Acta Cryst. (1995). D51, 896–903

On The *Ab Initio* Solution Of The Phase Problem For Macromolecules At Very Low Resolution: The Few Atoms Model Method

BY V. YU. LUNIN, N. L. LUNINA, T. E. PETROVA, E. A. VERNOSLOVA

Institute of Mathematical Problems of Biology, Russian Academy of Sciences, Pushchino, Moscow Region, 142292, Russia

AND A. G. URZHUMTSEV* AND A. D. PODJARNY†

UPR de Biologie Structurale, IGBMC, BP 163, 67404 Illkirch CEDEX, CU de Strasbourg, France

(*Received* 13 *January* 1995; *accepted* 10 *April* 1995)

Abstract

A method is proposed for the solution of the phase problem at very low resolution for macromolecules. It generates randomly a very large number of models, each consisting of a few (two to ten) pseudo-atoms. The corresponding amplitudes are used for selecting a subset of 'best' models by choosing those with the highest correlation with experimental values. The phases calculated from these 'best' models are analysed by a clusterization procedure leading to a few possible solutions, from which the correct one can be recognized by simple additional criteria. This method has been successfully applied to the neutron diffraction data of the AspRS–tRNA^{Asp} complex at 50 Å resolution and to data calculated from a model ribosome crystal at 60 Å resolution.

1. Introduction

The phase problem arises from the experimental difficulty of measuring phases in a diffraction experiment from molecular crystals. If no data other than the diffraction amplitudes from the native crystal are to be fitted, the phase problem is not solvable, since any set of phases will provide a density distribution. Therefore, additional information is necessary, which for the general case can be of two forms.

(1) Amplitudes from other crystals, related to the native amplitudes through a phase-dependent function; this is the case for the isomorphous-replacement method, anomalous scattering, multiwavelength anomalous scattering, solvent contrast and related methods where the phase is measured indirectly through the amplitudes.

(2) Information about the nature of the electron-density distribution, which limits the set of possible phases; for example, the strong constraint of atomicity has led to algorithms capable of providing the correct set of phases for small molecules.

Phases are currently obtained mostly through methods of type (1), *e.g.* isomorphous replacement. Methods of type (2) using general constraints, are currently being applied to improve an existing phase set (*e.g.* density modification, for review see Podjarny, Bhat & Zwick, 1987).

Several attempts are currently being made to develop methods of type (2) applicable to solve the phase problem *ab initio* for macromolecular crystals. One type of method is based on statistical approaches, which transform the known density constraint (*e.g.* atomicity, positivity, connectivity, flat solvent envelope, *etc.*) into a relationship between structure factors, without generating explicitly all possible densities which fit the constraints. An example is maximum-entropy methods (Bricogne, 1984, 1993; Navaza, 1985). These methods can also generate a very large quantity of individual solutions, and try to identify the correct one using both the information on diffraction amplitudes and on electron density. This can be carried out as an extension of standard direct methods, *e.g.* by generating randomly a large number of starting phase sets and applying direct methods to them (Yao Jia-Xing, 1981; Hauptman, 1994).

Another type of method, not based on statistical assumptions, is being developed in the very low resolution range (for review see Podjarny & Urzhumtsev, 1995), where the number of unknown phases is much smaller than at higher resolution ranges. One possibility is to search for a single solution, for example through the condensation of a large number of spheres (Subbiah, 1991), to generate a model which fits the observed amplitudes. A major problem in this approach is how to assure that the phase space is explored exhaustively, so the correct solution is not missed. This can be solved with a quasi-exhaustive investigation of all the possible density distributions, even in the absence of any model. Such an investigation has been carried out (Lunin, Urzhumtsev & Skovoroda, 1990) using the phases as variables and the histogram of the resulting density as the constraint.

* Permanent address: Institute of Mathematical Problems of Biology, Russian Academy of Sciences, Pushchino, Moscow Region, 142292, Russia.

† To whom correspondence should be addressed.

It showed the feasibility of a Monte Carlo approach to find the correct solution by sampling the space of possible solutions, where every point represents a full set of structure factors. However, several problems arose.

(1) The generated phase sets agreeing with the given criteria include both correct and incorrect solutions.

(2) The number of amplitudes that can be phased is very small, limiting strongly the resolution of the resulting image.

(3) Since the histograms are calculated in real space, the reflections in the inner core are necessary.

To address (1), a procedure was developed to identify possible solutions (Lunin *et al.*, 1990). This procedure works in a multidimensional space where every phase set is a point. Its application showed that the phase sets that agree with the given criteria are 'clustered' around a few separate points. This property is possibly linked to the fact that the low-resolution envelope depends mostly on a few phase invariants linked to reflections with strong amplitudes, and therefore the number of significant degrees of freedom is less than the total number of phases. However, since in this method the variables are the single phases, problem (2) remains unsolved.

2. The FAM method

To address the problems outlined above, a description of the electron density in terms of a few atoms model (FAM) was developed. In this approach, the variables are the positions of a very small number (<10) of large Gaussian scatterers or 'pseudo-atoms', and the fitness criterion is the agreement between structure-factor amplitudes calculated from these scatterers and the observed ones. Compared to the histogram method, the number of variables is significantly reduced. Moreover, as the fitness criterion is applied in reciprocal space, it is less sensitive to missing reflections. The *a priori* knowledge of the exact histogram is replaced by the determination of only two parameters, the number and size of the spheres.

