# An Overview of the wwPDB Remediation Project

**April 25, 2007**

The wwPDB Remediation Project is a collaboration among its members (RCSB PDB, MSD-EBI, PDBj, and BMRB) to integrate the uniformity and remediation data from each group into a single set of archival data files. This document describes the work that has been done and the schedule for testing and releasing these data.

## 1. Introduction

The evolution of experimental methods, functional knowledge of proteins, and methods used to process PDB data has introduced inconsistencies into the collection of data files within the archive. The wwPDB has worked together to remediate these **data** concentrating on:

- improving the detailed chemical description of non-polymer and monomer chemical components
- standardizing atom nomenclature for standard amino acids
- updating sequence database references and taxonomies
- resolving any remaining differences between chemical and coordinate sequence
- improving the representation of viruses
- verifying primary citation assignments

Improvements in chemical description and nomenclature changes are documented in a revised **chemical components dictionary**. The major new features of this dictionary include:

- addition of model and idealized coordinates
- chemical descriptors (e.g. SMILES and InChI) and systematic names
- stereochemical assignments
- IUPAC nomenclature for standard amino acids and nucleotides (*Pure & Appl. Chem*., 70, 117-142, 1998), with the exception of the well-established convention for C-terminal atoms OXT and HXT
- more conventional atom labeling
- removal of redundant ligands

## 2. Background

The wwPDB members have historically worked independently to fix errors and inconsistencies in PDB entries. These fixes were independently incorporated in the different database systems maintained by each group. The RCSB PDB further applied many of its corrections to the public mmCIF and PDBML data files but not to the data in PDB file format. With the exception of a modest number author requested revisions, the PDB files remain substantially unchanged from their original release.

Why the PDB archive has evolved in this peculiar way is a complicated question. When the RCSB PDB first addressed the remediation issues in 1998, it was with the intention of providing a uniform and consistent content across all formats. It was surprising and very disappointing to find that many PDB users at the time strongly objected to any changes in the released PDB entries, even if these changes addressed serious but correctable errors (*e.g.*, consistency between chemical and coordinate sequence). As a result of this prevailing attitude toward changes in PDB format entries, the RCSB PDB released its corrections in a new set of mmCIF format data files and left the data in PDB file format unchanged.

Since that initial release of mmCIF data, new data items and uniformity corrections have been added to the released mmCIF data files. More recently the content of the mmCIF data files has been translated into the content equivalent XML/PDBML format data files.

In 2003, the wwPDB was formed with the primary mission of coordinating the management of the PDB archive to guarantee that a single set of data files would be available worldwide. In 2005, the wwPDB agreed to collaborate on the production of a new version of the PDB archive that would include the corrections and updated content from all groups in mmCIF, PDBML and PDB formats.

## 3. Remediation project scope

This section describes the scope and impact of the changes introduced by the wwPDB remediation project. While this effort has been a large undertaking, it is important to note the limitations of the effort. The focus of the remediation project has been to address problems that limit the utility of the archive as a whole. This has primarily involved addressing: inconsistencies in chemical nomenclature and labeling, internal inconsistencies in macromolecular sequence within entries, consistent representation of large assemblies, and missing or incomplete citation information.

The underlying primary experimental data, coordinate data, and reported experimental descriptions have not been changed by this project. Section 3.9 enumerates a number of other issues that have not been addressed in this project.

### 3.1 Chemical description of non-polymer and monomer chemical components

The existing PDB dictionary of chemical components has been significantly extended. The dictionary now contains about 8000 definitions. The new content is described in more detail in Section 4 and Appendix A. Using this new dictionary, the chemical identities of the monomers (13M+) and ligands (168K) in all entries have been rechecked and corrected.

### 3.2 Standardization of polymer amino acid and nucleic acid nomenclature

More standard and/or conventional nomenclature has been used in the new chemical dictionary. The detailed changes in nomenclature are described in Section 4. Where changes in nomenclature have been made, these changes have been applied to all released entries.

### 3.3 Verification of primary citations

All primary citations have been rechecked. Citations formerly marked as *To Be Published* have researched and either the citation has been identified or marked as *Not Published.* PubMed identifiers have been provided where available.

### 3.4 Sequence database references and taxonomies

Sequence database references and all associated difference records have been checked and/or updated along with associated taxonomy information for ~61K sequences. UniProt references have been used where possible. Sequence database correspondences were verified in December 2006. Future changes in these correspondences will be tracked in auxiliary data files.

