Journal of
Applied
Crystallography

ISSN 0021-8898

Experiences with a new translation-function program

M. Fujinaga and R. J. Read

J. Appl. Cryst. (1987). **20**, 517–521

J. Appl. Cryst. (1987). **20**, 517–521

Experiences with a New Translation-Function Program

By M. Fujinaga* and R. J. Read†

MRC Group in Protein Structure and Function, Department of Biochemistry, University of Alberta, Edmonton, Alberta, Canada T6G 2H7

(*Received* 15 *September* 1986; *accepted* 27 *July* 1987)

Abstract

The development of a new translation-function program is reported. It is one that uses a linear correlation coefficient to determine the correct position of an oriented molecule in the crystal cell. The method has been implemented in a computer program called *BRUTE*. The program can also refine the orientation of the model and accept a set of atoms with fixed positions. Comparison of the correlation coefficient with other translation functions indicates that it is comparable to or slightly better than the rest. The most important feature of the program is its ability to adjust the orientation of the model. This allows for errors in the orientation obtained from the rotation function to be corrected.

Introduction

The phase problem for a protein structure is most often overcome by the method of multiple isomorphous replacement (MIR) (*e.g.* Blundell & Johnson, 1976). An alternative approach is available if the unknown structure is related to a known one. This is the technique of molecular replacement (Rossmann, 1972). The known structure, superimposed on the unknown structure in the crystal cell, is used as an approximate model to derive phase information. Molecular replacement is thus concerned with finding the three rotational and three translational parameters that specify the orientation and position, respectively, of the molecule in the crystal cell with respect to the symmetry elements.

The three rotational parameters are usually determined from the rotation function proposed by Rossmann & Blow (1962). It is based on the idea of maximizing the agreement of the intramolecular vectors between the observed and calculated Patterson functions. Crowther (1972) proposed a fast algorithm which expands the Patterson function in terms of spherical harmonics, allowing for the calculation to be performed with the use of the fast Fourier transform (FFT) (Cooley & Tukey, 1965). The rotational parameters are obtained independently of the position of the model in the cell. Once the orientation is known, the translational parameters can be determined.

There are numerous translation functions in the literature (Tollin, 1966; Crowther & Blow, 1967; Hendrickson & Ward, 1976; Harada, Lifchitz, Berthou & Jolles, 1981; Langs, 1985). The function of Crowther & Blow (1967) is often used. It is similar to the rotation function in that the correlation of the Patterson functions based on the observed and the model structure factors is calculated with a product function. In this case, however, the model Patterson function consists of the intermolecular vectors of molecules related by a symmetry operation.

$$T(t) = \int_V P_{ij}(u, t) P_o(u)\, \mathrm{d}u$$

where P_{ij} is the Patterson function due to symmetry-related molecules i and j of the model and P_o is the observed Patterson. The intermolecular vector between molecules i and j is given by t. The integral is taken over the cell volume V. The expression is evaluated in reciprocal space by means of a FFT algorithm.

Another commonly used procedure is to calculate the R factor ($R = \sum \| F_o| - |F_c\| / \sum |F_o|$, where $|F_o|$ and $|F_c|$ are the observed and calculated structure-factor amplitudes, respectively) as the model is translated in the cell. The computation is not as efficient as for the translation function since FFT cannot be used to compute the whole translation map, but with increasing computer speeds this is not a serious drawback. On the other hand, the procedure is sensitive to errors in the relative scale of $|F_o|$ and $|F_c|$. Such an error may result in an incorrect solution. Nevertheless, many structures have been solved by this method (Rossmann, 1980).

More recently Harada *et al.* (1981) have introduced a function that combines a product function, similar to the one defined by Crowther & Blow (1967), and an overlap function that measures the amount by which

[*] Present address: Laboratory of Physical Chemistry, University of Groningen, Nijenborgh 16, 9747 AG Groningen, The Netherlands.

[†] Present address: Laboratory of Chemical Physics, University of Groningen, Nijenborgh 16, 9747 AG Groningen, The Netherlands.

