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

УДК: 577.3

The use of connected masks for reconstructing the single particle image from X-ray diffraction data

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

Institute of Mathematical Problems of Biology, Russian Academy of Sciences, Pushchino, Moscow Region, 142290, Russia

Abstract. The problem of reconstructing the image of a single macromolecular object from X-ray diffraction data can be formulated as a problem of the reconstruction of the 3D electron density distribution from the magnitudes of its Fourier transform. This problem can be reduced to a series of standard X-ray crystallography tasks, namely, the recovery of a periodic function from its Fourier coefficients (structure factors) magnitudes, which are determined in an X-ray diffraction experiment. In this work, a new approach to the solution of these tasks is suggested which is based on the use of connected binary masks as an approximation of the required electron density distribution. The approach includes the random generation of a great number of connected masks, the selection of the masks that are in agreement with an experimental and *a priori* information about the object, and the alignment and the averaging of the phase sets of the structure factors that correspond to the selected masks. The averaged phase values together with the experimentally determined magnitudes are used for the calculation of the Fourier synthesis, which is applied for the visualization of the object under study. The approach can be used in studies of both single particles and crystalline species; however, it holds the greatest promise for investigations of single objects. The results of testing the approach are presented.

Key words: X-ray crystallography, phase problem, XFEL, scattering by a single particle.

INTRODUCTION

X-ray diffraction is the main method of determining the atomic structures of biological macromolecules and their complexes. However, a serious restriction of the method is that a sample for the X-ray experiment has to be prepared as a monocrystal. This requirement is dictated by the fact that the intensities of the waves scattered by a single molecule are many orders of magnitudes less than the intensity of the incident X-ray beam, which strongly complicates their registration. The interference of waves scattered by a set of molecules that are in the nodes of the regular periodic grid leads, for some particular directions, to the multiple amplification of secondary waves, which makes their registration possible. The development of new powerful sources of X-rays, such as X-ray free electron lasers, allows one in the nearest future to talk about X-ray diffraction experiments with single big molecules or complexes of molecules [1, 2]. This possibility makes timely the development of approaches to the determination of the 3D structure of biological macromolecules based on X-ray diffraction data obtained in the experiment with a single molecule. In the present work, one possible approach is suggested.

The main obstacle for the structure determination from X-ray diffraction data is the phase problem. The experiment enables one to determine directly only the magnitude of the complex function, which is the Fourier transform of the function that describes the spatial

distribution of electrons in the object under study (the function of the distribution of the electron density). The solution of the phase problem (calculation of the phase values of the Fourier transform) makes it possible to restore the electron density distribution by calculating the inverse Fourier transform usually named as the Fourier synthesis of the electron density. It should be noted that the solution of the phase problem in the case of a single particle is a theoretically simpler task than in the case of a crystal. During X-ray diffraction through a crystal, one can measure the intensity of scattered waves only for a discrete set of specific directions (Bragg reflections), while, in a successful experiment with a single molecule, one can obtain the information about the intensity of scattering in all directions.

The accuracy of the information obtained by calculating the synthesis of the electron density depends on the number of reflections (in case of a crystal) or the size of the region of scattering angles (in case of a single molecule) included in the calculation. This accuracy is characterized by the resolution value and depends on the limiting values of the scattering angles, for which it was possible to register the intensity of scattered waves. A common practice in crystallography is a step-by-step progress in the determination of the electron density distribution from low to high resolution. It this work, we discuss approaches to the solution of the phase problem at the initial stage of the study, namely, in the zone of low and medium resolution (small angle scattering). The information obtained allows one to determine the general contours of the molecule and is a starting point for the further use of the methods for increasing the resolution. The authors of the present work suggested a set of approaches to the solution of the phase problem in the low resolution zone in the studies of crystal structures [3, 4]. Here, we attempt to extend these approaches [5] into the case of the scattering by a single molecule. As it will be shown below, the task of the recovery of the electron density distribution in case of the scattering by a single molecule can be considered as a series of usual crystallographic tasks relevant to imaginary crystals containing the molecule under study buried in a large volume of a solvent in the unit cell. An increase in the portion of the solvent in the imaginary cell potentially facilitates the solution of the phase problem but leads to a significant increase in the computing time.

The method is based on the use of the compactness and connectivity features of the region of high density values of the electron density in biological macromolecules. Earlier, the Monte Carlo type procedure was suggested [5], which consists of a random generation of phase sets and the calculation of the corresponding Fourier syntheses, the selection of those phase sets that provide the connectivity of the region of high electron density values in Fourier synthesis, the alignment and averaging of selected phase sets. Here, we investigate an alternative approach to the use of connectivity features. In the new procedure, just hypothetically connected regions are randomly generated. For the further work, those regions are selected, for which the magnitude of the Fourier transform reproduces well the values obtained in the experiment. The phase sets of the Fourier transform corresponding to the selected regions are further aligned and averaged, which leads to the required solution of the phase problem. This approach demands much less computing calculations and provides a higher quality of the resulting phase sets compared with the earlier used approach [5].

1. FUNDAMENTALS OF THE METHOD

1.1. Theoretical basis of the X-ray diffraction analysis

A principal scheme of an X-ray diffraction experiment is shown in Fig. 1. An object under study is placed in an X-ray beam, and the intensities of arising secondary waves diverging in all directions are registered. During the experiment, the object is rotated, which makes it possible to obtain at the output a set of two-dimensional roentgenograms corresponding to different orientations of the object relative to the initial beam. In the kinematic theory of diffraction, a primary beam is considered as a plane monochromatic electromagnetic wave, and the interaction of electrons only with this primary wave is considered. By the action of

t2

this incident wave, electrons of the object begin to oscillate and become the sources of new spherical waves. These waves are summed to form scattered waves.

A complex magnitude of a scattered wave $\mathbf{E}$ differs from the magnitude of the primary wave $\mathbf{E}_0$ by two multipliers:

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

The multiplier ε is equal to the portion of the energy flow of the initial wave that comes with the scattered wave to the detector upon the scattering by only one electron. This portion is very small (it can be estimated to be $\varepsilon \sim 10^{-12}$), which just presents the main problem during the registration of the scattered radiation. The complex multiplier $\mathbf{F}(\mathbf{s})$ is called the structure factor and is determined by the electron density distribution in the object that scatters.
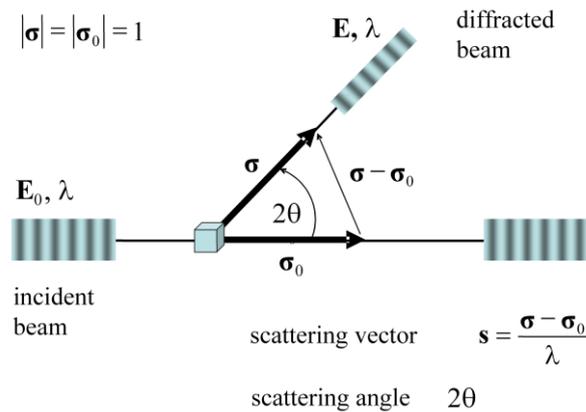

Fig. 1. Scheme of an X-ray diffraction experiment.

The value of the structure factor depends on the direction of the scattering (more exactly, on the vector

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

which is called the vector of scattering) and is related through the Fourier transform to the electron density distribution $\rho(\mathbf{r})$ in the object under study:

$$\mathbf{F}(\mathbf{s}) = \int \rho(\mathbf{r}) \exp\left[2\pi i(\mathbf{s},\mathbf{r})\right] dV_{\mathbf{r}} . \tag{3}$$

The intensity of the scattered wave registered in the experiment is proportional to the square of the structure factor magnitude, i.e., depends on the distribution of electrons in the object under study. This allows one to formulate the problem of determination of the distribution $\rho(\mathbf{r})$ from the set of experimentally determined intensities $\{I(\mathbf{s})\}$.