### 3.5 Resolution macromolecular sequence conflicts

Differences in entity sequence assignment between RCSB PDB and MSD-EBI have been resolved. Any remaining differences between the chemical sequence and the macromolecular sequence have also been resolved.

### 3.6 Assembly and 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 have been generalized to better represent experiments which 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 or helical symmetry. These additional definitions are documented in the PDB Exchange Dictionary in data categories pdbx_struct_entity_inst, pdbx_struct_oper_list, pdbx_struct_assembly, pdbx_struct_assembly_gen, pdbx_struct_asym_gen, pdbx_struct_msym_gen, pdbx_point_symmetry, and pdbx_helical_symmetry. These new data categories have been populated in remediated mmCIF and PDBML files, and have been checked with data for the existing 250+ virus entries. Corrected transformation matrices required to build full assemblies from deposited coordinates are included in remediated PDB format files.

### 3.7 Miscellaneous corrections and changes

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 incorrect alternate atom location labels, and the correction of miscellaneous typographical errors.

### 3.8 Detecting improbable data values

An analysis studied the distribution of many measured or computed values reported for data collection, phasing and refinement. This analysis revealed a number of systematic parsing errors resulting where missing values were treated as zero. These errors have been corrected. No other changes have been made to potential outlier values.

### 3.9 Issues NOT addressed by remediation

The following areas have not been addressed by the wwPDB remediation project:

- Aside from labeling, nomenclature, or typographical corrections primary experimental and coordinate data and associated experimental descriptions have not been changed.
- Free text remarks in pre-1999 files have not been fully parsed. Only those portions of remarks in older entries with regular structure have been translated into their corresponding data items.
- No scheme for systematically renaming non-polymer chemical components has been attempted.
- PDB chain identifiers have not been re-assigned for non-polymer groups or solvents

## 4. Improvements in the PDB Chemical Components Dictionary

A major focus of the wwPDB remediation project has been the reexamination of chemical definitions and instances of monomers and ligands. Improvements in chemical description and nomenclature changes are described in this section and in Appendix A.

The major new features of this dictionary are listed below.

### 4.1 Addition of *model* and *idealized* coordinates

Model heavy atom coordinates have been taken from examples in PDB entries. Hydrogen atom coordinates have been computationally modeled in most cases. Idealized coordinates have been obtained using standard molecular modeling packages (*e.g*. Molecular Networks Corina or OpenEye Omega). These packages calculate coordinates from a stereo-SMILES descriptor. Ideal coordinates have not been provided for chemical systems that are not well described by these two modeling packages (*i.e*. some organometallics and conjugated aromatic systems).

### 4.2 Addition of chemical descriptors and systematic names

SMILES and InChI chemical descriptors computed by several different software packages are provided. Computed systematic chemical names are also included. The software packages and associated version information are included with each descriptor.

### 4.3 Addition of stereochemical assignments

Absolute configuration is computed about each chiral center and bond. The Cahn-Ingold-Prelog system is used to distinguish R and S atom configurations and E and Z bond configurations. The assignments that are automated using software from CACTVS and OpenEye have proved reliable for tetrahedral centers. More complicated chemical systems have required manual inspection.

### 4.4 Identification of aromatic bonds

New chemical definitions use simplest possible covalent descriptions in terms of single, double and triple bonds. Pi, dative and delocalized bonds are not identified in the new chemical descriptions. This is largely because these bond types are not well supported by our cheminformatics tools. Separate flags are provided to identify aromatic bonds and atoms involved in aromatic bonds. The SMILES and InChI chemical descriptors include the identification of aromatic bonds.

### 4.5 IUPAC nomenclature for standard amino acid and nucleotides

Atoms names for standard amino acids and nucleotides now follow the recommendations described in *Pure & Appl. Chem.,* 70, 117-142, 1998 with the exception of the well-established convention for C-terminal atoms OXT and HXT. These nomenclature changes have been applied to polymeric chemical components only. No attempt has been made to extend systematic nomenclature to non-polymer chemical components. In cases where an atom name has been changed, the previous name is retained in an alternate data item (Appendix A).

### 4.6 Discrimination of DNA and RNA linking nucleotides and modifications

Deoxy and ribose nucleotides now have separate chemical definitions. The DNA forms are relabeled as DA, DC, DG, DT, and DI. 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.

