D426–D433 *Nucleic Acids Research, 2008, Vol. 36, Database issue* *Published online 11 December 2007*
doi:10.1093/nar/gkm937

Remediation of the protein data bank archive

Kim Henrick[1], Zukang Feng[2], Wolfgang F. Bluhm[3], Dimitris Dimitropoulos[1], Jurgen F. Doreleijers[4], Shuchismita Dutta[2], Judith L. Flippen-Anderson[2], John Ionides[1], Chisa Kamada[5], Eugene Krissinel[1], Catherine L. Lawson[2], John L. Markley[4], Haruki Nakamura[5], Richard Newman[1], Yukiko Shimizu[5], Jawahar Swaminathan[1], Sameer Velankar[1], Jeramia Ory[2], Eldon L. Ulrich[4], Wim Vranken[1], John Westbrook[2], Reiko Yamashita[5], Huanwang Yang[2], Jasmine Young[2], Muhammed Yousufuddin[2] and Helen M. Berman[2,*]

[1]MSD-EBI, EMBL Outstation-Hinxton, Cambridge CB10 1SD, UK, [2]RCSB Protein Data Bank, Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, 610 Taylor Road, Piscataway, NJ 08854-8087, USA, [3]RCSB Protein Data Bank, San Diego Supercomputer Center and the Skaggs School of Pharmacy and Pharmaceutical Sciences at the University of California, San Diego, 9500 Gilman Drive, Mailcode 0743, La Jolla, CA 92093, USA, [4]BioMagResBank, University of Wisconsin-Madison, Department of Biochemistry, 433 Babcock Drive, Madison, WI 53706, USA and [5]PDBj, Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan

Received September 13, 2007; Revised October 8, 2007; Accepted October 11, 2007

ABSTRACT

The Worldwide Protein Data Bank (wwPDB; wwpdb.org) is the international collaboration that manages the deposition, processing and distribution of the PDB archive. The online PDB archive at ftp://ftp.wwpdb.org is the repository for the coordinates and related information for more than 47 000 structures, including proteins, nucleic acids and large macromolecular complexes that have been determined using X-ray crystallography, NMR and electron microscopy techniques. The members of the wwPDB–RCSB PDB (USA), MSD-EBI (Europe), PDBj (Japan) and BMRB (USA)–have remediated this archive to address inconsistencies that have been introduced over the years. The scope and methods used in this project are presented.

INTRODUCTION

The Worldwide Protein Data Bank (wwPDB) consists of organizations that act as deposition, data processing and distribution centers for PDB data. The members are the Research Collaboratory for Structural Bioinformatics (RCSB PDB), Macromolecular Structure Data Bank at the European Bioinformatics Institute (MSD-EBI), Protein Data Bank Japan (PDBj) and the BioMagResBank (BMRB) (1). Since 1971, the PDB has been responsible for the collection, processing, archiving and distribution of biological macromolecular structural data (2). Over the last 36 years, the archive has grown from seven structures to now more than 47 000. During this same period, the methods used to determine structures, the size of individual structures and the rate at which they are being solved have all changed, as have the ways in which the archive is used.

The methods used to collect, curate and process the data also have evolved over time. Different tools have been used to collect the data including the earliest version of AutoDep developed at Brookhaven (3), a reengineered version developed at MSD-EBI (4) and ADIT used by RCSB PDB and PDBj (5,6). Over the years, data curation has become more and more automated, although expert curators still review the structures to ensure they are represented correctly. Finally, there have been subtle but definite changes in the PDB file format (7) and the definitions for the various records have been subject to different interpretations both by depositors and by curators. The result of all of these factors has been inconsistencies and outright errors in the data.

The wwPDB therefore undertook a project to remediate the entire archive. The scope of this remediation project has been to address problems that limit the utility of the archive as a whole. Thus, we have focused on the following areas: (i) improving the detailed chemical description and nomenclature of the monomer units of the biological polymers and small molecule ligands; (ii) resolving any remaining differences between the chemical and the macromolecular sequences, and updating

*To whom correspondence should be addressed. Tel: +1 732 445 4667; Fax: +1 732 445 4320; Email: berman@rcsb.rutgers.edu