Theoretically, the electron density distribution can be recovered from structure factor values through the inverse Fourier transform

$$\rho(\mathbf{r}) = \int \mathbf{F}(\mathbf{s}) \exp\left[-2\pi i(\mathbf{s},\mathbf{r})\right] dV_{\mathbf{s}} , \tag{4}$$

however, there are two obstacles. First, an X-ray experiment enables one to obtain the values of only structure factor magnitudes and does not allow one to measure phase values. Second, it follows from formula (2) that the experiment allows one to obtain the values of structure factor magnitudes only for the vectors s that satisfy the restriction $|\mathbf{s}| \le 2/\lambda$. Therefore, even with the known phases of the structure factors, the inverse Fourier transform enables one to

t3

obtain, instead of the exact distribution of electron density $\rho(\mathbf{r})$, only the Fourier synthesis of a finite resolution d

$$\rho(\mathbf{r}) = \int\limits_{|\mathbf{s}| \leq 1/d} \mathbf{F}(\mathbf{s}) \exp\left[-2\pi i(\mathbf{s},\mathbf{r})\right] dV. \qquad (5)$$

In this case, a possible resolution d cannot be less than half of the wavelength $\lambda/2$.

Let the object under study be a crystal; that is, there is a great number of identical similarly oriented molecules placed in the nodes of the crystal lattice $\mathfrak{R}$ built on the basis $\{\mathbf{a},\mathbf{b},\mathbf{c}\}$. A shift of the molecule position in the space by vector $\mathbf{u}$ leads to an appearance of an additional phase multiplier $\exp\left[2\pi i(\mathbf{s},\mathbf{u})\right]$ in its structure factor. Suppose that vector $\mathbf{u}$ runs through all nodes of lattice $\mathfrak{R}$, and additional conditions for vector $\mathbf{s}$ are satisfied:

$$(\mathbf{s},\mathbf{a}) = h, (\mathbf{s},\mathbf{b}) = k, (\mathbf{s},\mathbf{c}) = l, \quad \text{where } h,k,l \text{ are integers.} \qquad (6)$$

In this case, all additional multipliers in the structure factors that correspond to single molecules will be equal to 1; and a multiple amplification of the scattered wave due to the summation of identical contributions from a great number of molecules in the crystal will happen. The intensity of this wave becomes big enough to be registered in the experiment. If conditions (6) are not satisfied, the phases of additional complex multipliers are almost random, and the structure factors corresponding to single molecules give zero as a result of summation. In this case, the intensity of the scattered wave becomes to be too small to be detected experimentally. Conditions (6) are called the diffraction conditions (Laue-Bragg-Wulff), and the corresponding scattered waves are called the reflections. Below, we will talk about the magnitudes and phases of the structure factors corresponding to these or other reflections bearing in mind structure factors involved in equation (1) for these waves.

In the case of a crystal, the formulas for the calculation of structure factors and electron density have the form:

$$\mathbf{F}(\mathbf{s}) = \int\limits_{V} \rho(\mathbf{r}) \exp\left[2\pi i(\mathbf{s},\mathbf{r})\right] dV_{\mathbf{r}}, \qquad (7)$$

$$\rho(\mathbf{r}) = \sum_{\mathbf{s} \in \mathfrak{R}'} \mathbf{F}(\mathbf{s}) \exp\left[-2\pi i(\mathbf{s},\mathbf{r})\right], \qquad (8)$$

where V is a parallelepiped built on vectors $\{\mathbf{a},\mathbf{b},\mathbf{c}\}$ (the unit cell of the crystal), and $|V|$ is the volume of the unit cell. The summation in (8) occurs over all nodes of the integer grid built on the vectors of the basis $\{\mathbf{a}^*,\mathbf{b}^*,\mathbf{c}^*\}$, which is conjugated to the basis $\{\mathbf{a},\mathbf{b},\mathbf{c}\}$:

$$\mathfrak{R}' = \left\{ h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*, \quad \text{where } h,k,l \text{ are integers} \right\}. \qquad (9)$$

Note that these are exactly the same scattering vectors for which the diffraction conditions (6) are satisfied, that is, whose structure factor magnitudes can be measured in the experiment.

1.2. Scattering by a single particle. Reduction to crystallographic problems

Let us go back to the case of scattering by a single molecule. Let $\rho_0(\mathbf{r})$ describes the electron density distribution in a single molecule, and $\mathbf{F}^{sp}(\mathbf{s})$ are the corresponding structure factors calculated according to (3). Let us choose an arbitrary basis $\{\mathbf{a},\mathbf{b},\mathbf{c}\}$ in the space. Let the vectors be chosen big enough for the molecule to be positioned completely in the

parallelepiped built on these vectors. Among all structure factors $\{\mathbf{F}(\mathbf{s})\}$, we will select a subset corresponding to the nodes of the integer grid

$$\mathfrak{R}'_{abc} = \left\{ h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*, \text{ where } h, k, l \text{ are integers} \right\}, \tag{10}$$

where $\{\mathbf{a}^*, \mathbf{b}^*, \mathbf{c}^*\}$ is the basis conjugated to the basis $\{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$. Consider, finally, an imaginary crystal with the unit cell parameters $\{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$ inside which there is one molecule with electron density distribution $\rho_0(\mathbf{r})$ inside the molecule and the density outside the boundaries of the molecule equal to zero. Structure factors for this crystal have the form:

$$\mathbf{F}^{cryst}(\mathbf{s}) = \int_{V_{abc}} \rho_0(\mathbf{r}) \exp\left[2\pi i(\mathbf{s}, \mathbf{r})\right] dV_{\mathbf{r}} = \int \rho_0(\mathbf{r}) \exp\left[2\pi i(\mathbf{s}, \mathbf{r})\right] dV_{\mathbf{r}} = \mathbf{F}^{sp}(\mathbf{s}). \tag{11}$$

Thus, the setting of the Fourier transform for a single particle on an integer grid $\mathfrak{R}'_{abc}$ is equivalent to the setting of structure factors for an imaginary crystal structure in which the particle is in the crystal unit cell with the parameters $\{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$. And, correspondingly, the problem of the recovery of electron density from the magnitudes of the Fourier transform $\left|\mathbf{F}^{sp}\right|(\mathbf{s})$ can be reformulated as the problem of the recovery of the electron density distribution in the imaginary cell from structure factor magnitudes $\left\{ \mathbf{F}^{cryst}(\mathbf{s}): \ \mathbf{s} \in \mathfrak{R}'_{abc} \right\}$.

Because of this, a vast majority of methods of the solution of the phase problem in protein crystallography could be applied to the determination of the structure of a single particle. It should be noted that, with this reduction of the problem to the problem of the macromolecular crystallography, we use only a part of the potentially available information, i.e., only the Fourier transform magnitudes taken in the nodes of the grid (10). Both the basis $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ and the lengths a, b, c of the unit cell can be selected by a variety of ways. The data on the scattering by a single particle contain much more information than a usual crystallographic problem; and this gives more possibilities for the solution of the phase problem.

Although, the formulation of the problem suggested in the previous section is formally similar to the ordinary crystallographic problem, it has an essential difference. With large values of a, b, c parameters, we obtain a unit cell in which only a small part is occupied by an unknown electron density, and density values in the greater part of the unit cell are known (equal to zero). This makes the problem of the structure determination overdetermined [6] and makes possible its solution. A significant obstacle for the solution is that we do not know in which particular points the density is equal to zero. If this information (a mask of the molecule region) is obtained by some way, different iteration procedures can be applied for the solution of the phase problem [6–8]. In this paper, we discuss the way to obtain this mask. It should be mentioned that, by increasing the cell parameters, we involve a great deal of experimental information (a greater quantity of reflections) in the work and deal with a greater number of structure factors of the imaginary crystal (with the resolution being the same). On the other side, this increase leads to an increase in the portion of the solvent in the unit cell, which increases the information redundancy and can facilitate the solution of the phase problem.