RNA forms remain labeled as A, G, C, U, I.

### 4.7 More conventional atom labeling

Atom naming has been complicated due to the historical format restrictions that require the atom element symbol to be right justified in the second character position of the 4-character atom name. The first character position in atom names with a single letter atomic symbol such as carbon and hydrogen have included numbers or other symbols. In other cases white-space characters have been included within an atom names, or labels were chosen that obscured chemical identity of the atom (*e.g.* AC for C). In all of these situations, the resulting atom names appear non-conventional, non-intuitive, or at best awkward.

In the new chemical definitions, the following changes have been made to move the atom naming to a more conventional state. Atom names uniformly begin with their element symbol. In PDB format (Appendix B), heavy atom names follow the traditional PDB justification rules. Any 4-character names for atoms with 1-character element symbols have been compressed. Hydrogen atoms names all begin with "H" and are not subject to the justification rule. Therefore the PDB element column or mmCIF type_symbol data item should be used to determine the element type, rather than using the atom name.

This is the extent of atom re-labeling. These changes primarily affect hydrogen atoms and a limited number of chemical definitions with atypical atom names. The remaining atom names are largely unchanged.

The new detailed chemical component definitions should be used to obtain chemical bonding and stereochemical information. Aside from the standard polymer cases (Section 4.5), PDB atoms names do not reliably convey chemical structure information.

### 4.8 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 that were deemed incorrect or were better represented in other ways (e.g. metal hydrates) have also been obsoleted.

### 4.9 Additional definitions for protonated amino acids and nucleotides

Additional chemical definitions have been created for amino acids for protonation states of nucleic acids. These definitions document the nomenclature for the additional protons not specified in the standard definitions. These definitions have been given component identifiers exceeding the standard 3-character length. The one-letter and three-letter codes within these definitions adopt those of their parent standard forms. The additional definitions are maintained in the companion amino acid variants dictionary available at http://remediation.wwpdb.org/downloads.html.

## 5. Schedule for Testing and Release of Remediated Data Files

### 5.1 Testing History

The revised chemical component dictionary and example uniformity data files were released to a small group of PDB users at the beginning of December 2006. These early testers now have access to the complete set of remediated data files.

A public site describing the remediation project including dictionary and example files opened in January 2007 at http://remediation.wwpdb.org/. The latter site is linked from http://www.wwpdb.org, and will serve as the public portal describing the remediation project.

The remediated data files will be made available for broader testing as described below.

### 5.2 PDB FTP site transition for remediated data files

Remediated data files will be available for testing by FTP download in April 2007 at ftp://ftp-remediated-v3.rcsb.org. This site has the same organization as the current PDB ftp area.

Testing of remediated data files will continue until July 2007. Between April and July, data files containing both current and remediated nomenclature will be produced. These data will be used to load both the current production and remediated ftp sites. This dual operation will continue until July, when only data files containing remediated nomenclature will be produced. At this time, the current ftp site will be frozen, and the ftp.rcsb.org address redirected to the ftp site containing remediated data files.

#### 5.2.1 Changes to FTP delivery in April 2007

- A new ftp site (ftp-remediated-v3.rcsb.org) will provide the remediated data files for testing.
    - Aside from the changes in individual files, the site will be organized in the same manner as the current production ftp site. A VERSION-V3 & README-V3 file will be added in the top ftp directory to distinguish the content of the site.
    - Derived data files, such as the results of sequence clustering and putative functional assemblies, will reflect the changes from the remediated data.
    - Theoretical models and obsolete structures have not been included in the remediation project and these files will remain unchanged.
- An alias domain name has been created for the current RCSB ftp site, ftp-original-v2.rcsb.org. Prior to July, the ftp servers ftp.rcsb.org and ftp-original-v2.rcsb.org will provide the same content.

- Released entries will be added to both ftp sites (ftp-remediated-v3.rcsb.org and ftp-original-v2.rcsb.org). Files released delivered on ftp-remediated-v3.rcsb.org will follow the new remediated nomenclature conventions and use gzip compression while files delivered on ftp-orginal-v2.rcsb.org (ftp.rcsb.org) will follow current PDB nomenclature conventions and compression.

