The Academy of Sciences of the USSR
Pushchino Research Centre
Research Computing Centre

PREPRINT

V. Yu. LUNIN

USE OF THE ELECTRON-DENSITY-SYNTHESIS HISTOGRAMS FOR SOLVING OF THE PHASE PROBLEM IN PROTEIN CRYSTALLOGRAPHY

PUSHCHINO· 1991

The Academy of Sciences of the USSR
Pushchino Research Centre
Research Computing Centre

PREPRINT

V. Yu. LUNIN

USE OF THE ELECTRON-DENSITY-SYNTHESIS HISTOGRAMS FOR SOLVING OF THE PHASE PROBLEM IN PROTEIN CRYSTALLOGRAPHY

PUSHCHINO· 1991

Summary

This issue deals with a new approach to solving of the phase problem in protein crystallography. It has been recently revealed, that histograms (spectra of frequencies of different electron density values), corresponding to protein electron-density distributions have a specific shape. The shape is sensible to errors in structure factor phases and can be an indicator of correctness of phase determination. This preprint contains a review of the investigations connected with elaboration and application of the new source of information on proteins conducted at Research Computing Centre, USSR Academy of Sciences, (Pushchino, Moscow Region, 142292, USSR).

1. INTRODUCTION

X-Ray structure analysis (XSA) is considered attentively by many researchers concerned with the structure of matter at the atomic level, because it is the main method, that allows to determine object's structure at the atomic resolution. This method allows to determine the three-dimension location of all the atoms of the matter and thus to answer the questions of the molecular structure and its functioning, formulated in terms of geometrical characteristics (distances between atoms, bond lengths, bond or dihedral angles) or characteristics, connected with geometry (surface charge distribution, "admissible" surfaces) and so on.

The principal feature of the X-ray experiment is incompleteness of the data. The experiment allows to measure only the values F_s of modules of complex coefficients (the structure factors) in decomposition of the electron density distribution into Fourier series

$$\rho(r) = \frac{1}{V_{cell}} \sum_s F_s \, e^{i\varphi_s} \, e^{-2\pi i(s,r)} \quad . \qquad (1)$$

The determination of the phases φ_s of the structure factors constitutes the central problem of XSA, namely the "phase problem". Success of the whole work on structure determination is essentially conditioned by the accuracy of solving the phase problem.

Since X-ray experiment do not solve the phase problem directly, additional information on the object is required. At present there exist two kinds of methods, widely used in practice for solving the phase problem. "The direct methods of phase determination" are used to determine the structure of small

molecules (100-150 atoms). The main additional information for these methods is "atomicity" of the object. When it goes to the biological molecules , a method of isomorphous replacement can be applied. This method uses as an additional information on the object (native protein in our case) data of additional X-ray experiments with close substances (isomorphous derivatives) differing from the native protein by local additions ("heavy" atoms) not distorting native structure. In this case all the gravity of the phase problem solving removes into the field of biochemistry. Obtaining of isomorphous derivatives is a serious problem that not always can be solved. Even in cases, when derivatives are successfully obtained, isomorphism can take place only approximately since attached heavy atoms can make some distortions in native structure. It leads to errors when calculating phases, and as a result, to complication of the problem of interpretation of the electron density distribution in structure terms (up to full uninterpretability). Analogous problems arise when only one isomorphous derivative is obtained (for unambiguous solution of the phase problem at least two derivatives are required).

All these problems have recently stimulated considerable efforts aimed at searching for additional sources of information on macromolecular structure as well as methods of this information treating to solve two problems:

i) getting a more interpretable synthesis (correction of phases, determined with errors, and, possibly, determination of phase values that were not determined before);

ii) solution of the phase problem for macromolecules in cases, when heavy atom derivatives are absent.

It has been recently revealed (Lunin,1986; Lunin 1988; Luzzati, Mariani & Delacroix,1988; Harrison,1988; Zhang & Main,1990), that histograms (spectra of frequencies of different electron density values), corresponding to protein electron-density distributions have a specific shape. The shape is sensible to errors in structure factor phases and can be an indicator of correctness of phase determination. This preprint contains a review of the investigations connected with elaboration and application of the new source of information on proteins conducted at Research Computing Centre, USSR Academy of Sciences, (Pushchino, Moscow Region, 142292, USSR).

The presented results were obtained by Lunin V.Yu., Urzhumtsev A.G., Vernoslova E.A., Skovoroda T.P., Vernoslov S.E.,

El'kin Yu.E. The author is grateful to O.B.Ivanova for her help in preparing the manuscript.

2. HISTOGRAM OF THE FINITE RESOLUTION ELECTRON-DENSITY SYNTHESIS IS A NEW SOURCE OF INFORMATION ON THE PROTEIN CRYSTALS.

Trying to specify the phases of structure factors, various researchers made a lot of attempts to use some restrictions on the range of possible values of the electron density distribution function as an additional *a-priori* information on the object under analysis. Examples of such restrictions are:

i) $\rho(r) \geq 0$ (left side restriction on the range of possible values);

ii) $\rho_{min} \leq \rho(r) \leq \rho_{max}$ (both sides restriction on the range of possible values);

iii) $\rho(r) = \{ 0 \text{ or } 1 \}$ (finite set of possible values)

and so on.

2.1 Histogram, corresponding to the electron density distribution.

The base of the below expounded methods is an attempt to take into account not only restrictions on the values, which $\rho(r)$ may take in the unit cell, but also the frequency of each possible value. The most direct practical approach to present this information is as follows. Let us introduce a uniform grid in the unit cell V, and let $\{\rho_j\}_{j=1}^{N}$ be the set of values calculated at the grid points. Let us subdivide the interval (ρ_{min}, ρ_{max}) into K equal parts (bins) and determine how frequent are occurrences of ρ_j in each of the bins

$$\hat{\nu}_k = n_k / N \quad , \quad k = 1,\ldots,K \ .$$

Here n_k is the number of the grid points with values ρ_j, belonging to the K-th bin, that is the number of such ρ_j, that

$$\rho_{min} + (j-1)\frac{\rho_{max} - \rho_{min}}{K} \leq \rho_j \leq \rho_{min} + j\frac{\rho_{max} - \rho_{min}}{K} \ ;$$

N is the total number of the grid points. We call the set (distribution) of frequencies $\{\hat{\nu}_k\}_{k=1}^{K}$ the histogram, corresponding to the function $\rho(r)$.

Sometimes it's more convenient to deal with the normalized histogram

$$\nu_k = n_k / (\Delta_k N) , \quad k = 1,\ldots,K , \qquad (2)$$

where Δ_k denotes the length of k-th bin. In this case the probability of ρ-value to belong to k-th bin (for a random choice of the grid point in the unit-cell V) is $\nu_k \Delta_k$.

Frequencies, calculated from (2) depend strictly speaking not only on $\rho(r)$, but also on the grid and on the way of bins proposing. To get rid of this dependence one can introduce measure on the range of the analyzing function more precisely. Let us define the cumulative function for $\rho(r)$

$$N(t) = \frac{1}{|V|} \text{mes} \{ r : \rho(r) \le t \}$$

and the cumulative function density:

$$\nu(t) = \frac{d}{dt} N(t) = \frac{1}{|V|} \frac{d}{dt} \text{mes}\{r : \rho(r) \le t \}$$

(Here and further $\{r: A\}$ is the part of the unit cell occupied with points r , satisfying condition A ; mes S is the volume of the set S , $|V|$ = mes V is the volume of the unit cell V). Functions N(t) and $\nu(t)$ depend only on $\rho(r)$ but not on the choice of the grid and division of the real axis into bins. The value $\nu(t)\Delta t$ (when Δt is small enough) is a probability, that (for the random choice of the point in the region V) the value $\rho(r)$ belongs to the interval $(\rho,\rho+\Delta t)$. It's easy to see, that normalized frequencies ν_k , calculated from the formula (2) are approximate values of $\nu(t)$ at the points t_k , corresponding to the middles of bins

$$\nu(t_k) = \lim_{\Delta_k \to 0,\ N \to \infty} \nu_k .$$

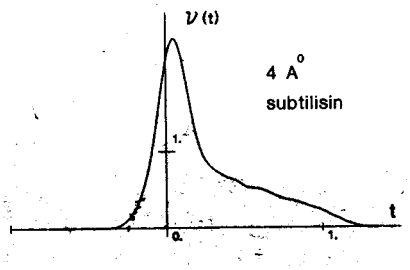
We shall further use a term "histogram" to denote both the sets of frequencies $\{\nu_k\}_{k=1}^K$ or $\{\hat{\nu}_k\}_{k=1}^K$, and the cumulative function density $\nu(t)$. Use of $\nu(t)$ is more convenient when considering theoretical questions while in practice it's more convenient to operate with frequencies.

2.2 Histogram of a finite-resolution Fourier synthesis.

Fig.1 shows the typical histogram, corresponding to the middle-resolution electron density synthesis (4Å in our case) for a protein. Its shape is typical for histograms corresponding to the electron density syntheses for the proteins.

