J. Appl. Cryst. (1997). **30**, 1022–1025

MOLREP: an Automated Program for Molecular Replacement

ALEXEI VAGIN[a]* AND ALEXEI TEPLYAKOV[b]

[a]*Department of Chemistry, University of York, Heslington, York YO1 5DD, England, and* [b]*EMBL, c/o DESY, Notkestrasse 85, D-22603 Hamburg, Germany. E-mail: alexei@yorvic.york.ac.uk*

(*Received 26 February 1997; accepted 6 May 1997*)

Abstract

MOLREP is an automated program for molecular replacement which utilizes effective new approaches in data processing and rotational and translational searching. These include an automatic choice of all parameters, scaling by Patterson origin peaks and soft resolution cut-off. One of the cornerstones of the program is an original full-symmetry translation function combined with a packing function. Information from the model already placed in the cell is incorporated in both translation and packing functions. A number of tests using experimental data proved the ability of the program to find the correct solution in difficult cases.

1. Introduction

The molecular replacement (MR) method of crystal structure determination (Rossmann, 1972) has developed significantly in recent years. This reflects increased interest in the method because of the rapidly growing number of protein structures that are available as search models. There are several program packages for MR using various algorithms for the six-dimensional search (reviewed by Turkenburg & Dodson, 1996). The most commonly used are *AMoRe* (Navaza, 1994), *X-PLOR* (Brünger, 1990) and *MERLOT* (Fitzgerald, 1988).

Here we present the program *MOLREP* which is a fully automated program for MR. It is characterized by several new approaches in data processing and rotational and translational searching. As part of the package *BLANC* (Vagin, 1982; Vagin, Murshudov & Strokopytov, 1997), it has been used successfully for solving the structures of a number of proteins (Borisova *et al.*, 1986; Zaitsev, Vagin, Nekrasov & Moshkov, 1986; Pavlovsky *et al.*, 1989; Teplyakov *et al.*, 1992).

2. Description of the program

Starting with the X-ray data for the unknown structure and a search model, *MOLREP* goes through all stages of MR automatically. It chooses all required parameters, selects the solution and produces the model properly oriented and positioned in the unit cell. On the other hand, calculation of the rotation function (RF), translation function (TF) and rigid-body refinement can be run separately using three independent programs which allow manual input of a number of parameters.

Structure factors are calculated for each orientation/position of the model by the fast Fourier transformation of its electron density. An alternative is to interpolate their values from a table of structure factors calculated on a fine grid, as in *AMoRe*. The price for the higher accuracy of direct calculations is the CPU time. However, the difference becomes noticeable only when testing many RF solutions. One cycle of MR (RF + TF) for a medium-sized protein (see §3.1) at 3 Å resolution takes about 10 min on SGI Challenge L for both *MOLREP* and *AMoRe*.

MOLREP exhibits the following new features in data processing in contrast with the other programs.

(i) Scaling of the observed and calculated structure factors is based on the scaling of the corresponding Patterson origin peaks (Rogers, 1965). This method is advantageous when only low-resolution data are available. In such a case, the estimation of the overall B factor from the Wilson plot may be inaccurate. Scaling by Patterson synthesis is also useful for the cross RF where there are different cells for the search model and for the unknown structure.

(ii) The low-resolution cut-off is applied in the form of a Gaussian $F = F_{\rm obs}\{1 - \exp[-B_{\rm off}(\sin\theta/\lambda)^2]\}$, where $B_{\rm off} = (2d_{\rm max})^2$. The common practice is to exclude low-resolution reflections from calculations because they are corrupted by the bulk solvent contribution. However, this results in the series termination effect and introduces systematic errors in the electron density, particularly at the surface of the molecule. The 'soft' low-resolution cut-off similar to the one described is used in some refinement programs as a way of modelling disordered solvent. We propose to apply it in all calculations and not only for scaling $F_{\rm obs}$ and $F_{\rm calc}$. This will effectively reduce the weight of low-resolution terms, which is preferable to excluding those terms completely. Application of the 'soft' cut-off increased the signal-to-noise ratio in a test described below.

The rotational search is performed using the fast RF of Crowther (1972). To reduce the noise, the RF is calculated for three orthogonal orientations of the model and averaged over them. The radius of the Patterson

© 1997 International Union of Crystallography
Printed in Great Britain – all rights reserved

sphere is derived from the size of the search model and is usually twice the radius of gyration. However, the Patterson radius should not exceed half of the unit-cell dimension by more than 25%, otherwise the contribution of intermolecular vectors becomes significant. The shape of the model is taken into account at the stage of the translational search. For elongated molecules, the resolution range is proportionally increased to include low-resolution terms so that the TF is less sensitive to errors in the model orientation.