Table 1. Histidine Variants in the Companion Amino Acids Variants Dictionary

CODE	Variant
HIS	HISTIDINE
HIS_LEO2	L-HISTIDINE C-TERMINAL DEPROTONATED FRAGMENT
HIS_LEO2H	L-HISTIDINE C-TERMINAL PROTONATED FRAGMENT
HIS_LEO2H_DHD1	L-HISTIDINE-C-TERMINAL PROTONATED FRAGMENT/WITH SIDE CHAIN DEPROTONATED ND1
HIS_LEO2H_DHE2	L-HISTIDINE-C-TERMINAL PROTONATED FRAGMENT/WITH SIDE CHAIN DEPROTONATED NE2
HIS_LEO2_DHD1	L-HISTIDINE-C-TERMINAL DEPROTONATED FRAGMENT/WITH SIDE CHAIN DEPROTONATED ND1
HIS_LEO2_DHE2	L-HISTIDINE-C-TERMINAL DEPROTONATED FRAGMENT/WITH SIDE CHAIN DEPROTONATED NE2
HIS_LFOH	L-HISTIDINE FREE NEUTRAL
HIS_LFOH_DHD1	L-HISTIDINE-FREE NEUTRAL/WITH SIDE CHAIN DEPROTONATED ND1
HIS_LFOH_DHE2	L-HISTIDINE-FREE NEUTRAL/WITH SIDE CHAIN DEPROTONATED NE2
HIS_LFZW	L-HISTIDINE FREE ZWITTERION
HIS_LFZW_DHD1	L-HISTIDINE-FREE ZWITTERION/WITH SIDE CHAIN DEPROTONATED ND1
HIS_LFZW_DHE2	L-HISTIDINE-FREE ZWITTERION/WITH SIDE CHAIN DEPROTONATED NE2
HIS_LL	L-HISTIDINE - LINKING EMBEDDED FRAGMENT
HIS_LL_DHD1	L-HISTIDINE-LINKING EMBEDDED FRAGMENT/WITH SIDE CHAIN DEPROTONATED ND1
HIS_LL_DHE2	L-HISTIDINE-LINKING EMBEDDED FRAGMENT/WITH SIDE CHAIN DEPROTONATED NE2
HIS_LSN3	L-HISTIDINE N-TERMINAL PROTONATED FRAGMENT
HIS_LSN3_DHD1	L-HISTIDINE-N-TERMINAL PROTONATED FRAGMENT/WITH SIDE CHAIN DEPROTONATED ND1
HIS_LSN3_DHE2	L-HISTIDINE-N-TERMINAL PROTONATED FRAGMENT/WITH SIDE CHAIN DEPROTONATED NE2

Note: The Chemical Component Dictionary is accompanied by a companion dictionary of amino acid variants that provides additional nomenclature information for the protonation states of standard amino acids in N-terminal, C-terminal and free forms. This dictionary also includes common side chain protonation states. It is similar to residue variants used in modeling software such as Charmm (42) and Amber (43).

sequence database references and taxonomies; (iii) improving the representation of viruses; and (iv) verifying primary citation assignments. We also addressed miscellaneous errors, some REMARKS, and structure factor and NMR restraint data. Coordinates have not been changed.

The impact of this work on the data files and dictionaries produced by the wwPDB are described in the following sections.

CHEMICAL DESCRIPTIONS: THE CHEMICAL COMPONENT DICTIONARY

A major portion of the wwPDB remediation project has been devoted to improving the chemical description and nomenclature used in the annotation of macromolecular structure data. This work has been incorporated into a new reference dictionary called the Chemical Component Dictionary. Key features include:

- Model and idealized coordinates
- Chemical descriptors (e.g. SMILES (8) and InChI (9)) and systematic names
- Stereochemical assignments and aromatic bond assignments
- IUPAC nomenclature for standard amino acids and nucleotides (10) with the exception of the well-established convention for C-terminal atoms OXT and HXT
- More conventional atom labeling
- Removal of redundant ligands
- Additional description of protonation states

The remediated dictionary of chemical components provides a richer and more accurate description of each molecule. The more detailed chemical definitions have been used to recheck the assignments of the monomer (13M+) and non-polymer (170K+) molecules in the PDB archive. While this chemical reference dictionary has been used in the remediation of each PDB entry, much of the information in this dictionary is not directly incorporated within individual remediated entries. In particular, the expressivity of the chemical description within PDB format CONECT records is very limited. PDB users are encouraged to take direct advantage of the content of the new chemical dictionary.

Additional chemical definitions have been created for amino acids in different states of protonation. These definitions document the nomenclature for the additional protons not specified in the standard definitions. The additional definitions are maintained in a Companion Amino Acids Variants Dictionary that provides complete molecular definitions of the protonated amino acids (Table 1).

The Chemical Component Dictionary provided the basis for the remediation of all monomer units and small molecule ligands in the PDB files. The impact of the new chemical definitions is seen in the atoms names, atom types, residue names and residue assignments.

The dictionaries and detailed descriptions of the improved description of chemical components are available for download from http://www.wwpdb.org.