The fundamental histogram property, determining its further application is its sensibility to errors in structure factor phases and lack of some structure factors when calculating

Fig.1 Histogram of the 4Å-resolution electron density synthesis for subtilisin



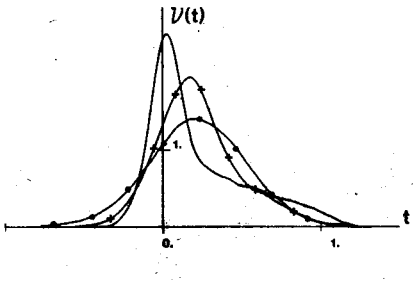
synthesis (Lunin,1986; Lunin,1988). Fig.2 shows how is the histogram shape changed under the influence of two factors: replacement of the exact phase values of structure factors by random ones and elimination of about 18% of reflections near axis 1 of the reciprocal space from the synthesis. Due to its sensitivity to errors, histogram can be expected to serve as an indicator of correctness of phase determining.

Fig.2 Influence of the errors in structure factors on the histogram of the Fourier synthesis

——— exact modules and phases;

–•––•– exact modules, random phases;

–+––+– 18% of reflections are eliminated from the synthesis.



In practice we deal with finite-resolution syntheses. (We call the sum (1) the synthesis at a resolution d_{min} if the sum is composed of all the items, corresponding to the grid points s of the reciprocal space with $|s| \leq 1/d_{min}$). Fig.3 shows how does the change of the the Fourier synthesis resolution influence the histogram shape. This picture results in two important consequences:

i) when speaking of the histogram of the Fourier synthesis we should realize clearly synthesis of what resolution do we

mean;

ii) the full possible histogram information on the electron density distribution is the set of histograms, corresponding to the syntheses of different resolutions.

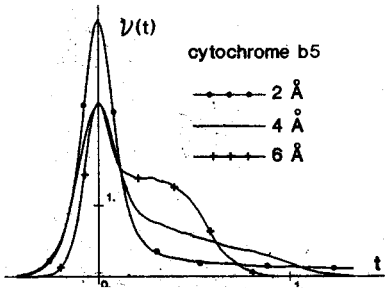


Fig.3. Dependence of the Fourier synthesis histogram on synthesis resolution.

More detailed analysis reveals also the histogram sensitivity to the average value of atoms' temperature factors. (Lunin & Skovoroda,1991).

3. HISTOGRAM PREDICTION FOR PROTEINS WITH UNKNOWN SPATIAL STRUCTURES.

3.1 Empirical Histogram Model.

Analysis of histograms corresponding to syntheses of one and the same resolution for different proteins makes it clear (fig 4)

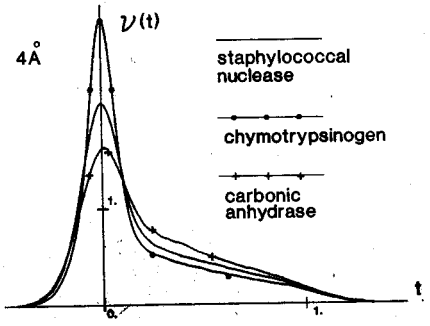


Fig.4. Histograms of Fourier syntheses for different proteins

that these histograms, though posses similar shapes, don't coincide closely. Thus the histogram, corresponding to a protein with known structure can not be directly used as the standard histogram for the other proteins.

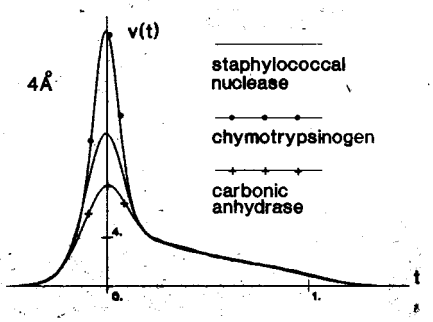


Fig.5. Normalized histograms of Fourier syntheses for the same proteins as at fig.4.

However, the graphs coincide better if we turn to the "normalized volumes", by applying additional renormalization to the histograms :

$$v_k = \nu_k |V| \; / \; F_{000} \; , \qquad k=1,\ldots,K,$$

$$v(t) = \nu(t) \; |V| \; / \; F_{000} \; .$$

Here F_{000} is the number of electrons in the unit cell, $|V|$ is the unit cell volume. (For small Δt the value $v(t)\,\Delta t$ is the volume of the part of the unit cell (due one electron) where the values of $\rho(r)$ lie in the interval $(t,t+\Delta t)$. As fig.5 shows, the normalized histograms have the similar plots, corresponding to middle, large and smallest values of ρ . (It should be stressed, that since we analyze the finite resolution syntheses, we inevitably have points in the unit cell with negative values of $\rho(r)$ and, moreover, these points are concentrated in the molecular region, in general (Urzhumtsev, Lunin & Luzyanina, 1989)).

Fig.5 makes it possible to suggest a hypothesis that for such ρ-values, which can be found in the molecular region only, "normalized volume" is the same for all the proteins and can be described by standard (the same for all the proteins) distribution $v^o(t)$. The distribution $v^o(t)$ varies with resolution of the synthesis. One can't unambiguously determine the values of $v^o(t)$ for t , close to zero directly from the graphs in fig.5. Such the values are met with not only in the

molecular region. However, more precise analysis of the graphs $v(t)$ for different proteins allows to determine values of the standard distribution $v^o(t)$ for all t.

Analysis of functions $v(t)$, corresponding to the crystals of different proteins, allows to notice that the central peak heights at $v(t)$ graphs are directly proportional to corresponding $|V|/F_{ooo}$ values, that is the "freer" the protein molecules are placed in the crystal cell, the higher peak is. This observation allows to suggest a hypothesis that the volume of the region in intermolecular space in which values of $\rho(r)$ belong to an interval $(t,t+\Delta t)$, is directly proportional to the whole volume of the intermolecular space.

Above formulated hypotheses result in the following empirical model of the normalized values distribution

$$v(\rho) = v^o(\rho) + (|V|/F_{ooo} - \int_{-\infty}^{\infty} v^o(x)\, dx)\, q^o(\rho), \qquad (3)$$

or, that's the same, in the histogram model

$$\nu(\rho) = \frac{F_{ooo}}{|V|}\, v^o(\rho) + (1 - \frac{F_{ooo}}{|V|} \int_{-\infty}^{\infty} v^o(x)\, dx)\, q^o(\rho) . \qquad (4)$$

Here $v^o(\rho)$ is the same for all the proteins function, describing distribution of $\rho(r)$-values inside the molecule region; and $q^o(\rho)$ is the same for all the proteins function, describing distribution of $\rho(r)$-values in the intermolecular region.

The discrete analogs of the formulae (3)-(4) are the expressions

$$v_k = v_k^o + (|V|/F_{ooo} - \sum_{j=1}^{K} v_j^o\, \Delta_j)\, q_k^o \qquad (5)$$

$$\nu_k = \frac{F_{ooo}}{|V|}\, v_k^o + (1 - \frac{F_{ooo}}{|V|} \sum_{j=1}^{K} v_j^o\, \Delta_j)\, q_k^o . \qquad (6)$$

Here functions $v^o(t)$ and $q^o(t)$ are replaced by the sets of their values $v_k^o = v^o(t_k)$ and $q_k^o = q^o(t_k)$ at the bin's middles t_k.

3.2 The calculation of the standard distributions.

To determine the standard distributions $v^o(t)$ and $q^o(t)$ a set of proteins ("base protein set") with known atomic structures was selected from among Protein Data Bank (table 1). Atomic models were used to calculate protein's structure factors.

Table 1.

protein	accuracy of histogram prediction Q_h d = 4Å	d = 10Å
base protein set		
Carbonic anhydrase	0.022	0.098
Chymotrypsinogen	0.033	0.051
Cytochrome b5	0.053	0.106
HIPIP	0.028	0.135
B-J Protein	0.023	0.074
Insulin	0.024	0.408
Lysozyme	0.025	0.064
Mioglobin	0.056	0.039
Neurotoxin	0.088	0.212
Ovomucoid	0.030	0.145
Phospholipase	0.019	0.091
Plastocyanin	0.032	0.090
Prealbumin	0.020	0.063
Proteinase A	0.023	0.130
Ribonuclease	0.038	0.162
Staphilococcal nuclease	0.044	0.066
proteins, not getting into the set		
Ubiquitin	0.045	0.176
Crambin	0.044	0.076
Avian pancreatic polypeptide	0.051	0.128
Rubredoxine	0.076	0.170
Concanavalin	0.064	0.073

$$Q_h = \sum_{k=1}^{K} | \nu_k^o - \nu_k^c | \Delta_k ,$$

ν_k^o – values of frequences for exact synthesis;
ν_k^c – values of frequences calculated from (6).

