Desy Summer Project

Maria Pakhnova, Moscow Institute of Physics and Technology, Dolgoprudny

group: FS-CFEL-1 supervisor: Yefanov Alexander

September 7, 2018

Abstract

Today there is a huge interest for observation three-dimensional structure of biological macromolecules. Scientists from all of the world are working on the addition of databases on the structure of different proteins. Fortunately, the existing technologies, based on synchrotron radiation, allows us to achieve the necessary resolution and observe tiny details. This project is devoted to identification the structure of protein Phycocyanin and to testing several indexing methods that are used to process the diffraction patterns.

Contents

1 Introduction

I was working on two different projects. The first project was connected with the study of the protein phycocyanin and with the processing of diffraction images obtained on PETRA III. The main aims of this project were

- reveal different structures and symmetries depending on crystallization conditions

- discover new interesting structures and save the conditions for further investigation

- find pattern and condition that corresponds the highest resolution

Phycocyanin is a rare natural pigment, existing in cyanobacteria. It is a blue powder. This protein is widely used in food, cosmetic, chemical industry and in pharmacology, as it has anti-inflammatory properties. It is a very rare pigment-protein, which has fluorescent properties. Under the influence of light can produce a strong red glow. That's why it is used in the diagnosis of cancer.
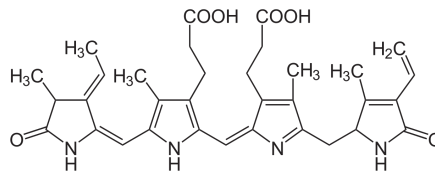
Figure 1: Chemical structure of phycocyanobilin

My second project was connected with comparing different indexing methods that were used in CrystFEL suite of programs. (asdf, mosflm, xds, dirax, xgandalf) And reveal the best ones that suits the particular situation.

2 Theory

Synchrotron radiation is electromagnetic radiation emitted by relativistic charged particles (electrons, positrons, protons) as they travel along curved paths under the action of a constant magnetic field. Synchrotron radiation is produced in storage rings where electrons or positrons are kept circulating at constant energy. In a storage ring the synchrotron radiation is produced either in the bending magnets needed to keep the electrons in a closed orbit, or in another devices such as wigglers or undulators situated in the straight sections of the storage ring. In these devices an alternating magnetic field forces the electrons to follow oscillating paths rather than moving in a straight line. As the result they emit radiation.

$$I \sim a^2, \tag{1}$$

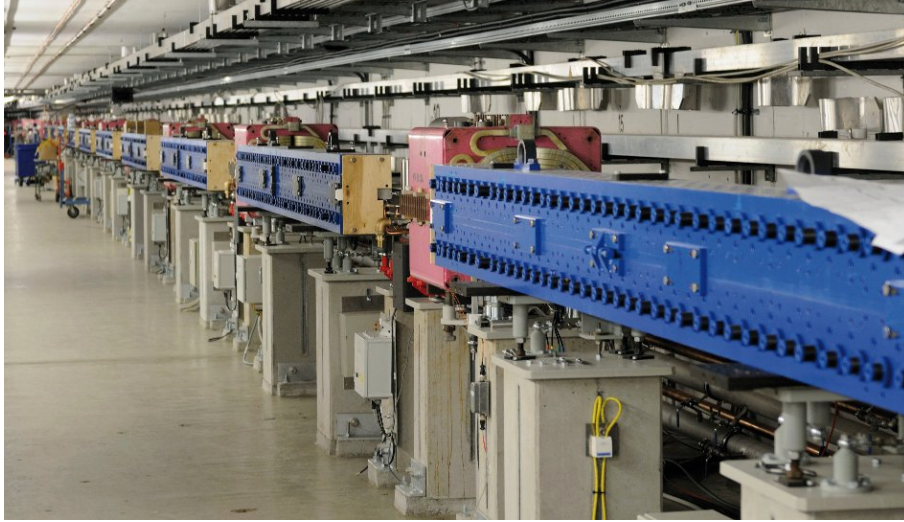I - intensity of radiation, a - acceleration of the particle