1.3. Phase problem. *Ab initio* methods of the solution

One of the central problems of the crystal structure determination is a loss in the experiment of "half" of the necessary information. For the recovery of the electron density using (8), both the magnitudes and phases of structure factors are necessary; an experiment allows one to obtain only values of the magnitudes. The lack of the information has to be compensated for by some other information about the object. As a rule, either the

experimental data obtained in additional experiments (with isomorphous derivatives or at different wavelengths of X-rays) or the known structure of a homologous protein are used as an additional information. A special group of methods are so-called direct or *ab initio* methods in which additional general information not relevant to a particular object under investigation is used. An example of this information is the connectivity of the regions of high electron density values [4, 9]. For using this property, a Monte-Carlo type procedure was developed, and the complex of the programs GENNEM was created in which this procedure is implemented. The procedure consists of a few steps. In the first step, a hypothetical set of possible values of structure factor phases is randomly generated. The phases with experimental structure factor magnitudes are used to calculate the Fourier synthesis of the electron density. From the Fourier synthesis, a region of high density values is determined (a cut-off level is a parameter of the method), and the number of the connectivity components of this region is determined. If the number of these components is no greater than the given number, the phase set generated is considered as "admissible" and is selected for further analysis. The generation of the phase sets is repeated until the number of selected admissible phase sets reaches a given number. At the second step, the phase sets selected are "aligned" [10, 11]. The necessity of the alignment is dictated by the fact that the structure factor magnitudes do not fix the choice of the origin. The electron density distribution of the initial object and that of the same object but shifted by some vector have identical sets of structure factor magnitudes (but different sets of phases). Therefore, with the random generation of the phase sets, a situation may arise that two phase sets generated would lead to the same (or very similar) images of the molecule but formally different phase values. This difference can be eliminated by a shift of the second image (and corresponding transformation of the phases). After the alignment of all phase sets, the phase values for each structure factor are averaged, and an indicator of the dispersion of the phase value in different sets (a figure of merit) is calculated. The mean values and the figures of merit (as correcting multipliers) are used to calculate Fourier series (8), which is a current approximation to the solution of the problem of determining the electron density in the object. The approximation found can be used for the modification of the first step of the procedure. Random phase values can be then be generated with allowance for the information about their most probable values. The procedure can be repeated several times, with varying individual probability distributions upon the generation of the values of structure factor phases until the attainment of convergence.

An additional criterion for the selection within this procedure can be restrictions upon the molecule size; thus, it is required that the region of high values was within the parallelepiped of the given size.

1.4. Using masks of the molecule region for the solution of the phase problem

In case of low and middle resolution, the main information extracted from the electron density distribution $\rho(\mathbf{r})$ is usually a mask of the molecule region (characteristic function). This mask is usually defined as a region of the highest values on the Fourier synthesis of the electron density (referred below to as the region of high values, RHV). With a given cut-off level ρ_{crit}, we define a mask of the region as a binary function

$$\rho^{mask}(\mathbf{r}) = \begin{cases} 1, & \text{if } \rho(\mathbf{r}) \geq \rho_{crit} \\ 0, & \text{if } \rho(\mathbf{r}) < \rho_{crit} \end{cases}. \tag{12}$$

An alternative way to define a RHV is to preset the desired size of the high values region and choose the cut-off level ρ_{crit} so that function (12) would distinguish the region of the desired size. The desired volume of the region can be specified by different ways. This can be, for example, the absolute volume of the region in the three-dimensional space (in Å3) or the number of grid nodes if the region is built on some grid in the unit cell. This can be also the

relative volume – that is the ratio of the region volume to the volume of the unit cell. In some cases, it is convenient to characterize the size of the region by a volume that corresponds to a particular structure unit. For this purpose, we will use a specific volume, which is calculated as an average volume per one amino acid residue of the protein structure being studied.

At the initial steps of the structure investigation, an object of the search can be not the electron density distribution itself but a binary mask that describes best the molecule. Moreover, the property of the binarity per se can be used as an additional information about the object under study for the reconstruction of phase values [12].

In this work, we suggest a new approach to the *ab initio* structure determination of biological molecules at low and middle resolution. The approach can be applied in cases of both crystalline samples and a single molecule, but its computing advantages become particularly significant in studies of single molecules.

The object of the search is a mask of the molecule region given at some grid in the unit sell

$$\{b_{mnp}\}, \quad 0 \le m < N_x, 0 \le n < N_y, 0 \le p < N_z . \tag{13}$$

Here, N_x, N_y, N_z are the numbers of grid nodes on the sides of the unit cell. At the first step, connected sets of points consisting of a given number of points are randomly generated. To each set of generated points Ω, the binary function

$$\rho^{mask}(\mathbf{r}) = \begin{cases} 1 & \text{for } \mathbf{r} \in \Omega \\ 0 & \text{for } \mathbf{r} \notin \Omega \end{cases} \tag{14}$$

is constructed. From this function, the magnitudes and phases of structure factors are calculated. If the calculated magnitudes are close enough to the experimental values (for example, have a correlation above the given value), then the corresponding phase set is considered as admissible and is kept for further work. The generation is repeated until the necessary number of admissible phase sets is selected. The next two steps of the work, the alignment and averaging, are accomplished as described above in section 1.3. Upon the selection of generated masks, additional requirements of the selection can be imposed. These requirements are formulated in term of the characteristics of masks; for example, the extension of the masks in different directions is limited or the required radius of the inertia of masks is given. At the beginning of the work with the structure, the addition of nodes to the mask being generated can occur with an equal probability for all grid nodes. With the advance of the work, the Fourier synthesis obtained after the averaging of selected phase sets can be converted into the distribution of probabilities for the presence in the mask of particular grid nodes. After this, the generation of the mask at the next circle can be performed with allowance for this probability distribution.

2. RESULTS

2.1. Test object

The atomic model of the trimer of the membrane protein AcrB [13] consisting of 3129 amino acid residues (23811 non-hydrogen atoms) was used. The overall view of the trimer is shown in Fig. 2.

Two test sets of complex structure factors corresponding to two variants of an "imaginary" crystal structure were calculated. The AcrB molecule was assumed to be placed in the rectangular unit cell with dimensions 120, 120, and 150. Å in the first case (the small unit cell) and 180, 180, 225 Å in the second case (the extended unit cell). In both cases, the space group was assumed to be *P*1 (the absence of crystallographic symmetry). The first case resembles an ordinary task of the crystal structure determination in the presence of a large volume of the solvent (about 75%). The second case corresponds to an imaginary crystal

t7

(with an abnormally large volume of the solvent) considered in the single particle X-ray study. The structure factor magnitudes are considered as known experimentally determined $F^{obs}(\mathbf{s})$ values and referred below as "experimental" values. The calculated phases were assumed to be exact phase values $\varphi^{exact}(\mathbf{s})$ and were used only for the control of the results. During tests, they were considered as unknown.
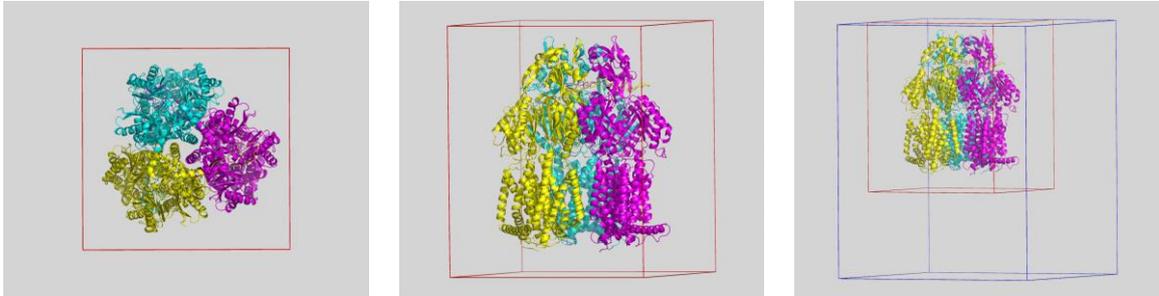

Fig. 2. Overall structure of the AcrB protein. Three monomers comprising the molecule are shown by different colors. The boundaries of a small and an extended unit cell are shown in red and blue.

