Realising the Living PDB and how raw diffraction data and its metadata can help

Workshop on *Metadata for raw data from X-ray diffraction and other structural techniques*

August 22, 2015 Rovinj, Croatia

Tom Terwilliger (Los Alamos National Laboratory; Diffraction Data Deposition Working Group)
Gerard Bricogne (Global Phasing)

Background

Deposition of data allows reanalysis:

"...in the long term the community will probably have to face the issue of whether the structural database should be static or dynamic. As methodology improves, it seems likely that re-refinement of older models (either on a case-by-case basis, or as one large-scale project) might provide better models and, hopefully, increase our understanding of the chemistry and biology of the molecules under study."

--Kleywegt, Harris, Zou, Taylor, Wählby, Jones (2004). The Uppsala Electron Density Server. Acta Cryst D60, 2240.

Deposition of data and metadata allows model improvement:

> "...improvements in crystallographic software and validation tools, combined with the deposition of X-ray data into the PDB, have enabled the development of automated rerefinement protocols... which can improve most structure models compared with their initially deposited form...
>
> ...greater sophistication can be achieved [from deposition of] metadata...
>
> ...users of the PDB and software developers will greatly benefit...as it will turn what was previously a static archive of frozen models into a repository of self-improving results....
>
> Depositors will also benefit ... because it will make their structural results more futureproof leading to more citations and to higher visibility
> "

--Joosten, Womack, Vriend, Bricogne (2009). Re-refinement from deposited X-ray data can deliver improved models for most PDB entries. Acta Cryst. D65, 176–185.

The current paradigm

The PDB is the definitive repository of macromolecular structures

- The crystallographic community uses the PDB as an archive of models and data
- The biological community uses the PDB as its window into structures
- Other repositories would need to add major value to structures to be widely viewed

The current paradigm, cont...

One-time structure determination and validation is standard

- Determine structure
- Deposit data and model (interpretation of data)

Possibility of updating structures exists but infrequently used

- Original depositor can update a structure (removing original)
- Other researchers can (in principle) re-analyze and deposit another interpretation

The current paradigm, cont...

Historical focus has been on the model, not the data

- Data are used to obtain and validate model
- Model is used by the community
- There is only rare reference to the data
- A PDB "entry" is a model, with data and metadata added

Why should there be continuous improvement of structures?

Reinterpretation is feasible

- PDB-REDO has demonstrated the feasibility of systematic, continuous reinterpretation of the data in the PDB
- Software and algorithmic improvements will extend this capability
- Standardized validation procedures exist to evaluate model quality
- Availability of raw images will extend reinterpretation even further

Why should there be continuous improvement of structures? (cont...)

Reinterpretation is desirable

- Minor errors (or major ones) can be fixed
- Consistency among related structures can be improved by using consistent methods
- New structural information (loops, ligands) can be obtained
- New formalism describing a crystallographic structure (multi-model interpretations) can be employed
- Models can be optimized to provide maximal information about specific features (a particular inter-atomic distance)
- Joint refinement of groups of structures can improve all of them

Why isn't continuous improvement already fully in place?

Cultural issues

- The interpretation is customarily "owned" by the person who solved the structure
- The depositor invested major effort to obtain the structure
- "Improvements" in the model may be looked at as criticism of the depositor

Practical issues

- Which model should a person look at?
- When is a new model "improved?"
- The original depositor may have had access to critical information or employed specialized approaches
- Deposition is a big job

What might continuous improvement of macromolecular structures include?

Large-scale efforts

- Systematic optimization of all structures in the PDB
- Redetermination of groups of related structures
- Redetermination of groups of structures focusing on specific questions

Small-scale efforts

- Optimization of structures that are of interest to specific investigators

Terwilliger T. & Bricogne G. (2014). Acta Cryst. D70, 2533-2543.

What is needed for successful continuous improvement of macromolecular structures?

Validated processed data and metadata

- Data and considerable metadata are already collected by the PDB
- Ideal would be for deposited data and metadata to be validated by full re-determination of the structure

Unmerged data

- Raw diffraction data
- Raw diffraction images
- Multiple X-ray wavelengths
- Data from heavy-atom derivatives

What is needed for successful continuous improvement of macromolecular structures? (cont...)

Multiple views of the PDB

- The views of the archive of structures presented by the PDB can have a major influence on biological interpretation of structure
- The model that is appropriate may vary because our models are not full descriptions of what is in the crystals
- The appropriate model will depend on the question being asked (What is the distance between these atoms? What is the buried surface area? What is the electrostatic potential? What is the difference between these two structures?)

A way to choose the right model for the question being asked

- Continuous improvement will need be accompanied by a mechanism for retrieving models appropriate for a variety of uses
- Choice of model might include validation metrics, resolution, method used for determining model, model parameterization

What is happening already?

Progress towards versioning in the PDB

- Discussion of the idea of maintaining a single PDB identifier with versions
- The data will then be fixed and associated with a single PDB code
- New versions will all relate to the same data

Validation of data in the PDB

- Each class of data in the PDB (X-ray, EM, etc) has a Validation Task Force
- The PDB is developing tools to validate structures
- The same information can be used as a basis for serving up the most appropriate structure for each purpose

Deposition of raw images making extensive reanalysis possible

- SBGrid Data Bank (http://data.sbgrid.org/) to deposit raw diffraction data
- Wladek Minor's group is making images available (http://www.proteindiffraction.org/; see this workshop)