Figure 2: Damping wiggler in PETRA III storage ring

1) For a relativistic electron, virtually all radiation is directed along its velocity and concentrated in a cone with a complete solution. Hence the synchrotron radiation is highly collimated.

$$I \sim 2\gamma, (10^{-4} - 10^{-3} radian) \tag{2}$$

2) The second property of synchrotron radiation associated with the Doppler effect is its spectral distribution. The main part of the emitted energy is shifted to higher frequencies, and the maximum falls on the frequency, corresponding to the X-ray radiation.

$$\omega_{max} \sim \omega_c \gamma^3, \tag{3}$$

ω_c - cyclotron frequency of electron rotation in a circular orbit

Diffraction of scattered X-rays as they pass through the crystal simplistically can be thought of as the reflection of rays from the crystallographic planes of translucent and the subsequent interference of the reflected waves. Condition of positive interference of rays, reflected from adjacent planes, corresponds to the well-known Bragg-wolf formula:

$$2dsin\theta = n\lambda, \tag{4}$$

d - interlayer distance, λ - wavelength
Maximum wavelength at which diffraction is possible:

$$d \geq \lambda/2, \tag{5}$$

To fulfil the diffraction condition, the node of inverse lattice must fall on the sphere of radius $1/\lambda$, called the Ewald sphere or the sphere of reflection. If the inverse lattice is rotated relative to the Ewald sphere, more nodes of the inverse lattice will fall into the

reflecting position. In practice, this means that the crystal must be rotated to obtain a complete diffraction pattern.

Finally, directing the beam of radiation at different angles to the investigated protein, we get diffraction patterns. They can be used to find the lattice parameters and restore the structure of crystallized protein.
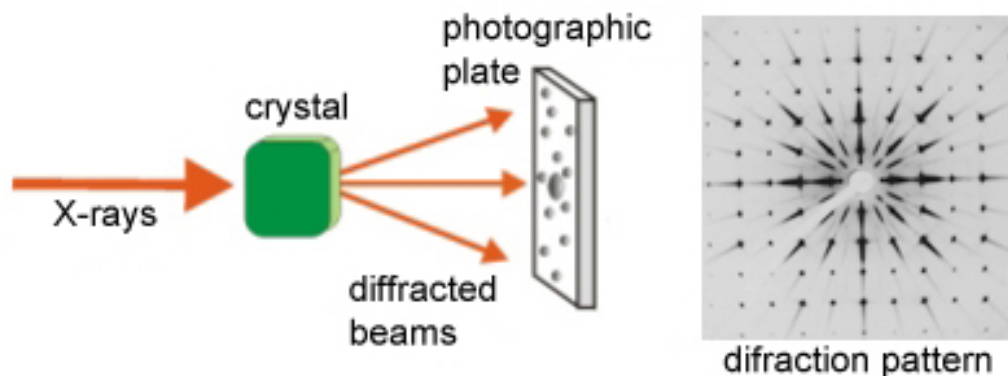

