Mathematical Biology and Bioinformatics
2015. V. 10. № Suppl. P. t56–t72. doi: 10.17537/2015.10.t56

The translation of the original article
2015 Lunina N.L., Petrova T.E., Urzhumtsev A.G., Lunin V.Y.
Matematicheskaya biologiya i bioinformatika. 2015. V. 10. No 2. P. 508–525. doi: 10.17537/2015.10.508.

========== **MATHEMATICAL MODELING** ==========

УДК: 577.3

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

Lunina N.L.[1], Petrova T.E.[1], Urzhumtsev A.G.[*2,3], Lunin V.Y.[1]**

[1]*Institute of Mathematical Problems of Biology, Moscow Region, Pushchino, Russia*
[2]*Institut de Génétique et de Biologie Moléculaire et Cellulaire, Illkirch, France*
[3] *Université de Lorraine, Vandoeuvre-lès-Nancy, France*

Abstract. Advances in the methodology of the X-ray diffraction experiments leads to a possibility to register the rays scattered by large isolated biological particles (viruses and individual cells) but not only by crystalline samples. The experiment with an isolated particle provides researchers with the intensities of the scattered rays for the continuous spectrum of scattering vectors. Such experiment gives much more experimental data than an experiment with a crystalline sample where the information is limited to a set of Bragg reflections. This opens up additional opportunities in solving underlying problem of X-ray crystallography, namely, calculating phase values for the scattered waves needed to restore the structure of the object under study. In practice, the original continuous diffraction pattern is sampled, reduced to the values at grid points in the space of scattering vectors (in the reciprocal space). The sampling step determines the amount of the information involved in solving the phase problem and the complexity of the necessary calculations. In this paper, we investigate the effect of the sampling step on the accuracy of the phase problem solution obtained by the method proposed earlier by the authors. It is shown that an expected improvement of the accuracy of the solution with the reducing the sampling step continues even after crossing the 'Nyquist limit' defined as the inverse of the double size of the object under study.

Key words: *X-ray crystallography, phase problem, XFEL, single particle diffraction.*

[*]sacha@igbmc.fr
[**]lunin@impb.psn.ru

1. X-RAY DIFFRACTION BY CRYSTALS AND SINGLE PARTICLES

1.1. X-ray diffraction experiment

The goal of the first stage of solving the structure of a biological object by X-ray diffraction is to determine the function that describes the distribution of scattering electrons in the sample being studied. After determining this function, at least approximately, its interpretation results in a preliminary atomic model of the object. At the last stage of the structure determination, the parameters of the model are refined to minimize the discrepancy between the experimental data and the theoretical scattering pattern corresponding to the model.

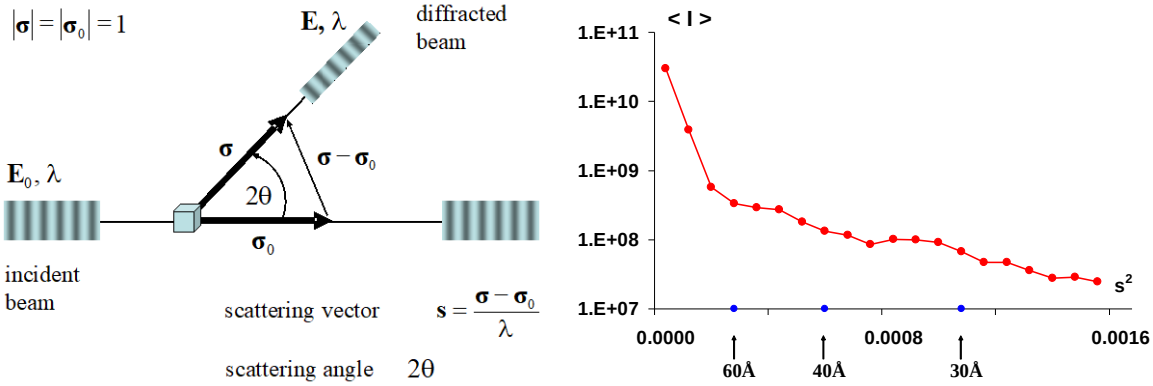

Fig. 1. Left: the scheme of an X-ray diffraction experiment. Right: The dependence of the intensity of reflections on resolution. For thin resolution shells in the reciprocal space, the mean intensity versus squared shell radius $s^2 = (2\sin\theta/\lambda)^2$ is shown. The plot corresponds to the test object PS-I in *c*1 cell (see Sec. 3).

The scheme of an X-ray diffraction experiment is shown in Figure 1. In this experiment, the sample is placed on the path of a primary X-ray wave with the wavelength λ and the direction $\boldsymbol{\sigma}_0$. A detector measures the intensity of secondary waves specified by directions $\boldsymbol{\sigma}$. The secondary waves are the result of a superposition of spherical waves (of the same wavelength λ) emitted by oscillating electrons excited by the primary wave. The complex amplitude $\mathbf{E}$ of a secondary wave differs from the amplitude $\mathbf{E}_0$ of a primary wave by two principal factors:

$$\mathbf{E} = \varepsilon \mathbf{F}(\mathbf{s})\mathbf{E}_0 . \tag{1}$$

The factor ε does not depend on the atomic structure of the object and is defined mostly by the part of the primary wave energy flow that comes with the wave scattered by one electron in the direction $\boldsymbol{\sigma}$. This part is extremely small (the factor ε in (1) can be as small as 10^{-12}), which creates the main difficulty to register the scattered radiation. The complex factor $\mathbf{F}(\mathbf{s})$, which is named the structure factor, depends on the distribution of electrons $\rho(\mathbf{r})$ in the object and experimental conditions (the direction of the primary and secondary waves and the wavelength). In the absence of anomalous scattering, these quantities are related by the equations:

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

$$\mathbf{s} = \frac{\boldsymbol{\sigma} - \boldsymbol{\sigma}_0}{\lambda} . \tag{3}$$

The quantities $F(\mathbf{s})$ and $\varphi(\mathbf{s})$ are called the magnitude and the phase of the structure factor, and $\mathbf{s}$ is the scattering vector.

Experimentally measured secondary wave intensity is proportional to the square of the structure factor magnitude with the coefficient C, which is common for all structure factors

$$I(\mathbf{s}) = CF^2(\mathbf{s}). \tag{4}$$

An X-ray diffraction experiment allows one to obtain the structure factor magnitudes. At the same time, the phase values cannot be measured in the standard X-ray diffraction experiment. The retrieval of the phase values is the central problem of the X-ray structure analysis, the so-called phase problem.

1.2. Scattering by a crystalline sample

The use of a crystalline sample in an X-ray diffraction experiment has a dual effect on the scattering pattern. On the one hand, for a discrete subset of scattering vectors, the waves scattered by different copies of the object come to the detector with the same phase, which leads to a multiple increase in the intensity of the secondary wave so that the intensity becomes measurable. These waves are called Bragg reflections. For other scattering vectors, the waves from different object copies come with chaotic phase shifts leading to their mutual annihilation. A very low intensity of these waves, in practice, cannot be measured.

For a crystalline sample, the electron density distribution is a periodic function, and can be represented as a Fourier series:

$$\rho(\mathbf{r}) = \frac{1}{|\mathbf{V}_{cell}|}\sum_{\mathbf{s}\in\Re'}\mathbf{F}(\mathbf{s})\exp[-2\pi i(\mathbf{s},\mathbf{r})]. \tag{5}$$

The summation here is over all nodes of the integer lattice $\Re' = \{h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*,\ h,k,l - \text{integers}\}$ in the scattering vector space (reciprocal space lattice). This lattice is built on the basis of vectors $\{\mathbf{a}^*,\mathbf{b}^*,\mathbf{c}^*\}$, which form the conjugate basis to the basis $\{\mathbf{a},\mathbf{b},\mathbf{c}\}$ of a crystal unit cell. Vectors $\{\mathbf{a},\mathbf{b},\mathbf{c}\}$ are linearly independent periods of the function $\rho(\mathbf{r})$. They are usually selected so that the volume of the unit cell (a parallelepiped based on these vectors) is minimal possible.