Then electron density syntheses and corresponding exact histograms were calculated. Standard distributions $\{v_k^o\}_{k=1}^K$ and $\{q_k^o\}_{k=1}^K$ were determined in accordance with the requirement of the best agreement between theoretical histograms, determined from (6) and exact histograms for the base proteins. More precisely, we minimized the value

$$Q = \sum_{J=1}^{J} \sum_{k=1}^{K} \frac{N^{(j)} P_{ooo}^{(j)} \Delta_k}{|V|^{(j)}} \frac{(v_{k,t}^{(j)} - v_k^{(j)})^2}{v_k^{(j)}} \Longrightarrow \min \qquad (7)$$

under additional normalizing conditions

$$\sum_{k=1}^{K} v_k^o \Delta_k = 0 \ , \qquad \sum_{k=1}^{K} q_k^o \Delta_k = 1 \ , \qquad (8)$$
$$\sum_{k=1}^{K} t_k v_k^o \Delta_k = 1 \ , \quad \sum_{k=1}^{K} t_k q_k^o \Delta_k = 0 \ .$$

The Lagrange multipliers method was used for these purposes. (Here $F_{ooo}^{(j)}$, and $|V|^{(j)}$ are the number of electrons and the volume of the unit cell, $N^{(j)}$ is the number of the grid points, $\{v_k^{(j)}\}_k$ are theoretical values, calculated from the formula (5), $\{v_k^{(j)}\}_{k=1}^{K}$ are exact values for j-th basic protein. Weight multipliers in (7) represent passing over from values v_k to the numbers n_k of the grid points, whose ρ-values belong to the k-th bin.

Fig.6 shows the graphs of the standard distributions $v^o(t)$ and $q^o(t)$, corresponding to the resolution 4Å.

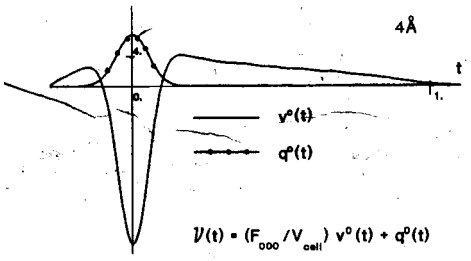


Fig.6. Standard distributions $v^o(t)$ and $q^o(t)$.

3.3 The histogram prediction.

After the standard distributions $\{v_k^o\}_{k=1}^K$ and $\{q_k^o\}_{k=1}^K$ (for the given resolution d_{min}) have been determined one can predict a histogram (for the same resolution d_{min}) for arbitrary protein if parameters V and F_{ooo} of the protein are known. Formulae (4) or (6) do this prediction.

Fig.7. shows exact and "theoretical" (calculated from $\{v_k^o\}_{k=1}^K$ and $\{q_k^o\}_{k=1}^K$) histograms for protein mioglobin (the worst agreement among basic proteins). Fig.8 shows actual and predicted histograms for protein concanavalin (not included into the basic set). Table.1 points out the quality of the agreement between the "theoretical" and the exact histograms at resolutions 4Å and 10Å.

Fig.7. The exact (———) and predicted (-•---•-) histograms of the Fourier synthesis for mioglobin at the 4Å resolution.

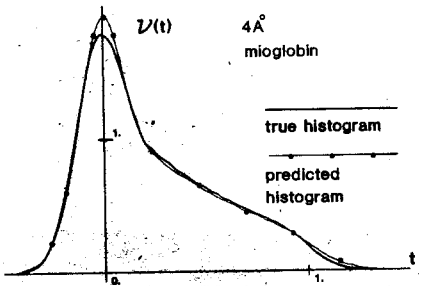
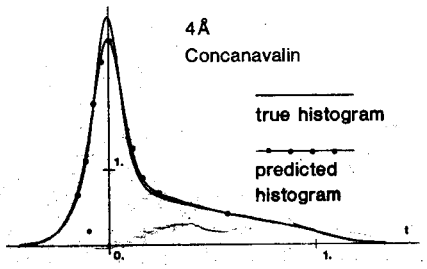


Fig.8. The exact (———) and predicted (-•---•-) histograms of the Fourier synthesis for concanavalin (not included into the base protein set) at the 4Å resolution.



3.4 The histogram prediction for low-resolution syntheses.

The above procedure of histogram prediction gives an acceptable for practice accuracy and appears to be available when dealing with the middle-resolution and high-resolution syntheses. However, when passing to the low-resolution syntheses, accuracy of the prediction decrease. This fact compelled us (Lunin,1988) (when working with the low-resolution synthesis) to use some other procedure, based on the application of either the atomic model of a homologous protein or the model, composed of the parts of different protein models and similar to the model of the analyzing protein in dimensions and outlines (such an information can be received from electron microscopy, for example). In this case the problem of histogram prediction can be solved in two stages:

i) "placing" of the homologous atomic model without

overlapping in the protein unit cell;

ii) calculation of the histogram for this hypothetical model (that is calculation of the structure factors from the atomic model, then, calculation of the Fourier synthesis at the necessary resolution and the histogram for it).

Tests have shown that histograms, calculated in such a way depend weakly on the possible changes of the model packing and can be used in practice.

4. USE OF THE HISTOGRAMS FOR STRUCTURE FACTORS RETRIEVAL.

To obtain an accurate image of molecule at the finite-resolution synthesis (1), we should use exact values of all the structure factors of desired resolution. In practice there is always something that prevents this, either due to unknown phases or even to unknown phases and modules for some of the structure factors. An impression can arise that exclusion of some hundreds (or even tens) of items from the sum (1) can't distort the synthesis (taking into account that the total number of items in (1) tends to thousands and tens of thousands). But it's not so. "Systematic" exclusion of even a small number of reflections from calculation of the synthesis can essentially worse its interpretability. An example of such a phenomenon is shown in fig.9, where about 18% of items (concentrated along the crystallographic axis 1) were excluded from synthesis calculation. Another typical example of incompleteness of the structure factor set is the absence of low-angle reflections. But it is just these reflections that are "responsible" for outlines of the molecule.

4.1 Statement of the problem of structure factors retrieval.

Assume, we are faced with the problem of calculation the Fourier synthesis at a finite resolution

$$\rho(r) = \frac{1}{|V|} \sum_{|s| \le s_{max}} F(s)\, e^{i\varphi(s)}\, e^{-2\pi i(s,r)} , \qquad (9)$$

but a part of the necessary values of structure factors are unknown. Let S_d be a set of indexes s, corresponding to the structure factors with known modules F_s^o and phases φ_s^o, and S_u be a set of indexes s, corresponding to the structure factors with either module or phase unknown. To calculate the synthesis (9) one should attach them (i.e. unknown structure

factors) certain values. The most common way, namely to exclude these reflections from the synthesis (that is to make the corresponding structure factors zero) can result in marked distortions of the synthesis.

We can try to select the values, we are going to attach to unknown structure factors more well-grounded. It appears to be possible if we dispose of some additional information on the properties which the synthesis $\rho(r)$,(i.e. the synthesis we'd like to obtain) should posses. In this case we can determine unknown structure factors so that $\rho(r)$ may meet this additional requirements in full measure.

Assume, we know the histogram $\{\nu_k^o\}_{k=1}^K$,the synthesis (9), calculated with proper values of all the structure factors possesses. We shall call it the standard histogram. Then for each trial set of unknown structure factors we can examine how does it agree with this histogram, by making the following chain of calculations:

i) introduce a grid in the unit cell and calculate the trial synthesis values at the grid points

$$\rho_j^c = \rho^c(r_j) = \frac{1}{|V|} \sum_{s \in S_d} F^o(s)\, e^{i\varphi^o(s)}\, e^{-2\pi i(s,r_j)} + \tag{9a}$$

$$+ \frac{1}{|V|} \sum_{s \in S_u} F^c(s)\, e^{i\varphi^c(s)}\, e^{-2\pi i(s,r_j)} ;$$

ii) calculate the histogram $\{\nu_k^c\}_{k=1}^K$, corresponding to th obtained synthesis;

iii) compare how close are the standard histogram and a calculated one, for example, by using the criterion of histogram closeness of the following form

$$Q(\rho^c) = \frac{1}{K} \sum_{k=1}^{K} \frac{(\nu_k^c - \nu_k^o)^2}{\nu_k^o} . \tag{10}$$

It's reasonably to think that those trial phase set has a best agreement with the standard histogram $\{\nu_k^o\}_{k=1}^K$, for which the value (10) is minimal. So the problem of determination of unknown structure factors can be formulated (Lunin,1986; 1988) as one of minimization of function (10). (Values of frequencies ν_k^c depend on values ρ_j^c of trial synthesis, which in their turn are determined by values F_s^c and φ_s^c of trial structure factors).

Naturally, all other kinds of additional information on the

object (e.g. noncrystallographic symmetry, information on the region taken by disordered solvent and so on) can also be used in work. The most general approach in these cases is minimization of the compound criterion, where each of the items is "responsible" for realizing of one of the additional requirements.

