Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Cluster analysis for phasing with molecular replacement: a feasibility study

Andreas Buehler,[a,b]‡ Ludmila Urzhumtseva,[c] Vladimir Y. Lunin[d] and Alexandre Urzhumtsev[a,b]*

[a]Physics Department, Nancy University, 54506 Vandoeuvre-les-Nancy, France, [b]CEBGS – Institut de Génétique et de Biologie Moléculaire et Cellulaire, Département de Biologie et de Génomique Structurales, CNRS–ULP–INSERM, 1 Rue Laurent Fries, 67404 Illkirch, France, [c]Architecture et Réactivité de l'ARN, Université de Strasbourg, Institut de Biologie Moléculaire et Cellulaire, CNRS, 15 Rue René Descartes, 67084 Strasbourg, France, and [d]Institute of Mathematical Problems of Biology, Russian Academy of Sciences, 142290 Pushchino, Moscow Region, Russia

‡ Current address: Helmholtz Zentrum München, Institute for Biological and Medical Imaging, Building 56, Ingolstädter Landstrasse 1, D-85764 Neuherberg, Germany.

Correspondence e-mail: sacha@igbmc.u-strasbg.fr

Received 10 October 2008
Accepted 16 March 2009

Molecular replacement can fail to find a solution, namely a unique orientation and position of a search model, even when many search models are tested under various conditions. Simultaneous use of the results of these searches may help in the solution of such difficult structures. A closeness between the peaks of several calculated rotation functions may identify the model orientation. The largest and most compact cluster of such peaks usually corresponds to models which are oriented similarly to the molecule under study. A search for the optimal translation may be more problematic and both individual translation functions and straightforward cluster analysis in the space of geometric parameters such as rotation angles and translation vectors may give no result. An improvement may be obtained by performing cluster analysis of the peaks of several translation functions in phase-set space. In this case, the Fourier maps computed using the observed structure-factor magnitudes and the phases calculated from differently positioned models are compared. Again, as a rule, the largest and the most compact cluster corresponds to the correct solution. The result of the updated procedure is no longer a single search model but an averaged Fourier map.

1. Introduction

The molecular-replacement procedure (reviewed by Rossmann & Arnold, 2001) works with a known model that is similar to the unknown structure. In contrast to other phasing methods, the method not only gives a set of phase values but also directly gives a starting atomic model which is subsequently improved and refined. To find the solution, the position of the search model (its orientation and the coordinates of its centre in the unit cell) is varied. For each of these positions, the magnitudes of structure factors that are calculated from the search model are compared with the experimental data. The molecular replacement is based on the assumption that the calculated magnitudes are maximally similar to the experimental data when the model is close to the structure in the unit cell. In practice, the similarity of the magnitudes can be expressed in multiple ways using deterministic (for example, comparison of Patterson maps or their peaks) or statistical approaches (likelihood maximization) and can be used to identify the optimal model position. When the optimization problem has been solved, the phases of the corresponding calculated structure factors are used as an approximation to the unknown values. The experimental magnitudes associated with these phases can be used to calculate various maps. The search for the optimum of the