3. Approximation of density with a few atoms model

The quality of the approximation of the very low resolution structure factors with those calculated from a FAM was initially checked using data sets from the AspRS–$tRNA^{Asp}$ complex (in what follows this is called the AspRS complex; Moras *et al.*, 1983; Urzhumtsev, Podjarny & Navaza, 1994), RNAse Sa (Sevcik, Dodson & Dodson, 1991) and ribosomal factor G (Chirgadze *et al.*, 1991). One of these checks, using the cubic form of the AspRS complex, is described in detail below. The structure has been solved by molecular replacement (Urzhumtsev *et al.*, 1994) with the package *AMoRe* (Navaza, 1994), using X-ray data to 8 Å resolution and a model from another crystal form ($P2_122$) solved at 2.9 Å resolution (Ruff *et al.*, 1991). The cubic crystal form is particularly well suited for low-resolution work, due to the large unit cell (space group $I432$, a = 354 Å), the large solvent content (78%), and the compact shape of the complex. Three different H_2O/D_2O contrast neutron data sets ($d > 16$ Å) were collected (Moras *et al.*, 1983), corresponding to the full complex, the synthetase moiety and the tRNA moiety. The neutron diffraction data sets can be fitted correctly with the molecular-replacement model [$\mathrm{Corr}(F_{mod};F_{obs}) > 92\%$ at 20 Å].

A 50 Å resolution map calculated with 31 observed amplitudes and model phases shows a clear ovoidal peak corresponding to the position of the synthetase dimer. Symmetry-related peaks are joined by 'arms' corresponding to the tRNA's. To test the validity of the few atoms approximation, a four-spheres model was built by putting two atoms in the position of the synthetase and two atoms near the tRNA 'arms'. This model could be refined to an R factor of 16% at 50 Å resolution, corresponding density correlation $\mathrm{Corr}(\rho_{mod};\rho_{true})$ = 71%. It should be noted that a three-spheres model has a higher R factor (24%) but the 50 Å resolution map reproduces almost equally well the correct one, $\mathrm{Corr}(\rho_{mod};\rho_{true})$ = 69%, with the third atom corresponding to the 'arms' joining the synthetase positions. Therefore, the observed data at 50 Å can be quite accurately reproduced by a three- or four-spheres model. Note that the positions of the pseudo-atoms do not necessarily agree with the centres of gravity of the individual molecules.

4. Exhaustive searches with a multiple-spheres model

4.1. *One-sphere searches*

The few atoms modelization for the AspRS complex was first used in a molecular-replacement context by Podjarny, Rees *et al.* (1987) who performed a one-sphere search using the neutron diffraction data with an adaptation of the program *TRAN* (J. Nachman, personal communication). This single-sphere search depends on three parameters: data resolution, $F/\sigma(F)$ cut-off and sphere size. After a trial-and-error procedure, it was found that a resolution of 50 Å, a value of $F/\sigma(F)$ of 15 and a Gaussian sphere with $\sigma(r)$ = 20 Å were optimal. Varying these parameters caused a variation of noise peaks on special positions but the correct solution remained stable.

These searches show a property of the chosen subset of phase space that appears consistently. The generated points can be classified in two types: the correct solution and several spurious ones. Both types agree with the search criterion (in this case, the amplitude correlation) and it is necessary to apply an independent check (in this case, the assumption that the correct solution is not on a special position) in order to identify the correct solution.

4.2. *Two-spheres searches*

These results were extended recently to two-spheres searches, at 50 Å resolution, with $F/\sigma(F)$ cut-off of 10 and with $\sigma(r)$ = 20 Å, using the diffraction from the full complex. Special programs were developed for this purpose. A clear signal appeared, corresponding to one sphere being in the centre of the molecule and the second in a packing interaction. The two-spheres exhaustive searches represent the computing limit of current algorithms working on laboratory workstations. To introduce more spheres it is necessary to sample the solution space differently, for example by random checking of points instead of a systematic search.

5. Monte Carlo searches and cluster analysis

The following technique is proposed.

(1) Few atoms models (FAM's), consisting of a small number (two to ten) of equally large spheres, are generated in large quantities by randomly choosing the centre of every sphere. The only parameters to be varied are the size of the spheres and their number.

(2) Structure factors are calculated for every FAM. The resulting sets of amplitudes and phases represent a sampling of solution space. The quantity of FAM's should be large enough so that this sampling is representative.

(3) The generated sets of structure factors are filtered according to the correlation of calculated and observed amplitudes. Those sets with amplitude correlation larger than a given threshold are kept for further analysis (without discrimination between them). It should be noted that at this moment every FAM is represented only by its set of structure factors, and not by the original coordinates.

(4) The sets of structure factors selected in (3) are grouped using a clusterization technique (see the *Appendix*). This technique identifies the regions inside the space of variables which are densely populated by kept solutions and produces their average inside each region.

In general, it is necessary to have additional criteria (other than just the amplitude correlation) in order to choose the correct solution after step (4).

