Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Multiple rotation function

Alexandre Urzhumtsev* and Ludmila Urzhumtseva

LCM3B, UPRESA 7036 CNRS, Faculté des Sciences, Université Henry Poincaré, Nancy I, 54506 Vandoeuvre-lés-Nancy, France

Correspondence e-mail: sacha@lcm3b.uhp-nancy.fr

Received 22 February 2002
Accepted 18 September 2002

A simultaneous analysis of several rotation functions allows identification of the model orientation in situations when a single rotation function fails to find the answer. Multiple rotation functions can be obtained by the usual modification of the search model or by variation of the resolution at which the function is calculated. A specially suitable case is a search with several NMR models. When the orientation of the model becomes available, its position can be found much more easily; low-resolution data can help in such a search. Many difficult cases of molecular replacement can be solved by new tools discussed in the article.

1. Introduction

Molecular replacement (MR) is one of the key methods for determination of the three-dimensional structure of macromolecules by X-ray crystallography (Rossmann, 1972, 1990). This method becomes even more important with the development of structural genomics, when the number of known structures (potential search models) will increase drastically, thus making it easier to resolve new ones. This widespread method is based on the *similarity assumption*: the search model is sufficiently close to the model of the crystal under study (or to a large enough part of it) so that it best reproduces the experimental structure-factor magnitudes when it is placed at the correct position.

This best fit to the experimental data is estimated by some numerical criterion, a function of six parameters: three rotational and three translational. Originally, the search for the answer was performed in two consecutive steps, rotation and translation searches, thus essentially reducing the CPU time. The price for this separation of the problems is an eventual risk of missing the answer if a wrong model orientation is chosen corresponding to one of highest peaks according to a criterion for the orientation search, the so-called rotation function. Fast translation functions (Navaza, 1994; Navaza & Vernoslova, 1995) allow the checking of many model orientations in the translation search and drastically reduce this risk. Recently, procedures for a direct search in six-dimensional space became available (Chang & Lewis, 1997; Kissinger *et al.*, 1999). Nevertheless, such perfection of the optimization does not resolve the problem when the search model has a poor similarity with the molecule under study: in such a situation, the similarity assumption is not true and the global optimum of the search criterion does not correspond to the correct model position. In other words, these new search techniques improve the minimization procedure but not the criterion of the search.

Similar difficulties appear when the search model is composed of several rigid groups whose relative position is unknown. In this multibody problem, the search for the global minimum of the criterion is much more complicated although not impossible (see for example, Glykos & Kokkinidis, 2000, 2001), but its success is also based on the correctness of the similarity assumption.

An essential feature of MR which has been used occasionally is that, as a rule, not a single but several search models are available and a variety of parameters and data sets is involved. For example, the diffraction data can be selected at different resolution limits to calculate rotation and translation functions (Urzhumtsev & Podjarny, 1995) or to compare rotation functions under different conditions. Similarly, several models, obtained by different modifications of the same model or determined by the NMR method, can be used in a search. Simultaneous use of this information creates a basis for more robust procedures.

In the following, we discuss how this variety of data can be used to solve some difficult cases of MR. In particular, we show how a simultaneous use of multiple rotation functions allows the correct model orientation to be found for poor models.

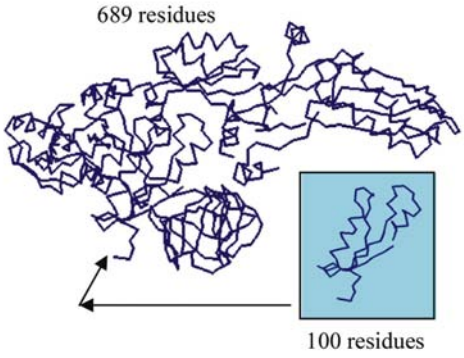
2. Rotation function and multiple models

2.1. Traditional approach

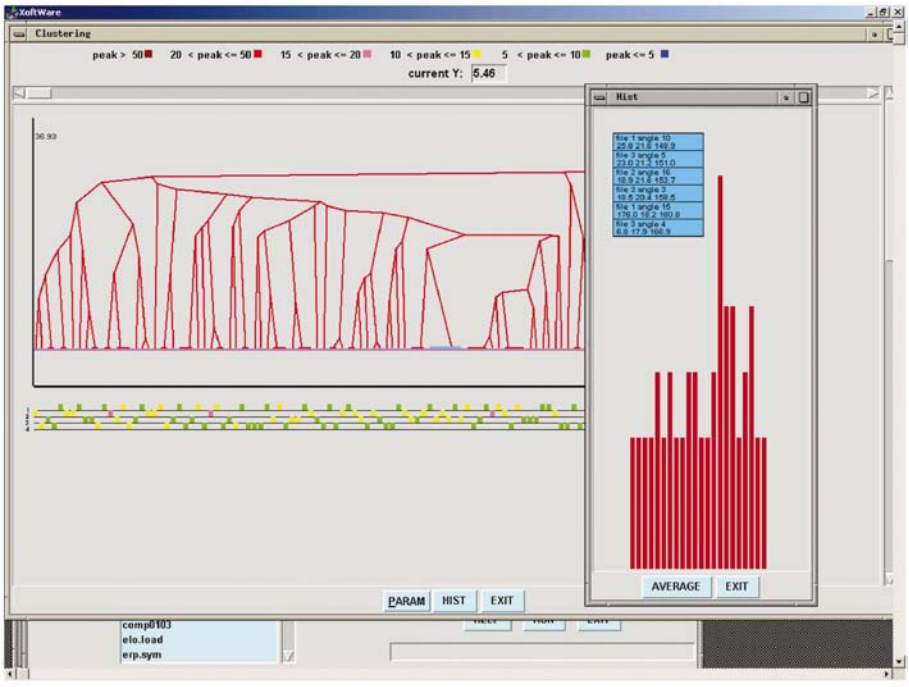
Traditionally, when the rotation function does not give an evident solution, the calculations are repeated varying the model (the whole model, the main-chain model, the C^{α} model, a model with deleted loops *etc.*), the set of structure factors (for example, varying the resolution) or the parameters (for example, the integration radius). The traditional goal is to find a combination of the parameters – model and data – such that it provides the researcher with a clear peak in the rotation function. Modern computers and programs such as *AMoRe* (Navaza, 1994) do not need this peak to be the highest one if it can be identified later by the translation search (when the similarity assumption is held). Unfortunately, this search does not always provide an unambiguous answer.

