Acta Crystallographica Section D
**Biological Crystallography**
ISSN 0907-4449

# Molecular replacement with *MOLREP*

**Alexei Vagin[a]* and Alexei Teplyakov[b]*‡**

[a]Structural Biology Laboratory, University of York, Heslington, York YO10 5YW, England, and [b]University of Maryland Biotechnology Institute, Rockville, MD 20850, USA

‡ Present address: Centocor R&D Inc., Radnor, PA 19087, USA.

Correspondence e-mail: alexei@ysbl.york.ac.uk, ateplyak@its.jnj.com

*MOLREP* is an automated program for molecular replacement that utilizes a number of original approaches to rotational and translational search and data preparation. Since the first publication describing the program, *MOLREP* has acquired a variety of features that include weighting of the X-ray data and search models, multi-copy search, fitting the model into electron density, structural superposition of two models and rigid-body refinement. The program can run in a fully automatic mode using optimized parameters calculated from the input data.

Received 9 July 2009
Accepted 13 October 2009

A version of this paper will be published as a chapter in the new edition of Volume *F* of *International Tables for Crystallography*.

## 1. Introduction

Molecular replacement (MR) is one of the two principal methods of crystal structure determination (Rossmann, 1972). About two-thirds of the X-ray structures recently deposited in the Protein Data Bank (Berman *et al.*, 2000) have been determined by MR. With the rapid growth in the number of protein structures available as search models, the use of the method will increase even further in the future. Traditionally, MR has been implemented as a three-dimensional rotational search followed by a three-dimensional translational search. With recent advances in computing a combined six-dimensional search is becoming feasible; however, to date the most popular program packages for MR use various algorithms for a (3+3)-dimensional search. These include *AMoRe* (Navaza, 2001), *Phaser* (McCoy *et al.*, 2007) and the MR implementation in *CNS* (Brünger *et al.*, 1998). Together with *MOLREP*, they cover over 95% of the structures solved by MR.

*MOLREP* is an automated program for molecular replacement that utilizes a number of original approaches to rotational and translational search and data preparation. *MOLREP* was initially developed (Vagin, 1989) as a fast and efficient alternative to the few existing program packages and has subsequently been enriched with new algorithms and adapted to various operating systems. *MOLREP* was originally part of the program suite *BLANC* (Vagin, 1982; Vagin *et al.*, 1998) and was later included in the *CCP*4 suite (Collaborative Computational Project, Number 4, 1994). *MOLREP* is a component of several MR pipelines, including *BALBES* (Long *et al.*, 2008), *MrBUMP* (Keegan & Winn, 2008) and the JCSG pipeline (Schwarzenbacher *et al.*, 2008), that utilize the entire PDB for fully automatic structure determination. Since the first publication describing the program (Vagin & Teplyakov, 1997) it has been cited nearly 2000 times. In this update, we highlight new features of the program, which include weighting of the X-ray data and search models, multi-copy search, fitting the model into electron density and structural superposition of two models by using the spherically

doi:10.1107/S0907444909042589

averaged and phased translation function and rotation function.

## 2. Program operation

The philosophy behind *MOLREP* was to provide a tool such that most if not all of the tasks in MR could be performed automatically, or at least the optimal choice of parameters at each step of the process would be performed by the program. To make the program user-friendly, the interface was organized so that the user can run the program in a dialogue mode. The prompts appear as self-explanatory questions with a choice of possible answers and a list of keywords to define certain parameters. The batch mode using a command file or a command string is also available, as well as the *CCP4i* GUI.

*MOLREP* can perform a variety of tasks that require rotational and/or positional search: standard MR, multi-copy search, fitting a model into electron density, heavy-atom search and model superposition. The arsenal of rotation (RF) and translation (TF) functions includes self-RF, cross-RF, locked cross-RF, phased RF, full-symmetry TF, phased TF, spherically averaged phased TF and packing function (PF). The program is general for all space groups.

MR with *MOLREP* is not only fast and efficient but is also largely automatic. The minimum input provided by the user includes the X-ray data and the atomic model.