4.2 Quasihistograms.

Minimization of the criterion (10) is a hard computational problem , because no methods, based on the information on derivatives can be applied. The matter is in the fact that for "small variations" of F_s and φ_s , values ρ_j^c (calculated from (9^a)), though change a few, remain inside the same bins as before. So, values of frequencies ν_k^c remain unchanged for small variations of trial values of structure factors and all the derivatives of the criterion (10) are equal to zero. That is why in practice a few different criterion of quality of structure factor trial set was applied.

The frequencies $\{\nu_k^c\}_{k=1}^K$, calculated in accordance with formula (2) can also be determined by the formula:

$$\nu_k^c = \frac{1}{N} \sum_{j=1}^{N} \lambda^*(t_k - \rho_j) \quad ,$$

where

$$\lambda^*(t) = \begin{cases} 1/\Delta & \text{for } |t| \le \Delta/2 \ , \\ 0 & \text{for } |t| > \Delta/2 \ , \end{cases}$$

Δ is the length of the bins, t_k are the middles of the bins.

"Bad" properties of the criterion (10) result from the fact that the values of ν_k^c are calculated by means of piecewise-constant function $\lambda^*(t)$.

Introduce (Lunin,1988)

DEFINITION: Let $\lambda(t)$ be an arbitrary function such that:

$$\int_{-\infty}^{\infty} \lambda(\tau)\, d\tau = 1 \quad .$$

By **quasifrequencies** (connected with the function $\lambda(t)$) we mean the values, calculated from the formula:

$$\tilde{\nu}_k = \frac{1}{N} \sum_{j=1}^{N} \lambda\,(t_k - \rho_j) \quad ,$$

A set of quasifrequencies $\{\tilde{\nu}_k\}_{k=1}^K$ we call a quasihistogram.

If the function $\lambda(t)$ continuously differentiable (or, at least, piecewise-continuously), quasifrequencies depend smoothly

on the values of modules and phases of structure factors, that were used for calculating $\rho(r)$. This makes it possible to use a more convenient criterion of the trial phase set quality of the following form.

Assume, we know the standard quasihistogram $\{\tilde{\nu}_k^o\}_{k=1}^K$, corresponding to the function $\rho(r)$ to be found. We determine a criterion of quality for trial set of unknown structure factors $\{ F^o(s)\exp(i\varphi(s)) \}_{s\in S_u}$ in the form :

$$\tilde{Q} = \frac{1}{K} \sum_{k=1}^{K} \frac{(\tilde{\nu}_k^c - \tilde{\nu}_k^o)^2}{\tilde{\nu}_k^o} . \qquad (11)$$

Now the problem of determination of unknown structure factor values can be formulated as the problem of minimization of the criterion (11). To minimize this criterion a special program, realizing algorithms of steepest descent, fast Fourier transform, and fast differentiation was written. To calculate quasifrequencies the piecewise-linear functions of the following form:

$$\lambda_{æ}(t) = \begin{cases} -(1/æ^2)\ |t| + 1/æ & \text{for } |t| \le æ , \\ 0 & \text{for } |t| > æ . \end{cases} \qquad (12)$$

were used.

The main idea of quasifrequencies introducing is that we don't refer a contribution of a grid point to a single bin any more, but distribute it over some neighboring bins. In this case values of contributions are sensible to the ρ_j changing and thereby make quasifrequencies sensible to small variations of trial structure factors. It's shown, that if standard frequencies are known, the standard quasifrequencies can be calculated from the formula:

$$\tilde{\nu}_k = \frac{1}{N} \sum_{j=1}^{N} \lambda\ (t_k - \rho_j) \approx \int_{-\infty}^{\infty} \lambda(t-\tau)\ \nu(\tau)\ d\tau \ . \qquad (13)$$

Formula (13) shows also that turning to quasifrequencies leads to a certain smoothing of the starting frequencies distribution. The main idea of $\tilde{Q}$ changing relative to the criterion (10) is passing over from comparison of histograms of trial and actual syntheses to comparison of some average histogram characteristics.

4.3 Test retrieval of structure factors for subtilisin.

To check-up the efficiency of the above approach a set of tests was conducted (Lunin 1986; 1988). The test object was an atomic model of subtilisin (Wright, Alden & Kraut,1969), placed in a 73x64x48 Å unit cell in space group $P2_12_12_1$. Atomic coordinates were used to calculate the values of structure factors and the synthesis at a resolution 4Å (Fig.9[a]). The synthesis was used to calculate the quasifrequencies (an interval (-0.5, 1.5) was divided into 30 bins and function $\lambda(t)$ of form (I2) with $\varkappa=5$ was applied).

After that, a situation of lack of information on some of structure factor modules was simulated. About 18% of structure factors (352 out of 2104) were declared to be unknown and an attempt to determine them by minimizing (11) was made. The set of lacking reflections S_u was obtained during one of the real X-ray experiments (the corresponding reflections were not obtained on the technical ground). The most part of elements of S_u were concentrated near the axis 1 of the reciprocal space.

Further tests were conducted in two modifications. In one of them it was assumed that we know only values of structure factors $F^o(s)\exp(i\varphi^o(s))$ for the set S_d and a standard quasihistogram $\{\tilde{\nu}_k^o\}_{k=1}^K$. The problem was to restore both modules and phases of unknown structure factors. In the second modification it was assumed that for $s\in S_u$ only phases are unknown, but modules are known and the problem was to restore unknown phase values.

The 1-st test was devoted to an attempt at restoring both phases and modules of structure factors with $s\in S_u$. The starting values for these structure factors were equal to zero. One of the sections of the starting synthesis (i.e. synthesis constructed from incomplete set of reflections) is shown at Fig.9[b]. As a result of 10 cycles of minimization the value of the criterion (11) has dropped from $0.3\text{x}10^{-2}$ to $0.5\text{x}10^{-5}$. Unknown phase values were determined with an average error 37^o. Value of R-factor for restored values of the structure factors modules was 0.46. Fig.9[c] shows a section of synthesis, constructed with restored values of unknown structure factors. There is a marked progress as compared with the starting picture.

In the second test it was assumed that the modules of structure factors for the set S_u were known and the problem was to restore the phase values. The starting phase values (when minimizing criterion (11)) were the results of the 1-st test.

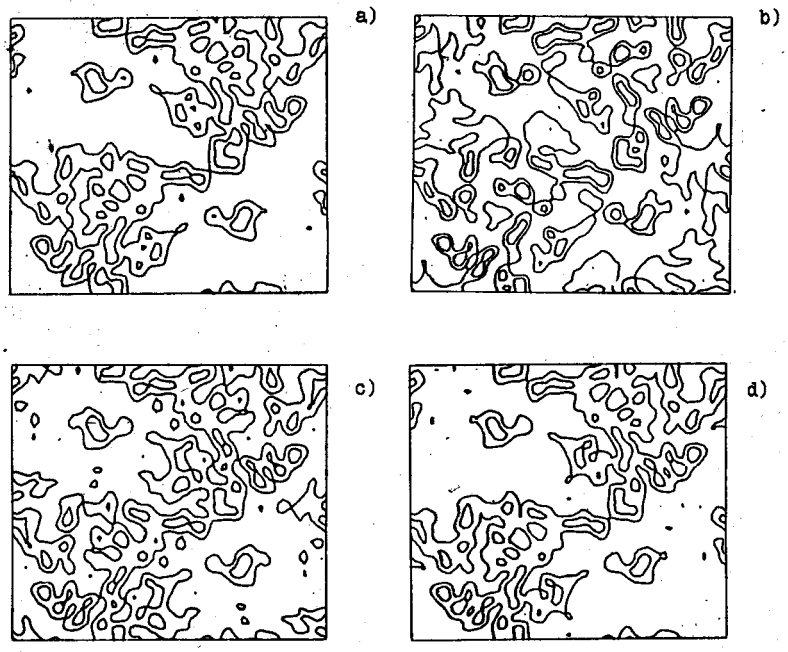


Fig.9. Sections z = 12/48 of syntheses at a resolution 4Å calculated from subtilisin model : a) exact synthesis; b) starting synthesis (about 18% of reflections are eliminated from the synthesis); c) restored modules and phases of eliminated reflections; d) restored phases (modules were known) of eliminated reflections.

Five cycles of minimization resulted in the value of minimizing criterion 0.6×10^{-6} and average value of phase error 33^{o} . Fig.9[d] shows a section of synthesis that was calculated by using restored values of structure factor phases (values of modules were exact).

4.4 Determination of lost structure factors for the "dry" form of γ-crystalline IIIb.

The above procedure of structure factors retrieval was applied to analyze the dry form of γ-crystalline IIIb. The

structure of protein γ-crystalline IIIb from calf's eye lens has been investigated at the laboratory of doctor Chirgadze at the Protein Research Institute USSR Academy of Sciences and at the laboratory of professor T.Blandell in Birkbeck-college (England). The protein crystals belong to the space group $P2_12_12_1$ with the unit cell parameters 58.7 x 69.5 x 116.9 Å .The structure of γ-crystalline was refined at a resolution 2.5 Å (Chirgadze et.al.,1986). Another diffraction set of "dried" protein was collected at a resolution up to 1.9 Å (Chirgadze et.al.,1989). It has a smaller unit cell: 57.38 x 70.13 x 115.4 Å. For the different modifications the discrepancy in the data over 2.5 Å area appeared to be:

$$R = 2\,\frac{\Sigma_s |\,F_{wet} - F_{dry}\,|}{\Sigma_s |\,F_{wet} + F_{dry}\,|} = 0.255 \quad .$$

On the technical ground the initial "dry" set lacked of considerably many reflections (the area up to 4 Å showed only 2834 out of 4224 possible reflections). An attempt at restoring some of the lost data, by making use of the information on the histogram of electron density synthesis was made.