CHANGES TO THE PDB COORDINATE ENTRIES

Atom and residue naming

Atom names in the polymer chains (ATOM records in the PDB file format) in the remediated data files directly reflect the nomenclature changes in the chemical dictionary. These names uniformly begin with their atom type symbol, including hydrogen atoms. Names beginning with numbers and unusual atom names have been changed accordingly. Atom types are provided for every

atom (i.e. ATOM record columns 77–78), so prior atom name justification conventions should no longer be assumed in reading atom names. As with the Chemical Component Dictionary, names for standard amino acids and nucleotides follow IUPAC recommendations (10) with the exception of the well-established convention for C-terminal atoms OXT and HXT. These nomenclature changes have been applied to standard polymeric chemical components only.

In the remediated entries, the atom names in the Companion Amino Acids Variants Dictionary have been used to describe protonated molecules; however, the extended residue names are not used. The proton names are assigned to the standard residue (i.e. HIS).

Residue assignments have all been rechecked against the new and more detailed chemical reference dictionary. A residue assignment was changed in the remediated entry if it was inconsistent with chemical connectivity and/or stereochemistry of its prior assignment, or the prior assignment was obsoleted.

DNA and RNA nucleotides now have separate chemical definitions. The DNA and RNA nucleotides are distinguished with the DNA forms relabeled as DA, DC, DG and DT. The nucleotide atom nomenclature has been standardized, and the format of the ATOM record provides explicit atom type information. Modified nucleotides formerly identified as using the 'plus-nucleotide' syntax have been relabeled with the particular 3-letter code corresponding to the full-modified nucleotide definition (Table 2).

The impact of the changes in the Chemical Component Dictionary on ligands (HET groups) in PDB entries consisted of removing redundant definitions, absorbing small modifying functional groups into complete components, and removing definitions with ambiguous chemical descriptions. More than 170 000 ligands in the data files were checked against the dictionary, and as a result 7700 names changed and 330 component definitions were obsoleted. The obsolete chemical components remain in the dictionary with an identifying status of 'OBS'. Beyond ensuring that atom names begin with their type symbol, no attempt was made to extend systematic nomenclature to non-polymer chemical components.

Table 2. RNA and DNA atom names in the remediated and unremediated files

RNA				DNA			
Remediated		**Unremediated**		**Remediated**		**Unremediated**	
A	OP3	A	O3P	DA	OP3	A	O3P
A	P	A	P	DA	P	A	P
A	OP1	A	O1P	DA	OP1	A	O1P
A	OP2	A	O2P	DA	OP2	A	O2P
A	O5'	A	O5*	DA	O5'	A	O5*
A	C5'	A	C5*	DA	C5'	A	C5*
A	C4'	A	C4*	DA	C4'	A	C4*
A	O4'	A	O4*	DA	O4'	A	O4*
A	C3'	A	C3*	DA	C3'	A	C3*
A	O3'	A	O3*	DA	O3'	A	O3*
A	C2'	A	C2*	DA	C2'	A	C2*
A	O2'	A	O2*				
A	C1'	A	C1*	DA	C1'	A	C1*
A	N9	A	N9	DA	N9	A	N9
A	C8	A	C8	DA	C8	A	C8
A	N7	A	N7	DA	N7	A	N7
A	C5	A	C5	DA	C5	A	C5
A	C6	A	C6	DA	C6	A	C6
A	N6	A	N6	DA	N6	A	N6
A	N1	A	N1	DA	N1	A	N1
A	C2	A	C2	DA	C2	A	C2
A	N3	A	N3	DA	N3	A	N3
A	C4	A	C4	DA	C4	A	C4
A	HOP3	A	3HOP	DA	HOP3	A	3HOP
A	HOP2	A	2HOP	DA	HOP2	A	2HOP
A	H5'	A	1H5*	DA	H5'	A	1H5*
A	H5''	A	2H5*	DA	H5''	A	2H5*
A	H4'	A	H4*	DA	H4'	A	H4*
A	H3'	A	H3*	DA	H3'	A	H3*
A	HO3'	A	H3T	DA	HO3'	A	H3T
A	H2'	A	H2*	DA	H2'	A	1H2*
A	HO2'	A	2HO*				
				DA	H2''	A	2H2*
A	H1'	A	H1*	DA	H1'	A	H1*
A	H8	A	H8	DA	H8	A	H8
A	H61	A	1H6	DA	H61	A	1H6
A	H62	A	2H6	DA	H62	A	2H6
A	H2	A	H2	DA	H2	A	H2

The remediated and unremediated residue name and atom name are given for the linked adenosine residue in RNA and DNA.