A case of a special interest, molecular replacement with NMR models, has recently been reviewed (Chen *et al.*, 2000; Chen, 2001). In this situation, a large number of atomic models, about 20–30, are available for the searches with no *a priori* preference for any of them; at the same time, the quality of these models is not always sufficient to find the solution.

Numerous difficult cases which cannot be resolved by traditional MR suggest that the search strategy should be changed when the similarity assumption is not held. In particular, this means that one can (i) use not the same but different functions for different problems (*e.g.* traditional separated analysis of the rotation and translation problems



(*a*)



(*b*)

Figure 1
Multiple rotation function: application to experimental data of EFG. (*a*) Search EFG model in comparison with the complete model. (*b*) Copy of the screen during the session of the program *COMPANG* when comparing several rotation functions for the EFG N-terminal model (see text). Parallel lines with squares below the cluster tree show the rotation peaks in different rotation functions with their height indicated by colour; the peaks are reordered to simplify the tree representation. The merging of two close peaks into a cluster is represented by two intersecting lines in the tree; the height at which these lines intersect corresponds to the distance between the peaks; the closer the peaks, the lower the point of intersection. A variable cluster threshold $D_{\rm min}$ for which the cluster size is calculated is indicated by a pink line above the zero level. The diagram of the size of clusters is shown in the inserted window. The correct orientation corresponds to the largest cluster (shown in light blue in the cluster tree); it can be noted that the distance between this cluster and the highest rotation peaks (pink squares) is quite large, indicating these peaks to be spurious. Initial rotation angles (from the *AMoRe* or1.s files) corresponding to model orientation for this cluster are shown in the blue window.

using features of each of these steps), (ii) use not the same but different data sets for rotation, translation and rigid-body refinement steps and (iii) look not for the global minimum of a single function but for relatively good values of many of them.

2.2. Multifunctional approach and consistency assumption

If a strong signal in the rotation function does not appear with any combination of parameters, several rotation functions can be analysed together. The idea behind this approach is that if we use the same model or if the models are oriented similarly before the search, then the peak for the correct model orientation, when it exists, always has the same coordinates; at the same time, spurious peaks (or at least most of them) are distributed more or less randomly and vary with the model and with the set of reflections. In other words, the signal in the rotation function may be weak but *is consistent* among the functions.

In such an approach, when we try to reduce (eventually, to a single possibility) the number of the model orientations, MR becomes more powerful owing to the substitution of the similarity assumption by a much weaker *consistency assumption*: the model is (or the models are) good enough to systematically give a signal (not necessary strong) in the rotation functions.

A simple and fast procedure for comparison of several rotation functions has been developed. Its application to a number of 'difficult' data sets shows that such a substitution of the similarity assumption, of the search goal and of the search strategy (including improvements in the translation search which are also discussed below) indeed solves some difficult cases of the MR problem. An alternative method of conducting a search with several models, based on the maximum-likelihood methodology and realised in the frame of the program *BEAST*, has been suggested recently by Read (2001).

3. Multiple rotation-function analysis

3.1. Main principles

A manual comparative analysis of multiple rotation functions is difficult for several reasons, including the following.

(i) Close model orientations can correspond to two triplets of rotation (Eulerian) angles that are quite different at first sight.

(ii) The presence of symmetry operations complicates even further the determination of close orientations.

(iii) Preliminary reorientation of the model before the search by some programs such as *AMoRe* (Navaza, 1994) makes direct comparison of the lists of peaks from different rotation functions useless.

An algorithm to compare automatically several rotation functions has been developed. All peaks from available rotation functions are taken together and the orientations which are closer each to other than a chosen cluster threshold $D_{\min}$ are considered 'to be the same within the given limit'. The size of all such groups (clusters) of similar orientations can be calculated and shown in a diagram; the hypothesis is that the cluster corresponding to the correct orientation is the most populated and analysis of the size of clusters therefore allows the answer to be found.

3.2. Main steps of the procedure

The suggested algorithm compares several rotation functions. The results of each rotation search are provided as a list of rotation peaks. If several search models are tested, they must be superimposed before calculating corresponding rotation functions. The principal steps of the comparative procedure are the following.

(i) The joined list of peaks is prepared as a concatenation of peaks from individual rotation functions.

(ii) For each pair of the peaks in this joined list, a 'distance' between corresponding model orientations is calculated, taking symmetry operations into account.

(iii) In order to study the distribution of peaks in space, a clustering procedure is applied which takes the matrix of distances calculated at the previous step as the input; the results are represented by a cluster tree.

(iv) For a chosen angle value, the orientations which are closer to each other than this value are considered to be identical and the groups (clusters) of such close orientations are defined. Naturally, the number and the composition of clusters vary with this angle value (the cluster threshold).

(v) The number of peaks inside each cluster is calculated; the largest cluster has a high chance of corresponding to the correct model orientation.

(vi) The rotation parameters corresponding to the chosen cluster are provided and can be converted to the rotation matrix either using the same package or, for example, using *CONVROT* (Urzhumtseva & Urzhumtsev, 1997).

3.3. Technical aspects of the multiple rotation-function analysis

3.3.1. Distance between two orientations. The distribution of points in multidimensional space can be studied by clustering techniques which have previously been used in crystallography for the solution of the phase problem (Lunin *et al.*, 1990, 1995). Based on the closeness of the points, clustering techniques merge them into groups, called clusters. The key tools of such an analysis are the distance $D(p_m, p_n)$ between two points p_m and p_n and the distance $D(C_k, C_l)$ between clusters.

When a point p_m represents an orientation expressed, for example, by three Eulerian angles $(\alpha_m, \beta_m, \gamma_m)$, the distance between two such points is less natural than for points represented by their Euclidian coordinates and many definitions can be introduced. For example, a one-to-one correspondence can be established between an orientation and the projection of the corresponding radius-vector on a sphere of unit radius. The distance between two orientations can then be defined as the length of the shortest arc on this sphere between their projections. This length is equal to the effective

angle of rotation from one of these orientations to another and therefore this effective angle can be used by itself as a distance.