We started with a synthesis at a resolution up to 4 Å, constructed from 2852 reflections with the coefficients

$$F_{dry}(s)\,\exp[\,i\varphi_{wet}(s)\,]\ , \qquad (14)$$

where $F_{dry}(s)$ are modules of the structure factors of the second ("dry") modification, $\varphi_{wet}(s)$ are phases calculated from the refined atomic model of the first modification. Fig.10 shows some sections of the synthesis calculated from these values of the structure factors. It should be stressed that quality of synthesis depends not only on the lack of some necessary reflections, but also on a certain errors in phases, because they corresponded to the 1-st modifications, but not to the second one.

Then we tried to determine the lacking structure factors (and phases, and modules) from the condition of minimum (11). Standard frequencies were determined from the procedure, expounded in section 2. The values of standard quasifrequencies in expression (11) were calculated from the values of frequencies in accordance with the formula (13). Some sections of the synthesis calculated with added restored structure factors are shown in Fig.10.

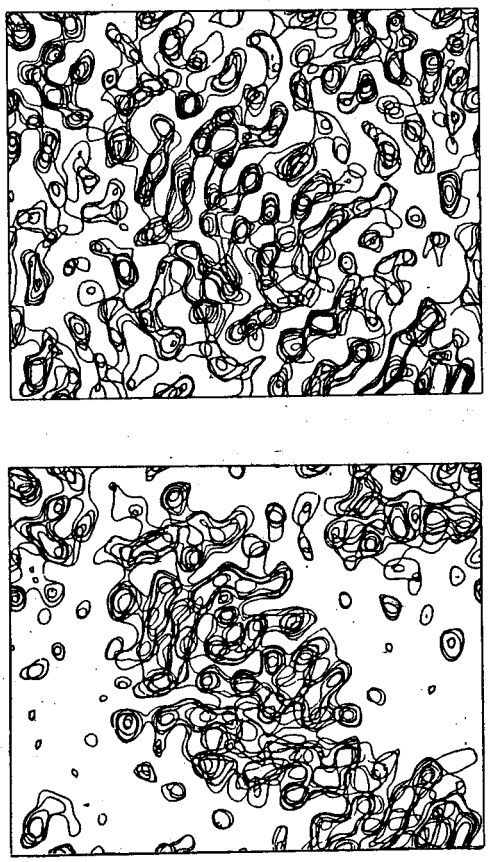


Fig.10. Syntheses at the resolution 4Å for γ-crystallin IIIb : a) the synthesis with the coefficients (14), (some reflections are absent); b) restored values of structure factors are included into the synthesis.

5. USE OF THE HISTOGRAMS IN THE PROBLEM OF PHASE REFINEMENT.

In the paper (Lunin & Vernoslova,1991) we made an attempt to find out which is the place, the procedures of phase refinement, based on the data, obtained from the histogram of the Fourier

synthesis, take as compared with other methods of phase refinement. The analysis was based on the three main characteristics of the phase refinement method:

i) what is the additional information on the object, our method require;

ii) how is it (the additional information) transformed into mathematical form:

iii) what is the calculation procedure we use to determine phases.

5.1 Additional information as presented in the form of the equation $\rho(r)=\tau[\rho](r)$.

In many cases (Lunin,1985) the additional information on the properties of the electron density syntheses can be expressed in mathematical terms as the property of the electron density distribution function to remain unchanged when a certain transform of the function is made

$$\rho(r) = \tau[\rho](r) \quad . \qquad (15)$$

Here τ is a transform (specially picked up), depending on what is the additional information we use. Thus, for example, presence of the local (noncrystallographic) symmetry is equivalent to the equation (15) with the transform $\tau(\rho)$, making by the electron density averaging in symmetrically connected points. The other kinds of additional information (Sayre equations, nonnegativeness of $\rho(r)$, known molecular boundaries, finite set of the function values and so on) can be presented in the analogous way.

5.2 Iterative approach to phase determining by using the equation $\rho=\tau[\rho]$.

Equation (15) is equivalent to the following system of equations for structure factors of the function $\rho(r)$:

$$F(s) = | \mathcal{F}_s\{ \tau[\frac{1}{|V|} \sum_u F(u)\, e^{i\varphi(u)}\, e^{-2\pi i(u,r)}] \}| \qquad (16)$$

$$\varphi(s) = \arg\{ \mathcal{F}_s\{ \tau[\frac{1}{|V|} \sum_u F(u)\, e^{i\varphi(u)}\, e^{-2\pi i(u,r)}] \}\} \qquad (17)$$

Here $\mathcal{F}_s\{ v \}$ is the s-indexed structure factor corresponding to the function $v(r)$, $|z|$ is the module, and $\arg\{z\}$ is the phase of a complex number z.

Considering modules of structure factors $\{F(s)\}_s$ to be

known from the X-ray experiment, we can regard the system of equations (16)-(17) as one, available for determining or refinement of phases $\{\varphi(s)\}_s$.

The great number of methods of phase refinement are based on the iterative procedure of solving the phase part of those equations by the simple iterations method (Lunin,1985). The radial part is simply ignored. Many works on phase refinement are based (evidently or not) on this iterative procedure and vary only in forms of $\tau[\rho]$ transformation.

It is shown in the paper (Lunin & Vernoslova, 1991), that the property of the electron density distribution to have a prescribed histogram can also be presented in the form (15). In this case, transformation $\tau[\rho]$ is made in two stages. First, a modifying function $\lambda_\rho(t)$ (own for each of the possible functions $\rho(r)$) is constructed as a solution of the equation:

$$N^{ex}(\lambda_\rho) = N_\rho(t) \ , \qquad (18)$$

Here N^{ex} and N_ρ are cumulative functions, corresponding to the exact Fourier synthesis and the trial one $\rho(r)$ respectively. Then the modification realizes:

$$\rho(r) \dashrightarrow \rho^m(r) = \lambda_\rho(\rho(r)) = \tau[\rho](r). \qquad (19)$$

It's shown that solution of equations (17) (corresponding to this transformation) by the method of simple iterations is a base of recently suggested methods of phase refinement by applying histograms. They are histogram specification (Harrison,1989) and histogram matching (Zhang & Main,1990).

It's also shown in the paper (Lunin & Vernoslova,1991), that in the case when phases of structure factors contain some errors, transform (18)-(19), restoring proper histogram is realized with modifying function $\lambda_\rho(t)$, which is very close to the function $3\rho^2-2\rho^3$, widely used in electron density modifications. It means that the "classical" method of electron density modification can be regarded as one, using (in a hidden form) specificity of the histogram, corresponding to proper Fourier synthesis.

6. DIRECT LOW-RESOLUTION PHASING.

In this section a new approach to direct solution of the phase problem for low angle reflections is proposed (Lunin, Urzhumtsev & Vernoslova,1990). The approach uses the histogram, corresponding to the electron density synthesis, as an indicator

of correctness of phase values.

The procedure can be divided into three stages. First, one generates a large number of phase sets and select those variants, whose electron density synthesis histograms are close to the prescribed standard. Then the set of the admissible variants is studied by the cluster analysis methods. Inside the set one picks out a subset grouped about the supposed solution of the phase problem. At the third stage one averages the phase sets inside the picked up subset (i.e. cluster) to "extract" some possible phase problem solutions.

The application of the above procedure can be illustrated by the following test example.

6.1 Model structure.

For test purposes we simulated a dimer built from two atomic models of carboxypeptidase and located in a 76x106x116 Å unit cell in a space group $P2_12_12_1$. The test forestalled the work with the elongation factor G (Chirgadze et.al.,1983), therefore the parameters of the unit cell of the elongation factor G were taken to construct a dimer model with the equivalent molecular weight. The dimer model was used to calculate the structure factors, with modules, simulated empirically obtained values $\{F^{ex}(s)\}$, whereas phases were used only to check the answer. The test consisted in determination of the phases of 29 low-angle reflections at a resolution 30 Å. The histogram $\{\nu_k^{ex}\}_{k=1}^K$, corresponding to the synthesis calculated with the exact values of modules $\{F^{ex}(s)\}$ and phases $\{\varphi^{ex}(s)\}$ was considered to be known.

The generated phase sets were analyzed in accordance with two criterions:

i) a criterion of histogram closeness, indicating how close are the histogram of synthesis, calculated with generated phases and the standard histogram $\{\nu_k^{ex}\}_{k=1}^K$;

ii) a criterion of Fourier synthesis closeness, indicating how close are the values of generated phases to correct ones.