It should be noted that the AcrB molecule has a symmetry axis of the third order. The symmetry was not taken into account either in the choice of the unit cell or during the retrieval of the values of structure factor phases. Therefore, the manifestation of this symmetry on molecule images obtained served as a confirmation of the correctness of the solution.

The *ab initio* determination of the structure factor phases was performed in a resolution zone of 25 Å. For the control, different indicators were calculated also in the zones of lower resolution of 40 and 30 Å. The distribution of the number of the structure factors in the resolution zones is shown in Table 1.

Table 1. Distribution of the number of structure factors in the resolution zones

	Dimensions of the unit cell [Å]	Resolution zone (d_{min}) [Å]		
		40	30	25
		Number of structure factors		
Small unit cell	120, 120, 150	69	170	288
Extended unit cell	180, 180, 225	247	555	976

2.2. Control criteria

The choice of the most optimal criteria for the comparison of the sets of the phases, or structure factor magnitudes, or Fourier syntheses of the electron density is a subject of the separate discussion in protein crystallography. In this paper, we use two of the most popular criteria. For the comparison of structure factor magnitudes calculated from a binary mask and their experimental values $\left\{F^{obs}(\mathbf{s})\right\}$, we use the non-centric correlation coefficient:

$$CM = \frac{\sum_{\mathbf{s}} F^{obs}(\mathbf{s}) F^{calc}(\mathbf{s})}{\sqrt{\sum_{\mathbf{s}} \left(F^{obs}(\mathbf{s})\right)^2 \sum_{\mathbf{s}} \left(F^{calc}(\mathbf{s})\right)^2}} . \tag{15}$$

The criterion for the comparison of two phase sets (for example, *ab initio* solution and the exact phases) is defined in two steps. At the first step, we define a "formal" correlation coefficient of two phase sets as:

$$FCP\left(\{\varphi_1(\mathbf{s})\},\{\varphi_2(\mathbf{s})\}\right)=FCP\left(\rho_1(\mathbf{r}),\rho_2(\mathbf{r})\right)=$$

$$\frac{\int\rho_1(\mathbf{r})\rho_2(\mathbf{r})dV_{\mathbf{r}}}{\sqrt{\int(\rho_1(\mathbf{r}))^2dV_{\mathbf{r}}\int(\rho_2(\mathbf{r}))^2dV_{\mathbf{r}}}}=\frac{\sum_{\mathbf{s}}\left(F^{obs}(\mathbf{s})\right)^2\cos\left(\varphi_1(\mathbf{s})-\varphi_2(\mathbf{s})\right)}{\sum_{\mathbf{s}}\left(F^{obs}(\mathbf{s})\right)^2}. \qquad (16)$$

Here, $\rho_1(\mathbf{r})$ and $\rho_2(\mathbf{r})$ are the Fourier syntheses calculated with experimental magnitudes $F^{obs}(\mathbf{s})$ and the phase sets being compared. (The term $\mathbf{F}(0)$ is omitted in these calculations.) It should be mentioned that this criterion is specific for a particular X-ray experiment because it is calculated using the experimental structure factor magnitudes.

The above formal comparison of phase sets is not always acceptable in the tasks of the reconstruction of phase values from the structure factor magnitudes. This is because the structure factor magnitudes do not fix the origin of the coordinates. To be more precise, two functions, $\rho_1=\rho(\mathbf{r})$ and $\rho_2=\rho(\mathbf{r}-\mathbf{u})$, which differ by a shift at the vector $\mathbf{u}$, have the same structure factor magnitudes but different phase values

$$\varphi_2(\mathbf{s})=\varphi_1(\mathbf{s})+2\pi(\mathbf{s},\mathbf{u}). \qquad (17)$$

These two functions provide one and the same image of the structure and have to be considered, in the frame of the structure determination task, as equivalent. During the random generation of masks or phases, a case is possible that the images being generated (or Fourier synthesis calculated from phases being generated) differ by only a shift or become very similar after the application of the properly selected shift. In the frame of our task, these phase sets have to be considered as close. Therefore, as the second step in the determination of the level of closeness, we introduce the alignment, i.e. the selection of a vector $\mathbf{u}$ that would provide a maximal value of the formal criterion (16):

$$CP\left(\{\varphi_1(\mathbf{s})\},\{\varphi_2(\mathbf{s})\}\right)=\max_{\mathbf{u}}FCP\left(\rho_1(\mathbf{r}-\mathbf{u}),\rho_2(\mathbf{r})\right)=$$

$$\max_{\mathbf{u}}\frac{\sum_{\mathbf{s}}\left(F^{obs}(\mathbf{s})\right)^2\cos\left(\varphi_1(\mathbf{s})+2\pi(\mathbf{s},\mathbf{u})-\varphi_2(\mathbf{s})\right)}{\sum_{\mathbf{s}}\left(F^{obs}(\mathbf{s})\right)^2}. \qquad (18)$$

In the space group P1, the maximization is performed over all vectors of the shift $\mathbf{u}$. A set of possible vectors of the shift can be limited in the case of the presence of crystallographic symmetry [11].

One more transformation of the function of the electron density distribution that does not lead to a change in the structure factor magnitudes is the transition to an enantiomer; the functions $\rho(\mathbf{r})$ and $\rho(-\mathbf{r})$ have the same set of the structure factor magnitudes. At low and medium resolutions, the choice of a correct enantiomer is usually impossible. Therefore, in the procedure of the alignment, an additional possibility of the change of the enantiomer is introduced in cases that such a change is not limited due to the symmetry (for example, when one works with the group *P*1):

$$CP\left(\{\varphi_1(\mathbf{s})\},\{\varphi_2(\mathbf{s})\}\right)=\max_{\kappa=\pm1}\max_{\mathbf{u}}FCP\left(\rho_1(\kappa\mathbf{r}-\mathbf{u}),\rho_2(\mathbf{r})\right). \qquad (19)$$

t9

2.3. Averaging of the phase sets. Weighted Fourier synthesis

In ab initio phase determination procedures discussed in this work, the result of the first step is a population of the phase sets chosen. The simplest processing procedure of the phase sets is their averaging, which is applied to the phase sets previously aligned relative to each other [4]. To be more precise, the averaging is applied to the Fourier syntheses calculated using experimental structure factor magnitudes $F^{obs}(\mathbf{s})$ and different phase sets. This leads to a "weighted" Fourier synthesis which can be calculated as

$$\rho(\mathbf{r}) = \frac{1}{|V|} \sum_{\mathbf{s}} m(\mathbf{s}) F^{obs}(\mathbf{s}) \exp\left[i\varphi^{best}(\mathbf{s})\right] \exp\left[i2\pi(\mathbf{s},\mathbf{r})\right]. \tag{20}$$

For each structure factor, the best phase $\varphi^{best}(\mathbf{s})$ and the indicator of the reliability of its determination (the figure of merit) are determined:

$$m(\mathbf{s}) \exp\left[i\varphi^{best}(\mathbf{s})\right] = \frac{1}{M} \sum_{j=1}^{M} \exp\left[i\varphi_j(\mathbf{s})\right]. \tag{21}$$

It is easily seen that $m(\mathbf{s})$ characterizes the scatter of phase values in different sets relative to the averaged phase value $\varphi^{best}(\mathbf{s})$:

$$m(\mathbf{s}) = \frac{1}{M} \sum_{j=1}^{M} \cos\left(\varphi_j(\mathbf{s}) - \varphi^{best}(\mathbf{s})\right). \tag{22}$$