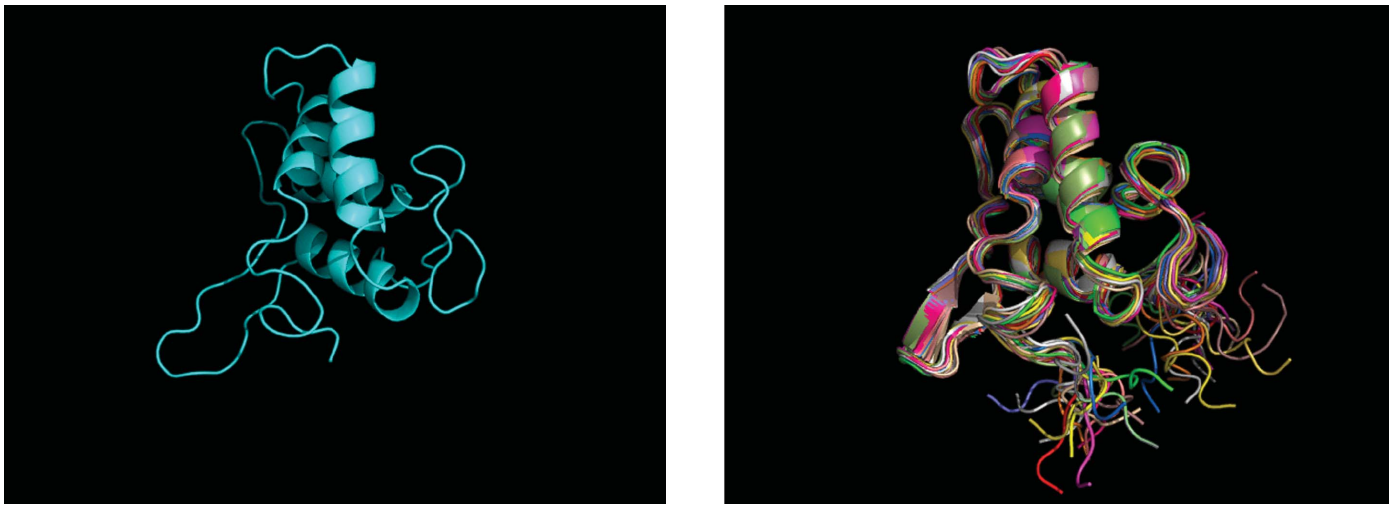


Figure 1
Ribbon view of (*a*) the CHFI final structure and (*b*) 20 superimposed NMR models used for molecular-replacement searches (these and other molecular images were produced using *PyMOL*; DeLano, 2002).

target function may be performed either directly in six-dimensional space or subsequently in three-dimensional rotational and in three-dimensional translational space (to simplify the presentation, we only discuss molecular replacement with a single independent molecule in the asymmetric unit).

However, when the search model is incomplete or significantly different from the structure, the method often fails to solve the phase problem and it is not possible to obtain an atomic model. For such a search model, the main molecular-replacement assumption is no longer justified and the global optimum of the target function generally corresponds to an incorrect position of the model. Therefore, improvements in the optimization procedures (*e.g.* Kissinger *et al.*, 1999), while extremely important in general, do not solve this problem. The use of low-resolution data which are less sensitive to model errors (Urzhumtsev & Podjarny, 1995; Fokine *et al.*, 2003) may either be insufficient to hide model imperfections or too strong to lose the features of the model. Different procedures for automatic building of new and more appropriate search models (see, for example, Suhre & Sanejouand, 2004; Keegan & Winn, 2007; Lebedev *et al.*, 2008) have led to important progress. Recent advances in maximum-likelihood-based procedures (Read, 2001; Storoni *et al.*, 2004), which statistically take into account model imperfections, have significantly extended the limits within which molecular replacement remains efficient.

Another possibility is to change the molecular-replacement strategy itself. For a search model of poor quality, the optimal model position may not correspond to the global optimum of the target. The peak indicating the correct position generally exists, but is weak. Variation of the model and the target (for example, resolution of the rotation and translation searches) changes the search results; however, the peak for the solution often remains the same for all the searches. As a consequence, one may expect to identify the solution by this persistence of the signal.

When looking for the 'optimal position', one further point is important. For poor models (*i.e.* those that differ significantly from the structure under study), the notion of the 'optimal position' may not be well defined in the usual geometric terms. For example, a model in one position may correspond better to one molecular domain and the same or a different model in a second position to another. In such cases, comparison of Fourier maps may be a more appropriate measure of the closeness of solutions than a straightforward comparison of translation parameters.

2. Multiple rotation function

2.1. Basic definitions

Finding a good model orientation is a necessary condition for success with conventional consecutive molecular replacement. Often, and especially for difficult cases, a single rotation function does not give an answer and the search is therefore repeated under different conditions and with different models. It can occur that neither of these finds the correct molecular orientation. At the same time, the answer may be indicated by the most persistent orientation when the results of several such searches are available and the peaks of these rotation functions are taken together (Urzhumtsev & Urzhumtseva, 2002). This persistent signal can be recognized by cluster analysis in rotation-angle space. In the procedure *COMPANG* developed previously for this goal, the distance between two orientations $p_{\rm m}$ and $p_{\rm n}$ is defined as the corresponding effective rotation angle from $p_{\rm m}$ to $p_{\rm n}$. Obviously, for nontrivial space groups all symmetry-related orientations should be taken into account. An important factor in the cluster-analysis procedure is the definition of the distance between two clusters. *COMPANG* defines it as the minimal distance between all one-to-one orientations, one from each cluster.

2.2. Application of multiple rotation function

To test the multiple rotation-function analysis, we applied it to the experimental data of corn Hageman factor inhibitor (CHFI; Behnke *et al.*, 1998) as follows. The solution of this structure by molecular replacement was previously reported to be difficult (Chen *et al.*, 2000). We briefly introduce this example here (for details, see Urzhumtsev & Urzhumtseva, 2002) because its results are used below for the translation searches. In *AMoRe* (Navaza, 1994), the rotation function was calculated with the default protocol for each of 20 available NMR models (Strobl *et al.*, 1995; Fig. 1) using experimental structure-factor magnitudes $\{F_{\rm obs}\}$. These calculations were repeated in different resolution zones. None of these functions, taken one by one, allowed the identification of the correct model orientation.