The TF used in the package has been described elsewhere (Vagin, 1989). The basic equations are given in the Appendix. This TF originates from the $T2$ function of Crowther & Blow (1967) corrected by Harada, Lifchitz, Berthou & Jolles (1981). $T2$ is usually calculated as the sum of individual T_{jk} functions, each of which corresponds to a pair of crystallographic symmetry operators j and k

$$T2(s) = \sum_{j\neq k} T_{jk}(s) = \sum_{j\neq k} \int P^o(r) P^c_{jk}(s, r)\, \mathrm{d}r$$

where P^o and P^c are the observed and calculated Patterson functions, respectively.

The principal advantage of $T2$ over $T1$ is the simultaneous use of all symmetry operators resulting in a single peak with an improved signal-to-noise ratio. This gives the position of the model in the unit cell whereas $T1$ indicates the position relative to a symmetry element. *MOLREP* exhibits two main differences compared with other TF programs. First, the TF is calculated as the product of T_{jk} functions rather than their sum. This increases the contrast of the TF. Second, the translational search is coupled with a packing function (PF) to remove false maxima which correspond to interpenetrating molecules

$$Q(s) = -\sum_{j\neq k} \int \rho_j(r, s)\rho_k(r, s)\mathrm{d}r$$

where $\rho_j(r,s)$ is the electron density of the search model calculated for the symmetry operator j. Thus, the TF used here is the following

$$T(s) = \prod_{j\neq k} T_{jk}(s) \times Q(s).$$

The TF is computed using the fast Fourier technique. The algorithm for calculating the PF is similar to that used for the TF and performed by the same subroutine. Both TF and PF allow incorporation of the second model already placed in the cell (see Appendix). The orientation and position of the model obtained from the RF and TF may be refined by the conventional rigid-body multidomain refinement incorporated in *MOLREP*. The minimized function is the difference between the observed and calculated structure factors. The best solution is indicated by the highest correlation coefficient

$$\mathrm{CC} = \frac{\langle |F_o||F_c| - \langle |F_o|\rangle\langle |F_c|\rangle\rangle}{(\langle |F_o|^2 - \langle |F_o|\rangle^2\rangle\langle |F_c|^2 - \langle |F_c|\rangle^2\rangle)^{1/2}}.$$

The refinement of the orientation of the model prior to the translational search is performed in space group $P1$ and has proved to be very useful in some cases.

3. Tests

The program was tested on a number of cases with experimental X-ray data and gave satisfactory results. Here we present three tests: protein structure with a low sequence similarity to the search model, the MR solution of a complex using one of the components and the structure of an RNA hexamer.

3.1. *Test 1*

Inorganic pyrophosphatase from *Thermus thermophilus* (PDB code 2PRD; Teplyakov *et al.*, 1992) was crystallized in space group $R32$ with one 20 kDa monomer in the asymmetric unit. The search model was constructed from the structure of the 32 kDa monomer of the yeast enzyme (1YPP; Bernstein *et al.*, 1977) by removing extra parts of the sequence. The case is difficult for MR because of the low level of amino acid identity of the two enzymes (only 20%). The three-dimensional superposition of the structures gives an r.m.s. deviation of 2.0 Å for all 170 common Cα atoms. 40 Cα atoms, *i.e.* almost a quarter of the structure, deviate by more than 2 Å. The case was tested using *MOLREP*, *AMoRe* and *X-PLOR*. The RF was calculated in the resolution range 15–3 Å with the radius of the Patterson sphere of 25 Å (these parameters gave the best results for both *MOLREP* and *AMoRe*). The correct orientation was found by *MOLREP* and *AMoRe* as the top peak of the RF. The contrast, defined as the ratio of the first (correct) peak to the next (incorrect) peak, was 1.30 in *MOLREP* and 1.11 in *AMoRe*. The correct solution was not among the first 300 peaks of the *X-PLOR* RF (these were also checked using PC refinement). The translational search performed by *MOLREP* and *AMoRe* identified the correct solution as the top peak of the TF, with the contrast 1.52 and 1.36, respectively.

3.2. *Test 2*

This example demonstrates the ability of the program to find a solution when only a small fraction of the structure is used as a search model. The 'unknown' structure was the complex of Fab with lysozyme (2IFF; Bernstein *et al.*, 1977; Chacko *et al.*, 1995), crystallized in space group $P2_1$ with one molecule in the asymmetric unit. The search model was that of lysozyme (1HEM), which constituted 24% of the total mass of the complex. A single peak which was twice the height of the others

appeared on both the RF and the TF calculated with *MOLREP* and indicated the correct solution.

