LABORATORY OF BIOPHYSICS FOR ADVANCED

Experimental exercises for III year of the First cycle studies
Field: “Applications of physics in biology and medicine”
Specialization: “Molecular Biophysics”

X-RAY CRYSTALLOGRAPHY

PROTEIN STRUCTURE DETERMINATION

(ex. 36)

HUMAN CAPITAL
NATIONAL COHESION STRATEGY

UNIVERSITAS VARSOVIENSIS

EUROPEAN UNION
EUROPEAN
SOCIAL FUND

Project co-financed by European Union within European Social Fund

Content

1. Principles of X-ray crystallography

Study of the objects with electromagnetic radiation, requires radiation to have a wavelengths similar to dimensions of the object. Atoms dimensions and bond lengths are typically in the range of 1 to 3,5Å. 1Å is dimension of hydrogen atom, 2,4 – 3,5Å are lengths of hydrogen bonds, 1,3 - 1,6Å is a typical length range of C-C bonds in proteins [1]. To study objects of that size X-rays radiation has to be used.

But, the result of crystallographic experiment, is not picture of atoms, it is electron density map. This is because, the electromagnetic radiation interact with matter through its electromagnetic field. The intensity of scattered radiation is proportional to charge to mass ratio. As the electrons are few thousands times lighter than protons and atomic nuclei they interact much stronger with radiation. Moreover the velocity of electrons in the atoms is many times higher than the speed of changes of electromagnetic field of the X-ray, therefore what is observed are not single electrons, but time-averaged distribution of electron density in molecule. However, as electrons are tightly localised around the nuclei and bonds, the electron density map, gives good picture of molecule itself.

Diffraction on single molecule is extremely week, thus difficult to detect, and measure above the noise level (the scattering of water and air molecules). Crystal contains large number of molecules ordered in space, so scattered radiation is coherent in phase. Thus the constructive interference occurs, the waves amplitudes can add up and reflection intensity rise to the measurable level – crystal acts as an amplifier.

Of course, if the waves adds up in some direction, due to interference, it has to cancel out in many other directions, thus the diffraction pattern of crystal is not continuous, in fact this is an array of spots.

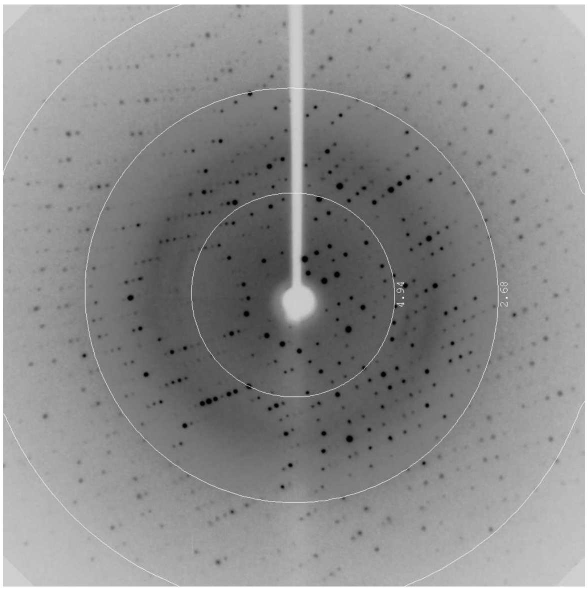


Fig. 1: Diffraction image of protein crystal

1.1. Diffraction on crystals

When a wave scatter on the electrons in the crystal scattered waves interfere with each other. Dependently on the relative distances of the electrons and the angle of incidence the waves can add up, cancel out, or something in between. Which of this will happen, depends of the total distance between source and detector. If the pathlenght of diffracted waves differ by multiple of wavelength, the waves will be in phase, they will add up, and the as the result of interference the wave will amplify; if the pathlenght differ by multiple of wavelength plus half wavelength, they will cancel out. The condition for constructive interference, can be obtain quite easily. Think of diffracted waves, like if they were diffracted from the plane passing through atoms, this plane is called Bragg's plane, and it behaves like mirror reflecting radiation.

When a parallel beam is reflected from the mirror, the incidence angle equals reflection angle, the same happens for Bragg's planes. So if the incident beam is in phase, the reflected beam is also in phase, independently where they hit the plane. Scheme given at Fig. 2 explains why.

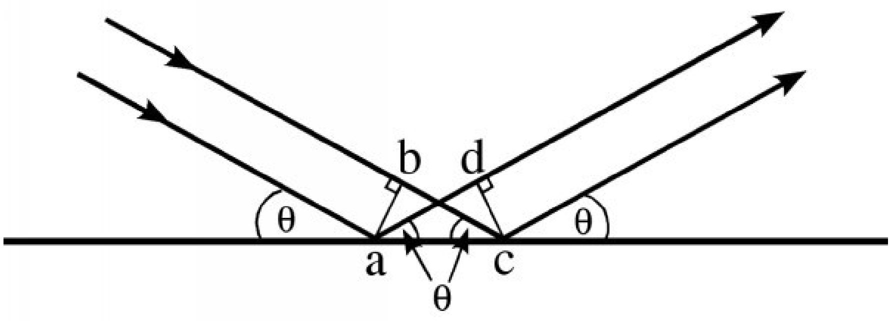


Fig. 2: Diffraction on the plane[2]

The incident beam is in phase, that means that in points a and b incident waves are in phase. Distance bc is equal to distance ad, because triangles abc and acd are congruent. Thus at points c and d waves have the same phase, and the reflected beam is in phase.

If waves scattered on plane have to have the same pathlenght to be in phase, waves scattered on different planes have to have different pathlenght, and the difference in pathlenghts has to be equal to multiple wavelength. Bragg's law gives the relationship between spacing of Bragg's planes to give constructive interference.