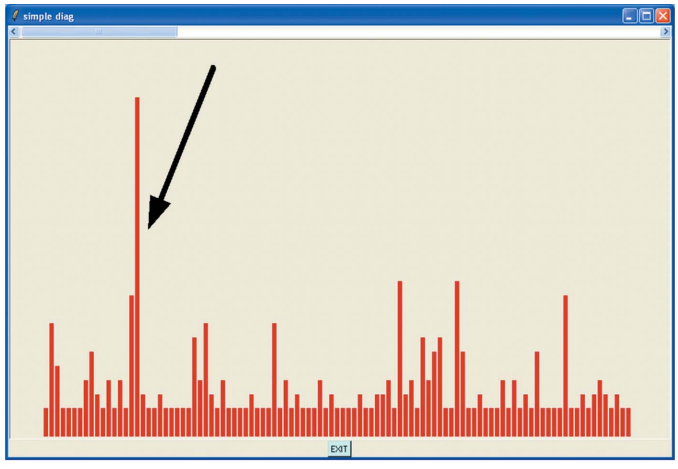
The 30 highest peaks from each of the 20 rotation functions were then selected, taken together and studied by cluster analysis in the space of model orientations (§2.1). An angular cutoff level defines whether two orientations (or their smaller clusters) are considered to be a single cluster or not; this parameter was varied with the tests. The analyzed feature was the relative size of different clusters. Fig. 2 shows the results of the cluster analysis for rotation functions calculated at a resolution of 4–10 Å. The largest cluster always indicates the correct orientation. For large cutoff levels (8° and larger) the noise peaks start merging and the signal decreases. With too small cutoff levels (smaller than 1–2°) no significant clusters can be seen. When rotation functions are calculated at lower resolution, for example at 5–10 Å, the model orientations are defined less accurately and the signal also becomes lower.

3. Multiple solutions in the translation problem

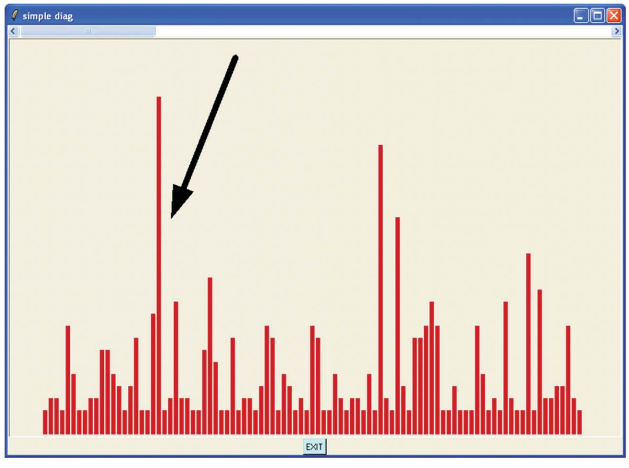
3.1. Conventional translation functions

Translation functions for CHFI were calculated with *AMoRe* (Navaza & Vernoslova, 1995). A straightforward molecular-replacement search in the default mode with the top peaks from the rotation functions gave no result for each of the 20 NMR models (different combinations of the resolution for the rotation and translation functions were tried).

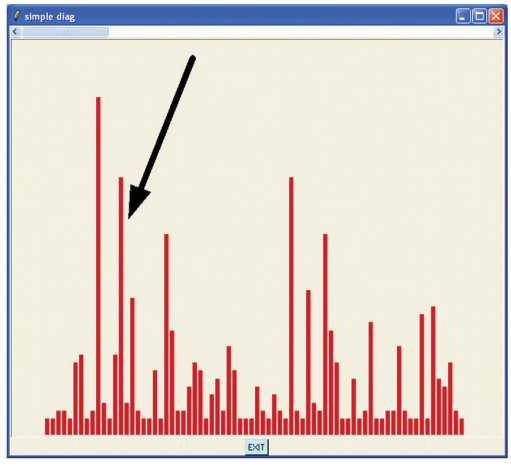
The translation search was then performed at a resolution of 5–15 Å with three groups of models. In the first test, which was performed as a control check, the search NMR models were in the best possible orientations as found by their optimal superimposition with the known answer. In the second test, the model orientations were taken from the best cluster in the multiple rotation function calculated at a resolution of 4–10 Å and with a relatively low cluster cutoff level equal to 4.0° (§2). The third test was similar to the second test but the orientations were taken from the set of rotation functions calculated at resolution 5–10 Å and the cluster cutoff was relatively high at 6.5°, so that the model orientations were less accurate.



