WORLDWIDE
wwPDB
PROTEIN DATA BANK

Description of Changes and Corrections for PDB File Format Version 4.0

Provisional Document
April 12, 2011

The wwPDB has reviewed the PDB archive and created a new set of corrected files that will be released in June 2011.

This document describes the review and resulting changes and corrections.

Version 4.0 PDB entries will follow the PDB Exchange Dictionary v.4.0 (http://mmcif.pdb.org/dictionaries/mmcif_pdbx_v40.dic/Index/index.html).

TABLE OF CONTENTS

1. Biological assemblies

Problem

6126 entries were identified with missing or incomplete computational annotation of biological assemblies.

Approach

Missing computational assembly data was obtained from curated PQS results as well as PISA where these were available. Coordinate sets for predicted biological assemblies were created for each corrected entry. Molecular images of the assemblies were reviewed for the predicted assemblies. Improbable assemblies were flagged from the visual inspection and removed from the entries.

Results

Biological assembly predictions were updated in 5837 entries. 5581 entries were updated with both curated PQS and PISA results. 178 entries were updated with PISA results only and 78 entries were updated with PQS results only. Assemblies could not be predicted for 47 entries.

2. Residual B factors

Problem

The ATOM records in PDB format files produced by the REFMAC program where TLS refinement was used in many cases contain residual rather than full isotropic B-values. The type of anisotropic displacement parameters (ADPs) deposited in entries obtained using TLS refinement with the REFMAC program was verified. 7676 entries refined using REFMAC with TLS were evaluated.

Approach

Each entry in this set was first assumed to contain only residual B-values for each atom site. Using these values, an attempt was made to back-calculate a new set of full B-values using the TLSANL program. A single round of refinement was performed using these back-calculated B-values. The refinement statistics (i.e. R-values) were compared with the depositor-reported statistics. Closer reproduction of reported results using back-calculated full B-values was interpreted to mean that the entry was likely to contain residual B-values.

Detailed statistical analyses were also performed on the distribution of back-calculated refinement statistics and the differences in predicted and calculated values. These analyses did not reveal evidence of additional factors aside from TLS to account for the differences in reported refinement statistics, thus supporting the assignment procedure employed.

In some cases, adjustments were made to residue ranges defining TLS groups to correct problems introduced by relabeling residues or polymer chains during PDB data processing and annotation. A factor contributing to the problems is the ambiguous manner in which REFMAC interprets overlapping residue ranges. This issued was verified with the program developer, which then allowed to properly represent TLS groups as sets of non-overlapping residue ranges.

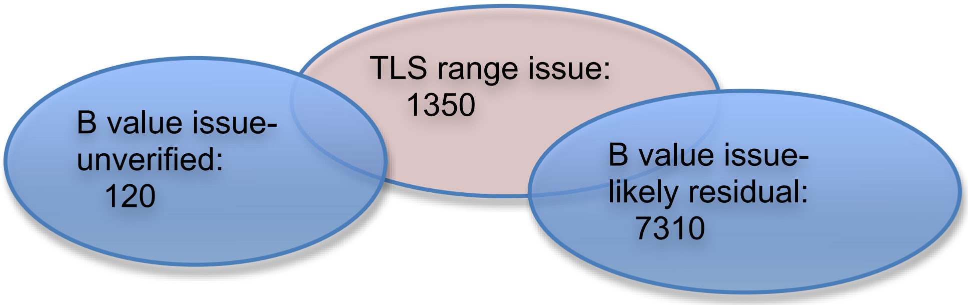
Entries have been flagged with a new mmCIF data item and additional text in REMARK 3 identifying the deposited B-value content as “LIKELY RESIDUAL”.

Example:

```
mmCIF/PDBx format:

_refine.pdbx_TLS_residual_ADP_flag  'LIKELY RESIDUAL'

PDB format:

       REMARK   3  B VALUES.
       REMARK   3   B VALUE TYPE : LIKELY RESIDUAL
```

Results

7310 entries were found to contain residual B-values using the back-calculation procedure. Based on the back-calculation procedure and other information in the deposited entries, 212 entries were found to contain full B-values. No changes were made in these entries. The procedure could not be applied to 120 entries. These entries have been labeled as "UNVERIFIED". In 1350 entries some correction was also made to the residue ranges defining TLS groups. Most of the entries were identified as "likely residual" after correction was made to the TLS ranges. A total of 7464 unique entries were corrected for TLS range and/or tagged as "likely residual" or as having unverified B-values.


TLS range issue: 1350
B value issue- unverified: 120
B value issue- likely residual: 7310

	Total
TLS range issue	1350
B value issue- likely residual	7310
B value issue- unverified	120
Total unique entries with TLS range and B value issues	7464
Full B values	212
Total unique entries reviewed	7676

Table 1: Entries involved in the B-value and TLS remediation.