3.3. *Test 3*

Nucleic acids are difficult objects for MR because of their internal symmetry. We have tested the program on the RNA hexamer CGCGCG, which crystallized in space group $P6_122$ with one duplex in the asymmetric unit. The molecules pack head-to-tail and form continuous helices along the twofold axes. The structure was solved by MR using *AMoRe* (Rypniewski, 1997). The search model was the *B*-form RNA duplex UAUAUA (1RNA). The correct solution was selected at the stage of the translational search (it corresponded to the eighth peak of the RF), whereas the RF was insensitive to the correct orientation: 20 top peaks were approximately of equal height. The RF in *MOLREP* identified the correct solution as the top peak though with a very low contrast (1.05). This orientation resulted in the TF peak being 60% higher than the others.

These nontrivial tests present *MOLREP* as a reliable, fast and completely automatic program. In some cases, *MOLREP* demonstrated higher accuracy and a better contrast compared with *AMoRe*, which is commonly recognized as the best in the field. We believe that in difficult cases one program may have an advantage over the other due to the different algorithms and it is therefore wise to try both of them.

4. Distribution

The program *MOLREP* is written in standard Fortran77 and can be run under Unix, VMS and Windows. It is available free as part of the package *BLANC* or independently *via* anonymous FTP from the account ftp.yorvic.york.ac.uk. All inquiries about the program should be addressed to A. Vagin (alexei@yorvic.york.ac.uk).

APPENDIX *A*
Translation function and packing function for one and two models

The notation is given in Table 1.

Table 1. *Notation*

h, H	Reciprocal-space vectors
r	Position in the unit cell
s	Translation vector in real space
u	Patterson-space vector
V	Volume of the unit cell
N	Number of symmetry operators
$F_{\rm obs}(h)$	Observed structure-factor amplitude
$F(h)$	Structure-factor amplitude calculated for the model in the initial position
$P(u,s)$	Patterson function calculated for the model in position s
$P_{\rm obs}(u)$	Patterson function for the unknown structure
$\rho(r,s)$	Electron density of the model in position s
$\rho_j(r,s)$	Electron density of the model related to the original model by the jth symmetry operator
$[R_j]$	Rotation matrix component of the jth symmetry operator
d_j	Translation vector component of the jth symmetry operator
$T_{jk}(s)$	Translation function for symmetry operators j and k

(*a*) The electron density for the model in position s

$$\rho(r,s) = \sum_{j=0}^{N-1} \rho_j(r,s) = \sum_j [V^{-1} \sum_h F(h[R_j]) \exp(-2\pi i h r) \times \exp(2\pi i h [R_j] s)]$$

The Patterson function for the model in position s

$$P(u,s) = \int \rho(r+u,s)\rho(r,s)\mathrm{d}r = \sum_j \sum_k \int \rho_j(r+u,s)\rho_k(r,s)\mathrm{d}r = \sum_j \sum_k \int P_{jk}(u,s)$$

$$P(u,s) = V^{-1} \sum_j \sum_k \sum_h [F(h[R_j])F^*(h[R_k]) \exp(2\pi i h d_j) \times \exp(-2\pi i h d_k) \exp(-2\pi i h u) \exp(2\pi i h [R_j] s) \times \exp(-2\pi i h [R_k] s)]$$

The translation function for one model

$$T(s) = \int P_{\rm obs}(u)P(u,s)\mathrm{d}u = \sum_j \sum_k \int P_{\rm obs}(u)P_{jk}(u,s)\mathrm{d}u = \sum_j \sum_k T_{jk}(s)$$

The translation function presented as Fourier series

$$T(s) = V^{-1} \sum_H A_H \exp(-2\pi i H s)$$

with coefficients

$$A_H = \sum_j \sum_k \sum_h [F^2_{\rm obs}(h)F(h[R_j])F^*(h[R_k]) \exp(2\pi i h d_j) \times \exp(-2\pi i h d_k)]$$

where summation is over those h for which $H = h([R_k] - [R_j])$.

(*b*) The electron density and translation function for two independent models in the unit cell

$$\rho(r, s_1, s_2) = \rho_1(r, s_1) + \rho_2(r, s_2)$$

$$T(s_1, s_2) = \sum_j \sum_k T_{jk}(s_1, s_2)$$

The translation function presented as Fourier series

$$T(s_1, s_2) = V^{-1} \sum_{H1} \sum_{H2} A_{H1} A_{H2} \exp(-2\pi i H_1 s_1) \times \exp(-2\pi i H_2 s_2)$$

with coefficients

$$\begin{aligned} A_{H1,H2} = & \sum_j \sum_k \sum_h F^2_{\rm obs}(h) F_1(h[R_j]) F_1^*(h[R_k]) \\ & \times \exp(2\pi i h d_j) \exp(-2\pi i h d_k) \\ & + \sum_j \sum_k \sum_h F^2_{\rm obs}(h) F_2(h[R_j]) F_2^*(h[R_k]) \\ & \times \exp(2\pi i h d_j) \exp(-2\pi i h d_k) \\ & + 2 \sum_j \sum_k \sum_h F^2_{\rm obs}(h) F_1(h[R_j]) F_2^*(h[R_k]) \\ & \times \exp(2\pi i h d_j) \exp(-2\pi i h d_k) \end{aligned}$$

where summation in the first term is over those h for which $H_1 = h([R_k] - [R_j])$ and $H_2 = 0$, summation in the second term is over those h for which $H_1 = 0$ and $H_2 = h([R_k] - [R_j])$, summation in the third term is over those h for which $H_1 = -h[R_j]$ and $H_2 = h[R_k]$.