(*a*)



(*b*)



(*c*)

Figure 2
Size of clusters of similar model orientations for 20 NMR models of CHFI. Each bar represents one cluster; its height is proportional to the cluster size. A set of the highest peaks of 20 rotation functions is analyzed together. All functions were calculated in the resolution zone 4–10 Å. The three images correspond to a different choice of the angular cutoff distance that defines the separation of clusters: (*a*) 4.0°, (*b*) 6.5°, (*c*) 8.0°. The correct model orientation belongs to the cluster indicated by the arrow.

In all three tests the translation function had a long list of peaks of roughly similar height. The model with the best value of the search criterion (the correlation between the observed structure-factor magnitudes and those calculated from the model) was distant from the correct solution. Some translation peaks did correspond to a model position close to the solution; for example, such a peak was among the top peaks for the models in the optimal orientations (the artificial situation of test 1). However, it was not easy to identify these peaks in the lists, especially in tests 2 and 3 with approximate model orientations.

3.2. Translation searches and multiple peaks

We supposed that if individual translation functions fail to find the solution then a simultaneous analysis of several of them could help, similar to the multiple rotation-function approach. Unfortunately, a direct comparison of translation peaks by closeness of atomic positions was inefficient (we have previously tried numerous variants of this method). There are several reasons that may explain this. For different models taken in different orientations, such a measure is not always straightforward. A different choice of the origin and the presence of symmetry-related molecules cause further confusion. More importantly, search models may fit the electron density of the crystal under study in different ways.

An example is RNA molecules with a pseudo-helical symmetry that are often packed in 'columns' and for which the corresponding electron-density maps at medium and low resolution show continuous helices. For such crystals the rotation function predicts the direction of the helix well but not the model orientation around it. Models oriented differently around the helix can be inserted relatively well into the continuous helical density, giving equally persistent peaks of similar size in the translation function (see, for example, Ogihara *et al.*, 1997; we made the same observations when

