Acta Cryst. (1998). D**54**, 726–734

On the *Ab Initio* Solution of the Phase Problem for Macromolecules at Very Low Resolution. II. Generalized Likelihood Based Approach to Cluster Discrimination

V. Y. Lunin,[a] N. L. Lunina,[a] T. E. Petrova,[a] A. G. Urzhumtsev[b] and A. D. Podjarny[b]*

[a]*Institute of Mathematical Problems of Biology, Russian Academy of Sciences, Pushchino, Moscow Region, 142292, Russia, and* [b]*UPR de Biologie Structurale, IGBMC, BP 163, 67404 Illkirch CEDEX, CU de Strasbourg, France. E-mail: podjarny@igbmc.u-strasbg.fr*

(*Received 13 May 1997; accepted 22 September 1997*)

Abstract

The multisolution strategies for direct phasing at very low resolution, such as the few atoms model technique, result in a number of alternative phase sets, each of them arising from a cluster of closely related models. Use of a Monte-Carlo type computer procedure is suggested to choose between the possible phase sets. It consists of generating a large number of pseudo-atom models inside the mask defined by a trial phase set and the use of histograms of magnitude correlation to evaluate the masks. It is shown that the procedure may be considered as a generalization of the statistical maximum-likelihood principle and may be used as a powerful supplementary tool in the likelihood-based approaches to the phase problem solution.

1. Introduction

The problem of *ab initio* phasing for macromolecules and their complexes is in the focus of attention for the last decade. By *ab initio* phasing we mean in this paper the attempts to solve macromolecular structures based on X-ray (or neutron) native diffraction data only, without isomorphous derivatives or anomalous scattering data. It is now possible to separate two main directions to attack this problem. The first one is to use from the very beginning a close-to-atomic-resolution set of structure factors and develop classical direct methods (Woolfson & Yao, 1990; Sheldrick *et al.*, 1993; Weeks *et al.*, 1994). The other way is to start from very low resolution and use some features of macromolecular objects to solve the low-resolution phase problem, followed by density modification and refinement methods to improve resolution (Podjarny *et al.*, 1987; Lunin *et al.*, 1990, 1995; Bricogne & Gilmore, 1990). We are concerned in this paper with the problem of very low resolution phasing only.

We have developed earlier some methods of *ab initio* low-resolution phasing (Lunin *et al.*, 1990, 1995) and showed that they allow reduction of the phase problem to a small number of alternative solutions. We recall this method briefly in §3. Some additional criteria were suggested to reduce this number (Urzhumtsev *et al.*, 1996) for special cases. In §4 we propose a general way to rank these possible solutions and to refine and extend phases. We start from an empirical procedure and present results of tests in §5. In §6 we highlight the connections between this empirical procedure and a maximum-likelihood based choice of the prior, proposed by Bricogne & Gilmore (1990). We discuss in §7 the advantages and disadvantages of both approaches.

2. Definitions

We define in this section some general concepts which will be used afterwards without additional explanation.

2.1. *Control criterion: phase correlation*

When testing phasing methods, we call a control criterion some measure of the quality of the current phase set $\{\varphi_{\mathbf{h}}\}$. In this paper we use mostly one type of control criteria based on the comparison of current phases with the true ones $\{\varphi_{\mathbf{h}}^{ex}\}$, namely the phase correlation coefficient (Carbo *et al.*, 1980; Read & Moult, 1992; Lunin & Woolfson, 1993),

$$\mathrm{PC}(\{\varphi_{\mathbf{h}}\}, \{\varphi_{\mathbf{h}}^{ex}\}) = \sum_{\mathbf{h}} (F_{\mathbf{h}}^{o})^2 \cos(\varphi_{\mathbf{h}}^{ex} - \varphi_{\mathbf{h}}) \Big/ \sum_{\mathbf{h}} (F_{\mathbf{h}}^{o})^2, \quad (1)$$

where $\{F_{\mathbf{h}}^{0}\}$ are experimental magnitudes. This value can be calculated only in test cases, when the true phases are known. It must be noted that the experimental magnitudes themselves do not fix the origin and enantiomorph unambiguously. So in *ab initio* phasing we must either specify some phases fixing origin/enantiomorph or produce the phase alignment (Lunin & Lunina, 1996) before calculating (1). It is supposed further that such an alignment is always performed before the phase comparison.

The phase correlation coefficient is not the only way to compare current phases with the true ones. For example, the mean absolute phase error or some weighted mean phase error may be used for the same purposes. We will therefore call 'phase correlation' any control criterion based on the comparison of trial and exact phase values. An example of a control criterion

which is not a phase correlation is the model trapping value, defined as the number of atoms of the true (refined) model inside the mask region, which is calculated using the trial phases and the observed magnitudes (Lunin & Lunina, 1998).

2.2. *Selection criterion: magnitude correlation*

We call a selection criterion an estimate of a trial phase set quality which can be calculated either on the basis of the experimental magnitudes only or with the use of some prior information. This prior information is of a general type, *e.g.* electron-density histograms (Lunin *et al.*, 1990), connectivity (Wilson & Agard, 1993; Baker *et al.*, 1993; Bystroff *et al.*, 1993), atomicity (Woolfson & Yao, 1990; Weeks *et al.*, 1994), *etc.*, and does not require the knowledge of the true phases. It must be possible to calculate this criterion value for every trial phase set in a real *ab initio* phasing process; the appropriate choice of the criterion is the key point for the success of the method. It should be noted that so far no criterion has been found to choose the correct solution unambiguously.