While the exhaustive searches described above are a particular case of low-resolution molecular-replacement methods (Urzhumtsev & Podjarny, 1995, and references therein), the FAM method is of a more general character.

(*a*) The internal geometry of the model is completely unrestrained, allowing for sphere overlap; the imposed model is not a series of separate spheres, but a compact molecular region surrounded by a flat solvent region.

(*b*) The output of FAM is not the position of the model (as in molecular replacement) but the phases associated with a group of them; FAM models which are very different in atomic positions but lead to similar phase sets are associated in a 'cluster', and the final result is a single averaged phase set for this 'cluster'.

6. Applications of Monte Carlo searches to the case of the AspRS complex

6.1. *FAM model generation*

The algorithm described above was tested using the neutron diffraction data set from the AspRS complex. These tests were conducted at 50 Å (31 reflections), where the experimental data are practically complete. Different numbers of pseudo-atoms were tried, ranging from one to ten. To measure the effectiveness of the FAM technique in a test case where the answer is known, the correlation $\mathrm{Cor}(\varphi,\varphi_{\mathrm{true}})$ for the different phase sets was calculated after every step. A more detailed description of this calculation, including the effects of symmetry, is given in the *Appendix*.

Fig. 1(*a*) shows the histogram of phase correlation after generation of one-, two- and ten-pseudo-atom models, compared with that of a random generation of independent phases with uniform probability. The random distribution is essentially a measure of the volume of phase space at a given distance in correlation with the correct solution.

The FAM distributions follow the random one for values of $\mathrm{Cor}(\varphi,\varphi_{\mathrm{true}}) < 0.45$, but for the one- and two-atom cases become significantly larger when $\mathrm{Cor}(\varphi,\varphi_{\mathrm{true}}) > 0.45$. Note that this effect diminishes as the number of atoms increases; it is largest for one-atom FAM's, remains important for two-atoms FAM's, but almost disappears for ten-atoms FAM's, as in this latter case the number of variables (30 atomic coordinates) approaches the number of phases (31 phases). This is an example of the fact that the filtering power of this first stage of FAM, as compared with just random phase sets, depends on its ability to generate good phase sets with fewer degrees of freedom.

An interesting characteristic of the distribution of points after FAM generation is its bimodality, clearly seen for the one-atom FAM generation (Fig. 1*a*). This bimodality corresponds to phase sets where the pseudo-atom is either in the molecular region, with a positive correlation, or in the solvent region, with a negative density correlation. However, since the FAM method does not generate models which reproduce closely the large and convoluted solvent region, there are no peaks for $\mathrm{Cor}(\varphi,\varphi_{\mathrm{true}})$ values close to -1. A possible way of eliminating this bimodality (when the correct solution is known) is to maximize the correlation in each point with respect to the transformation $\rho = -\rho$ as shown in Fig. 1(*b*). The resulting curve has only positive correlations, and its highest end remains the same as before.

6.2. *Model selection by amplitude correlation*

The filtering of the phase sets from the FAM modelization is further enhanced by choosing the solutions with a large correlation in F. This is shown in Fig. 1(*c*) for the two-atoms FAM selection, with cut-offs at 0.74 (corresponding to 1σ deviation from the mean

amplitude correlation) and 0.81 (corresponding to 2σ). This last curve, corresponding to 1179 possible points out of 500 000 generated, shows peaks for high values of $\mathrm{Cor}(\varphi,\varphi_{\mathrm{true}})$ indicating the creation of a densely populated zone of phase sets near the correct one.

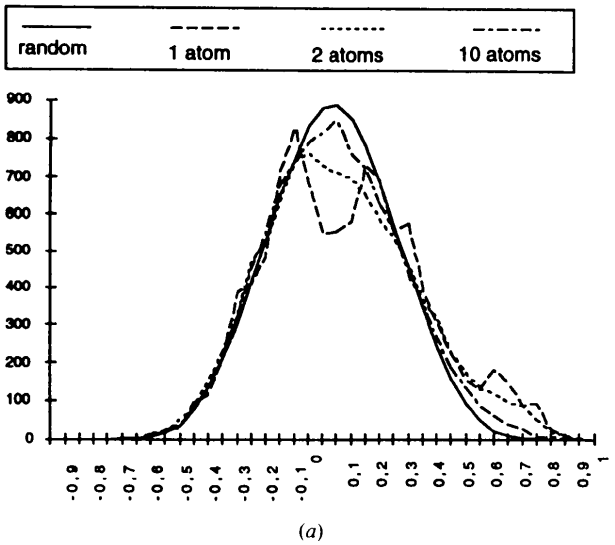
6.3. *Clusterization*

In the general case when the correct solution is not known, the identification of the dense zones after filtering is carried out by the clusterization procedure described in the *Appendix*, which leads to a 'cluster tree'. Fig. 2 shows several representations of the results of the application of this procedure to the data described in §6.2.