Bragg reflection appears if the scattering vector belongs to the reciprocal space lattice $\mathbf{s}\in\Re'$ (the Bragg-Laue-Wulff condition). The corresponding structure factors can be represented as

$$\mathbf{F}(\mathbf{s}) = F(\mathbf{s})\exp[i\varphi(\mathbf{s})] = N\int_{\mathbf{V}_{cell}}\rho(\mathbf{r})\exp[2\pi i(\mathbf{s},\mathbf{r})]dV_{\mathbf{r}}, \tag{6}$$

where N is a factor common to all Bragg reflections. It is proportional to the number of copies of the object in the crystal and enhances many times the magnitude of the secondary wave. The problem of the retrieval of the distribution $\rho(\mathbf{r})$ may be reformulated as the problem of the determination of the phase values necessary to calculate the distribution (5).

In crystal studies, an X-ray experiment provides one with only half of the information needed to reconstruct the electron density distribution by formula (5), i.e. the structure factor magnitudes for scattering vectors $\mathbf{s}\in\Re'$. The restoration of the missing half of the information, i.e. the structure factors phases, requires additional information. Usually, it comes from complementary experiments carried out in modified conditions or from the general properties of the object.

The need to use crystalline samples in the experiment is caused by the fact that the secondary waves scattered by only one copy of the object are too weak to be experimentally measured. Until recently, the experimental equipment did not allow one to register the

scattering of isolated biological molecules. However, the development of new powerful impulse radiation sources, X-ray free electron lasers (XFEL), and the progress in detector technology give grounds to expect that this difficulty will be overcome. The first experimental data sets obtained in diffraction experiments with individual biological particles (viruses and whole cells) have been already announced [1-6]. Although being of fairly low resolution and limited to a few two-dimensional frames, they give promise to obtain three-dimensional diffraction data sets in the near future. Note that we are talking about three-dimensional data sets recorded for individual particles and not one-dimensional scattering curves obtained using the method of small-angle X-ray scattering [7] when the recorded data are the result of the averaging over a large number of particles oriented in different ways.

One more problem complicating the calculation of the electron density distribution is that the summation in the formula (5) is over an infinite series, while the experiment allows one to measure only a finite set of structure factor magnitudes. The number of magnitudes obtained experimentally depends primarily on the quality of the crystal. It is usually characterized by the resolution $d_{\min}$ calculated as $d_{\min} = 1/s_{\max}$, where $s_{\max}$ is the maximal length of scattering vectors for the terms of series (5) included into the summation. Geometrically, the value $d_{\min}$ is equal to the minimum length of the period for the Fourier harmonics $\exp[2\pi i(\mathbf{s},\mathbf{r})]$ included in the calculation of (5), and defines the minimal size of visually distinct details on the corresponding contour maps. The partial sums of the Fourier series (5) are called in crystallography the Fourier synthesis of electron density. A theoretical limit of the resolution of this sum is restricted by $\lambda/2$, where λ is the wavelength used in the experiment. For wavelengths close to 1.0 Å, a value commonly used for current experiments in synchrotrons, this limit approaches 0.5 Å. However, for the majority of structures deposited in the bank of protein structures PDB [8], the experiments have allowed collecting a set of data only with a resolution of about 2 Å. This is explained by the difficulties in obtaining high quality crystals of biological macromolecules.

The concept of the resolution is used to characterize not only the whole set of structural factors involved in the calculation of the Fourier synthesis but also individual structure factors; in the latter case, it is defined as $d = 1/s$. Note that the length of the scattering vector $\mathbf{s}$ is related to the scattering angle as $s = 2\sin\theta/\lambda$. Therefore, high-resolution structure factors (low d values, high s values) correspond to the scattering at large angles. Conversely, low-resolution data (high d values, low s values) are referred to as small-angle scattering. When working with large biological objects, it is common practice to raise gradually the resolution of the data set used in the calculation. In this paper we restrict ourselves to the discussion of the first phase of the study, i.e. imaging a macromolecular object at a resolution of about 25 Å, which allows us to determine the shape of the macromolecular complex and that of the domains or molecules it consists of.

1.3. Single particle scattering

Theoretically, an X-ray diffraction experiment with an isolated particle allows measuring the moduli of the Fourier transform of the electron density distribution in the individual particle

$$F_{sp}(\mathbf{s}) = \int_{\mathbf{R}^3} \rho_{sp}(\mathbf{r}) \exp[2\pi i(\mathbf{s},\mathbf{r})] dV_{\mathbf{r}} \qquad (7)$$

for any value of the scattering vector $\mathbf{s}$ within the resolution limit $1/s \le \lambda/2$, and not only for a discrete set $\Re'$ of Bragg vectors, as in the case of a crystalline sample. Thus, the experiment with isolated particles can give a significantly larger amount of information than the

Mathematical Biology and Bioinformatics. 2015. V. 10. № Suppl. doi: 10.17537/2015.10.t56

experiment with crystal samples and thereby greatly facilitate the solution of the phase problem.

The finite dimensions of the particle allow one to reduce the problem of the retrieval of the electron density distribution $\rho_{sp}(\mathbf{r})$ to a standard crystallographic problem [9]. Let us consider an imaginary crystal with a sufficiently large unit cell containing the particle under study. Let's denote the basis of this cell by $\{\mathbf{a},\mathbf{b},\mathbf{c}\}$ and the conjugate basis (the basis of the reciprocal space) by $\{\mathbf{a}^*,\mathbf{b}^*,\mathbf{c}^*\}$ and introduce the periodic function $\rho_{cryst}(\mathbf{r})$ coinciding with $\rho_{sp}(\mathbf{r})$ inside the unit cell and extended to the whole space with the periods $\{\mathbf{a},\mathbf{b},\mathbf{c}\}$. The electron density distribution $\rho_{cryst}(\mathbf{r})$ is completely determined by a set of the structure factors with the scattering vectors $\mathbf{s}$ belonging to the reciprocal space lattice $\Re'_{\mathbf{abc}}$ built on the conjugate basis vectors $\{\mathbf{a}^*,\mathbf{b}^*,\mathbf{c}^*\}$. The values of the corresponding structure factors

$$\mathbf{F}_{cryst}(\mathbf{s}) = \int\limits_{\mathbf{V_{abc}}} \rho_{cryst}(\mathbf{r})\exp[2\pi i(\mathbf{s},\mathbf{r})]dV_{\mathbf{r}} = \int\limits_{\mathbf{R}^3} \rho_{sp}(\mathbf{r})\exp[2\pi i(\mathbf{s},\mathbf{r})]dV_{\mathbf{r}} = \mathbf{F}_{sp}(\mathbf{s}) \qquad (8)$$

coincide with the values of the Fourier transform (7) of the single particle distribution taken for the same vectors $\mathbf{s}$. Therefore, reducing the set of available Fourier transform magnitudes to a discrete grid $\Re'_{\mathbf{abc}}$ $\{F_{sp}(\mathbf{s}); \mathbf{s}\in\Re'_{\mathbf{abc}}\}$, we come to the problem of reconstructing the imaginary crystal density distribution $\rho_{cryst}(\mathbf{r})$ from the magnitudes of its structure factors. Formally speaking, we are faced with the same problem as before: the experiment gives only a half of the information necessary to calculate the density distribution $\rho_{cryst}(\mathbf{r})$ by formula (5). However, a significant difference is that, now, choosing a sufficiently large cell, we have the task of retrieval of the function that is equal to zero for the most part of the unit cell. (Although it remains unknown, for which particular points in the cell this function is equal to zero.) In crystallographic terms, this can be formulated as the presence of a large amount of solvent in the unit cell. These results in the redundancy of the experimental data relative to the number of values to be identified [10]. This property is widely used in crystallographic practice for the refinement of phase values. Clearly, the choice of the imaginary unit cell is rather arbitrary, and changing the unit cell parameters, we change the amount of experimental data involved in the work within the resolution zone used.