The value, characterizing difference between histograms was:

$$Q_h = Q_h(\{\nu_k^c\}, \{\nu_k^{ex}\}) = \sum_{k=1}^{K} | \nu_k^c - \nu_k^{ex} | \Delta_k . \qquad (20)$$

We call it the distance between histograms $\{\nu_k^c\}$ and $\{\nu_k^{ex}\}$. Of course, other measures of histogram closeness may be introduced, such as (10), for example, and so on. Our tests have not revealed

any serious advantage of one over the other.

The aim of solving the phase problem is to produce an interpretable synthesis. Equal phase errors in weak and strong reflections result in very different synthesis defects. This is especially appreciable when the synthesis is calculated with a small number of structure factors. That is why, when comparing phase sets, one should take into account whether these phases correspond to weak or strong reflections. Examples of weighted criteria of the phase set closeness are: the correlation coefficient

$$\hat{C}(\rho^c,\rho^{ex}) = \sum_s F^2(s)\cos(\varphi^c(s)-\varphi^{ex}(s)) \;/\; \sum_s F^2(s)$$

(its maximum value is 1 at $\rho^c = \rho^{ex}$, minimum is -1 at $\rho^c = -\rho^{ex}$, and the mean value is 0), and the criterion of closeness between syntheses

$$\hat{Q}_s(\rho^c,\rho^{ex}) = \{\int_V [\rho^c(r)-\rho^{ex}(r)]^2 dV_r \;/\; \int_V [\rho^{ex}(r)]^2 dV_r\}^{1/2} =$$

$$= (2 - 2\hat{C})^{1/2}$$

(its minimum value is 0 at $\rho^c = \rho^{ex}$, maximum is 2 at $\rho^c = -\rho^{ex}$, and the mean value is $\sqrt{2}$).

When solving the phase problem *ab initio*, we should bear in mind that the phase sets should be reduced to the same origin before comparison. The matter is in the fact that all the functions of the form

$$\rho^c_{t,æ}(r) = \rho^c(æ\, r + t) ,$$

(where t is an arbitrary vector, $æ = \pm 1$) will have the same set of structure factor modules and the same histogram $\{\nu^c_k\}$. Therefore to compare the two Fourier syntheses ρ^c and ρ^{ex}, we should shift ρ^c into the coordinate system where it is as close to ρ^{ex} as possible, (and, possibly, turn to enantiomorph). We define the "crystallographic" distance between ρ^c and ρ^{ex} (or, equivalently, the weighted crystallographic distance between the phase sets $\{\varphi^c(s)\}$ and $\{\varphi^{ex}(s)\}$) to be:

$$Q_s = \min_{t\in T} \min_{æ=\pm 1} \hat{Q}_s(\rho^c_{t,æ}, \rho^{ex}) \qquad (21)$$

(here T is a set of possible shifts of the origin). If $\rho^{ex}(r)$ has a symmetry group, distinct from P1, the set T of possible shifts may consist of a finite number of variants. For example, for the group $P2_12_12_1$ we should check 16 variants of origin

and enantiomorph to calculate Q_s.

6.2 The first stage. Selection of the admissible variants.

Table 3 shows distribution of the values Q_h and Q_s for 400 000 phase sets (29 reflections in each set) generated by randomizer (requirements of symmetry of the group $P2_12_12_1$ were naturally taken into account). First of all we can see from the table that phase sets, resulted in the histograms, closest to the standard one ($Q_h < 0.10$) include variants both close to the exact set and very far from it ($Q_s \sim 1.0$). This means, particularly, that a good histogram does not guarantee the correct synthesis.

A more thorough analysis of Table 3 allows to notice that the variants with good histograms are divided into two groups: one consists of the variants with $Q_s \sim 0.5$, and the other consists of the variants with $Q_s \sim 1.0$. As requirements to the quality of histograms lower (Q_h increase), a number of variants increase, the deviation of values Q_s from the mean grow inside the groups and the groups coalesce. Such a picture allows to infer, that there should exist at least two different phase sets resulting (when modules of structure factors are prescribed) in a prescribed histogram.

Table 3. The distribution of the values Q_h and Q_s, for trial phase sets (there given the number of sets for which the values Q_h and Q_s belong to corresponding intervals; the values Q_h and Q_s are calculated from the formulae (20) and (21) relatively).

Q_h \ Q_s	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2
0.075	1	0	0	0	0	0	0	0	0	0
0.100	1	7	14	6	4	0	5	27	10	0
0.125	5	29	86	103	42	0	44	305	119	0
0.150	1	70	284	359	160	3	138	1290	689	2
0.175	4	84	531	846	428	24	451	3384	2146	9
0.200	1	93	701	1523	872	110	848	6301	4556	71
0.225	1	52	727	2139	1680	269	1502	9732	7571	191
0.250	0	32	685	2812	2852	569	2123	12326	10144	312
0.275	0	27	610	3265	4398	1220	2746	13828	11593	508
0.300	0	6	486	3594	5114	2235	2876	13752	11107	553
0.325	0	2	311	3645	7310	3298	2740	12483	9511	477
0.350	0	2	186	3186	8633	4763	2385	10359	7000	303
0.375	0	0	78	2504	8842	5995	1929	7784	4781	149
0.400	0	0	52	1723	8476	7495	1465	5157	2887	49
0.425	0	0	26	1145	7424	8358	1383	2950	1716	26
0.450	0	0	8	651	6212	8566	1520	1492	891	10
0.475	0	0	3	377	4527	8204	2001	658	453	1

6.3 The second stage. The cluster analysis of the set of admissible variants.

Since in reality the standard histogram is predicted with a certain error, all the phase sets, for which Q_s is not too large, should be considered as not contradicting to the standard histogram. In our tests we can regard as admissible, for example, 39 variants, ensuring value $Q_h(\rho_j^c, \rho^{ex}) < 0.1$. In reality the exact phase values are unknown, so the values of the criterion of closeness $Q_s(\rho_j^c, \rho^{ex})$ between the generated phases and the exact ones can not be calculated. However, one can calculate the intervariant distances between admissible phase sets $Q_s(\rho_j^c, \rho_k^c)$. The procedure of cluster analysis allows (basing on the analysis of matrix of intervariant distances between the admissible variants) to make a notion of how are these variants distributed in many-dimensional "configurational" space: whether they form one (or a few) compact groups or dissipate evenly all over the space. The procedure of cluster analysis consists in joining together close variants (those with $Q_s(\rho_j^c, \rho_k^c) < \varepsilon$). It is clear that, if ε increases, the number of variants in each of the clusters increases, but the number of clusters decreases. Fig.11 illustrates process of cluster organization (the order in which the variants are shown in Fig.11 is chosen proceed from the simplicity of tree representation; it is, of course, not the order in which they were generated). The analysis was made by using the program P1M of the software package BMDP (Dixon,1977).

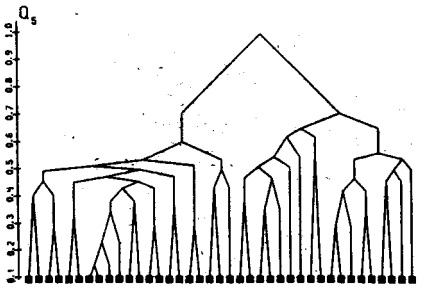


Fig.11 Organization of admissible variants into clusters in the test with model protein.

Fig.11 shows that 39 admissible variants are separated into two clusters: cluster A consisting of 21 variants and cluster B consisting of 18 variants. It appeared so (table 4), that

cluster A included variants at distances Q_s (from the exact solution) varied from 0.23 to 0.66 , and cluster B included variants at distances $Q_s > 0.89$. It should be emphasized that this division was made with the use of the intervariant distances matrix only and took no account of how far the variants actually were from the exact solution.

6.4 The third stage. Averaging of variants inside the cluster.

To choose the phase set that is the solution of the phase problem, the "centre of gravity" was chosen for each of the clusters. More precisely, the figure of merit $m(s)$ and the "best" phase $\varphi^{best}(s)$ for every reflection were determined in each of the clusters in accordance with the formula:

$$m(s)\, e^{i\varphi^{best}(s)} = \frac{1}{M} \sum_{j=1}^{M} e^{i\varphi_j(s)} \quad .$$

Here M is the number of variants in the cluster (it is 21 for cluster A) and $\varphi_j(s)$ is the value of the s-indexed phase in the j-th phase set. Naturally, all the phase sets were reduced to the same coordinate system and enantiomorph before averaging. For this purpose, one of the cluster variants was taken as the frame of reference and the others were shifted to those coordinate systems that ensured a minimal possible distance $\hat{Q}_s$ from the frame of reference. The synthesis $\rho_A(r)$ was calculated with modules $\{F^{ex}(s)\}$ and phases $\{\varphi^{best}(s)\}$, obtained in such a way. Analogous procedure of averaging was applied to 18 variants of cluster B and the synthesis $\rho_B(r)$ was calculated in the same way. Fig.12 shows the maps of electron density distribution in one of the sections of the unit cell, corresponding to syntheses $\rho_A(r)$ and $\rho_B(r)$ as well as to the synthesis $\rho^{ex}(r)$, calculated with exact phase values.