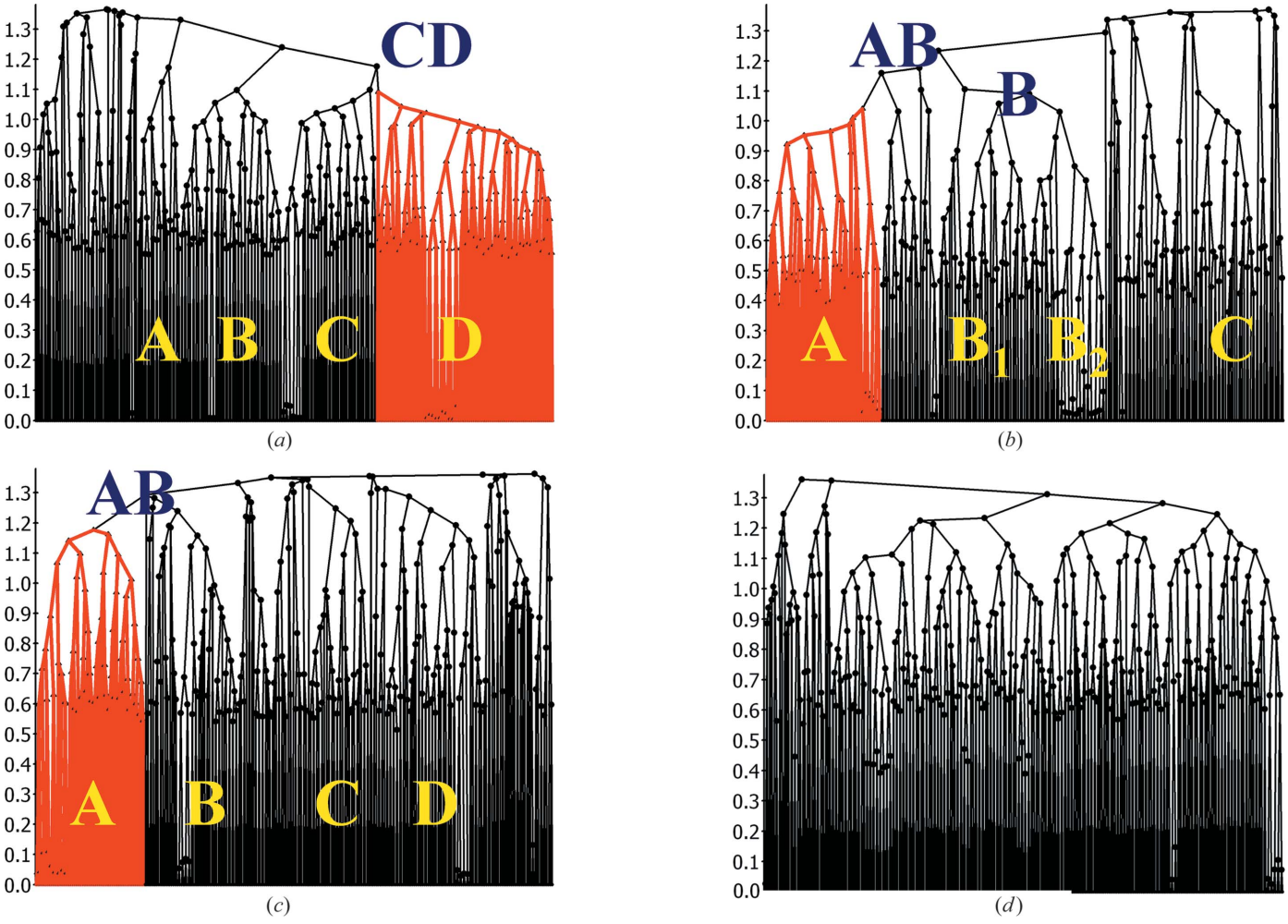


Figure 3
Cluster trees for the phase sets obtained after multiple translation-function analysis with the 20 NMR models of CHFI. The selected (correct) cluster is highlighted in the trees in (*a*)–(*c*). Letters indicate individual clusters or their groups as referred to in Table 1 and in the text. The translation search was performed with four different groups of model orientations. (*a*) Models in the best possible orientations. (*b*) Models in approximate orientations obtained from the multiple rotation-function analysis at 4–10 Å with the clusters selected with cutoff 4.0°. (*c*) Models in less accurate orientations obtained from the multiple rotation-function analysis at 5–10 Å with the clusters selected with cutoff 6.5°. (*d*) Models from a wrong orientation cluster. No compact and large cluster can be identified.

Table 1
Characteristics of the principal clusters of the cluster trees shown in Fig. 3.

The correct cluster is shown in bold.

Cluster	A	B	C	D	AB	CD	All
Test 1: ideal orientation							
N_{variants}	26	61	57	**106**	—	163	312
CorrP	0.15	0.33	0.38	**0.67**	—	0.64	0.60
⟨FOM⟩	0.54	0.53	0.53	**0.57**	—	0.49	0.31
Test 2: good orientation							
N_{variants}	**61**	74	33	—	179	—	273
CorrP	**0.62**	0.33	0.05	—	0.54	—	0.52
⟨FOM⟩	**0.61**	0.56	0.61	—	0.42	—	0.29
Test 3: imprecise orientation							
N_{variants}	**65**	50	56	61	120	—	308
CorrP	**0.55**	0.30	0.08	0.13	0.52	—	0.51
⟨FOM⟩	**0.49**	0.47	0.32	0.35	0.35	—	0.17

solving several crystals of the ribosomal decoding A sites using data provided by J. Kondo and E. Westhof and making further complementary tests).

These considerations show that at the medium and low resolution typical for molecular replacement the presence of multiple peaks in the translation (and rotation) functions may be natural for the problem, especially when searching using models of poor quality. These multiple peaks may differ significantly if we compare them in terms of geometric parameters. However, they become quite close if we change the measure and compare them by the similarity of the electron density that the corresponding models produce. This completely changes the type of output of the molecular replacement. Traditionally, molecular replacement results in an (atomic) model in a particular position; as a consequence, one obtains a trial set of atomic coordinates (that may be incomplete and with significant errors) and not only the structure-factor phases calculated from this model. Molecular replacement with electron density or with envelopes (see, for example, Urzhumtsev & Podjarny, 1995) does not generate an atomic model but still results in a single position of the search object in the unit cell. This new strategy suggests that in difficult cases we abandon the idea of identifying a single position of the search object (this unique position may simply be undefined for models that differ significantly from the structure) and look only for a phase set. This phase set is used, together with the experimental structure-factor magnitudes, to calculate a map which is then interpreted as in other phasing methods.