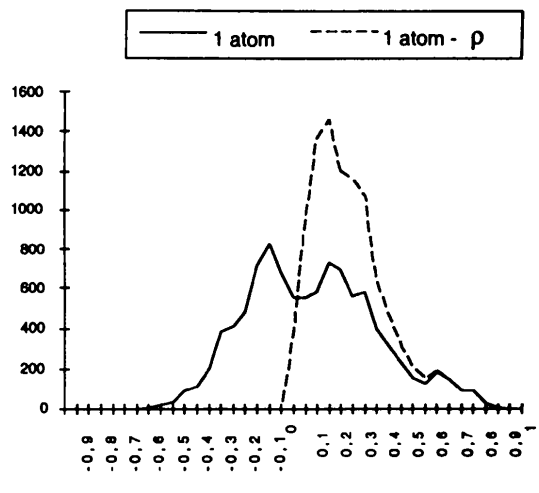
The untreated 'cluster tree' is shown in Fig. 2(*a*), and its labeled interpretation in Fig. 2(*b*). There are three main clusters, indicated as C11, C12 and C13 (Fig. 2*b*). Investigation of these points allows the clear choice of one of them (C12), based on the criterion that the core of the density should not be on symmetry elements (here, on the dyad axis $x = y$, $z = 0$). C12 is the largest cluster, and it can be further divided into two points, C21 and C22; again, one point is clearly better than the other by the same criterion. In this way, the correct pathway could be followed without any comparison with the correct density. When the process was finished, the syntheses corresponding to the 'cluster' phase sets were compared with the correct one at 50 Å resolution. By going down two levels of the cluster tree, the correlation with the correct synthesis increases from 77 to 93%; however, going one level further does not increase this value anymore.

As shown in Fig. 2(*c*), these clusters were already seen before the clusterization procedure in the histogram of $\mathrm{Cor}(\varphi,\varphi_{\mathrm{true}})$ after selection in amplitude correlation, where they appeared as separate peaks at different distances from the correct solution. The figure shows the decomposition of this histogram (without normalization) into separate ones, corresponding to the different clusters. A clear correlation between the three main peaks and the three main clusters C11, C12 and C13 is seen. Furthermore, the cluster C12 shows a finer peak structure, corresponding to its partition into smaller ones; the correct cluster C22 is also indicated.

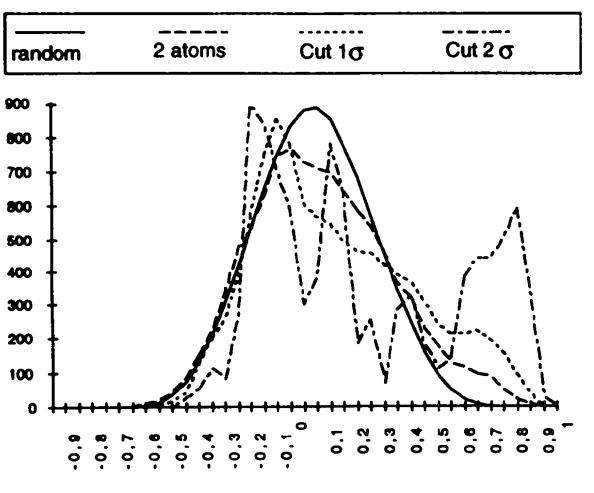
For this favourable case, the compact portion of the clusters explains most of the models, and at a given distance from the correct solution one of them dominates the histogram. Note, however, the overlap of the peaks



(*a*)



(*b*)



(*c*)

Fig. 1. Histograms of the number of points with a given value of the correlation $\mathrm{Cor}(\varphi,\varphi_{\mathrm{true}})$ for different conditions of FAM generation. The correlations obtained with random phases is also shown in (*a*) and (*c*). In all cases, the histograms are normalized to the same area. (*a*) FAM's with different number of atoms: one, two and ten; (*b*) one-atom FAM's with and without the transformation $\rho \rightarrow -\rho$; (*c*) two-atom FAM's with different cut-offs in amplitude correlation: none, 1σ and 2σ.

with correlation of 0.15, between the cluster C11 and the tail of cluster C12. This situation is schematized in Fig. 2(*d*).

The process of clusterization has, therefore, been successful in recognising the correct solution. It is interesting to see how the merging and weighting of phase sets improves their quality. This is accomplished by two mechanisms: weighting down of structure factors whose phases are not collinear between different models, and cancellation of errors (the error of the average of several measurements is less than the average of the absolute value of the errors of each measurement). This is most evident for cluster C00. The phase correlation of the cluster is 77%, while the average phase correlation of all models is only 26%. Likewise, the phase correlation of cluster C22 is 93%, while the average phase correlation for the intervening models is 77%.

As noted above, this averaging inside the cluster differentiates the FAM method from molecular replacement because the goal is not to identify the correct solution as a single point, but to generate a manifold of points near the correct solution. Due to the approximations in the model, none of these points corresponds exactly to the correct solution, but the average approaches it very closely.

Fig. 3 shows the 50 Å resolution map calculated with the phases and figures of merit from the correct cluster superposed with the correct model. This figure shows that the predicted envelope follows closely the borders of the model, and has therefore useful phasing information.