Table 4 lists mean values of the figures of merit and the phase errors for the phases $\{\varphi^{best}(s)\}$, made by averaging in clusters A and B . One can see from this table, that cluster, corresponding to the true solution (cluster A), has a larger mean figure of merit and a smaller dissipation of variants about the mean variant, than cluster B , corresponding to the false solution of the phase problem.

6.5 Test direct phasing for cytochrome b5 .

The object of the next testing was cytochrome b5

Table 4. Characteristics of clusters, picked out during the test with dimer model structure.

	Cluster A	Cluster B
Number of variants in cluster	21	18
Distances Q_s between cluster elements and exact phase problem solution		
min	0.23	0.89
max	0.66	1.12
average	0.45	0.97
$< m >_s$	0.52	0.41
$< Q_s(\rho^{best}, \rho_j) >_j$	0.42	0.54
$Q_s(\rho^{best}, \rho^{ex})$	0.34	0.95
$C(\rho^{best}, \rho^{ex})$	0.94	0.55
$< \lvert\varphi^{best} - \varphi^{ex}\rvert >_s$ (deg.)	40	71

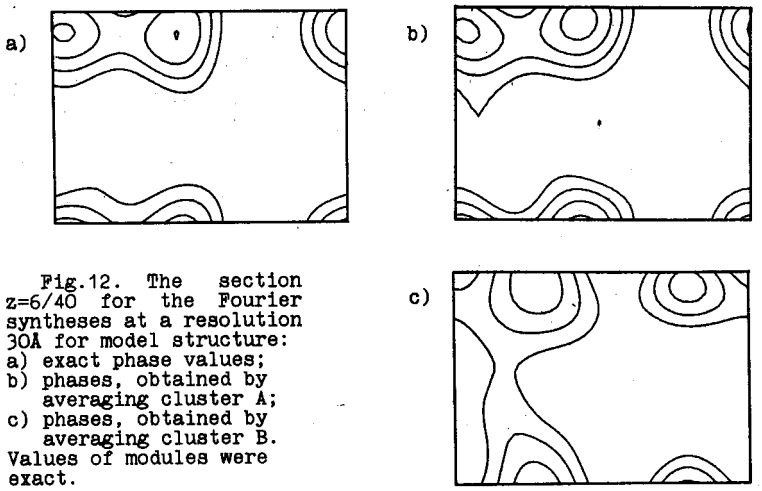


Fig.12. The section z=6/40 for the Fourier syntheses at a resolution 30Å for model structure:
a) exact phase values;
b) phases, obtained by averaging cluster A;
c) phases, obtained by averaging cluster B.
Values of modules were exact.

(Mathews,Levine & Argos,1971), 65X46X30 Å unit cell, $P2_12_12_1$ space group. Atomic coordinates, taken from the Protein Data Bank, were used to calculate exact values of structure factors; the structure factors were used to calculate the Fourier synthesis at the resolution 13.6Å (29 reflections) and to construct the standard histogram. Then the following problem was

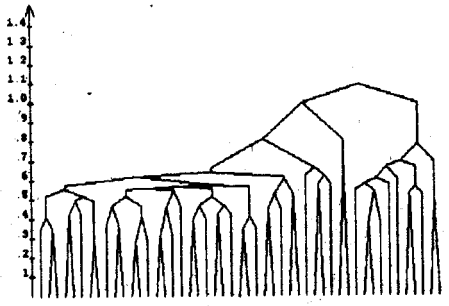
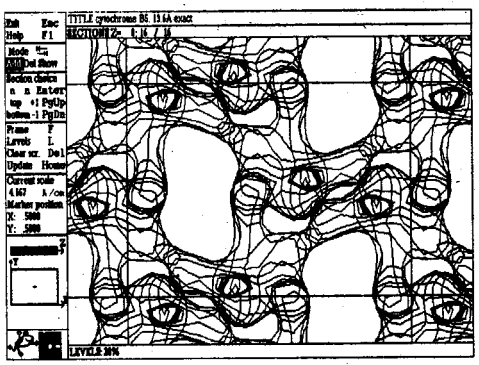


Fig.13. Organization of variants into clusters in test with cytochrome.



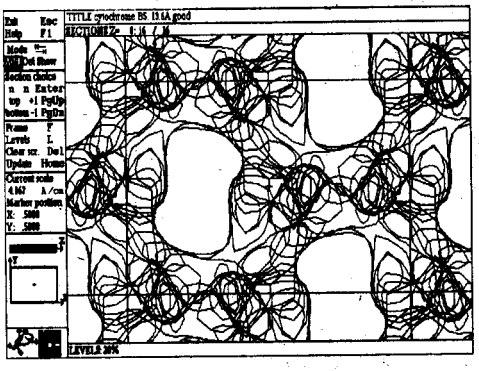


Fig.14. Exact Fourier synthesis for cytochrome and synthesis, calculated with the phases determined *ab initio* The projection along z-axis is shown.

formulated: to determine phase values of structure factors of the range 13.6 Å , by using for this purpose only modules of structure factors and the standard histogram.

For this purpose we generated 500 000 random phase sets and separated 49 of them resulting in the histogram close to the standard one ($Q_h < 0.1$). Fig 13 illustrates the process of cluster formation. The best of them contains solution of the phase problem with average phase error 32° and the correlation coefficient C = 0.92 . Fig.14 shows the maps of the electron density distribution for exact synthesis and synthesis with phases, determined by means of the above method.

6.6 Test direct phasing for Bence-Jones protein.

This protein crystallizes in space group $P2_12_12$ in a 55x52x43 Å unit cell (Furey et al.,1979). For the protein we determined phases of 25 low-angle reflections (at a resolution 16 Å). The process of testing was analogous to one with cytochrome. 100 000 random phase sets were generated and 488 variants were separated for further analysis. Fig 15 shows the process of cluster formation. Singling out a cluster and averaging its elements allowed to get a solution with the average phase error 41° and a correlation coefficient C = 0.93 . Fig 16 shows the maps of electron density distribution, calculated with phases, determined by the suggested method.

So, the results of the tests show that the following procedure:

i) generating random phase sets and selecting those with synthesis histogram close to the prescribed one;

ii) organizing the chosen variants into clusters on the basis of the matrix of intervariant distance Q_s;

iii) averaging the variants inside every cluster

leads to a small number of possible solutions, including solution sufficiently close to the true one.

6.7 Phase determination for elongation factor G at the resolution 30 Å.

The spatial structure of the elongation factor G from *Thermus Thermophilus* are investigated under the leadership of Yu.N. Chirgadze at the protein Research Institute and the Research Computing Centre in Puschino. The protein is

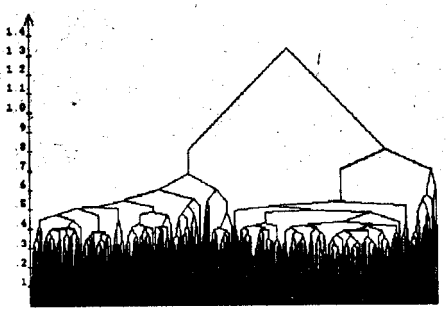


Fig.15. Organization of variants into clusters in the test with the Bence-Jones protein.

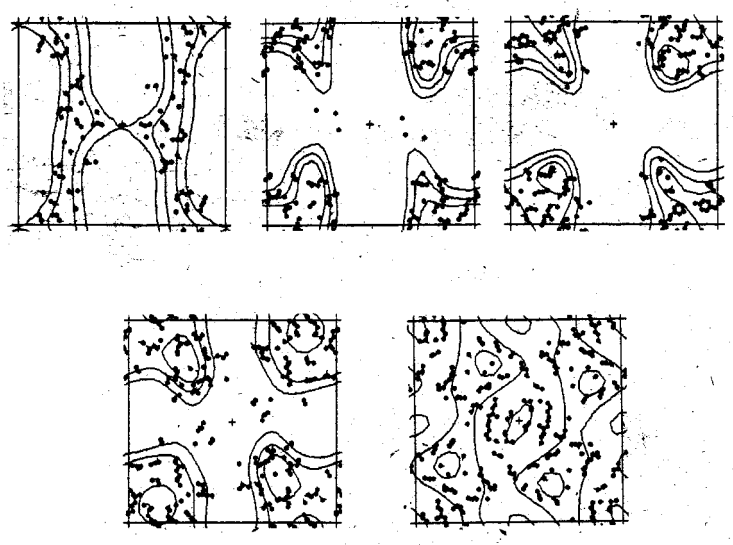


Fig.16. Sections of the Fourier synthesis for Bence-Jones protein, calculated with phases, determined *ab initio*. There are shown atom's positions. (Levels, separating 10, 30 and 50% of the unit volume are shown).

crystallized in space group $P2_12_12_1$ in a 76x106x116 Å unit cell (Chirgadze et.al.,1983).