Figure 3: X-ray diffraction on a crystal

3 Measurements

Firstly, samples were prepared for the measurement. They were crystallized. Individual protein molecules packed in a repeating array, held together by non covalent interactions. Such crystals are convenient for the molecular structure by using X-ray diffraction. Crystallisation is a rather complicated technique. At all we got nearly 100 samples, that were crystallized under different conditions.
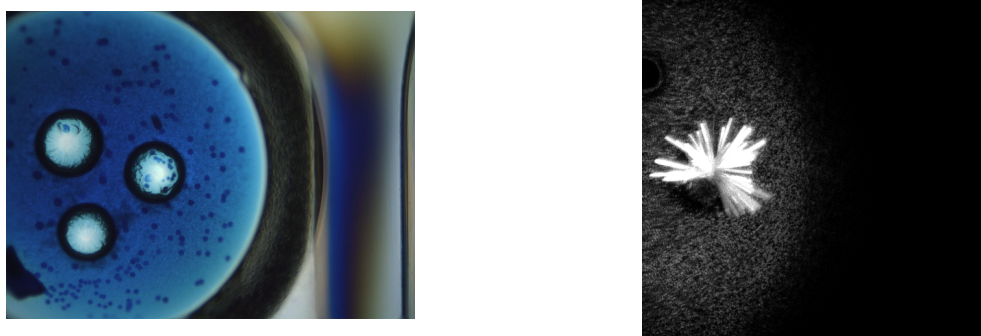

Figure 4: Frozen crystals, the cubic crystals look more roundish

Then using a unique synchrotron light source - PETRA III we collected diffraction frames in the reciprocal space for each sample. In result we received about 300 pictures to be
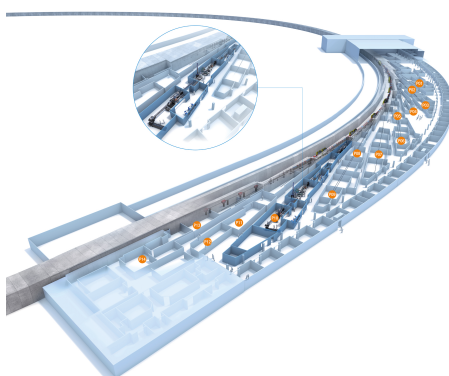
5

processed.

Figure 5: Experimental stations at the accelerator PETRA III

4 Software

The following software packages were used to work with a huge amount of data.

4.1 X-ray Detector Software

XDS processes single-crystal monochromatic diffraction data recorded by the rotation method. It's a good program with an intuitive interface. It gains a lot of .cbf files - crystal images obtained at different angles. As a result it determines the space group number and unit cell parameters of each pattern.

At the beginning of our work it was necessary to ensure that the detector parameters are correct. During the experiment distance between the detector and crystal could be changed several times. There were certain ways to overcome this problem: take pictures of a known sample (silicon), use ADXV program or begin working with the most successful frames. With ADXV (a program to graphically display and analyse 2D area detector data) we found the sample, that had the best resolution, and carried out further settings using it for other crystals. Samples with other initial parameters of the experiment were processed manually. So we got several results:

- The best resolution was about 1.2 A. (ph-level 6.5; 0.5% 40% v/v Ethylene glycol; 20 % w/v; 0.1 M Imidazole; MES; Monohydrate acid; 0.25% 3% w/v Sodium-L ascorbate, 3% w/v Choline Chloride, 3% v/v D-Panthenol, 3% w/v Pyridoxine hydrochloride, 3% w/v Thiamine hydrochloride)

- The resolution doesn't really depend on the ph-level of the mix.

- Most of the samples had Monoclinic and Trigonal structures, but there were found several Hexagonal (resolution = 2.3 A) and cubic(resolution = 7.4 A) examples. These unexpected samples will be carefully measured and examined again

The final table detailing the characteristics of each processed sample, including precipitants and the pH-level of mixture, is attached in Finaltable.xlsx.

4.2 CrystFEL

CrystFEL is a suite of programs for processing diffraction data. CrystFEL comprises programs for indexing and integrating diffraction patterns, scaling and merging intensities, simulating patterns, calculating figures of merit for the data and visualising the results.

