Mathematical Biology and Bioinformatics
2020. V. 15. № 1. P. 57–72. doi: 10.17537/2020.15.57

========== MATHEMATICAL MODELING ==========

Mask-Based Approach in Phasing and Restoring of Single-Particle Diffraction Data

Lunin V.Y., Lunina N.L., Petrova T.E.

Institute of Mathematical Problems of Biology RAS, Keldysh Institute of Applied Mathematics of Russian Academy of Sciences, Pushchino, Moscow Region, 142290 Russia

Abstract. The development of experimental techniques, in particular the emergence of the X-ray free-electron lasers, allows one to register the scattering from an isolated particle and, thereby, opens a door to the study of a fine three-dimensional structure of non-crystalline biological objects by X-ray diffraction methods. The possibility to measure non-Bragg reflections makes experimental data mutually dependent and essentially simplifies the structure solution. The sampling of experimental scattering data to a sufficiently fine grid makes the structure determination equivalent to phasing of structure factor magnitudes for a 'virtual' crystal with extremely large solvent content. This makes density modification phasing methods especially powerful supposing the object envelope is known. At the same time, such methods may be sensitive to the accuracy of the predefined envelope and completeness of experimental data and may suffer from non-uniqueness of the solution of the phase problem. The mask-based approach is a preliminary phasing method that performs random search for connected object envelopes possessing of the structure factor magnitudes close to the values observed in X-ray experiment. The alignment and averaging of the phase sets corresponding to selected putative envelopes produce an approximate solution of the phase problem. Beside the estimation of unknown phase values this approach allows one to estimate the values of structure factor magnitudes lost in the experiment, e.g. those corresponding to beam-stop shade zone or to oversaturated reflections.

Key words: *biological macromolecules, single-particles, X-ray scattering, X-ray free electron lasers, phase problem, magnitude retrieval, effective resolution.*

1. INTRODUCTION

It exists a popular belief that if one wants to know the structure of a biological macromolecule, then the only problem is to prepare its crystal. People engaged in computing methods do their best to support this statement, developing more and more powerful and user-friendly programs for structure solution [1–6]. However, this does not eliminate the need for crystal preparation. Another permanent problem is the resolution limit of the collected data that again relates to the quality of the prepared crystals. In the last decade, the development of the X-ray experiment, in particular, the commissioning of X-ray free-electron lasers, has opened the door for approaches that would possibly allow to overcome these difficulties and determine three-dimensional structures of non-crystallized objects [7–12]. In this paper we touch upon one of the problems that arise, namely, solving the structure, under the assumption that the experimental data has been collected. The main content of this article was presented at the 32nd European Crystallographic Meeting in Vienna, Austria, August 18–23, 2019 [13].

2. X-RAY DIFFRACTION EXPERIMENT

Two following sections present the basic information about the X-ray diffraction experiment and properties of the diffraction pattern. More details can be found in [14]. In the routine diffraction experiment, both with a crystal or an isolated particle, the sample is exposed

to a primary X-ray beam and the energies of the new X-ray beams (electromagnetic waves) scattered in different directions are registered by some detector. The kinematic theory of diffraction describes the scattered waves as the superposition of the secondary spherical waves emitted by the oscillating electrons that were excited by the primary wave. These diffracted waves are traditionally called reflections. This term was inspired by the study of crystal samples in which the diffracted waves were interpreted as the result of the primary wave reflection in the "crystal planes" according to the geometrical optics. In the framework of the kinematic diffraction theory, the result of an experiment is determined by the electron distribution within a sample described by the density distribution function $\rho(\mathbf{r}), \mathbf{r} \in \mathbf{R}^3$, and the measured energy is

$$E(\boldsymbol{\sigma}_0, \boldsymbol{\sigma}) = \varepsilon E_0 \left|\mathbf{F}(\mathbf{s})\right|^2 \quad . \tag{1}$$

Here $\boldsymbol{\sigma}_0$ and $\boldsymbol{\sigma}$ are unit length vectors that specify the directions of the propagation of the primary and scattered waves, E_0 is the energy of the primary wave, the constant ε is the combination of physical constants and parameters of the experiment that do not depend on the sample structure. The vector s is defined as

$$\mathbf{s} = \frac{\boldsymbol{\sigma} - \boldsymbol{\sigma}_0}{\lambda} \quad , \tag{2}$$

where λ is the wavelength of the primary wave. This vector is referred to as scattering vector and plays a significant role in the theory of diffraction. We call the reciprocal space the three-dimensional space formed by scattering vectors corresponding to different combinations of directions $\boldsymbol{\sigma}_0$ and $\boldsymbol{\sigma}$. Potentially, the experiment allows one to record the diffracted waves corresponding to the scattering vectors that fill the sphere $|\mathbf{s}| \le 2/\lambda$ in the reciprocal space. Nevertheless, in practice, the zone of observed reflections is much smaller; it depends on the quality of the sample, the energy of the primary beam, the sensibility of the detector, etc.

The quantities $\left|\mathbf{F}(\mathbf{s})\right|^2$, referred to as intensities $I(\mathbf{s})$, are the squares of the magnitude of the complex-valued Fourier transform of the electron density distribution

$$\mathbf{F}(\mathbf{s}) = \int_{\mathbf{R}^3} \rho(\mathbf{r}) \exp\left[i 2\pi \mathbf{s} \cdot \mathbf{r}\right] dV_{\mathbf{r}} \ , \ \mathbf{s} \in \mathbf{R}^3 \ , \tag{3}$$

where the dot denotes the scalar (dot) product of two vectors. The values $I(\mathbf{s})$ depend on the structure of the sample only and may be obtained (in a relative scale) directly in the diffraction experiment.

Under the routine experiment conditions the value ε is extremely small and can be estimated as 10^{-24}. This makes the experimental recording of reflections very difficult. Until now, the only way to obtain experimental values of the intensity of reflections was to prepare a sample in the form of a single crystal. In this case, for a particular discrete set of scattering vectors (Bragg reflections), the intensities of the scattered rays increase proportionally to the square of the number of elementary cells in the crystal. At the same time, the use of the crystal leads to a loss of information for all other reflections. The development of X-ray experiment technique (in particular, the commissioning of X-ray free electrons lasers, XFEL [7–12]) has led to the possibility of registration of the scattering by isolated (non-crystallized) objects, although for a very limited zone $|\mathbf{s}| \le s_{\min}$ of the reciprocal space.

The case of a single particle may be converted to the standard crystallographic case, if we assume that the particle is immersed into a virtual crystal unit cell *V* that is large enough for the particle to fit into the cell, so that outside the cell the density $\rho(\mathbf{r}) = 0$. Let {**a**, **b**, **c**} be the

basis of this cell. For simplicity, below we suppose that the unit cell is cubic with the edge a_{cell}. Inside the cell, the distribution of the electron density may be represented as Fourier series

$$\rho(\mathbf{r}) = \frac{1}{|V|}\sum_{\mathbf{s}\in\mathfrak{R}'}\mathbf{F}_V(\mathbf{s})\exp[-\mathrm{i}2\pi\mathbf{s}\cdot\mathbf{r}],\ \mathbf{r}\in V. \tag{4}$$

Here $|V|$ is the unit cell volume and the Fourier coefficients (the structure factors) $\mathbf{F}_V(\mathbf{s})$ are nothing but the values of the Fourier transform (3). The set $\mathfrak{R}'$ of scattering vectors, along which the summation in (4) is performed, is the grid in the reciprocal space formed by vectors that satisfy the conditions

$$\mathbf{s}\cdot\mathbf{a} = h, \mathbf{s}\cdot\mathbf{b} = k, \mathbf{s}\cdot\mathbf{c} = l, \quad h,k,l\ -\ \text{integers}, \tag{5}$$

i.e. have integer coordinates in the basis $\{\mathbf{a}^*, \mathbf{b}^*, \mathbf{c}^*\}$ that is conjugate to $\{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$. We call corresponding reflections as Bragg reflections. The identity (4) is not valid for the points $\mathbf{r}$ that lie~~s~~ outside V. We consider the electron density distribution $\rho^{cryst}(\mathbf{r})$ defined by the identity (4) for all point $\mathbf{r}\in\mathbf{R}^3$ as the virtual crystal. This distribution is periodic with the periods {a, b, c} and coincides with $\rho(\mathbf{r})$ in the cell V. The problem we are faced with is the standard problem of biological crystallography, that is, to restore electron density in the unit cell of this virtual crystal starting from the intensities of Bragg reflections, or, equivalently, from the magnitudes |F(s)| of structure factors. This problem is a problem of the phasing of structure factors corresponding to Bragg reflections, i.e. of the retrieval of the values of their phases φ(s), which are necessary to calculate the synthesis (4), but are lost in the diffraction experiment.