This paper will consider a particular case when the trial phases are calculated from trial atomic models. In this case model magnitudes may be used to define a selection criterion, *e.g.* as

$$\mathrm{MC}(\{F_{\mathbf{h}}^c\},\{F_{\mathbf{h}}^o\}) = \sum_{\mathbf{h}}(F_{\mathbf{h}}^c - \langle F^c\rangle)(F_{\mathbf{h}}^o - \langle F^o\rangle)\Big/ \left[\sum_{\mathbf{h}}(F_{\mathbf{h}}^c - \langle F^c\rangle)^2 \sum_{\mathbf{h}}(F_{\mathbf{h}}^o - \langle F^o\rangle)^2\right]^{1/2}, \quad (2)$$

or, as

$$\mathrm{MC}(\{F_{\mathbf{h}}^c\},\{F_{\mathbf{h}}^o\}) = \sum_{\mathbf{h}} F_{\mathbf{h}}^c F_{\mathbf{h}}^o \Big/ \left[\sum_{\mathbf{h}}(F_{\mathbf{h}}^c)^2 \sum_{\mathbf{h}}(F_{\mathbf{h}}^o)^2\right]^{1/2}. \quad (3)$$

We will call magnitude correlation (MC) any type of selection criterion based on the comparison of calculated and observed magnitudes. Obviously, other types of magnitude correlation, besides (2) or (3) are possible, *e.g.* the usual R factor. An example of a selection criterion which is not a magnitude correlation is the closeness of the standard electron-density histogram to the one corresponding to the Fourier synthesis calculated with the use of trial phases and the observed magnitudes (Lunin, 1993).

We will say that structure factors calculated from a trial atomic model are ω-correlated if the corresponding magnitude correlation value, MC, is higher then some prescribed level ω. We will call the structure factors well magnitude correlated (WMC) if they are ω-correlated with a high enough value of ω, to be defined in each particular case.

2.3. *Cluster centroid phases*

For a collection (cluster) of phase sets $\{\varphi_{\mathbf{h}}^j\}, j = 1,...,M$ we define the centroid phases $\varphi_{\mathbf{h}}^{\mathrm{best}}$ and the figures of merit $m_{\mathbf{h}}$ as

$$m_{\mathbf{h}} \exp[i\varphi_{\mathbf{h}}^{\mathrm{best}}] = \frac{1}{M}\sum_{j=1}^{M} \exp[i\varphi_{\mathbf{h}}^j]. \quad (4)$$

It is supposed that the best alignment of all the phase sets to a reference set was performed before averaging in (4).

2.4. *Few atoms model (FAM)*

We call a few atoms model a model composed of a relatively small number of equal pseudo-atoms. These atoms are modeled either with a Gaussian shape,

$$\rho(\mathbf{r}) = \left(\frac{4\pi}{B}\right)^{3/2} \exp\left(-\frac{4\pi^2 r^2}{B}\right), \quad (5)$$

(usually with an artificially large B value) or with a constant non-zero density inside a sphere of given radius. We will refer further to the number of atoms in an asymmetric unit as the number atoms in an FAM, supposing that all the symmetry-related atoms are included in the model. The simplest example of an FAM is the single-huge-atom model which contains one pseudo-atom in the asymmetric unit.

2.5. *Test object*

We use in the tests described below the same AspRS–$\mathrm{tRNA}^{\mathrm{Asp}}$ complex data described in a previous paper (Lunin *et al.*, 1995). The cubic crystal form of AspRS–$\mathrm{tRNA}^{\mathrm{Asp}}$ complex is particularly well suited for low-resolution work, because of the large unit cell (space group $I432$, a = 354 Å), the large solvent content (78%), and the compact shape of the complex. The structure was solved by the molecular-replacement method (Urzhumtsev *et al.*, 1994) using X-ray data to 8 Å resolution and a model from another crystal form ($P2_12_12$) solved at 2.9 Å resolution (Ruff *et al.*, 1991). The neutron diffraction low-resolution (∞–20 Å) data set (Moras *et al.*, 1983) was used in tests described below as the experimental magnitudes. These magnitudes can be fitted correctly (MC = 0.92) with the molecular-replacement model. Molecular-replacement model phases were considered in the tests as the exact ones. All the tests were performed at 40 Å resolution (49 reflections).

3. The outlines of FAM approach

3.1. *Specific features of low-resolution phasing*

Two major ideas were originally proposed to find very low resolution (VLR) phases (or molecular region masks) which lead sometimes to unpredictable results.

Table 1. *The result of reciprocal-space refinement of randomly generated 100-atoms models against experimental data at 40 Å resolution (49 reflections)*

Magnitude (MC) and phase (PC) correlations are defined according to (2) and (1), respectively.

	Before refinement		After refinement	
Variant	MC	PC	MC	PC
a	0.50	0.53	1.00	0.54
b	0.50	−0.48	0.99	−0.44
c	0.17	0.04	1.00	0.06
d	0.64	0.24	0.98	0.24
e	0.54	−0.20	1.00	−0.22
f	0.40	−0.04	1.00	−0.05

The first one consists of modeling the molecule at VLR as a sphere ('flat' or 'Gaussian') and obtain the VLR phases from such a single-huge-atom (SHA) model (Podjarny *et al.*, 1987; Harris, 1995; Andersson & Hovmöller, 1996). The molecular volume can usually be estimated, defining the sphere radius [or B value for a Gaussian sphere (5)], and the only problem is to find the position of its centre. To do this we can scan the unit cell and compare for every position the values of magnitudes calculated from SHA-approximation with the observed ones. The best agreement of the two sets of magnitudes will give us the optimal position for the sphere center.