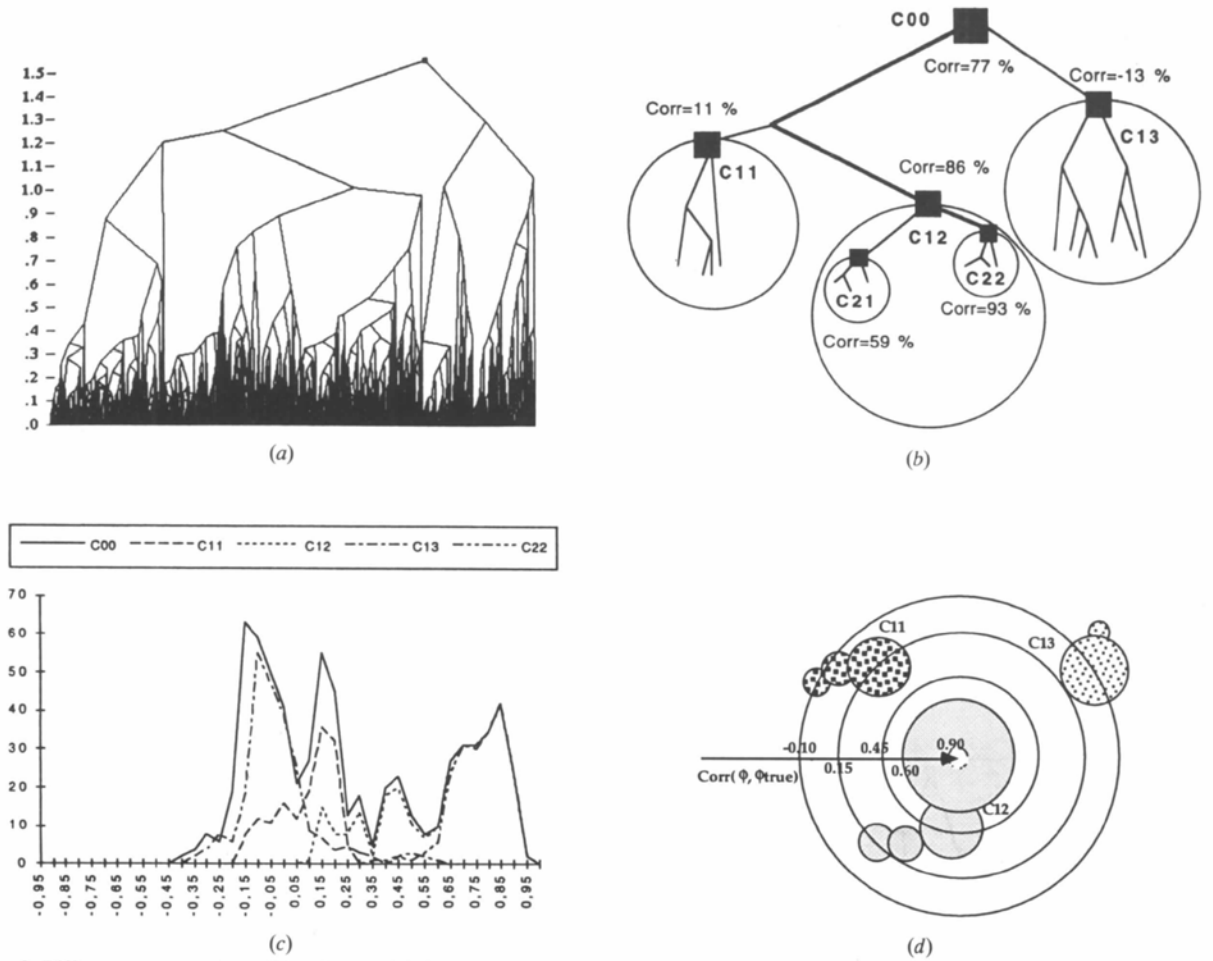


Fig. 2. Different representations of the clusters: (*a*) the untreated cluster tree, generated as described in the *Appendix*; (*b*) a labelled scheme of the cluster tree, indicating the main nodes and the corresponding $\mathrm{Cor}(\varphi,\varphi_{\mathrm{true}})$ values; (*c*) analysis of curve (2σ) of Fig. 1(*c*); for this analysis, the $\mathrm{Cor}(\varphi,\varphi_{\mathrm{true}})$ histograms were generated independently for each one of the main clusters; the peaks of these partial histograms correspond to the peaks of the main histogram, showing that the clusters appear in the main histogram independently of the clusterization procedure; note the overlap in $\mathrm{Cor}(\varphi,\varphi_{\mathrm{true}})$ values of the smaller parts of C12 with the main parts of C11; (*d*) a schematic two-dimensional view of the shape of the clusters; each one of the first level clusters (C11, C12 and C13) is represented by a series of circles, corresponding to the major peaks of the corresponding curve in Fig. 2(*c*); note that the central circle, corresponding to C22, is empty at the centre, indicating that no single FAM model reproduces the exact solution; the phases obtained by averaging this circular crown will be close to the centre, $\mathrm{Cor}(\varphi,\varphi_{\mathrm{true}}) = 93\%$.

7. Application to a ribosome model crystal

A case where finding an envelope is of great practical importance is that of the ribosome particle. To check the possibility of its application, the FAM method was tested with simulated ribosome data based on a 28 Å electron microscopy image of the 50S particle from *Bacillus stearothermophilus* (Berkovitch-Yellin, Wittmann & Yonath, 1990). For the purpose of the current tests, the corresponding envelope was packed simulating proper crystal contacts in the tetragonal crystal lattice of the 50S particle from *Thermus thermophilus* (space group $P4_12_12$, $a = b = 495$, $c = 196$ Å; Volkmann *et al.*, 1990).

