Lossy compression on Crystallography images

Utsab Khakurel[1], Chun Hong Yoon[2]

[1]Department of Electrical Engineering and Computer Science, Howard University, 2400 Sixth St NW, Washington, DC 20059, USA

[2]Linac Coherent Light Source, SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025, USA.

[+]Contact: yoon82@slac.stanford.edu utsab.khakurel@bison.howard.edu

Introduction

There are millions of diffracted crystallography images obtained from LCLS at SLAC. They end up taking a lot of space and consequently cost a lot of money. Compressing these images is a need for LCLS-II. Lossless compression doesn't save a lot of memory so lossy compression is the better option.
The goal of this project is to perform lossy compression on diffraction image files and still preserve the science behind the experiment.

Keywords: compression, LCLS, crystallography, images, memory

Research

This compression project was built on the experiments run on the Streptavidin crystals.

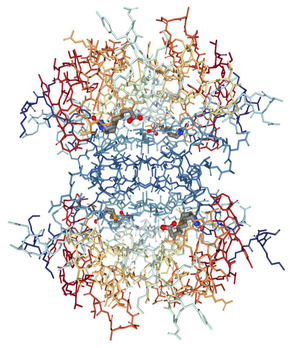

Figure 1: SFX structure of corestreptavidin-selenobiotin complex

The scattering of X-rays from atoms produces a diffraction pattern, which contains information about the atomic arrangement within the crystal.
Every images received from the detector pad are not the diffracted images.

There are three primary steps involved to test if the compressed image files have the data we need or not, which are explained further in next section.

Process Overview:

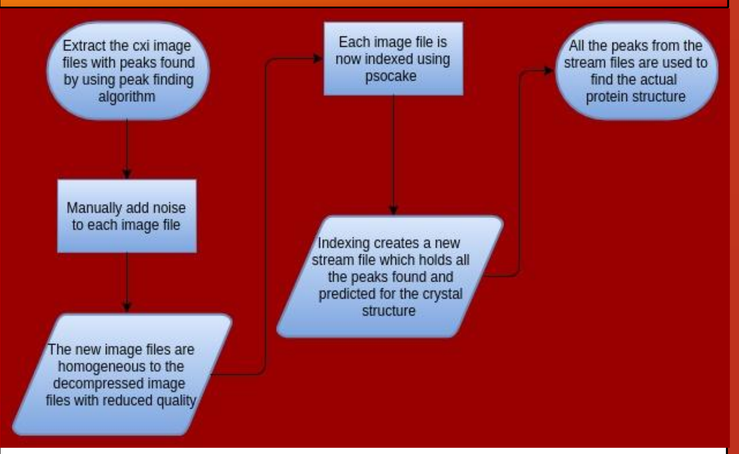

Figure 2: Flowchart representing the phases of obtaining the crystal structure

Step 1: Noise addition

The interference of scattered diffracted lights results on a peak formation. The position of the diffraction peaks are determined by the distance between parallel planes of atoms.

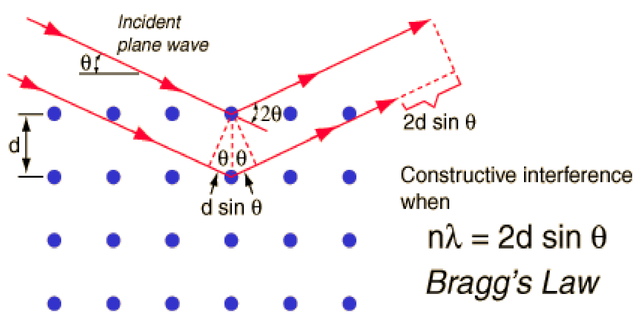

Figure 3: Bragg's Law defines the condition for the peak formation

The images from the detector pad are supposed to be compressed by a software eventually but meanwhile for the project, the images are extracted from the source directory and 30 ADU uniform random noise is being added to each image file manually with python script. The noise added image files are homogeneous to the decompressed image files we obtain from the lossy compressed ones. Noise is added to the peaks in the diffracted images.