The first part of this proposition is reasonable, *i.e.* low-resolution phases can usually be approximated by SHA ones (Podjarny *et al.*, 1987; Andersson & Hovmöller, 1996), but the best agreement of magnitudes may be obtained for sphere centres quite different from the ones producing good SHA phases. Fig. 1. displays a distribution of sphere centers with respect to two criteria, namely the correlation of SHA-calculated and observed magnitudes (2) and phase correlation coefficient (1). It shows that for our test case the sphere positions resulting in good magnitudes usually give quite bad phases. Similar pictures were obtained for other objects. Special efforts are necessary in such a search (Podjarny *et al.*, 1987) to avoid false solutions, *e.g.* identify these false solutions as being placed in special positions in the unit cell.

The other idea (Subbiah, 1991, 1993) is to take a large number of randomly placed atoms and refine their positions against experimental magnitudes. The refined set of atoms is expected to produce good low-resolution phases. Fig. 2. shows magnitude and phase-correlation values for 100 FAM's consisting of 100 random atoms each. Attempts to refine such models lead to equally well correlated magnitudes, but do not produce any phase improvement (Table 1).

These examples illustrate the two main features of low-resolution phasing, which can be obtained when using other selection criteria also (Lunin *et al.*, 1990, 1995).

The best values of the selection criterion may be coupled with bad control criterion values and *vice versa*.

Local refinement may significantly improve selection criterion values without any improvement of phases.

3.2. *Redefinition of the problem of low-resolution phasing*

These two pessimistic inferences may be partially compensated by the observation (Lunin *et al.*, 1990, 1995) that the phase sets (variants) corresponding to good selection criterion values do not fill a 'configuration space' randomly, but form a small number of compact regions in configuration space, one of which is close enough to the true solution. This may be interpreted as stating that the true solution corresponds to one selection criterion optimum, but not necessarily to the best one. Therefore, we redefine the problem of low-resolution phasing as the task of studying and describing all the main regions in multi-dimensional configuration space corresponding to good selection criterion values.

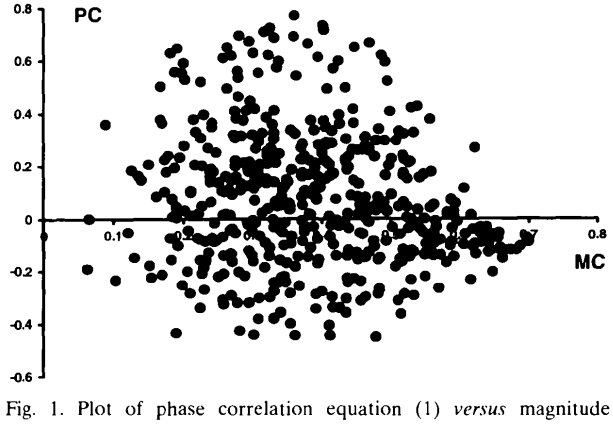

Fig. 1. Plot of phase correlation equation (1) *versus* magnitude correlation (2) for the calculated structure factors of 1000 randomly generated models. Each point represents a single one-atom model.

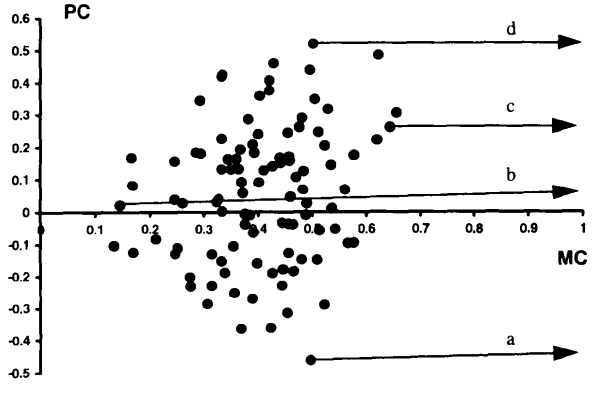

Fig. 2. Plot of phase correlation (1) *versus* magnitude correlation (2) for the calculated structure factors of 100 randomly generated models. Each point represents a single 100 atoms model. The arrows show the result of the reciprocal-space refinement of some models (Table 1).

The multidimensional cluster analysis technique was found to be a useful tool for such an analysis.

3.3. *The few atoms models approach*

Even for low resolution the high number of dimensions of the phase sets configuration space is too large to perform an exhaustive search. To overcome this difficulty a Monte-Carlo type procedure was suggested to study the configuration space (Lunin *et al.*, 1990). The few atoms model (FAM) method allows a further reduction in the dimensionality (Lunin *et al.*, 1995). In this approach we sample the configuration space by considering only the phase sets corresponding to pseudo-atomic models, composed of a relatively small number (usually less than ten) of artificially large atoms. A single FAM experiment consists in the random generation of a large number of FAM's and the selection of the phase sets (which are called variants) corresponding to well magnitude correlated FAM's. This stage is formally similar to the multi-peak search incorporated into the *AMoRe* package (Navaza, 1994), but differs in having a larger number of selected phase sets. During the second stage the distribution of the selected variants in the multidimensional configuration space is investigated, and the variants are grouped in a small number of clusters, each consisting of closely related phase sets. The centroid phases (4) calculated for every cluster give a small number of alternative solutions to the phase problem.