The input reflection file may be in *CCP*4 MTZ format, *BLANC* format or CIF format. An electron-density map or an electron-microscopy (EM) reconstruction (in *CCP*4 or *BLANC* format) may substitute for the structure factors. In this case, the map is converted to structure factors and phases. Crystal symmetry information is taken from the reflection file.

The input model is usually provided as an X-ray structure in the PDB format, but may also be an NMR ensemble. The ensemble is utilized either as a single entity (structure factors are calculated from the entire ensemble) or as individual members (MR is carried out for each structure). A set of homologous structures can be used in a similar way to the NMR ensemble. An electron-microscopy image or electron-density map can also substitute for the search model.

*MOLREP* can run with a minimum input using default parameters and parameters calculated from the input data. However, it may be useful to control the program through the use of a wide range of keywords to override the defaults. The keywords are described in the documentation supplied with the software.

The output of the program is a PDB file with the atomic model ready for refinement and a text file with details of the calculations. Additionally, an XML file suitable for communication between different programs in a pipeline is generated.

## 3. Preparation of the search model

The search model is automatically modified in various ways in order to try these variants for MR (Lebedev *et al.*, 2008). The options include (i) a polyalanine model, (ii) a model with modified atomic *B* factors increased proportionally to the atom accessibility and (iii) a model with the corrected amino-acid sequence according to the alignment of the search and target sequences. The alignment is carried out by the program when the target sequence is supplied. It takes into account the three-dimensional structure of the search model and weights the buried residues more than the surface residues. It also positions insertions and deletions away from the secondary-structure elements. Based on this alignment, certain atoms in the side chains and entire residues in the deletion regions are removed from the search model.

## 4. Preparation of the X-ray data

Proper treatment of the X-ray data may dramatically improve the results of MR and structure refinement. Several unique and conventional approaches for data scaling and correction have been implemented in *MOLREP*.

Anisotropic correction of the experimental data is applied when the X-ray intensity fall-off with resolution varies substantially with direction. The X-ray data are fitted to the isotropic Gaussian derived from the origin peak of the Patterson function (Rogers, 1965). The procedure is implemented in reciprocal space according to Blessing & Langs (1988).

Scaling of the observed and calculated structure factors is based on the scaling of the height and width of the corresponding Patterson origin peaks (Rogers, 1965; Blessing & Langs, 1988). This method is advantageous when only low-resolution data are available. In such a case, the estimation of the overall *B* factor ($B_{\rm over}$) from the Wilson plot may be inaccurate. Scaling by Patterson is also of value for the cross-RF, where different cells are used for the search model and for the unknown structure.

To define the weighting scheme for the X-ray data, *MOLREP* estimates a number of parameters for the search model. Two of them, the radius of gyration of the model ($R_{\rm g}$) and the sequence similarity to the target protein ($\Omega$), are translated into the parameters of a low-pass filter and a high-pass filter (Gonzalez & Woods, 2002), which define the resolution dependence of the reflection weights in the RF and TF. It is assumed that the intensities of the low-resolution reflections mostly depend on large-scale details of the crystal and are not very sensitive to the internal features of the molecules. In contrast, the high-resolution reflections define the structural details that are smaller than the differences between the search and the target molecule. The exact measure of this difference is unknown until the structure is solved, but it correlates with the known sequence similarity (Chothia, 1992). The low-pass filter helps to reduce noise in the RF and TF by dampening high-resolution reflections and thus blurring the Patterson map. The low-pass filter is applied to structure factors as an additional *B* factor,

$$F_{\rm new} = F \times \exp[-B_{\rm add}(\sin\theta/\lambda)^2]. \qquad (1)$$

The value of $B_{\rm add}$ is based on the sequence similarity and normalized to the Patterson origin peak according to an empirical formula,

$$B_{\rm add} = (B_{\rm lim} - B_{\rm res}) \times (1 - \Omega)^4 + B_{\rm res} - B_{\rm over} \qquad (2)$$

where $B_{\rm lim}$ is the maximum allowed addition to the $B$ factor when no similarity exists ($\Omega = 0$). $B_{\rm lim} = 8\pi^2 u^2$, where $u$ is the mean atomic displacement (1.1 Å). $B_{\rm res}$ is determined from the maximum resolution $R_{\rm max}$ as

$$B_{\rm res} = 2R^2_{\rm max}. \qquad (3)$$

The high-pass filter dampens low-resolution reflections and thus sharpens the Patterson map. The high-pass filter is applied to structure factors as

$$F_{\rm new} = F \times \{1 - \exp[-B_{\rm off}(\sin\theta/\lambda)^2]\}, \qquad (4)$$

where $B_{\rm off} = 2\pi^2(R_{\rm g}/6)^2$.