The redundancy of experimental data due to the presence of a large "void volume" in the unit cell has been used for decades to solve the phase problem in biological crystallography and optics; the basic methods are iterative procedures such as solvent flattening, density modification, and hybrid input-output algorithm [11-13]. Recently, we have proposed an alternative approach based on procedures of random search strengthened by restrictions of connectivity and binarity of the particle region [14-15].

Theoretically, increasing the size of the unit cell of an imaginary crystal provides an unlimited increase in the experimental information used to find the phases. However, in practice we face with certain restrictions. First, the minimal size of the reciprocal lattice grid is limited by the technical characteristics of the detector (e.g., the size of the detector pixel). Second, an increase in the number of structure factors involved in the work results in a significant increase in the complexity of calculations and places high requirements on the computing resources. Finally, the potential redundancy of data can indeed be helpful up to some resolution limit while its use cannot be extended to a higher resolution. The aim of this work was to study the dependence of the quality of the solution of the phase problem on the size of the unit cell of an auxiliary imaginary crystal.

2. THE USE OF CONNECTED MASKS IN SOLUTION OF THE PHASE PROBLEM IN MACROMOLECULAR CRYSTALLOGRAPHY

Our method of solving the phase problem is described in [14-15]. Briefly, it can be summarized as follows. A grid is introduced into the unit cell of an imaginary crystal, and the number of grid points inside the particle is estimated from the particle volume. At the first stage of the work, a large number of finite connected sets of grid points (masks) of a preset size are generated randomly. The connectivity of a set of grid points is determined on the basis of some neighbouring rules. In our tests, for every grid point six adjacent grid points were considered as its neighbours. To every generated mask, a binary characteristic function of this set is defined and its structure factors magnitudes and phases are calculated. When these calculated magnitudes are close to the experimental ones, the set of calculated phases is considered as 'admissible' and stored for further use. The generation continues until the assigned number of admissible phase sets accumulates (100 in our tests). At the next step, the phase sets stored are aligned by applying the shift of the origin and/or changing the enantiomer to make the corresponding Fourier syntheses as close as possible. The aligned phases and the experimental magnitudes are averaged, and the phase set obtained is considered to be the result of the first stage of the solution of the phase problem [14].

The non-centred correlation coefficient:

$$CM\left[d_{\max}, d_{\min}\right] = \frac{\sum_{\mathbf{s}\in\mathbf{S}} F_{calc}(\mathbf{s}) F_{exact}(\mathbf{s})}{\sqrt{\sum_{\mathbf{s}\in\mathbf{S}} F_{calc}^2(\mathbf{s}) \sum_{\mathbf{s}\in\mathbf{S}} F_{exact}^2(\mathbf{s})}}. \qquad (9)$$

was used as the criterion of closeness of structure factor magnitudes calculated from the mask to the experimental ones. The summation in (9) is performed over scattering vectors $\mathbf{s} \in \mathbf{S} = \left\{\mathbf{s} : 1/d_{\max} < s \le 1/d_{\min}\right\}$.

At the first stage, each mask is constructed stepwise, starting with a randomly selected point. At each step, a new point is added to the mask; it is selected with the equal probability from the boundary points of the already constructed part of the mask.

The second stage (iterative phase refinement) differs from the fist one by two features [15]. First, as before, the mask is build starting from a random point, but the choice of the point to be added to the mask is performed from the boundary points in accordance with some prior probability distribution $P_{prior}(\mathbf{r})$. In our tests, the prior distribution was built on the basis of Fourier synthesis $\rho(\mathbf{r})$ calculated using the experimental structure factor magnitudes and phases found in the previous iteration cycle. We defined this distribution in the exponential form:

$$P_{prior}(\mathbf{r}) = C \exp\left[\frac{\ln t}{\rho_{\max} - \rho_{\min}} \rho(\mathbf{r})\right], \qquad (10)$$

where C is the normalizing factor (to have the sum of probabilities equal to 1), $\rho_{\max}$ and $\rho_{\min}$ are the maximal and minimal values in synthesis $\rho(\mathbf{r})$, and t is a parameter of the method named 'the contrast'. It is easy to see that t is equal to the ratio of the maximal and minimal probabilities in (10).

Second, several mask selection criteria were used together at the stage of the phase refinement. A feature of macromolecular biological objects is a sharp decrease in the structure factor magnitudes with an increase in the scattering angle (or, what is the same, with an increase in the scattering vector length s) (Fig. 1). Therefore, the value of the correlation coefficient (9) can be strongly influenced by a few strong low resolution reflections weakly depending on high resolution reflections. To overcome this difficulty preventing accurate

phase determination at higher resolutions, the selection of the masks was performed using simultaneously several criteria (9) calculated in resolution zones 60-25, 40-25, 30-25 Å, which led to the successive exclusion of strong reflections of the low resolution from the calculation of control criterion (9).

3. TEST OBJECT

To study the potential of this approach, the known trimeric structure of the cyanobacterial photosystem I [16] (PDB code 1JBO [8]) was used as a test object. This trimer contains 36 protein chains and 381 co-factors, which amounts to about 72 thousands of non-hydrogen atoms. The trimer has the molecular weight of 1068 KDa and external dimensions approximately equal to 200 × 200 × 100 Å. Figure 2 shows the overall structure of the trimer.

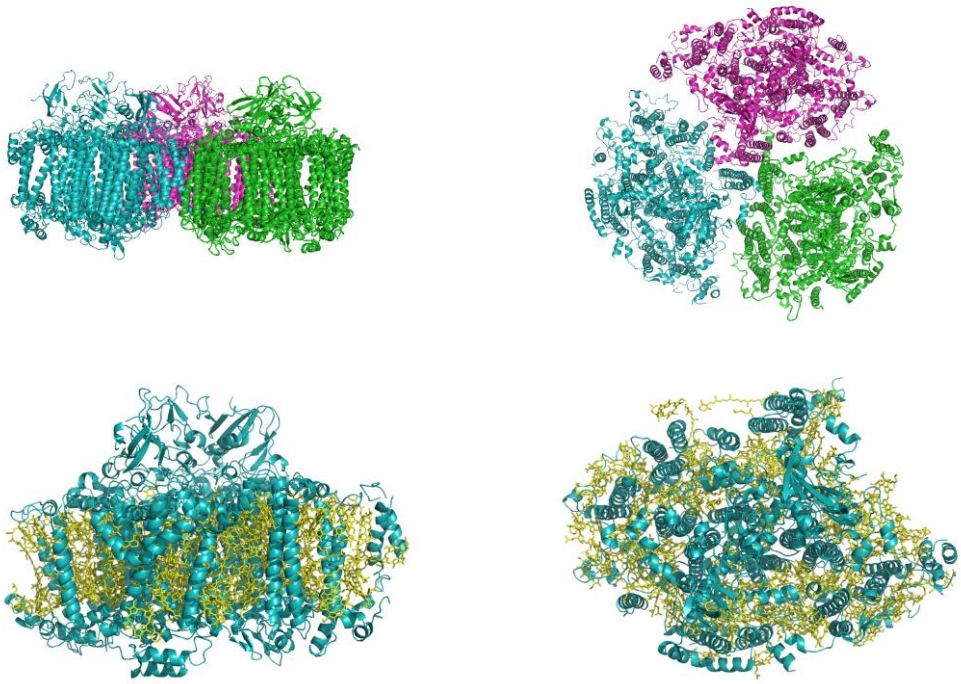

Fig. 2. The protein moiety of the trimer of PS-I (top row; different colours correspond to different monomers) and the structure of the monomer (bottom row; proteins and co-factors).

