Data sharing comes to structural biology

New archives for raw X-ray crystallography and cryo-EM data will accelerate progress in structural biology.

The accepted community standard in structural biology is that authors of a paper describing a 3D macromolecular structure must submit model coordinates to the Protein Data Bank (PDB) and provide the accession code. A structure model, however, is just that—an investigator's interpretation of the experimental data. Although the PDB systematically performs some validation checks for model quality, the only way to ensure total transparency is to make the raw experimental data available.

Access to raw data is also essential for training and testing new software tools for data processing and analysis, for developing metrics to assess the quality of results and for teaching the next generation of scientists. When raw data are available to peer reviewers, potentially embarrassing mistakes can often be caught before papers are published. Sharing raw data expedites scientific progress: other groups can reprocess raw data sets with new software tools to generate new insights, while avoiding the redundancy of generating data on the very same system.

As editors, and as editors of a methods journal in particular, we commend the development of two new archives, now open to the structural biology community, hosting the raw data underlying 3D macromolecular structure models. The Structural Biology Data Grid (SBDG; https://data.sbgrid.org) is an archive mainly for X-ray diffraction image data (as well as for a few other data types) supporting structures in journal publications (Meyer *et al.*, 2016). EMPIAR (http://www.ebi.ac.uk/pdbe/emdb/empiar/), an archive for electron microscopy (EM) image data supporting structures in the Electron Microscopy Data Bank (EMDB), is described on page 387 of this issue.

Fundamentally, raw data sets from both X-ray crystallography and cryo-EM used for high-resolution macromolecular structure determination consist of series of 2D images. A typical X-ray diffraction data set is about 5 gigabytes, and a large cryo-EM data set could top out at an astounding 10 terabytes. It is only in the past few years, especially with the advent of inexpensive cloud storage systems, that hosting and sharing data sets of such enormity has even been technically realistic.

But the advances are not just technical; they are also cultural, part of a movement across scientific fields to enhance transparency and reproducibility in research. As noted in our Method of the Year 2015 feature, many in the rapidly growing cryo-EM field, concerned about the quality of published work, have been calling for making raw data available. Nevertheless, given the competitive nature of this field, there are understandably dissenters on this position who worry about releasing their precious data sets into the wild to be potentially exploited by other groups that put no time, effort or money into their generation.

For a raw data archive to be useful, submitting and retrieving data sets should be as painless as possible, and the archive must have the capacity for growth. EMPIAR, a project of the PDB in Europe (PDBe), aims to make uploading terabyte-sized EM data sets a single-click operation. Its present capacity is in the petabyte range, but its growth must be managed, according to the EMPIAR staff. Access to data sets in the SBDG is facilitated by the Data Access Alliance, a voluntary organization of data-storage providers with diverse funding sources. Data sets are replicated in centers located in the United States, Sweden, China and Uruguay, which helps minimize data loss and enables local access to large data sets that can be challenging to download. The hope is that this data-grid model will be more sustainable in the long term than a traditional repository.

To ensure that these archives will be funded far into the future, they need support from the structural biology community—not just through the deposition of raw data, but through demonstrations that show how such resources propel the field forward. EMPIAR and the SBDG join the established Biological Magnetic Resonance Data Bank (BMRB; http://www.bmrb.wisc.edu/), which for more than two decades has hosted biomolecular NMR spectral data, mostly in the form of machine-readable tables of chemical shifts (though it also has the capacity to host raw NMR spectra). Though it has been a decisively important resource for the NMR community, it has not been without funding challenges, as detailed in a 2012 special issue of *Nature Structural & Molecular Biology*. Funders must devise long-term solutions for repositories with proven impact on scientific fields.

Despite our enthusiasm and support for SBDG, EMPIAR and other data-sharing efforts, we are not mandating raw X-ray or cryo-EM data deposition at this time (our current requirements are here). We need first to be sure that a new data archive is reasonably stable, is nearly painless to use, has the capacity to handle an influx of depositions and, most important, is serving the needs and desires of the research community well. We will continue our discussions on this matter within Nature Research Group in collaboration with the structural biology community, and we welcome feedback from our readers.