Similarly, if the continuous diffraction pattern $I(\mathbf{s}), \mathbf{s}\in\mathbf{R}^3$ has been sampled to a regular grid specified by $\{\mathbf{a}^*, \mathbf{b}^*, \mathbf{c}^*\}$ sampling steps, we can consider the intensities in the grid points as ones corresponding to the Bragg reflections for the virtual crystal with the conjugate periods $\{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$ and the unit cell V. The fundamental difference from the standard crystallographic situation is that, for a single particle, we have a considerable freedom in choosing the unit cell size, or, what is the same, of the sampling step. It is only necessary to have the dimensions of the cell large enough to contain the particle inside.

Once the phase values of the structural factors are somehow found they are used to approximate the electron density distribution by formula (4). The result of this calculation is usually referred to as the Fourier synthesis of the electron density. The concept of 'resolution' is used as a characteristic of the completeness of the set of reflections included in the calculation. In biological crystallography, the concept of resolution is initially introduced for an individual reflection $\mathbf{s}$ as the period value of the corresponding Fourier harmonic $\exp[-i2\pi\mathbf{s}\cdot\mathbf{r}]$ in series (4) considered as the function on $\mathbf{r}$. This period is equal to

$$d = \frac{1}{|\mathbf{s}|} = \frac{\lambda}{2\sin\theta}, \tag{6}$$

where 2θ is the angle between the direction $\boldsymbol{\sigma}_0$ of the primary beam and the direction $\boldsymbol{\sigma}$ of the reflected beam, and λ is the wavelength. It is customary to say that the data set has been collected or the synthesis (4) was calculated with the resolution $d_{\min}$, if they include all reflections with $|\mathbf{s}| \le s_{\max} = d_{\min}^{-1}$. We call below the set of reflections defined as $|\mathbf{s}| \le s_{\max}$ the resolution zone, and the set defined by $s_{\min} < |\mathbf{s}| \le s_{\max}$ the resolutions shell. The other concepts of the resolution that accounts for the accuracy of phase values and visual properties of the 'topographical map' corresponding to the Fourier synthesis are considered as well [15–17].

3. ANALYTICAL PROPERTIES OF A SINGLE PARTICLE DIFFRACTION PATTERN

A decrease in the sampling step increases the amount of observed data involved in the work. However, this data may be highly dependent. If the virtual unit cell is large enough to contain inside the support of the distribution $\rho(\mathbf{r})$, i.e., all points where the density takes non-zero values, then each Fourier transform value is the linear combination of the Bragg structure factors

$$\mathbf{F}(\mathbf{u}) = \sum_{\mathbf{s}\in\mathfrak{R}'} \mathbf{sinc}(\mathbf{u}-\mathbf{s})\mathbf{F}(\mathbf{s}),\ \mathbf{u}\in\mathbf{R}^3\ , \tag{7}$$

with the predefined (non-depending on the function $\rho(\mathbf{r})$) coefficients

$$\mathbf{sinc}\left(h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*\right) = \frac{\sin\pi h}{\pi h}\cdot\frac{\sin\pi k}{\pi k}\cdot\frac{\sin\pi l}{\pi l}\,. \tag{8}$$

(In the last formula, we suppose the origin be placed in the center of the unit cell). Formula (7) plays an important role in the theory of information and is usually associated with the names of V. Kotelnikov, E. Whittaker, C. Shannon, and H. Nyquist, although the early mentions of it can be found in the mathematical works of E. Borel [18, 19]. The perspectives of using this formula in biological crystallography were discussed in the works of D. Sayre and G. Bricogne [20–22]. It follows from (7) that the functions **F**(**s**) and $\rho(\mathbf{r})$ are fully defined by the values of structures factors corresponding to the set of Bragg reflections related to the minimal virtual unit cell that still contains the particle inside.

A similar formula is valid for the intensity function $I(\mathbf{s}) = \left|\mathbf{F}(\mathbf{s})\right|^2$ on assumption that the virtual unit cell is large enough to contain the support of the autocorrelation function defined by

$$A(\mathbf{r}) = \rho * \breve{\rho} = \int_{\mathbf{R}^3} \rho(\mathbf{u})\rho(\mathbf{u}-\mathbf{r})\,dV_{\mathbf{u}}\,,\ \mathbf{r}\in\mathbf{R}^3. \tag{9}$$

Supposing $\operatorname{supp} A(\mathbf{r}) \subset V$, we get

$$I(\mathbf{u}) = \sum_{\mathbf{s}\in\mathfrak{R}'} \mathbf{sinc}(\mathbf{u}-\mathbf{s})\,I(\mathbf{s})\,,\ \mathbf{u}\in\mathbf{R}^3\ . \tag{10}$$

A sufficient condition to have this formula valid is that the virtual cell dimensions are twice as large as the particle size, or, what is the same, the sampling step is twice as small as the inverse of the particle diameter. Such sampling step value is often referred to as the Nyquist limit.

It follows from (10) that the diffraction pattern $I(\mathbf{s})$ is fully defined by the values of intensities for the set of the Bragg reflections corresponding to the minimal possible choice of the virtual unit cell that still contains the support of the autocorrelation function. This can create the illusion that the sampling step below the Nyquist limit does not introduce new experimental information. This is not the case in practice, since formulas (7) and (10) imply the summation over an infinite set of reflections. In the case that the information for a part of the Bragg reflections is missing or errors are present, the measurement of an additional non-Bragg intensity leads to the equation linking unknown Bragg intensities [23].

The particular importance of the intermediate points is emphasized by the fundamental mathematical result, namely, by the Schwartz's-Paley-Wiener theorem. This theorem states that Fourier transform of a function localized in a finite domain is an entire holomorphic function. The word 'entire' means, in particular, that the Taylor series for the intensity function derived at some point $\mathbf{s}_0$ converges at any other point in the reciprocal space and, therefore, determines the intensity at all points in the three-dimensional space. The coefficients of the expansion in

the Taylor series are the partial derivatives of the intensity function. To calculate these derivatives, it is sufficient to know the function in an arbitrarily small neighborhood of the point $\mathbf{s}_0$ only. Thus, we come to an amazing theoretical conclusion: it is enough to know all the intensity values in a small limited three-dimensional area $|\mathbf{s}-\mathbf{s}_0|<\delta$ in order to restore all the other values not being restricted by the experimental resolution limit. Unfortunately, the described above extension procedure is not feasible in practice. Nevertheless, this theoretical result makes reasonable the following task: knowing the values of intensities on a sufficiently fine grid in a limited part of the reciprocal space, to extend these values into a higher resolution zone. This task is not solved yet and presents a challenge for applied mathematics methods.