0021-8898/87/060517-05$01.50

the atoms of symmetry-related molecules of the model overlap. They began with a correlation coefficient defined as

$$C' = \sum |F_o|^2|F_c|^2[\sum |F_o|^4 \sum |F_c|^4]^{-1/2}$$

and then derived a quantity that is more easily computed. Their function is

$$T'(r) = TO(r)/O(r)$$

where

$$TO(r) = \sum |F_o|^2|F_c|^2/\sum |F_o|^4,$$
$$O(r) = \sum |F_c|^2/n \sum |F_m|^2,$$

n is the number of symmetry operations and F_m is the contribution to the structure factor of one molecule. *TO* is proportional to a product function and *O* is an overlap function. The function is maximal when the agreement between the intermolecular vectors of the observed and calculated Patterson functions is large and the overlap among the molecules is small. Harada *et al.* (1981) have shown that the function can be evaluated with the FFT.

The correlation coefficient

We were inspired by the work of Harada *et al.* (1981) to use the correlation coefficient for the solution of the translation problem. We chose to work with the standard linear correlation coefficient defined as

$$C = \sum (|F_o|^2 - \overline{|F_o|^2})(|F_c|^2 - \overline{|F_c|^2})$$
$$\times [\sum (|F_o|^2 - \overline{|F_o|^2})^2 \sum (|F_c|^2 - \overline{|F_c|^2})^2]^{-1/2}.$$

Like the *R* factor, it is basically a measure of the agreement between observed and calculated quantities. Unlike the *R* factor, it is scaling insensitive, as replacement of $|F_o|$ by $k|F_o|$ (k = arbitrary constant) gives the same value. Scaling insensitivity is very important when high-resolution data are not available and an accurate scale factor cannot be obtained. The calculation of this function is time consuming as it is not in a form suitable for the use of FFT and the correlation coefficient for each trial position must be evaluated separately.

Using Parseval's theorem (Bracewell, 1965), one can show that C' defined by Harada *et al.* (1981) is equal to the corresponding quantity for Patterson functions, *i.e.*

$$C' = \int_V P_oP_c \mathrm{d}u \left[\int_V P_o^2\,\mathrm{d}u \int_V P_c^2\,\mathrm{d}u\right]^{-1/2}$$

and, similarly, for summation over a narrow range of resolution, C can be expressed in terms of origin-removed Patterson functions,

$$C = \int_V P'_oP'_c \mathrm{d}u \left[\int_V P'^2_o\,\mathrm{d}u \int_V P'^2_c\,\mathrm{d}u\right]^{-1/2},$$

where P'_o and P'_c are origin-removed Patterson functions. If the $|F_{000}|^2$ term is not included, the mean values of the Patterson functions are zero so that both C and C' are correlation coefficients between Patterson functions. In either case, overlapping atoms will lead to an increase in both the numerator and the denominator of the function so that the denominator compensates for the increase in the numerator.

Alternatively, the correlation coefficient can be interpreted in reciprocal space as a measure of the phase error. A group in Madras has worked out probability distributions for a pair of related structures (Srinivasan & Parthasarathy, 1976). They considered the case of comparing an observed structure with a partial model with errors. The conditional probability distribution for the phase error $\alpha = \alpha_o - \alpha_c$, given the normalized structure-factor amplitude, is

$$P(\alpha; |E_o|, |E_c|) = K \exp [(2\sigma_A|E_o||E_c| \cos \alpha)/(1-\sigma_A^2)]$$

where

$$K = \{2\pi I_0[2\sigma_A|E_o||E_c|/(1-\sigma_A^2)]\}^{-1},$$

$$\sigma_A = \sigma_1 D,$$

$$\sigma_1^2 = \left(\sum_j^M f_j^2\right)\bigg/\left(\sum_j^N f_j^2\right)$$ with M and N the numbers of atoms in the model and the observed structure, respectively,

$$D = \langle \cos (2\pi \mathbf{h}.\Delta\mathbf{r}_j)\rangle$$

$\Delta\mathbf{r}_j$ = coordinate error

$I_0(X)$ = zero-order modified Bessel function (Watson, 1958).

This is a unimodal distribution with the maximum at $\alpha = 0$ and the width determined by σ_A. The distribution becomes sharper as σ_A becomes larger.