When using the weighted Fourier syntheses (20), we will also apply, along with the phase correlation coefficient, the weighted correlation coefficient CP_w, which is the result of the optimization over all possible phase shifts

$$CP_w\left(\{\varphi_1(\mathbf{s})\}, \{\varphi_2(\mathbf{s})\}\right) = \max_{\kappa=\pm 1} \max_{\mathbf{u}} FCP_w\left(\rho_1(\kappa\mathbf{r} - \mathbf{u}), \rho_2(\mathbf{r})\right), \tag{23}$$

where instead of the formal correlation coefficient (16), the weighted formal correlation coefficient is used:

$$FCP_w\left(\{\varphi_1(\mathbf{s})\}, \{\varphi_2(\mathbf{s})\}\right) = \frac{\sum_{\mathbf{s}} m(\mathbf{s})\left(F^{obs}(\mathbf{s})\right)^2 \cos\left(\varphi_1(\mathbf{s}) - \varphi_2(\mathbf{s})\right)}{\sqrt{\sum_{\mathbf{s}}\left(F^{obs}(\mathbf{s})\right)^2 \sum_{\mathbf{s}}\left(m(\mathbf{s})F^{obs}(\mathbf{s})\right)^2}}. \tag{24}$$

2.4. The averaging of random phase sets. The efficiency of *ab initio* procedures

At low resolution, the correlation coefficient CP as a criterion of success should be used with caution. The point is that the sets of the structure factor magnitudes for biological macromolecules often have anomalously large values for a relatively small number of reflections in the low resolution zone (Fig. 3). This is why the images on low resolution maps can be determined mostly by a small number of these strong reflections. In this case, high CP values can indicate a consensus of phase values only for a small number of dominant reflections. In addition, it is necessary to take into account that when random and exact phase values are compared, then the intuitively expected zero value of the correlation coefficient can appear only when the formal criterion FCP is used. In case of the preliminary alignment of the phase sets being compared, the values of the correlation coefficient can be much higher. The distributions of the values of the CP correlation coefficient calculated for randomly generated phase sets are shown on Fig. 4, and the characteristics of these distributions are given in Table 2.

The averaging of the preliminary aligned phase sets leads to greater values of the correlation coefficient CP_w (Tables 3, 4). Thus, for a test object in the resolution zone of 40 Å, the correlation coefficient value of 0.7 can be obtained simply by the averaging of randomly generated phase sets. This base value has to be taken into account when evaluating the success of any procedure of the solution of the phase problem. The procedure can be considered successful only if it allows one to obtain higher values of the correlation coefficient than those allowed by the averaging of random phase sets.
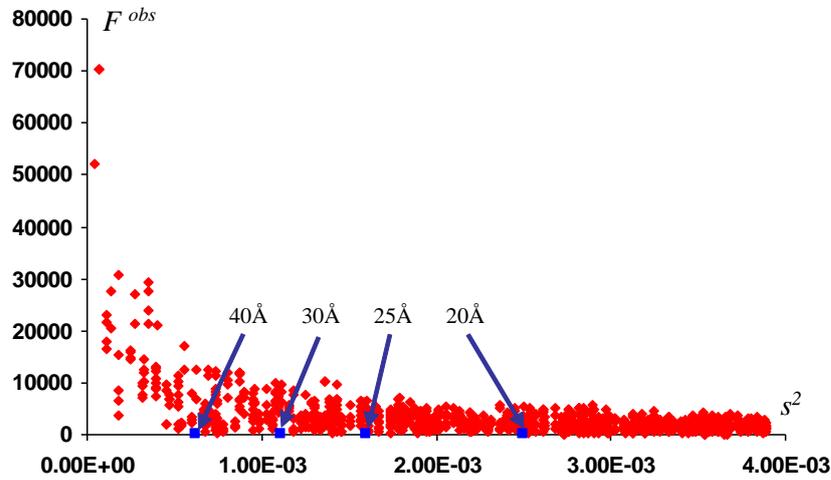

Fig. 3. Structure factor magnitudes for AcrB (the small cell case). For each reflection, the structure factor magnitude is shown as a function of s^2.
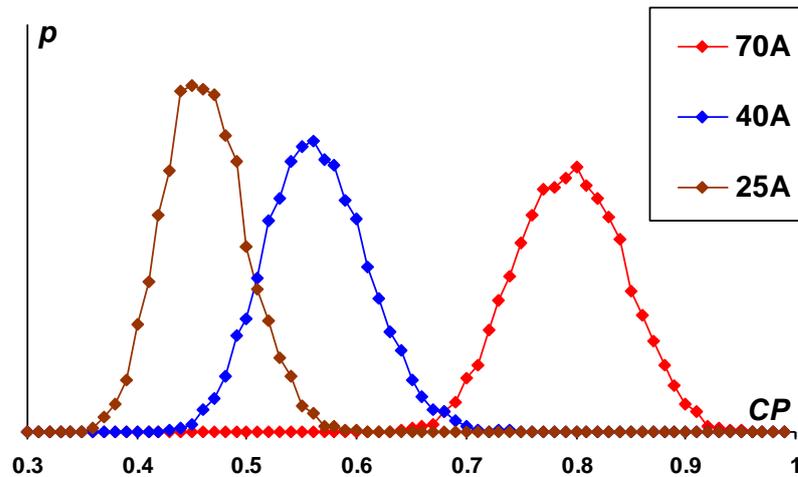

Fig. 4. Empirical distributions of the correlation coefficient CP of exact and randomly generated phases of the structure factors in different resolution zones (AcrB, the small cell case)**.**

Table 2. Values of the correlation coefficient CP for randomly generated and exact phase values (AcrB, the small cell case)

Resolution d_{min} [Å]	Number of reflections	CP				
		min	max	ave	sigma	ave+3sigma
40	69	0.4457	0.7739	0.5791	0.0464	0.7183
30	170	0.3863	0.6971	0.512	0.0412	0.6355
25	288	0.3585	0.6433	0.4767	0.0379	0.5904

t11

Table 3. Values of the correlation coefficient CP_w for exact phases and phases obtained by the averaging of 100 randomly generated phase sets (AcrB, the small cell case, five independent tests *a–e*)

Resolution d_{min} [Å]	Variant				
	a	*b*	*c*	*d*	*e*
	CP_w				
40	0.696	0.713	0.679	0.713	0.718
30	0.646	0.653	0.643	0.626	0.665
25	0.625	0.632	0.623	0.647	0.644

Table 4. Values of the correlation coefficient CP_w for exact phases and phases obtained by the averaging of 100 randomly generated phase sets (AcrB, the extended cell, five tests *a-e*)

Resolution d_{min} [Å]	Variant				
	a	*b*	*c*	*d*	*e*
	CP_w				
40	0.50	0.54	0.55	0.56	0.49
30	0.48	0.52	0.52	0.54	0.46
25	0.46	0.40	0.50	0.52	0.45

A comparison of Tables 3 and 4 shows that the quality of averaged random phases decreases with an increase in the cell dimensions of the imaginary crystal.

2.5. Generation of random phase sets. The criterion of finiteness for the region of high electron density values. Restriction of the number of connected components

Here, the tests were arranged in the following way. In each test, a desired volume V_{HDR} of the region of high electron density values was set. A great number (up to hundred millions) of random phases was generated. For every phase set, the Fourier synthesis was calculated using these phases and the set of the experimental structure factor magnitudes. A region of high values of the given volume V_{HDR} was selected, and it was checked whether this region can be placed entirely in the crystal cell without crossing the border. If this condition was satisfied, the checked phase set was considered as admissible and was selected for further analysis. The generation of phase sets was continued until 100 admissible phase sets were selected. Then, these 100 phase sets were aligned and averaged. The accuracy of the phase sets thus obtained is given in Tables 5 and 6. For each preset value of volume V_{HDR}, the test was repeated five times with different starting constants for the generator of random numbers. It is necessary to point out that at big V_{HDR} values, the calculation procedure required considerable processor time (about a full day).

