MONITORING AND CATALOGUING THE PROGRESS OF SYNCHROTRON EXPERIMENTS, DATA REDUCTION, AND DATA ANALYSIS AT DIAMOND LIGHT SOURCE FROM A USER’S PERSPECTIVE


Abstract
The high data rates produced by the latest generation of detectors, more efficient sample handling hardware and ever more remote users of the beamlines at Diamond Light Source require advanced data reduction and data analysis tools to maximize the potential benefit to scientists. Here we describe some of our experiment data reduction and analysis steps, including real time image analysis with DIALS, our Fast DP and xia2-based data reduction pipelines, and Fast EP phasing and DIMPLE difference map calculation pipelines that aim to rapidly provide feedback about the recently completed measurements. SynchWeb, an interface to an open source laboratory information management system called ISPyB (co-developed at Diamond and the ESRF), provides a modern, flexible framework for managing samples and visualizing the data from all of these experiments and analyses, including plots, images, and tables of the analysed and reduced data, as well as showing experimental metadata, sample information.

INTRODUCTION
The availability of fast, continuous readout detectors (e.g. Pilatus3, Dectris[1]), highly automated sample handling systems such as BART robots[2], and intense X-ray beams at synchrotron sources have been crucial in increasing the rate of sample throughput and diffraction data collection at modern macromolecular crystallography (MX) beamlines. Non-automated evaluation of the many thousands of diffraction images that are generated per measurement in such experiments is no longer feasible, necessitating automated data reduction and analysis techniques combined with a laboratory information management system (LIMS) (SynchWeb [3]) that provides the user with the information necessary to evaluate the progress of their experiment while at the beamline.

SAMPLE EVALUATION
A typical macromolecular crystallography experiment on a beamline begins with visually locating a single crystal or using grid scanning[4] to locate samples or subsamples whose diffracting ability are assessed directly. For experiments where heavy atoms are incorporated for phasing purposes[5], a fluorescence scan may be performed to determine the optimal parameters for anomalous data collections.

STRATEGY
To determine the data required for an experiment, a screening data collection is typically performed with images at 0, 45, and 90 degrees. EDNA[6], a Python pipeline wrapping underlying software, is used on a multi-core computing cluster to calculate optimal data collections for several situations including native, anomalous, gentle (for radiation-sensitive samples), single anomalous diffraction (SAD), and considering measured flux. If applicable to the beamline, kappa alignment parameters (using XOAlign[7]) may also be calculated. Results of the strategy calculations are available in SynchWeb (Figure 1) and the desired EDNA strategy may also be retrieved when setting up a data collection using the Generic Data Acquisition software used at Diamond (GDA) [8].
Figure 1: SynchWeb display of alignment (XOAlign[7]) and strategies (EDNA[6] and Mosflm[9]) of thermolysin.

Figure 2: A normal crystal has no decrease in number of spots or worsening resolution after thousands of collected images (left), radiation damage causes a reduction in spots over time (center), and mis-centered crystals cause a region with no diffraction (right). In all images, numbers of diffraction spots in an image are indicated in yellow (all spots) or blue (Bragg diffraction spots), while red indicates resolution of the image.

PER-IMAGE DATA ANALYSIS

While most data analysis occurs after the data collection has finished and the data has been reduced, a running plot of the observed diffraction of the current sample can be monitored by analysing a subset of the diffraction images. By applying spot finding algorithms in DIALS [10] to 200-500 diffraction images distributed uniformly throughout the data set, data collection pathologies may be detected by the user by monitoring the trends during the data collection. Examples of a plot from a normal crystal and some pathological cases are shown in Figure 2, and include radiation damage (decrease in the number of spots visible after e.g. 360 degrees of exposure to the X-ray beam) and crystal miscentering (there is diffraction at the start and end of an experiment, but diffraction completely disappears during the experiment as the sample is no longer illuminated).

DATA REDUCTION

As the raw diffraction images are too numerous to manually inspect, data quality measures derived from data reduction are typically the better method of assessing the quality of the resulting data. As soon as the data collection has completed, data reduction is triggered to run the software packages Fast DP [11] and xia2 [12] on multi-core computing clusters.

Fast DP runs XDS [13] for the majority of its initial indexing and multiple integration steps. Pointless [14] and Aimless [15] are used to determine the correct point group and merge the data. Fast DP is designed to run
quickly and can complete reduction of a single dataset of up to 3600 images from Pilatus3-6M detectors within a minute depending on data quality.

xia2 has been developed as the more thorough data reduction system. xia2 runs multiple data reduction processes with the possibility to run several different software packages simultaneously. xia2 can be run utilizing Labelit [16] to index using 3 images (3d mode) or XDS and all frames for indexing (3dii mode). xia2 may also be run with DIALS being used for indexing (dials mode). Results from xia2 are typically available within 10 minutes of starting the data reduction process.

Results from all of the data reduction processes that complete successfully, including space group, unit cell parameters, completeness, resolution, signal-to-noise ratio and the reduced datasets, are stored in the ISPyB database and may be viewed by using the SynchWeb page.

Initial conditions to performing this analysis include high data completeness (> 80%), significant differences in anomalous pairs (\(dI/\sigma(dI) \geq 1\) if data extend to equal or worse than 2A resolution, or \(dI/\sigma(dI) \geq 0.8\) if data extend to better than 2A resolution). If these conditions are met, Fast EP is run [11]. This software pipeline consists of SHELXC, D, and E [17] run on a computing cluster with multiple possible values of spacegroup and number of sites are tested first in SHELXD, with the best atom positions used for phasing and solvent flattening with multiple solvent fractions in SHELXE. The best results are then used to calculate electron density maps. Putative heavy atom locations and occupancies, figure of merit, correlation coefficients, and electron density maps may all be viewed in the SynchWeb interface.

**DATA ANALYSIS – DIFFERENCE MAPS**

To determine whether ligands are bound to proteins, users may provide an atomic coordinate file (PDB format[18]) that can be used for difference map calculation performed by DIMPLE [11]. The coordinates that best match the reduced data point group and unit cell parameters undergoes REFMAC5 [19] rigid body refinement. Further Phaser [20] molecular replacement may be performed if the initial model was not similar enough in orientation to the correct structure. An image of the largest region of difference map density with the underlying molecular model is rendered and is shown in SynchWeb. (Figure 4)

**CONCLUSION AND EXTENSION TO OTHER SCIENTIFIC METHODS**

Improvements in all phases of Sample Handling, Data Collection, Reduction, and Analysis for MX beamlines have resulted in doubling of samples handled per shift between 2010 and 2015 on beamline I03 at Diamond (Katherine McAuley, personal communication).

Other domains of science are experiencing similar advances that will require better and more automated methods of analysis so that users are able to interpret...
experiments. The underlying ISPyB database schema used by SynchWeb has been made more generic to allow sensible storage of data from these scientific domains. Two particular sciences where SynchWeb is currently used at Diamond for experiment evaluation are tomography and powder diffraction (Figure 5).

ACKNOWLEDGMENT

Thanks to Patrick Collins for the miscentered crystal per-image-analysis image and Data Reduction SynchWeb image, Scientific Software for implementing and improving the software pipelines, IT Support for maintaining the cluster computing systems, and CCP4 for providing the effort for supporting and improving Dimple.

REFERENCES