The goal of our study was to investigate the influence of imaginary crystal cell dimensions on the possibility of solving the phase problem and the accuracy of the phases found. As already mentioned, if the intensities of the waves scattered by an isolated particle are available for a continuous region of scattering vectors, then the researcher has a certain freedom to define an imaginary crystal cell. However, the selection of too large cell parameters may lead to an unacceptable growth in the complexity of the calculations. Therefore, one has to find a compromise between the complexity of calculations and the quality of the results. Below, we discuss the results of *ab initio* determination of structure factor phase values for three variants of the unit cell dimensions (Table 1). The tests were conducted for a cubic cell, assuming that we know nothing about the shape of the object. Compared to the Nyquist limit (defined as the inverse of the double size of the object, i.e. $1/400$ Å^{-1}) for diffraction intensities sampling in the reciprocal space, the sampling step was greater than this limit for cell *c*1, equal to it for cell *c*3, and 1.5 less for cell *c*4. Table 1 summarizes the parameters of cells and the number of structure factors in different resolution shells in.

For each cell, the structure factor magnitudes and phases were calculated from the atomic coordinates for the resolution zone ∞-25 Å. The calculated values of magnitudes were further

considered as the experimentally obtained values. The phase values calculated from the model were not involved in solving the phase problem and were used only to check the results. To estimate the closeness of different phase sets, the map correlation coefficient was calculated for different resolution shells

$$CP\left[d_{\max}, d_{\min}\right]\left(\{\varphi(\mathbf{s})\}\right) = \frac{\sum_{\mathbf{s}\in\mathbf{S}} F_{exact}^2(\mathbf{s})\cos\left(\varphi_{exact}(\mathbf{s}) - \tilde{\varphi}(\mathbf{s})\right)}{\sum_{\mathbf{s}\in\mathbf{S}} F_{exact}^2(\mathbf{s})}. \quad (11)$$

The summation here is performed over structure factors with $\mathbf{s}\in\mathbf{S} = \{\mathbf{s} : 1/d_{\max} < s \le 1/d_{\min}\}$, $F_{exact}(\mathbf{s}), \varphi_{exact}(\mathbf{s})$ are the exact values of the structure factor magnitudes and the phases (those calculated from atomic coordinates), and $\{\tilde{\varphi}(\mathbf{s})\}$ is the phase set obtained by the best alignment of the set $\{\varphi(\mathbf{s})\}$ with respect to $\{\varphi_{exact}(\mathbf{s})\}$ phases [14-15].

Table 1. Dimensions of the unit cells used and the number of reflections for different resolution shells

	Unit cell		
	*c*1	*c*3	*c*4
Cell parameters [Å]	250 × 250 × 250	400 × 400 × 400	600 × 600 × 600
"Solvent content" (%)[1]	91.59	97.95	99.39
Grid for mask generating	30 × 30 × 30	48 × 48 × 48	72 × 72 × 72
No. reflections in resolution shells ∞-25 Å	2071	8538	28877
∞-30 Å	1234	4921	16700
∞-40 Å	510	2082	7061
∞-60 Å	152	618	2084
60-40 Å	358	1464	4977
40-30 Å	724	2839	9639
30-25 Å	837	3617	12177
redundancy[2]	0.23	0.96	3.25

[1]The solvent content was defined as $(1.0\text{-}1.23\cdot Mw/V_{cell})\cdot 100$ %, where Mw is the molecular mass of the trimer, and V_{cell} is the unit cell volume.

[2]The redundancy was defined as $N_{ref}/(3\cdot Np)$, where N_{ref} is the number of reflections in the resolution zone ∞-25 Å, Np is the number of grid nodes in the region with the specific volume equal to 1.6 Å^3/Da.

The test object is a trimer and has a three-fold symmetry axis. In our tests, this symmetry was not taken into account. Of course, taking this symmetry into consideration could facilitate the solution of the phase problem and lead to more accurate phase values (see e.g. [17]). However, since the purpose of the tests was to check the efficiency of the technique based on the use of connected masks, we did not use additional methods of phase refinement to avoid their influence on the result. On the contrary, a spontaneous manifestation of the three-fold symmetry in the final electron density maps is an additional proof of the correctness of the solution of the phase problem.

In the mask generation procedure, the grid step in the unit cell was taken equal to $d_{\min}/3$, and it was equal to $d_{\min}/4$ in the phase alignment procedure, i.e. 8.3 and 6.25 Å, correspondingly. The contrast value in distribution (10) was taken as 10^6. At each step of phasing, the construction of random masks was performed until 100 admissible masks were selected. The alignment of phase sets was performed in the resolution zone ∞-25 Å.

4. RESULTS

An essential parameter of our method is the mask size or, equivalently, the expected volume of the particle. This characteristic can be expressed in different units. In this paper, we estimate the mask size either by the number of mask points or by the specific volume, defined as the ratio of the volume to the molecular mass of the particle. In crystallography, when calculating the solvent content, it is a common practice to take the specific volume of the molecular region equal to 1.23 Å^3/Da [18–19]. However, this estimate is not optimal at low resolution as the borders of the molecule images are significantly smoothed at low resolution Fourier synthesis. To find the optimal estimate, different values of the size of the mask were tested.

Another significant parameter is the cut-off value of the correlation coefficient (9), which determines the selection of masks. In our tests, this cut-off value was set either directly or in a relative scale (as z-score)

$$z_{crit} = \frac{CM_{crit} - \langle CM \rangle}{\sigma_{CM}}. \qquad (12)$$

Here $\langle CM \rangle$ and σ_{CM} are the mean value and the standard deviation of CM, respectively, calculated over a large number of generated variants. Table 2 shows the results of the initial phasing for different combinations of the mask size and the selection cut-off.

Table 2. The quality of the averaging of the structure factor phases corresponding to admissible masks selected from randomly generated masks on the basis of magnitude correlation (9). The values of phase correlation (11) are shown for resolution shells: ∞–60, ∞–40, ∞–30, ∞–25 (top row) and 60–40, 40–30, 30–25 Å (bottom row). The cut-off z_{crit} was defined for the resolution shell ∞–25 Å. The unit cell was *c*1

*CP**100		Mask volume: specific [Å^3/Da]/No. of points					
		1.0 1845	1.2 2214	1.4 2583	1.6 2952	1.8 3321	2.0 3691
z_{crit}	No	66/61/59/57 18/24/04	66/60/57/56 –01/28/18	69/64/59/57 15/04/00	73/67/64/62 16/26/–02	75/69/63/62 17/–10/00	77/72/69/67 28/21/14
	1.5	72/68/64/62 21/15/13	73/66/62/61 –02/09/11	82/74/68/67 –06/–05/12	81/75/72/70 22/22/11	81/75/71/68 22/13/–09	81/75/70/68 21/05/05
	2.0	75/68/65/63 11/12/14	81/73/68/66 03/–02/–04	83/75/70/68 –02/05/10	79/72/68/66 06/16/11	81/75/71/69 18/18/09	80/75/71/69 22/22/04
	2.5	76/69/66/64 08/21/14	83/77/73/71 20/18/01	85/76/71/69 –10/05/09	87/82/75/73 33/–10/–09	83/77/73/71 24/17/04	83/75/71/69 –01/19/10
	3.0	79/73/70/68 13/25/13	88/81/76/74 16/08/04	83/76/70/68 16/–05/05	88/83/78/76 30/27/08	85/79/75/73 16/25/02	82/76/71/69 19/08/01

The first row of the table corresponds to the test with no mask selection. This means that the phase sets corresponding to the first 100 randomly generated masks were aligned and averaged to produce the resulting phases. As was shown previously [14–15], even in this case, the resulting phase values, are closer to the true ones than the phase sets generated randomly. Testing different mask sizes showed that the optimal size of the mask is in the range of the specific volumes 1.4–1.8 Å^3/Da, which is consistent with the results obtained earlier for other objects [14–15]. At the second step of phasing (phase refinement), the specific volume was fixed as 1.6 Å^3/Da.