The FAM technique was applied to structure factors calculated from this simulated crystal in the resolution range from 60 to 500 Å. Fig. 4(*a*) shows the cluster tree for the case of five atoms, 1 000 000 generations, with 90 solutions having $\mathrm{Cor}(F) > 0.85$. Fig. 4(*b*) shows the exact map at the resolution of 60 Å, Fig. 4(*c*) shows the maps (also at 60 Å resolution) corresponding to the largest clusters and Table 1 gives their evaluation. The tree is dominated by a single large cluster, of radius close to 1.0 (corresponding to 60° phase error). Therefore, for these calculated data, the amplitude cut-off alone is enough to produce a first image. Note that, due to the merging of structure factors inside a cluster, higher resolution structure factors tend to be weighted down. This leads to an 'effective' resolution $d_{\mathrm{eff}} = 110$ Å of the map for the largest cluster (d_{eff} is a conditional limit defined as follows: for $d < d_{\mathrm{eff}}$ mean value of the figure of merit is less than 0.5). Dividing this cluster into smaller ones leads to an improvement of the image, leading in two levels to a map with $d_{\mathrm{eff}} = 60$ Å and $\mathrm{Cor}(\varphi,\varphi_{\mathrm{true}}) = 0.86$, and going one level further does not substantially improve the image. As in the case of the AspRS complex, the correct solution can be easily identified by imposing a minimum of density on particle contacts.

Table 1. *Evaluation of the different syntheses for the model ribosome crystal shown in Fig. 4*

Level	Cluster	$C(\varphi,\varphi_{\mathrm{true}})$ (%)	D_{eff} (Å)	No. of points	Notes
0	C00	73	110	89	
1	C11	57	70	48	
	C12	80	65	41	Best
2	C21	60	60	26	
	C22	46	60	22	
	C23	68	60	24	
	C24	86	60	17	Best
3	C31	75	60	5	
	C32	87	60	2	Best

Similar maps, with $\mathrm{Cor}(\varphi,\varphi_{\mathrm{true}})$ values ranging from 0.84 to 0.88, can be obtained by varying the number of trial models from 500 000 to 2 000 000, the number of atoms from five to nine and the amplitude correlation

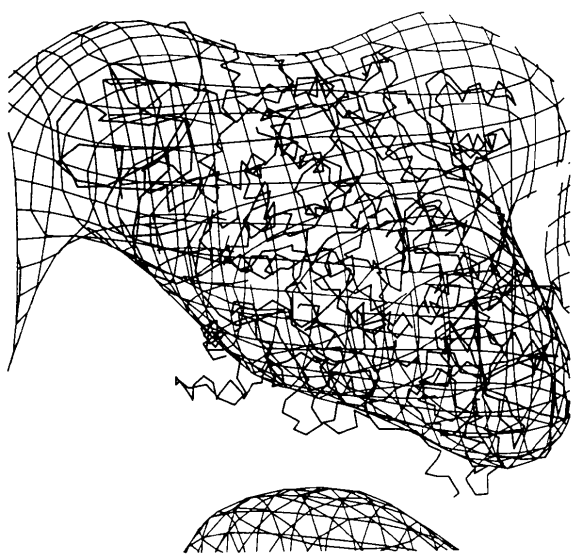


Fig. 3. Overlap of the 50 Å resolution synthesis calculated with FAM phases with the atomic model of the AspRS complex in the cubic cell. The shape of the density follows the model closely. The additional densities correspond to symmetry-related molecules (not shown).

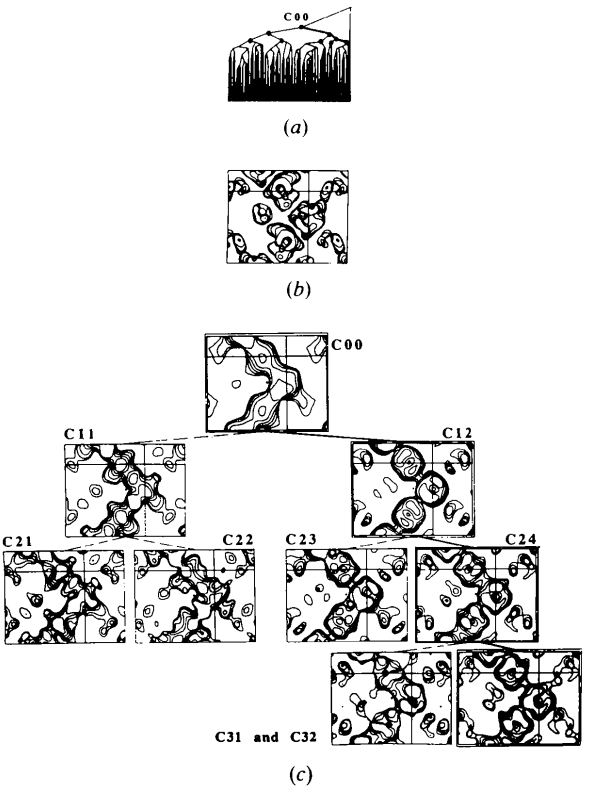


Fig. 4. Cluster tree for the model of the ribosome crystal and corresponding maps. The 10^6 FAM's were obtained with five atoms. After an amplitude cut-off of 0.85, 90 points were left. (*a*) Cluster tree obtained as described in the *Appendix*. Note that one point is clearly outside of the main cluster C00. (*b*) Exact synthesis at 60 Å resolution. (*c*) Maps corresponding to the nodes of the first four levels (0 to 4) of the cluster tree. The correct path is marked by thick lines, and the best synthesis at each level is framed by a double line. The evaluation of these synthesis is shown in Table 1.

