Protein Data Bank Contents Guide:

Atomic Coordinate Entry Format Description

Version 3.0, December 1, 2006

Updated to Version 3.01 March 30, 2007

1. Introduction

The Protein Data Bank (PDB) is an archive of experimentally determined three-dimensional structures of biological macromolecules that serves a global community of researchers, educators, and students. The data contained in the archive include atomic coordinates, bibliographic citations, primary and secondary structure, information, and crystallographic structure factors and NMR experimental data.

This guide describes the "PDB format" used by the members of the worldwide Protein Data Bank (Berman, H.M., Henrick, K. and Nakamura, H. (2003) Announcing the worldwide Protein Data Bank. Nat Struct Biol, 10, 980). Questions should be sent to info@wwpdb.org

This version of the PDB file format has been used in the wwPDB to integrate uniformity and remediation data into a single set of archival data files. This document describes the small number of differences between version 3.0 and the preceding version 2.3 formats. The complete details of the PDB format version 2.3 can be found at http://www.wwpdb.org/docs.html.

2. Title Section

This section contains records used to describe the experiment and the biological macromolecules present in the entry: HEADER, OBSLTE, TITLE, CAVEAT, COMPND, SOURCE, KEYWDS, EXPDTA, AUTHOR, REVDAT, SPRSDE, JRNL, and REMARK records. The changes in records in this section are described below.

REMARK 4

Remark 4 indicates the version of the PDB file format used to generate the file. Version 3.0 files will include a version remark like the following:

```
         1         2         3         4         5         6         7
1234567890123456789012345678901234567890123456789012345678901234567890
REMARK   4
REMARK   4 1ABC COMPLIES WITH FORMAT V. 3.0, 1-DEC-2006
REMARK   4
REMARK   4 THIS FILE IS A TEST VERSION.
REMARK   4
REMARK   4 THIS IS THE REMEDIATED VERSION OF THIS PDB ENTRY.
REMARK   4 REMEDIATED DATA FILE REVISION 3.100 (2007-03-17)
```

REMARKs 102-199 Nucleic Acids

The text remarks for nucleic acids will reflect the standardization of nomenclature for the polymer nucleotides described in later sections. In particular, the polymeric deoxyribonucleotides are represented by 2-letter codes DC, DG, DA, and DT to distinguish these from their ribonucletide counterparts. The asterisk character in the in saccharide atom names is replaced by the single prime character. The text of REMARK 105 is correspondingly changed as follows.

REMARK 105

Remark 105 is mandatory if nucleic acids exist in an entry.

Template

```
         1         2         3         4         5         6         7
1234567890123456789012345678901234567890123456789012345678901234567890
REMARK 105
REMARK 105 THE PROTEIN DATA BANK HAS ADOPTED THE SACCHARIDE CHEMISTS
REMARK 105 NOMENCLATURE FOR ATOMS OF THE DEOXYRIBOSE/RIBOSE MOIETY
REMARK 105 RATHER THAN THAT OF THE NUCLEOSIDE CHEMISTS.  THE RING
REMARK 105 OXYGEN ATOM IS LABELLED O4' INSTEAD OF O1'.
```

3. Primary Structure Section

The primary structure section of a PDB file contains the sequence of residues in each chain of the macromolecule. Embedded in these records are chain identifiers and sequence numbers that allow other records to link into the sequence.

The changes in the records in this section result from the standardization of nomenclature the standard nucleotides and nucleotide modifications.

SEQRES

SEQRES records contain the amino acid or nucleic acid sequence of residues in each chain of the macromolecule that was studied.

The ribo- and deoxyribonucleotides in the SEQRES records are now distinguished. The deoxy- forms of these residues are now identified with the residue names DA, DC, DG, DT, and DU. Modified nucleotides in the sequence are now identified by separate 3-letter residue codes. The use of the *plus* character prefix to label modified nucleotides (e.g. +A, +C, +T) is no longer used.

MODRES

The MODRES record provides descriptions of modifications (e.g., chemical or post-translational) to protein and nucleic acid residues. Included is a mapping between residue names given in a PDB entry and standard residues.

Modified nucleotides in the sequence are now identified by separate 3-letter residue codes. The use of the *plus* character prefix to label modified nucleotides (e.g. +A, +C, +T) is no longer used.

4. Heterogen Section

The heterogen section of a PDB file contains the complete description of non-standard residues in the entry. Changes in the detailed chemical descriptions of non-polymer chemical components are described in the PDB Chemical Components dictionary,

http://remediation.wwpdb.org/downloads/Components-rel-alt.cif.