In practical terms, when $p_m = (\alpha_m, \beta_m, \gamma_m)$ and $p_n = (\alpha_n, \beta_n, \gamma_n)$ are two orientations found from a rotation function, the corresponding matrices M_m and M_n represent the rotation of the same initial model into these two orientations. As a consequence, the matrix of the rotation from the orientation p_n to the orientation p_m is calculated as the product $M_m M_n^{-1}$. If the distance between the orientations p_n and p_m is defined as the corresponding effective rotation angle, then

$$d(p_m, p_n) = \kappa = \arccos\{[\mathrm{trace}(M_m M_n^{-1}) - 1]/2\}. \quad (1)$$

It is obvious that $d(p_m, p_n) = d(p_n, p_m)$. The definition (1) is easily generalized for cases when the orientation is defined not by Eulerian angles but by any other parameters (for a list of definitions, see Urzhumtseva & Urzhumtsev, 1997). Substitution of the distance (1) by another definition changes the distance matrix and the cluster tree (see below). However, the cluster tree calculated with any other definition such that

$$d'(p_m, p_n) = f[d(p_m, p_n)], \quad (2)$$

where f is a monotonically increasing function, conserves the cluster topology which is the main subject of further analysis.

3.3.2. Role of symmetries. If the space group contains several symmetry operations R_k, $k = 1, \ldots, K$, each peak p_m in the rotation function represents a group of symmetrically linked orientations $R_k p_m$, $k = 1, \ldots, K$ and it is natural to define the distance between two peaks as the minimal distance value calculated for all symmetrically related pairs of orientations corresponding to these peaks,

$$D(p_m, p_n) = \min_{k,l}[d(R_k p_m, R_l p_n] = \min_k[d(R_k p_m, p_n)], \quad (3)$$

with the last equality owing to the group properties. It is important to note that if a non-crystallographic rotation is present in the crystal, its order and the axis direction can be defined independently using the self-rotation function. This non-crystallographic rotation can be also considered at the step of the distance calculation (3), allowing identification of the pairs of angles linked by this symmetry and enforcing the signal.

3.3.3. Clustering and the cluster tree. When the distance between all pairs of points (N points in total) is calculated, the clustering procedure searches for the two closest points and merges them into a cluster. This cluster acts further as a new point instead of two merged points, thus reducing by one the total number of points; however, the distance between a point and a cluster or between two clusters needs to be defined (see below). After the procedure is repeated $N - 1$ times, all points are merged into a single cluster. This process can be shown in a cluster tree. Initially, every point is represented by a node on the abscissa axis. Two points merged at a distance D are represented by two lines issuing from the corresponding nodes and intersecting at a height D. This new node with ordinate D represents the corresponding cluster and is the starting point for the line when this cluster is merged with another cluster. The points on the abscissa axis are usually reordered to avoid intersecting lines in the tree.

After the cluster tree is built, all points from a cluster below some level $D_{\min}$ can be considered as indistinguishable points at the precision $D_{\min}$.

3.3.4. Distance between two clusters. The clustering procedure needs to extend the definition of the distance (3) for clusters. If C_n is a cluster composed from the orientations $p_{n1}, p_{n2}, \ldots, p_{nK}$, a distance $D(p_m, C_n)$ between this cluster and the orientation p_m can be defined as

$$D(C_n, p_m) = D(p_m, C_n) = \min_k[D(p_m, p_{nk})]. \quad (4)$$

Alternative definitions are possible, for example

$$D'(p_m, C_n) = D(p_m, \langle p_{nk}\rangle), \quad (5)$$

where $\langle p_{nk}\rangle$ is the geometric centre of the cluster C_n. Generally speaking, cluster trees calculated with definitions (4) and (5) can have some differences in their topology and can lead to slightly different results. While the definition (5) is quite attractive, its use for clustering is more time-consuming than the use of (4) and the latter is used in the current version of the described procedure.

Similarly to (4), the distance between two clusters C_m and C_n is defined as the minimal distance between all pairs of orientations, one from cluster C_m and the second from cluster C_n,

$$D(C_m, C_n) = \min_i[D(p_{ml}, C_n)], \quad (6)$$

where $C_m = \{p_{m1}, p_{m2}, \ldots, p_{mL}\}$. The definitions (4 and (6) are used to recalculate the distance matrix at each step of merging two points or clusters into a new cluster.

3.3.5. Cluster study. After the cluster tree is calculated, some cluster threshold $D_{\min}$ is chosen (which can be varied by the user) and all orientations which are closer to each other than this threshold are considered to belong to the same cluster. In other words, they indicate the same model orientation within the chosen precision. The cluster tree is analysed from left to right and the size of the clusters found is represented in the same order in a diagram, the subject of further analysis. The goal is to find the most populated cluster which is believed to correspond to the correct answer. The coincidence (or closeness) of higher rotation peaks can be more significant that the coincidence of lower peaks; therefore, a contribution of every peak to the size of the cluster may be weighted, for example, by the height of the contributing peaks.

4. Solving difficult problems with multiple rotation function

4.1. General description of the tests

The described procedure has been tested first with a synthetic case and then with several cases where the structure could not be solved previously by conventional MR procedures or where such procedures had major difficulties. The case of NMR models is specially suited to such common analysis of several rotation functions. In the tests discussed below, the NMR models were taken as they are in the PDB

Table 1
Rotation-function analysis for the N-terminal end of EFG.

Peaks for the largest cluster are given. The correct solution is (27.6, 21.9, 148.3).