The results obtained by XDS were verified by CrystFEL. At the end, two variants of processing diffraction patterns gave the same result, which confirms the accuracy of findings.
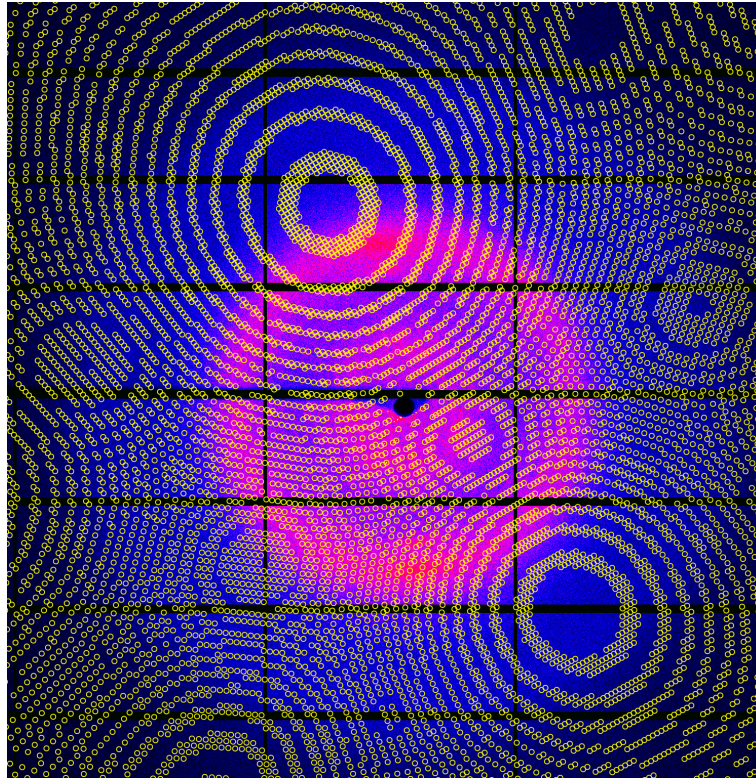

Figure 6: Checking indexed by CrystFEL and real peaks (cubic structure, input data from XDS iteration)

5 Testing indexers

My second project was connected with testing various indexing method and identifying the most optimal in different situations. We worked with several types of indexers: mosflm, dirax, asdf, xds, xgandalf. We paid attention to the following parameters of the each indexer:

- time of indexing

- correctness of indexing

- the minimum number of Bragg peaks that are needed to index the sample and reveal its parameters

- ability of working with different types of noise, how it influence the correctness of indexing

This project is rather big, it needs to be continued. At the moment, several tasks have already been performed. Here you can see the list of programs, that were written to cope with the following challenges.

Table 1: Python programs and bash scripts for testing indexers

code name	input/output parameters	comments
list_maker.sh	data files/list of data files	creates a list of file names to process
h5py_display.py	h5 file/data and groups	shows the specific structure of h5 file
add_few_peaks_groups.py	h5 file/.h5 file	creates datasets that has from 5 to 50 brag peaks closest to the centre of the detector
random_add_few_peaks _groups.py	h5 file/.h5 file	creates datasets that has from 5 to 50 brag peaks chosen randomly
run_all_indexers.py	file list/stream	automatically runs all indexers
StreamFileParser.py	stream/stream	gets information from the stream file
streamViewer.py	stream/photo	returns photo of indexed Bragg peaks
keep_peak_interval _indexed.py	stream/list	creates a list of files that have from 50 to 150 indexed crystals
rotation_angle.py	2 matrix/angle	gives the angle of rotation between the two matrix of crystal location
equivalentBasis.py	matrix/matrix	rotates the matrix arrangement of the crystal so that the normals were pointing in the same direction
create_comparison_plots.py	stream/plots	return plots that display the quality of indexing methods

It is necessary to mention that the main problem was to learn how remove Bragg peaks from .h5 file observing the correct frame of reference. Firstly, we tried to create new datasets with the correct amount of data in old files using peaks from h5 files.(peakinfo-raw) But sometimes the location of the selected peaks fell on the detector, so they did not participate in indexing. That's why we firstly indexed all peaks and then deleted peaks from recognised ones from stream.

