

MS41 Automation in data collection and processing

MS41-04

Data reduction in protein crystallography

M. Galchenkova¹, **A. Tolstikova**², **O. Yefanov**¹, **H. Chapman**¹

¹DESY, CFEL - Hamburg (Germany), ²DESY - Hamburg (Germany)

Abstract

Area X-ray detectors became bigger (having more megapixels) and faster (measuring more frames per second). This allows to measure dynamical processes in protein crystals with high resolution and below 1ms time scale. The price to pay is the amount of data that such detectors generate. Unfortunately, storage volume is growing much slower. Therefore, there is an increasing gap between the data generation volume and the possibility to store it.

One of the most disk consuming type of experiments is macromolecular crystallography (MX) and especially serial crystallography (SX). Such experiments often require many Mpix detectors and usually can be done with very high speed – modern facilities produce enough photons to measure at 1kHz rate. Unfortunately, the lossless compression rate of such diffraction patterns is rather poor due to the high background. In standard MX experiments at synchrotrons, due to the well-established processing pipeline, only the averaged intensities of the Bragg peaks are kept. Unfortunately for FELs and synchrotron SX experiments such luxury is not possible yet – reprocessing of raw data can improve the result a lot. Recently we have shown that reprocessing of the data measured 10 years ago at LCLS led to resolution improvement from 3.5Å to 2.5Å.

We have tested different approaches for SX data reduction: compressing with different lossless algorithms, binning the data, saving only hits, quantization, etc. Data from different experiments at synchrotrons and FELs with various detectors and different samples were used. Checking the resulting statistics of compressed data (like CC*/Rsplit, Rfree/Rwork, anomalous signal) we have demonstrated that the volume of the measured data can be greatly reduced (10-100 times!) while the quality of the resulting data was kept almost constant. Some compression strategies, tested on SX and MX datasets, can be applicable to other type of experiments.