Examples of obsolete heterogroup names

The various hydrated magnesium ions (MO1, MO2, MO3, MO4, MO5 and MO6) have been split into an MG (magnesium ion) and the appropriate number of water molecules. In a similar manner, other examples are now obsolete het-groups. KO4 has been split into a potassium ion (K) and four water molecules (HOH), while het-group 543 has been split into a CA (calcium ion), an EOH (ethanol molecule) and six water molecules. Some 64 such groups were made obsolete. Other groups have been superceded to give a single unique het-group name in the PDB collection, including: LTR, now TRP (L-Tryptophan); FCY, now CYS (cysteine); NEV and NIV, replaced by NVP (Nevirapine); and GS4, replaced by SGC (4-thio-beta-D-glucopyranose). More than 180 such groups were made obsolete. Where possible, single atom or small groups have been replaced by complex single compound entries. These include making the ethyl group (ETH) obsolete and creating new hetgroups where previous PDB entries contained an ETH linked to another het-group.

Sequence and taxonomy

Some inconsistencies between the chemical and the coordinate macromolecular sequences were largely resolved in data files deposited before 1998 when the first set of mmCIF data files were released in 2000 (11). Remaining differences between the chemical and the macromolecular sequence have been resolved through the remediation project. All of these changes have been applied to the remediated files in PDB format. The remediated data files deposited pre-1998 reflect many changes in SEQRES and ATOM records that were required to resolve inconsistencies. Typical changes included: assignment of poly-ALA sequences to the

corresponding amino acids in the chemical sequence, reassignment of chain identifiers to correspond to complete chemical sequences, correcting terminal atom nomenclature at internal gaps and providing non-blank labels for all polymer chains.

Sequence database references and all associated difference records have been checked and/or updated along with associated taxonomy information for ~61 K sequences. UniProt (12) references have been used where possible. Sequence database correspondences were verified in December 2006. To maintain these correspondences in the future, the PDB will use the mapping data from the Structure Integration with Function, Taxonomy and Sequence (SIFTS) initiative (13).

Virus representation

The representation of viruses and large assemblies has been extended to better describe existing and anticipated entries of this type. The description of the deposited and experimental coordinate frames, symmetry and frame transformations has been generalized to better represent experiments that do not exclusively use crystallographic symmetry. This description has been properly decoupled from the description of non-crystallographic symmetry (NCS) exploited within a crystallographic structure determination. A simplified notation has been adopted to express the symmetry generation of assemblies from deposited coordinates and a standard set of matrix operations describing either point, helical or crystallographic symmetry.

Errors in archived transformation matrices required to build full assemblies from the deposited coordinates for the existing 280+ virus structures were identified by inspection of images generated with the multiscale model module of UCSF Chimera (14). Corrected matrices were obtained from the Virus Particle Explorer database (VIPERdb, http://viperdb.scripps.edu) (15) or the Protein Quaternary Structure server (PQS, http://pqs.ebi.ac.uk) (16). The corrected transformation matrices are included in remediated PDB format files.

In addition, transformations to crystal frame were collected from author text remarks or primary citations, or they were extracted from SCALE records for ~210 icosahedral virus crystal structures. NCS operations defining crystal asymmetric units were determined and crystal packing was inspected using the crystal contacts module of UCSF Chimera. Entries with structure factors were validated with SFCHECK (17). For structures deposited in the crystal frame, NCS operations are provided in the MTRIX records; for structures deposited in other frames, a text description of how to build the crystal asymmetric unit is provided in REMARK 285.

Primary citations

All primary citations have been rechecked. Citations formerly marked as *To Be Published* have been researched and either the citation has been identified or marked as *Not Published*. PubMed identifiers have been provided where available. The PubMed identifiers only appear in the remediated mmCIF and PDBML files.

Miscellaneous improvements in consistency

To improve the overall consistency and accuracy of the archive, a variety of individual corrections have been applied. These include beamline names, synchrotron facility names, source organism, method names, elimination of singleton alternate atom location labels, diffraction wavelength, computing methods and the correction of miscellaneous typographical errors. The latter includes correcting misspellings and nonstandard usage, resolving of duplicated identifiers (e.g. author residue numbers, entity and citation identifiers) and properly distinguishing null values from zero.

Free text PDB REMARKS have generally not been remediated and have not been incorporated in the remediated PDB entries. These remarks remain in a legacy remark category *data_PDB_remark* in the mmCIF and PDBML remediated files. These remarks can also be viewed in the original entries that will always be preserved.

The following PDB remarks are constructed from text templates using data items in the mmCIF/PDBML entry file: 2, 3, 4, 100, 200, 210, 215, 220, 225, 230, 240, 245, 247, 250, 265, 280, 290, 300, 350, 375, 465, 470, 500, 525, 900. These PDB remarks are reports constructed from the individual data items in the more structured mmCIF/PDBML data files. While the information presented in the PDB remarks directly corresponds to the content of the mmCIF/PDBML data files, the content of the PDB remark may not be comprehensive. The mmCIF/PDBML files should be used to obtain the most complete view of a data entry. For instance, X-ray data collection details in REMARK 200 may be found in the mmCIF/PDBML data categories in refln_group category group, and X-ray refinement details in REMARK 3 may be found in the data categories in the refine_group category group.

