An Overview of the 2008 wwPDB Remediation

March 11, 2009

Introduction

To maintain high standards in data curation and processing, the members of the worldwide Protein Data Bank (wwPDB) collaborate to standardize annotation procedures. Documentation of the procedures, formats, and related data dictionaries used in data annotation are available at the wwPDB website (www.wwpdb.org).

In consultation with our advisory committees, standing task forces, and community experts, a new version of the PDB archive has been created. Version 3.2 of the PDB archive will be released on March 17, 2009. The full description of this format can be found at: http://www.wwpdb.org/documentation/format32/v3.2.html

Summary of changes

New features

- A split identifier (SPLIT) indicates that an entry is part of a single structure made up of several PDB entries.
- A place holder (NUMMDL) indicates the number of models (previously only reported for NMR structures in REMARK 210)
- A place holder (MDLTYP) specifies minimized average model and entries containing only Cα and P atoms.
- Additional records (DBREF1/DBREF2) contain sequence references when the accession codes or sequence numbering does not fit the DBREF format
- REMARKs 475/480 indicate Zero occupancy residues/atoms
- Metal coordination information is provided in REMARK 620

Enhancements

- Database references: the source organism as listed in the NCBI Taxonomy database is indicated by the Taxonomy ID. PubMed IDs and DOIs are available for the primary citations in PDB format and in mmCIF and XML formats
- Biological assemblies: The quaternary assembly as calculated by PISA/PQS as well as author provided biological unit are included in the files (REMARK 350).
- Binding sites: SITE records define any residues that interact with ligands and metal ions, based on distance. Author provided information is also included. An

evidence code has been added to identify whether the SITE records are software calculated or author provided.

- Electron microscopy and NMR templates have been updated and standardized.
- Small molecule chemistry: The Chemical Component Dictionary (http://www.wwpdb.org/ccd.html) has been enhanced with consistent chemical and systematic naming, re-generation of SMILES strings, and chirality checks.

Miscellaneous corrections

- Sequence corrections
- Beamline wavelength corrections
- Proper handling on NCS and TLS
- Atom nomenclature
- Flag on microheterogeneity containing entries
- Connectivity check
- Re-capture of important non-standard legacy REMARKs into current standard remarks

Known Issues

Certain aspects of the files could not be changed. These include:

- Ambiguous geometry on ligands: Ligands with distorted geometry that cannot be matched to any existing chemical component have not been changed. However, the geometry issue is flagged in the cif file (_pdbx_struct_chem_comp_diagnostics).
- Microheterogeneity: When two different residues are located at the same position within one polymer chain, only one sequence is reported. These two residues are flagged as microheterogeneity in the cif file (_entity_poly_seq_hetero and _pdbx_poly_seq_scheme.hetero) and in PDB formatted files with SEQADV.
- Inhomogeneous models: Some entries with multiple models do not have the same chemistry among these models. In the remediated set, these could not be changed. In future entries, it is required that the chemistry matches among these models.
- Entries containing peptide inhibitors or antibiotics are currently being reviewed. These remediated entries will be released at a later date.