Eric Fanchon for helpful discussions. This work was supported in part by a grant (GM-34102) from the US National Institute of Health.

References

BLOW, D. M. & CRICK, F. H. C. (1959). *Acta Cryst.* **12**, 794-802.
BLUNDELL, T. L. & JOHNSON, L. N. (1976). *Protein Crystallography.* New York: Academic Press.
GRAVES, B. J., HATADA, M. H., HENDRICKSON, W. A., MILLER, J. K., MADISON, V. S. & SATOW, Y. (1990). *Biochemistry*, **29**, 2679-2684.
GUSS, J. M., MERRITT, E. A., PHIZACKERLEY, R. P., HEDMAN, B., MURATA, M., HODGSON, K. O. & FREEMAN, H. C. (1988). *Science* **241**, 806-811.
HATADA, M. H., MILLER, J. K., GRAVES, B. J., HENDRICKSON, W. A. & SATOW, Y. (1989). Am. Crystallogr. Assoc. Abstr. Ser. 2, Vol. 17, 89.
HENDRICKSON, W. A. (1979). *Acta Cryst.* A**35**, 245-247.
HENDRICKSON, W. A. (1985). *Trans. Am. Crystallogr. Assoc.* **21**, 11-21.
HENDRICKSON, W. A., HORTON, J. R. & LEMASTER, D. M. (1990). *EMBO J.* In the press.
HENDRICKSON, W. A. & LATTMAN, E. E. (1970). *Acta Cryst.* B**26**, 136-143.
HENDRICKSON, W. A., LOVE, W. E. & KARLE, J. (1973). *J. Mol. Biol.* **74**, 331-361.
HENDRICKSON, W. A., PÄHLER, A., SMITH, J. L., SATOW, Y., MERRITT, E. A. & PHIZACKERLEY, R. P. (1989). *Proc. Natl Acad. Sci. USA*, **86**, 2190-2194.
HENDRICKSON, W. A., SMITH, J. L., PHIZACKERLEY, R. P. & MERRITT, E. A. (1988). *Proteins*, **4**, 77-88.
HORTON, J. R. & HENDRICKSON, W. A. (1989). Proc. Am. Crystallogr. Assoc. Meet. Abstr. Ser. 2, Vol. 17, p. 122.
KARLE, J. (1980). *Int. J. Quant. Chem. Symp.* **7**, 357-367.
KARLE, J. (1989). *Acta Cryst.* A**45**, 303-307.
MURTHY, H. M. K., HENDRICKSON, W. A., ORME-JOHNSON, W. H., MERRITT, E. A. & PHIZACKERLEY, R. P. (1988). *J. Biol. Chem.* **263**, 18430-18436.
OGATA, C. M., HENDRICKSON, W. A., GAO, X., PATEL, D. J. & SATOW, Y. (1989). Proc. Am. Crystallogr. Assoc. Meet. Abstr. Ser. 2, Vol. 17, p. 53.
YANG, W., HENDRICKSON, W. A. & CROUCH, R. J. (1989). Proc. Am. Crystallogr. Assoc. Meet. Abstr. Ser. 2, Vol. 17, p. 114.

Acta Cryst. (1990). A**46**, 540-544

Direct Low-Resolution Phasing from Electron-Density Histograms in Protein Crystallography

BY V. YU. LUNIN, A. G. URZHUMTSEV AND T. P. SKOVORODA

Research Computing Centre, USSR Academy of Sciences, Pushchino, Moscow Region 142292, *USSR*

(*Received* 12 *February* 1990; *accepted* 6 *March* 1990)

Abstract

An approach to direct phasing of low-resolution reflections is proposed. It is based on the generation of a large number of phase sets and selection of those variants whose electron-density-synthesis histograms are close to a prescribed standard. Classifying them into clusters and averaging them inside every cluster restricts their number to one to three usually, in which a phase set close to the standard is contained. The best variant can be recognized by the properties of its cluster. Test phasing of 29 low-resolution reflections has resulted in a correlation coefficient of 0·94 and a mean phase difference of 40° compared with the true phases.