STRUCTURE FACTOR AND NMR RESTRAINT DATA

Many issues with structure factor data files have been addressed through a collaboration with the developers of the Uppsala Electron Density Server (EDS) (18).

Nomenclature standardization for NMR restraint files in the current PDB archive has been done as part of the NMR Restraints Grid Project, a collaboration with the Collaborative Computing Project for NMR and Bijvoet Center for Biomolecular Research. NMR restraint data files with atom nomenclature corresponding to remediated PDB data files will be available by the end of 2007.

FORMATS

The focus of the remediation project has been to address certain data consistency issues within entries and to bring all of the files in the archive to the current level of each of the PDB data formats (PDB, mmCIF/PDBx and PDBML). While the content of certain records may reflect changes from remediation, the syntax and organization of this information is largely the same as for new entries processed by PDB. Some changes in content may affect the way in which existing records are used; these issues for particular formats are discussed below.

PDB format

The record structure of the PDB format is essentially unchanged by the remediation project. The format prior to the remediation project was documented in the PDB V2.3 contents guide (19). The small number of format differences for the remediated entries are documented in the PDB V3.0.1 contents guide (http://www.wwpdb.org/docs.html). There are a few issues related to the use of the remediated files that may require attention of software developers. These include:

- Standardization of hydrogen atom nomenclature has required clarifying historical conventions in the justification of atom names in PDB ATOM records. These conventions were used to convey atom type information in early PDB format entries in which the element symbol was not included. The remediated entries uniformly include atom type information in columns 77–78. Using the justification of the atom name to derive atom type information is now strongly discouraged.
- DNA nucleotide residues are differentiated from RNA nucleotides in the remediated data files. DNA residues are now preceded by the letter 'D' (e.g. DA, DC and DG). Nucleotide modifications in the remediated files are now fully described as complete chemical components. The prior practice of identifying a nucleotide modification of with a preceding 'plus' character is not used.
- To distinguish PDB files containing the remediated nomenclature from previous files, REMARK 4 has been updated to reflect the format version 3.0 and a notation that the file has been remediated.

mmCIF/PDBx format

The remediated data files introduce no change in the syntax of mmCIF format data files (20). The following issues may require the attention of software developers:

- The maximum line length used in writing the remediated data files has been extended such that each atom record in the *atom_site* category is written in a single line.
- Additional auditing information is included in each remediated file. The underlying dictionary name, location and version are included in category *audit_conform*. Version information for each mmCIF data file is included in category *pdbx_version*.

The definitions of the data items included in the remediated mmCIF files is described in the PDB exchange dictionary version 1.045 (PDBx) (21) (http://mmcif.pdb.org/dictionaries/ascii/mmcif_pdbx.dic) This version of the dictionary incorporates some improvements in the consistency of data typing, corrections to category key structure, and miscellaneous corrections in definitions, examples and enumerations. The details of the changes are described in the dictionary history.

PDBML-XML

The remediated PDBML-XML (22) files are translated from mmCIF remediated data files and reflect the content changes described in the previous section. The revised PDBML XSD schema also includes all the changes in the PDBx version 1.045. The changes in category key structure (e.g. *citation_author*, *refine_ls_restr_ncs*) and some data type changes may require attention in parsing software.

METHODS

Chemical dictionary

The approach to improving the chemical description in the PDB relied heavily on improving and verifying the Chemical Component Dictionary. A chemical component description consists of a representative 3-dimensional model taken from the archive along its associated atom nomenclature, covalent bonding and stereochemistry.

This work involved extracting all instances of each chemical component from the archive and verifying the chemical assignments. Since prior chemical definitions did not include detailed stereochemical assignments, these were first verified relative to the molecular name. New stereochemically specific chemical definitions where created in cases where multiple enantomeric forms had previously been assigned to the same component identifier.

The preliminary screening of chemical component definitions took advantage of the stereochemical assignments used by MSDCHEM (23) obtained using the CACTVS (24) chemical informatics toolset. The stereochemical and aromatic bond assignments for the complete chemical dictionary were later rechecked using CACTVS tools and the OpenEye OEChem tools (25). Software assisted assignments of stereochemistry and aromaticity are limited to the chemical systems for which these tools were developed (e.g. primarily tetravalent organic systems). Saccharide components were also checked using the GlycoSciences PDB-care software tool (26). Improved description of chemical components involving metal coordination or dative bonding is ongoing.

A set of computationally modeled coordinates was provided in each component definition if a satisfactory set of coordinates could be obtained using either CORINA (27) or OpenEye Omega (28) packages. Systematic chemical names and chemical descriptors were also included in each component definition. Systematic names were computed using ACDLabs ACD/Name batch naming software (29) and OpenEye Lexichem (25). Stereo SMILES descriptors were computed using both CACTVS and OpenEye tools, and InChI descriptors were computed using software distributed by this IUPAC project.