### 5.2.2 Changes to FTP delivery in July 2007

- At a date to be announced in July, the FTP delivery of PDB data will be changed to serve the remediated data.
- The contents of the site with unremediated data (ftp-original-v2.rcsb.org) will be frozen and supported for the foreseeable future. A mid-year snapshot of this archive will be added to snapshots.rcsb.org.
- The address ftp.rcsb.org will be redirected to ftp-remediated-v3.rcsb.org and will serve the remediated data with the new nomenclature.
- After this change is made, entries will only be released in the new nomenclature from ftp.rcsb.org.

# Appendix A – Chemical Component Dictionary Content

## Introduction

The purpose of this document is to describe how the wwPDB represents the polymer and non-polymer small molecule chemical constituents of macromolecular systems. The wwPDB maintains chemical descriptions of small molecule chemical components in the wwPDB Chemical Components Dictionary. This dictionary currently contains ~8000 chemical definitions which are packaged in mmCIF and PDBML formats. The information describing each chemical component is organized in a hierarchy of data categories. Each category contains a collection of related data items. At the top of the hierarchy is the CHEM_COMP category holding the identity and a few key features of the component as a whole. Linked to this category are additional data categories listing atom and bond information, chemical descriptor and identifiers associated with each component. The contents of these data categories are presented in the following sections.

## Category CHEM_COMP

This category provides the top-level description of each chemical component including essential molecular features, common names, standard identifier codes, status and revision information, and source information for representative and idealized coordinate data. The following data items are currently included in this category:

### id

This is the identifier code for the chemical component. For typical components this is a 3-letter code always expressed in uppercase.

### name

The chemical name selected at annotation time most commonly used to represent this chemical component.

### pdbx_synonyms

A semi-colon delimited list of alternative chemical names for the chemical component. These are often common names or commercial names.

### one_letter_code , three_letter_code

For standard polymer components, the one-letter and three-letter code identifiers for the component. In cases where the component identifier exceeds 3-character limit imposed by the PDB format, the three-letter code is used for the component identifier. This situation arises in the description of various protonation states for amino acids and nucleotides.

### type

If the component is a polymer residue then the type of polymer bonding is specified (e.g. D-peptide linking, DNA linking, RNA linking, …). If the component does not participate in polymer bonding the type is given the value, non-polymer.

**pdbx_type**
An alternative chemical type code that is required by RCSB software.

**formula**
The chemical formula following the Hill ordering conventions.

**formula_weight**
The formula weight (g/mol) for the chemical component.

**pdbx_formal_charge**
The total formal charge for the chemical component derived from the sum of atomic formal charges (ie. item charge in category CHEM_COMP_ATOM)

**pdbx_ambiguous_flag**
A Y/N flag indicating the detailed chemical description of the component are not well defined, or that the component cannot be properly defined at level of chemical description used in this chemical dictionary.

**mon_nstd_parent_comp_id**
This item holds the component identifier code of the standard polymer component from which the current component could be derived. In cases where the modified polymer component is sufficiently changed such that no standard parent component can be deduced, then no value is given.

**pdbx_initial_date, pdbx_modified_date**
These audit items provide the dates of creation and last modification (YYYY-MM-DD).

**pdbx_release_status**
Specifies the release status (e.g. REL, HOLD, OBS). The code REL indicates that the component is observed in a released entry. The code HOLD is used to indicate that the component is observed in an entry that is awaiting release. The status code OBS identifies components that have been replaced or owing to redefinition are no longer used.

**pdbx_replaced_by**
For the case of an obsolete component, this item provides the component identifier of the corresponding superceding component if one exists.

**pdbx_replaces**

For the case of a superceding component, this item provides the component identifier of the corresponding obsolete component. If this component definition replaces multiple definitions, a comma-separated list of obsolete component identifiers is given.

**pdbx_model_coordinates_details, pdbx_ideal_coordinates_details**
The source information and special details for the model and ideal coordinates included as part of this chemical definition in category CHEM_COMP_ATOM.

**pdbx_model_coordinates_missing_flag, pdbx_ideal_coordinates_missing_flag**
A Y/N flag indicating that the model or ideal coordinates in this chemical definition are either incomplete or missing.

**Category CHEM_COMP_ATOM**
This category is used to list the atoms in this component and the names, features, and representative coordinates for these atoms. The following data items are currently included in this category:

**comp_id**
This is the identifier code for the chemical component and is a reference to (must match) the identifier code (id) in category CHEM_COMP.

**atom_id**
This item provides a unique name for each atom within this component. The atom name also conforms to PDB atom name formatting conventions (V2.3).