1. Introduction

In previous years histograms corresponding to finite-resolution electron-density syntheses were shown to be a useful tool in macromolecular structure-factor determination (Lunin, 1986, 1988; Lunin & Skovoroda, 1990; Luzzati, Mariani & Delacroix, 1988; Mariani, Luzzati & Delacroix, 1988) and refinement (Harrison, 1988; Zhang & Main, 1990). Some methods of histogram prediction were suggested for proteins with unknown spatial structure. In this paper we discuss how histograms may be used to phase low-resolution reflections directly.

The idea of the approach is very simple on the face of it. One generates many (*e.g.* random) trial phase sets and separates those which lead to histograms close to the predicted one. It would be reasonable to expect that if the number of generated phase sets is large enough, one necessarily finds a variant close to the true one, which can be identified by a 'good' histogram of the corresponding synthesis. The actual situation is much more complicated, and there may exist several different phase sets leading to histograms close to the prescribed one. Since these histograms can always possess errors, we should consider all such variants as admissible.

Cluster-analysis methods permit a more thorough study of the set of admissible variants. These are classified into subsets grouped about different solutions of the phase problem, which then are averaged to 'extract' some (two or three) possible phase-problem solutions. In our tests one of these extracted variants was found to be sufficiently close to the true

0108-7673/90/070540-05$03.00

solution. Furthermore, it was possible to identify the cluster corresponding to the true solution because of its compactness.

Test phasing of 29 low-resolution reflections for a model structure of two molecules of carboxypeptidase A (Rees, Lewis & Lipscomb, 1983) resulted in a correlation coefficient of 0·94 and a mean phase difference of 40° compared with the true phases. The developed procedure was successfully used to phase 30 Å resolution reflections for an elongation factor G (Chirgadze, Nikonov, Brazhnikov, Garber & Reshetnikova, 1983).

A similar approach was applied by Luzzati, Mariani & Delacroix (1988; Mariani *et al.*, 1988) in their investigations of ordered phases of lipid–water systems. They checked all possible phase sets (for the centrosymmetric case) and chose the one resulting in the best histogram. However, since the possible number of sets grows exponentially with the number of reflections, we had to proceed in another direction, which required less computation when searching for the variants. Also, because of possible errors in the predicted histogram we refused to seek an absolute minimum in the histogram discrepancy and introduced averaging in every cluster.

2. Does a good histogram guarantee the correct structure-factor phases?

2.1. *Formulation of the problem and notation*

Let

$$\rho^{ex}(\mathbf{r}) = |V|^{-1} \sum_{|\mathbf{s}| \le s_{max}} F^{ex}(\mathbf{s}) \exp[i\varphi^{ex}(\mathbf{s})] \times \exp[-2\pi i(\mathbf{s}, \mathbf{r})] \quad (1)$$

and $\{F^{ex}(\mathbf{s})\}$, $\{\varphi^{ex}(\mathbf{s})\}$ be the sets of structure-factor amplitudes and phases. We assume that all the amplitudes are known and are used to calculate the syntheses below.

Let a function $\rho(\mathbf{r})$ be calculated at N grid points of a unit cell. Assume that the interval (ρ_{min}, ρ_{max}) of possible ρ values is divided into K parts (bins). We determine the frequency that the values of $\rho(\mathbf{r})$ occur in the bins to be

$$\nu_k = n_k/N, \qquad k = 1, 2, \ldots, K,$$

where n_k is the number of grid points with ρ values belonging to bin k. The set $\{\nu_k\}$ of these frequencies is called here a histogram. Different approaches have been used to predict histograms of protein electron-density syntheses (Lunin, 1988; Lunin & Skovoroda, 1990; Zhang & Main, 1990). In this paper we suppose that the histogram $\{\nu_k\}$ corresponding to the synthesis $\rho^{ex}(\mathbf{r})$ is known and call it the 'standard'.

The problem we are studying in this paper consists in finding structure-factor phases provided the amplitudes $\{F^{ex}(\mathbf{s})\}$ and standard histogram $\{\nu_k^{ex}\}$ at a resolution of $d_{min} = 1/s_{max}$ are known.

We say that some trial phase set $\{\varphi^c(\mathbf{s})\}$ results in the histogram $\{\nu_k^c\}$, by which the histogram corresponding to the synthesis calculated from the amplitudes $\{F^{ex}(\mathbf{s})\}$ and phases $\{\varphi^c(\mathbf{s})\}$ is meant. The first question we try to answer is: can we be sure that the trial phase set $\{\varphi^c(\mathbf{s})\}$ is close to the exact one $\{\varphi^{ex}(\mathbf{s})\}$ if the trial histogram $\{\nu_k^c\}$ is close to the standard? To answer this question, we should define more formally what are close histograms and close phase sets.