The improved chemical component definitions were then used to recheck the assignments of each non-polymer, modified amino acid or modified nucleotide component instance in the PDB archive. This work involved extracting the coordinates of each component,

deriving the chemical connectivity of the component, and comparing this to the chemical dictionary. This process was driven by the DOHLC data processing program that uses BALI (30) and OpenBabel (31,32) for bond assignment and subgraph matching software from the CCP4 Coordinate Library (23).

Integration of remediated data

To manage and track data files during the remediation project a CVS archive was created for released PDB entries as of March 2006. The CVS repository was built from the mmCIF versions of these released entries.

Because the changes in sequence and taxonomy manifest the greatest change in the organization of an entry, these remediation corrections were integrated first. This work and other integration operations were performed using tools adapted from the RCSB PDB's data processing and annotation software suite (5,6,33). These tools perform edits in the macromolecular sequence and propagate these changes consistently throughout the entry. Sequence database correspondences and updated taxonomy information were also updated at this point.

After revisions in macromolecular sequence were applied, changes in component-level (modified residue and ligand) nomenclature were reintegrated into the remediated entries. Primary citation data, revised virus representations, corrections to experimental and other data items were then integrated.

Atom-level nomenclature changes were performed in the final software translation step prior to creating remediated files in PDB and PDBML formats. This was done in order to allow atom nomenclature to be refined during the course of the project. Beginning in December 2006, remediated data files in PDB, mmCIF/PDBx and PDBML formats along with supporting dictionaries were provided for public review.

Testing and validation

After all of the content and corrections were integrated into the remediated data files, these files were rechecked for consistency. Each of the wwPDB partners has contributed to this final validation of the remediated data files by applying their respective data processing and database tools to this task.

Using PDBx as a reference, each of the remediated mmCIF files was rechecked. This dictionary-level testing identifies inconsistencies in controlled vocabularies, boundary conditions and relationships between common identifiers. Similar checks of this type were performed on the PDBML data files using the XML schema translated from the PDBx. Checks for atom and residue nomenclature consistency were also performed against the Chemical Component Dictionary.

Data files were loaded into several relational database systems with different table schema. These loading operations provided further tests of data type, controlled vocabulary, boundary value and referential integrity. Loading data within a native XML database system provided additional complementary diagnostics.

During the public review of the remediated data, we benefited greatly from diagnostics contributed from PDB users who exercised the remediated data files in the application area of visualization, crystallographic phasing and refinement, docking, and homology modeling. Questions and comments about the remediated data should be sent to info@wwpdb.org.

Software support

In producing the remediated PDB data files, every effort was made to minimize the impact of the remediation on existing software applications. However, in order to support community standard nomenclature, Version 3.0 of the PDB Format was introduced. While adopting more standard nomenclature greatly simplifies the use and comparison of PDB data in most respects, many existing software applications have been developed to cope with the eccentric historical nomenclature.

As described in the previous section on 'Testing and Validation', the remediation project has included active participation from PDB users and software developers. The wwPDB maintained an informational website and mail server during the last year of the project to provide project information to earlier adopters and testers. The wwPDB also hosted a workshop for software developers at the 2007 American Crystallographic Association's annual meeting to address data representation issues that became highlighted during the remediation project.

By the time the remediated data files replaced the existing entries in August 2007, many widely-used visualization programs such as OpenRasMol, Chimera, PyMol, JMol, WebMol, KiNG, the Molecular Biology Toolkit, jV (formerly known as PDBjViewer) and Discovery Studio Visualizer were already compatible with the remediated PDB data format (34–41). wwPDB and user-contributed tools are also available to translate between the nomenclatures used in old and remediated data formats. A current list of applications reported as compatible with the remediated data files and related conversion software tools is available at http://remediation.wwpdb.org/software.html. All of the wwPDB deposition sites continue to accept depositions with either nomenclature.

FTP

The remediated data and data annotated and released by members of the wwPDB are available for download from ftp://ftp.wwpdb.org. This site is updated on a weekly basis.

A snapshot of the unremediated PDB archive (as of July 31, 2007) is available at ftp://ftp.rcsb.org. This site has been frozen, and will not be updated.