4. Clusters selection and refinement procedures

The FAM approach to an *ab initio* solution of the phase problem results usually in a number of clusters with no obvious preference of one with respect to the others. Different additional considerations may be used in the selection procedure; *e.g.*, the corresponding electron-density distributions may be calculated and visually analyzed for every cluster (Urzhumtsev *et al.*, 1996). We present here a procedure of cluster comparison, which has a common basis with the maximum-likelihood-based choice of prior distributions (Bricogne & Gilmore, 1990), but is more straightforward and simpler in computer implementations. A comparison of these approaches is made in §7. Another feature of the suggested procedure is the possibility of refining and extending cluster phase sets. This part of the procedure has a common basis with the use of conditional probability distributions and is discussed in §6 below.

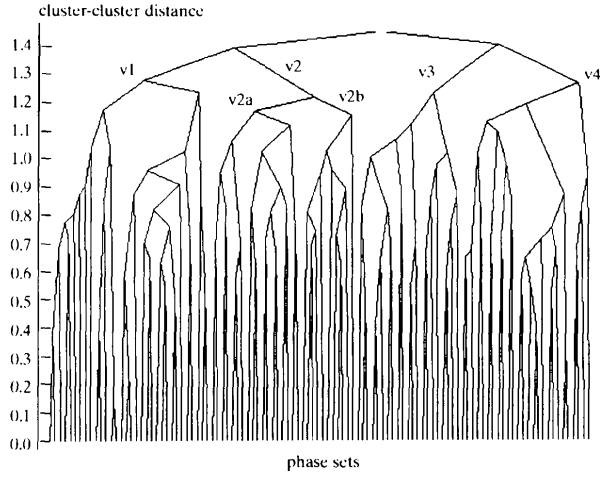

Fig. 3. Cluster tree for the phase sets calculated from the best magnitude correlated two-atoms models. Every node corresponds to merging two clusters into a larger one. The cluster–cluster distance is calculated as the average value of the variant–variant distance (7).

4.1. *Mask test*

To obtain the mask of the molecular region for a cluster of phase sets, a Fourier synthesis is calculated with the cluster centroid phases (4) and the observed magnitudes, and the points with highest density values are selected. The cut-off density level is defined so that the mask has a volume close to or slightly larger than the one of the molecular region. A series of FAM models, restrained to be in the mask region, is then generated a larger number of atoms (usually around 100). The histogram (or the corresponding cumulative function) for magnitude correlation values is then calculated.

The main idea of cluster comparison is that in general the number of WMC variants will be more when generating FAM's into the true molecular region, than into some arbitrary one. The plots of cumulative functions may be used to give qualitative estimates for the correctness of the different clusters (see §5 below). Special care has to be taken when masks lie on top of symmetry axes. To have a numerical estimate of the mask quality we can calculate the frequency for the selection criterion MC to be higher than the prescribed level ω

$$R_{\text{mask}} = \frac{\{\text{the number of FAM's with MC}(\{F_{\mathbf{h}}^{c}\}, \{F_{\mathbf{h}}^{0}\}) \geq \omega\}}{\{\text{the total number of generations}\}}. \quad (6)$$

The cut-off level ω may be chosen on the basis of the histogram corresponding to FAM generation in the whole unit cell, *e.g.* as the 'mean plus r.m.s.d' value of MC. The purpose of the test is to choose the right cluster; however, there might be more than one 'good' cluster (see §5).

4.2. *Mask refinement*

After the best mask regions are chosen as outlined in the mask test section, we can attempt phase refinement

and extension by repeating the envelope based FAM-generating procedure as follows.

(*a*) Generating FAM's inside molecular region only.

(*b*) Selecting ω-correlated variants.

(*c*) Calculating the mask average phases as the centroid phases for the chosen set of variants.

These new phases allow the calculation of a new mask region; the iteration of this procedure may be considered as cluster refinement, during which the resolution can be gradually increased.

5. Test results

The procedure described above has been applied to the neutron data of the $tRNA^{Asp}$–Asp RS complex (§2.5). Firstly, 1000 two-atoms FAM's (B = 12 000 Å^2) were generated into the whole unit cell and about 100 of them revealed magnitude correlations, equation (3), better then 0.8 (which is approximately the 'one plus sigma' value for the magnitude correlation). The corresponding phase sets were selected for a further analysis. Fig. 3 illustrates the distribution of the selected phase sets (variants) in the configuration space. This tree shows the order in which the clusters are merged together, on the basis of the variant–variant distance defined as

$$\mathrm{dist}(\{\varphi^1_{\mathbf{h}}\},\{\varphi^2_{\mathbf{h}}\}) = \left[\sum_{\mathbf{h}} |F^o_{\mathbf{h}}\exp(i\varphi^1_{\mathbf{h}}) - F^o_{\mathbf{h}}\exp(i\varphi^2_{\mathbf{h}})|^2 \Big/ \sum_{\mathbf{h}}(F^o_{\mathbf{h}})^2\right]^{1/2} \quad (7)$$

and the cluster–cluster distance defined as the average value of the variant–variant distance for all the pairs of different clusters members. The best phase alignment must be performed before calculating (7).