Table 3. The results of phase refinement. The values of phase correlation (11) are shown for resolution shells: ∞–60, ∞–40, ∞–30, ∞–25 (top row) and 60–40, 40–30, 30–25 Å (bottom row). The cut-off z_{crit} was defined for resolution shell ∞–25 Å. The unit cell was $c1$

*CP**100 /No. of masks build	Start phase set to build distribution (10)				
	$r1$	$r2$	$r3$	$r4$	$r5$
Start	87/82/75/73 33/–10/–09 14087	83/77/71/69 16/01/–14 18469	84/78/73/71 16/08/02 22039	87/80/75/73 17/11/–03 21979	83/76/71/69 07/10/05 19074
Step 1	95/90/83/80 46/–12/–13 513	89/83/78/75 20/18/–05 633	90/85/78/76 32/00/–03 555	92/87/82/80 36/28/15 625	90/83/76/74 12/–03/02 544
Step 2	94/90/84/82 47/12/01 105	91/84/80/77 24/22/02 103	91/86/80/77 35/08/–05 100	94/89/85/83 44/32/15 111	92/85/78/76 16/00/05 104
Step 3	96/92/85/82 50/–05/–01 114	91/85/81/78 26/24/07 121	91/86/81/78 37/14/–04 110	94/90/86/84 47/36/17 105	92/85/79/77 20/02/05 162
Step 4	95/90/85/83 49/18/06 181	91/85/81/79 27/26/08 116	92/87/82/79 39/19/–02 135	96/92/87/85 48/24/09 111	93/86/80/78 22/04/05 245
Step 5	95/91/85/83 50/21/09 159	92/86/81/79 28/26/09 209	92/87/82/80 40/22/01 137	95/91/87/85 51/38/18 224	93/87/81/78 24/06/05 678
Step 6	95/91/86/84 50/23/11 332	92/86/81/79 28/26/10 301	92/87/83/80 41/24/04 226	97/93/88/86 53/28/12 119	93/87/81/79 26/07/06 1434
Step 7	95/91/86/84 51/24/12 608	92/86/81/79 29/25/11 333	92/88/83/81 41/25/06 289	97/93/88/86 54/29/12 317	94/87/82/79 27/09/06 1242
Step 8	95/91/86/84 52/25/14 1241	92/86/82/79 29/24/11 582	93/88/83/81 43/25/08 362	96/92/88/86 55/41/20 276	94/87/82/80 28/11/06 1104
Step 9	95/91/86/84 53/25/15 2155	92/86/82/79 29/23/11 819	93/88/83/81 44/24/09 570	97/93/89/87 56/33/15 664	94/88/82/80 28/13/06 1112
Step 10	95/91/86/84 56/25/14 1843	92/86/82/79 29/21/11 1269	93/88/83/81 44/24/10 774	96/93/89/87 57/44/22 1220	94/88/82/80 29/14/06 2224

The results of the phasing depend on the random number generator and may differ for different initializations of this generator. To study the process of phase refinement more thoroughly, five preliminary phase sets $r1-r5$ were obtained independently. In the five runs, different seeds were used to initialize the random number generator, but the mask size and cut-off correlation values were the same (1.6 Å^3/Da and $z_{crit}=2.5$ for the criterion $CM[\infty-25]$, correspondingly). The first row in Table 3 shows the quality of these five initial phase sets. Then, 10 refinement cycles were applied to each of the five solutions as described in Section 2.

For phase refinement, the phase values obtained in the previous cycle were used to build the Fourier synthesis and prior distribution (10); the contrast value was fixed as $t=10^6$ for all refinement cycles. Four criteria, namely $CM[\infty,25]$, $CM[60,25]$, $CM[40,25]$, $CM[30,25]$ were used to select the masks. The criteria choice corresponded to a gradual elimination of strong low resolution reflections. The cut-off values of the criteria for each cycle were taken as the mean values of corresponding criteria for the masks selected in the previous refinement cycle. The exception was the first cycle where the cut-off values were determined as the average values of the criteria for the masks generated in a trial run. The results of the phase

refinement are summarized in Table 3. Figure 3 shows changes in the cut-off values of the selection criteria during refinement.

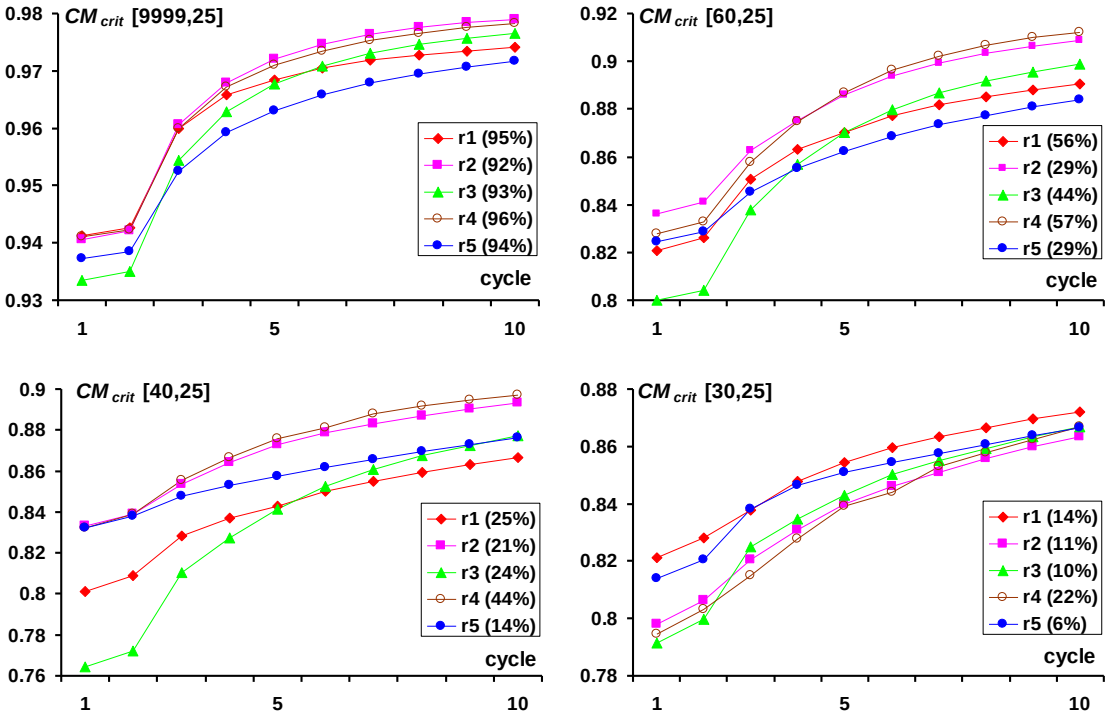

Fig. 3. Change in the cut-off value during phase refinement. Figures in the boxes correspond to the achieved phase correlation (in accordance with Table 3). The unit cell was *c*1. Five independent runs were done.

The tests with cells *c*3 and *c*4 were done similarly to tests with cell *c*1. The results of the tests are summarized in Tables 4–6 and Figures 4 and 5.

Table 4. The quality of the phase sets obtained by averaging the structure factor phases corresponding to admissible masks selected from randomly generated masks on the basis of magnitude correlation (9). The values of phase correlation (11) are shown for the resolution shells: ∞–60, ∞–40, ∞–30, ∞–25 (top row) and 60–40, 40–30, 30–25 Å (bottom row). The cut-off z_{crit} was defined for resolution shell ∞–25 Å. The unit cell was *c*3

*CP**100		Mask volume: specific [Å^3/Da]/No. of points					
		1.0 1846	1.2 2215	1.4 2584	1.6 2954	1.8 3322	2.0 3691
z_{crit}	No	72/67/65/63 21/27/06	76/69/64/62 01/–15/04	77/70/67/65 02/25/04	79/73/69/67 03/13/04	72/68/65/64 25/22/12	75/69/65/64 07/10/04
	1.5	80/73/70/68 00/19/–05	82/76/70/68 17/–07/–11	83/77/74/71 19/22/–05	86/81/76/74 26/07/–07	87/82/78/75 32/19/–15	87/82/76/74 31/–05/–09
	2.0	81/75/71/69 –01/23/11	86/80/75/73 14/10/–03	86/81/77/75 24/23/13	86/81/77/75 27/19/–04	88/84/79/76 32/19/–13	87/82/78/76 33/22/–13
	2.5	80/75/71/69 17/28/–01	86/81/77/74 22/22/–06	87/82/77/75 24/15/–01	88/84/78/76 33/05/–14	88/83/78/76 29/18/–12	90/86/81/78 41/08/–12
	3.0	86/80/76/74 10/21/03	89/82/78/76 16/15/01	89/84/79/76 30/13/–05	90/85/80/77 34/11/–12	90/86/81/79 39/21/–05	90/86/81/78 36/21/–13