4. SOLVENT FLATTENING AND PROJECTION METHODS

In crystals of biomolecules, a significant part of the volume (50% on the average) is occupied by a solvent. Similarly, in the single particle study, we call the part of the virtual cell not occupied by the particle under study as the solvent region. The latitude in selecting of the virtual cell dimensions allows one to increase significantly the solvent content. For example, the doubling of the cell size as compared to the diameter of a spherical particle results in an estimated solvent content of 93 %. The presence of a large amount of solvent in the virtual unit cell suggests the use of the 'solvent flattening' method to solve the phase problem. This method belongs to a wide class of the 'density modification' methods that have been developing in biological crystallography since the early 70s [21, 22, 24–26]. The solvent flattening approach [21, 27] (suggested independently in optics applications under the name “the phase retrieval algorithm” [28]) can be described briefly as follows. Let’s suppose that we know what area of the unit cell is occupied by the particle (molecule region) or its complement (solvent region). One starts with some set of structure factor phases, combines them with the observed values of the magnitudes, and calculates Fourier synthesis (4). The resulting synthesis does not satisfy the requirement to be equal to zero in the solvent region. Therefore, its values in the solvent area are set to zero. Now, the modified distribution is equal to zero in the solvent area, but has wrong magnitudes of structure factors. Therefore, one takes only phases from it, combines them with the observed values of magnitudes, and calculates Fourier synthesis again. This procedure is repeated iteratively. This method is a part of many crystallographic software systems.

From the mathematical point of view, this approach belongs to the class of projection methods [29, 30] and can be described as follows. Let’s consider the configuration space of all electron density distributions. Let’s select two classes of functions (manifolds). The first manifold is the functions that vanish in the solvent region. The second manifold is the functions possessing the values of the structure factor magnitudes equal to the observed ones. Our goal is to find a point that belongs to both manifolds. The second manifold is nonlinear and non-convex, and therefore, the two manifolds may not intersect or may have several intersection points. Our iterative procedure consists in switches between these manifolds; the movement each time occurs to the closest point at the alternative manifold. The transition from some point to the closest one at the target manifold is called the projector. The convergence of this procedure can be slow. To accelerate it, the following technique can be applied. After a point has been projected to the manifold, the procedure does not stop; the point continues moving in the same direction by the same step. It seems as if we come to the mirror image of the starting point in this manifold. Such operations are called reflectors. Formally, the reflector is linked to the projector by the equality $R = 2P–I$, where I is the identity operator. Using this equality we can easily calculate the reflector that corresponds to the reduction of the solvent density to zero. This is a well-known operation of the solvent density flip [31, 32]. This operator does not change the density in the molecule region, but alternates the sign of the density in the solvent region. Similarly, the reflector corresponding to the substitution of the calculated magnitudes in Fourier synthesis by the observed ones is nothing but the calculation of the well-known

$\left(2F^{obs} - F^{calc}, \varphi^{calc}\right)$ combined Fourier synthesis. The use of reflectors can greatly accelerate convergence, but makes the process less stable. To accelerate convergence while maintaining stability, various combinations of these four operators with different weights can be used, which leads to a wide variety of schemes called projection methods. A list of the popular schemes may be found in [29]. We remind that all of them are combinations of four crystallographic operations, namely, solvent density zeroing or flip, and $\left(F^{obs}, \varphi^{calc}\right)$ or $\left(2F^{obs} - F^{calc}, \varphi^{calc}\right)$ Fourier syntheses. These and some other schemes are implemented in the Hawk program [33], which provides opportunities for the interactive control of schemes and parameters with visual control of the iterations. Application of Hawk program leads to amazingly good results when working with test data with the known answer. In this case, after a short training session, it is possible to find combinations of parameters that lead to the correct result. The situation becomes more complicated in the case of the imperfect data sets and unknown answer. This is of particular importance when working at very low resolution, when there is no clear idea of how the desired object should look like at this resolution.

Additional difficulties are introduced by the fundamental ambiguity of the solution of the phase problem. An arbitrary shift of the found density distribution in the space or a transition to the enantiomer ($\rho(\mathbf{r}) \rightarrow \rho(-\mathbf{r})$) leads to a distribution that has the same magnitudes of the structure factors, but different phase values, i.e. to a new solution to the problem. Therefore, the question arises as to what solution the iterative procedure converges. To exercise more complete control, global search methods can be used as a preliminary analysis allowing for more comprehensive-exploration of the configuration space of density distributions.

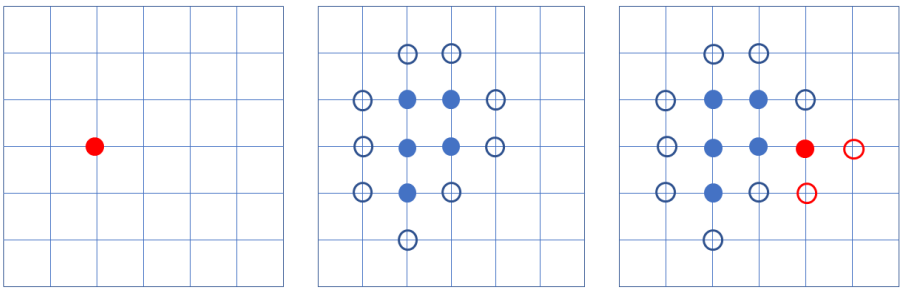
The solvent flattening approach supposes that the molecule region is known in advance. This region is often refined iteratively by the filtering of the current noisy Fourier synthesis, starting sometimes from a randomly phased one. A possible way of the filtering is the double-step procedure developed in [27, 34–36]. At the first step, every point in the unit cell gets a weight that reflects the confidence that this point belongs to the molecule region. In the simplest mode, this weight is equal to one if the synthesis value at this point exceeds some cutoff and is zero in the other case. In a more advanced mode, the weight value equal to one is assigned both to the highest and lowest values of the synthesis. At the second step, the weight value is replaced by the result of averaging of the weights in a vicinity of this point. The obtained smoothed distribution is used then to build the molecule region as the region of specified volume containing the points of maximal averaged weight.

5. MASK-BASED PHASING

The mask-based phasing procedure [37, 38] may be used for two goals. First, this is a procedure of low-resolution phasing, which gives the values of the structure factor phases and their reliability indexes. Second, it may be considered as a procedure for the *ab initio* search of a connected molecule mask consistent with the observed data. The demand of the connectivity of the mask has its own phasing power and enhances the phasing. The approach is based on the random generation of a large number of connected masks of the molecule region. A mask is considered acceptable if the structure factor magnitudes calculated from it are close to the observed values. Acceptable masks are stored. The generation process continues until the appointed number of acceptable masks (100 in out tests) is selected. The structure factor phases corresponding to the selected masks are aligned and averaged. The found phases constitute the output of the phasing step and may be used to calculate Fourier synthesis or to form the input for a phase refinement step.

A mask is a binary function defined usually on some grid in the unit cell. We say that the mask is connected if any two points of the mask can be connected by a path that passes from neighbor to neighbor nodes inside the mask. Figure 1 illustrates the way of a random mask generating. An important characteristic of the mask is its size. We specify the size of the mask

by the specific volume, which is calculated as the volume of the mask region per one Dalton of the molecular mass of the object under study. In the evaluation of solvent amount in the unit cell, the specific volume of the molecule region is usually supposed to be 1.23 $Å^3Da^{-1}$ [39, 40].



a) **b**) **c**)

Fig. 1. The mask generating procedure. The random mask is built point by point. **a**) A seed point is taken randomly. **b**) The boundary points (open circles) are shown for a partial mask. **c**) To extend the mask, one of the boundary points is selected randomly (the red filled circle). The set of the boundary points is corrected (red open circles).