3. Peptide inhibitors/antibiotics

Problem

Small peptide inhibitors and antibiotic molecules have not been uniformly represented throughout the PDB archive. These molecules have been variously described as short polymers, as single molecules, and in some cases as a collection of molecules connected by LINK records. The detailed chemical annotation of these molecules has also been inconsistent.

Approach

The representation of these molecules has been standardized according to the following scheme. Oligopeptides that are natural products or small synthetic oligopeptides containing predominately standard amino acids and standard peptide linkages have been represented as polymers. Other small oligopeptide molecules containing substantially modified amino acids or non-standard peptide linkages have been represented as single molecules.

Sequences of natural products were matched to available sequence databases (UniProtKB for gene products and Norine for non-ribosomal products). Structures of evolutionarily related families of antibiotics were represented consistently. Consistent and informative keywords were added to the files to facilitate their identification. The sequence composition of the synthetic molecules was also checked and updated to match the literature. In some cases, the original authors had to be consulted to resolve discrepancies.

Results

1029 entries containing small polypeptide and antibiotic molecules have been reviewed. In 459 entries the small oligopeptide molecule is now represented as a single molecule, and in 570 entries the peptide-containing molecule has been represented as a polymer. Entries in which the molecular representation has been changed contain additional REMARK records documenting the nomenclature changes.

The polymer sequence information for the consolidated single molecules has been preserved in the wwPDB Chemical Component Dictionary (http://www.wwpdb.org/ccd.html) and in REMARK 630 of the PDB file format. A new reference file analogous to the Chemical Component Dictionary has been created to describe the chemical structure and biological functions of the oligopeptide molecules. This new reference file also contains the detailed documentation that has been used to standardize the representation and nomenclature of these molecules and to link their chemical description to other chemical and biological data resources.

4. X-ray entries in nonstandard crystal frame

Problem

A small number of non-virus containing X-ray entries have been deposited in a non-standard coordinate orientation. For these entries, the transformation between Cartesian and fractional coordinates cannot be computed from the crystallographic cell constants alone.

Approach

Non-virus X-ray entries with coordinates not oriented in the standard crystallographic frame have been flagged with a data item and REMARK as shown in the following example:

mmCIF/PDBx format:

```
_pdbx_database_remark.id     285
_pdbx_database_remark.text
;THE ENTRY COORDINATES
ARE NOT PRESENTED IN THE STANDARD CRYSTAL FRAME.
;
```

PDB file format:

```
REMARK 285
REMARK 285 THE ENTRY COORDINATES
REMARK 285 ARE NOT PRESENTED IN THE STANDARD CRYSTAL FRAME.
```

Results

148 non-virus X-ray entries were identified as deposited in non-standard orientation and these were flagged with the new PDB file format REMARK.

5. Entries with incomplete coordinate sets

Problem

There are 3 entries in which the deposited coordinates are incomplete, containing only one residue or a few atoms, which have to be combined with another entry to generate the complete model.

Approach

The missing coordinates corresponding to these entries have been located in other PDB entries. The coordinate data for these entries has been combined to fill in the missing portions of the incomplete entries.

Results

The coordinates in these 3 entries are now complete. These entries were marked with REVDAT indicating a change in ATOM and added a detailed description in REMARK 3. The three entries and their parent entries are:

1GRH - parent 1GRG
1TN1 - parent 1TN2
9LYZ - parent 6LYZ

6. Polymers containing nonstandard polymer linkages

Problem

Previously, an effort was made to identify and merge duplicate residues (i.e., residues with identical chemistry and different 3-letter codes). In this process, a number of amino acids that appeared to be redundant but were in fact residues involved in a non-standard linkage, were accidentally merged with the corresponding standard residues.

For example, the sequence and connectivity for the current PDB entry 1AT6 is shown below. In this entry, the highlighted residue ASP is linked through a side-chain peptide bond rather than through the main chain. This non-standard link shown in Figure 1 was not reflected in the entry. In the original, pre-remediation entry, this ASP residue was properly represented by the non-standard amino-acid residue, IAS.

```
SEQRES   1    129  LYS VAL PHE GLY ARG CYS GLU LEU ALA ALA ALA MET LYS
SEQRES   2    129  ARG HIS GLY LEU ASP ASN TYR ARG GLY TYR SER LEU GLY
SEQRES   3    129  ASN TRP VAL CYS ALA ALA LYS PHE GLU SER ASN PHE ASN
SEQRES   4    129  THR GLN ALA THR ASN ARG ASN THR ASP GLY SER THR ASP
SEQRES   5    129  TYR GLY ILE LEU GLN ILE ASN SER ARG TRP TRP CYS ASN
SEQRES   6    129  ASP GLY ARG THR PRO GLY SER ARG ASN LEU CYS ASN ILE
SEQRES   7    129  PRO CYS SER ALA LEU LEU SER SER ASP ILE THR ALA SER
SEQRES   8    129  VAL ASN CYS ALA LYS LYS ILE VAL SER ASP GLY ASN GLY
SEQRES   9    129  MET ASN ALA TRP VAL ALA TRP ARG ASN ARG CYS LYS GLY
SEQRES  10    129  THR ASP VAL GLN ALA TRP ILE ARG GLY CYS ARG LEU
```