For phasing with cell *c*3, the parameters of the method (mask size and cut-off level) were tested first (Table 4). The results obtained are close to those observed in the test with the cell

*c*1: the optimal size of the mask varies in the range 1.6–1.8 Å^3/Da, and the accuracy of the preliminarily found phase sets increases with increasing cut-off value. At the same time, in general, the accuracy of the phase problem solutions was higher than it was for the same parameters in the tests with the cell *c*1. As before, five independent preliminary solutions $r1-r5$ were generated using the same parameter values 1.6 Å^3/Da and $z_{crit}=2.5$, but with different seeds of the random number generator. The same parameters were used in phasing with the cell *c*4, where testing the preliminary phasing was omitted. Despite the fact that the best results in Tables 3 and 4 correspond to the cut-off value $z_{crit}=3.0$, the smaller value $z_{crit}=2.5$ was used to produce preliminary solutions $r1-r5$. This was done for two reasons. First, using the cut-off $z_{crit}=3.0$ leads to an essential growth of computational cost, especially for the cell *c*4. Second, with real objects, the strength of the selection should be consistent with the precision of experimental data and should not be excessively high.

Table 5. The results of phase refinement. The values of phase correlation (11) are shown for the resolution shells: ∞–60, ∞–40, ∞–30, ∞–25 (top row) and 60–40, 40–30, 30–25 Å (bottom row). The cut-off z_{crit} was defined for resolution shell ∞–25 Å. The unit cell was *c*3

*CP**100/No. of masks build	Start phase set to build distribution (10)				
	*r*1	*r*2	*r*3	*r*4	*r*5
Start	87/81/77/75/ 23/23/07 36490	88/82/76/73/ 17/–12/–08 37336	87/81/74/73/ 18/–14/12 39327	87/81/76/74/ 22/02/05 35142	87/83/79/77/ 40/28/10 39069
Step 1	94/90/86/84/ 42/35/19 559	94/88/82/79/ 24/–00/–12 508	93/87/81/79/ 27/–09/19 529	92/88/83/81/ 35/12/16 537	94/91/87/85/ 55/42/20 573
Step 5	96/94/91/89/ 63/55/29 562	95/91/87/84/ 46/21/–03 154	95/91/85/82/ 45/05/–00 100	95/91/87/85/ 50/34/26 1250	96/93/91/89/ 65/56/28 163
Step 10	97/94/92/91/ 68/62/34 799	96/92/88/86/ 53/33/02 1000	95/91/86/83/ 52/13/00 1032	95/91/87/85/ 50/35/21 1292	96/94/91/90/ 69/59/36 871

Table 6. The results of phase refinement. The values of phase correlation (11) are shown for the resolution shell: ∞–60, ∞–40, ∞–30, ∞–25 (top row) and 60–40, 40–30, 30–25 Å (bottom row). The cut-off z_{crit} was defined for resolution shell ∞–25 Å. The unit cell was *c*4

*CP**100/No. of masks build	Start phase set to build distribution (10)				
	*r*1	*r*2	*r*3	*r*4	*r*5
Start	90/85/81/79 25/28/09 34190	89/83/78/76 18/–02/09 38570	91/85/79/77 27/–16/09 35081	89/83/78/76 18/15/07 35394	91/86/82/80 24/25/20 32130
Step 1	96/92/88/86 46/37/06 505	95/90/84/82 40/02/07 561	94/90/85/83 47/13/20 438	94/89/85/83 37/16/21 524	97/93/89/87 47/34/29 517
Step 5	98/95/92/90 66/51/22 334	96/92/88/85 54/21/02 133	95/92/89/86 62/36/20 101	96/93/89/87 57/37/26 1425	98/96/94/93 73/57/51 141
Step 10	98/96/94/92 73/57/38 169	96/93/88/86 58/27/07 557	95/92/89/87 62/36/17 405	94/90/88/86 49/55/32 42780	99/97/95/94 79/65/56 420

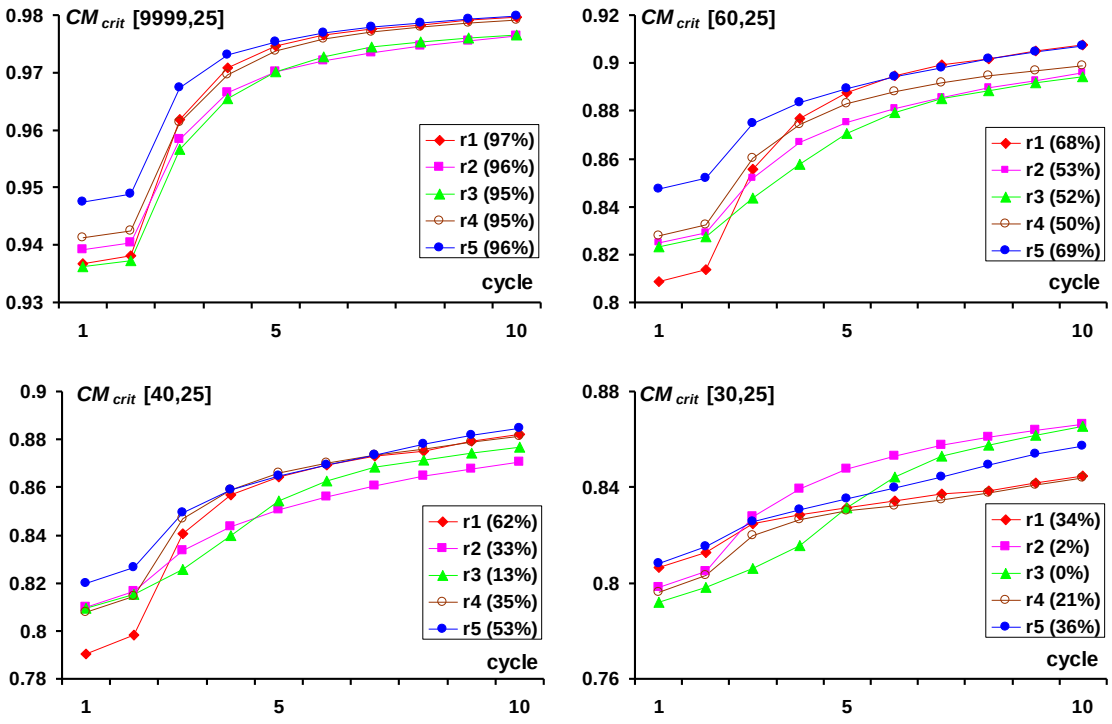

Fig. 4. Change in the cut-off value in the course of phase refinement. Figures in the boxes correspond to the achieved phase correlation (in accordance with Table 5). The unit cell was *c*3; five independent runs.