Hauptman (1982), working on a related problem, came up with an identical distribution for a pair of normalized structure-factor amplitudes. He went on to show that

$$\sigma_A^2 = \langle(|E_1|^2 - \langle|E_1|^2\rangle)(|E_2|^2 - \langle|E_2^2|\rangle)\rangle$$
$$\times [\langle(|E_1|^2 - \langle|E_1|^2\rangle)^2\rangle$$
$$\times \langle(|E_2|^2 - \langle|E_2|^2\rangle)^2\rangle]^{-1/2}$$
$$\simeq C.$$

If the summation is performed over a narrow range of resolution, then a correlation coefficient calculated for $|F|$ will be the same as that calculated for $|E|$. Under this condition, finding the position for the molecular model in the unit cell that maximizes the correlation coefficient is equivalent to minimizing the phase error. Moreover, this interpretation gives some meaning to the actual value obtained for C. The score for the correctly positioned molecule is expected to be proportional to the fraction of the crystal content the model represents (σ_1^2). In addition, C will be dependent on the factor D which in turn is dependent on the degree of homology between the model and the unknown structure. Once a convincing solution is found, the value of C (hence σ_A) gives an indication of the accuracy of the phases that can be obtained from the model.

In both of the above interpretations of the correlation coefficient is the condition that a narrow range of resolution of data is used. This condition is usually satisfied because only a restricted resolution of data is used to minimize the computation time. However, this suggests that if a wider resolution range of data were used, it might be better to work with $|E|$'s.

The calculation of correlation coefficients has been implemented in a program called *BRUTE* (for the brute-force technique of finding a solution). It moves the search model over a grid of points in the crystal cell. At each point the symmetry-related positions are generated and the structure factors are calculated. The amplitudes of these calculated structure factors are then compared with the observed values using the correlation coefficient, C, as well as the conventional R factor. The calculation of structure factors, F_c, is done rapidly by the use of molecular scattering factors (Lipson & Cochran, 1957).

Let

$$\mathbf{x}_{jk} = R_j \mathbf{x}_{0k} + \mathbf{T}_j$$

where R_j, $\mathbf{T}_j$ are the rotation matrix and the translation vector, respectively, of the symmetry operation j of the space group, and $\mathbf{x}_{0k}$ are the coordinates of the atom k of an oriented search model in an asymmetric unit (k varies from 1 to the number of atoms in the model, NATM). Then for a shift $\boldsymbol{\Delta}$ in the coordinates,

$$F(\mathbf{h}) = \sum_{j}^{\mathrm{NSYM}} G_j(\mathbf{h}) \exp 2\pi i(\mathbf{h}.R_j\boldsymbol{\Delta}),$$

where NSYM is the number of symmetry operations,

$$G_j(\mathbf{h}) = \sum_{k}^{\mathrm{NATM}} f_k \exp 2\pi i(\mathbf{h}.\mathbf{x}_{jk})$$

$$= \text{molecular scattering factor}$$

and

$$f_k = \text{atomic scattering factor.}$$

The molecular scattering factors, G_j, are calculated once for the first grid point and stored for use at subsequent grid points. Therefore the overall calculation time becomes essentially independent of the number of atoms in the model and is a function of the number of symmetry operations, the number of reflections and the number of grid points.

In addition to the basic computations described above, the program incorporates some very useful features. First, it allows for the inclusion of a set of atoms with fixed positions. Their contributions to the structure factors are added to the part due to the moving set of atoms. This is useful when the orientation and the position of a part of a molecule are both known. Secondly, the program can make adjustments in the orientation of the search model. A rotational search can be made by rotating the model about each axis of an orthogonal set at regular intervals. The resulting rotations will sample the space evenly for small changes in the angles. When combined with a translational search a six-dimensional search is possible. However, for each new orientation, a whole set of molecular scattering factors G_j must be recalculated so that the computation time becomes prohibitively long if the 6D search includes more than a few orientations near the rotation function peak. Rabinovich & Shakked (1984) consider packing of the molecule to reduce the number of trial points in a multi-dimensional search using the R factor. We have found that rotational parameters can be refined before any translational searches are done, by specifying only the $P1$ symmetry and adjusting the angular parameters for maximum correlation. Such a calculation is in fact not very different from the rotation function.

The program has been written for the Floating Point Systems 164 Attached Processor (FPS164) which is driven at the University of Alberta by the host computer, an Amdahl 5870. The FPS164 is a parallel-pipeline machine which is capable of fast operations on long arrays. *BRUTE* was written to maximize vector usage and takes advantage of the assembler subroutines supplied with the system for performing the vector and matrix operations. A version of the program written totally in standard Fortran also exists.

The calculation time for the translational search for pepsinogen (James & Sielecki, 1986), space group $C2$, was 275 s on the FPS164 for 53 × 45 points with 1565 reflections. The same search running totally on the Amdahl 5870 took 726 s.

