wwPDB/CCDC/D3R Ligand Validation Workshop

Center For Integrative Proteomics Research Rutgers, The State University of New Jersey

July 30/31 2015

Group C: Academic and Industrial Computational Chemistry Software Specialists

Participants

Co-Chairs:
Jeff Blaney: Genentech
David Case: Rutgers

Members:
TN Bhat: NIST
Evan Bolton: NCBI
Daniel Cheney: Bristol-Myers Squibb
Thomas Darden: OpenEye Scientific
Thomas Holder: Schroedinger
Anthony Nichols: OpenEye Scientific
Terry Stouch: University of Maryland, Baltimore County
Radka Svobodova: Masaryk University Brno

Scribes:
Suzanna Ward: Cambridge Crystallographic Data Centre
Jasmine Young: RCSB Protein Data Bank

1. **Identification and searching of ligands**
    a. Provide unambiguous way to find ligands
        i. Recommend or require depositors provide full valence description of the *input* molecule including explicit hydrogens, fully specified stereochemistry in SDF or isomeric SMILES format (machine-readable, not just a JPEG or GIF)
        ii. Using the explicit ligand provided by depositor, register it with a canonical representation (atom ordering, tautomeric form, etc.)
        iii. Provide standard InChI and InChIKey as a way to provide inter-database chemical identity linking (e.g., between CSD, PubChem, ChEMBL, PDB)
        iv. Put query molecule through same canonicalization procedure to ensure consistent search and retrieval
    b. Provide the interpreted bound state chemical valence protonization form and cross link it to the input canonical form (e.g., create a variant dictionary that shows the variations per canonical form)
    c. At deposition time require the depositor to classify small molecule entities (e.g., solvent, ion, important ligand, etc.) using a controlled vocabulary (e.g., annotate in the mmCIF output file coordinate section what is the bound ligand and its coordinates)
    d. Build into the QA procedure a single number metric on the quality of atom position (e.g., is it in electron density or not?) .. use integrated electron density map (OMIT)
    e. Flag electron density that is there but atom residing in it is not known

2. **QA control and validation report**
    a. Recommended pushing QA/QC responsibilities to depositors during deposition: set higher metrics/criteria. Provide validation report clearly indicating ligand problems and/or providing an interactive tool for depositor to obtain with machine-readable description.
    b. Provide better visibility of the validation report for the ligand
    c. Recommended inclusion of CCDC's atom-pair contact search/stats report into PDB, similar to Mogul for torsions.
    d. Add a figure of merit that includes the local protein environment. For example, compare phases and observed amplitudes with and without bound ligand (e.g., see TN Bhat & Colin Groom paper published in the 90s)
    e. Make it easy for end users of structures to visualize electron density. Pre-calculate a pre-contoured graphics object (e.g. VRML) for density in the ligand binding site region so you can toggle on and off the electron density. Atoms without occupancies should be represented differently too (for example, make it easy to visualize electron density OMIT map around the ligand atoms)
    f. If there are crystal contacts around the ligand, they should be reported. (e.g., if a crystal contact is within 6 Å of the ligand, it should be annotated)
    g. Promote the existence of a machine readable version of the validation report in XML format (consider JSON format variant)

3. **Chemical component dictionary**
    a. Provide different versions in an archive. (e.g., also include version date and timestamp)
    b. Mark entries in the PDB awaiting remediation based on changes to the CSD
    c. Make the dictionary more accessible - link to CSD from mmCIF

4. **Community-based annotation**
    a. Create a moderated Wiki for each entry to promote crowd-sourcing corrections to ligands and/or revisions and annotations of structures in general
    b. Ask for everyone else's re-refinements that have already been done, such coordinates already exist in many cases within industry.
    c. Start versioning entries – have a structure of record and enable subsequent updates
    d. Consider if criteria for how much change are required for a new submission or alternatively open this up to community.
    e. Having to publish is too much of a barrier, could every pdb entry have its own wiki – pdb control structure of record, community able to clean & share structures. Public game – 3D print molecular models for cleaned up structures.
    f. Moderated wiki with ability to comment – quick & easy feedback mechanism
    g. Should flag and fix them but it doesn't need to be the pdb that fixes them, it could be the community.

5. **Journal Policy**
    a. Journal should make submission of PDB validation report mandatory as part of manuscript submission, especially the journals that publish most of the co-crystal structures, e.g., J Med Chem.
    b. Suggest archiving workflow to help understand the process.
    c. Should journals also require mmCIF and structure factors be provided to referees during peer review process? [Potential problems and resistance were noted.]
    d. Have a list of potential reviewers that journals can access or recommend that crystallographers are used to review the crystal data alongside the publication reviewers
    e. PDB should provide a validation report specific to ligands (or emphasize the ligand aspect more strongly) and recommend journals require these for peer review process

6. Other recommendations

a. A wholesale careful re-refinement of PDB archive is *not* recommended.
b. Re-refine individual structure that results in peer-reviewed publication
c. Flag entry with quality confidence: the likelihood of the model
d. Recommend community deposit improved structures to the PDB: *may not show significant improvement on the protein structure, but can significantly improve ligand torsion angles and bond angles*.
e. Work out how removal of ligand affects R-Factor; use R-omit (from 1994 publication) or something similar
    i. Add a normalization if coverage in the database is poor. Ligand-specific correlation.
    ii. Journals and PDB should make this value mandatory just like R factor.
    iii. Should be considered as future validation report in deposition to the PDB.
    iv. Ask crystallographic community what other parameters should be reported (figure of merits for ligands fitting electron density).