Secondly, the four top clusters were chosen for the mask test. They are marked as v1–v4 in Fig. 3. For each of them, a weighted Fourier syntheses was calculated with the cluster centroid phases and the observed magnitudes. The masks were then built by selecting the highest density values occupying 30% of the unit-cell volume, and they were tested by 100 generations of 100-atoms (B = 3000 Å) FAM's inside them. The distributions of the magnitude correlation values (3) are shown at Fig. 4; the largest number of WMC variants is obtained for the best mask and the smallest number for the worst mask. The two other clusters, with a PC value close to zero, have an intermediate distribution of MC values. Fig. 5 illustrates an attempt to use the same test to split the best cluster v2 into two smaller ones. Here we do not see such a clear picture as that of Fig. 4, and furthermore the distribution for the worse cluster v2b seems better than the one for the best cluster v2a. Therefore, the mask is sensitive when analyzing variants with very different quality, but may be misleading when

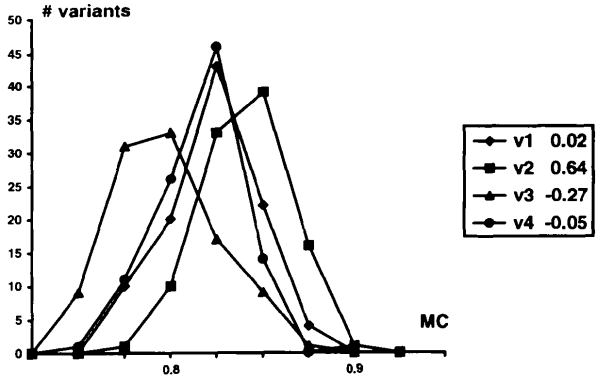

Fig. 4. The distribution of the magnitude correlation (3) for structure factors calculated from 100-atoms models generated into masks corresponding to different cluster nodes of the tree presented in Fig. 3. The values in the legend are the phase correlation (1) calculated for cluster centroid phases.

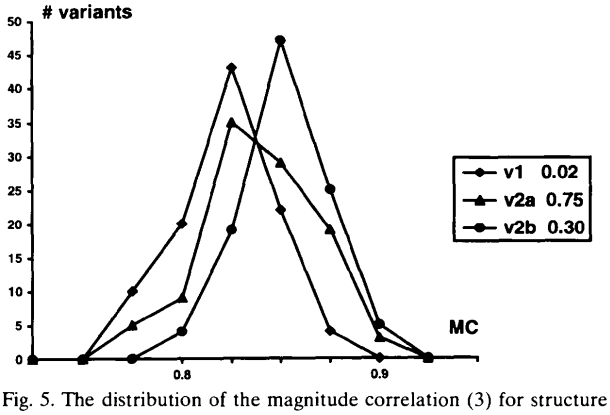

Fig. 5. The distribution of the magnitude correlation (3) for structure factors calculated from 100-atoms models generated into masks corresponding to different cluster nodes of the tree presented in Fig. 3. The values in the legend are the phase correlation (1) calculated for cluster centroid phases.

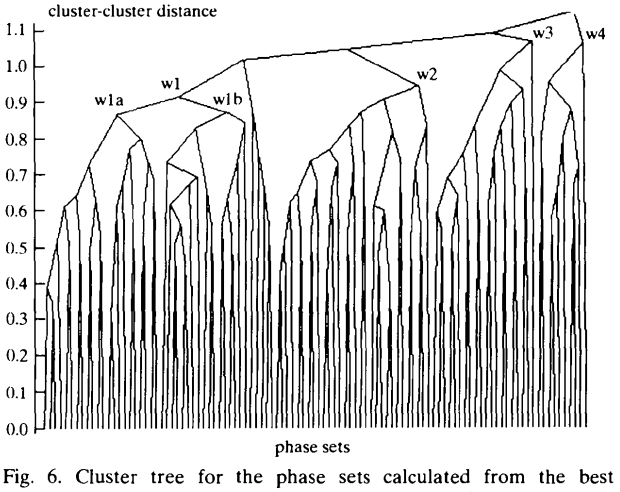

Fig. 6. Cluster tree for the phase sets calculated from the best magnitude correlated five-atoms models generated into the mask v2. Every node corresponds to merging two clusters into a larger one. The cluster-cluster distance is calculated as the average value of the variant–variant distance (7).

applied to clusters of close quality composed of a small number of phase sets each.

The mask corresponding to the cluster v2 was used at the next step to generate 1000 FAM models of five atoms each, restricting their positions to be in the mask region only. The five clusters chosen for the mask tests are marked in Fig. 6, and Fig. 7 shows the test results. Again, the best cluster has the best histogram, but the other clusters cannot be ranged unambiguously on the basis of these tests.

In the final stage the mask corresponding to the best cluster was used to generate 1000 FAM models of 100 atoms each, and the centroid phases for the 100 best WMC variants were calculated. Fig. 8 shows the sections of corresponding Fourier synthesis together with the C_α atoms of the synthetase dimer and the P atoms of the two tRNA's.

6. Theoretical analysis

Now we consider the FAM approach from a different point of view. We begin with the cluster discrimination procedure and finish with the mask refinement one.