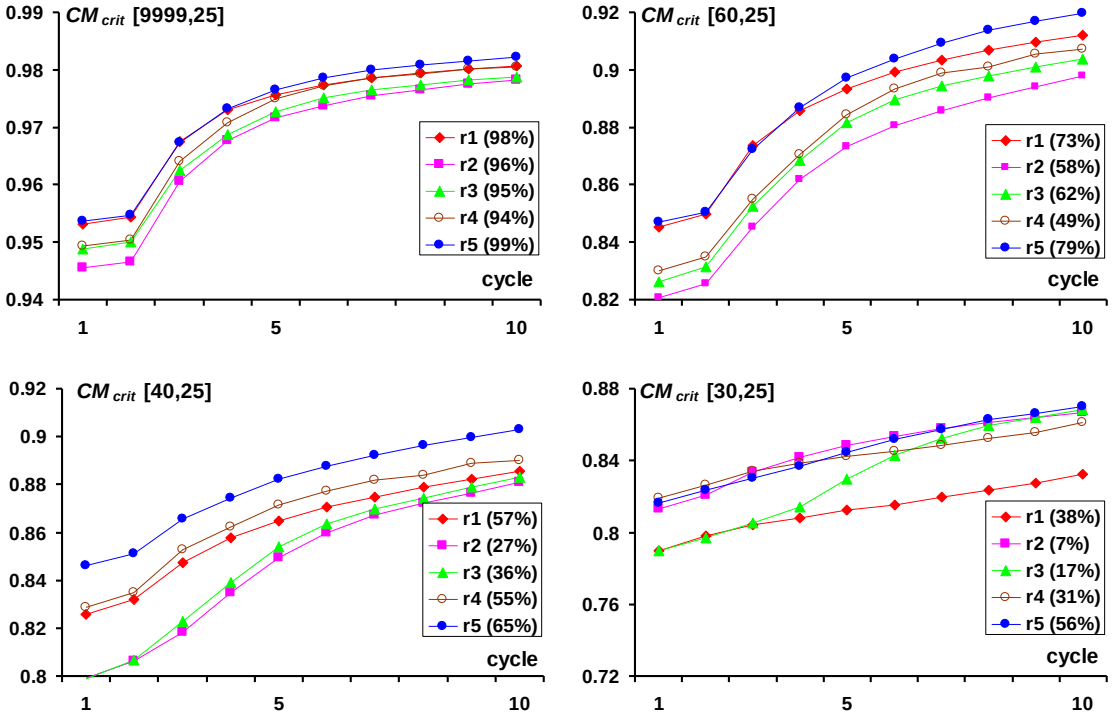

Fig. 5. Change in the cut-off value in the course of phase refinement. Figures in the boxes correspond to the achieved phase correlation (in accordance with Table 6). The unit cell was *c*4; five independent runs.

DISCUSSION AND CONCLUSIONS

Figure 6 shows the accuracy of phase problem solutions obtained at different stages and for different choice of an auxiliary imaginary crystal cell. It follows from the sections above and from Tables 3, 5 and 6 that increasing the size of an imaginary cell improves the results. Furthermore, this improvement goes beyond the "Nyquist limit" (defined as the inverse to the double size of the particle). Figure 6 and Table 3 also show that, even with the cell size below the Nyquist limit, the procedure provides reasonable information about the phase values, although at a lower resolution than for larger unit cells. The possibility to obtain phase information, even with the limited cell size, is due to the fact that the introduction of additional information such as, binarity, connectivity, finite dimensions of the particle, etc. into the process of phasing allows one to escape false solutions.

Figure 7 shows phase correlation (11) in different shells in the reciprocal space for the best solutions obtained with different cells. Along with Tables 5 and 6, this confirms that increasing imaginary cell parameters and involving additional experimental information improve the accuracy of the phases. The plot corresponding to the best solution (*r*5 for cell *c*4) repeats the shape of the plot for the averaged intensities (Fig. 1). This reflects the general trend of the known *ab initio* phasing methods: more reliable determination of phase values for stronger reflexes. Figure 7 (right) shows the value of the phase correlation for the expanding resolution shells $\infty - d_{\min}$ in the reciprocal space. It confirms the same trend for improving the accuracy of solutions with increasing the cell size.

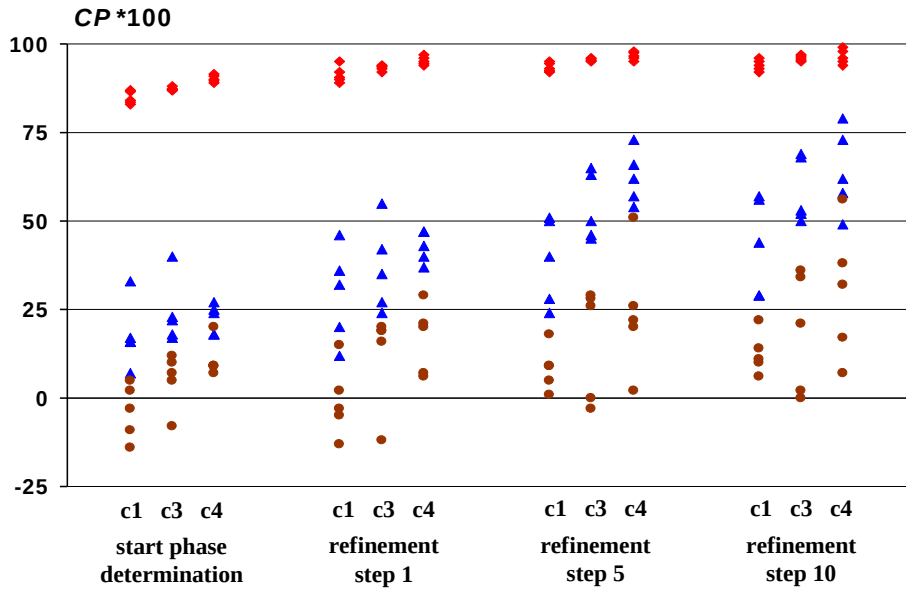

Fig. 6. Phase correlations (11) of current and exact phases for different stages of phasing for different cells. Correlation coefficients *CP*[∞–60] are shown by red diamonds, *CP*[60–40] by blue triangles, *CP*[30–25] by brown circles. For a particular stage and cell 5, similar markers correspond to five independent solutions (*r*1–*r*5).

Figures 8 and 9 show an image of the PS-I trimer, corresponding to the best solution (*r*5 for cell *c*4), in comparison with the image corresponding to the exact synthesis at 25 Å resolution.

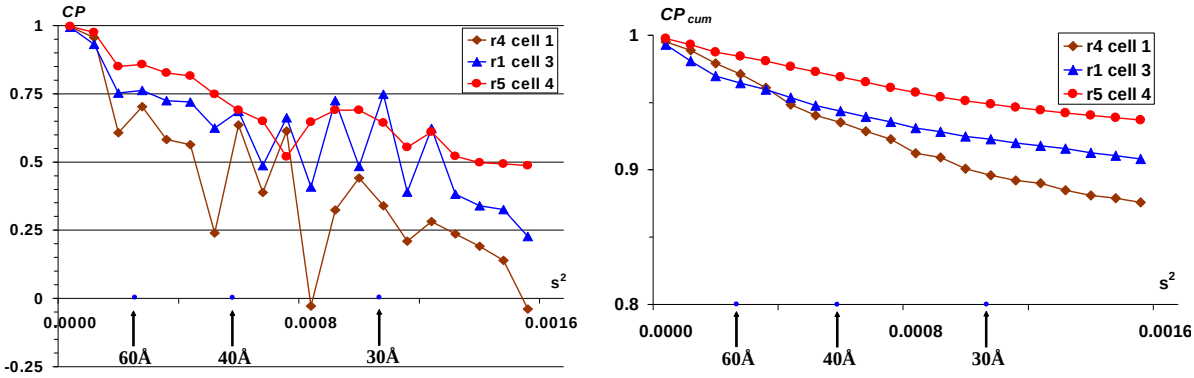

Fig. 7. Phase correlations with the exact phases for the best solutions for three different unit cells. Left: phase correlation coefficients (11) calculated for thin spherical shells in the reciprocal space *vs*. squared shell radius $s^2 = (2\sin\theta/\lambda)^2$. Right: phase correlation coefficients (11) calculated for extending spheres in the reciprocal space *vs*. squared sphere radius.