Table 5. Accuracy of the phase values obtained by averaging the randomly generated phase sets leading to a finite region of high values on the Fourier synthesis. The phase correlation CP_w values calculated in different resolution zones (40, 30, 25 Å) are given for five independent tests (the small cell case)

	Volume of high density region: specific [Å³ per residue]/relative/number of points						
	50	75	100	125	150	175	200
	0.072	0.109	0.145	0.181	0.218	0.254	0.290
	334	501	668	834	1002	1168	1336
CP_w*100 40/30/25Å	69/65/63	78/72/70	76/71/69	81/75/73	84/78/75	87/81/78	
	74/69/67	76/72/69	81/75/73	82/77/74	82/77/74		
	75/70/67	76/71/68	79/74/71	80/75/72	86/80/77	86/80/77	
	74/69/66	78/73/70	80/75/72	83/78/75	86/80/77		
	75/70/68	75/68/66	81/76/75	83/78/75	85/79/75	86/80/78	

t12

Table 6. Accuracy of the phase values obtained by averaging the randomly generated phase sets leading to a finite region of high values on the Fourier synthesis. The phase correlation CP_w values calculated in different resolution zones (40, 30, 25 Å) are given for five independent tests (the extended cell case)

	Volume of high density region: specific [Å³ per residue]/relative/number of points						
	75 0.0322 436	100 0.0429 582	125 0.0636 727	150 0.0644 873	175 0.0751 1018	200 0.0859 1164	225 0.0966 1309
CP_w*100	53/51/49	55/53/51	50/48/46	62/59/57	65/62/60	73/69/67	75/71/69
	57/54/53	55/52/50	54/51/50	64/61/60	66/63/61	72/68/66	74/70/68
40/30/25Å	53/51/49	65/62/60	66/63/61	67/64/62	67/64/62	72/69/66	73/70/67
	56/53/52	56/54/52	63/60/58	65/62/60	67/64/62	70/67/65	73/69/67
	49/47/46	51/49/48	61/58/56	61/58/56	63/59/58	69/66/64	73/69/67

An analysis of Tables 3 and 4, as well as Tables 5 and 6, allows one to make following conclusions. First, the procedure described in this section makes it possible to obtain a higher quality of averaged phases than in case of phases obtained by averaging randomly generated phases without any preliminary selection. Therefore, this procedure can be considered as a working ab initio procedure of the determination of the phases. Second, an increase in the cell volume with the retention of the specific volume of the region of high values having a physical sense leads to a considerable decrease in the quality of the phase sets obtained. In other words, the procedure considered seems to be more suited for the work with "real" crystals than with "imaginary" crystals in studies of single particles.

In the selection of generated phase sets, the attempt to impose additional restrictions on the number of connectivity components in the region of high values led only to an inconsiderable increase in the quality of averaged phase sets.

2.6. Generation of random connected masks. Averaging without selection

The goal of this series of tests was to examine how useful, in terms of the phase problem solution, the restrictions such as the connectivity and binarity of the region of high values may be. The test was performed in the following way. In the unit cell, a uniform grid with a step approximately equal to the third of the nominal resolution of 25 Å (i.e., about 8.3 Å) was introduced. Each test was performed for a given volume V_{HDR} of the region of high values on the Fourier synthesis of the electron density, with the number of the grid nodes inside the mask N_{mask} corresponding to this volume (Tables 7, 8). A hundred of random connected masks consisting of N_{mask} nodes were generated. Each mask was used for the calculation of the structure factor magnitudes and phases. These 100 phase sets were aligned and averaged. The accuracy of the phase values thus obtained is given in Tables 7, 8. For each given value of the volume V_{HDR}, the test was repeated five times with different start constants of the generator of random numbers.

Table 7. Results of the averaging of phase sets calculated from randomly generated connected masks. The phase correlation CP_w values calculated in different resolution zones (40, 30, 25 Å) are given for five independent tests (the small cell case)

	The volume of the mask: specific [Å³ per residue]/relative/number of points						
	50 0.072 334	75 0.109 501	100 0.145 668	125 0.181 834	150 0.218 1002	175 0.254 1168	200 0.290 1336
CP_w*100	61/57/55	70/65/63	77/73/70	77/72/69	85/80/77	85/80/77	85/80/77
	64/60/57	70/65/63	75/71/68	81/77/74	83/78/75	85/79/77	85/80/77
	66/61/58	72/67/65	78/74/71	80/75/72	84/80/77	87/82/79	84/79/76
40/30/25Å	63/59/57	72/68/66	78/73/71	80/76/73	82/77/74	86/81/78	87/81/78
	62/57/55	71/65/63	75/70/68	81/76/73	87/82/79	83/78/75	86/80/77

Table 8. Results of the averaging of phase sets calculated from randomly generated connected masks. The phase correlation CP_w values calculated in different resolution zones (40, 30, 25 Å) are given for five independent tests (the extended cell case)

	The volume of the mask: specific [Å³ per residue]/relative/number of points						
	50 0.072 334	75 0.109 501	100 0.145 668	125 0.181 834	150 0.218 1002	175 0.254 1168	200 0.290 1336
CP_w*100 40/30/25Å	61/57/55	70/65/63	77/73/70	77/72/69	85/80/77	85/80/77	85/80/77
	64/60/57	70/65/63	75/71/68	81/77/74	83/78/75	85/79/77	85/80/77
	66/61/58	72/67/65	78/74/71	80/75/72	84/80/77	87/82/79	84/79/76
	63/59/57	72/68/66	78/73/71	80/76/73	82/77/74	86/81/78	87/81/78
	62/57/55	71/65/63	75/70/68	81/76/73	87/82/79	83/78/75	86/80/77

An analysis of the tables allows one to make the following conclusions. First, the procedure described above makes it possible to obtain phases of higher quality than the quality of averaged randomly generated phase sets used without any preliminary selection. Thus, this procedure can be considered as a working ab initio procedure for phase determination. Second, a comparison of Tables 5 and 7 shows that, in the case of a small cell, the quality of the phase sets obtained as a result of the averaging is approximately the same in both cases, when phases are randomly generated and the selection of the variants is based on the demand of the finiteness of HDR, and when connected masks are generated and the phases calculated from the masks are directly averaged. However, the computational expenses with the second approach are considerably lower. Third, in the case of the averaging of phases obtained from connected masks, the extending of unit cell dimensions leads to a considerable increase in the quality of the solution of the phase problem. Thus, the new approach implements the advantages when working with a single particle, which are attained due to the possibility of using more experimental information (a greater number of structure factor magnitudes, which are included in the calculation when one works with a large unit cell of an imaginary crystal).

2.7. Generation of random connected masks. Selection based on the correlation of structure factor magnitudes

In this series of tests, an additional requirement to randomly generated masks was included, namely the correspondence of the structure factor magnitudes calculated from the masks to the experimental ones.

The test was performed in the following way. In the unit cell, a uniform grid with a step approximately equal to the third of the nominal resolution of 25 Å (i. e., about 8.3 Å) was introduced. Each test was carried with two parameters preset: the volume V_{HDR} of the region of high values on the Fourier synthesis of the electron density and the required accuracy of the correspondence of structure factor magnitudes calculated from the mask to the experiment. A great number of random connected masks consisting of N_{mask} nodes were generated. Each generated mask was used to calculate structure factor magnitudes and phases. For the calculated structure factor magnitudes, the coefficient of their correlation with experimental values (15) was calculated. If this coefficient was higher than a given threshold value, the generated phase set was considered as acceptable and was kept for further analysis. The generation was continued until 100 acceptable phase sets were selected. The selected phase sets were aligned and averaged.

