Group B – Discussion
July 30, 2015 –

Chairs: Paul Adams: Lawrence Berkeley Laboratory Stephen Soisson: Merck

Members:

Jeffrey Bell: Schroedinger Gerard Bricogne: Global Phasing Paul Emsley: MRC Laboratory of Molecular Biology Wladek Minor: University of Virginia Garib Murshudov: MRC Laboratory of Molecular Biology Randy Read: University of Cambridge/wwPDB X-ray Validation Task Force Chair Tom Terwilliger: Los Alamos National Laboratory/ACA&IUCr Dale Tronrud: Oregon State University Greg Warren: OpenEye Scientific

Questions:

1. 1a) What are current best practices for selecting an initial target ligand atomic model(s) for co-crystal structure refinement from X-ray diffraction data? -- Combined with -- What are current best practices for validating the ligand(s) coming from such a co-crystal structure refinement?

    a. Can the starting atomic model be adequately described without also describing the conformational restraints *or* the fit to experimental density.
    b. A poor starting model will impact the final structure determination.
    c. Considerable effort is devoted to creating a starting model using chemical informatics tools, this process is not currently necessarily driven by the experimental data from the co-crystal. Selecting a particular starting conformation may introduce a bias of the particular computational tool used or small molecule crystal data.
    d. One protocol for selecting starting bond distances and angles includes using CSD/Mogul combined with ab initio QM. This is packaged in Global Phasing's Grade server.
    e. If the target molecule is not represented in the CSD or well represented via Mogul it would not typically be separately solved to generate an experimental structure.
    f. There is variable coverage of possible structural topologies in CSD influencing analyses derived from this data source.
    g. Dependence on existing experimental database structures may introduce a bias.
    h. Dealing with chemical variation or ambiguity introduces additional complexity in defining the target. This can be magnified by the chemical environment when the ligand is in complex or subject to other sources of environmental stress (e.g. radiation damage).
    i. There is a general difficulty in representing ambiguity of the chemical structure.
    j. There is a distinction between a priori ambiguity in assigning chemical identity and the typical representation of occupancy

weighted alternate conformations.

k. Shape of density and local interacting environment should be included in the assessment.
l. A priori chemical and biological information should be included along with any ligand metadata.

1) (1b) What are the best practices for generating restraints for modeling and refinement?
   a) Multiple tools exist to create restraints although some tools require commercial licensing
   b) High quality QM/MM methods are available and can be leveraged to construct restraints and starting models.
   c) Restraints should be supplied at deposition and validation protocols must consider these restraints in structure assessment.
   d) Recommendations can help shape a future improved data pipeline which includes: generating quality restraints and incorporating these properly during validation and assessment. The pipeline protocol should provide for well-justified alternative choices.
   e) The representation and validation of restraints must be accessible to non-expert users. Explicit target values can be substituted for functional approaches (e.g. QM/MM) where needed.
   f) Some chemical systems cannot be modeled purely using experimental data sources (e.g. radicals).
   g) Buster Reports evaluates all of: agreement with Grade (QM/Mogul) restraints, fit with experimental density (orthogonal ED views), real space ED statistics, and comparison of experimental structure data alone.
   h) Current practice may include the use of one of many existing tools to create restraint dictionary. These restraints will largely influence the structural outcomes.
   i) It is important to consider alternative models in ED fitting (e.g. biological target versus crystallization artifact).
   j) What is the scope of validation and how are biologically important ligands distinguished from *less important* molecules.

2) What new data pertaining to X-ray co-crystal structures should be required for PDB depositions going forward?
   a) Restraints and the details of the origin of the restraint data
   b) An *omit* map
   c) Depositors' coefficients for the *interpreted* map
   d) Depositors' ED view supporting the ligand fit
   e) Is there a single choice for *a best practice* in map protocol
   f) Should depositors' be required to respond to issues of a certain severity raised in the validation report. Could the manner of the response be formalized further. The criteria requiring a depositor comment should be defined. This information should be accessible in downstream data mining. Responses to flagged validation issues could be collected at

deposition and these responses could be mandatory. Getting the depositors responses into the validation report in time for the peer review may still be challenging in some cases. By requiring more comprehensive chemical and restraint information at deposition, more complete validation reports may be generated at an earlier stage in the deposition process.

g) The community should be further encouraged to require validation reports during per peer review.
h) Was the ligand soaked into the crystal or endogenous. The details of the sample preparation should be provided (e.g. purity, soaking treatment such as duration and solvents)
i) Other measurements on the sample (e.g. various spectroscopic data, mass spec) or other assessments of purity.
j) Identify ligands of interest – How should validation be applied differently to *interesting* and *less interesting* ligands. Differentiating the interesting ligands may provide a useful filter of some downstream analyses for non-expert users. Ligands of interest are difficult to assess retrospectively.
k) Provenance details need to be clearly assigned on all ligand metadata. Shared community libraries should be clearly named and versioned. A registry of shared resources should be published to the community to avoid duplication of efforts.