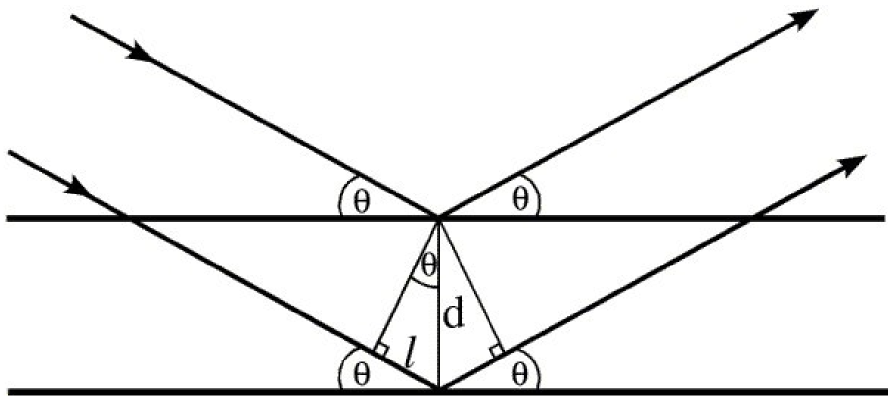


Fig. 3: Diffraction on the Bragg's planes[2]

Difference of pathlenght of waves diffracted on different planes is equal to $2l$ (fig. 3) and it depends of the incident angle $l = d \sin\theta$, thus the Bragg's law is given as:

$$n\lambda = 2d\sin\theta \qquad (1)$$

According to Bragg's law, the higher the angle of diffraction, the smaller distance between Bragg's planes has to be, to keep the pathway difference equal wavelength. That means, that the higher angle, the smallest details can be “seen” in diffraction experiment.

This inverse proportionality causes that diffraction data is usually analysed in reciprocal space. The bigger distance between objects in reciprocal space the closest they are in real space, the diffraction is reflected at higher angles, and more sensitive to smaller details.

As mentioned above, the crystal act as an amplifier thanks to interference of scattered radiation. For constructive interference to happen, scattering on all unit cells has to be in phase, thus the Bragg's planes has to go through the same points in every unit cell of the crystal.

If the objects on Bragg's plane scatter in phase, objects placed of planes scatter out of phase, and the phase shift is proportional to the distance from the Bragg plane. Thus a single diffraction experiment allows to calculate relative distances of objects from Bragg's planes. If all the objects are on the Bragg's planes, the diffraction image is one spot. If half of the objects is on the Bragg's plane, and the other half on the parallel planes exactly in half of the distance between Bragg's planes, this two sets of object will scatter out of phase, and phase shift will be 180°, so there will be destructive interference. The diffraction image will not change. If the other set of objects will be in any other distance that $d/2$ the phase shift will be somewhere between zero and 180°, the diffracted wave will add contribution to interference. On the diffraction image the intensity if the spot will change.

The relationship between diffraction image and the object the radiation scattered on is given by Fourier transform. Assuming that the electron density is a mathematical function, the diffraction image is a Fourier transform of electron density. Most of the mathematical function have their inverse function, like sinus and arc sinus, the Fourier transform is not an exception. That means, that electron density is an inverse Fourier transform. So, just calculate?[1]

Unfortunately its not that simple. To compute electron density amplitudes and phases of diffracted waves have to be known. But during experiment, the number of photons scattering on the detector are measured. This number gives intensity of scattered wave, and intensity is proportional to amplitude of the wavelength, but there is no experimental method allowing to measure phases of scattered waves. This is so called The Phase Problem[2]

1.2. The Phase Problem

Besides obtaining well diffracting protein crystal, The Phase Problem, is often second bottleneck in protein crystallography.

Understanding, why phases “disappear” during diffraction experiment, the wave theory of light is not enough, the quantum-mechanic theory is necessary. In quantum mechanic, the probability that a photon will be reflected in certain direction is given by a square of amplitude of the scattered wave.

$\Psi(x,t)=A\cdot e^{2\pi i(\nu t-x/\lambda)}$ - wave function of photon

$P_{a<x<b}=\int_a^b \Psi(x,t)\cdot\Psi^*(x,t)dx=\int_a^b |\Psi(x,t)|^2 dx$ - probability that photon happened to be in the area of $x\in(a,b)$

$|\Psi(x,t)|^2=A\cdot e^{2\pi i(\nu t-x/\lambda)}\cdot A\cdot e^{-2\pi i(\nu t-x/\lambda)}=A^2$

That's why, there is only intensity of reflected waves on the diffraction image.

As the phases can not be measured directly, they has to be obtained from indirect measurements. There are two general methods of obtaining phase: by guessing, and be disturbing structure thus disturbing diffraction image.

Heavy atoms, as they have more electrons, scatter radiation stronger. Replacement of, for example, sulphur atom in SH grup with mercury atom, changes diffraction images. Comparison of diffraction images of “native” protein and its isomorphous derivative allows obtaining some information about phases of wave scattered on the object.

This method is used in two variants:

– Multiple Isomorphous Replacement (MIR) – demands creating few derivatives with different heavy atoms

– Multiple Anomalous Dispersion (MAD) – demands creating only one isomorphous derivative, but diffraction is measured at few wavelengths.

[1] Kevin Cowtan's page: *The Interactive Structure Factor Tutorial* provides a nice look on how inverse Fourier transform works http://www.ysbl.york.ac.uk/~cowtan/sfapplet/sfintro.html

[2] Why phases are so important? Look at another page of Kevin Cowtan: *Kevin Cowtan's Book of Fourier* http://www.ysbl.york.ac.uk/~cowtan/fourier/fourier.html

“Guessing” phases demands having already solved structure of very similar protein. Calculation of diffraction image of known structure, including amplitudes and phases of scattered radiation is not a problem. Phases from that model protein are transferred as initial phases of structure in question. Then in a recursive process, the phases are adjusted to recreate measured electron density. This method is called Molecular Replacement.

1.3. Molecular replacement

Molecular Replacement (MR) is a technique using a model molecule, to solve The Phase Problem.

How to obtain a model? It's probably the simplest if a protein in question is a mutant of protein which structure was previously solved, this structure can be a model. For a wild type protein crystallised for the first time, a similar protein has to be found, a protein from the same family, or a homologous protein from another specie (for example human and bovine PNP are very similar).

