COMPUTER PROGRAMS

J. Appl. Cryst. (1994). **27**, 122–124

Programs for translation searches with two independent models simultaneously.

By A. G. URZHUMTSEV* and A. D. PODJARNY, 15, *Rue Descartes, UPR de Biologie Structurale, Institut de Biologie Moleculaire et Cellulaire du CNRS,* 67084 *Strasbourg, France*

(*Received* 2 *February* 1993; *accepted* 9 *June* 1993)

Abstract

Programs to carry out translation searches for the molecular replacement method using *two* independent models simultaneously have been developed. Both correlation-coefficient distributions and packing functions can be calculated.

I. Introduction

The translation problem in the molecular-replacement method is traditionally a difficult one. The usual tools for its solution are translation functions, packing analyses and amplitude correlation-coefficient searches [a review of this problem and list of references can be found in Fitzgerald (1991)]. For each of them, a function of the position of the search model in the unit cell is calculated. However, for the case of several (for example, three) independent molecules in the asymmetric unit, these functions depend on the proper positioning of *all* corresponding models. Therefore, standard one-model functions can easily fail to solve the problem.

II. Description of the programs

1. *Correlation-coefficient search function – CORR2M*

The program calculates the distribution of the correlation-coefficient values. These values are similar to those reported previously by Harada, Lifchitz, Berthou & Jolles (1981) and Fujinaga & Read (1987).

$$C(\mathbf{t}_1, \mathbf{t}_2) = [\sum (F_{\rm calc} - \langle F_{\rm calc}\rangle)(F_{\rm obs} - \langle F_{\rm obs}\rangle)] \times \{[\sum (F_{\rm calc} - \langle F_{\rm calc}\rangle)^2] \times [\sum (F_{\rm obs} - \langle F_{\rm obs}\rangle)^2]\}^{-1/2} \quad (1)$$

is a function of the positions $\mathbf{t}_1$ and $\mathbf{t}_2$ of *two* independent models. Here, the $F_{\rm obs}(\mathbf{s})$ are observed structure-factor moduli, the $F_{\rm calc} = F_{\rm calc}(\mathbf{s}; \mathbf{t}_1, \mathbf{t}_2)$ are structure-factor moduli calculated for the models at positions $\mathbf{t}_1$ and $\mathbf{t}_2$, the summation is carried out over the given reflection set $S = \{\mathbf{s}\}$ and $\langle F\rangle$ is the mean value of F over the same data set.

The program uses the structure factors calculated for the initial positions of the models but not the atomic models themselves. Input structure factors $\{\mathbf{F}_1(\mathbf{s})\}$ and $\{\mathbf{F}_2(\mathbf{s})\}$ should be calculated for each of the two models placed in the unit cell without any symmetry, *i.e.* in $P1$. These model positions are the initial ones to which the model shifts, $\mathbf{t}_1$ and $\mathbf{t}_2$, are applied. Thus, the calculated structure factors are obtained from $\mathbf{F}_1(\mathbf{s})$ and $\mathbf{F}_2(\mathbf{s})$, the shifts $\mathbf{t}_1$ and $\mathbf{t}_2$, the symmetry matrix G_m and the symmetry translation vector $\mathbf{u}_m$ (for each of the crystallographic or noncrystallographic symmetry operations) by

$$\mathbf{F}_{\rm calc}(\mathbf{s}; \mathbf{t}_1, \mathbf{t}_2) = \sum_{k=1,2} \sum_{m=1,M} \mathbf{F}_k(G_m^*\mathbf{s}) \exp [2\pi i(\mathbf{s}, G_m\mathbf{t}_k + \mathbf{u}_m)]. \quad (2)$$

2. *Packing function – PACK2M*

The program calculates packing functions of different types, using either the percentage of empty volume or the percentage of overlapping volume.

The function is calculated following the method of Hendrickson & Ward (1976). For every molecule, a 1/0 mask is calculated in a given grid. These masks are added, giving the number of molecules contributing to each grid point. The resulting grid is analyzed according to the chosen packing function, for example, the percentage of points where two or more molecules contribute is the relative volume of overlap. This procedure is repeated for each translation. Even for a coarse grid (*e.g.* 5 Å), this calculation estimates reasonably the allowed and forbidden regions.

3. *Cross-correlation function – CROS2M*

The program calculates the correlation

$$Q(\mathbf{t}_1, \mathbf{t}_2) = [\sum (C_1 - \langle C_1\rangle)(C_2 - \langle C_2\rangle)]\{[\sum (C_1 - \langle C_1\rangle)^2] \times [\sum (C_2 - \langle C_2\rangle)^2]\}^{-1/2} \quad (3)$$

between two distributions, $C_1(\mathbf{t}_1, \mathbf{t}_2)$ and $C_2(\mathbf{t}_1, \mathbf{t}_2)$ (the results of the programs described above) and performs a statistical analysis of the result. The program may also be applied to a single distribution. In this case, the program does not calculate (3) but only performs the statistical analysis of the input distribution $C_1(\mathbf{t}_1, \mathbf{t}_2)$, assuming the second distribution to be uniform.