There are no character/column format changes in the records in this section; however, the definition of a PDB HET group is revised owing to the change in nomenclature for the standard deoxyribonucleotides as described in the following section.

HET

HET records are used to describe non-standard residues, such as prosthetic groups, inhibitors, solvent molecules, and ions for which coordinates are supplied. Groups are considered HET if they are not part of a biological polymer described in SEQRES and considered to be a molecule bound to the polymer, or they are a chemical species that constitutes part of a biological polymer that is not one of the following:

- not one of the standard amino acids, and
- not one of the ribonucleic acids (C, G, A, T, U, and I), and
- not one of the deoxyribonucleic acids (DC, DG, DA, DT, DU and DI)
- not an unknown amino acid or nucleic acid where UNK is used to indicate the unknown residue name.

HET records also describe chemical components for which the chemical identity is unknown, in which case the group is assigned the hetID UNL (Unknown Ligand).

5. Secondary Structure Section

The secondary structure section of a PDB file describes helices, sheets, and turns found in protein and polypeptide structures.

There are no changes in the formats of the records in this section.

6. Connectivity Annotation Section

The connectivity annotation section allows the depositors to specify the existence and location of disulfide bonds and other linkages.

There are no changes in the formats of the records in this section.

7. Miscellaneous Features Section

The miscellaneous features section may describe features in the molecule such as environments surrounding a non-standard residue or an active site. Other features may be described in the remarks section but are not given a specific record type so far.

There are no changes in the formats of the records in this section.

8. Crystallographic Coordinate Transformation Section

The Crystallographic Section describes the geometry of the crystallographic experiment and the coordinate system transformations.

There are no changes in the formats of the records in this section.

9. Coordinate Section

The Coordinate Section contains the collection of atomic coordinates as well as the MODEL and ENDMDL records.

ATOM/HETATM

The ATOM records present the atomic coordinates for standard residues. They also present the occupancy and temperature factor for each atom. Non-polymer chemical coordinates use the HETATM record type. The element symbol is always present on each ATOM record; segment identifier and charge are optional.

The character/column format of the ATOM/HETATM records is not changed. Changes in ATOM/HETATM records result from the standardization atom and residue nomenclature. This nomenclature is described in electronic form in the PDB Chemical Components Dictionary, which may be downloaded at http://remediation.wwpdb.org/downloads/Components-rel-alt.cif.

10. Connectivity Section

This section provides information on chemical connectivity. LINK, HYDBND, SLTBRG, and CISPEP are found in the Connectivity Annotation section.

There are no changes in the formats of the records in this section.

11. Bookkeeping Section

The Bookkeeping Section provides some final information about the file itself.

There are no changes in the formats of the records in this section.

12. Nomenclature

Atom and residue nomenclature has been standardized in a variety of ways in PDB version 3.0 data files. All changes in nomenclature are documented in the electronic chemical components dictionary,

http://remediation.wwpdb.org/downloads/Components-rel-alt.cif.

The changes in nomenclature include:

- **IUPAC nomenclature for standard amino acid and nucleotides.** Atoms names follow the recommendations of described in *Pure & Appl. Chem.,* 70, 117-142, 1998. with the exception of the well-established convention for C-terminal atoms OXT and HXT. In this and other cases where an atom name has been changed, the previous name is retained in an alternate name in the PDB Chemical Components dictionary.
- **Discrimination of DNA and RNA linking nucleotides and modifications.** Deoxy- and ribose nucleotides now have separate chemical definitions with the DNA forms relabeled as DA, DC, DG, DT, DI and DU. Modified nucleotides formerly identified as using the *"plus-nucleotide"* syntax (e.g. +C, +G) have been relabeled with the particular 3-letter code corresponding to the full chemical description of the modified nucleotide.
- **More conventional atom labeling for non-polymer atoms.** In the new chemical definitions the following changes have been made to move the atom naming to a more conventional state.
    - Atom names begin with their element symbol
    - Heavy atom names follow the traditional PDB justification rules in which the atom element symbol is right justified in the second character position of the 4-character atom name. 4-character names for atoms with 1-character element symbols have been compressed to 3 characters.
    - Hydrogen atoms names all begin with "H" and are not subject to the justification rule.
- **Removal of redundant and deprecated ligands.** In cases where the same monomer or ligand had been defined using different identifiers, the most common identifier has been retained and the others have been marked as obsolete. Definitions which were deemed incorrect or better represented in other ways (e.g. metal hydrates) have also been obsoleted.