If none of this can be found, a model protein can be build, starting from an aminoacid sequence. There are a lot of databases containing protein sequences (UniProt, PIR, etc) and a lot of biophysical tools, allowing search these databases for similar proteins and proteins fragments(PHYRE2, MSD, Blocks). When appropriate proteins are found, the structural databases have to be searched for three dimensional structure of the model or its fragments. MR is possible thanks to growing number of solved protein structure, in fact this is a positive feedback, the more structure is solved, the more possible templates for MR exist, so the MR can be more often used, giving rise to the number of solved structure, which can be templates for next structure to solve. At present, about 70% of structures deposited in Protein Data Bank (PDB) had been solved with MR[3].

As determining three-dimensional structure from diffraction images is problematic (The Phase Problem!), the inverse calculation is straightforward, as all the necessary data is known: coordinations of atoms and quantum-mechanical description of diffraction phenomenon. So it is possible to calculate diffraction pattern from any known structure. The diffraction pattern is usually presented as a set of structure factors. Structure factors represent waves scattered on studied molecule, so as well as waves, are described by amplitude and phase. These structure factors, with amplitudes and phases, have to be calculated for the model molecule.

The next step is to place the model into unit cell of studied crystal. That means determining position and rotation. And again, there is no analytical method to do that. All possible positions and rotations of molecule in the unit cell have to be sampled, in order to find parameters, that the best reconstruct measured diffraction pattern. To place model in the unit cell six parameters has to be figured out, three angles describing rotation and three translation vectors describing position, if there are N protein molecules in the unit cell, that means search for 6N parameters. Probing 6N-dimensional space is time consuming task, but it can be simplified, as rotation and translation are independent. The search can be divided into two searches: 3N-dimensional probing of rotational space and 3N-dimensional probing of translational space. First a rotational search is performed, the best solution or solutions are saved, and only they are used to search within translational search. In this way, the best fit is found. The structure factors phases from the model are transferred to corresponding experimental structure factors, and initial electron density map can be computed.

1.4. Electron density maps

If the atoms in crystal would be still, all the molecules in exactly the same conformation and crystal lattice – ideal, then all the waves scattered on the lattice would be in phase and the only limitation to resolution would be the radiation wavelength. But the atoms experience thermal motions, proteins are quite flexible molecules and crystal lattices have defects. Another problem is this, that diffraction of proteins is rather weak, it means, the signal to noise ratio is relatively low(the profiles of diffracted spots are wide and shallow). What's more the smaller details are to observe, the higher angle of radiation and the weaker the intensity of scattered wave. Because all of this, typical resolution of electron density maps are to low to see separated atoms, its rather electron density pipeline of atoms and bonds.

1.5. *Fitting and refining structure*

Refinement is never done, it is only postponed!

As electron density maps very rarely has atomic resolution, fitting a model is somewhat arbitrary. This also causes that solving a structure of protein only with electron density map is usually impossible. In crystallography, when solving protein structure, usually all available information is used.

The most basic information about protein is its amino acid sequence. Without that information assigning amino acids to specific electron density blobs would be impossible. When building and refining model hundreds of constraints are used, obtained from thousands of previously solved protein structures. These constrains gives restrictions for bond lengths and angles. Building main polypeptide chain demands keeping the peptide bond planar, then angles ψ and θ can adopt only certain range of values, and of course bond lengths adopt only certain range of values also. When modelling amino acids side chains databases of its conformations and frequency occurrence are used.

Refining is a recursive process of adjusting, computing discrepancy, more adjusting, computing discrepancy, …. up to the moment of achieving established level of agreement. Refinement is never really done, it's just stopped, at some arbitrary chosen point.

1.6. *Validation*

Overfitting is very common!

The last step of solving a structure is validation of obtained structure. Fitting and refining as somewhat arbitrary, in diffused blobs of electron density it is possible to fit almost everything(this is called overfitting), thus there is a necessity of tools, that could validate the structure independently, without human bias.

Firstly the quality of fit has to be verified against experimental data. For that purpose the factors of discrepancy *R* are calculated. They are defined as average difference between structure factors computed from model F^{calc} and experimental structure factors $F^{\exp}$ according to the following expression:

$$R=\frac{\sum_{hkl}\left\|F_{hkl}^{\exp}\right|+\left|F_{hkl}^{calc}\right\|}{\sum_{hkl}\left|F_{hkl}^{\exp}\right|}$$

The smaller *R* is the better model fit the data. Structures with $R < 0{,}2$ are considered well refined. To validate, if refined parameters are statistically fully covered in experimental data, the factor R_{free} is calculated. This factor is calculted for the 5% of data excluded from refinement, thus not human biased. R_{free} shouldn't be much larger than R .

Secondly the geometrical correctness of structure is checked. Bond lengths and angles should be in certain ranges, typical for proteins. Obtained structure is compared against databases. The planarity of peptide bond is checked, the ψ i θ angles in main polypeptide chain are checked, and very often Ramachandran plot is given, conformations of amino-acids side chains are checked. Calculation of van der Waals contact allow to evaluate if atoms are not to close.

Thirdly there is “environmental” check. Hydrophobic amino-acids should be rather inside the protein, interacting with other hydrophobic amino-acids, polar amino-acids should rather be on the surface of the protein interacting with water or other polar amino-acids.

2. *Crystals*

In the ideal crystal molecules are regularly ordered in space creating three-dimensional net, called crystallographic lattice. Unit cell is primary structural element of crystal repeating in space and creating that crystallographic lattice. Unit cell is parallelepiped characterised by three edges of lengths: a, b, c and three angles between those edges: α, β, γ . Lengths and directions of edges of unit cell are defined as three vectors: $\vec{a}$, $\vec{b}$, $\vec{c}$, these vectors span a crystallographic real space. Position of every point in that space can be described by the translation vector: $\vec{r}'=\vec{r}+n\vec{a}+m\vec{b}+l\vec{c}$. Points of coordinates (na, mb, lc) are called nods of the lattice (n,m,l are any integers).