6.1. *Generalized likelihood*

Let us now convert the mask region into an *a priori* probability distribution, which has a constant value inside the mask region and is equal to zero outside it. Supposing that the atomic coordinates are randomly distributed in accordance with this prior, we can define the likelihood of the prior as the probability (or, more precisely, as the value of the probability density function) of getting a set of structure-factor magnitude values equal to the experimental ones. The theoretical calculation of this probability is difficult and different approximations are necessary (Bricogne, 1984; Bricogne & Gilmore, 1990). We can significantly simplify the task

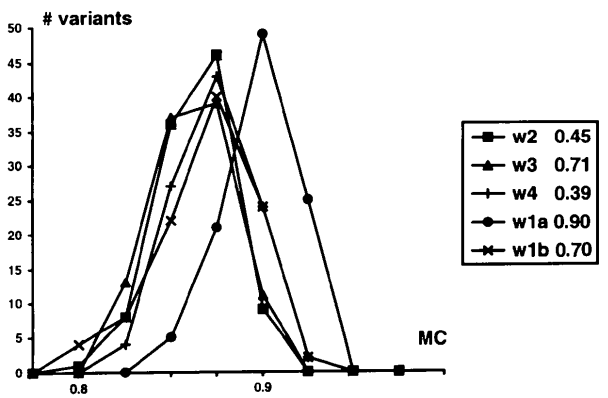

Fig. 7. The distribution of the magnitude correlation (3) for structure factors calculated from 100-atoms models generated into masks corresponding to different cluster nodes of the tree presented in Fig. 3. The values in the legend are the phase correlation (1) calculated for cluster centroid phases

if we define the generalized likelihood 'as the probability of getting a set of structure-factor magnitude values which are approximately equal to the experimental ones,

$$\mathrm{GL}_\omega = P[\mathrm{MC}(\{F_\mathbf{h}\}, \{F_\mathbf{h}^o\}) \geq \omega]. \tag{8}$$

This definition depends on the chosen cut-off level ω and coincides with the usual likelihood value for $\omega = 1$. The last formula allows the calculation of an approximate value of GL by a Monte-Carlo type simulation as,

$$\mathrm{GL}_\omega \simeq \frac{\{\text{the number of FAM's with } \mathrm{MC}(\{F_\mathbf{h}\}, \{F_\mathbf{h}^o\}) \geq \omega\}}{\{\text{the total number of generations}\}}. \tag{9}$$

This last value is just a numerical estimate of the cluster quality (6) suggested above. Therefore, our procedure of cluster discrimination may be considered as a generalization of the maximum-likelihood principle (Cox & Hinkley, 1974). The higher the value of ω, the closer is GL_ω to the ordinary likelihood value, but more generations must be carried out in the Monte-Carlo simulation to get a reliable result. Another consideration preventing us from using too high an ω value is that our mask-prior hypothesis may be too crude to expect precise magnitude values. Therefore, the approximate likelihood estimate (9) may be more robust in cluster discrimination.

6.2. *Cluster average and conditional distribution for phases*

It is usual practice to use conditional distributions $P(\varphi|F = F^o)$ to estimate phase values. Following the considerations outlined above, we can generalize this definition as,

$$P^{\mathrm{gen}}(\varphi) = P[\varphi|\mathrm{MC}(\{F_\mathbf{h}\}, \{F_\mathbf{h}^o\}) \geq \omega] \tag{10}$$

and use a Monte-Carlo simulation procedure to calculate these distributions empirically, separating the ω-correlated variants. The FAM experiments have revealed that such distributions are generally multimodal. The division into clusters may be considered further as an attempt to isolate regions of a particular mode in a multimodal distribution.

To be more precise, we consider a more general situation. The usual approach is to consider the probability for the phase $\varphi_\mathbf{h}$ provided that the particular magnitude $F_\mathbf{h}$ is known. The FAM generation and the separation of the variants with an MC value of 1.0 allows us (at least, theoretically) to estimate the multidimensional conditional distribution for all the phases provided that all the magnitudes have taken the observed values. To calculate empirically an analog of a multidimensional histogram for the variants with an MC value of 1.0 we must have an enormously large number

of generations. To reduce this number to a tractable one, we calculate empirically the distribution for every particular phase φ_h provided that all the magnitudes are close enough to the experimental values, which is a multimodal one. To reduce multimodality we separate variants into clusters and calculate restricted distribu-