2.2. *Criteria of histogram discrepancy*

To describe the discrepancy between two histograms, we use the value

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

We call this the distance between histograms $\{\nu_k^c\}$ and $\{\nu_k^{ex}\}$. [For all syntheses of a given resolution we take the same interval (ρ_{min}, ρ_{max}) and fix its decomposition into bins.] Of course, other measures of histogram closeness may be introduced, such as

$$Q_\chi = \sum_{k=1}^{K} (n_k^c - n_k^{ex})^2/n_k^{ex}.$$

Our tests have not revealed any serious advantage of one over the other.

2.3. *Criteria of phase-set discrepancy*

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 one should take into account not only differences between phases but also the values of the corresponding amplitudes when comparing phase sets. Examples of weighted phase-discrepancy criteria are the correlation coefficient

$$\hat{C}(\rho^c, \rho^{ex}) = \sum_{\mathbf{s}} F^2(\mathbf{s}) \cos[\varphi^c(\mathbf{s}) - \varphi^{ex}(\mathbf{s})] \Big/ \sum_{\mathbf{s}} F^2(\mathbf{s})$$

(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 'distance' between syntheses

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

(its minimum value is 0 at $\rho^c = \rho^{ex}$, maximum is 2 at $\rho^c = -\rho^{ex}$, and the mean value is $2^{1/2}$).

When solving the phase problem *ab initio*, we should keep in mind that phase sets should be reduced

Table 1. *The distribution of the values Q_h and Q_s for test phase sets (the numbers of variants are given)*

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	12 326	10 144	312
0·275	0	27	610	3265	4398	1220	2746	13 828	11 593	508
0·300	0	6	486	3594	5114	2235	2876	13 752	11 107	553
0·325	0	2	311	3645	7310	3298	2740	12 483	9511	477
0·350	0	2	186	3186	8633	4763	2385	10 359	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

to the same origin and enantiomorph before comparison. All functions of the form

$$\rho^c_{\mathbf{t},\kappa} = \rho^c(\kappa\mathbf{r}+\mathbf{t}), \mathbf{t}\in\mathbf{T}, \kappa = \pm 1$$

(with T the set of all vectors in the unit cell) will have the same set of structure-factor amplitudes and the same histograms. Therefore, to compare $\rho^c(\mathbf{r})$ with $\rho^{ex}(\mathbf{r})$ we should shift $\rho^c(\mathbf{r})$ into the coordinate system where it is as close to $\rho^{ex}(\mathbf{r})$ as possible. We define the 'crystallographic' distance between syntheses $\rho^c(\mathbf{r})$ and $\rho^{ex}(\mathbf{r})$ [or, equivalently, the weighted crystallographic distance between phase sets $\{\varphi^c(\mathbf{s})\}$ and $\{\varphi^{ex}(\mathbf{s})\}$] to be

$$Q_s = \min_{\mathbf{t}\in\mathbf{T}} \min_{\kappa=\pm 1} \hat{Q}_s(\rho^c_{\mathbf{t},\kappa}, \rho^{ex}).$$

If $\rho^{ex}(\mathbf{r})$ has a nontrivial symmetry group, the set T of possible shifts may consist of a finite number of vectors. For example, for the group $P2_12_12_1$ we should check 16 variants of origin and enantiomorph to calculate Q_s.

2.4. *Test structure*

For test purposes we used a dimer built from two atomic models of carboxypeptidase A (Rees *et al.*, 1983) and located without self-intersections in a unit cell with $P2_12_12_1$ symmetry. This test was connected with the investigation of the elongation factor G (Chirgadze *et al.*, 1983), therefore the size of the test object and the unit-cell parameters (76 × 106 × 116 Å) were the same as for factor G. The dimer model was used to calculate amplitudes and phases of structure factors. In the test the amplitudes played the role of the known $\{F^{ex}(\mathbf{s})\}$ values, unlike the phases that were used only to check the answer.