Sets of parallel Bragg's planes are described by Miller indices: h, k,l. Values of Miller indices are given by the points of intersections of axis coordinate system with this one of the set of planes, which is the closest to the origin of coordinate system and goes through the nodes of lattice(Fig. 4)

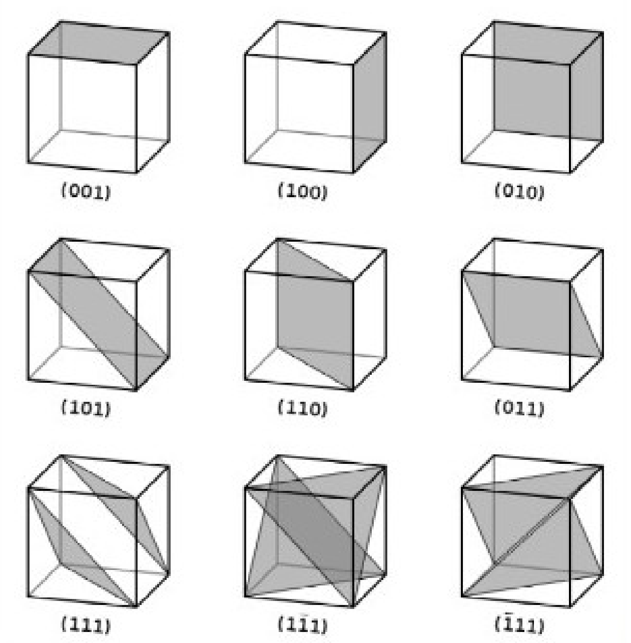


Fig. 4: Miller indices for basic crystallographic planes in cubic unit cell[4]

Crystallographic lattice in reciprocal space is also defined by edges of the unit cell: $\vec{a}^*, \vec{b}^*, \vec{c}^*$ and angles between this edges: $\alpha^*, \beta^*, \gamma^*$. The relationship between vectors spanning real and reciprocal space are as follows:

$$\vec{a}^* \cdot \vec{a} = \vec{b}^* \cdot \vec{b} = \vec{c}^* \cdot \vec{c} = 1$$

$$\vec{a}^* \cdot \vec{b} = \vec{a}^* \cdot \vec{c} = \vec{b}^* \cdot \vec{c} = 0$$

This means that $\vec{a}^*$ is perpendicular to $\vec{b}$ and to $\vec{c}$ and it's length equals $1/a$.

Point described in a real space with vector $\vec{r}' = n\vec{a} + m\vec{b} + l\vec{c}$, is described with vector $\vec{H} = h\vec{a}^* + k\vec{b}^* + l\vec{c}^*$ in the reciprocal space. The symmetry of real space is preserved in reciprocal space, not only geometry but also the intensity of scattered waves.

Relations between edges lengths and angles of unit cell defines seven crystal classes(Fig. 5): cubic, tetragonal, orthorhombic, hexagonal, trigonal, monoclinic, triclinic.

Primitive lattice, build of primitive cells (P) have nods only in the apexes of cell. Nonprimitive lattices can have more nods: on all faces (*face-centered,* F), on two opposite faces (side-*centered,* C) or in centre of the unit cell volume *(body-centered,* I). Seven crystal classes together with four type of unit cells creates fourteen Bravais lattices(Fig. 5).

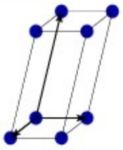
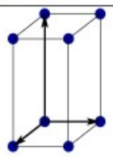
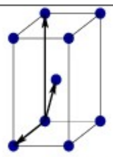
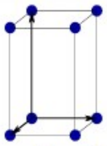
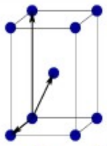
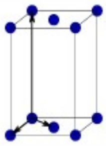
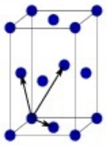
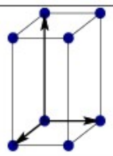
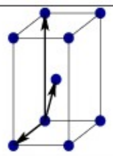
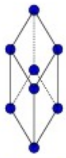
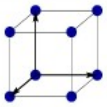
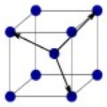
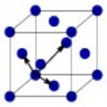
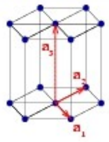
Bravais lattice	Parameters	Simple (P)	Volume centered (I)	Base centered (C)	Face centered (F)
Triclinic	$a_1 \neq a_2 \neq a_3$ $\alpha_{12} \neq \alpha_{23} \neq \alpha_{31}$				
Monoclinic	$a_1 \neq a_2 \neq a_3$ $\alpha_{23} = \alpha_{31} = 90°$ $\alpha_{12} \neq 90°$				
Orthorhombic	$a_1 \neq a_2 \neq a_3$ $\alpha_{12} = \alpha_{23} = \alpha_{31} = 90°$				
Tetragonal	$a_1 = a_2 \neq a_3$ $\alpha_{12} = \alpha_{23} = \alpha_{31} = 90°$				
Trigonal	$a_1 = a_2 = a_3$ $\alpha_{12} = \alpha_{23} = \alpha_{31} < 120°$				
Cubic	$a_1 = a_2 = a_3$ $\alpha_{12} = \alpha_{23} = \alpha_{31} = 90°$				
Hexagonal	$a_1 = a_2 \neq a_3$ $\alpha_{12} = 120°$ $\alpha_{23} = \alpha_{31} = 90°$				

Fig. 5: Bravais lattices

In protein crystal only few symmetry operations are allowed: translations, axes and screw axes. Because all amino acids in proteins are L-amino-acids operation like mirrors or inversion are impossible.

In molecular crystals unit cell are often quite big and contain more than one protein molecule, that means that there is another element of symmetry, smaller than the unit cell, it's called asymmetric unit (ASU). ASU copied in space according to symmetry operations given by crystal group of symmetry fills the unit cell.

Protein crystals contains significant amount of water. Water molecules are alike inside protein molecules, on their surfaces creating hydration shell, and between protein molecules as unordered bulk occupying even 50-60% of the volume of the crystal.