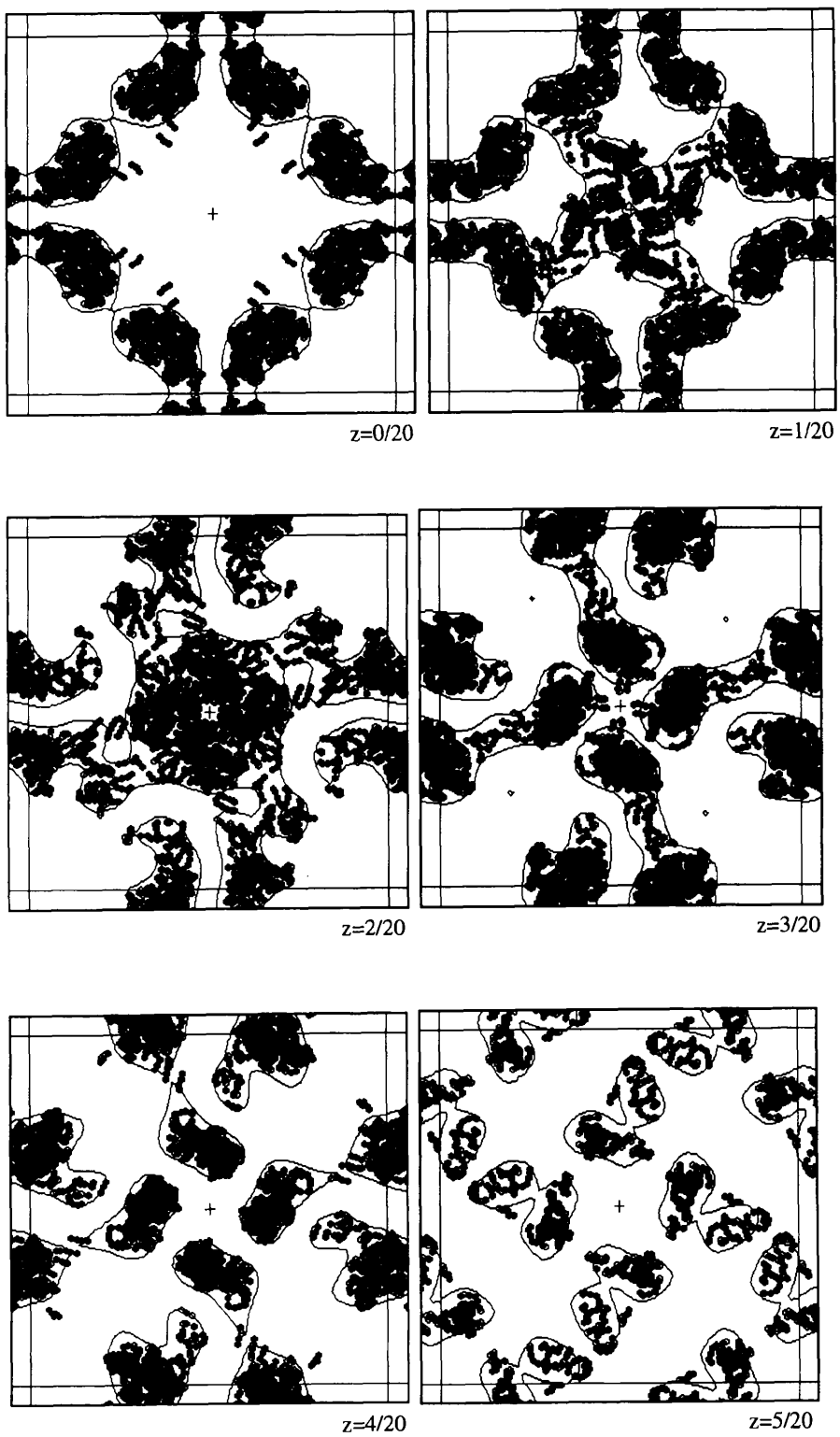

Fig. 8. The sections of the asymmetric part of the unit cell representing the Fourier map calculated with the best cluster phases. C_α atoms of the synthetase molecule and P-atoms of the tRNA molecule are marked.

tions corresponding to these clusters. The synthesis calculated with the centroid phases of these distributions corresponds to the cluster average. It must be noted that this restricted distributions may be again multimodal ones. An attempt to take the ones that are almost unimodal is close to the attempt to find reliable phases on the basis of Student's test (Bricogne & Gilmore, 1990).

7. Discussion

The problem of calculating the joint probability distribution (j.p.d.) for a set of structure factors, provided the atomic coordinates are distributed randomly, is one of the basic problems of mathematical crystallography. This problem has been the focus of attention for about half of century (Wilson, 1949; Karle & Hauptman, 1953; Woolfson, 1954; Bertaut, 1955; Klug, 1958; Bricogne, 1984) and numerous attempts were made to find a practical way to calculate the j.p.d. The main results were obtained at the cost of additional simplifications, the main ones being.

(*a*) All the atomic positions are supposed to be independent random variables, making it extremely difficult to take into account some stereochemical restrictions.

(*b*) The resulting formulas usually give the main term of an asymptotic expansion and it is difficult to estimate properly the effect of the other terms.

(*c*) The Gaussian approximation (the central limit theorem of the theory of probabilities) is used; it works properly for small deviations from the mean and less so for the large ones (strong reflections), which are usually of particular interest.

(*d*) The saddle point approximation (Bricogne, 1984) provides a more suitable tool to calculate probabilities for the large deviations, but at the cost of the solution of a set of non-linear simultaneous equations for every trial structure-factor set.

Additional difficulties appear when calculating the likelihood function from the j.p.d. as the marginal probability distribution for the magnitudes only. This requires the integration of the j.p.d. with respect to the phases, which is a serious problem even in the case of the multidimensional Gaussian approximation.

The proposed approach, based on the generalized likelihood functions coupled with Monte-Carlo simulations, shows a possible way to avoid these difficulties.

(i) It is simple to introduce additional stereochemical restrictions into the FAM simulation process; *e.g.* it is possible to generate a current atom position depending on the positions of previously generated atoms.

(ii) It is not necessary to have some large parameters (*e.g.* the number of atoms) and look for asymptotic formulae.

(iii) The Gaussian approximation and the problem of the likelihood calculation do not appear in this case.

On the other hand, empirical Monte-Carlo calculations have obvious disadvantages.

(i) The procedure is very time-consuming and, therefore, strongly limited by the available computer power.

(ii) There is no theoretical analysis.

(iii) There are a number of parameters to be determined, such as the cut-off level when calculating generalized likelihood, *etc.*, and the results may be sensitive to them.