The filter parameters derived automatically from the search model proved to work well in some difficult MR test cases. However, the user may want to define them explicitly when, for example, high sequence similarity between the search and the target protein is in discord with the significant differences in their structures.

## 5. Rotational search

The rotational search is performed using the RF of Crowther (1972), which utilizes the fast Fourier transform (FFT) technique. The number of spherical harmonics used in the calculations is limited to 100. Spherical harmonics with $l = 0$ are not included in order to eliminate the effect of the origin Patterson peak. The default radius of the integration sphere is derived from the size of the search model and is usually two times larger than the radius of gyration. However, the Patterson radius should not significantly exceed half of the minimal unit-cell dimension because the contribution of intermolecular vectors becomes significant.

The RF solutions may be refined prior to positional search using a rigid-body technique. The refinement is performed in space group $P1$ and the outcome is evaluated by the correlation coefficient.

## 6. Positional search

The full-symmetry TF (Vagin, 1989) originates from the T2 function of Crowther & Blow (1967) corrected by Harada *et al.* (1981). It simultaneously uses all symmetry operators, resulting in a single peak with an improved signal-to-noise ratio which directly gives the position of the model in the unit cell. In addition, the TF is coupled with a PF to remove false maxima which correspond to interpenetrating molecules. Both the TF and PF allow the incorporation of a second model already placed in the cell. The TF solution may be subjected to rigid-body refinement incorporated in *MOLREP*. Noncrystallographic symmetry may be imposed on the model in order to restrain the refinement.

Pseudo-translation is automatically detected from analysis of the Patterson map. A significant off-origin peak gives the pseudo-translation vector, which is used to modify structure factors in the TF calculation (Navaza *et al.*, 1998). Alternatively, the pseudo-translation vector may be supplied by the user. A pseudo-translational copy of the model is added at the end of the positional search.

## 7. Multi-copy search

The MR method has been extended to a simultaneous search for multiple copies of the macromolecule in the unit cell (Vagin & Teplyakov, 2000). The central point of this approach is the construction of a dimer search model from the properly oriented monomers using a special TF. This model is then used for a positional search with a conventional TF. The method does not impose any limitation on the oligomeric structure of the protein, either on the number of monomers or on their relative location, *i.e.* pure rotational symmetry is not required. In principle, the monomers may even be of different types. In the case of two monomers, which we call a dyad, the multi-copy search takes place as follows.

(i) The RF is calculated for a monomeric search model. $N_{\rm R}$ highest peaks of the RF are used to produce a set of oriented monomers. All possible $N_{\rm R}(N_{\rm R} + 1)/2$ pair combinations of these monomers constitute a set of putative dyads.

(ii) For each putative dyad, the intermolecular vector relating the two monomers of the dyad is determined using a special TF. The special TF is a phased TF, which treats the Patterson function as electron density, and the search model is described by the structure factors $F_1F^*_2$. The solution of the special TF is the dyad vector. $N_{\rm T}$ top solutions will be considered to ensure that the correct dyad is not missed. Therefore, the total number of dyads selected for the final positional search will be $N_{\rm T}N_{\rm R}(N_{\rm R} + 1)/2$.

(iii) A positional search for each dyad is performed using a conventional TF. The results are estimated on the basis of the TF value and a correlation coefficient.

The dyad search can be extended to a triad search by including a third monomer in the search model. The phased TF will then be used to find three vectors describing the triad. However, the dyad search seems to be sufficient in most cases, as the main problem is usually the location of the first pair of monomers. When this task is fulfilled, the search can be repeated for a third monomer or a second dyad with the first being fixed.