Resolution limits (Å)	Sequential No. of the peak close to the solution	Peak parameters (Eulerian angles) α, β, γ (°)	Height of the peak	Height of the 1st peak	Height of the 2nd peak
4–10	10	25.8, 21.6, 148.9	10.0	13.2	12.4
5–10	5	23.0, 21.2, 151.0	11.3	14.1	13.1
4–15	16	18.9, 21.6, 153.7	13.4	18.5	15.7
5–10	3	18.5, 20.4, 158.5	11.3	14.1	13.1
4–10	15	176.0, 18.2, 180.8	9.8	13.2	12.4
5–10	4	6.8, 17.9, 166.9	11.3	14.1	13.1

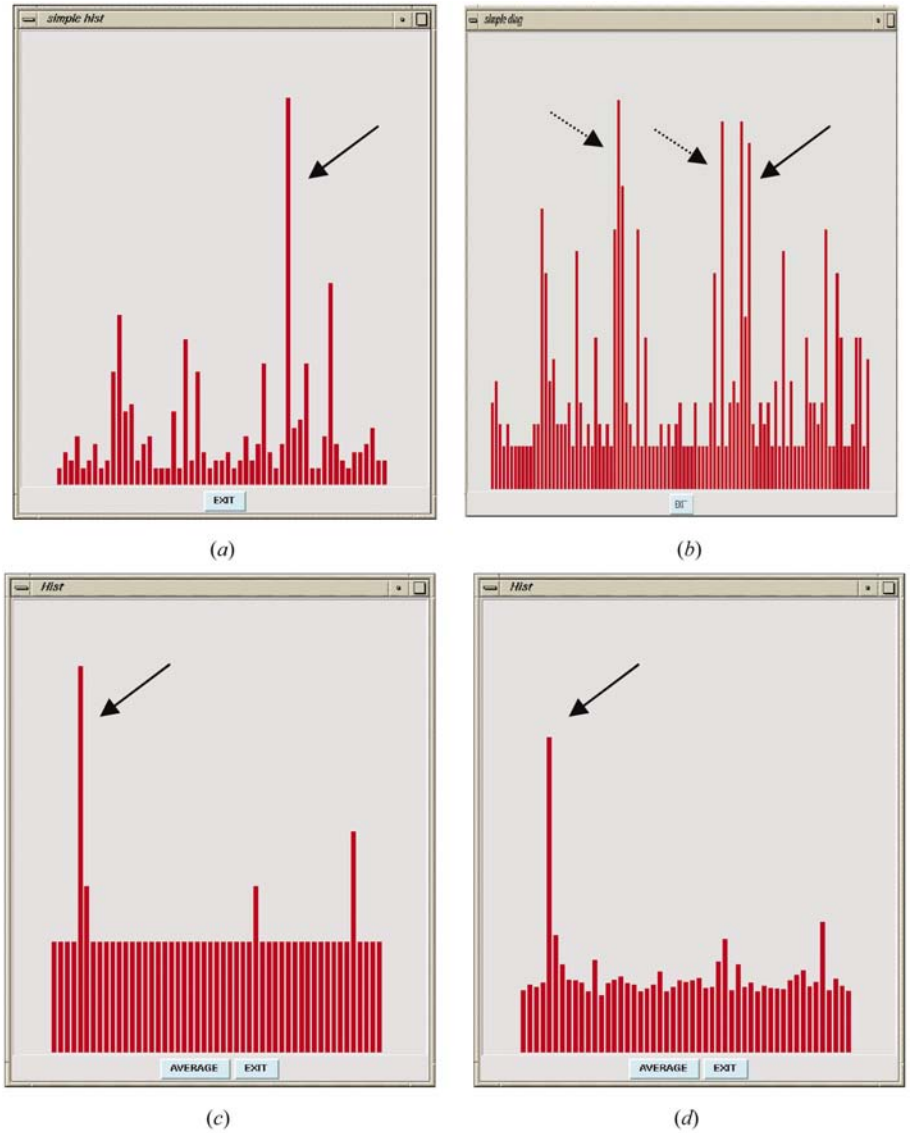


Figure 2
Multiple rotation function: application to the experimental data of E*r*-1 and of CHFI. In both cases 20 NMR models were used and in both cases the highest peak in the diagram of the cluster size corresponds to the correct orientation. (*a*) E*r*-1 rotation functions, cluster threshold 9°; the correct cluster is indicated by an arrow. (*b*) As (*a*) but the cluster threshold is equal to 5°; false peaks appeared and the correct cluster has been split into three close subclusters (marked with arrows). (*c*) CHFI rotation function; cluster size is calculated with unweighted peaks; the cluster for the correct orientation is indicated by an arrow. (*d*) As (*c*) but the peaks are weighted by their height when calculating the cluster size.

(Bernstein *et al.*, 1977), all temperature factors were assigned to be equal to 20 Å^2, no model modification has been performed and no optimal protocol (Chen *et al.*, 2000) applied. All rotation functions were calculated with *AMoRe* (Navaza, 1994).

In all examples reported below, the experimental structure-factor magnitudes have been used. The diffraction data for elongation factor G (subsequently referred to as EFG) were available to one of the authors (AU), who participated in the initial structural analysis of this protein (Lunin *et al.*, 1990; Chirgadze *et al.*, 1991; Urzhumtsev, 1991). The experimental data for the pheromone E*r*-1, corn Hageman factor inhibitor (subsequently referred to as CHFI) and thioredoxin *h* were kindly provided by the principal investigators of the corresponding projects.

4.2. Elongation factor G

In this first series of tests, a common situation was simulated where the model is very incomplete and does not give a strong signal in the rotation function. The N-terminal end (the first 100 residues from the 689 in the complete model; Fig. 1*a*) of a large protein, elongation factor G (Aevarsson *et al.*, 1994), was used as the search model. Corresponding crystals belong to space group $P2_12_12_1$ and have unit-cell parameters a = 75.6, b = 106.0, c = 116.6 Å. The rotation functions were calculated using the same model but varying the resolution range: 4–15, 4–10, 4–8, 5–10 Å. Individual rotation functions do not allow identification of the solution (Table 1). At the same time, joint analysis of the rotation peaks at a cluster threshold $D_{\rm min}$ of 5°, as described above, gives several clusters composed of two or three peaks, three clusters composed of four peaks and a single cluster composed of six rotation peaks (Fig. 1*b*). This latter, being notably larger than others, corresponds as expected to the correct orientation. This cluster is stable for a large range of the parameter $D_{\rm min}$. Table 1 shows that the closeness of some angles of this cluster is not evident at first sight when

Table 2
Translation search at 4–8 Å followed by rigid-body refinement for one of the NMR models of the E*r*-1 in the orientation defined by the multiple-function analysis as (116.2, 73.3, 209.9).