threshold from 0.85 to 0.83. The topology of the cluster tree tends to be independent of the number of accepted solutions.

8. The retrieval of 'inner core' amplitudes

This method depends on the availability of the 'inner core' of reflections, which are generally very difficult to measure. Previous experience in density-modification techniques (Rayment, 1983; Urzhumtsev, 1991) has shown that if an approximate envelope can be obtained, the inner core of reflections can be retrieved from this envelope. Extension of this idea to the FAM method suggests that if the proper cluster could be identified from data without the 'inner core' reflections, the FAM models inside the clusters could provide information about them.

This possibility was tested using the ribosome model data. In this test, all reflections were calculated from FAM models, but the criteria of checking by amplitude correlation with observed data and the clusterization procedure were applied using only the reflections in the 60–150 Å resolution range. For this calculated case with exact amplitude values, these reflections were enough to make FAM converge to the correct solution, and the generated FAM models were able to retrieve the phases in the 150–500 Å resolution range. The possibility of generalizing this result to experimental data with amplitude errors is being analysed.

9. Concluding remarks

The results shown above prove that using only the amplitudes the FAM method can produce a small number of images, from which the correct one can be picked unambiguously using simple additional criteria (*e.g.* packing). Therefore, it provides an *ab initio* method for solving the phase problem at very low resolution, from which a detailed envelope is obtained. It is important to note that the phase prediction carried out by averaging all solutions inside the correct cluster is significantly better than the one obtained by a single FAM model.

This method works optimally with the correct choice of number and size of spheres, complexity of image and resolution. Once a solution is found for the 'inner core' of reflections, it is necessary to extend it to higher resolution. Several alternatives are being tested for this extension, for example to limit the number of models to be searched by simple criteria in real space (*e.g.* generation inside one prescribed region), as well as to use a limited part of phase space (*e.g.* inside one cluster). Furthermore, this methodology is not restricted to spheres, but can be extended to different modelizations.

The authors thank Drs P. Dumas, D. Moras, J. Navaza, B. Rees, M. Roth, N. Volkmann and A. Yonath for useful discussions. The contrast variation neutron data of the AspRS complex used in this work were collected by Drs M. Roth, A. Lewitt-Bentley and D. Moras. We thank them for making these data available to us. This work has been supported by grant 94-04-12844 of the Russian Foundation for Fundamental Research, by grant RMZ000 from the International Science Foundation, by the CNRS through the UPR 9004, by the Institut National de la Santé et de la Recherche Médicale, by the Centre Hospitalier Universitaire Régional and by a joint collaboration CNRS–Russian Academy of Sciences. VYL was supported by ACA-USNCCr.

APPENDIX

The process of clusterization is carried out as a function of a parameter δ measuring the distance between points in the space of phase sets. Initially, $\delta = 0$ and each point is an individual cluster; δ is then continuously increased, and cluster pairs are iteratively merged as the distance between them becomes less than δ (Lunin *et al.*, 1990). This distance, *e.g.* between points j and k, is determined in terms of the correlation of the corresponding densities as,

$$D_{jk} = \kappa^*\{\int[\rho_j(r)-\rho_k(r)]^2 d^3r\}^{1/2} = (2-2C_{jk})^{1/2}, \quad (A1)$$

where the density correlation is,

$$C_{jk} = \{\int[\rho_j(r)\rho_k(r)]d^3r\}/(\{\int[\rho_j(r)]^2 d^3r\} \times \{\int[\rho_k(r)]^2 d^3r\})^{1/2}, \quad (A2)$$

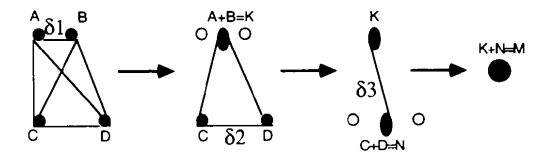
(Lunin *et al.*, 1990) and can be directly calculated in terms of structure factors (Read, 1986; Lunin & Woolfson, 1993). This distance is 0 when the points are the same ($C_{jk} = 1$), 2 when the points are exactly opposite ($C_{jk} = -1$) and $2^{1/2}$ when they are not related ($C_{jk} = 0$). A D_{jk} value of 1.0 corresponds to a density correlation of 0.5 (weighted phase error of 60°). It is important to note that this correlation C_{jk} depends on the choice of the origin (and enantiomorph, if this operation is admissible in the particular space group). Therefore, for every point all corresponding correlations should be checked and the maximum kept as a true one. When the two clusters consist of more than one point, the distance between clusters is defined as the minimum of all pairwise distances between points in different clusters.

