**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

**Meeting notes**

General topics

Challenges in the ligand representation at the PDB:

1. Ligand is represented only at neutral state
2. Chemical description only refers to product representation, but not to the input ligand. Lack of reaction annotation.
3. Metal containing compounds- coordination partners can be swapped with amino acids in the protein environment

* Depositors are not required to enter small molecule w/ rigorous standards. PDB apparently lacks a small molecule registration system. Should we recommend a long-term goal of adopting HELM (would need more than just HELM) for a general, extensible system?
* Recommend enforcing depositors to provide chemical description for both input and output ligands during deposition. Common FAQ from annotators to depositors should be pushed to depositors’ end.
* Long discussion re: recommending PDB, CSD, and PubChem adopting consistent canonical representation, ideally SMILES, or at least provide tools for canonical exchange and integration. This would mean we don’t worry about atom names for small molecules. We would retain standard PDB residue and atom names for common residues, cofactors, solvents.
* Colin recommended requiring depositors to provide all-atom ligand covalent 2D structures, w/ explicit bond order. mmCIF as provided by PDB doesn’t specify bond order, but mmCIF is capable of this. PDB only stores bond orders in ligand dictionary; should be in mmCIF.
* We also discussed using extensibility of mmCIF to identify 2D chemical structure of the input ligand(s) and output ligand(s) with appropriate identification. In most cases they are the same, but this will handle cases where chemistry occurs in the binding site.  This will also make it easy for users to identify what are the bound ligands, which is currently often a difficult task.
* Discussion re: validation. Most journals now require papers to include official PDB validation. We should recommend/require all journals reporting co-xtal structures to require submission of PDB validation report.
* Recommended inclusion of CCDC’s atom-pair contact search/stats report into PDB, ala Mogul for torsions.
* TN Bhat & Colin Groom published a paper in the 90s on a metric for ligand quality, based on comparing phases with and without bound ligand. Other such metrics exist. We should recommend that protein crystallographers pick one or two such metrics for inclusion in the validation report, ideally as a single ‘figure of a merit’.
* Discussion seems to cover two aspects: 1) the incompleteness or error in the deposited data and how PDB validates and catches the errors; 2) how ligand should be represented and how to make the data useful to the users.
* Targeting a subset of small molecules for remediation is not a long term solution. A general long-term solution is a way to approach.
* Move away from three letter ligand codes and have longer more meaningful codes– pdb format was restrictive but in mmCIF this can be extended. Three letter codes could be more useful
* Software needs to recognize tautomers are the same when searching and integrating across databases
* Looking at one chemical class at a time is not a long term solution we need to represent small molecules in a standard why & be able to search for them so need a more general solution
* HELM is one example which provides a framework that could be extended – molecule registration system to represent any molecules – PubChem could do that when it is applicable.

Specific topics

1. *What are current best practices for selecting an initial target ligand atomic model(s) for co-­‐crystal structure refinement from X-­‐ray diffraction data?*
2. *What are current best practices for validating the ligand(s) coming from such a co-­‐crystal structure refinement?*
3. *What new data pertaining to X-­‐ray co-­‐crystal structures should be required for PDB depositions going forward?*
4. *What information should accompany journal submissions reporting X-­‐ray co-­‐ crystal structure determinations? What supplementary materials should accompany publication of X-­‐ray co-­‐crystal structure determinations?*

* 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. JMC.
* Suggest archiving workflow to help understanding the process.
* Should journals also require mmCIF and structure factors be provided to referees during peer review process?
* 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
* PDB should provide a validation report specific to ligands and recommend journals require these for peer review process

1. *What do you recommend be done with existing X-­‐ray co-­‐crystal structures in the PDB archive?*

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

Further recommendations:

* Crowd source the fixing of ligands
* Classify how suspect the ligands are based on a number of parameters.
* Ask for everyone else’s re-refinements that have already been done, so coordinates already exist.
* Start versioning entries – have a structure of record and enable subsequent updates
* Consider if criteria for how much change are required for a new submission or alternatively open this up to community.
* 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 – molecular models for so many cleaned up structures.
* Moderated wiki with ability to comment – quick & easy feedback mechanism
* Should flag and fix them but it doesn’t need the pdb that fixes them it could be the community.
* To clean ligands need to look at:
  + Ligand strain
  + Local measure of electron fit
  + rsr
  + R-Omit
  + Crystal contacts, missing chemistry

Chemical component dictionary:

* Should be versioned with different versions available in an archive.
* Use a text based version control system so you can rewind.
* Top line of file should tell you when it as created.
* Mark entries in the pdb awaiting remediation based on changes to the CCD
* Make the dictionary more accessible - link to CCD from mmCIF

1. *What do you recommend be done to improve descriptions of ligand chemistry in the PDB archive?*

* The organizations from this workshop to agree on one canonical description for representation. Need to ensure if registering a structure no matter how you depict it the same structure is picked up. A way to transform pubchem to CSD to pdb etc.
* Use canonical representation as registration and link the experimental data back from there.
* Suggest having longer and meaningful ligand identifier rather than using three letter code.
* There are software to predict tautomers.
* Provide automated tool to fix tautomer is not the scope of the PDB.
* Ligands should be deposited with all the hydrogens (regardless they are zero occupancies) and with explicit bond order to help understand the chemistry. Concern: may result close contacts during refinement.
  + Could drop H’s before final refinement – but make sure they are deposited so pseudo force fields don’t have issues
  + Would need to consider disordered water H atoms
* How to link ligand identifier to the protonation variants? mmCIF can be extended to provide the reference.
* How can PDB distinguish co-solvent vs important ligand?
* Describe both input ligand and output ligand, and annotate the difference. SMILES and sdf might be best to do this. Also consider marking point that changed from input & output.
* Add a flag to mark the ambiguity atom/position: may be understandable for crystallographers, but may not be useful for non-structural biologists. Providing difference ligand density maps at different contours for visualization may be a better way to go.
* Can PDB provide histogram distribution on the atom position or the likely atom type?
* PDB should put similarity search results out there and keep it updated
* Need same method between databases to generate SMILEs etc. databases need to work together to agree on one canonical description for representation of chemical structure.
* Routinely add bond order to mmCIFs for all new pdb entries
* Each tautomer needs a sub-id code – they currently do
* Consider having a separate archive just for the ligands?
* Allow depositor to be able to identify important ligands in a structure so they can be rank ordered
* Integrate the variant dictionary as part of the structure deposition?
* In cases where chemistry structure is not known then could put an any atom – a flag in an mmCIF
* If coordinates are unknown and you know chemistry- omit them and annotate them.
* Need a better way to annotate or mark zero occupancy/missing part of the structure so users don’t just use the coordinates- need to provide visualization tools so that atoms are annotated for medicinal chemists.
* Make it easy for end users of structures to visualise electron density – pre calculate just a graphics object so you can toggle on and off the electron density – GUI writers to provide visualization, atoms without occupancies should be represented differently too
* PDB need to supply data to make it easy for people to visualize electron density