2.5. *Connection between histogram and phase-discrepancy factors*

The test consisted in generating a large number (400 000) of random phase sets at a resolution of 30 Å (29 reflections), calculating the values Q_h and Q_s for every set and analysing the variant distribution with respect to these two parameters. Table 1 shows some results of the test. We can see that the phase sets closest to the standard histograms ($Q_h < 0{\cdot}1$) include variants both close to the exact set and very far from it ($Q_s \sim 1{\cdot}0$). So the answer to the question with which § 2 was begun is negative - a good histogram does not guarantee a correct synthesis.

3. Selection of particular phase-problem solutions from a set of admissible variants

3.1. *The cluster analysis*

A more thorough analysis of Table 1 allows classification of all variants with good histograms into two groups: those which are close to the exact phase set ($Q_s \sim 0{\cdot}5$) and those with $Q_s \sim 1{\cdot}0$. When differences Q_h between the calculated and the standard histograms increase, the groups and the deviation of values Q_s from the mean grow, and, when Q_h becomes large enough, the groups coalesce. Such a picture allows us to infer that there should exist at least two different phase sets resulting [together with the prescribed amplitudes $\{F^{ex}(\mathbf{s})\}$] in a good histogram (the variants with $Q_s \sim 1{\cdot}0$ may show an even greater diversity of phase sets). Cluster analysis may give here a clearer picture.

For a more precise analysis we took 39 variants $\rho_1, \rho_2, \ldots, \rho_{39}$ with good histograms ($Q_h < 0{\cdot}1$) and calculated the matrix of distances $Q_s(\rho_j, \rho_k)$ between them. The procedure of cluster analysis is to join together close variants [those with $Q_s(\rho_j, \rho_k) < \varepsilon$ for a given ε]. It is clear that, if ε increases, the number of such clusters decreases but the number of variants in each of them increases. Fig. 1 illustrates the process of cluster organization. (The order in which the variants are shown in Fig. 1 is a simplified tree representation; it is, of course, not the order in which

they were generated.) The analysis was made by the *P1M* routine (Dixon, 1977).

3.2. *Picking out a particular solution*

Fig. 1 reasonably suggests that all 39 variants should be separated into two clusters *A* and *B*, consisting of 21 and 18 variants, respectively. In fact (see Table 2), cluster *A* only included variants close to the true solution (Q_s varied from 0·23 to 0·66) and cluster *B* included phase sets dissimilar to the true one ($Q_s > 0{\cdot}89$). It should be emphasized that this division was made with the use of the intervariant distances matrix $\{Q_s(\rho_j, \rho_k)\}$ only and took no account of how far the variants actually were from the exact solution.

To choose a 'representative member' from a cluster, the variants were averaged in the cluster. For every structure factor we defined the 'best' phase $\varphi^{\text{best}}(\mathbf{s})$ and the figure of merit $m(\mathbf{s})$ to be

$$m(\mathbf{s}) \exp[i\varphi^{\text{best}}(\mathbf{s})] = M^{-1} \sum_{j=1}^{M} \exp[i\varphi_j(\mathbf{s})].$$

Here M is the number of variants in the cluster (it is 21 for cluster *A*) and $\varphi_j(\mathbf{s})$ is the value of the s-indexed phase in the jth phase set. Naturally, all phase sets were reduced to the same coordinate system before averaging. For this purpose, one of the cluster variants was taken as basic, ρ^{bas}, and for others coordinate systems and enantiomorphous modifications were varied so as to produce a minimal possible value of the distance $\hat{Q}_s(\rho^{\text{bas}}, \rho_j)$. The syntheses ρ_A^{best} and ρ_B^{best} were calculated from the exact amplitudes $\{F^{\text{ex}}(\mathbf{s})\}$ and phase sets $\{\varphi_A^{\text{best}}\}$ and $\{\varphi_B^{\text{best}}\}$, resulting from averaging in clusters *A* and *B*.