To run this procedure, it is necessary to specify the criteria by which the acceptable masks are selected. The simplest type of this criterion is the correlation-type criterion

$$CM_{(d_{\max},d_{\min})} = \frac{\sum_{\mathbf{s}} F^{obs}(\mathbf{s}) F^{mask}(\mathbf{s})}{\sqrt{\sum_{\mathbf{s}} \left(F^{obs}(\mathbf{s})\right)^2 \sum_{\mathbf{s}} \left(F^{mask}(\mathbf{s})\right)^2}}, \tag{11}$$

where the summation extends over the reflections with $d_{\max}^{-1} < |\mathbf{s}| \le d_{\min}^{-1}$. A generated mask is considered as acceptable if this value exceeds the specified level CM^{crit}. More sophisticated measures of the quality of the mask are possible as well, for example, the likelihood of the mask [41, 42]. In the case of testing of a method using an object with the known structure, the true phases are known, and the found phases can be compared with the true ones. A popular criterion of phase similarity is the correlation coefficient of two electron density distributions calculated with the same observed set of structure factor magnitudes, but with different phases [43]. This Map Correlation Coefficient can be calculated, as well, in terms of structure factors as the weighted sum of cosines of phase differences

$$CP_{(d_{\max},d_{\min})} = \frac{\sum_{\mathbf{s}} \left(F^{obs}(\mathbf{s})\right)^2 \cos\left(\varphi^{true}(\mathbf{s}) - \varphi^{calc}(\mathbf{s})\right)}{\sum_{\mathbf{s}} \left(F^{obs}(\mathbf{s})\right)^2}, \tag{12}$$

where the calculation of the Fourier syntheses and summation extends to the reciprocal space region $d_{\max}^{-1} < |\mathbf{s}| \le d_{\min}^{-1}$.

It should be kept in mind that, in view of the fundamental ambiguity of the phase problem solution, in the course of random mask generation we may find acceptable masks that seem differently, but become similar after a shift of the origin or the change of the enantiomer. Therefore, before the averaging or formal calculation of the correlation, the sets of phases should be aligned taking into account the trivial ambiguity group, which consists of all shifts and the enantiomer change. After the structure factors corresponding to the selected masks have been aligned, the 'best' phases $\varphi^{best}(\mathbf{s})$ and their 'figure of merits' $m(\mathbf{s})$ are defined for all reflections as

$$m(\mathbf{s})\exp\left[i\varphi^{best}(\mathbf{s})\right]=\sum_{j=1}^{M}\exp\left[i\varphi_j(\mathbf{s})\right]. \tag{13}$$

Here $\varphi_j(\mathbf{s})$ are structure factor phases corresponding to *j*-th of *M* selected masks.

Figure 2 shows the quality of phase sets obtained after the alignment and averaging of the phase sets corresponding to 100 masks selected on the basis of the magnitudes correlation [37]. Different curves correspond to different cut-off values for magnitude correlation when selecting the masks. The quality of phases grows with the increase in the selection stiffness. It is worthy of noting that the averaging of structure factor phases for connected masks of the proper volume has some phasing power even if no selection was applied at all. The tests performed with different values of the specific volume of masks show that the specific volume, optimal for the phasing at very low resolution, slightly exceeds the customary estimate 1.23 $Å^3Da^{-1}$.

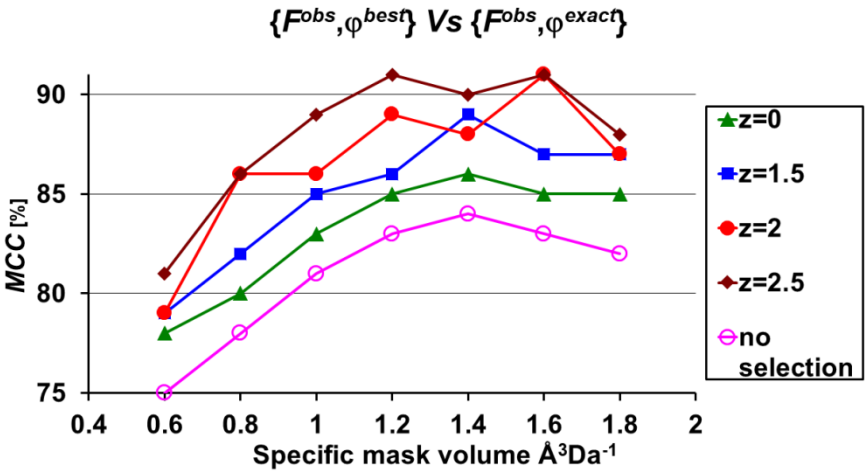
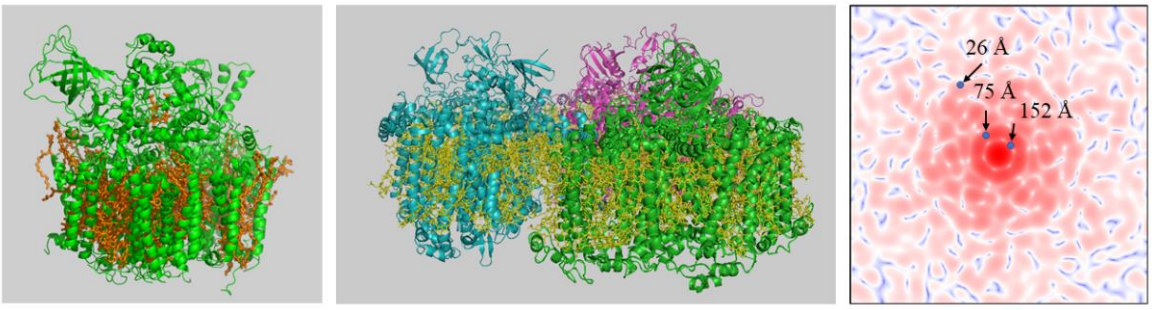


Fig. 2. The results of the test phasing of simulated data for monomeric Photosystem II (Fig. 3). Map correlation of the found phases and the exact ones in 25 Å resolution zone is shown as the function of the mask specific volume. The final phases $\varphi^{best}(\mathbf{s})$ were obtained by the averaging (13) of the phases for the masks selected with the cutoff level CM^{crti} = MEAN(*CM*) + *z*·RMSD(*CM*), where the mean and RMSD of *CM*-value were calculated for all generated masks. (See [37] for details).



a) **b**) **c**)

Fig. 3. The test objects: **a**) Monomeric Photosystem II ([44], Pdb code 3kzi, *MW* = 300 kDa). **b**) Trimer of Photosystem I ([45], PDB code 1jbo, *MW* = 1068 kDa). **c**) One of the sections of simulated data *I*(**s**) for the trimer of Photosystem I.

The found phase values can be further refined and the set of phases may be expanded [37, 41, 46]. The construction of the mask is executed point by point (Fig. 1). The selection of the next point is performed from the boundary points of the already constructed partial mask. This choice can be made with equal probabilities, or taking into account the predefined probabilities.

The prior probabilities of belonging to the molecule region may be defined for the cell points on the basis of Fourier synthesis $\rho(\mathbf{r})$ obtained at the previous step of phasing

$$P(\mathbf{r}) = C\exp\left[\kappa\rho(\mathbf{r})\right], \tag{14}$$

where parameter κ defines the desired sharpness of the probability distribution. This forms the basis for iterative phase refinement and extension. Figure 4 shows the progress in phase refinement and extension. The graphs show the correlations of the phases obtained at different stages of the refinement, calculated in spherical resolution shells in the reciprocal space. Figure 5 shows the images of the particle shown by Fourier syntheses at 25 Å resolution. The syntheses were calculated with the exact phases and phases obtained at the last step of the phasing procedure. The shown surfaces correspond to the specific volume 1.23 $Å^3Da^{-1}$.