Although there are no limitations on the relative position and orientation of the two monomers constituting a dyad, the search space can be limited by imposing restraints on the oligomeric structure, *e.g.* by defining the pure rotational symmetry. This feature may be particularly useful in the case of molecular dimers and higher oligomers. Prior knowledge of the molecular symmetry may not only reduce the computational time but may also facilitate the search by selecting functionally meaningful solutions. This can be achieved by defining the angular relation between the monomers, *e.g.* as a 180 ± 10° rotation. Alternatively, the search may be reduced to the use of several top peaks of the self-RF that define a set of possible relative orientations of the monomers in a dyad.

## 8. Fitting the model into electron density

This 'real-space' search is employed when some phase information is available either from experimental data (MIR or MAD) or from a partial MR solution, *e.g.* when one domain of a multi-domain protein has been located in the unit cell. Positioning a model into electron density (or into an EM reconstruction) is carried out in three steps.

(i) The model is positioned in the cell using the spherically averaged TF (Vagin & Isupov, 2001). The locally spherically averaged experimental electron density is compared with that calculated from the model and all possible positions are tabulated.

(ii) For each selected position, the orientation of the model is determined using the local phased RF.

(iii) The orientation and position of the model is verified and refined using the phased TF.

It should be noted that a full six-dimensional search, although time-consuming, may in many cases work better. However, compared with traditional (3+3)-dimensional searches the fitting based on the spherically averaged phased TF may be advantageous, particularly when the search model is small. In general, the procedure works better for globular models.

The same procedure is used for the superposition of atomic models. One model is converted to the electron-density map and another model is fitted to this map using the spherically averaged and phased TF and RF. No sequence information or secondary-structure assignment is required for such a superposition. The procedure may be useful for proteins with very low sequence similarity. It is also suitable for nonprotein models such as nucleic acids or carbohydrates.

## 9. Distribution

The program *MOLREP* is written in standard Fortran 90 and can be run under Linux, Unix, Mac OSX and Windows. It is available free to academic users as a standalone version or as part of the *CCP*4 suite. The program, examples and documentation may be downloaded from the author's web site (http://www.ysbl.york.ac.uk/~alexei/molrep.html) or from the CCP4 web site. All inquiries about the program should be addressed to Alexei Vagin (alexei@ysbl.york.ac.uk).

We thank Andrey Lebedev for his contribution to the program and helpful discussions. This work was supported by CCP4, the Biotechnology and Biological Sciences Research Council (BBSRC) and EU Biotech.

## References

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

Blessing, R. H. & Langs, D. A. (1988). *Acta Cryst.* A**44**, 729–735.

Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.

Chothia, C. (1992). *Nature (London)*, **357**, 543–544.

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.

Crowther, R. A. (1972). *The Molecular Replacement Method*, edited by M. G. Rossmann, pp. 173–183. New York: Gordon & Breach.

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

Gonzalez, R. C. & Woods, R. E. (2002). *Digital Image Processing.* Upper Saddle River, New Jersey: Prentice Hall.

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

Keegan, R. M. & Winn, M. D. (2008). *Acta Cryst.* D**64**, 119–124.

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

Long, F., Vagin, A. A., Young, P. & Murshudov, G. N. (2008). *Acta Cryst.* D**64**, 125–132.

Navaza, J. (2001). *Acta Cryst.* D**57**, 1367–1372.

Navaza, J., Panepucci, E. H. & Martin, C. (1998). *Acta Cryst.* D**54**, 817–821.

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *J. Appl. Cryst.* **40**, 658–674.

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

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

Schwarzenbacher, R., Godzik, A. & Jaroszewski, L. (2008). *Acta Cryst.* D**64**, 133–140.

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

Vagin, A. A. (1989). *CCP4 Newsl. Protein Crystallogr.* **29**, 117–121.

Vagin, A. A. & Isupov, M. N. (2001). *Acta Cryst.* D**57**, 1451–1456.

Vagin, A. A., Murshudov, G. N. & Strokopytov, B. V. (1998). *J. Appl. Cryst.* **31**, 98–102.

Vagin, A. & Teplyakov, A. (1997). *J. Appl. Cryst.* **30**, 1022–1025.

Vagin, A. & Teplyakov, A. (2000). *Acta Cryst.* D**56**, 1622–1624.