In most cases, protein crystals are not mono-crystals, they are composed of mono-crystalline domains, which are not ideally aligned, but slightly shifted and distorted relative to each other. Parameter that describes the level of distortion is called mosaicity.

3. Diffractometers

There are various types of diffractometers, from so called pocket diffractometers to X-rays beam created in synchrotrons. Every diffractometer has to contain X-ray source, detector, to register diffraction images, goniometer, to allow positioning the crystal, and cryo-system, to keep crystal in safe temperature of 100K. There is a large variety of ways, this parts can be done. This brief introduction is not designed to be a lecture on X-ray diffraction equipment. The main parts are presented for the *SuperNova* diffractometer by *Oxford Diffraction*, an instrument which works in Department of Biophysics.

3.1. Source

The source is a microfocus sealed tube X-ray generator with unmovable copper anode.

The X-ray source is a vacuum tube containing cathode and anode. The cathode is heated by high intensity current, the heat causes cathode to emits electrons, which are accelerated by strong electric field towards anode. The electric field is generated by high voltage(30-150kV) power source connected across cathode and anode. Electrons collide with anode, carrying enough energy to strike out electrons from atomic inner shells. Electrons coming back to its shells emit few characteristic wavelengths. In protein crystallography the Kα line is used, of the wavelength of 1,5418Å.

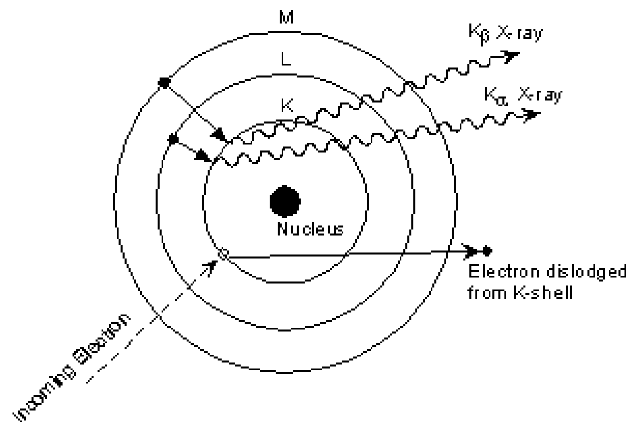


Fig. 6: Interaction of electrons with the atoms of anode

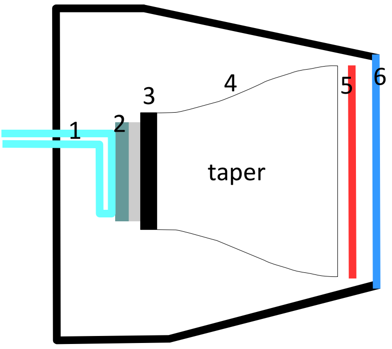
Only about 1% of energy absorbed by anode is emitted in the form of radiation. The rest change into heat, and cooling anything in the vacuum is not easy task. There are two solutions that are used: rotating anode and microfocusing of electron beam. When anode rotates, the electron beam collides with different parts of anode, not allowing to overheat any part of it. In case of microfocusing, the electron beam is precisely collimated, and collide with very small area of anode, of diameter of tens of micrometers, this allow reduce intensity of electron beam and therefore heating of anode. Another advantage of microfocusing is better collimation of X-ray beam emitted from anode, reduced power consumption and prolonged X-ray source lifetime.

The X-ray beam is additionally collimated by optical microfocusing system, giving finally beam of diameter of tens to few hundreds micrometers.

3.2. Detector

For the X-ray detection a scintillation counter combined with CCD camera is used.

Scintillator is substance, which hit by quantum of ionisation radiation produces a photon of visible radiation. Behind the scintillator there is taper, an optical element, which "scales" image from scintillator into CCD matrix, which is smaller. To decrease noise level from CCD matrix this is cooled to -40°C. This system also works in vacuum. The beryllium window in front of scintillator seals the vacuum compartment, beryllium is chosen because is well transparent for the X-ray radiation.

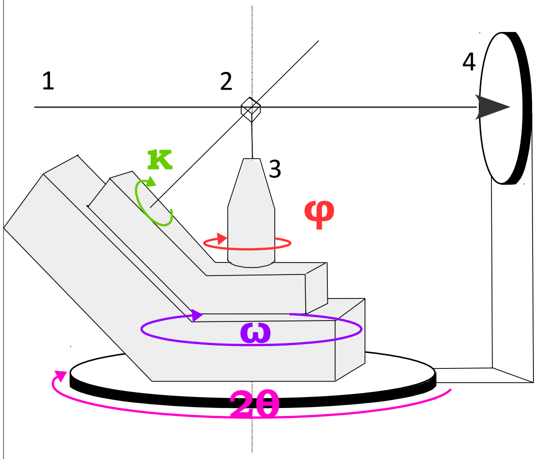


1: cooling of Peltier's element
2: Peltier's element
3: CCD matrix
4: taper
5: scintillator
6: beryllium window

Fig. 7: Scheme of detector

3.3. Goniometer

Protein crystal is mounted on the head of the goniometer. This devices allow to align crystal precisely in relation to X-ray beam and detector, and rotate crystal into almost every position, to collect waves scatter into whole sphere. The goniometer showed in Fig. 8 is so called 4-circle kappa goniometer, there are three axes of rotation for the crystal (the angles of rotation φ, κ and ω are shown on the scheme) and one axis of rotation for the detector 2θ. Goniometer is equipped with precise steeper motor which allows adjustment of crystal position with accuracy of about 10µm.



1: incident X-ray
2: crystal
3: goniometer head
4: detector

Fig. 8: 4-circle kappa goniometer

Through whole the experiment crystal is in the jet of gaseous nitrogen at the temperature 100K.