**alt_atom_id**
This item provides an alternative unique name for each atom within this component. The atom name also conforms to PDB atom name formatting conventions (V3.0).

**type_symbol**
This item gives the element symbol for the atom.

**charge**
This item gives the atomic formal charge.

**pdbx_aromatic_flag**
A Y/N flag indicating that the atom is a member of an aromatic bond.

**pdbx_leaving_atom_flag**
A Y/N flag indicating that the atom may not be observed in an instance of this component owing to the formation of a covalent bond or to ionization.

**pdbx_stereo_config**

This item gives the Cahn-Ingold-Prelog stereochemical configuration (R/S) for a chiral atom. The value of "N" is given for non-chiral atoms.

**model_Cartn_x, model_Cartn_y, model_Cartn_z**
The Cartesian coordinates for the atom. These coordinates are taken from an instance of the chemical component in a PDB entry. The selection details for these coordinates are described in item pdbx_model_coordinates_details in the CHEM_COMP category.

**pdbx_model_Cartn_x_ideal, pdbx_model_Cartn_y_ideal, pdbx_model_Cartn_z_ideal**
The Cartesian coordinates for the atom. These coordinates are determined computationally. The computational details are described in item pdbx_ideal_coordinates_details in the CHEM_COMP category.

**pdbx_align**
The value of pdbx_align gives the character indentation required to justify the atom_id value within the 4-character field reserved the atom name in PDB format.

**pdbx_ordinal**
This is an integer index provided to order the atoms listed in this category.

**Category CHEM_COMP_BOND**
This category is used to list the bonded atoms in this component and the features of these bonds. The following data items are currently included in this category:

**comp_id**
This is the identifier code for the chemical component and is a reference to (must match) the identifier code (id) in category CHEM_COMP.

**atom_id_1**
This is the first atom identifier in the bonded pair. This item is a reference to (must match) an atom identifier for this component in the CHEM_COMP_ATOM category.

**atom_id_2**
This is the second atom identifier in the bonded pair. This item is a reference to (must match) an atom identifier for this component in the CHEM_COMP_ATOM category.
**value_order**
This item specifies the bond order for the bond using one of the codes SING, DOUB, TRIP, or QUAD. The description of delocalized or Pi bonds are not currently supported.

**pdbx_aromatic_flag**
A Y/N flag indicating that the bond is aromatic.

**pdbx_stereo_config**
This item gives the Cahn-Ingold-Prelog stereochemical about a double bond. The E /Z (trans/cis) notation is used to describe the configuration about the double bond; otherwise, a value of “N” is given.

**pdbx_ordinal**
This is an integer index provided to order the bonds listed in this category.

**Category CHEM_COMP_DESCRIPTOR**
This category contains a list of chemical descriptors and associated derivation information for the component. For instance, the SMILES and molecular fingerprints may be included here along with the details of the software used to produce these descriptors. The following data items are currently included in this category:

**comp_id**
This is the identifier code for the chemical component and is a reference to (must match) the identifier code (id) in category CHEM_COMP.

**descriptor**
The text of the descriptor.

**type**
This type identifies the particular descriptor specified (e.g. SMILES, SMILES_CANONICAL, InChI).

**program, program_version**
The name of the software application and associated version used to produce the identifier where appropriate.

**Category CHEM_COMP_IDENTIFIER**
This category contains a list of identifiers and associated derivation information for the component. For instance, common and systematic names or identifiers used by chemical databases may be included along with the details of the software used to produce these identifiers (where appropriate). The following data items are currently included in this category:

**comp_id**
This is the identifier code for the chemical component and is a reference to (must match) the identifier code (id) in category CHEM_COMP.

**Identifier**
The text of the identifier.

**type**

This type identifies the particular identifier specified (e.g. SYSTEM_NAME, COMMON_NAME, PUBCHEM Identifier, CAS REGIISTRY NUMBER).

**program, program_version**
The name of the software application and associated version used to produce the identifier where appropriate.

## Appendix B – PDB File Format V 3.0 Description

The following document describes the small number of differences between version 3.0 and the preceding version 2.3 formats. The complete details of these formats can be found at http://www.wwpdb.org/docs.html

### 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

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 in this section.

### 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
```

## REMARKs 102-199 Nucleic Acids

The text the 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, 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 deoxyribocleic 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. These nomenclature changes are also described in Section 12.

## 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.