ACKNOWLEDGEMENTS

The contributions of all of the wwPDB staff members are gratefully acknowledged. Special thanks goes to the many PDB users who tested the remediated data and provided comments, especially Dan Bolser,

Alexandre M.J.J. Bonvin, Tommy Carstensen, Roland Dunbrack, Howard Feldman, Dave Howorth, Miron Livny, Eric Pettersen, the Richardson Lab at Duke University and Clemens Vonrhein. The RCSB PDB is operated by Rutgers, The State University of New Jersey and the University of California, San Diego. It 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, the National Center for Research Resources, the National Institute of Biomedical Imaging and Bioengineering, National Institute of Neurological Disorders and Stroke and the National Institute of Diabetes and Digestive and Kidney Diseases. The EMBL-EBI MSD group gratefully acknowledges the support of the Wellcome Trust, the EU (FELICS, EXTENDNMR, EuroCarbDB and 3DEM), the BBSRC, the MRC and EMBL. PDBj is supported by grant-in-aid from the Institute for Bioinformatics Research and Development, Japan Science and Technology Agency (BIRD-JST), and the Ministry of Education, Culture, Sports, Science and Technology (MEXT). The BMRB is supported by NIH grant LM05799 from the National Library of Medicine. Funding to pay the Open Access publication charge was provided by NSF DBI 03-12718.

Conflict of interest statement. None declared.

REFERENCES

1. Berman,H.M., Henrick,K. and Nakamura,H. (2003) Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.*, **10**, 980.
2. Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F. Jr., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.
3. Lin,D., Manning,N.O., Jiang,J., Abola,E.E., Stampf,D., Prilusky,J. and Sussman,J.L. (2000) AutoDep: a web-based system for deposition and validation of macromolecular structural information. *Acta Cryst. D*, **D56**, 828–841.
4. Keller,P.A., Henrick,K., McNeil,P., Moodie,S. and Barton,G.J. (1998) Deposition of macromolecular structures. *Acta Crystallogr. D Biol. Crystallogr.*, 1105–1108.
5. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
6. Dutta,S., Burkhardt,K., Bluhm,W.F. and Berman,H.M. (2005) Using the tools and resources of the RCSB Protein Data Bank. *Current Protocols in Bioinformatics*, 1.9.1–1.9.40.
7. Westbrook,J. and Fitzgerald,P.M. (2003) The PDB format, mmCIF formats and other data formats. In Bourne,P.E. and Weissig,H. (eds), *Structural Bioinformatics*, John Wiley & Sons, Inc., NJ, Hoboken, pp. 161–179.
8. Weininger,D. (1988) SMILES 1. Introduction and encoding rules. *J. Chem. Inf. Comput. Sci.*, **28**, 31.
9. © The International Union of Pure and Applied Chemistry. (2005) IUPAC International Chemical Identifier (InChI) (contact: secretariat@iupac.org)
10. Markley,J.L., Bax,A., Arata,Y., Hilbers,C.W., Kaptein,R., Sykes,B.D., Wright,P.E. and Wüthrich,K. (1998) Recommendations for the presentation of NMR structures of proteins and nucleic acids. 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. NMR*, **12**, 1–23.
11. Bhat,T.N., Bourne,P., Feng,Z., Gilliland,G., Jain,S., Ravichandran,V., Schneider,B., Schneider,K., Thanki,N., Weissig,H. *et al.* (2001) The PDB data uniformity project. *Nucleic Acids Res.*, **29**, 214–218.
12. The UniProt Consortium. (2007) The universal protein resource (UniProt). *Nucleic Acids Res*, **35**, D193–D197.
13. Velankar,S., McNeil,P., Mittard-Runte,V., Suarez,A., Barrell,D., Apweiler,R. and Henrick,K. (2005) E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res.*, **33**, D262–D265.
14. Novoselov,K.P., Shirabaikin,D.B., Umanskii,S.Y., Vladimirov,A.S., Minushev,A. and Korkin,A.A. (2002) CHIMERA: a software tool for reaction rate calculations and kinetics and thermodynamics analysis. *J. Comput. Chem.*, **23**, 1375–1389.
15. Shepherd,C.M., Borelli,I.A., Lander,G., Natarajan,P., Siddavanahalli,V., Bajaj,C., Johnson,J.E., Brooks,C.L. III and Reddy,V.S. (2006) VIPERdb: a relational database for structural virology. *Nucleic Acids Res.*, **34**, D386–D389.
16. Henrick,K. and Thornton,J.M. (1998) PQS: a protein quarternary file server. *Trends Biochem. Sci.*, **23**, 358–361.
17. Vaguine,A.A., Richelle,J. and 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 Crystallogr. D Biol. Crystallogr.*, **55**, 191–205.
18. Kleywegt,G.J., Harris,M.R., Zou,J.Y., Taylor,T.C., Wahlby,A. and Jones,T.A. (2004) The Uppsala electron-density server. *Acta Crystallogr. D Biol. Crystallogr.*, **60**, 2240–2249.
19. Callaway,J., Cummings,M., Deroski,B., Esposito,P., Forman,A., Langdon,P., Libeson,M., McCarthy,J., Sikora,J., Xue,D. *et al.* (1996) Protein data bank contents guide: atomic coordinate entry format description. Brookhaven National Laboratory.
20. Fitzgerald,P.M.D., Westbrook,J.D., Bourne,P.E., McMahon,B., Watenpaugh,K.D. and Berman,H.M. (2005) Definition and exchange of crystallographic data. In Hall,S.R. and McMahon,B. (eds), *International Tables for Crystallography*, Springer, Dordrecht, The Netherlands, Vol. G, pp. 295–443.
21. Westbrook,J., Henrick,K., Ulrich,E.L. and Berman,H.M. (2005) Definition and exchange of crystallographic data. In Hall,S.R. and McMahon,B. (eds), *International Tables for Crystallography*, Springer, Dordrecht, The Netherlands, Vol. G, pp. 195–198.
22. Westbrook,J., Ito,N., Nakamura,H., Henrick,K. and Berman,H.M. (2005) PDBML: the representation of archival macromolecular structure data in XML. *Bioinformatics*, **21**, 988–992.
23. Golovin,A., Oldfield,T.J., Tate,J.G., Velankar,S., Barton,G.J., Boutselakis,H., Dimitropoulos,D., Fillon,J., Hussain,A., Ionides,J.M. *et al.* (2004) E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res.*, **32**(Database issue), D211–D216.
24. Ihlenfeldt,W., Takahasi,Y., Abe,H. and Sasaki,S. (1994) Computation and management of chemical properties in CACTVS: An extensible networked approach toward modularity and flexibility. *J. Chem. Inf. Comp. Sci.*, **34**, 109–116.
25. OpenEye Scientific Software Inc. (2007) OpenEye OEChem version 1.5, www.eyesopen.com Santa Fe, NM, USA
26. Lutteke,T. and von der Lieth,C.W. (2004) pdb-care (PDB carbohydrate residue check): a program to support annotation of complex carbohydrate structures in PDB files. *BMC Bioinformatics*, **5**, 69.
27. Gasteiger,J., Rudolph,C. and Sadowski,J. (1990) Automatic generation of 3D-atomic coordinates for organic molecules. *Tetrahedron Comp. Method*, **3**, 537–547.
28. OpenEye Scientific Software Inc. (2007) OpenEye Omega version 2.2.1, www.eyesopen.com Santa Fe, NM, USA
29. Advanced Chemistry Development, I. (2007) ACD/Name Batch, version 9.0, Toronto ON, Canada, www.acdlabs.com
30. Hendlich,M., Rippmann,F. and Barnickel,G. (1997) BALI: automatic assignment of bond and atom types for protein ligands in the Brookhaven Protein Databank. *J. Chem. Inf. Comp. Sci.*, **37**, 774–778.
31. Guha,R., Howard,M.T., Hutchison,G.R., Murray-Rust,P., Rzepa,H., Steinbeck,C., Wegner,J. and Willighagen,E.L. (2006) The blue obelisk-interoperability in chemical informatics. *J. Chem. Inf. Model*, **46**, 991–998.
32. The Open Babel Package. (2006) Version 2.0.1 http://openbabel.sourceforge.net/