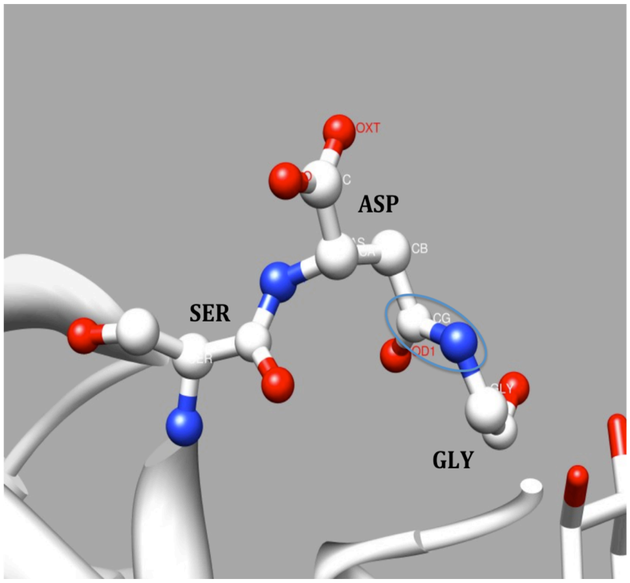


Figure 1. In entry 1AT6, the peptide bond between ASP and GLY involves the side-chain carboxylate group of the ASP instead of the main-chain carboxylate. This bond is circled in blue in the 3D diagram. The ASP residue is highlighted in yellow in the PDB SEQRES records above.

There are 58 entries where such non-standard linkage information was removed.

```
Residue
Change     Entries with nonstandard linkages
---------   ---------------------------------------------
 ASP was IAS  1at6 1c9p 1dlg 1DY5 1ejc 1ejd  1eyn 1jg3 1lsq

                   1q3g 1rtu 1ryw 1ybg 2ftm 2jv0 2ftl 2fi5

                   2fi4 2z2c 3ahs 3iss 3kqa 3kqj 3kr6 3lth

 DGL was FGA  148l 1ay3 1fjm 1p4n 1wco 2nym 2nyl 2ie3 2eax 2iae

                   2npp 3d2y 3d2z 3dw8 3gkr 3itb

 ILG was GGL  1aqv 1aqw 1aqx 1eva 1evb 1evc 1evd 1gac 1lcm 2gsr
```

Approach

Restore the original definitions of the non-standard chemical components and add more detailed linkage information to their chemical component definitions in item chem_comp.type. The list of linking types includes:

chem_comp.type	Description
L-gamma-peptide, C-delta linking	Iso-peptide linking L-gamma peptide
D-gamma-peptide, C-delta linking	Iso-peptide linking D-gamma peptide
L-beta-peptide, C-gamma linking	Iso-peptide linking L-beta peptide
D-beta-peptide, C-gamma linking	Iso-peptide linking D-beta peptide

The residues that have been added thus far are: IAS, FGA, GGL, and ACB. Others are being identified.

The advantage of this approach is that the PDB sequence explicitly reflects non-standard behavior, and this is reflected in the PDB SEQRES, SEQADV, and MODRES records. Collectively these records inform the PDB user (and software) that there is a special situation at this residue. Because this information is explicit in the sequence, it also informs users of PDB FASTA sequence data files.

Results

The sequences in the 58 PDB entries have been updated with the appropriate restored residues containing the detailed linking feature. SEQADV, MODRES records have been similarly updated.

7. Hybrid X-ray/neutron diffraction structures

Problem

There are 54 released entries that were solved using more than one experimental method. Although the experimental method and data-collection details have been captured for both methods, the relationship between the method and the data collection details is not properly represented in the PDBx dictionary.

Approach

The PDBx exchange dictionary has been extended to include the additional data items to handle hybrid X-ray/neutron diffraction methods. The key identifier for the diffraction dataset (i.e., diffrn_id and pdbx_diffrn_id) has been added to categories where data collection statistics and refinement statistics are stored (i.e. REFLNS and REFINE).

Results

The PDBx mmCIF files in the PDB archive have been updated for these categories.

8. Partial occupancy

Problem

In the 2009 remediation, occupancies were corrected in 490 X-ray and neutron entries. A mistake was made in 104 of these entries: for atoms with alternate conformer labels and with summed total occupancy less than 1.0, the occupancies were re-scaled as 1.0/n, where n is the number of conformers.

Approach

The originally deposited occupancies of the affected atoms were restored and the remediation was then carried out properly, via:

- Atoms with multiple conformations but identical coordinates and B-values were merged and their occupancies were summed.
- Atoms which now have (total) occupancies <= 1.0 were left as deposited.
- Atoms with (total) occupancies > 1 were rescaled proportionally to a sum of 1.0

Results

The occupancies have been corrected in these entries.