Fig. 2 shows sections corresponding to syntheses $\rho_A^{\text{best}}(\mathbf{r})$, $\rho_B^{\text{best}}(\mathbf{r})$ and $\rho^{\text{ex}}(\mathbf{r})$. Table 2 lists mean values of figures of merit and phase differences for phase sets made by averaging in clusters *A* and *B*. One can see from this table that cluster *A* (corresponding to the true solution) has a large mean figure of merit

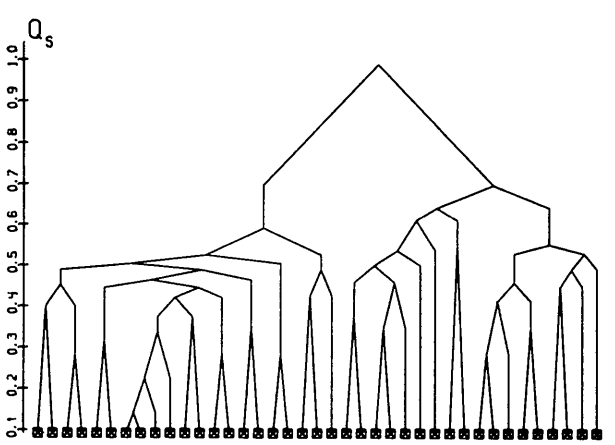

Fig. 1. Organization of admissible variants into clusters in the test.

Table 2. *Characteristics of test clusters*

	Cluster *A*	Cluster *B*
Number of variants	21	18
Distance Q_s of cluster elements from the true phase-problem solution		
min.	0·23	0·89
max.	0·66	1·12
average	0·45	0·97
$\langle m \rangle_s$	0·52	0·41
$\langle Q_s(\rho^{\text{best}}, \rho_j) \rangle_j$	0·42	0·54
$Q_s(\rho^{\text{best}}, \rho^{\text{ex}})$	0·34	0·95
$C(\rho^{\text{best}}, \rho^{\text{ex}})$	0·94	0·55
$\langle \lvert\varphi^{\text{best}} - \varphi^{\text{ex}}\rvert \rangle_s$ (°)	40	71

and a small mean value for the distance between variants in the cluster and the averaged one, compared with cluster *B*.

Our tests with other objects gave similar results. So we may infer that

(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 distance matrix; and

(iii) averaging the variants inside every cluster restricts the number of possible phase-problem solutions. The true solution is sufficiently close to one of the selected variants and can be identified by its cluster's properties.

4. Direct phasing of 30 Å reflections for the elongation factor *G*

The spatial structure of the elongation factor *G* from *Termus termophilus* has been investigated at the Protein Research Institute and the Research Computing Centre in Pushchino (Chirgadze *et al.*, 1983). The

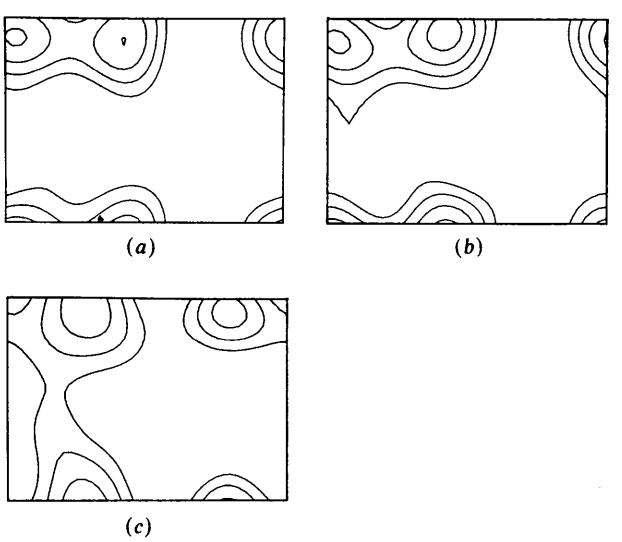

Fig. 2. The section $z = 6/40$ for syntheses: (*a*) ρ^{ex}, (*b*) ρ_A^{best}, (*c*) ρ_B^{best}.

protein crystals belong to space group $P2_12_12_1$ and have unit-cell parameters about $76\times106\times116$ Å.

4.1. *Standard histogram simulation*

The method suggested for histogram prediction by Lunin & Skovoroda (1990) gives acceptable results at medium and high resolution. However, the quality of predicted low-resolution histograms is not always satisfactory. That is why we simulated 30 Å resolution histograms by another method similar to the use of a homologous model (Lunin, 1988).