Table 9. Results of the averaging of phase sets calculated from randomly generated connected masks having a given level of correlation (25) of structure factor magnitudes. The phase correlation CP_w values calculated in different resolution zones (40, 30, 25 Å) are given for five independent tests (the small cell case). The best result is highlighted

CP_w* 100		Mask volume: specific [Å³ per residue]/relative/number of points						
40/30/25 Å		50 0.072 334	75 0.109 501	100 0.145 668	125 0.181 834	150 0.218 1002	175 0.254 1168	200 0.290 1336
Correlation of magnitudes z_{crit} (25 Å)	1.5	60/57/54	69/64/62	72/67/64	80/75/72	87/82/79	88/83/80	87/82/79
	2.0	61/57/55	68/63/61	73/68/66	83/78/75	88/83/80	**90/84/81**	89/83/81
	2.5	59/55/53	65/60/57	73/68/66	81/76/73	88/83/80	89/83/80	89/83/79
	3.0	59/55/53	62/58/56	71/66/64	81/76/73	88/83/79	83/78/74	

Table 10. Results of the averaging of phase sets calculated from randomly generated connected masks having a given level of correlation (25) of structure factor magnitudes. The phase correlation CP_w values calculated in different resolution zones (40, 30, 25 Å) are given for five independent tests (the extended cell case). The best results are highlighted

| CP_w* 100%		Mask volume: specific [Å³ per a.a.]/relative/number of points						
40/30/25 Å		50 0.0215 291	75 0.0322 436	100 0.0429 582	125 0.0636 727	150 0.0644 873	175 0.0751 1018	200 0.0859 1164
Correlation of magnitudes z_{crit} (25 Å)	1.5	72/69/67	78/74/71	86/82/79	89/85/83	91/86/84	91/87/85	90/86/83
	2.0	72/68/66	77/73/71	85/81/79	90/86/84	91/87/85	92/88/85	91/87/84
	2.5	71/68/66	77/73/70	88/84/82	91/87/84	92/88/85	92/88/86	90/86/83
	3.0	71/69/67	81/77/74	89/85/83	92/88/85	**94/90/87**	**94/90/87**	93/88/86
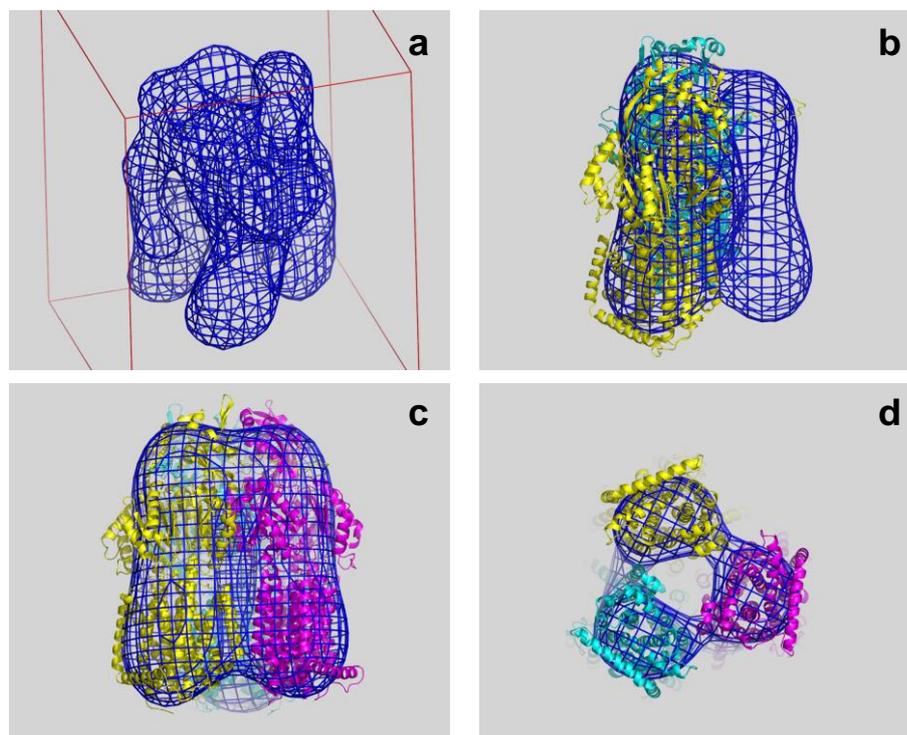
Fig. 5. AcrB model and the regions of high electron density values on unweighted Fourier synthesis calculated with experimental values of structure factor magnitudes and ab initio determined phases: a – resolution 25 Å, cut-off level 3σ; b – resolution 40 Å, cut-off level 3σ, one of the monomers of the model is not shown; c – resolution 40 Å, cut-off level 2σ; d – resolution 40 Å, cut-off level 3σ.

t15

The accuracy of the phase values obtained in this way is given in Tables 9 and 10. In these tables, the rigidity of the selection of acceptable masks is given in terms of the normalized z_{CM} score, which was determined in the following way. As a preliminary step, a great number (in our tests, 10000) of random connected masks of the given size were generated. For each mask, the coefficient of the correlation between the calculated and the exact structure factor magnitudes was calculated. For 10000 values of CM, the mean value $\langle CM \rangle$ and the mean standard deviation σ_{CM} were calculated. Then, when the mask generation was performed with the selection, the calculated CM value for each set of structure factor magnitudes was transformed to a normalized value

$$z_{CM} = \frac{CM - \langle CM \rangle}{\sigma_{CM}}, \tag{25}$$

which just was compared with the given threshold value z_{crit}.

The tables show that the introduction of the selection according to the correspondence of the calculated structure factor magnitudes to the experiment significantly improves the quality of the resulting phases compared with that attained in the averaging procedure without selection. Fig. 5 shows the images of the object obtained from the Fourier synthesis calculated with experimental structure factor magnitudes and the best phases found.

ADDITIONAL NOTES AND CONCLUSIONS

The testing of the new approach to the ab initio solution of the phase problem of the biological crystallography showed its applicability in both traditional studies of crystal species and investigations of single biological macromolecular objects.

The approach is based on the Monte-Carlo type procedure [2, 3], which consists of a few steps. At the first step, an object of the search is binary masks serving as an approximation of the region of high values of the electron density function in the object under study. For this purpose, a great number of connected binary masks are randomly generated, and for each of them, the structure factor magnitudes and phases are calculated. The mask (and the phases corresponding to it) is considered as admissible if the level of the correspondence of the structure factor magnitudes calculated from the mask and the experimental values exceeds a given limit. The generation is performed until the preset number of admissible phase sets are obtained. A criterion of the closeness of the structure factor magnitudes may be the coefficient of the magnitude correlation, although other criteria may be used (for example, the statistical likelihood [16]). Along with the criterion of the closeness of structure factors, the criteria of another type may be used at the step of the selection of masks, which are related to their expected parameters or are based on the results of other experiments (e.g., small angle X-ray diffraction or neutron scattering). Random masks can be generated either on assumption of the equal probability of all grid nodes in the unit cell or with the use of a priori preferences established at previous steps of the work. In case of large unit cell values and rigid restrictions on the degree of correspondence of the mask to experimental data, the generation of the required number of admissible masks can demand a significant amount of the computer time. However, the natural parallelism of the selection process allows one to significantly decrease time expenses in the work on computers with parallel architecture.

At the second step, the alignment and, in the simplest case, the averaging of the phase sets are performed. A more accurate procedure involves using the methods of the cluster analysis, which allows one to distinguish, among selected variants, compact clusters for further averaging inside these clusters. It should be noted that the averaging and cluster analysis are performed using special metric relevant to a specific X-ray experiment. The phase values

obtained as a result of the averaging together with structure factor magnitudes determined experimentally are used to build Fourier synthesis.
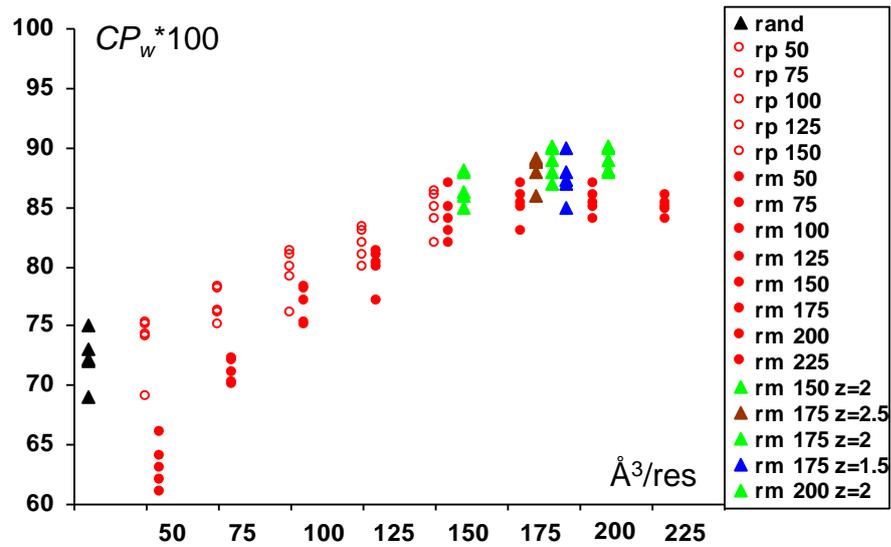