The processed data was taken from CrystFEL tutorial. However, all programs are universal, they can work with any other data files. Firstly, we created the list of files and launched indexamajig. Then we chose the most successful files, that had from 50 to 150 found peaks. Then we created new datasets in h5-files, that consisted of 5 to 50 Bragg peaks (leave only peaks that are close to the detector or randomly). Ran all indexers for each dataset and displayed the results.

The correctness of indexing was checked using the visualization package streamViewer. As a result, for each stream, containing from 5 to 50 Bragg peaks, the following characteristics were counted:

- indexing rate = number of correct indexed crystals / number of chunks

- number of wrong indexed crystals (the number of cases when the matrix of the crystal arrangement in the stream containing, for example, 5 (6,7,...50) peaks was rotated by more than 2 degrees in relation to the matrix of the crystal from the stream containing all peaks)

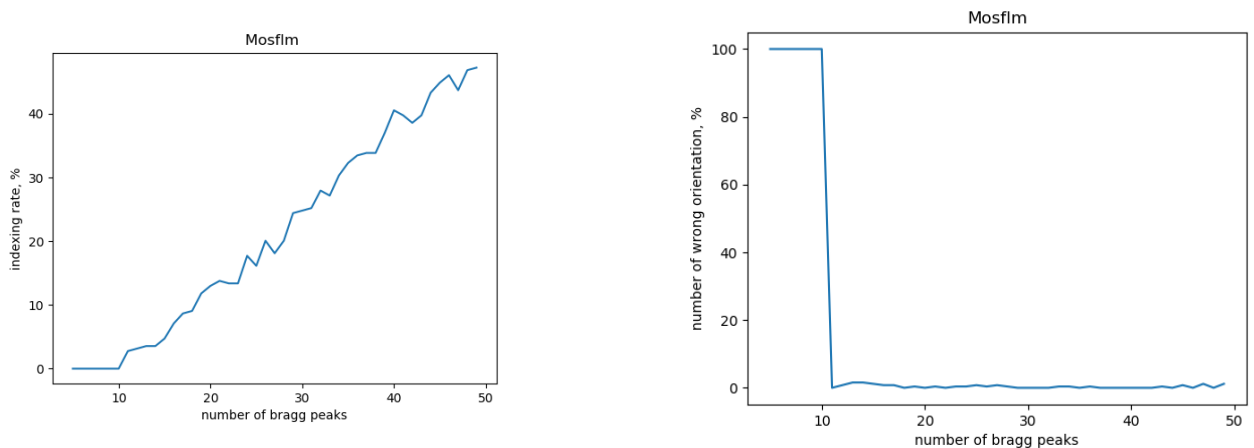The resulting plots looked something like this:

Figure 7: Characteristics of particular indexer (in this instance Mosflm)

This project is still under development and will be continued after my departure. ☹

9

6 Results

The results of this work are:

1. received on PETRA III diffraction images of Phycocyanin, that was crystallised under different conditions (personally attend)

2. wrote programs for automatic processing of a large number of files, received using XDS and CrystFEL:
   - parser files obtained in the course of working with xds
   - automatically starting indexing all files with 60 threads parallel
   - formation of Excel-table with the characteristics of each sample

3. revealed patterns with the best resolution and different structures

4. wrote a program to filter data in .h5 files

5. compared the efficiency of various indexing methods

7 Acknowledgements

References

[1] Elements of Modern X-ray Physics *Jens Als-Nielsen, Des McMorrow*

[2] What's synchrotron radiation? *MSU*

[3] CrystFEL tutorial

[4] Fundamentals of Crystallography. *C. Giacovazzo, H.L. Monaco, D. Viterbo, G.Gilli, G. Zanotti, M. C. (n.d.), (1992)*

[5] Fundamentals of X-ray diffraction *MSU Bogdan T. V.*

[6] Learning python *Mark Lutz*

[7] HDF5 documentation *Andrew Collette, the HDF Group*