4. Measurement of crystals diffraction patterns

4.1. Preparing crystals

Highly energetic photons of X-ray radiation cause degradation of protein crystals. “Home” or “pocket” diffractometers have quite week beams, so measurements takes a lot of time, usually few or even several minuter for one frame, the synchrotron beams are much stronger, so taking an image is shorter, but overall destruction of crystal – similar. As long as diffraction experiments were conducted at room temperature, it took few crystals to collect full dataset, up to the discovery, that glassification of crystals extend their lifetime in the X-ray beam. Nowadays crystals are fished from the drops with special loops, flash-cooled in liquid nitrogen. During diffraction experiments crystals are in the jet of gaseous nitrogen at the temperature 100K. This solution also has disadvantages, as freezing increases mosaicity of crystals.

4.2. Data collection

Collecting full dataset, means measure all of the scattered waves, it may demand collecting data from the full sphere(when crystal has no rotational symmetry) but usually its just a fraction of that, because of rotational symmetry of the crystal. The diffraction pattern repeats every 180, 90 or 60° dependently of the group of symmetry of crystal lattice, so its enough to collect that fraction of full sphere with reasonable margins.

Therefore, for designing reasonable data collection, first the inner symmetry of crystal has to be determined. Pre-experiment consist of collecting at least two images, often spaced by 90°, this is enough for initial assessment of group of symmetry, shape and dimension of unit cell. Based on this data, with the help of software designed for that purpose the full experiment is planned, and subsequently data is collected.

Collecting diffraction images is the last experimental step in protein structure determination, all the subsequent steps are computational.

5. Solving protein structure – computational stage

Analysis of diffraction data is a multi-stage process. There is a lot of software and software packages for that purpose. The most used are CCCP4 and Phenix, and they are both free-ware!

In this exercise the CCP4 package will be used.

On the scheme on the right, a typical work-flow is shown.

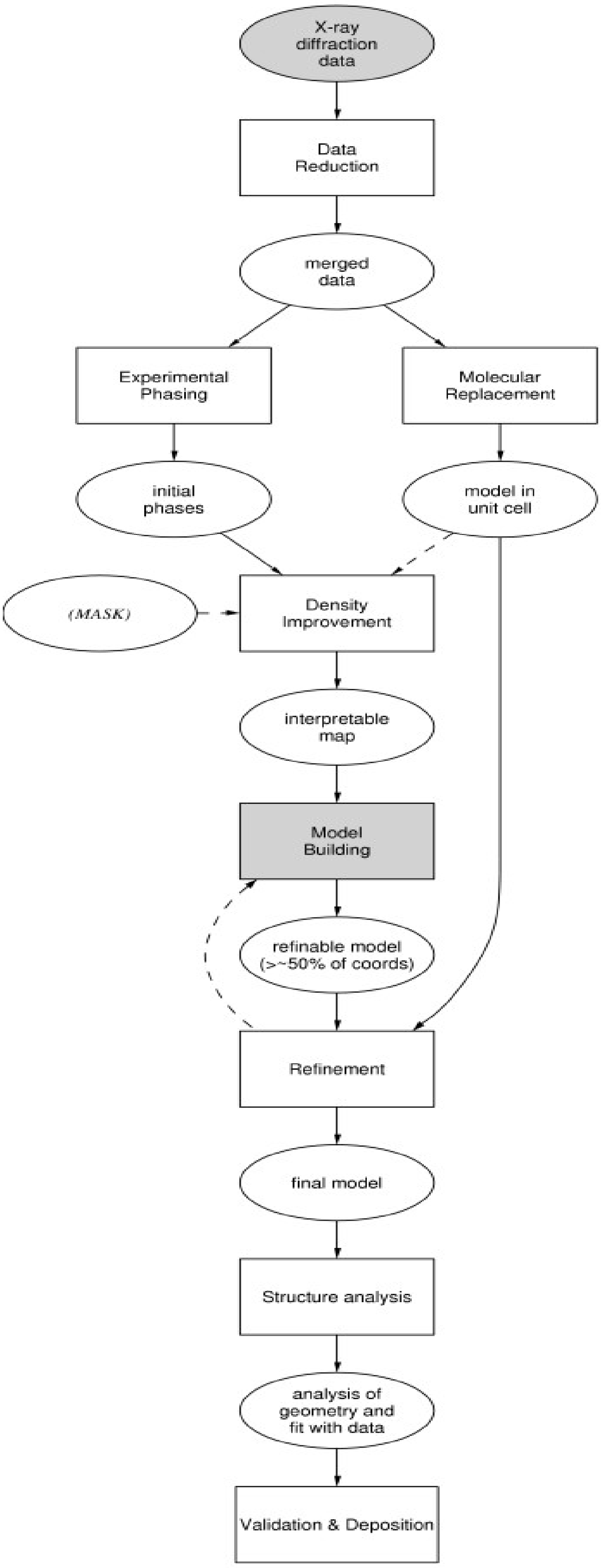


Fig. 9: Typical work-flow of analysis of protein diffraction data [6].

5.1. Data reduction

Data reduction is stage in which large quantity of data, diffraction images, will be transformed into one table containing structure factors amplitudes with corresponding Miller indices hkl assigned.

5.1.1. Indexing

The aim of indexing diffraction images is to find all the "measurable" spots on the images and identify on which Bragg's planes the radiation scattered. The Bragg's planes are described by their Miller indices hkl. On all images the darkness of pixels is analysed, the groups of pixels which are darker than the background are recognised as spots. For each the three-dimensional Gauss profile is fitted, from this profile the intensity I and its error(standard deviation $sigI$) is calculated. As the same diffracted reflections appears on many subsequent images, there are many spots assigned to the same hkl triplet.

The catalogue of all spots with corresponding hkl indices are written in file with mtz extension. The file contains five columns h, k, l, I , $sigI$, information about unit cell dimensions and possible symmetry groups.

For data measured on the synchrotron beams, there is program *iMosflm* in CCP4 package for indexing images.

For data measured on the *SuperNova* diffractometer there is a dedicated program created by the manufacturer: *CrysAllisPro*.

5.1.2. Scaling & Merging

The catalogue of "spots" obtained in the previous step, has to be transformed into data, which finally allow to plot electron density map. Therefore reflections from parallel Bragg's planes have to be "gathered" into one "spot", one intensity. And the structure factors amplitudes have to be calculated from intensities.