The correct orientation is (113.3, 77.2, 200.3) and the position is (0.3151, 0.0, 0.4892). Appropriate solutions are indicated *. Peaks 1 and 4 are immediately eliminated by packing considerations.

Peak No.	α, β, γ (°)	Molecular position	Correlation coefficient	Intermolecular distance (Å)
1	113.1, 77.9, 202.0	0.4260, 0.0, 0.4493	49.2	7.0
2**	110.8, 74.6, 207.7	0.3209, 0.0, 0.4936	37.5	14.6
3*	114.2, 76.6, 203.6	0.3823, 0.0, 0.4902	35.4	12.9
4	113.7, 77.9, 204.8	0.4714, 0.0, 0.3263	30.7	7.1
5	112.8, 77.7, 207.6	0.0837, 0.0, 0.3635	27.7	13.2
6	113.0, 72.9, 210.2	0.2043, 0.0, 0.4097	26.8	12.4

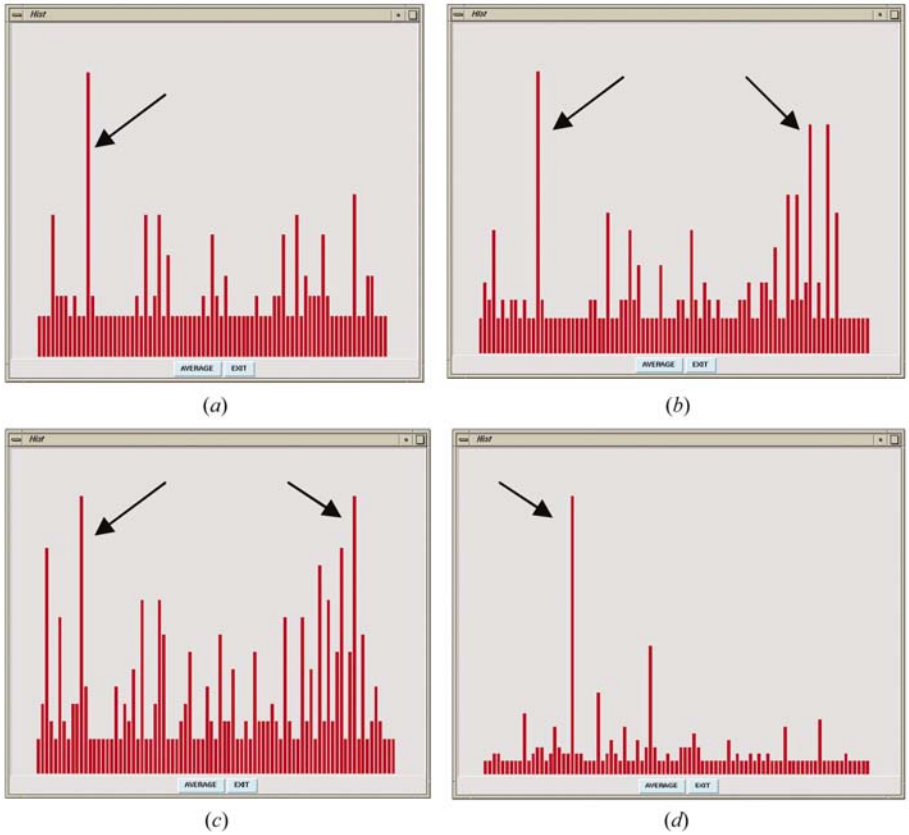


Figure 3
Multiple rotation function: application to the experimental data of thioredoxin *h*. The diagram of the cluster size for the rotation functions calculated with 23 NMR models at the resolution 4–15 Å. (*a*) Cluster threshold 3.5°; the highest peak, indicated by an arrow, corresponds to molecule *B*. (*b*) Cluster threshold 4.0°; the highest peak corresponds to molecule *B* and the peak for molecule *A* is the second from the right; peaks are indicated by arrows. (*c*) Cluster threshold 4.5°; the two highest peaks, indicated by arrows, correspond to molecules *B* and *A*. (*d*) The same as (*b*), but the non-crystallographic symmetry is included in the list of symmetry operations; the highest peak, marked by an arrow, corresponds to the orientations of the molecules *A* and *B* linked by this symmetry.

these angles are taken directly as they are in the rotation-function files.

4.3. E*r*-1 pheromone

The second series of tests was performed with experimental data from E*r*-1 (Anderson *et al.*, 1996), called by the authors 'a challenging case for protein crystal structure determination'. This example was identified as a limiting case for searches with NMR models, where the optimized protocol was crucial in finding the solution (Chen *et al.*, 2000). This protein is formed by three practically parallel α-helices and crystallizes, with a very dense packing, in space group *C*2, with unit-cell parameters a = 53.91, b = 23.08, c = 23.11 Å, β = 110.4°. While in the previous test with EFG a set of rotation functions was obtained with the same model but with different sets of data, in this and in all following tests the same data but different models were used.

Similarly to the previous report (Anderson *et al.*, 1996), neither of the rotation functions calculated with all NMR models (Mronga *et al.*, 1994) at resolutions of 3–8 Å and 4–8 Å highlight the solution. In fact, the orientations close to the correct one were in the list of rotation peaks and the translation functions calculated with them also contained the correct position; however, it was not possible to recognize the answer among wrong variants with similar or even better values of the criteria.

Rotation functions calculated at 3–8 Å resolution for all 20 NMR models taken directly from the PDB (Bernstein *et al.*, 1977) were studied. Multiple analysis with these functions shows several clusters when the cluster threshold $D_{\rm min}$ is about 8–9° or higher; one of these clusters is much larger than others (Fig. 2*a*) and corresponds to the correct answer. When the cluster threshold $D_{\rm min}$ decreases to 5°, this cluster splits into three subclusters (Fig. 2*b*), probably owing to a slight rearrangement of the three helices in the models. In this case, the interactivity of the cluster choice was much more important than for EFG.