4. *Parameters of the programs*

All three programs are written in the same manner and have the same output and similar control data.

* Permanent address: Institute of Mathematical Problems of Biology, Russian Academy of Sciences, Pushchino, Moscow region 142292, Russia.

Common parameters of the functions. For each of the two models, the translation grid searched by *CORR2M* and *PACK2M* is defined by two points: the dimension of the search and the step size. If the search is one-dimensional, these points define the line of the search and its initial and final points. If the search is multidimensional, these points define diametrically opposite vertices of a parallelipiped with sides parallel to the coordinate planes of the unit cell. For the latter case, the grid planes are parallel to the coordinate ones. Any combination of one- and multidimensional searches is possible for the two models.

It should be mentioned that this grid has no relation to the grid used for packing calculations, which is defined independently in the control data.

The program *CROS2M* needs no parameters to calculate a cross correlation or to carry out a statistical analysis of the combined distribution. But in order to calculate the model positions corresponding to the cross-correlation peaks, the program asks for the translation-search parameters.

The symmetry operations of the space group are introduced for *CORR2M* and *PACK2M* throught the control data in a convenient symbolic form. This provides the user with the possibility of easy inclusion of non-crystallographic symmetry if necessary.

Specific parameters of the functions. The resolution of the data used for the correlation-coefficient search by *CORR2M* is an essential parameter. The functions may, therefore, be calculated in different resolution shells and the results jointly analyzed and compared.

To define the packing function by *PACK2M*, the type of packing function and the composition of the model have to be chosen. The packing function is defined by choosing the level of overlap and the criterion of the comparison ('less than or equal to', 'equal to' or 'greater than or equal to' this level). The grid for molecular-mask analysis is defined by its step and limits. The model is taken either as a whole or with the inclusion or exclusion of some types of atoms, which can be done by pointing to the corresponding labels.

5. *Program output*

The main result of each of these programs is a two- to six-dimensional distribution and the main output file contains this distribution. To facilitate its mapping with standard plotting programs, the file is prepared in a convenient format of the well known program *GHC650* (Cohen, 1977) and can easily be transferred to any other format.

A simple statistical analysis of the calculated distribution is carried out by the programs themselves. They calculate the main statistics like the mean value, the standard deviation and the limits of the distribution (in absolute values and standard deviations). The programs also calculate the histogram and determine the highest points of the distribution.

A fast symbolic map presentation of the distribution can be obtained directly using the programs.

6. *Computational resources and practical aspects*

The programs keep all necessary data in memory in a single buffer array. The dimension of this array can be varied; it must be large enough to have the full map of one calculated distribution and, simultaneously, the set of structure factors for each of the symmetry-related copies of the models (for the correlation-coefficient function) or the complete list of atoms and the mask grid (for packing-function calculation).

Table 1. *CPU time required for PACK2M (in minutes on the DEC Station AXP* 3000-500) *for* (3 + 3)- *and* (3 + 1)- *dimensional searches with* N_{step} *points per dimension*

The calculations were done for the test structures placed in a unit cell of 200 × 200 × 200 Å. The last string corresponds to the test done in a unit cell of 100 × 100 × 100 Å. The grid step for the molecular mask was 5 Å. Cases with four and eight symmetry operations were considered; in order to measure only the effect of the additional symmetry, one quarter of the unit cell was used for the mask calculation and analysis in both cases.

Number of search points (models 1, 2) in the form			Number of atoms per molecule (four symmetry operations)			Number of atoms per molecule (eight symmetry operations)		
N_{step1}^{dim1}	×	N_{step2}^{dim2}	500	1000	2000	500	1000	2000
5^3	×	5^3	2.7	3.1	3.9	4.3	5.2	6.6
10^3	×	5^3	21.3	23.9	30.2	34.5	42.4	52.8
10^3	×	10^3	175.1	190.3	244.9	297.5	323.2	418.7
10^3	×	10^1	1.7	2.0	2.5	2.8	3.2	4.2
10^3	×	10^1	45.0	67.5	113.5	80.0	128.7	221.0

Table 2. *CPU time required for CORR2M (in minutes on the DEC Station AXP* 3000-500) *for* (3 + 3)- *and* (3 + 1)- *dimensional searches with* N_{step} *points per dimension*

Number of search points (models 1, 2) in the form			Number of reflections (four symmetry operations)			Number of reflections (eight symmetry operations)		
N_{step1}^{dim1}	×	N_{step2}^{dim2}	500	1000	2000	500	1000	2000
5^3	×	5^3	1.4	2.2	4.0	1.9	3.4	6.5
10^3	×	5^3	4.8	9.5	20.1	7.5	16.5	32.8
10^3	×	10^3	33.5	69.4	155.2	55.4	114.8	265.8
10^3	×	10^1	1.1	1.4	2.5	1.4	2.0	3.3

The computation time T_C of *CORR2M* can be roughly estimated as

$$T_C = CN_{obs}N_{sym}N_{posit}, \tag{4}$$

where N_{obs} is the number of obseved amplitudes included in the search, N_{sym} is the number of symmetry operations, N_{posit} is the number of checked pairs of model positions and C is a constant. A similar estimation of the computation time T_P of *PACK2M* gives

$$T_P = (C_1N_{atom} + C_2N_{grid})N_{sym}N_{posit}, \tag{5}$$

where N_{atom} is the number of atoms included in the search model, N_{grid} is the size of the grid used to calculate a model mask and C_1 and C_2 are constants.