To predict a histogram for the elongation factor G an approach, suggested in section 3.4 was applied. The dimer model mentioned above may be located in the unit cell differently. We chose three variants of model packing and calculated histograms for the corresponding 30 Å syntheses. These histograms were sufficiently close to one another (distances Q_h between the histograms did not exceed 0.1). The phase sets were generated in four ways : with the use one from the three histograms and with the use of the histogram, averaged over the three versions. All four variants gave similar results. Below we give a brief description of the work with the averaged histogram.

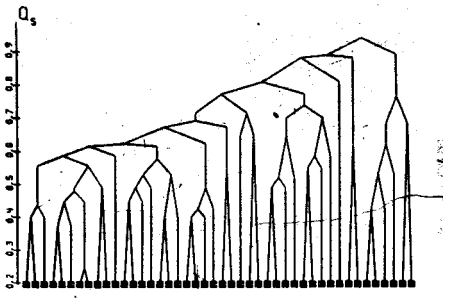


Fig.17. Organisation of variants into clusters for the elongation factor G.

We generated 500 000 random phase sets and separated 44 of them resulting in best correlation with simulated histogram (Q_h < 0.125). Fig.17. illustrates the process of cluster formation. Averaging variants in the selected cluster we have got phases with a mean figure of merit of 0.54 . The deviation of variants in the cluster from the mean value was Q_s = 0.46 .

Fig.18 shows a unit cell projection along the crystallographic axis x . The obtained synthesis agrees with the results, obtained by other methods.

7. COMPUTING PROBLEMS

7.1 The fast differentiation algorithm.

Many of the above approaches come in the end to minimization of a complicate criterion of a trial phase set quality. Notice,

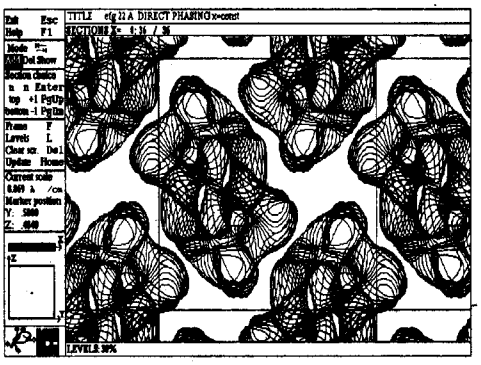


Fig.18. Fourier synthesis for the elongation factor G at a resolution 30Å calculated with phases, determined *ab initio*. The projection of the unit cell along the axis x is shown

that calculation of each of the criterion values requires much CPU time. But the most serious problem, connected with criterion minimization is calculation of the minimizing function's gradient. (We calculate it, because it's just gradient that determines the direction of variables shifting to diminish the aim function value). Thus, if we use different formulae to calculate the derivatives, we have to spend n times more CPU time than for one function value calculation. Taking into account, that calculation of one criterion value requires some minutes and the number of variables is large enough (e.g. it may reach tens of thousands when dealing with refinement of atomic structure), one can consider the problem to be dissoluble. However, it was proved (Kim,Nesterov & Cherkassky,1984), that for any function, possessing arbitrary number of variables, one can construct an algorithm, requiring equal time to calculate all the components of the gradient and one value of the function f(x). This fact is of great methodological importance for problems of XSA. It follows that, when locally refining the object's structure, one can use any property of the object, available for calculation with the computer, as a criterion of model's correctness. Realization of the general idea in reference to the problems connected with determination of spatial structures by XSA methods is expounded in (Lunin & Urzhumtsev,1985; Lunin,1985). In particular, the created algorithms allowed to realize the approaches, we suggested in the previous sections.

The crystals of analyzing substance possess nontrivial symmetry, in general. So, one can essentially economize CPU time, using this symmetry. The symmetry can be accounted in process of fast calculation of minimizing function gradient.

REFERENCES

Chirgadze Yu.N., Nikonov S.V., Brazhnikov E.V., Garber M.B. & Reshetnikova L.S. "Crystallographic study of elongation factor G from Thermus thermophilus HB8", J.Mol.Biol., 1983, 168, 449-450.

Chirgadze Yu.N., Nevskaya N.A., Fomenkova N.P., Nikonov S.V., Sergeev Yu.V., Brazhnikov E.V., Garber M.B., Lunin V.Yu., Urzhumtsev A.G. & Vernoslova E.A. "The structure of γ-crystallin III b from calf lens at 2.5 Å resolution" , Docl.Acad.Nauk.SSSR, 1986, 290, 492-495 (Russian).

Chirgadze Yu.N., Nevskaya N.A., Vernoslova E.A., Urzhumtsev A.G., Lindley P. & Bibby M. "Structure refinement of "dry" crystal form of calf eye lens γ-crystallin IIIb at 1.9 Å resolution", Twelfth European Crystallographic Meeting, 1989, Moscow, Collected Abstracts, vol.2, 363.

Dixon W.J., editor "Biomedical Computer Programs, P-Series", 1977.

Harrison R.W. "Histogram Specification as a Method of Density Modification", J.Appl.Cryst., 1988, 21, 949-952.

Furey W.J., Wang B.C., Yoo C.S., Sax M. "Phase Extension and Refinement of Bence-Jones Protein RHE (1.9Å)", Acta Cryst., 1979, A35, 810-817.

Kim K.V., Nesterov Yu.E. & Cherkassky B.V. "The estimate of the cost of gradient computation", Docl.Acad.Nauk SSSR, 1984, 275, 1306-1309 (Russian).

Lunin V.Yu. "Use of the Fast Differentiation Algorithm for Phase Refinement in Protein Crystallography", 1985, Acta Cryst., A41, 551-556.

Lunin V.Yu. "Use of the Information on Electron Density Distribution in Proteins", Preprint, 1986, Pushchino, USSR (Russian).

Lunin V.Yu. "Use of the Information on Electron Density Distribution in Macromolecules", 1988, Acta Cryst., A44, 144-150.

Lunin V.Yu "The lost structure factors retrieval at X-ray study of macromolecules structures", Docl.Acad.Nauk SSSR, 299, 363-366 (Russian).

Lunin V.Yu., Vernoslova E.A. "Frequencies-Restrained Structure Factor Refinement. II. Comparison of Methods", Acta Cryst., 1991, in press.

Lunin V.Yu., Skovoroda T.P. "Frequencies-Restrained Structure Factor Refinement. I. Histogram simulation ", 1991, A Cryst., A47, 45-52.

Lunin V.Yu., Urzhumtsev A.G. "Program Construction for Macromolecule Atomic Model Refinement Based on the Fast Fourier Transform and Fast Differentiation Algorithms", 1985, Acta Cryst., A41, 327-333.

Lunin V.Yu., Urzhumtsev A.G., Skovoroda T.P. "Direct Low-Resolution Phasing from Electron-Density Histograms in Protein Crystallography", Acta Cryst., 1990, A46, 540-544.

Luzzati V., Mariani P. & Delacroix H. "X-ray crystallography at macromolecular resolution : a solution of the phase problem", Macromol.Chem.,Macromol.Symp., 1988, 15, 1-17.

Main P. "The use of Sayre's Equation with Constraints for the Direct Determination of Phases", Acta Cryst., 1990, A46, 372-377.

Main P. "A Formula for Electron Density Histograms for Equal-Atom Structures", Acta Cryst., 1990, A46, 507-509.

Mariani P., Luzzati V. & Delacroix H. "Cubic Phases of Lipid-containing Systems. Structure Analysis and Biological Implications", J.Mol.Biol., 1988, 204, 165-189.

Mathews F.S., Levine M., Argos P. "The structure of calf liver cytochrome b5 at 2.8Å resolution", 1971, Nature New Biol., v.233, 15-16.

Podjarny A.D. & Yonath A. "Use of Matrix Direct Methods for Low-Resolution Phase Extension for tRNA", Acta Cryst., A33, 655-661.

Urzhumtsev A.G., Lunin V.Yu., Luzyanina T.B. "Bounding a Molecule in a Noisy Synthesis", 1989, Acta Cryst., A45, 34-39.

Wright C.S., Alden R.A. & Kraut J. "Structure of subtilisin BPN' at 2.5Å resolution", Nature, 1969, 221, 235-242.

Zhang K.Y.J. & Main P. "Histogram Matching as a New Density Modification Technique for Phase Refinement and Extension of Protein Molecules", Acta Cryst., 1990, A46, 41-46.

Zhang K.Y.J. & Main P. "The use of Sayre Equation with Solvent Flattening and Histogram Matching for Phase Extension and Refinement of Protein Structures", Acta Cryst., 1990, A46, 377-381.

CONTENTS

4.04.91 г. Тир.150 Зак.3269Р Уч.-изд.л. 2,3 Изд.№ 90
Отпечатано с оригинала-макета на ротапринте в ОНТИ ПНЦ АН СССР