Merging of spots is conducted as weighted integration. Crystal is not spherical, and spots, even from the same hkl planes have different intensities on different images, and the background around them has different levels as the X-ray beam goes through layers of different thickness. Weights are assigned to spots by analysing spot intensity and level of background around them[5].

During all the stages of analysis the quality of data is assessed by comparison of computed averaged values with experimental data, various R coefficients are computed:

$R_{merge}=\frac{\sum_h \sum_l |I_{hl}-\langle I_h\rangle|}{\sum_h \sum_l \langle I_h\rangle}$ - measures discrepancy between measured values, has a tendency to grow with the growing number of measured reflections.

$R_{meas}=\frac{\sum_h \left(\frac{n_h}{n_h-1}\right)\sum_l |I_{hl}-\langle I_h\rangle|}{\sum_h \sum_l \langle I_h\rangle}$ - this factor doesn't depend on number of measured spots, is somewhat corrected R_{merge} n_h is the number of measured reflections of h indice

$R_{p.i.m.}=\frac{\sum_h \left(\frac{1}{n_h-1}\right)\sum_l |I_{hl}-\langle I_h\rangle|}{\sum_h \sum_l \langle I_h\rangle}$ - this factor assess the quality of averaged data

There are many R factors that are computed, because each of them has some advantages and disadvantages.

Scaling & Merging is a few step process, there will be needed following programs *Pointless*, *Scala*, *cTruncate*. Each of this program needs as an input mtz file from previous step of analysis.

5.1.2.1. Pointless

Pointless point out space group which the best fit to the data. The spots have to be scaled & merged in proper group.

5.1.2.2. Scala

Scala merges and scales intensities of reflections into one intensity for spots scattered on set of parallel Bragg's planes.

At this stage of analysis the 5% of measured data is separated, it will serve as control dataset. The set of all measured reflections, is randomly divided into 20 sets, in such a way that distribution of reflections versus resolution is similar in all dataset. One of these sets is chosen as control dataset. In the mtz file a new column will appear: R_{free}

From that moment the analysis will be performed on this two datasets independently. This 5% of data is not taken into account for solving protein structure, but it allows on better statistical control of computations. The R factors calculated for both dataset should be similar, values of these factors for R_{free} dataset will be higher, but they should keep the same tendencies as the main dataset. Significant discrepancies between R factors for main dataset and control dataset is a sign, that something goes wrong in conducted analysis.

5.1.2.3. cTruncate

From reflections intensities *cTruncate* computes the amplitudes of structure factors.

For ideal experimental data amplitude of structure factor is square root of intensity of reflection, but as data bears an experimental errors, better estimation of amplitudes of structure factors, especially when the intensities of spots are low, is the mean probability distribution. *cTruncate* uses this kind of algorithm to compute amplitudes of structure factors. Moreover it analyses dataset in order to find twinning.

To mtz file *cTruncate* ads two more columns: amplitudes of structure factors F with errors computed as standard deviations $sigF$.

At this point mtz file contains all the data that can be extracted from diffraction images: amplitudes of structure factors with errors.

5.2. Solving The Phase Problem by Molecular Replacement

5.2.1. Model building

CCP4 contain few tools, which allow to edit sequence and structure, build model from fragments or domains found in different databases, for example *Chainsaw*, *Modeller*

In this case, the model will be simply structure deposited in PDB.

5.2.2. Preparing data for MR

5.2.2.1. Cell Content Analysis(Matthews coefficient)

First step is to determine how many protein molecules contains asymmetric unit(ASU), that means calculating Matthews coefficient. This coefficient specify mean distances of atoms in protein crystal in Å^3/atom, thus allowing to compute approximate volume of protein molecule. By comparing this volume with volume of unit cell program can suggest how many protein molecules has to be in ASU, to fill unit cell according symmetry group operations.

For proteins, mean value of Matthews coefficient is equal 2,35Å^3/atom.

As an input the mtz file has to be given, as it contains information of unit cell and symmetry group in its header, and information that will allow to compute protein volume e. g. its sequence or molecular mas.

As an output a table with Matthews coefficients will be displayed.

5.2.2.2. Analyse data for MR

This program computes Patterson function, determines its maxima and calculates value of B factor based on the Wilson plot.

The Patterson's function, can be visualised as a Patterson's map. This map, is somehow similar to electron density map, but its maxima are not at the positions of atoms, but in positions corresponding to relative distances between atoms. If, in the unit cell, there is an atom at position $\vec{x}_1$ and another atom in position $\vec{x}_2$, on the Patterson's map there will be two peaks in positions $\vec{x}_1 - \vec{x}_2$ and $\vec{x}_2 - \vec{x}_1$. The heights of peaks on the Patterson's map is proportional to product of peaks on the electron density map. If there is N atoms in the unit cell, there should be N peaks of electron density. On the Patterson's map, each of this N atoms is linked by a vector with every N atom (with itself too), that gives N^2 vectors. N of these vector connects atoms with themselves, that will generate a large peak at position zero. The rest of the N^2-N peak will surround that peak.

The Patterson's map is difficult to interpret, but as for molecular replacement, there is only few information needed out of it. The highest peak on Patterson's map is normalised to 100, if height of none of the surrounding peaks does not exceed 20% of the highest peak, there is no pseudotranslation in the unit cell. That means, that there is no additional molecules, besides this already had been assigned in the ASU.

The B factor is the effective diameter of electron density cloud of an atom, that diameter is a result of thermal vibration. This value can be useful during molecular replacement and refinement, it allows to impose additional constrains.

5.2.2.3. Molecular replacement

There are several programs in the CCP4, which can perform molecular replacement: *Phaser*, *MolRep*, *MrDUMP*, *Balbes*, *Amore*.

As the input the mtz file with amplitudes of structure factors and the pdb with the model molecule has to be given.

As an output the new pdb file with reoriented molecule will be generated.

5.3. Fitting & Refinement

One should bear in mind that a macromolecular refinement against high resolution data is never finished, only abandoned.
George Sheldrick (2008), Acta Cryst. D 64,112–122.

