_res_high	REMARK 3 BIN RESOLUTION RANGE HIGH : 2.096
_refine_ls_shell.number_reflns_all	REMARK 3 BIN RESOLUTION RANGE LOW : 2.151
_refine_ls_shell.number_reflns_R_work	REMARK 3 REFLECTION IN BIN (WORKING SET) : 999
_refine_ls_shell.R_factor_R_work	REMARK 3 BIN R VALUE (WORKING SET) : 0.141
_refine_ls_shell.R_factor_all	REMARK 3 BIN FREE R VALUE SET COUNT : 51
_refine_ls_shell.number_reflns_R_free	REMARK 3 BIN FREE R VALUE : 0.228
_refine_ls_shell.R_factor_R_free	
_refine_ls_shell.wR_factor_R_work	
_refine_ls_shell.pdbx_total_number_of_bins_used	
2.151 2.096 1092 999 0.141 0.145 51 0.228 0.129 20	
2.210 2.151 1027 970 0.152 0.156 57 0.238 0.136 20	
.....TRUNCATED.....	
79.057 9.313 190 147 0.225 0.222 14 0.187 0.333 20	

Figure 3
An example of the correspondence between mmCIF file and PDB formatted data. The left, middle and right columns are the crystallographic data names, the data in mmCIF format and the corresponding PDB formatted data.

Table 2
Software locations and documentation.

Software	Location and tutorial for web version	Location and installation instructions for standalone version
pdb_extract	http://pdb-extract.rutgers.edu	http://deposit.pdb.org/software/PDB_EXTRACT/, http://www.ccp4.ac.uk/ (pdb_extract as part of *CCP*4)
ADIT	http://deposit.pdb.org/adit	http://deposit.pdb.org/software/ADIT
PDB validation	http://deposit.pdb.org/validate	http://deposit.pdb.org/software/VAL

necessary. Clicking the submit buttons in both frames produce two mmCIF files: one containing the structure factors and the other containing details of the structure including the coordinates. Fig. 3 shows the data-item correspondence between the merged mmCIF and the header section of the PDB file.

The mmCIF files produced here should be uploaded into the *ADIT* web interface for validation. The user can then manually add the non-electronically captured information such as the author names, citation information, deposition status and submit the completed file to the PDB.

3.2. Using the *pdb_extract* script interface

The first step here is to create a data-template file (called `data_template.text`) and script-input file (called `log_script.inp`) from the final coordinate file using the command `extract -cif cns.cif`. The protein sequence extracted from the coordinate file is written to the `data_template.text` file. It should be examined and corrected as necessary and non-electronically produced information such as author names and citation information can be included here. The `log_script.inp` file is then edited to enter the names of the applications used (*HKL*2000, *SOLVE*, *RESOLVE* and *CNS*) and the appropriate log and output files (as described above). The name of the data-template file (`data_template.text`) is also included in the `log_script.inp` file. After completing both these files, the program is run using the command `extract -ext log_script.inp`. This produces two files which are similar to those generated by the *pdb_extract* web interface (see §3.1). However, if any non-electronically generated information was included in `data_template.text`, they are carried over to the output mmCIF file.

The two mmCIF files produced by *pdb_extract* may either be uploaded into a standalone workstation version of *ADIT* for validation of the files. Alternatively, if the standalone version of *pdb_extract* has been installed, the mmCIF file containing structural details and coordinates can be validated using the command `validation-v8 -f example.mmCIF -o 2 -public -exchange -adit`. In both cases a validation report is generated, which should be carefully examined. Any errors reported here should be corrected before final deposition using *ADIT*. Alternatively, the validated mmCIF files can be also be submitted to the PDB *via* ftp or e-mail.

4. Conclusions

Procedures and tools have been developed by the RCSB PDB to facilitate data deposition to the PDB archives. Information about structure determination can be automatically extracted from many crystallographic applications and merged into mmCIF format files ready for validation and deposition. This process produces more complete and reliable files, while reducing the human effort involved in data deposition and data processing. The procedures described here lend themselves to single-structure depositions as well as to high-throughput depositions of multiple structures.

5. Access and documentation

The source code for *pdb_extract*, the *PDB validation suite* and *ADIT* are available under an Open Source license. Table 2 shows the location of the software, documentation and web servers described here. *pdb_extract* is also found in the *CCP*4 package (version 5.0 and above).

The RCSB Protein Data Bank (RCSB PDB) is operated by Rutgers, The State University of New Jersey, the San Diego Supercomputer Center at the University of California, San Diego (SDSC/UCSD) and the Center for Advanced Research in Biotechnology (CARB)/UMBI/NIST – three members of the Research Collaboratory for Structural Bioinformatics (RCSB). The RCSB PDB is supported by funds from the National Science Foundation, the National Institute of General Medical Sciences, the Office of Science, Department of Energy, the National Library of Medicine, the National Cancer Institute and the National Center for Research Resources, the National Institute of Biomedical Imaging and Bioengineering and the National Institute of Neurological Disorders and Stroke. RCSB PDB is a member of the wwPDB.