Results and discussion

The performance of the correlation-coefficient search in *BRUTE* has been compared with those of an R-factor search and the translation functions of Crowther & Blow (1967) and Harada *et al.* (1981). The structures used in the comparison were all solved by *BRUTE*. These are two serine proteases, tonin (Fujinaga & James, 1987) and *Streptomyces griseus* trypsin (SGT) (Read, 1986, Read, Brayer, Jurasek & James, 1984), and the aspartyl protease zymogen, pepsinogen (James & Sielecki, 1986). A summary of these structures and the models used in the molecular

Table 1. *Structures used for the tests*

	Space group and cell parameters (Å, °)	Search model	Sequence homology*	Structural homology†
Tonin (235a.a.)‡	$P4_32_12$ $a = 48·64$ $b = 48·64$ $c = 201·2$	Trypsin (223a.a.)	40%	0·89 (191a.a.)
SGT (223a.a.)	$C222_1$ $a = 72·04$ $b = 50·86$ $c = 120·4$	Trypsin (223a.a.)	33%	0·86 (168a.a.)
Pepsinogen (370a.a.)	$C2$ $a = 105·8$ $b = 43·40$ $c = 88·60$ $\beta = 91·4$	Penicillopepsin (323a.a.)	35%	1·63 (275a.a.)

* Percentage of identical residues.
† R.m.s. deviations (Å) of equivalent α-carbon atoms after superposition by algorithm of Rossmann & Argos (1975) with probability cut-off of 0·005. Number in brackets is the number of residues considered equivalent.
‡ a.a. = amino acids.

replacement is given in Table 1. In each case the search model has relatively low sequence homology with the unknown structure. The results of the translation searches are shown in Table 2 where the highest and second-highest points (or lowest in the case of *R* factor) are given as the number of standard deviations above (or below) the mean. The ratio of these two peaks is given as the signal-to-noise ratio (*S*/*N*).

It is difficult to compare the results of Crowther & Blow's translation function, in which three different planes are calculated, with those of the other methods. Each plane gives two of the three coordinates and there is considerable information in the consistency among the solutions obtained from all the planes. Usually the first consistent solution ranks far above the second. For pepsinogen, only one plane is computed since the space group is polar, so comparison for this case is straightforward. For SGT, two different orientations are used. Orientation 1 is that obtained from the rotation function using 2·8 Å resolution data and is 6° away from the correct orientation. Orientation 2 is from the rotation function that used 3·5 Å resolution data and is 3° away from the correct orientation. In all cases except for SGT orientation 1, all the translation functions give the correct answer. The small number of trials presented here does not allow one to say conclusively that one translation function performs better than the rest. However, it would seem that the *R*-factor search performs most poorly.

For SGT orientation 1, all the functions tested give an incorrect solution but the low signal-to-noise ratio in each of the cases makes the solution untrustworthy. The result for this case can be looked at from two different points of view. The first is that current translation functions do not always work, or that they should be more robust in the presence of model orientation errors. The second point of view is that the

Table 2. *Comparison of various translation functions*

Data between 4 and 8 Å resolution were used for all the runs. The peak heights are expressed as the number of standard deviations above (or below for *R* factor) the mean. The signal-to-noise ratio (*S*/*N*) is calculated as the ratio of the highest (lowest) and the second-highest(lowest) peaks.

	Crowther & Blow (1967)			*R* factor	Harada *et al.* (1981)	*BRUTE* (correlation coefficient)
Tonin						
1st	4·6	5·1	4·0	3·6	6·3	10·4
2nd	3·8	3·4	3·5	2·1	3·8	6·0
S/*N*	1·2	1·5	1·1	1·7	1·7	1·7
Pepsinogen						
1st		5·2		3·8	5·3	7·2
2nd		3·9		2·4	2·5	3·5
S/*N*		1·3		1·6	2·1	2·1
SGT						
Orientation 1†						
1st	3·8*	3·1*	1·5*	1·5*	2·9*	4·3*
2nd	3·7	3·1	1·5	1·5	2·5	4·0
S/*N*	1·0	1·0	1·0	1·0	1·2	1·1
Orientation 2‡						
1st	4·2*	3·5	4·6	1·7	3·6	5·7
2nd	3·8	3·3	3·6	1·6	3·2	4·5
S/*N*	1·1	1·1	1·3	1·1	1·2	1·3

* Incorrect solution
† From rotation function using 2·8 Å resolution data (6° error).
‡ From rotation function using 3·5 Å resolution data (3° error).

rotation function does not always work, or that it can be too imprecise and inaccurate to allow the translation function to work. This is contrary to the popular belief (*e.g.* Harada *et al.*, 1981) that the rotation function is reliable whereas the translation functions are not.