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). We phased structure factors three times, each time using a new one from the three histograms to separate phase sets. Then we generated phase sets with the use of the histogram averaged over the three versions. All four runs gave similar results. A brief description of the fourth run is given below. A more detailed report on the *G*-factor structure determination will be published separately (Chirgadze, 1990).

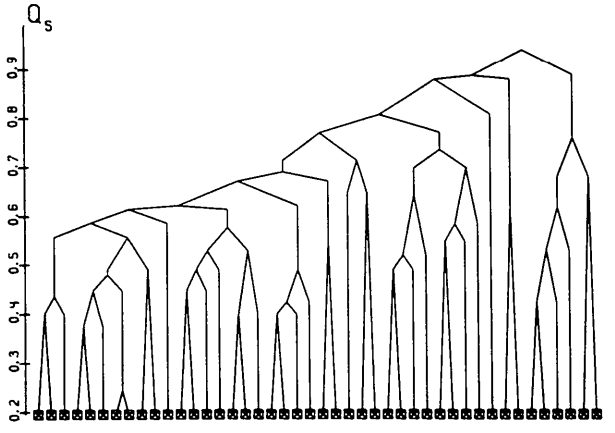

Fig. 3. Organization of admissible variants into clusters for the elongation factor *G*.

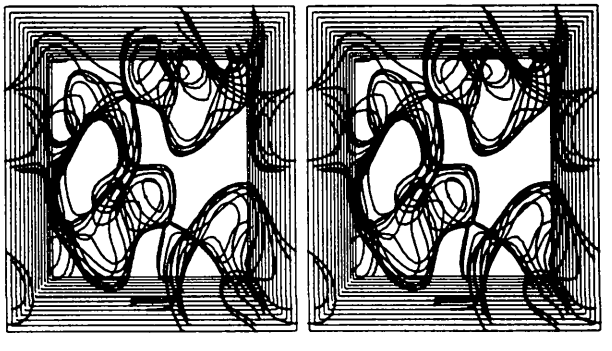

Fig. 4. A fragment of a directly phased 30 Å electron-density synthesis for the elongation factor *G*.

4.2. *Generation of variants and cluster analysis*

In this run we generated 500 000 random phase sets and separated 44 of them resulting in best histograms ($Q_h < 0{\cdot}125$). Fig. 3 illustrates the process of cluster formation. Unlike Fig. 1, here we could separate confidently only one cluster of 18 phase sets. Our attempts to organize other clusters failed because of small numbers of variants and large spread among them. Averaging 18 variants in the cluster obtained resulted in phases with a mean figure of merit of 0·54. The deviation of variants in the cluster from the mean value was $Q_s = 0{\cdot}46$.

Fig. 4 shows a fragment of electron-density synthesis for the elongation factor *G* at a resolution of 30 Å. This synthesis agrees with the results obtained by other methods (Chirgadze, 1990).

The authors are grateful to G. N. Borisyuk for valuable discussions and for her help in using the BMDP package and to O. M. Liguinchenko for her assistance in preparing the manuscript.

References

CHIRGADZE, YU. N. (1990). In preparation.
CHIRGADZE, YU. N., NIKONOV, S. V., BRAZHNIKOV, E. V., GARBER, M. B. & RESHETNIKOVA, L. S. (1983). *J. Mol. Biol.* **168**, 449-450.
DIXON, W. J. (1977). Editor. *Biomedical Computer Programs P-Series.* Univ. of California Press.
HARRISON, R. W. (1988). *J. Appl. Cryst.* **21**, 949-952.
LUNIN, V. YU. (1986). *Use of the Information on Electron Density Distribution in Proteins*, Preprint. Pushchino, USSR.
LUNIN, V. YU. (1988). *Acta Cryst.* A**44**, 144-150.
LUNIN, V. YU. & SKOVORODA, T. P. (1990). Submitted to *Acta Cryst.*
LUZZATI, V., MARIANI, P. & DELACROIX, H. (1988). *Macromol. Chem. Macromol. Symp.* **15**, 1-17.
MARIANI, P., LUZZATI, V. & DELACROIX, H. (1988). *J. Mol. Biol.* **204**, 165-169.
REES, D. C., LEWIS, M. & LIPSCOMB, W. N. (1983). *J. Mol. Biol.* **168**, 367-387.
ZHANG, K. Y. J. & MAIN, P. (1990). *Acta Cryst.* A**46**, 41-46.