33. Westbrook,J., Feng,Z., Burkhardt,K. and Berman,H.M. (2003) Validation of protein structures for the Protein Data Bank. *Meth. Enz.*, **374**, 370–385.
34. Sayle,R. and Milner-White,E.J. (1995) RasMol: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374.
35. Bernstein,H.J. (2000) Recent changes to RasMol, recombining the variants. *Trends Biochem. Sci.*, **25**, 453–455.
36. DeLano,W. (2002) The PyMOL Molecular Graphics System on World Wide Web http://www.pymol.org
37. Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org/
38. Walther,D. (1997) WebMol–a Java-based PDB viewer. *Trends Biochem. Sci.*, **22**, 274–275.
39. Davis,I.W., Arendall,W.B. III, Richardson,D.C. and Richardson,J.S. (2006) The backrub motion: how protein backbone shrugs when a sidechain dances. *Structure*, **14**, 265–274.
40. Moreland,J.L., Gramada,A., Buzko,O.V., Zhang,Q. and Bourne,P.E. (2005) The molecular biology toolkit (MBT): a modular platform for developing molecular visualization applications. *BMC Bioinformatics*, **6**, 21.
41. Kinoshita,K. and Nakamura,H. (2004) eF-site and PDBjViewer: database and viewer for protein functional sites. *Bioinformatics*, **20**, 1329–1330.
42. Brooks,B.R., Bruccoleri,R.E., Olafson,B.D., States,D.J., Swaminathan,S. and Karplus,M. (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, **4**, 187–217.
43. Weiner,P. and Kollman,P. (1981) Amber. *J. Comput. Chem.*, **2**, 287–303.