Test calculations (Tables 1 and 2) show that a complete search in a large unit cell for two macromolecules in general positions is feasible but needs significant CPU time. Therefore, any additional information that limits the region of the search is useful. In particular, if one of the molecules is in a special position, *e.g.* on a symmetry axis, a one-dimensional search is enough for it. The CPU time can also be reduced by a proper choice of the data and the strategy of the search. For example, the correlation search can be started at very low resolution, where it can be done with a large step, and then scanned with a finer step around the

possible solutions found. In addition, this search can be done with the strongest reflections only. In order to decrease CPU time for the packing search, the main-chain atoms alone may be used. Because the packing function is quite smooth, it can be calculated in a very coarse grid (5 Å or even larger). All this means that, even in a general case, the search can be done in an acceptable time.

The programs are written in Fortran and have no computer-specific features. Source and executable modules are available from the authors on request.

III. General notes

The presence of several independent molecules in the asymmetric unit is common and can be a source of problems for structure solution. The programs described here provide a tool for the solution of the translation problem for such cases.

The programs for the two-model search were developed before the fast algorithm for the calculation of the packing function and correlation-coefficient distributions was published (Stubbs & Huber, 1991) and, unfortunately, do not use it. However, available up-to-date hardware allows the use of the programs even for very large structures.

The accuracy of the result depends on the search step but also on rotation-function errors, differences between the search and real models *etc.* Therefore, decreasing the search step does not necessarily improve the solution. Refinement programs with a large convergence of radius for a rigid-body/rigid-groups model, like those of Brünger, Kuriyan & Karplus (1987), Urzhumtsev, Lunin & Vernoslova (1989) and Castellano, Oliva & Navaza (1992), can follow these search programs in order to improve the orientation and position of each model, which also allows the checking of several possible solutions to find the right one.

The authors thank Drs D. Moras, P. Dumas and J.-C. Thierry for discussions and their continuous interest in the work. They are also grateful to Dr B. Rees for discussions during the work and his careful reading and improvement of the manuscript. This work was supported by the CNRS through funding of the Unite Propre de Recherche (UPR) de Biologie Structurale. AGU was an EMBO fellow.

References

Brünger, A. T., Kuriyan, J. & Karplus, M. (1987). *Science*, **235**, 458–460.
Castellano, E. E., Oliva, G. & Navaza, J. (1992). *J. Appl. Cryst.* **25**, 281–284.
Cohen, G. (1977). Personal communication.
Fitzgerald, P. M. D. (1991). *Crystallographic Computing* 5, edited by D. Moras, A. D. Podjarny & J.-C. Thierry, pp. 333–347. Oxford Univ. Press.
Fujinaga, M. & Read, R. J. (1987) *J. Appl. Cryst.* **20**, 517–521.
Harada, Y., Lifchitz, A., Berthou, J. & Jolles, P. (1981). *Acta Cryst* A**37**, 398–406.
Hendrickson, W. A. & Ward, K. B. (1976). *Acta Cryst* A**32**, 778–780.
Stubbs, M. T. & Huber, R. (1991). *Acta Cryst* A**47**, 521–526.
Urzhumtsev, A. G., Lunin, V. Yu. & Vernoslova, E. A. (1989). *J. Appl. Cryst.* **22**, 500–506.

J. Appl. Cryst. (1994). **27**, 124–127

DIMS – a direct-method program for incommensurate modulated structures.*

By Fu Zheng-qing and Fan Hai-fu,† *Institute of Physics, Chinese Academy of Sciences, Beijing* 100080, *People's Republic of China*

(*Received* 26 *April* 1993; *accepted* 30 *June* 1993)

Abstract

A direct-method program, *DIMS* (direct methods for incommensurate modulated structures), has been written to solve the phase problem of incommensurate structures with one-dimensional modulation. The program uses conventional structure factors instead of normalized structure factors in the phase derivation. It derives phases for satellite reflections by making use of the known phases of main reflections, which can be calculated from the known basic structure or high-resolution electron-microscope images of the sample, or derived by a conventional structure-analysis method involving only main reflections. Two types of phase relationship are used in the program. The first type consists of one main reflection and two satellites of the same order. This type of relationship is used to relate phases of satellites belonging to the same order with phases of main reflections. The second type of relationship consists of three satellites belonging to at least two different orders. This type of relationship is used to link satellites of different orders. X-ray as well as electron diffraction data from a dozen incommensurate modulated structures have been used to test the program. Two typical examples are described.

Introduction

Incommensurate modulated structures are an important category of crystal structures. They do not have three-dimensional periodicity but can be regarded as a three-dimensional hypersection of a higher-dimensional periodic structure (de Wolff, 1974; Janner & Janssen, 1977). A number of multidimensional least-squares programs, *REMOS*,

* Supported in part by the National Natural Science Foundation of China.

† To whom correspondence should be addressed.