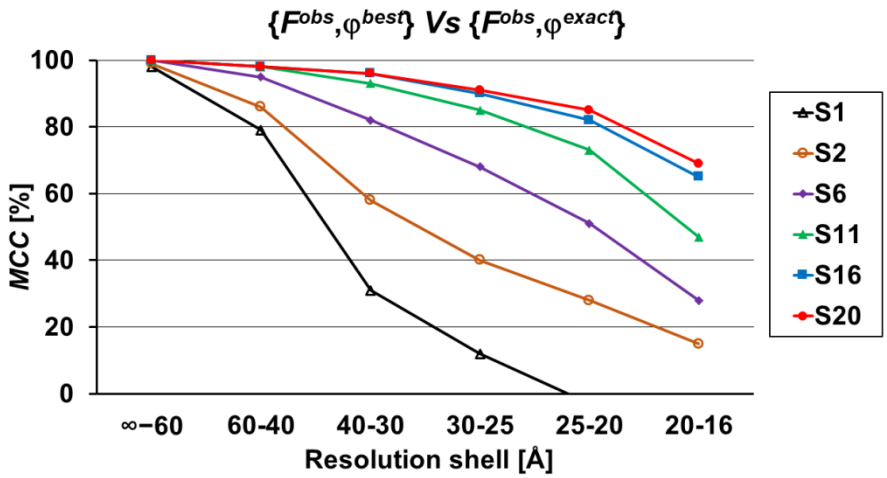
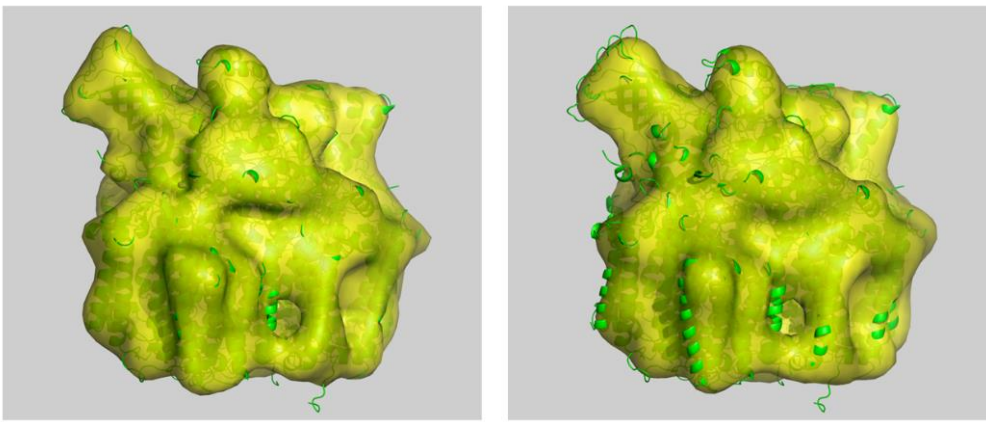


Fig. 4. The results of the phase refinement and resolution extension in the test phasing of simulated data for monomeric Photosystem II (Fig. 3). Map correlations of the found phases and the exact ones in successive resolution shells are shown for different steps (S1–S20) of the work. The phases $\varphi^{best}(\mathbf{s})$ were obtained by the averaging (13) of the phases for the masks selected at the corresponding refinement step. (See [41] for details).



a) **b**)

Fig. 5. The images of the particle. 25 Å resolution The Fourier syntheses were calculated with the exact structure factor magnitude and different sets of phases at 16 Å resolution. The both surfaces correspond to the same specific volume 1.23 $Å^3Da^{-1}$. The protein part of the complex is shown as a cartoon. **a**) The synthesis was calculated with the exact phases. **b**) The synthesis was calculated with the phases obtained at the step S20 of the phasing. (See [41] for details).

6. THE RESTORING OF THE LOST REFLECTIONS

When working with observed data, it often occurs that the lowest-resolution data are lost and some strong reflections are "oversaturated". The values of structure factor magnitudes of such reflections, as well as their phases, can be restored within the framework of the proposed procedure and used further for the calculation of electron density syntheses. When we calculate the value of the selection criterion, we naturally can take into account only those values that were measured in the experiment. However, for selected masks, we can calculate the magnitudes and phases for all required reflections. Therefore, the lost structure factor magnitudes can be estimated by the values of the mask magnitudes averaged through the selected masks. An example is shown in Figure 6. In this test with the trimer of photosystem I, 182 reflections of the central zone up to resolution of 136 Å and the 66 strongest among the remaining reflections were considered unknown and were excluded from the phasing. The diagram shows the values for a resolution-ordered list of initially lost reflections. Blue dots correspond to exact values. Open circles correspond to restored values. The standard R-factor between the restored and exact values of the structure factors magnitudes was 7 %, the mean phase error in the found phase-values was 12 degrees.

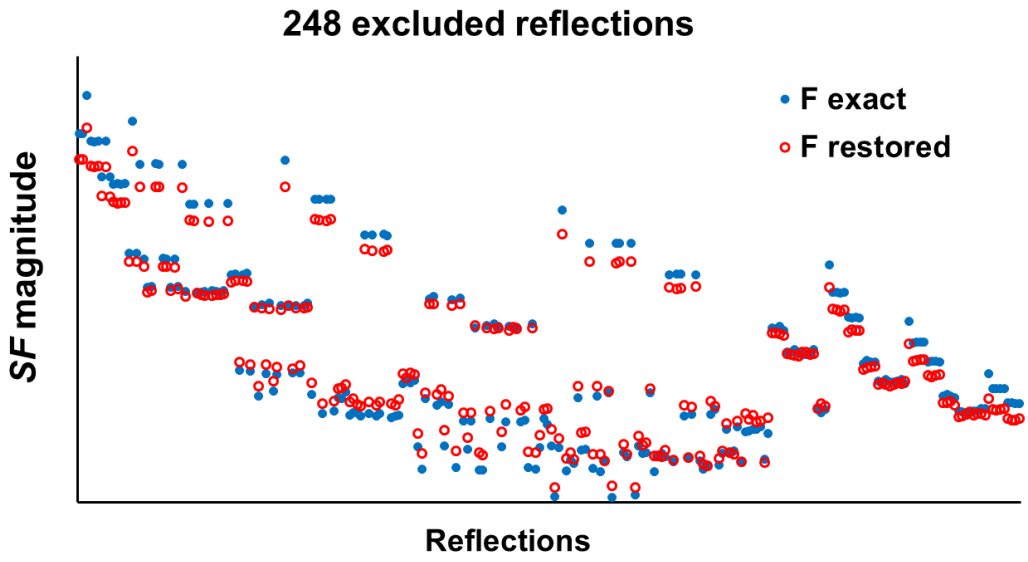


Fig. 6. The results of the retrieval of unknown structure factor magnitudes in the test with simulated data for the trimer of Photosystem I (Fig. 3). The restored and exact values of the magnitude are shown for 248 reflections declared as unknown in the test. The reflections are presented in the resolution order (increasing |**s**| value order).

Table 1. Map Correlation Coefficient in resolution shells and an estimated resolution

Step	Resolution shell (Å)						Estimated resolution (Å)	
	∞–60	60–40	40–30	30–25	25–20	20–16	by *MCC*[*]	by *MTF*[**]
S6	100	95	82	68	51	28	20	40
S11	100	98	93	85	73	47	20	30
S20	100	98	96	91	85	69	16	25
# refl.	85	170	363	436	1028	2026		

[*]Map Correlation Coefficient.
[**]Model Trapping Function

7. MODEL-TRAPPING FUNCTION AS AN ESTIMATE OF THE SYNTHESIS QUALITY

There is a common, though seriously criticized [47, 48], practice of considering phases in some resolution shell as reliable if correlation coefficient between two independently obtained solutions exceeds 50 %. Table 1 shows the shell correlation of the true phases with three solutions obtained at different stages of phasing for the photosystem II test structure. Based on this table, the resolution of the syntheses obtained at steps S6 and S11 can be evaluated as 20 Å, and as 16 Å for step S20. However, the answer will be different if we try to estimate resolution by the usefulness of the synthesis for building a model of the structure [41].