5.3.1. Refmac5

For fitting and refining the *Refmac 5* will be used.

Using the maximum likelihood algorithm, the model molecule will be fitted into electron density, and first few rounds of restrained refinement will be performed. Then for the adjusted model electron density map F and differential electron density map ΔF will be computed. This maps are defined as follows:

$$F = 2 \cdot F_{obs} - F_{calc}$$
$$\Delta F = F_{obs} - F_{calc}$$

The refining algorithms implemented in *Refmac5* are not enough, and refinement is still done by human being. The two maps calculated in *Refmac5* are of great help. Analysing just electron density map, it is very difficult to find areas which need improvement. That is why also differential electron density map is computed. It allows to find easily regions where model doesn't fit well to electron density.

As an input mtz file with structure factors, and pdb file with model positioned in MR has to be given.

As an output, there will be new pdb file with coordinates of initially refined molecule, and new mtz file with four new columns containing amplitudes and phases of electron density map(WT – amplitudes, PHWT - phases) and differential electron density map(DELFWT i PHDELWT). Moreover there are calculated several parameters allowing assessing quality of fit and refinement.

5.3.2. Coot

This program is used for manual refining of the structure. *Coot* allows to plot in 3D both electron density maps. Intuitive colouring of differential electron density map makes easy to find regions which need refinement. The green regions of that map mean, that there is lacking electron density in the model, in comparison to experimental electron density map, the red regions mean the there are to much electrons in the model.

Coot has many tools which allow to modify positions and shapes of amino acids and polypeptide chain, and many validating tools which help to find region which need refinement.

After refining same regions, it is good to save changes, and perform again analysis in *Refmac5*, the new electron density maps will be calculated, the regions in which adjustments of the model were done well will disappear from differential electron density map, and more subtle problems will become visible.

In that recursive process the model is step by step adjusted to electron density.

The last step of analysis is analysis of empty blobs of electron density. This is a procedure to add any possible ligands to the model. When all the “big” blobs are filled, and only little ones remains, most of the are probably water molecules. Adding the waters, either with help of *Coot* or *Refmac5* ends refinement.

5.4. Structure Validation

After making decision of stopping further refinement, the quality of obtained structure can be validated with *sfcheck*. There will be done analysis of structure against experimental data, how good is fit, the *R* factors will be computed. The geometrical correctness of the structure will be checked, and Ramachandran plot will be given.

5.5. Presentation of the results

Solved structures are usually deposited in PDB, which we don't do!

In the corresponding papers, the tables with same statistics are published, they usually looks more less like this:

Data collection	
Wavelength (Å)	0.975
Resolution range (Å)	28.6 - 1.9
Space group	P 61 2 2
Unit cell	120.8 120.8 239.1 90 90 120
Total reflections	704235
Unique reflections	80026
Multiplicity	8.8
Completeness (%)	99.4
Rmerge	0.25
Rmerge in top intensity bin	0.019
I/sigma(I)	21.6
Wilson B-factor	26.14
Refinement	
Rwork	0.163
Rfree	0.188
Number of atoms	6187
Protein residues	718
Water molecules	623
RMS(bonds)	0.008
RMS(angles)	1.10
Ramachandran favored (%)	96
Ramachandran outliers (%)	0.69
Average B-factor	31.10

6. Bibliography

[1] Bernahard Rupp, Crystallography 101, http://www.ruppweb.org/Xray/101index.html
[2] University of Cambridge, Cambridge Institute for Medical Research (CIMR), Protein Crystallography Course, http://www-structmed.cimr.cam.ac.uk/course.html
[3] P. Evans, A McCoy, An introduction to molecular replacement,
[4] Yuan Ming Huang, Solid State Physics: Miller indices, http://www.lcst-cn.org/Solid%20State%20Physics/Ch16.html
[5] P. Evans, Scalling & Assesment of data quality,
[6]CCP4 Software for Macromolecular X-Ray Crystallography http://www.ccp4.ac.uk/, http://ccp4wiki.org/

7. *Further reading*

- Bernhard Rupp *Crystallography 101* http://www.ruppweb.org/Xray/101index.html
- Protein Crystallography Course (University of Cambridge, Cambridge Institute for Medical Research (CIMR)) http://www-structmed.cimr.cam.ac.uk/course.html
- CCP4 Software for Macromolecular X-Ray Crystallography http://www.ccp4.ac.uk/, http://ccp4wiki.org/
- Basic Maths for Protein Crystallographers: http://www.ccp4.ac.uk/dist/html/pxmaths/index.html
- Zbigniew Dauter: Data collection strategies, Acta Crys D (1999) 55,1703-1717
- Philip Evans: Scalling & Assesment of data quality, Acta Cryst. D (2006). 62,72–82
- Philip Evans, Airlie McCoy *An introduction to Molecular Replacement* Acta Cryst. D(2008) 64, 1–10
- Garib N. Murshudov et. al. *REFMAC5 for the refinement of macromolecular crystal structures* Acta Cryst. D(2011). 67, 355–367

8. *Topics for preliminary test*

- Interacting of light and matter – diffraction, interference
- Crystals features and specific features of protein crystals.
- Symetry of crystals, and specificity of protein crystals.
- Stability of protein crystals in X-ray beams.
- What can be seen on diffraction images?
- Indexing of diffraction images
- The Phase Problem and methods of solving it.
- Molecular Replacement.
- Structure validation.

9. *Performing the exercise*

The aim of this exercise is to collect data and solve structure of human serum albumin (HSA).

The crystals of HSA will be grown previously. By visual assessment of crystal the selection of best crystal for diffraction experiment will be done. Then the crystal will be fished out, flash cooled in liquid nitrogen, mounted in diffractometer, and full data set collected. The anylisis of the data and solving structure will be conducted by means of CCP4 package for protein crystallography (http://www.ccp4.ac.uk/). The Phase Problem will be solved by Molecular Replacement.

HUMAN CAPITAL
NATIONAL COHESION STRATEGY

EUROPEAN UNION
EUROPEAN
SOCIAL FUND

Project co-financed by European Union within European Social Fund