3.3. Multiple translation searches

Developing this suggestion, we generated the models for a relatively large number of the highest peaks of all translation functions taken together. For each of the models translated by the vector $\mathbf{t}$, we calculated its structure factors $\{F_{\text{mod}}(\mathbf{t}) \exp(i\varphi_{\text{mod}}(\mathbf{t})\}$. For the computed phase sets, we found their optimal alignment over all possible choices of the unit-cell origin $\mathbf{u}$ for the given space group (Lunin & Lunina, 1996). This alignment was performed using the correlation

$$\text{CorrP}(\mathbf{t}_1, \mathbf{t}_2) = \max_{\mathbf{u}} \frac{\sum_{\mathbf{s}} F^2_{\text{obs}}(\mathbf{s}) \cos[\varphi_{\text{mod}}(\mathbf{s}, \mathbf{t}_1) - \varphi_{\text{mod}}(\mathbf{s}, \mathbf{t}_2 + \mathbf{u})]}{\sum_{\mathbf{s}} F^2_{\text{obs}}(\mathbf{s})}$$

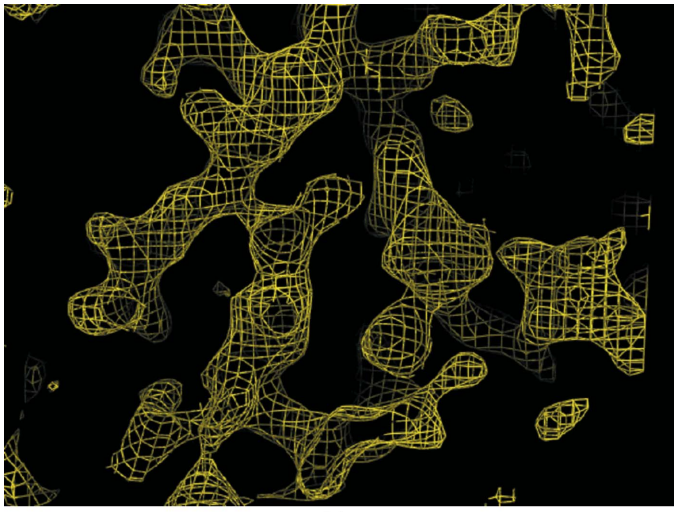
of maps (Lunin & Woolfson, 1993) calculated with the experimental structure-factor magnitudes $\{F_{\text{obs}}\}$ and generated phases $\{\varphi_{\text{mod}}\}$. The models relevant to the correct solution should reproduce the correct density more or less well and therefore should have close phase sets (the phase sets are close to the same unknown phase set $\{\varphi_{\text{exact}}\}$ and thus are close to each other). Similarly to the multiple rotation-function approach, we supposed that this solution gives a persistent signal among a large number of randomly distributed noise peaks. The persistence is measured not by the closeness of the atomic coordinates to each other, but by the map correlation CorrP. The cluster analysis identifies groups of close phase sets and the group for the correct solution is expected to be the largest group (many translation functions contribute to it) and the most compact. The compactness of the cluster may be characterized by its mean figure of merit (FOM). The larger the FOM, the more compact the cluster. When the best cluster has been identified, the resulting phase set is obtained by averaging individual phase sets inside this cluster (see, for example, Lunin *et al.*, 1990, 1995).

We started our tests from the easiest and, in practice, unrealistic case of the 20 NMR models in the best possible orientations (§3.1, test 1). All the highest peaks of the set of translation functions calculated previously were taken together; a set of structure factors was calculated for each of them using the corresponding models. Clustering of the calculated phase sets resulted in the tree shown in Fig. 3(*a*) with one cluster, marked D, being much more compact (with a low summit) and larger than other clusters (Table 1). Indeed, it corresponds to the correct solution.

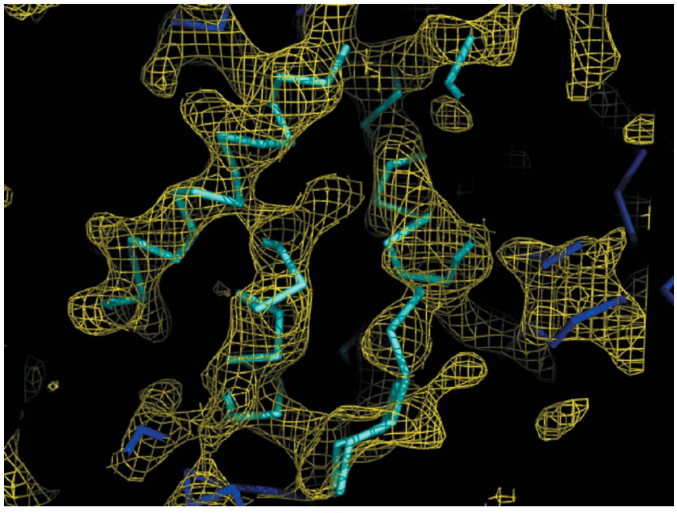
The situation was similar when studying the translation functions for the orientations in the cluster of approximate orientations (test 2; rotation functions at 4–10 Å, cutoff 4°). Here, cluster A is a single significant cluster at the level chosen (Fig. 3*b*). Cluster B is larger than A, but it is formed at a higher level and its components B_1 and B_2 are smaller than A.