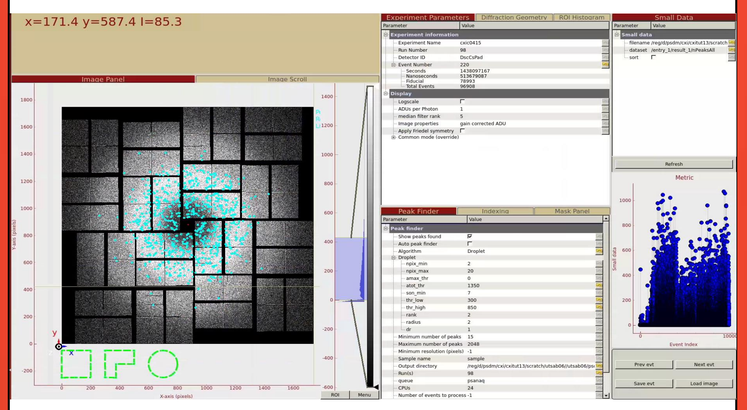

Figure 4: Peak finding using psocake on experiment run98, event 220 before noise addition

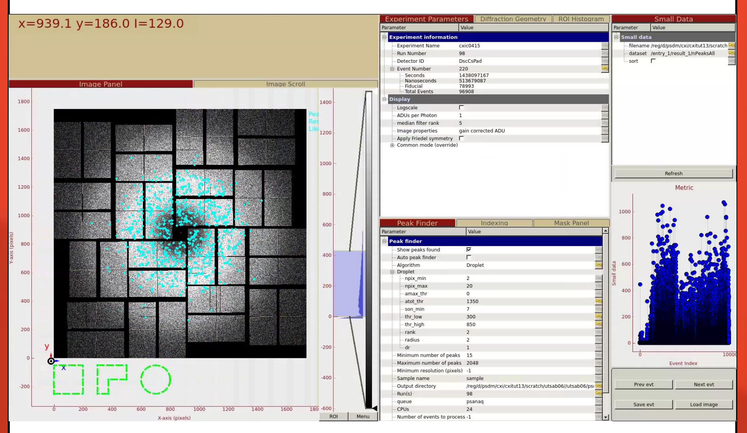

Figure 5: Peak finding using psocake on experiment run98, event 220 after noise addition

Step 2: Indexing

The peaks obtained from the new noise added image files will be used to predict more coordinates of the crystal lattice. This step is carried out to measure the molecular transform of the streptavidin crystal. After indexing, a stream file is created which contains all the coordinates of the peaks found and predicted.

Step 3: Phasing

Stream file containing all the observed and predicted coordinates is now used to calculate the coordinate of the 4 selenium atoms inside unit cell in streptavidin. Furthermore, these coordinates are used to solve the structure of a protein. The software used for this process are HKL2MAP, CCP4, PHENIX and COOT.

Before refinement begins, about 10% of the experimental observations are removed from the data set. Then, refinement is performed using the remaining 90%. The R-free value is then calculated by seeing how well the model predicts the 10% that were not used in refinement. Typical values are about 0.20.

$$R = \sum ||Fobs| - |Fcalc|| / \sum |Fobs|$$

Figure 6: Formula for R-free value where F(obs) is the observed dataset and F(calc) is the calculated dataset using the selected observed ones

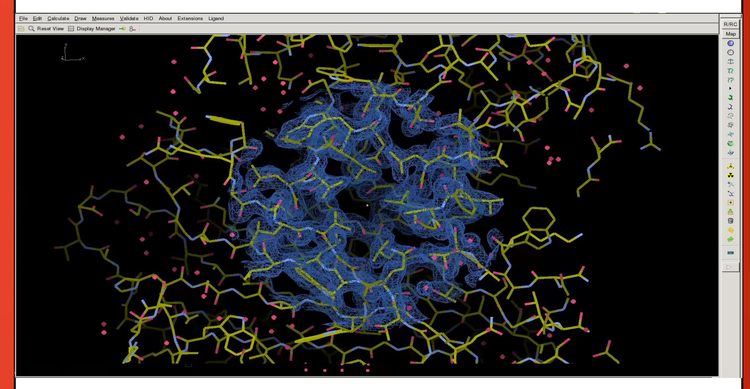

Figure 7: The known crystal structure(stick) is overlaid with the electron density(blue).

Conclusions

The indexing rate for the crystal decreased a lot after noise addition which is quite puzzling, hence the R-free value we obtained is close to 0.5 which is not a good value. But with more repetitions and further parameter tuning better results are certainly possible.

Acknowledgments

Use of the Linac Coherent Light Source (LCLS), SLAC National Accelerator Laboratory, is supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under Contract No. DE-AC02-76SF00515.

Date: 08/09/2017