Hence, the generalized likelihood Monte-Carlo based approach cannot be considered as replacing theoretical studies, but as providing an additional tool.

These numerically calculated j.p.d.'s can be successfully applied in phasing. In particular, as shown in §5, they are useful to distinguish between clusters, assuming that the correct one is more populated. A similar behavior was observed by Urzhumtsev & Podjarny (1995) when using the molecular-replacement method at very low resolution. This automated cluster analysis becomes essential when resolution increases, as the number of clusters is too large to be individually analyzed. Therefore, the combination of the FAM method and the detailed analysis of each mask shows the way to solve the phase problem at low resolution.

We thank Dino Moras, Bernard Rees, Jean Claude Thierry and Ada Yonath for their support and useful discussions. The neutron data of the $tRNA^{Asp}$–Asp RS complex were measured by Michel Roth and Dino Moras, and we thank them for making these data available. This work was supported by the joint collaboration program CNRS–Russian Academy of Sciences, by RFBR grant 97–04–48319, by the CNRS through the UPR 9004, by the Institut National de la Sante et de la Recherche Medicale and the Centre Hospitalier Universitaire regional. VYL was supported by a Travel grant of MENESRIP, France.

References

Andersson, K. M. & Hovmöller, S. (1996). *Acta Cryst.* D**52**, 1174–1180.
Baker, D., Bystroff, C., Fletterick, R. J. & Agard, D. A. (1993). *Acta Cryst.* D**49**, 429–439.
Bertaut, E. F. (1955). *Acta Cryst.* **8**, 537–543.
Bricogne, G. (1984). *Acta Cryst.* A**40**, 410–445.
Bricogne, G. & Gilmore, C. J. (1990). *Acta Cryst.* A**46**, 284–297.
Bystroff, C., Baker, D., Fletterick, R. J. & Agard, D. A. (1993). *Acta Cryst.* D**49**, 440–448.
Carbo, R., Leyda, L. & Arnau, M. (1980). *Int. J. Quant. Chem.* XVII, 1185–1189.
Cox, D. R. & Hinkley, D. V. (1974). *Theoretical Statistics*. London: Imperial College.
Harris, G. W. (1995). *Acta Cryst.* D**51**, 695–702.
Karle, J. & Hauptman, H. (1953). *Acta Cryst.* **6**, 131–135.
Klug, A. (1958). *Acta Cryst.* **11**, 515–543.
Lunin, V. Yu. (1993). *Acta Cryst.* D**49**, 90–99.

Lunin, V. Yu. & Lunina, N. L. (1996). *Acta Cryst.* A**52**, 365–368.
Lunin, V. Y. & Lunina, N. L. (1998). In preparation.
Lunin, V. Yu., Lunina, N. L., Petrova, T. E., Vernoslova, E. A., Urzhumtsev, A. G. & Podjarny, A. D. (1995). *Acta Cryst.* D**51**, 896–903.
Lunin, V. Yu., Urzhumtsev, A. G. & Skovoroda, T. P. (1990). *Acta Cryst.* A**46**, 540–544.
Lunin, V. Yu. & Woolfson, M. M. (1993). *Acta Cryst.* D**49**, 530–533.
Moras, D., Lorber, B., Romby, P., Ebel, J.-P., Giegé, R., Lewit-Bentley, A. & Roth, M. (1983). *J. Biomol. Struct. Dynam.* **1**, 209–223.
Navaza, J. (1994). *Acta Cryst.* A**50**, 157–163.
Podjarny, A. D., Rees, B., Thierry, J.-C., Cavarelli, J., Jesior, J. C., Roth, M., Lewitt-Bentley, A., Kahn, R., Lorber, B., Ebel, J.-P., Giegé, R. & Moras, D. (1987). *J. Biomol. Struct. Dynam.* **5**, 187–198.
Read, R. J. & Moult, J. (1992). *Acta Cryst.* A**48**, 104–113.
Ruff, M., Krishnaswamy, S., Boeglin, M., Poterszman, A., Mitschler, A., Podjarny, A., Rees, B., Thierry, J.-C. & Moras, D. (1991). *Science*, **252**, 1682–1689.
Sheldrick, G. M., Dauter, Z., Wilson, K. S., Hope, H. & Sieker, L. (1993). *Acta Cryst.* D**49**, 18–23.
Subbiah, S. (1991). *Science*, **252**, 128–133.
Subbiah, S. (1993). *Acta Cryst.* D**49**, 108–119.
Urzhumtsev, A. G. & Podjarny, A. D. (1995). *Acta Cryst.* D**51**, 888–895.
Urzhumtsev, A. G., Podjarny, A. D. & Navaza, J. (1994). *Jnt CCP4 ESF-EACBM Newslett. Protein Crystallogr.* **30**, 29–36.
Urzhumtsev, A. G., Vernoslova, E. A. & Podjarny, A. D. (1996). *Acta Cryst.* D**52**, 1092–1097.
Weeks, Ch. M., DeTitta, G., Hauptman, H., Thuman, P. & Miller, R. (1994). *Acta Cryst.* A**50**, 210–220.
Wilson, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.
Wilson, C. & Agard, D. A. (1993). *Acta Cryst.* A**49**, 97–104.
Woolfson, M. M. (1954). *Acta Cryst.* **7**, 61–64.
Woolfson, M. M. & Yao, J.-X. (1990). *Acta Cryst.* A**46**, 409–413.