References

Berman, H. M., Henrick, K. & Nakamura, H. (2003). *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). *Nucleic Acids Res.* **28**, 235–242.

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

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

Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.

Burley, S. K., Almo, S. C., Bonanno, J. B., Capel, M., Chance, M. R., Gaasterland, T., Lin, D., Sali, A., Studier, F. W. & Swaminathan, S. (1999). *Nature Genet.* **23**, 151–157.

Clowney, L., Jain, S. C., Srinivasan, A. R., Westbrook, J., Olson, W. K. & Berman, H. M. (1996). *J. Am. Chem. Soc.* **118**, 509–518.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Cowtan, K. (1994). *Jnt CCP4/ESF–EACBM Newsl. Protein Crystallogr.* **31**, 34–38.
Engh, R. A. & Huber, R. (1991). *Acta Cryst.* A**47**, 392–400.
Evans, P. R. (1997). *Jnt CCP4 /ESF–EACBM Newsl. Protein Crystallogr.* **33**, 22–24.
Feng, Z., Westbrook, J. & Berman, H. M. (1998). Report NDB-407. Rutgers University, New Brunswick, NJ, USA.
Furey, W. & Swaminathan, S. (1997). *Methods Enzymol.* **277**, 590–620.
Gelbin, A., Schneider, B., Clowney, L., Hsieh, S.-H., Olson, W. K. & Berman, H. M. (1996). *J. Am. Chem. Soc.* **118**, 519–528.
Haebel, P. W., Arcus, V. L., Baker, E. N. & Metcalf, P. (2001). *Acta Cryst.* D**57**, 1341–1343.
Harris, M. & Jones, T. A. (2002). *Acta Cryst.* D**58**, 1889–1891.
Hendrickson, W. A. (1991). *Science*, **254**, 51–58.
Henrick, K. (1998). *CCP4 Newsl. Protein Crystallogr.* **35**, 13–16.
Hooft, R. W. W., Vriend, G., Sander, C. & Abola, E. E. (1996). *Nature (London)*, **381**, 272–272.
Kissinger, C. R., Gehlhaar, D. K. & Fogel, D. B. (1999). *Acta Cryst.* D**55**, 484–491.
La Fortelle, E. de & Bricogne, G. (1997). *Methods Enzymol.* **276**, 472–494.
Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.
Li'ebecq, C. (1992). Editor. *Biochemical nomenclature and related documents: a compendium prepared for the Committee of Editors of Biochemical Journals*, 2nd ed. North Carolina: Portland Press.
Lovell, S. C., Davis, I. W., Arendall, W. B. III, de Bakker, P. I. W., Word, J. M., Prisant, M. G., Richardson, J. S. & Richardson, D. C. (2003). *Proteins*, **50**, 437–450.
Markley, J. L., Bax, A., Arata, Y., Hilbers, C. W., Kaptein, R., Sykes, B. D., Wright, P. E. & Wüthrich, K. (1998). *J. Biomol. NMR*, **12**, 1–23.
Murshudov, G. N., Vagin, A. A., Lebedev, A., Wilson, K. S. & Dodson, E. J. (1999). *Acta Cryst.* D**55**, 247–255.
Navaza, J. (1994). *Acta Cryst.* A**50**, 157–163.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Pflugrath, J. W. (1999). *Acta Cryst.* D**55**, 1718–1725.
Sheldrick, G. & Schneider, T. (1997). *Methods Enzymol.* **277**, 319–343.
Terwilliger, T. C. (2000). *Acta Cryst.* D**56**, 965–972.
Terwilliger, T. C. & Berendzen, J. (1999). *Acta Cryst.* D**55**, 849–861.
Tronrud, D. E. (1997). *Methods Enzymol.* **277**, 306–319.
Vagin, A. & Teplyakov, A. (2000). *Acta Cryst.* D**56**, 1622–1624.
Vaguine, A. A., Richelle, J. & Wodak, S. J. (1999). *Acta Cryst.*D**55**, 191–205.
Weeks, C. M., Blessing, R. H., Miller, R., Mungee, S., Potter, S. A., Rappleye, A., Simith, G. D., Xu, H. & Furey, W. (2002). *Z. Kristallogr.* **217**, 686–693.
Weeks, C. M. & Miller, R. (1999). *Acta Cryst.* D**55**, 492–500.
Westbrook, J., Feng, Z., Burkhardt, K. & Berman, H. M. (2003). *Methods Enzymol.* **374**, 370–385.
Winn, M. (1999). *CCP4 Newsl. Protein Crystallogr.* **37**.