MS053.O05

Macromolecular diffraction data fit for archiving

Andreas Foerster[1], Markus Mathes[1], Stefan Brandstetter[1], Clemens Schulze-Briese[1]
[1]*DECTRIS Ltd., Baden-daettwil, Switzerland*
E-mail: andreas.foerster@dectris.com

Macromolecular crystallography (MX) has always been at the cutting edge of data sharing and preservation. The Protein Data Bank was established in 1971 to archive structural models and make them available to the community. Submission of atomic coordinates to the PDB was made mandatory by most journals by the 1990s. Since the 2000s, the submission of structure factors, data one step closer to the experiment, is similarly required by most journals. Now, the idea of archiving the raw diffraction data is gaining ground [1].

Archiving of raw diffraction data faces two substantial challenges, disk space and annotation. With higher-powered synchrotrons and faster detectors, the rate of produced data is higher than ever before. Efficient lossless compression is critical for making the storage of raw data practicable and acceptable, but this is not enough to make it useful. To retrieve datasets and be able to process them, clearly defined metadata need to be associated with each dataset that describe instrument and experiment. Without sufficient metadata, raw data are just a huge pile of bytes.

Compression is always a trade-off between speed and compression ratio. Algorithms providing compression ratios close to what is limited by the information content of the data tend to be slow. One example is bzip2. Fast algorithms tend to yield poor compression. I will compare various compression algorithms and assess their appropriateness to MX data in a high-throughput data collection pipeline. LZ4 compression, optionally preceded by bit-shuffling (bs-LZ4), offers an excellent compromise between speed and compression ratio.

EIGER detectors are increasingly used at MX beamlines. Use of bs-LZ4 ensures high compression rates with no observable penalty in read or write times. EIGER takes a dataset-centric approach, writing data into HDF5 containers instead of individual images to files. Metadata are written according to the NeXus standard [2].

While the metadata known to the detector software at the point of data collection are saved with the data, they do not form a complete description of the experiment. They do not suffice for data processing and are thus not fit for archiving. I will discuss additional pieces of metadata that need to be added before the data can be archived. These metadata terms are specified in the NeXus format, and simple tools exist for their inclusion in existing datasets. Beamline scientists must make sure to hand their users only data containing these additional metadata items.

With fourth generation synchrotrons and faster detectors that record more pixels per image, the amount of raw data generated by X-ray crystallography is expected to increase dramatically in the next few years. The efforts to make these data available to the wider scientific community hinge on fast and effective compression and comprehensive metadata. EIGER provides both excellent compression and a mechanism of recording complete metadata, making it possible to exploit the full power of latest-generation light sources.

[1] Kroon-Batenburg, L. M. J. et al. (2017). IUCr J, 4, 1-13
[2] http://download.nexusformat.org/sphinx/classes/index.html

Keywords: raw data, metadata, compression