The choice of a reasonable cluster is also possible for the more difficult scenario of the set of relatively poor model orientations which was performed in test 3 (rotation functions at 5–10 Å; cluster cutoff 6.5°). Here, the signal is slightly weaker (the cluster is less populated and less compact) but it leaves no ambiguity in the choice of cluster A as the solution (Fig. 3*c*).

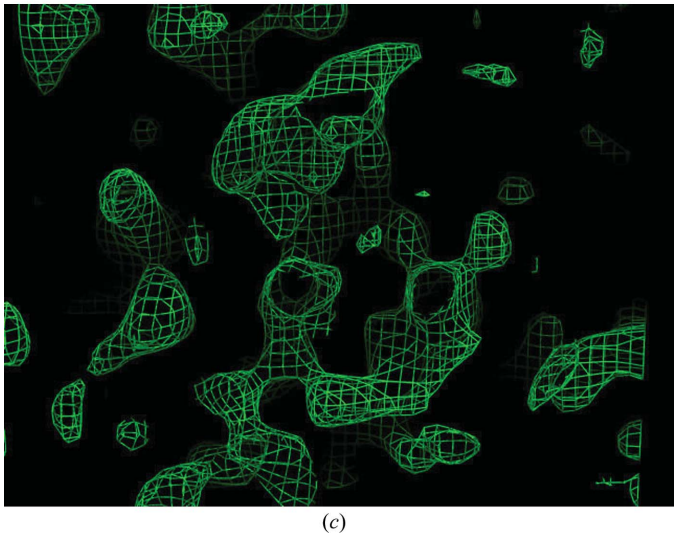
Several remarks can be made. Firstly, in all three cases the average phase set has a correlation to the exact values that is slightly higher than that for any individual phase set of this cluster. This can be compared with the widespread procedure of averaging results of several experimental measurements to obtain a best estimation of some value. Secondly, increasing the cluster size up to some level does not really decrease the phase quality (Table 1). This means that the contributions of a few extra phase sets mutually cancel and that in practical applications there is a certain freedom to choose the cluster.



(*a*)



(*b*)



(*c*)

Figure 4
(*a*) 5 Å resolution Fourier map calculated with experimental structure-factor magnitudes and phases from the correct cluster A (test 2; Fig. 3*b*). Continuous density is seen, especially for the helices. (*b*) The same map superimposed with the main chain of the correct model. (*c*) A similar image calculated with the phases from cluster B, an incorrect cluster from the same cluster tree.

However, when the 'height' of the cluster is increased, the mean FOM falls and the confidence in the result decreases.

Map calculation is an extra check for the quality of the choice of the cluster and the corresponding phase set. For example, if in test 2 one chose cluster B by mistake and not cluster A as the solution to the translation problem, its map does not show a 'protein-like' image (Fig. 4*c*), in contrast to the correct case (Fig. 4*a*). Also, if an incorrect cluster has been chosen at the previous step of the rotation studies, the translation functions should not find a persistent signal and a cluster tree for the translation searches would be more or less uniform, as the test calculations show (Fig. 3*d*).

4. Discussion

This study indicates several points that deserve special discussion.

Firstly, many crystallographic procedures are in some way reduced to the optimization of a single function and the result is a global or, more often, an appropriate local minimum (or maximum) point. It may occur that an individual target is not selective enough to give a result in this way. For example, in molecular replacement a search model may be of poor quality and thus incapable of accurately reproducing the experimental structure-factor magnitudes even when it is 'optimally positioned'. In such a situation, it is crucial to take information from several runs into account and to search for a persistent signal and not for the global optimum.

Secondly, when looking for the persistence of the signal, an appropriate measure should be used. In particular, in macromolecular crystallography, when performing a search with geometric objects and using reciprocal-space targets, it may be better to express the closeness of the peaks of the search function (or closeness of the models) in terms of the similarity of corresponding structure factors or Fourier maps and not in geometric units such as distances and angles.

Thirdly, molecular replacement, which is formally considered as a phasing method, traditionally results directly in an approximate atomic model and not only in a set of phase values. A failure to find such a model means failure of the method. The new approach does not require that a single best model position is found, thus simplifying the task. As a price, the result of the method is simply a phase set and molecular replacement becomes more similar to other phasing methods. To obtain this phase set, several translation functions can be used simultaneously; the structure factors are calculated for all highest peaks together and then treated by a cluster procedure. Interestingly, an average phase set may be more precise than any of the individual phase sets.

The current report does not go further than a feasibility study and leaves a number of open questions.