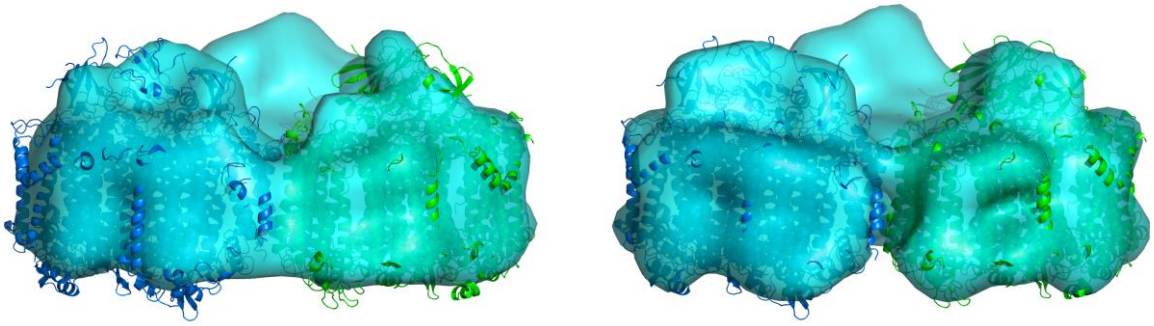

Fig. 8. Images of trimer PS-I as shown by 25 Å resolution Fourier synthesis maps. The regions corresponding to the specific volume 1.23 Å^3/Da are shown. Two of three monomers in the atomic model are shown as cartoon. Left: the best *ab initio* phases (*r*5 in cell *c*4) were used together with the observed magnitudes to calculate Fourier synthesis. Right: the exact phase values were used.

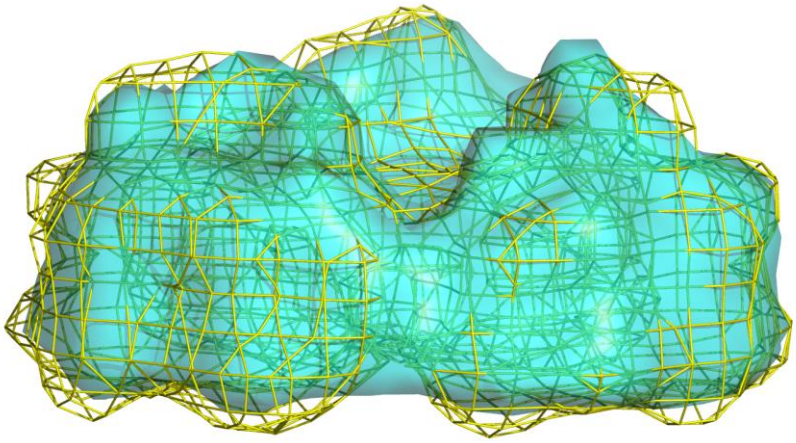

Fig. 9. Images of trimer PS-I. as shown by 25Å resolution Fourier synthesis maps. The regions corresponding to the specific volume 1.23 Å^3/Da are shown. Surface: the best *ab initio* found phases (*r*5 in *c*4 cell) were used together with the observed magnitudes to calculate Fourier synthesis. Mesh: the exact phase values were used.

In the case of a real object, the exact phase values are unknown. Hence, the value of the phase correlation (11) cannot be calculated and used to select the best solution from a series of independently obtained ones ($r1-r5$ in our tests). Some approaches to the selection of the best solution are discussed in [15]. Figure 5 suggests that the cut-off values achieved during refinement can serve as an additional indicator of the quality of the solution. Although this characteristic is not in a one-to-one correspondence with the quality of the solution, the best solution in case of the cell $c4$ could be identified on the basis of the plots shown in Figure 5. The cut-off values attained for solution $r5$ dominate the others in all resolution shells considered. The situation is less clear for cell $c3$ (Fig. 4). Here, the cut-off values of the two best solutions dominate only in the first three resolution shells. At the same time, it should be noted that for the fourth resolution shell the phase correlation coefficients for all five solutions are very low. It can be said here that the effective resolution for the phase sets is limited to 30 Å. A similar situation can be observed with the cell $c1$ (Fig. 3) where the effective resolution of the best solution is limited to about 40 Å. Thus, using the criterion discussed, one can distinguish the best solution if the solution is quite good; however, this criterion cannot properly indicate the hierarchy of weak solutions.

The results of the tests confirm that the suggested method solves the phase problem and demonstrate the advantage of using large imaginary cells. However, it should be noted that the tests were performed at limited computational capacity of laboratory computers. A further increase in the amount of experimental information involved would require a modification of the software including the use of computers with a parallel architecture. It should also be noted that the possibility of reducing the sampling step of the experimental data can be limited by the technical characteristics of the experimental equipment, such as the pixel size of the detector recording the intensities of scattered X-rays.

This work was supported by a grant of RFBR (project No. 13-04-00118).

REFERENCES

1. Miao J., Kirz J., Sayre D. The oversampling phasing method. *Acta Crystallographica Section D: Biological Crystallography*. 2000. V. 56. P. 1312–1315.
2. Thibault P., Elser V., Jacobsen C., Shapiro D., Sayre D. Reconstruction of a yeast cell from X-ray diffraction data. *Acta Crystallographica Section A: Foundations of Crystallography*. 2006. V. 62. P. 248–261.
3. Song C., Jiang H., Mancuso A., Amirbekian B., Peng L., Sun R., Shah S.S., Zhou Z.H., Ishikawa T., Miao J. Quantitative Imaging of Single, Unstained Viruses with Coherent X Rays. *Physical Review Letters*. 2008. V. 101. Article No. 158101. doi: 10.1103/PhysRevLett.101.158101.
4. Maia F.R.N., Ekeberg T., van der Spoel D., Hajdu J. *Journal of Applied Crystallography*. 2010. V. 43. P. 1535–1539.
5. Seibert M.M., Ekeberg T, Maia F.R.N.C., Svenda M., Andreasson J.,Jönsson O., Odic D., Iwan B., Rocker A., Westphal1 D. Single mimivirus particles intercepted and imaged with an X-ray laser. *Nature*. 2011. V. 470. P. 78–82.
6. Van der Schot G., Svenda M., Maia F.R.N.C., Hantke M., DePonte D., Seibert M.M., Aquila A., Schulz J., Kirian R., Liang M. et al. Imaging single cells in a beam of live cyanobacteria with an X-ray laser. *Nature Communication*. 2015. V. 6. Article No. 5704. doi: 10.1038/ncomms6704.
7. Feigin L.A., Svergun D.I. *Structure Analysis by Small-Angle X-ray and Neutron Scattering*. Shpringer. 1987.
8. Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E. The Protein Data Bank. *Nucleic Acids Research*. 2000. V. 28. P. 235–242. URL: http://www.rcsb.org/ (accessed 22.11.2015).

9. Sayre D. Some implications of a theorem due to Shannon. *Acta Crystallographica*. 1952. V. 5. P. 843.
10. Bricogne G. Geometric sources of redundancy in intensity data and their use for phase determination. *Acta Crystallographica Section A: Foundations of Crystallography*. 1974. V. 30. P. 349–405.
11. Bricogne G. Methods and programs for direct-space exploitation of geometric redundancies. *Acta Crystallographica Section A: Foundations of Crystallography*. 1976. V. 32. P. 832–847.
12. Fienup J.R. Reconstruction of an object from the modulus of its Fourier transform. *Optics Letters*. 1978. V. 3. № 1. P. 27–29.
13. Zhang K.Y.J., Cowtan K.D., Main P. Phase improvement by iterative density modification. In: *International Tables for Crystallography*. 2012. V. F. P. 385–400. doi: 10.1107/97809553602060000847.
14. 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*. 2015. V. 10. № Suppl. P. t1–t19. doi: 10.17537/2015.10.t1.
15. 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 Section D: Biological Crystallography*. 2016. V. 72. doi: 10.1107/S2059798315022652. (in press).
16. 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.
17. Rossmann M.G., Arnold E. Noncrystallographic symmetry averaging of electron density for molecular-replacement phase refinement and extension. In: *International Tables for Crystallography*. 2012. V. F. P. 352–363. doi: 10.1107/97809553602060000842.
18. Matthews B.M. Solvent content of protein crystals. *Journal of Molecular Biology*. 1968. V. 33. P. 491–497.
19. Weichenberger C.X., Rupp B. Ten years of probabilistic estimates of biocrystal solvent content: new insights via nonparametric kernel density estimate. *Acta Crystallographica Section D: Biological Crystallography*. 2014. V. 70. P. 1579–1588.

Received December 15, 2015.
Published December 23, 2015.