A Fourier synthesis is favorable for model building if the molecule region determined as the highest synthesis values covers the atomic positions, but does not include the parts occupied by the bulk solvent. Such a situation may occur for high-resolution syntheses, but is unattainable at low resolution. Figure 7 shows the regions of the same volume built as the regions of highest values $\{\mathbf{r}:\rho(\mathbf{r})>\rho_{crit}\}$ in the exact Fourier syntheses of different resolution. It can be seen, *e.g.*, at the resolution of 3 Å, that this region embraces the model well, really reproducing the region of the molecule. As the resolution of synthesis decreases, some of the atoms in the model begin to creep out of the region. The number of such atoms increases with the decrease of the resolution. We can estimate the quality of a mask as the approximation of the molecule region by the percentage of model atoms captured by this mask. This estimate may be calculated for regions of different specific volumes. We define the Model Trapping Function (*MTF*) as the function of specific volume κ as follows. Let the cutoff level ρ_κ be the value chosen for the function $\rho(\mathbf{r})$ and the specified value κ so that the specific volume of the region $\Omega_\kappa=\{\mathbf{r}:\rho(\mathbf{r})\geq\rho_\kappa\}$ is equal to κ. We define

$$MTF(\kappa)=\frac{\text{number of atoms in the }\Omega_\kappa\text{ region}}{\text{total number of atoms in the model}}\ . \tag{15}$$

The MTF shows how the percentage of trapped atoms of the model grows with the increase in the specific volume of the region. Figure 8 shows model trapping curves for the exact syntheses of different resolution calculated for the test model of monomeric photosystem I.

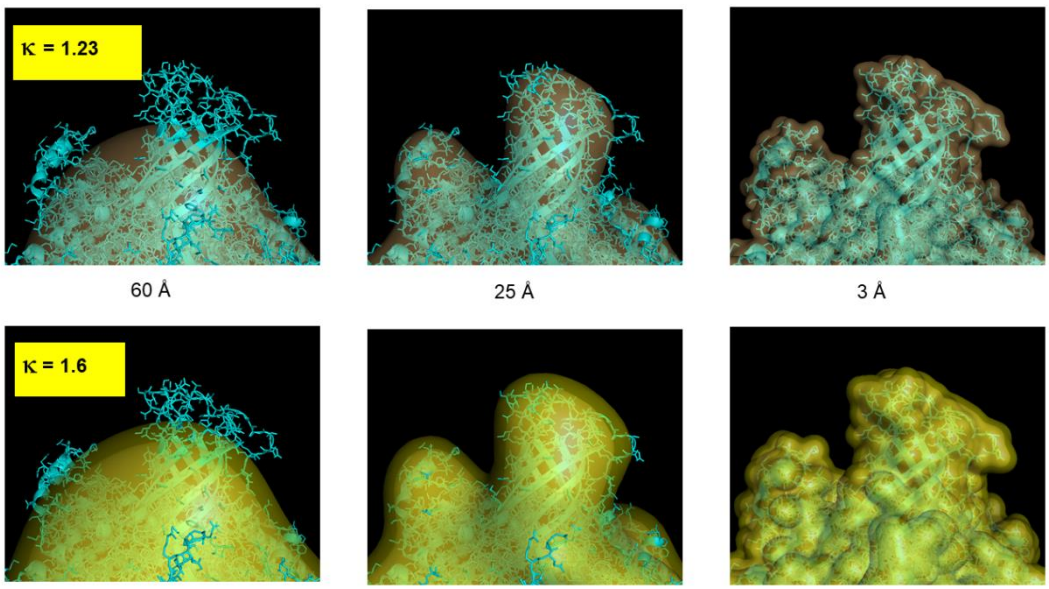


Fig. 7. Molecule region mask build with the use of the exact Fourier syntheses of increasing (from left to right) resolution. The masks corresponding to two different values κ of the specific volume are shown.

Now we can approach the evaluation of the results obtained at different steps of the test phasing on the new basis. For each set of phases obtained, for example, S6, we can calculate a model trapping curve and compare it with reference curves corresponding to the exact syntheses of different resolution. For example, Figure 9 shows that the curve for the step S6 is close to the curve corresponding to 40 Å resolution exact synthesis. This means that in terms of the quality of the molecule region revealing, this synthesis obtained at the step S6 is close to the synthesis of 40 Å resolution. A similar analysis for the steps S11 and S20 gives an assessment of the quality of the syntheses as 30 and 25 Å, respectively. Thus, an attempt to evaluate the result by the quality of the selection of the region of the molecule gives a significantly lower resolution than the 50 % rule. Table 1 indicates that this rather corresponds to the 90 % rule, which is radically different from the traditional assessment.

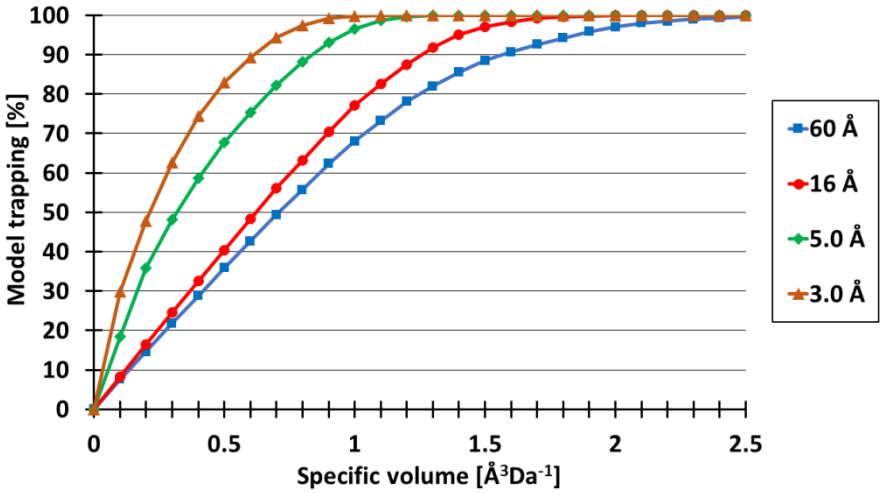


Fig. 8. The model trapping functions corresponding to the exact Fourier synthesis of different resolution calculated with the simulated data for monomeric photosystem I [41].

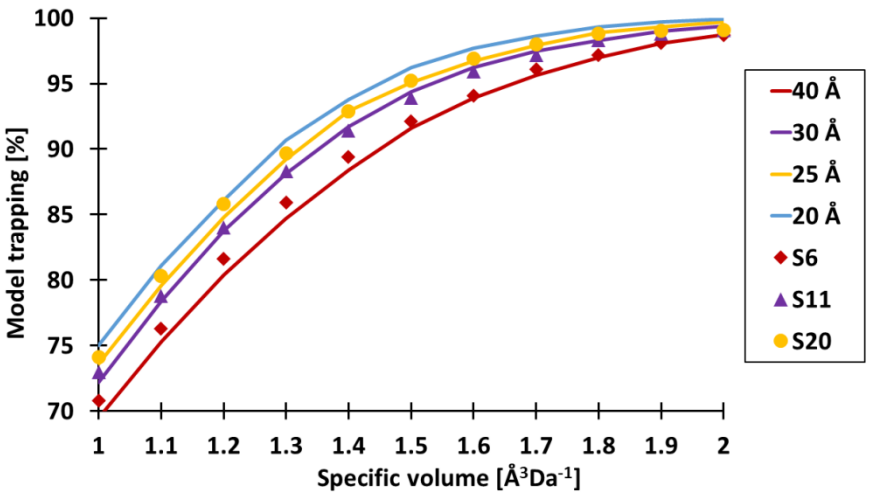


Fig. 9. The evaluation of the resolution for results of the test phasing of simulated data for monomeric photosystem I. The model trapping curves calculated for results of different steps in the phasing are shown by markers. The model trapping curves corresponding to the exact syntheses of different resolution are shown by solid lines [41].