A further conclusion can be drawn from the SGT example. It is possible to distinguish the correct orientation using the results of the translation functions. In a set of translation searches using models with different orientations, the best orientation is the one that gives the most unambiguous solution (*i.e.* highest S/N). The value of the correlation coefficient is also generally a good indication of a correct orientation. For orientation 2 of SGT the highest correlation was 0·22, whereas for orientation 1 it was only 0·13. However, it has been observed in one test case (unpublished results) that the orientation that resulted in the highest correlation coefficient gave an incorrect solution. In this case, the translation map with the highest S/N did give the correct solution but with a slightly lower value of the correlation.

The main advantage of *BRUTE* over the other programs is its ability to adjust the orientation. This allows structures to be solved even in the presence of error in the results of the rotation function. In addition, the ability of the program to accept contributions from partial structures whose positions are fixed is useful for cases where there are more than one subunit or domain (Cygler, Boodhoo, Lee & Anderson, 1987). The insensitivity to the scale of the data allows structure solution even when high-resolution data are not available and an accurate scale cannot be determined. Finally, because the correlation coefficient gives an absolute measure of the phase error, its value can be used to assess not only the correctness of the solution but also the quality of the model in providing phase information.

We acknowledge the support and the encouragement of this work by M. N. G. James. We also thank Anita Sielecki and Paula Fitzgerald for running the tests. This research was funded by the Medical Research Council of Canada grant to the MRC Group in Protein Structure and Function at the University of Alberta. MF and RJR acknowledge the Medical Research Council of Canada and the Alberta Heritage Foundation for Medical Research for financial support.

References

BLUNDELL, T. L. & JOHNSON, L. N. (1976). *Protein Crystallography.* London: Academic Press.

BRACEWELL, R. (1965). *The Fourier Transform and Its Applications.* New York: McGraw-Hill.

COOLEY, J. W. & TUKEY, J. W. (1965). *Math. Comput.* **19**, 297–301.

CROWTHER, R. A. (1972). In *The Molecular Replacement Method*, edited by M. G. ROSSMANN, pp. 173–178. *International Science Review*, Vol. 13. New York: Gordon & Breach.

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

CYGLER, M., BOODHOO, A., LEE, J. S. & ANDERSON, W. F. (1987). *J. Biol. Chem.* **262**, 643–648.

FUJINAGA, M. & JAMES, M. N. G. (1987). *J. Mol. Biol.* **195**, 373–396.

HARADA, Y., LIFCHITZ, A., BERTHOU, J. & JOLLES, P. (1981). *Acta Cryst.* A**37**, 398–406.

HAUPTMAN, H. (1982). *Acta Cryst.* A**38**, 289–294.

HENDRICKSON, W. A. & WARD, K. B. (1976). *Acta Cryst.* A**32**, 778–780.

JAMES, M. N. G. & SIELECKI, A. R. (1986). *Nature (London)*, **319**, 33–38.

LANGS, D. A. (1985). *Acta Cryst.* A**41**, 578–582.

LIPSON, H. & COCHRAN, W. (1957). *The Determination of Crystal Structures*, p. 235. London: Bell.

RABINOVICH, D. & SHAKKED, Z. (1984). *Acta Cryst.* A**40**, 195–200.

READ, R. J. (1986). PhD Thesis. Univ. of Alberta, Canada.

READ, R. J., BRAYER, G. D., JURASEK, L. & JAMES, M. N. G. (1984). *Biochemistry*, **23**, 6570–6575.

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

ROSSMANN, M. G. (1980). In *Theory and Practice of Direct Methods in Crystallography*, edited by M. F. C. LADD & R. A. PALMER, pp. 361–417. New York: Plenum Press.

ROSSMANN, M. G. & ARGOS, P. (1975). *J. Biol. Chem.* **250**, 7525–7532.

ROSSMANN, M. G. & BLOW, D. M. (1962). *Acta Cryst.* **15**, 24–31.

SRINIVASAN, R. & PARTHASARATHY, S. (1976). *Some Statistical Applications in X-ray Crystallography.* Oxford: Pergamon Press.

TOLLIN, P. (1966). *Acta Cryst.* **21**, 613–614.

WATSON, G. N. (1958). *A Treatise on the Theory of Bessel Functions.* Cambridge Univ. Press.