If the orientation of the model is chosen from the major cluster (the NMR model first in the PDB list was taken in the orientation corresponding to its peak in the cluster), the standard translation function calculated at 4–8 Å resolution and the intermolecular distance allows one immediately to identify the solution (Table 2); on the contrary, a 'brute' *AMoRe* translation search calculated for all rotation peaks completely hides this signal.

4.4. Corn Hageman factor inhibitor

Corn Hageman factor inhibitor (Behnke *et al.*, 1998) can be considered to be the most difficult case for molecular-replacement searches with NMR models; the solution could

not be found even with the optimal protocol (Chen *et al.*, 2000). This protein crystallizes in space group $P4_22_12$, with unit-cell parameters $a = b = 57.12$, $c = 80.24$ Å.

A joint analysis of the rotation functions calculated by *AMoRe* in the resolution range 4–15 Å with all 20 NMR models taken from the PDB (Strobl *et al.*, 1995) gives an extremely strong signal corresponding to the correct solution (Figs. 2*c* and 2*d*) when $D_{\min}$ varies between 2 and 4°. For higher values of this parameter, alternative clusters appear. The quality of the models is not good enough; the correlation-coefficient translation search at 4–15 Å resolution gives a peak outside the first ten and therefore does not allow the correct answer to be recognized. However, a search with the same function but calculated with all available low-resolution reflections and bulk-solvent contribution taken into account gives the highest correlation for the correct model position (Fokine & Urzhumtsev, 2002).

For a comparison, cluster analysis was performed when the peaks were weighted by their height. For these particular data, the weighting slightly increases the relative contrast of the signal but does not change the result qualitatively.

4.5. Thioredoxin *h*

One further series of tests was performed using experimental data from thioredoxin *h* from *Chlamydomonas reinhardtii* (Menchize *et al.*, 2001). This example differs from the previous two by the presence of two molecules in the asymmetric unit, thus allowing study of the usefulness of non-crystallographic symmetry in multiple rotation-function analysis.

Thioredoxin *h* crystallizes in space group $P3_121$, with unit-cell parameters $a = b = 49.45$, $c = 45.31$ Å. An attempt to solve this structure by conventional MR using the 23 available NMR models (Mittard *et al.*, 1997) failed. In contrast, clustering of peaks of the rotation functions, calculated in the standard resolution range 4–15 Å, is quite efficient. At the cluster threshold $D_{\min} = 3°$ one of the clusters is much larger than others; this cluster corresponds to the orientation of molecule *B*. The cluster for molecule *A* is small (Fig. 3*a*), but when $D_{\min}$ is increased to 4–5° it becomes the second in size (Figs. 3*b* and 3*c*), while the cluster for molecule *B* continues to be the largest. Naturally, such an increase in the cluster threshold merges more and more wrong orientations and the signal for the first peak (molecule *B*) loses its contrast.

When the known non-crystallographic symmetry is included into the list of symmetry operations, the initial cluster tree at $D_{\min} = 4°$, where the choice of the orientation of molecule *A* is slightly ambiguous (Fig. 3*b*), is replaced by another tree with a single dominating cluster which simultaneously shows the orientation of both molecules (Fig. 3*d*).

4.6. Role of the cluster threshold

The described examples show the key role of the cluster-threshold parameter in the determination of clusters. Interactivity in the choice of $D_{\min}$ is very important because it allows one to follow the variation of clusters with this parameter, for which no universal value can be recommended.

The examples studied show differing dependence of the results on the cluster threshold. For EFG, the cluster is the largest over quite a large threshold interval. For E*r*-1, at $D_{\min} = 5°$ there are two false clusters of a similar size to two clusters close to the correct solution and an unambiguous choice of the solution is impossible (although several tens of possibilities were reduced to only three variants including the correct one). At 9° three close subclusters merge to a single cluster of a much larger size than any other (Figs. 2*a* and 2*b*). This allows it to be chosen as the answer and further checks to be made on the rotations from both its largest subclusters, which in any case are quite close to each other. For CHFI, the situation is inverted. At a larger threshold the tree gives three large clusters, while at $D_{\min} = 2$–4° the correct cluster is the largest, with high contrast. Finally, thioredoxin *h* is a mixed case in which the peak for one molecule is very stable at large threshold limits and the cluster for the second molecule is significant only at about 4°. However, in this case the use of non-crystallographic symmetry allows a single cluster, very stable with respect to modification of $D_{\min}$, to emerge.

In conclusion, an interactive way of working is important for the procedure described above. The continuous variation of $D_{\min}$ allows one to identify the branches of the cluster tree which contain large clusters. A visual analysis of the tree gives an idea of the threshold values at which the clusters and their size can be analysed. In general, a cluster study with a $D_{\min}$ value of between 3 and 5° seems to give a good signal and this cluster threshold can be suggested as a starting value for interactive analysis. Excessive increase of $D_{\min}$ allows noise peaks to be merged and to produce a spurious signal. On the other hand, if the model is composed of several domains and they are slightly rearranged in the search model in comparison with the answer, increasing $D_{\min}$ to larger values can allow merging of peaks corresponding to optimal alignment of individual domains.

In favourable cases, such a choice of $D_{\min}$ indicates a single cluster around the solution, as was observed in all the reported tests. In the most difficult cases it can happen that a few clusters are obtained and not a single one; in any case, even with a non-optimal choice of $D_{\min}$ the number of possible solutions after the clustering procedure is very small (2–3), the correct orientation was among them in all tests (some further examples are not described here) and the variation of $D_{\min}$ allows the answer to be chosen, for example, by the stability of the corresponding cluster.

4.7. Peak weighting

A weighting of rotation peaks by their height was studied. While the idea that the merging of higher peaks is more significant than the merging of small peaks seems to be correct and slightly increased the signal for CHFI, in practice such cluster analysis is required in difficult cases in which the rotation peaks are rather small in comparison with spurious peaks and such weighting may work against the final goal.

Molecular-replacement packages usually produce a limited list of rotation-function peaks, selected by their height or by their contrast with the highest peak. Such a selection can be considered as a weighting of the peaks by a step function. A similar filtration can be further performed when the procedure starts the cluster analysis. It is difficult to give a universal recipe for the number of peaks which should be included from each rotation function: too short a list will exclude the signal which is expected to be weak, while too long a list gives too many noise peaks and increases the chance of spurious clusters being formed by them. Rotation functions for the difficult cases discussed above contained the correct (or close to correct) orientation in the second or third dozen of peaks, suggesting inclusion of 30–40 peaks from each rotation function, as is performed by default in *AMoRe* (Navaza, 1994). In general, this limit is one more parameter, variation of which can help in difficult cases.