8. CONCLUSIONS

Generally, the phasing of single particle diffraction data is similar to one in biological crystallography. The main difference is that the recording of a continuous diffraction pattern provides the data of a high redundancy, which significantly improve the phasing and allows

one to extend the diffraction pattern to regions in the reciprocal space not covered by the particular experiment. The redundancy of the data may be utilized by, e.g., density modification methods, Shannon-type interpolation formulas, mask-based approach or other methods of phasing and intensity extension. In developing and testing new approaches in macromolecular studies it is important to have adequate criteria to evaluate the success. Model trapping function seems to be more sensitive for macromolecular objects than the customary formal map correlation coefficient. In particular, it shows that usual 50% shell correlation rule when applied in macromolecular studies gives too optimistic estimate of electron density map quality.

REFERENCES

1. Adams P.D., Afonine P.V., Bunkóczi G., Chen V.B., Davis I.W., Echols N., Headd J.J., Hung L.-W., Kapral G.J., Grosse-Kunstleve R.W. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica D*. 2010. V. 66. P. 213–221. doi: 10.1107/S0907444909052925.
2. Winn M.D., Ballard C.C., Cowtan K.D., Dodson E.J., Emsley P., Evans P.R., Keegan R.M., Krissinel E.B., Leslie A.G.W., McCoy A. et al. Overview of the CCP4 suite and current developments. *Acta Crystallographica D*. 2011. V. 67. P. 235–242. doi: 10.1107/S0907444910045749.
3. Sheldrick G.M. A short history of SHELX. *Acta Crystallographica A*. 2008. V. 64. P. 112–122. doi: 10.1107/S0108767307043930.
4. Bricogne G., Vonrhein C., Flensburg C., Schiltz M., Paciorek W. Generation, representation and flow of phase information in structure determination: recent developments in and around SHARP 2.0. *Acta Crystallographica D*. 2003. V. 59. P. 2023–2030. doi: 10.1107/S0907444903017694.
5. Blanc E., Roversi P., Vonrhein C., Flensburg C., Lea S.M., Bricogne G. Refinement of severely incomplete structures with maximum likelihood in BUSTER-TNT. *Acta Crystallographica D*. 2004. V. 60. P. 2210–2221. doi: 10.1107/S0907444904016427.
6. Minor W., Cymborowski M., Otwinowski Z., Chruszcz M. HKL-3000: the integration of data reduction and structure solution - from diffraction images to an initial model in minutes. *Acta Crystallographica D*. 2006. V. 62. P. 859–866. doi: 10.1107/S0907444906019949.
7. Spence J.C.H. XFELs for structure and dynamics in biology. *IUCrJ*. 2017. V. 4. P. 322–339. doi: 10.1107/S2052252517005760.
8. Standfuss J., Spence J. Serial crystallography at synchrotrons and X-ray lasers. *IUCrJ*. 2017. V. 4. P. 100–101. doi: 10.1107/S2052252517001877.
9. Aquila A., Barty A., Bostedt C., Boutet S., Carini G., dePonte D., Drell P., Doniach S., Downing K.H., Earnest T. et al. The linac coherent light source single particle imaging road map. *Structural Dynamics*. 2015. V. 2. doi: 10.1063/1.4918726.
10. Ayyer K., Geloni G., Kocharyan V., Saldin E., Serkez S., Yefanov O., Zagorodnov I. Perspectives for imaging single protein molecules with the present design of the European XFEL. *Structural Dynamics*. 2015. V. 2. doi: 10.1063/1.4919301.
11. Daurer B.J., Okamoto K., Bielecki J., Maia F.R.N.C., Muhlig K., Seibert M.M., Hantke M.F., Nettelblad C., Benner W.H., Svenda M. et al. Experimental strategies for imaging bioparticles with femtosecond hard X-ray pulses. *IUCrJ*. 2017. V. 4. P. 251–262. doi: 10.1107/S2052252517003591.
12. Lunin V.Y., Lunina N.L., Petrova T.E. The biological crystallography without crystals. *Mathematical Biology and Bioinformatics*. 2017. V. 12. No. 1. P. 55–72. doi: 10.17537/2017.12.55.
13. Lunin V.Y. Mask-based approach to restoring and phasing single-particle diffraction data. In: *32nd European Crystallographic Meeting, Vienna, Austria, August 18-23*: Abstract Booklet. 2019. P. 138.

14. Lunin V.Y., Lunina N.L., Petrova T.E. Single particle study by X-ray diffraction: Crystallographic approach. *Mathematical Biology and Bioinformatics*. 2019. V. 14. No. 2. P. 500–516. doi: 10.17537/2019.14.500.
15. Urzhumtseva L., Klaholz B., Urzhumtsev A. On effective and optical resolutions of diffraction data sets. *Acta Crystallographica D*. 2013. V. 69. P. 1921–1934. doi: 10.1107/S0907444913016673.
16. Kucukelbir A., Sigworth F.J., Tagare H.D. Quantifying the local resolution of cryo-EM density maps. *Nature Methods*. 2014. V. 11. P. 63–65. doi: 10.1038/nmeth.2727.
17. Afonine P.V., Klaholz B.P., Moriarty N.W., Poon B.K., Sobolev O.V., Terwilliger T.C., Adams P.D., Urzhumtsev A. *Acta Crystallographica D*. 2018. V. 74. P. 814–840. doi: 10.1107/S2059798318009324.
18. Meijering E. A chronology of interpolation: from ancient astronomy to modern signal and image processing. *Proceedings of the IEEE*. 2002. V. 90. P. 319–342. doi: 10.1109/5.993400.
19. Kotel'nikov V.A. On the transmission capacity of 'ether' and wire in electric communications. *Physics-Uspekhi*. 2006. V. 49. № 7. P. 736–744. doi: 10.1070/PU2006v049n07ABEH006160.
20. Sayre D. Some implications of a theorem due to Shannon. *Acta Crystallographica*. 1952. V. 5. P. 843. doi: 10.1107/S0365110X52002276.
21. Bricogne G. Geometric sources of redundancy in intensity data and their use for phase determination. *Acta Crystallographica A*. 1974. V. 30. P. 395–405. doi: 10.1107/S0567739474010722.
22. Bricogne G. Methods and programs for direct-space exploitation of geometric redundancies. *Acta Crystallographica A*. 1976. V. 32. P. 832–847. doi: 10.1107/S0567739476001691
23. Lunin V.Y., Lunina N.L. Repairing of the diffraction pattern in the X-ray free electron laser study of biological particles. *Advanced Mathematical Models & Applications*. 2018. V. 3. P. 117–127.
24. Lunin V.Y. Use of the fast differentiation algorithm for phase refinement in protein crystallography. *Acta Crystallographica. A*. 1985. V. 41. P. 551–556. doi: 10.1107/S0108767385001209.
25. Podjarny A.D., Rees B., Urzhumtsev A.G. Density modification in X-ray crystallography. In: *Methods in Molecular Biology, Crystallographic Methods and Protocols*. Eds. Jones C., Milloy B, Sanderson M.R. Totowa, New Jersey: Humana Press, 1996. P. 205-226. (Methods in Molecular Biology, Vol. 56.). doi: 10.1385/0-89603-259-0:205.
26. Zhang K.Y.J., Cowtan K.D., Main P. Phase improvement by iterative density modification. In: *International Tables for Crystallography*. Vol. F. Eds. Arnold E., Himmel D.M., Rossmann M.G. Chichester: John Wiley and Sons, 2012. P. 385–400. doi: 10.1107/97809553602060000847.
27. Wang B.C. Resolution of phase ambiguity in macromolecular crystallography. *Methods in Enzymology*. 1985. V. 115. P. 90–111. doi: 10.1016/0076-6879(85)15009-3.
28. Fienup J.R. Reconstruction of an object from the modulus of its Fourier transform. *Optics Letters*. 1978. V. 3. N. 1. P. 27–29. doi: 10.1364/OL.3.000027.
29. Marchesini S. A unified evaluation of iterative projection algorithms for phase retrieval. *Rev. Sci. Instrum*. 2007. V. 78. Article No. 011301. doi: 10.1063/1.2403783.
30. Millane R., Lo V.L. Iterative projection algorithms in protein crystallography. I. Theory. *Acta Crystallographica A*. 2013. V. 69. P. 517–527. doi: 10.1107/S0108767313015249.
31. Abrahams J.P. Bias reduction in phase refinement by modified interference functions: introducing the γ-correction. *Acta Crystallographica D*. 1997. V. 53. P. 371–376. doi: 10.1107/S0907444996015272.
32. Oslányi G., Sütő A. *Ab initio* structure solution by charge flipping. *Acta Crystallographica A*. 2004. V. 60. P. 134–141. doi: 10.1107/S0108767303027569.