We did not optimize the targets used for the rotation and translation searches, but took the simplest ones in their default mode (Navaza, 1994; Navaza & Vernoslova, 1995). Obviously, their specific use or the application of advanced tools (see, for example, Read, 2001; Storoni *et al.*, 2004) may simplify some structure solutions; nevertheless, this does not solve all

molecular-replacement problems and leaves room for our suggestions.

We did not optimize the strategy. The same kind of search for a phase set without the determination of a single model may be applied directly in six-dimensional space and not subsequently for rotation and translation, as in the *FAM* phasing strategy (Lunin *et al.*, 1995, 1998). The preference for one or the other probably depends on the practical situation, *i.e.* whether the risk of missing approximately correct orientations is high or not. [In the *FAM* method of *ab initio* phasing starting from low resolution, one generates a very large number of simplified models composed of a few large Gaussian scatterers (large pseudo-atoms). The models for which structure-factor magnitudes correspond relatively well to experimental data are selected and the phase values of their structure factors are kept. The selected phase sets are then processed together to obtain Fourier maps, while the individual models may have no meaning.]

We did not analyze how to extract the maximum information from the new type of search. Translation searches are performed at a particular resolution (for example 5 Å, as in our example). However, when the translation peaks are selected, the model phase sets can be calculated at any resolution, even one that is higher than that used for the translation. This may be crucial in order to succeed in further structure solution.

We also did not analyze whether this method may be useful for crystals with several independent copies of the same molecule. Conceivably, such a search could place roughly half of the models at the position of the first molecule and half at the position of the second molecule, thus solving both problems simultaneously, but complicating the cluster analysis.

Lastly, the analysis of a cluster tree (Fig. 3) and identification of the principal cluster are not always simple tasks. Some approaches to formalize this procedure should be developed.

Answering these and other questions will require complementary studies.

The authors thank N. Lunina and O. Sobolev for help with programs for cluster analysis, M. Sander for initial tests with multiple translation functions, E. Westhof and J. Kondo for experimental data of the ribosomal decoding sites A and D. Moras for his support of the project. This work was supported by RFBR grant 07-04-00137.

References

Behnke, C. A., Yee, V. C., Le Trong, I., Pedersen, L. C., Stenkamp, R. E., Kim, S.-S., Reeck, G. R. & Teller, D. C. (1998). *Biochemistry*, **37**, 15277–15288.

Chen, Y. W., Dodson, E. J. & Kleywegt, G. J. (2000). *Structure*, **8**, R214–R220.

DeLano, W. L. (2002). *The PyMOL Molecular Graphics System.* DeLano Scientific, San Carlos, USA. http://www.pymol.org.

Fokine, A., Capitani, G., Grütter, M. G. & Urzhumtsev, A. (2003). *J. Appl. Cryst.* **36**, 352–355.

Keegan, R. M. & Winn, M. D. (2007). *Acta Cryst.* D**63**, 447–457.

Kissinger, C. R., Gehlhaar, D. K. & Fogel, D. B. (1999). *Acta Cryst.* D**55**, 484–491.

Lebedev, A. A., Vagin, A. A. & Murshudov, G. N. (2008). *Acta Cryst.* D**64**, 33–39.

Lunin, V. Yu. & Lunina, N. L. (1996). *Acta Cryst.* A**52**, 365–368.

Lunin, V. Y., Lunina, N. L., Petrova, T. E., Urzhumtsev, A. G. & Podjarny, A. D. (1998). *Acta Cryst.* D**54**, 726–734.

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.

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

Navaza, J. & Vernoslova, E. (1995). *Acta Cryst.* A**51**, 445–449.

Ogihara, N. L., Weiss, M. S., Degrado, W. F. & Eisenberg, D. (1997). *Protein Sci.* **6**, 80–88.

Read, R. J. (2001). *Acta Cryst.* D**57**, 1373–1382.

Rossmann, M. G. & Arnold, E. (2001). *International Tables for Crystallography*, Vol. *F*, edited by M. G. Rossmann & E. Arnold, pp. 263–292. Dordrecht: Kluwer Academic Publishers.

Storoni, L. C., McCoy, A. J. & Read, R. J. (2004). *Acta Cryst.* D**60**, 432–438.

Strobl, S., Muhlhahn, P., Bernstein, R., Wiltscheck, R., Maskos, K., Wenderlich, M., Huber, R., Glockshuber, R. & Holak, T. A. (1995). *Biochemistry*, **34**, 8281–8293.

Suhre, K. & Sanejouand, Y.-H. (2004). *Acta Cryst.* D**60**, 796–799.

Urzhumtsev, A. & Podjarny, A. (1995). *Acta Cryst.* D**51**, 888–895.

Urzhumtsev, A. & Urzhumtseva, L. (2002). *Acta Cryst.* D**58**, 2066–2075.