4.8. Use of non-crystallographic symmetry

In the presence of non-crystallographic rotation symmetry, cluster analysis can be used in various ways. While the order of the rotation is known exactly (except in some 'pathological' cases), the direction of the axis is known with a limited accuracy. When non-crystallographic symmetry with the correct direction of the axis is included in inter-peak distance calculations (4–6), one might expect that many peaks of the rotation function will be merged at a quite low cluster threshold, increasing the signal. If this does not happen, then the direction of the axis should be revised.

Several tests have been performed with thioredoxin *h* data in order to check the extent to which the direction of a twofold axis may be in error and yet still help to resolve the rotation problem. The exact rotation axis was found from the optimal superposition of two crystallographically independent molecules and an artificial error of 2, 4, 6 and 8° was introduced into the direction of the axis before its use in the multiple rotation-function analysis. These tests show that under the conditions described above, when a good signal appears at the cluster threshold of about 3–4°, the use of non-crystallographic symmetry with an error in the axis orientation of 4° or below increases the relative size of the cluster containing the correct answer. When the errors are larger, the use of non-crystallographic symmetry does not help, but also does not remove the signal, which exists in the cluster analysis without such symmetry.

Therefore, since an error within the 4° limit seems to be quite realistic for a well calculated self-rotation function, the general advice is to use this information for the multiple rotation-function analysis. If no significant change in the cluster tree happens in comparison with an initial cluster tree calculated without non-crystallographic symmetry, then the user is advised is to check the symmetry axis. Alternatively, several axes with the direction close to that from the self-rotation function can be tried and the direction which gives the highest 'reduction' of the tree can be considered to be the correct direction.

5. Consistency assumption and easier solution of MR problems

5.1. Principal reasons for MR failure

The consistency assumption (see §2.2), which requests much less from the search model (models) than the traditional similarity assumption (see §1), allows the reconsideration of MR. First of all, the separation of the search into two traditional steps, rotation and translation, can be advantageous for difficult situations and the knowledge of the model orientation plays a key role in the structure determination (this opinion is also supported by Glykos & Kokkinidis, 2001). Secondly, the current state of the art in low-resolution direct-phasing methods (for a review, see, for example, Lunin *et al.*, 2000) allows the molecular position in the unit cell to be found relatively easily and therefore the model to be positioned if its orientation is known.

Similarly to other crystallographic problems such as refinement of atomic models, the two main reasons for failure of the MR searches are the incompleteness of the model and too high errors in the relative arrangement of atoms inside the model (we suppose that the experimental data are measured correctly). A special case is the situation when an asymmetric part of the unit cell contains a large number of independent molecules whose position needs to be defined. These situations are discussed below one by one, taken as extreme cases.

5.2. Incomplete models

In this section it is supposed for simplicity that the search model corresponds exactly to part of the macromolecule under study. When the model is significantly incomplete, structure factors calculated from such a model have no reason to best fit the experimental data when the model is placed correctly (Afonine *et al.*, 2001).

However, it can be noted that for such an incomplete but exact model the rotation function calculated through the traditional comparison of Patterson maps will always have a peak for the correct model orientation. Naturally, this peak can be weak because of a small size of the model and can be buried in the noise, but it should be identified by the multiple rotation functions (see, for example, §4.2).

In order to cope with the incompleteness in the following translation search, missing atoms can be taken into account statistically (Afonine *et al.*, 2001). The magnitudes of structure factors calculated from the search model and completed by estimates for missing atoms can now be compared more correctly with the experimental magnitudes. Such statistical correction of the model is the basis of the maximum-likelihood approach for molecular replacement (Read, 2001).

5.3. Complete models with errors

The second extreme situation is represented by a complete model which has significant errors in atomic positions. Naturally, special attempts can be undertaken in order to improve the model. For example, some probabilistic criteria can be used which allow the imperfectness of the model to be taken

into account (Read, 2001; Afonine *et al.*, 2001). Otherwise, an improved model can be constructed when several NMR models are available (for reviews, see Chen *et al.*, 2000; Chen, 2001).

Again, for such a model with errors, its correct position will not necessarily give the best correspondence between the calculated and observed magnitudes for a set of structure factors taken at the traditional resolution of 4–15 Å. However, taken in a known approximate orientation, such a model positioned correctly gives a strong coincidence of low-resolution data (Urzhumtsev & Podjarny, 1995; Fokine & Urzhumtsev, 2002), whose magnitudes are less sensitive to errors in atomic coordinates and to typical errors (about 2–5°) in the model orientation. In other words, the models are more similar at a resolution of 10–15 Å and lower than they are at 4 Å. Traditionally, these low-resolution reflections are excluded from the MR searches because they are very strongly influenced by bulk solvent. It has been shown that for a known model orientation this bulk-solvent contribution can be taken into account efficiently and quickly; this modification drastically increases the signal in the translation function (Fokine & Urzhumtsev, 2002). Attention must be paid to the fact that these low-resolution data should not be used for the rotation analysis because they can displace and decrease the peak for the correct orientation.

Therefore, for such an inexact but relatively complete model, determination of the model orientation is the key point in the resolution of the whole MR problem. Unfortunately, the relative displacement of atoms makes the experimental and calculated Patterson maps dissimilar and decreases the signal in their optimal superimposition. Also, the model imperfection can be caused by a slight reorientation of the model domains in comparison with their position in the molecule under study; in this case, the rotation function can have several peaks relatively close each to other when one or another part of the calculated Patterson map fits best to the corresponding region of the experimental Patterson map. Both these complications in the orientation search can be overcome by the multiple rotation-function approach.