33. Maia F.R.N.C., Ekeberg T., Spoel D., Hajdu J. Hawk: the image reconstruction package for coherent X-ray diffractive imaging. *J. Applied Crystallography*. 2010. V. 43. P. 1535–1539. doi: 10.1107/S0021889810036083.
34. Urzhumtsev A.G. *The use of local averaging in analysis of macromolecule images at electron density distribution maps*: Preprint. Pushchino, 1985 (in Russ.).
35. Urzhumtsev A.G., Lunin V.Y., Luzyanina T.B. Bounding a Molecule in a Noisy Synthesis. *Acta Crystallographica A*. 1989. V. 45. P. 34–39. doi: 10.1107/s0108767388008955.
36. Marchesini S., He H., Chapman H.N., Hau-Riege S.P., Noy A., Howells M.R., Weierstall U., Spence J.H.C. X-ray image reconstruction from a diffraction pattern alone. *Phis. Rev. B.* 2003. V. 68. Article No. 140101(R). doi: 10.1103/PhysRevB.68.140101.
37. Lunin V.Y., Lunina N.L., Petrova T.E., Baumstark M.W., Urzhumtsev A.G. Mask-based approach to phasing of single-particle diffraction data. *Acta Crystallographica D.* 2016. V. 72. P. 147–157. doi: 10.1107/S2059798315022652.
38. Lunin V.Y., Lunina N.L., Petrova T.E. The use of connected masks for reconstructing the single particle image from X-ray diffraction data. *Mathematical Biology and Bioinformatics*. 2014. V. 10. № Suppl. P. t1–t19. doi: 10.17537/2015.10.t1.
39. Matthews B.W. Solvent Content of Protein Crystals. *Journal of Molecular Biology*. 1968. V. 33. P. 491–497. doi: 10.1016/0022-2836(68)90205-2.
40. Weichenberger C.X., Afonine P.V., Kantardjieff K., Rupp B. *Acta Crystallographica D.* 2015. V. 71. P. 1023–1038. doi: 10.1107/S1399004715006045.
41. Lunin V.Y., Lunina N.L., Petrova T.E., Baumstark M.W., Urzhumtsev A.G. Mask-based approach to phasing of single-particle diffraction data. II. Likelihood-based selection criteria. *Acta Crystallographica D.* 2019. V. 75. P. 79–89. doi: 10.1107/S2059798318016959.
42. Lunina N.L., Petrova T.E., Urzhumtsev A.G., Lunin V.Y. The Use of Connected Masks for Reconstructing the Single Particle Image from X-Ray Diffraction Data. III. Maximum-Likelihood Based Strategies to Select Solution of the Phase Problem. *Mathematical Biology and Bioinformatics*. 2018. V. 13. № Supl. P. t70–t83. doi: 10.17537/2018.13.t70.
43. Lunin V.Y., Woolfson M.M. Mean Phase Error and the Map Correlation Coefficient. *Acta Crystallographica D*. 1993. V. 49. P. 530–533. doi: 10.1107/S0907444993005852.
44. Broser M., Gabdulkhakov A., Kern J., Guskov A., Müh F., Saenger W., Zouni A. Crystal structure of monomeric Photosystem II from Thermosynechococcus elongatus at 3.6 Å resolution. *J. Biol. Chem*. 2010. V. 285. P. 26255–26262. doi: 10.1074/jbc.M110.127589.
45. Jordan P., Fromme P., Witt H.T., Klukas O., Saenger W., Krauß N. Three-dimensional structure of cyanobacterial photosystem I at 2.5 Å resolution. *Nature*. 2001. V. 411. P. 909–917. doi: 10.1038/35082000.
46. Lunina N.L., Petrova T.E., Urzhumtsev A.G., Lunin V.Y. The use of connected masks for reconstructing the single particle image from X-ray diffraction data. II. The dependence of the accuracy of the solution on the sampling step of experimental data. *Mathematical Biology and Bioinformatics*. 2015. V. 10. № Suppl. P. t56–t72. doi: 10.17537/2015.10.t56.
47. Van Heel M., Schatz M. Fourier shell correlation threshold criteria. *J. Struct. Biol*. 2005. V. 151. P. 250–262. doi: 10.1016/j.jsb.2005.05.009.
48. Van Heel M., Schatz M. Reassessing the Revolution's Resolutions. *bioRxiv*. 2017. Article No. 224402. doi: 10.1101/224402.

Received 19.11.2019.
Revised 19.02.2020.
Published 02.03.2020.

Восстановление модулей и расчет фаз для дифракционной картины изолированной частицы с использованием бинарных масок объекта

Лунин В.Ю., Лунина Н.Л., Петрова Т.Е.

Институт математических проблем биологии РАН - филиал Института прикладной математики им. М.В. Келдыша РАН, Пущино, Московская область, Россия

Аннотация. Развитие экспериментальной техники и, в частности, ввод в эксплуатацию рентгеновских лазеров на свободных электронах позволяют приблизиться к возможности регистрации рентгеновского рассеяния отдельной макромолекулярной частицей. Это открывает дорогу к определению методами рентгеновской дифракции структуры некристаллизованных макромолекулярных объектов. Возможность измерения интенсивностей не-Брэгговских рефлексов создает существенную избыточность экспериментальных данных, что существенно упрощает определение структуры объекта. Дискретизация непрерывной дифракционной картины на сетку с достаточно мелким шагом позволяет рассматривать проблему определения структуры как проблему определения структуры для "виртуального" кристалла с чрезвычайно большим относительным объемом растворителя в элементарной ячейке. В предположении, что область, занимаемая объектом в элементарной ячейке, известна, это позволяет ожидать высокой эффективности в решении фазовой проблемы итерационных методов, типа методов модификации электронной плотности. В то же время, итерационные методы чувствительны к точности задания области молекулы, неполноте экспериментальных данных и изначальной неединственности решения. Разработанный авторами метод предварительного решения фазовой проблемы осуществляет случайный поиск связных бинарных аппроксимаций распределения электронной плотности в объекте (масок области молекулы), воспроизводящих с достаточной точностью дифракционную картину, наблюдаемую в эксперименте. Выравнивание, в рамках группы эквивалентности решений фазовой проблемы, найденных масок с последующим усреднением позволяет получить приближенное решение фазовой проблемы. Помимо оценки неизвестных значений фаз структурных факторов разработанный подход позволяет восстанавливать фрагменты дифракционной картины (значения модулей структурных факторов), потерянные в эксперименте. Примерами таких фрагментов могут служить нерегистрируемая центральная зона рентгенограммы или области "переэкспонированных" (ввиду ограниченности рабочего диапазона детектора) рефлексов.

Ключевые слова: *биологические макромолекулы, изолированные частицы, рентгеновское рассеяние, рентгеновские лазеры, фазовая проблема, восстановление дифракционных данных, эффективное разрешение.*