Fig. 6. Phase correlation CP_w*100 in the resolution zone of 40 Å for averaged phase sets with the different size of the region of high values: rand – random phases without selection (section 2.4); rp – random phases with selection based on the finiteness property of the region of high values (section 3.5); rm – random masks without selection (section 2.6); rmz – random masks with selection based on the magnitude correlation at different selection levels (section 2.7). The specific volume of the region of high values is shown on the *X*-axis. For all conditions, the results of five independent experiments in the small cell are shown.
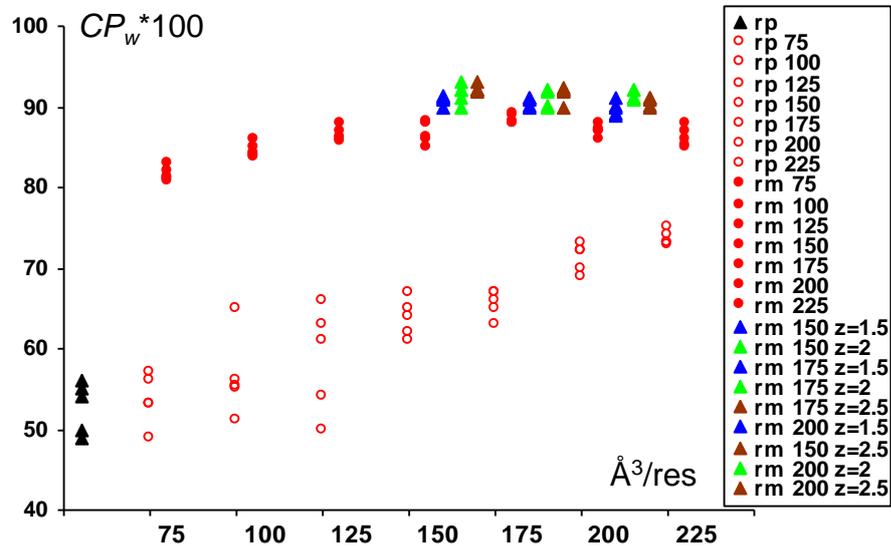

Fig. 7. Phase correlation CP_w*100 in the resolution zone of 40 Å for averaged phase sets with the different size of the region of high values: rand – random phases without selection (section 2.4); rp – random phases with selection based on the finiteness property of the region of high values (section 3.5); rm – random masks without selection (section 2.6); rmz – random masks with selection based on the magnitude correlation at different selection levels (section 2.7). The specific volume of the region of high values is shown on the X-axis. For all conditions, the results of five independent experiments in the extended cell are shown.

t17

The method was tested in two regimes. A "small" unit cell modeled a crystal sample with a large content of the solvent in the cell. An "extended" unit cell modeled a reduced task of structure determination using X-ray diffracting from a single particle.

The results of the testing allowed one to find optimal parameters for the procedure and showed that the procedure does make it possible to advance in the determination of the phases of the structure factors. The accuracy of the values found increases with increasing size of the unit cell, which makes this method especially promising for the investigation of single particles.

In the case of the traditional crystallographic task, a comparison of the new method with the one suggested earlier showed a compatible efficiency of the solution of the phase problem, with the computing efficiency of the new approach being higher. At the same time, the new method leads to a much better solution of the phase problem when one has to do with cells containing artificially increased portions of the solvent.

Fig. 6 and 7 illustrate the accuracy of determining the phases of structure factors with the use of different approaches and parameters of the method.

REFERENCES

1. Krupyanskii Yu.F., Balabaev N.K., Petrova T.E., Sinitsyn D.O., Gryzlova E.V., Tereshkina K.B., Abdulnasyrov E.G., Stepanov A.S., Lunin V.Y., Grum-Grzhimailo A.N.. Femtosecond X-ray free-electron lasers: A new tool for studying nanocrystals and single macromolecules. *Russian Journal of Physical Chemistry B*. 2014. V. 8. P. 445–456.

2. Sinitsyn D.O., Lunin V.Y., Grum-Grzhimailo A.N., Gryzlova E.V., Balabaev N.K., Lunina N.L., Petrova T.E., Tereshkina K.B., Abdulnasyrov E.G., Stepanov A.S., Krupyanskii Yu.F. New possibilities of X-ray nanocrystallography of biological macromolecules based on X-ray free-electron lasers. *Russian Journal of Physical Chemistry B*. 2014. V. 8. P. 457–463.

3. Lunin V.Y., Lunina N.L., Petrova T.E., Skovoroda T.P., Urzhumtsev A.G., Podjarny A.D. Low-resolution ab initio phasing: problems and advances. *Acta Crystallographica Section D: Biological Crystallography*. 2000. V. 56. P. 1223–1232.

4. Lunin V.Y., Urzhumtsev A.G., Podjarny A. *Ab initio* phasing of low-resolution Fourier syntheses. In: *International Tables for Crystallography.* Volume F. Crystallography of Biological Macromolecules. Second Edition. Ed. Arnold E., Himmel D.M., Rossmann M.G. John Wiley & Sons, 2011. P. 437–442.

5. Lunin V.Y., Lunina N.L., Urzhumtsev A.G. Connectivity properties of high-density regions and *ab initio* phasing at low resolution. *Acta Crystallographica Section D: Biological Crystallography*. 2000. V. 56. P. 375–382.

6. 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.

7. Fienup J.R. Reconstruction of an object from the modulus of its Fourier transform. *Optics Letters*. 1978. V. 3. № 1. P. 27–29.

8. Elser V. Solution of the crystallographic phase problem by iterated projections. *Acta Crystallographica Section A: Foundations of Crystallography*. 2003. V. 59. P. 201–209.

9. Baker D., Krukowski A.E., Agard D.A. Uniqueness and the *ab initio* phase problem in macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography*. 1993. V. 49. P. 186–192.

10. Lunin V.Y., Urzhumtsev A.G., Skovoroda T.P. Direct low-resolution phasing from electron-density histograms in protein crystallography. *Acta Crystallographica Section A: Foundations of Crystallography*. 1990. V. 46. P. 540–544.

11. Lunin V.Y., Lunina N.L. The Map Correlation Coefficient for Optimally Superposed Maps. *Acta Crystallographica Section A: Foundations of Crystallography*. 1996. V. 52. P. 365–368.

12. Lunin V.Y., Urzhumtsev A., Bockmayr A. Direct phasing by binary integer programming. *Acta Crystallographica Section A: Foundations of Crystallography*. 2002. V. 58. P. 283–291.

13. Murakami S., Nakashima R., Yamashita E., Yamaguchi A. Crystal structure of bacterial multidrug efflux transporter AcrB. *Nature*. 2002. V. 419. P. 587–593.

14. Urzhumtsev A., Afonine P.V., Lunin V.Y., Terwilliger T.C., Adams P.D. Metrics for comparison of crystallographic maps. *Acta Crystallographica Section D: Biological Crystallography*. 2014. V. 70. P. 2593–2606.

15. Lunin V.Y., Woolfson M.M. Mean Phase Error and the Map Correlation Coefficient. *Acta Crystallographica Section D: Biological Crystallography*. 1993. V. 49. P. 530–533.

16. Lunin V.Y., Afonine P.V., Urzhumtsev A.G. Likelihood-based refinement. I. Irremovable model errors. *Acta Crystallographica Section A: Foundations of Crystallography*. 2002. V. 58. P. 270–282.