(*c*) The translation function for two independent models when one model is fixed

$$T(s) = V^{-1} \sum_H A_H \exp(-2\pi i H s)$$

with coefficients

$$\begin{aligned} A_H = & \sum_j \sum_k \sum_h F^2_{\rm obs}(h) F_1(h[R_j]) F_1^*(h[R_k]) \\ & \times \exp(2\pi i h d_j) \exp(-2\pi i h d_k) \\ & + 2 \sum_j \sum_k \sum_h F^2_{\rm obs}(h) F_1(h[R_j]) F_2^*(h[R_k]) \\ & \times \exp(2\pi i h d_j) \exp(-2\pi i h d_k) \end{aligned}$$

where summation in the first term is over those h for which $H = h([R_k] - [R_j])$, summation in the second term is over those h for which $H = -h[R_j]$.

(*d*) The packing function is an overlapping function of electron densities for the symmetry-related models. The equations are analogous to those for the translation function

$$Q(s) = -\sum \sum_{j \neq k} \int \rho_j(r, s) \rho_k(r, s) {\rm d}r$$

or as Fourier series

$$Q(s) = V^{-1} \sum_H A_H \exp(-2\pi i H s)$$

with coefficients

$$\begin{aligned} A_H = & -\sum_{j \neq k} \sum \sum_h F(h[R_j]) F^*(h[R_k]) \exp(2\pi i h d_j) \\ & \times \exp(-2\pi i h d_k) \end{aligned}$$

where summation is over those h for which $H = h([R_k] - [R_j])$.

We thank Dr Wojtek Rypniewski for providing experimental data for the RNA structure. The work was supported by the EU BIOTECH grant BIO4CT-96-0189 (AV).

References

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

Borisova, S. N., Vagin, A. A., Nekrasov, Y. V., Pavlovsky, A. G., Strokopytov, B. V. & Vainshtein, B. K. (1986). *Kristallografiya*, **31**, 474–477.

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

Chacko, S., Silverton, E., Kammorgan, L., Smithgill, S., Cohen, G. & Davies, D. (1995). *J. Mol. Biol.* **245**, 261–274.

Crowther, R. A. (1972). *The Molecular Replacement Method*, edited by M. G. Rossmann, pp. 173–178. New York: Gordon & Breach.

Crowther, R. A. & Blow, D. M. (1967). *Acta Cryst.* **23**, 544–548.

Fitzgerald, P. M. D. (1988). *J. Appl. Cryst.* **21**, 273–278.

Harada, Y., Lifchitz, A., Berthou, J. & Jolles, P. (1981). *Acta Cryst.* A**37**, 398–406.

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

Pavlovsky, A. G., Sanishvili, R. G., Borisova, S. N., Strokopytov, B. V., Vagin, A. A., Chepurnova, N. K. & Vainshtein, B. K. (1989). *Kristallografiya*, **34**, 137–142.

Rogers, D. (1965). *Computing Methods in Crystallography*, edited by J. S. Rollett, pp. 117–148. Oxford University Press.

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

Rypniewski, W. (1997). Personal communication.

Teplyakov, A., Polyakov, K., Obmolova, G., Strokopytov, B., Kuranova, I., Osterman, A., Grishin, N., Smulevitch, S., Zagnitko, O., Galperina, O., Matz, M. & Stepanov, V. (1992). *Eur. J. Biochem.* **208**, 281–288.

Turkenburg, J. P. & Dodson, E. J. (1996). *Curr. Opin. Struct. Biol.* **6**, 604–610.

Vagin, A. A. (1982). PhD thesis, Institute of Crystallography, Moscow, Russia.

Vagin, A. A. (1989). *CCP4 Newsletter on Protein Crystallography* 29, edited by K. Henrick & K. S. Wilson, pp. 117–121. Warrington: Daresbury Laboratory.

Vagin, A. A., Murshudov, G. N. & Strokopytov, B. V. (1997). *CCP4 Newsletter on Protein Crystallography* 33, edited by M. Winn, pp. 25–30. Warrington: Daresbury Laboratory.

Zaitsev, V. N., Vagin, A. A., Nekrasov, Y. V. & Moshkov, K. A. (1986). *Kristallografiya*, **31**, 937–941.