3) 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?
   a) It is not current or common practice to provide coordinates and structure factors for journal review.
   b) Including more details within the PDB validation report may provide a short-term achievable approach to better informing the review process. For example, including orthogonal ED views for the ligand of interest in the validation report would be a highly desirable.
   c) Pressure should continue to be applied on journals to encourage coordinates and structure factor data to be made available for peer review.
   d) There is general agreement on the part of *reviewers* that access to coordinates and structure factors is desirable. However, there is a fundamental limitation in the peer review process making early access to primary data difficult. Solving this problem may be out of scope of this discussion.
   e) There should *adequate* experimental description provided either in the publication or in the PDB. For instance, a details description of sample preparation, any treatment applied to the sample, and the details of sample on which data was collected (e.g. video documentation).

4) What do you recommend be done with existing X-ray co-crystal structures in the PDB archive?
   a) PDB Dynamic is a vehicle to support multiple alternative interpretations of each entry. Handling reinterpretation of existing entries is currently

possible in PDB when accompanied by a supporting publication describing the reinterpretation. The entry containing the reinterpretation currently receives a new accession code.

b) The manner of versioning may require support for branching among versions with the potential for a diversity of molecular interpretations.

c) The extent of a revision that would require a new version must be defined (e.g. changes in xyz, chemical identity, sequence)

d) Identification of contributors with be based on digital signature (e.g. ORCID) in the future.

e) Changes in any annotations should be documented more clearly within revised entries.

f) Prior versions of validation reports should remain available for analysis.

g) It would be desirable to register for notifications of entry or ligand definition revisions.

h) The validation software should be available as a downloadable package or sufficient detail for users to reproduce validation diagnostics.

i) Some consensus tooling is required to evaluate the electron density fit over a molecular environment (e.g LLDF or altenative).

j) A more standard procedure should be created to address corrections in chemical assignments.

k) Maintaining PDB-unique 3-letter codes as ligand identifiers may not be the best forward practice. Using common descriptors such as SMILES or InChI may be a better choice. There is current plan to extend this up to 16 characters in PDBx/mmCIF/PDBML data files.

l) Managing versioning for mass re-refinements, even when results are published, would be cumbersome with the current PDB accessioning method.

5) What do you recommend be done to improve descriptions of ligand chemistry in the PDB archive?
   a) Include restraints corresponding to ligand definitions
   b) Improve the representation of chemical diversity (e.g. tautomers, protonation).
   c) Provide a better description of ambiguity, for cases of -
      i) Radiation damage – changes within the experiment (hydrogen vs radical)
      ii) Alternate conformations for complete or parts of molecule
      iii) Partially modeled ligands
      iv) Distinguishing between modeled and purely computed portions of structures can be done at the atom-level using an additional data items to identify calculated, modeled, … regions of structures. This hook could be used by validation and visualization tools provide the appropriate handling of these regions. Visualization tool support is identified as particularly important.
   d) Define classes of tautomers and protonation states with a familiar

representative for each class. Further elaboration of how tautomers will be identified in coordinate records is required.

Additional topics -

Estimating of Strain –

Energetically only possible within the context of the force-field that is used to construct the ligand model. This would be difficult to assess given range of methods in current use. Another approach would be to assess geometrical differences to experimentally based reference.

Tools are required to estimating the ligand environment in terms of energetically favorable or unfavorable interactions. CCDC ReliBase and WhatCheck could be investigated as possible solutions. This could include annotation of interacting chemical groups rather than atom-level interactions (i.e. ReliBase). This is an active area of research in the CC community. The level of detail in the assessment of interactions may be limited to clear mismatches.

A list of problematic entries/ligands should be created to fuel a coordinated community effort at improvement/remediation.