Crystallization Conditions in the Protein Data Bank

Deborah Harrus[1]

[1]EMBL-EBI

dharrus@ebi.ac.uk

When solving a protein structure using x-ray crystallography, getting a crystal is the major bottleneck of the pipeline. Unlike single component solutions for which it is possible to draw a phase diagram, the parameters affecting the crystallization of a protein solution are numerous, and we can only ever access an empirical protein crystallization phase diagram. Therefore, the workflow for obtaining a protein crystal mainly relies on a trial-and-error approach.

On one hand, many aspects of the x-ray crystallography technique are constantly improved: easier sample preparation and protein purification, better x-ray sources and detectors, robots for mounting crystals under the beam, streamlined data processing and validation, etc. On the other hand, the improvements around the crystallogenesis step focus on easing the trial-and-error approach: robots handling nanoscale volumes requiring less sample, diversification of the commercially available crystallization kits allowing a larger number of conditions to be tested, automated photography of the drops, image recognition of the drops' content, etc.

The Protein Data Bank (PDB) currently holds over 150,000 records of crystallization conditions. Providing the crystallization conditions is mandatory when depositing a structure solved using x-ray crystallography. The deposition interface currently allows the depositors to enter their crystallization details in a free-text field. The capture of this data is therefore made easy and quick, since the depositors can simply copy/paste one sentence or an excerpt of their material & methods text from their article.

Nevertheless, this data capture method has serious drawbacks. Firstly, it may not be clear what level of details is expected. For example, only some depositors provide details regarding the protein buffer, or the solution used for the crystal's cryo-protection. A lot of information around the crystal growth may thereby not be captured.

Additionally, the crystallization conditions data are currently captured and stored as a single string of text in the PDBx/mmCIF item _exptl_crystal_grow.pdbx_details. This makes it impossible to search the PDB for crystallization details, or do statistics or comparisons of this data from one entry to another.

By improving the capture of crystallization conditions at deposition using detailed PDBx/mmCIF categories, the data would be made both consistent and searchable. This could provide data to develop better machine learning methods to predict the crystallization zone for a protein of interest, therefore reducing the trial-and-error efforts. It could also allow exploring the physico-chemistry behind the crystallogenesis of proteins.

We will present our exploratory work on improving the data capture of crystallization conditions. Our main goals are:

- To establish a dictionary of chemical compounds, and to set a controlled vocabulary regarding all aspects of sample preparation for crystallization.

- To design new PDBx/mmCIF categories able to store all the details around the crystallization experiment, in collaboration with crystallization facilities.