At the very low resolution range, the model can correspond to the protein or to the solvent, and therefore the distance D_{jk} can also be optimized with respect to the transformation $\rho = -\rho$. In this case, one model is kept as a reference and all maps are explored both for maxima and minima. This option has been implemented in the last versions of FAM and has been applied for the calculation described in Fig. 1(*b*).

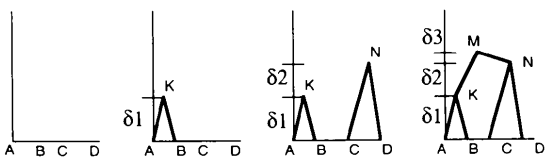
The scheme of the clusterization is shown in Fig. 5(*a*). Let us consider four points in solution space, *A*, *B*, *C* and *D*, and vary the parameter δ from 0 to 2 during the clusterization process. When $\delta = \delta 1$, the distance between *A* and *B*, these points are merged into a new cluster, *K*. When $\delta = \delta 2$, the distance between *C* and *D*, these points are merged into a new cluster, *N*. When $\delta = \delta 3$, the distance between *K* and *N*, these points are merged into a new cluster, *M*. Fig. 5(*b*) shows how this process can be graphically represented in a 'tree', in which the *X* axis shows the solution number, ordered by proximity to avoid crossing of lines, and the *Y* axis shows the distance δ. A new point, for example *K*, will have as *Y* coordinate the distance $\delta 1$, and an *X* coordinate between *A* and *B*. To illustrate the merging, lines are drawn from the original points to the new one. The process of clusterization continues in this way until all solutions have merged into a single point.

Once full clusterization is achieved, the tree can be analysed to obtain representative phase sets by averaging all phase sets in each cluster. The zero level corresponds to the overall average (point *M*). The first level corresponds to the splitting of this point into its components (points *N* and *K*). The next levels can be calculated similarly. For each level, the phase $\varphi(h,k,l)$ and its figure of merit $w(h,k,l)$ can be obtained by averaging all the original phases $\varphi_j(h,k,l)$, where the index j goes over all the points in the cluster, as follows,

$$w(h,k,l)\exp[i\varphi(h,k,l)] = \sum_j \exp[i\varphi_j(h,k,l)]. \quad (A3)$$



(*a*)



(*b*)

Fig. 5. Scheme of clusterization procedure.

References

BERKOVITCH-YELLIN, Z., WITTMANN, H. G. & YONATH, A. (1990). *Acta Cryst.* B**46**, 637–643.

BRICOGNE, G. (1984). *Acta Cryst.* A**40**, 410–445.

BRICOGNE, G. (1993). *Acta Cryst.* D**49**, 37–60.

CHIRGADZE, YU. N., BRAZHNIKOV, E. V., GARBER, M. B., NIKONOV, S. V., FOMENKOVA, N. P., LUNIN, V. YU., URZHUMTSEV, A. G., CHIRGADZE, N. YU. & NEKRASOV, YU. N. (1991). *Dokl. Akad. Nauk SSSR*, **320**, 488–491.

HAUPTMAN, H. (1994). Am. Crystallogr. Assoc. Annu. Meet., June 25–July 1, 1994, INFORUM, Atlanta, Georgia, p. 38.

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. (1985). *Acta Cryst.* A**41**, 232–244.

NAVAZA, J. (1994). *Acta Cryst.* A**50**, 157–163.

PODJARNY, A. D., BHAT, T. N. & ZWICK, M. (1987). *Ann. Rev. Biophys. Chem.* **16**, 351–373.

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.

PODJARNY, A. D. & URZHUMTSEV, A. G. (1995). *Methods Enzymol.* Submitted.

RAYMENT, I. (1983). *Acta Cryst* A**39**, 102–116.

READ, R. (1986). *Acta Cryst.* A**42**, 140–149.

RUFF, M., KRISHNASWAMY, S., BOEGLIN, M., POTERSZMAN, A., MITSCHLER, A., PODJARNY, A., REES, B., THIERRY, J.-C. & MORAS, D. (1991). *Science*, **252**, 1682–1689.

SEVCIK, J., DODSON, E. & DODSON, G. G. (1991). *Acta Cryst.* B**47**, 240–253.

SUBBIAH, S. (1991). *Science*, **252**, 128–133.

URZHUMTSEV, A. G. (1991). *Acta Cryst.* A**47**, 794–801.

URZHUMTSEV, A. G. & PODJARNY, A. D. (1995). *Acta Cryst.* D**51**, 888–895.

URZHUMTSEV, A. G., PODJARNY, A. D. & NAVAZA, J. (1994). *J. CCP4 ESF-EACBM Newslett. Protein Crystallogr.* **30**, 29–36.

VOLKMANN, N., HOTTENTRÄGER, S., HANSEN, H. A. S., ZAYTSEV-BASHAN, A., SHARON, R., YONATH, A. & WITTMANN, H. G. (1990). *J. Mol. Biol.* **216**, 239–241.

YAO JIA-XING (1981). *Acta Cryst.* A**37**, 642–644.