5.4. Large number of molecules in the unit cell

When the molecule consists of several domains whose relative orientation is unknown, one possible approach is to introduce more degrees of freedom, making the model flexible (Brünger, 1990; DeLano & Brünger, 1995), or to decompose it in several rigid groups. In this difficult case of MR a search with independent domains can be tried. A similar problem appears when the crystal contains several independent molecules in the asymmetric unit. Positioning of multiple models one by one is expected to be rather unsuccessful after some limit because, similar to the case of a very incomplete model, there is not much reason to place the first model correctly, which represents a very small part of the whole diffraction matter of the crystal. To overcome this obstacle, a multibody strategy can be applied (Glykos & Kokkinidis, 2000, 2001) in which a search for the positions of all molecules is carried out simultaneously. Again, this strategy supposes that the quality of the models is sufficiently high that the similarity assumption is held and that the search criterion has its optimum for the correct position and orientation of the models.

Alternatively, low-resolution *ab initio* phasing procedures (Lunin *et al.*, 2000) can be used. In general, these procedures do not need any model. As practical experience shows, low-resolution direct phasing is capable of finding the position and the shape of the molecule even when the exact number of molecules in the unit cell is unknown (N. Lunina & J. Müller, personal communication).

While low-resolution phasing procedures can sometimes even provide the secondary-structure elements (Lunina *et al.*, 2000; Chabriere *et al.*, 2001), they are currently more robust in the determination of the model position. Therefore, an independent determination of the model orientation using higher resolution data followed by low-resolution positioning can be helpful. This agrees also with the general strategy, which is to decompose the $6N$-dimensional search into separated $3N$ multirotation and $3N$ multitranslation problems (Glykos & Kokkinidis, 2000). In this situation of multiple models, the rotation-function analysis is similar to the analysis of an incomplete model: the signal can be weak but should appear in the rotation function and can be identified by the multiple rotation function.

6. Discussion

Separated searches in the orientation and translation stages of molecular replacement have the advantage of making use of the features of each of these steps. For the rotation search, a cluster analysis of multiple rotation functions can be useful in many practical situations, especially with NMR models. This approach has been successfully applied for several test examples of 'difficult structures'. Recently, a determination of a new protein structure, that of the *A* domain of complement factor B, has been reported where the use of the multiple rotation function was essential (A. Bhattacharya, personal communication). The role of the different parameters of the method, especially the cluster threshold, has been studied. When the orientation of the search model is defined, the position of the model can be found much more easily because spurious signals corresponding to wrong model orientations will not appear during the search. The efficiency of the translation search can be increased by including low-resolution data, especially if the bulk-solvent correction is taken into account. The molecular-replacement problems can also be resolved through direct phasing, which directly shows a rough molecular image in the cell.

The authors thank D. Teller and A. Aubry for experimental data, L. Torlay for technical help, LIPHA (Lyon) and CPER 'Lorraine' for financial support, M. Weiss and R. Read for useful comments, V. Y. Lunin for a very fruitful discussion of the project and the manuscript, C. Lecomte for his support of

the project and the referees for their constructive criticism. The authors are members of GdR 2417 CNRS.

References

Aevarsson, A., Brazhnikov, E., Garber, M., Zheltonosova, J., Chirgadze, Yu., Al-Karadaghi, S., Svensson, L. A. & Liljas, A. (1994). *EMBO J.* **13**, 3669–3677.

Afonine, P., Lunin, V. Yu. & Urzhumtsev, A. (2001). *CCP4 Newsl. Protein Crystallogr.* **39**, 52–56.

Anderson, D. H., Weiss, M. S. & Eisenberg, D. (1996). *Acta Cryst.* D**52**, 469–480.

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.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Brünger, A. T. (1990). *Acta Cryst.* A**46**, 46–57.

Chabriere, E., Lunina, N., Lunin, V. Y. & Urzhumtsev, A. (2001). Abstr. XXth Eur. Crystallogr. Meet. Krakow, p. 70.

Chang, G. & Lewis, M. (1997). *Acta Cryst.* D**53**, 279–289.

Chen, W. (2001). *Acta Cryst.* D**57**, 1457–1461.

Chen, W., Kleywegt, G. & Dodson, E. (2000). *Structure*, **8**, 213–220.

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. V. (1991). *Dokl. Acad. Nauk SSSR*, **320**, 488–491. (In Russian.)

DeLano, W. L. & Brünger, A. (1995). *Acta Cryst.* D**51**, 740–748.

Fokine, A. & Urzhumtsev, A. (2002). *Acta Cryst.* A**58**, 72–74.

Glykos, N. & Kokkinidis, M. (2000). *Acta Cryst.* D**56**, 169–174.

Glykos, N. & Kokkinidis, M. (2001). *Acta Cryst.* D**57**, 1462–1473.

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

Lunin, V. Yu., Lunina, N., Petrova, T., Skovoroda, T., Urzhumtsev, A. & Podjarny, A. (2000). *Acta Cryst.* D**56**, 1223–1232.

Lunin, V. Yu., Lunina, N., Petrova, T., Vernoslova, E., Urzhumtsev, A. & Podjarny, A. (1995). *Acta Cryst.* D**51**, 896–903.

Lunin, V. Y., Urzhumtsev, A. & Skovoroda, T. (1990). *Acta Cryst.* A**46**, 540–544.

Lunina, N., Lunin, V. Yu. & Urzhumtsev, A. (2000). *Acta Cryst.* A**56** (Supplement), s62.

Menchize, V., Corbier, C., Didierjean, C., Saviano, M., Benedetti, E., Jacquot, J.-P. & Aubry, A. (2001). *Biochem. J.* **359**, 65–75.

Mittard, V., Blackledge, M. J., Stein, M., Jacquot, J.-P., Marion, D. & Lancelin, J.-M. (1997). *Eur. J. Biochem.* **243**, 374–383.

Mronga, S., Luginbühl, P., Brown, L. R., Ortenzi, C., Luporini, P., Bradshaw, R. A. & Wütrich, K. (1994). *Protein Sci.* **3**, 1527–1536.

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

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

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

Rossmann, M. G. (1972). *The Molecular Replacement Method.* New York: Gordon & Breach.

Rossmann, M. G. (1990). *Acta Cryst.* A**46**, 73–82.

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

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

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

Urzhumtseva, L. & Urzhumtsev, A. (1997). *J. Appl. Cryst.* **30**, 402–410.